Realising the Potential of Rich Energy Datasets

by

Robert Ellis

A thesis submitted to

The University of Birmingham

for the degree of

Doctor of Philosophy

School of Engineering

College of Engineering and Physical Sciences

University of Birmingham 2016

Abstract

In the last twenty years the availability of vast amounts of data has enabled many industries to gain significant insight into numerous aspects of their operation whose trends and relationships with each other were previously unknown. The result is an unprecedented ability to predict operational needs, to evaluate performance of individuals or assets and prepare such industries for future uncertainties. The rail industry currently produces and has access to large amounts of data that are, in many cases, not used to their full potential. This thesis investigates the additional value that can be gained from such data. Additional value can be seen as the identification of previously hidden or unknown relationships between various inputs within a dataset and their use to provide benefits that are beyond the original use the data was intended for. This is termed 'Rich data'. Case studies that are discussed in this thesis show how energy and financial benefits can be gained for the rail industry by using existing data that were originally recorded for a relatively narrow scope.

The first case study demonstrates a novel method to identify and cluster distinct driver styles in use on a DC rail network. Using the optimal driver styles identified, improved 'driver cultures' were designed that are shown to provide up to 10% energy savings without the need for expensive in cab driver advisory systems.

The second case study details data taken from a full fleet that were used to develop a statistical method to identify the minimum amount of vehicles that required energy metering whilst still providing an accurate mean energy consumption estimate. The identification of this minimum amount was then used to validate the fleet size intended for partial fleet metering options for UK rail networks. This can offer significant meter installation cost reduction as well as provide an improved energy billing model to rail networks.

Acknowledgments

I would like to thank my supervisors Dr Stuart Hillmansen and Dr Pietro Tricoli for their support and input during my PhD. Pietro many thanks for your help early on, giving me a crash course in electrical engineering and helping to get my model and simulation making sense. Stuart thank you for all of the great suggestions, ideas and advice throughout, related to research, academia and raising a family!

I would like to thank Merseyrail and Ian Jones for the data that made my research possible, their assistance and many discussions about my findings and how they can be used. I would also like to thank ATOC and Neil Ovenden for the opportunity to conduct the PFM study.

Thank you Dr Paul Weston for your guidance, patience and help with my research. You have been there to tell me when I am doing well and also to point out when I am totally wrong! Thank you for letting me feel comfortable saying “I don’t understand” and always having the time to teach me so that the next time I do.

Thank you Dr Gemma Nicholson for your help and support during my PhD. You have a great ability to spot things that are good quality work, and things that are not. I always try to keep in my head “Does this count as good in Gemma’s books”.

I would like to thank Professors Clive Roberts and Felix Schmid. Clive thank you for giving a geologist with and environmental science masters the chance to be an Engineer with Rail group! Anything is possible! Thank you as well for the many opportunities working with the group has given me and hopefully will continue to do so in the future.

Felix thank you for the many opportunities I have had to work and help out in the wider rail group community. My involvement with various MSc classes, young student events, dinners, visitor tours and all the other activities that I have enjoyed but can’t recall has made me feel like a real valued member of the group.

A great big thanks to all of my colleagues and friends in BCRRE that have made my PhD rewarding and always interesting. A special mention for Mary Winkles, thank you so much for your support throughout my PhD, you have been a tremendous help.

Finally I would like to thank my family. Mum and Dad thank you for your support during my Masters and PhD study, I could not have done it without you. Thank you for your continued encouragement it is always welcome.

Thank you to Kayo for always being there for me　本当にありがとうございます。

最後に希春君、あなたは 将来いつもがんばって。ママとパパ　希春君大好きだよ！

Contents

List of Figures

List of Tables

1 Introduction

The rail industry could benefit substantially from the wide scale use of Big and Rich data analytics but, for a number of reasons there has not been the same uptake that has been seen in other transport and business sectors [1]. This is not to say that the rail industry is ignorant of these benefits, on the contrary there is a substantial amount of literature identifying areas where data analytics would be effective tools, these broadly cover energy monitoring, asset management, condition monitoring, passenger flow and customer behaviour [1, 2]. In most of these areas there currently exists a vast quantity of data that are collected and analysed which have also been done for as long as railways have existed [3]. The issue with data in the railways is that is has only recently begun to be optimised and used holistically; for example track circuit maintenance data and weather data are used for track circuit maintenance and weather forecasting only, rather than identifying possible correlation between heavy rain and problematic track circuits. The key idea of Big and Rich data analytics are that previously unknown trends or patterns can be identified using combinations of data types from many different sources. Utilising data in this way is still in its nascent stage in the rail industry [3], which is likely a result of high initial implementation costs or the lack of suitable data, understanding and processing tools. This is particularly true for smaller rail networks and networks which are well established but resistant to significant change.

1.1 Statement of the problem

This thesis aims to show, with case studies using real data, the potential benefits Rich energy data analytics can provide to the rail industry, which in turn can be used to develop a set of recommendations to help the relevant actors take steps to solve these problems.

The rail industry records and has access to a large amount of data, often this data is used solely for the purpose or project it was recorded for or in some cases not at all. There exists a gap in the use of multimodal datasets where massive amounts of information could provide greater insight into the characteristics of the dynamic railway environment, to improve operations, reduce energy consumption and increase safety, which at present is only in the early stages in the rail industry.

An example: Vehicle movement and control data are recorded on rail vehicles in much the same way as a black box data recorder, in the On Train Monitoring Recorder (OTMR) system. This is often only short term data as its main purpose is for enquiry during the event of an incident. Energy data is also recorded on many vehicles. However, there are very few examples that link vehicle movement and control to energy consumption, and as a result there is a significant missed opportunity to analyse, for example, driver performance effects on energy consumption. In an industry where condition monitoring is one of the most important aspects to maintaining the safety of rolling stock, infrastructure and assets it is not unreasonable to expect that the human part of the system should also be conditioned monitored.

1.2 Purpose of the research

The purpose of this research is to demonstrate the value that data analysis can provide for the rail industry. Specifically this refers to large datasets, in this case called Rich data, but also known as Big data. As the data being examined here is static, i.e. it will not change, and is historic it differs slightly from the common definitions of ‘Big data’ however, no universally accepted definition for this exists. In the event that methodologies designed in this thesis were adopted by the rail industry it is likely data could be used in real time to short term, and significantly larger in size.

A second consideration is that the purpose of this research is to demonstrate the additional value that existing data can provide. This is perhaps the most important point of the two. There is intrinsic value in both large datasets that are analysed in this research, and indeed in the vast majority of data that is recorded in the railway industry. This is a result of the data being purpose recorded, i.e. it was needed and so was taken. As a result the value of this type of data is created by the analysis that is made to it. Often hidden or otherwise unknown relationships between inputs can be discovered that were not intended when the data was taken. Applying this idea to this thesis means that analysis aims to find additional value in the datasets beyond their original use. By demonstrating this added value it is thought that a good case for large scale cross-environment data recording and analysis in the rail industry can be presented which would be a small step to the introduction of a smart, resilient and sustainable transport system that can impact in the sustainability, whole-system reliability and simplicity challenges set out by the Rail Technical Strategy [4].

1.3 Hypotheses

There is significant additional value from Rich datasets beyond their original purpose that can be used to make inferences that were not originally intended to be made when the recording systems were installed. These inferences could lead to energy savings and financial benefits. This general hypothesis is tested by two sub hypotheses:

- Is it possible to identify driver characteristics, behaviour or possibly individual drivers from a Rich energy dataset? And could such characteristics be used to provide energy savings?

- Is it possible to use historic energy data from a rail network to validate the use of a partially metered fleet, where only a portion of vehicles have electricity metering, to provide accurate electricity bill estimations? And what would the minimum size of such a fleet be and what level of statistical accuracy would estimate bills have?

1.4 Research design

The thesis contains four main sections which investigate the research questions stated in the hypothesis and are described below:

Chapter 2 – Literature Review

A literature review was carried out to:

- Describe the concepts of Rich and Big data, data analytics and present an overview of their use in a number of fields.
- Investigate historic, current and future roles of data analytics in the rail industry.

Chapter 3 – Merseyrail Energy Monitoring Project (MEMP)

The MEMP was a long term vehicle and substation monitoring experiment that was conducted by the University of Birmingham from 2010 to 2013. This chapter is the first of three that discuss the analysis of data accumulated during this period. Chapter 3 introduces the Merseyrail network and gives a background to the project. A simulator designed to model two routes on this network is presented with a number of performance metrics that can be used to identify driver behaviour. Section 3.3 shows preliminary findings from the analysis of experiment data; these findings were published in a journal paper [5] in 2015 and are summarised below

- Journey time and energy consumption are inversely proportional in simulated data.

- The same general trend is observable in experiment data but examples of low journey time low energy, and high journey time high energy are present.
- These demonstrate examples of efficient and inefficient driving.
- Driver performance can be classified if journey time and energy consumption are matched to control input.

Chapter 4 – Classifying simulated driver styles

This chapter describes the development of a methodology to identify and categorise driver styles using synthetic data generated from a modified version of the simulator designed in chapter 3. This is followed by descriptions of data clustering and vector decomposition steps known as Principal Component Analysis PCA.

Chapter 5 – Identification of a 'Driver Culture' and its effect on energy consumption on a DC rail network

This chapter shows the application of the analysis methodology designed in chapter 4 to experiment data taken from four Merseyrail routes. Driver styles for each route are identified and described. This is followed by an exploration of potential energy savings that could be made in the event that control reflected optimal and energy saving driver styles identified from the real data. Findings in this chapter were published in a conference paper [6], which was presented at the STech 2015 conference in Chiba, Japan. An extended version of this paper is currently in review stages for the International Journal of Rail Technology. The latter part of this chapter evaluates the analysis technique, discusses limitations and the possibility of future implementation.

Chapter 6 – ATOC Partial Fleet Metering

This is the second of two case studies where real data used for electricity billing is analysed to develop a statistical method to identify required fleet numbers for partial fleet metering (PFM). This case study was conducted with the association of train operating companies (ATOC) energy steering group and was designed to validate the use of PFM on non-metered rail networks. Results from this investigation were reported to ATOC in the autumn of 2015 and have since been used in electricity billing policy recommendations. Also a journal paper [7] is currently in preparation stages discussing the findings of this chapter.

Chapter 7 – Successive dataset validation

This chapter demonstrates hypothesis testing using ATOC data to provide:

- An historic dataset
- A selection of synthetic datasets

Tests are performed to show the similarity of possible future datasets to the historical dataset with statistical significance. This allows a future partially metered fleet to be compared to existing data to ensure, with a degree of confidence, that energy estimations are significantly similar over a number of years. In the event of changes to vehicle operational characteristics, where efficiency is improved for example, the test will show that a partially metered fleet is now significantly different meaning that historic data is no longer comparable, which would require a repeat of the analysis described in chapter 6 to establish a new historic dataset.

Chapter 8 – Conclusion

A review of the findings of both case studies is given and addresses whether results answer the research questions set in the hypothesis section. An evaluation of the analysis techniques and their limitations are presented. Finally a general conclusion is given to identify the key achievements of this research.

1.5 Original contribution

This section give an overview on the original contribution to research this thesis provides. There are five areas where original contributions were made a brief description of each is given and reference to the relevant chapter in the thesis.

1.5.1 Single Train Simulator Class 507/8 motor model

The Single Train Simulator (STS), described in detail in 3.2.1, is a piece of software developed by the University of Birmingham Centre for Rail Research and Education. The software can be used to accurately simulate single train journeys and output various, traction, power, speed and energy performance parameters. Its design is extremely customisable allowing for versatile use and options to design many different simulation scenarios. The software was developed prior to the start of this Thesis by Dr Stuart Hillmansen. The basic simulator calculates vehicle performance based on static input variables. This thesis describes the design and implementation of a dynamic motor model which approximates the function of the camshaft controlled rheostatic traction system that the Class 507/8 EMU operates with (described in detail in 3.2.2). This motor model provides both realistic vehicle performance and driver control that reflects observed in real data, to the authors knowledge there are no other simulation models of the class 507/8 EMU and only one other recent examples of research on camshaft controlled vehicles which is briefly described in 2.3. Secondly this is the first example of realistic traction and control models being implemented in the STS. The knowledge gained by during the development of this piece of software has allowed the author to develop additional motor, engine and traction package models that have been used in a variety of other research beyond the work in this PhD. This original contribution has expanded the capability of the STS making it a more versatile simulation tool that will continue to be used by BCRRE in future research.

1.5.2 Identification of driver style characteristics from real data

Data taken in the Merseyrail Energy Monitoring Project (MEMP), described in detail in chapter 3, is high resolution allowing for very detailed analysis of various vehicle parameters while in operation; driver command, vehicle speed and energy consumption can be seen in precise detail. Most data recording systems installed on rail vehicles record at once per minute or once every five minutes meaning that there is a significant amount of information lost between each recording. This thesis presents the first analysis and characterisation of driver styles taken from high resolution data, this is described in detail in 3.3 and lead to the publication of a journal paper [5]. The ability to characterise driver styles lead to the development of the Driver Culture analysis methodology which is described in the next section.

1.5.3 Driver Culture analysis methodology

This analysis uses techniques that are more commonly associated with image processing and big data analytics, it is described in detail in chapter 5. The specific research in this section i.e. the application of these techniques to the rail industry with energy, control and vehicle performance data and using driver style characteristics (described above) as feature vectors is novel. More importantly the general research area of human (i.e. driver) performance analysis using real data is a completely new field in the rail industry. Findings from the first Driver Culture analysis undertaken in this thesis lead to the presentation of a conference paper [6] as well as journal paper which is under review as of 2017. The author is currently involved in a project to develop driver training strategies that are based on real data analysis as well as applying for funding to implement the methodology on a larger rail network to develop its capabilities. It is hoped that as part of continuing increase in vehicle sensing, data recording and automated analysis (the digital railway) that Driver Culture will be a widely accepted and understood term and analysis of this type would commonplace.

1.5.4 Partial fleet metering methodologies

This research is described in detail in chapter 6. The error characteristic between a known energy consumption take from an entire vehicle fleet and an estimate energy consumption taken from data representing a portion of a rail vehicle fleet are analysed. The statistical error method used was developed by Dr Gemma Nicholson and Dr Stuart Hillmansen using simulation and real data from a small sample of trains. The original contribution of this research was to apply the error method to data taken from an entire fleet for the duration of a year. The amounts of data involved were significantly larger and required extensive pre-processing time. As well, due to the analysed fleet operating over multiple routes, approximately 40, two methodologies that approached whole network and route specific analysis had to be developed. As well as these novel methodologies, another major output of this research has been recommendations that have advised the Rail Delivery Group on how to offer partial fleet metering to Train Operating Companies (TOCs). The findings from the research determine the minimum number of vehicle that require energy metering refitment to give accurate energy bill estimates, which could potentially change the way in which billing is undertaken and reduce energy costs for TOCs. A journal article describing the findings is currently in preparation.

1.5.5 Partial and whole fleet hypothesis testing

This is the follow-on research to the partial fleet metering analysis described above. Any significant change to the operation of a rail vehicle fleet that increases or decreases energy consumption has varying effects on the mean energy consumption estimate, meaning that billing based on a partial fleet that has been determined from real data is likely to become inaccurate. Hypothesis testing, described in detail in chapter 7, is a standard statistical method that is commonly used in medical testing, opinion polls and economics. The application of hypothesis testing to rail vehicle energy data is and original contribution to research. Although it is known that energy consumption increase or decrease will have an impact of the energy bill estimate, the method allows simulation to test how large a change must be before significant changes take place. This can then be used to test successive years’ energy consumption against historic energy consumption to maintain that estimates are similar with known significance. This can protect TOCs and energy infrastructure owners against unfair billing or energy use.

1.6 Publications and awards

This section documents a number of publications and awards that works taken from this body of research has received:

08/09/2014 – Winner best presentation, Business and Science – Partner for success: BCRRE research seminar.

10/03/2015 – Winner best paper, IMechE Railway Division NW Centre Young Engineers' Presentation Competition.

11/05/2015 – Runner up, IMechE Railway Division Future of Rail Young Engineers Presentation Competition.

15/06/2015 – Publication, Observations of train control performance on a camshaft operated DC electrical multiple unit, published in Proc. IMechE

10/11/2015 – Conference Publication, Identification of a 'Driver Culture' and its effect on energy consumption on a DC rail network, presented at the International Symposium on Speed-up and Sustainable Technology for Railway and Maglev Systems, Chiba, Japan.

The following section documents other publications awards and achievements that were received during the PhD period but unrelated to the body of research:

08/07/2013 – Winner, best presentation, RRUKA Next Generation Rail Conference, graduates into rail ad campaign presentation.

28 – 30/06/2013 – Fuel cell narrow gauge locomotive build Project leader, IMechE Railway challenge 2013.

27 – 29/06/2014 – Fuel cell narrow gauge locomotive build Project leader, IMechE Railway challenge 2014.

10/11/2015 – Conference Publication, What happened to hydrogen?, presented at the International Symposium on Speed-up and Sustainable Technology for Railway and Maglev Systems, Chiba, Japan.

September 2015 – March 2016 – Route simulation design and analysis, Future Railway power train challenge – Hitachi, BCRRE and Fuel Cell Systems consortia.

September 2015 – March 2016 – Route simulation design and analysis, Future Railway power train challenge – Analysis work carried out for Gvolution lead consortia.

2 Literature Review

This chapter presents a general literature review that gives an introduction to Big data and its use in the rail industry. Each chapter provides more specific literature that is related to the case studies being discussed, analysis techniques and algorithms. As this chapter only provides a general background the reader may wish to advance to chapter 3, page 25 to commence with the Merseyrail case study, or to chapter 6, page 168 to commence with the ATOC case study.

2.1 What is Big data?

The aim of this research is to provide a persuasive argument for the use large amounts of data in the rail industry, and, as described in the introduction chapter, there is an opportunity to demonstrate the additional value that can be gained by inclusive analysis of the data beyond its original intent. There is substantial interest and numerous examples of research on this type of data use in the rail industry but initial costs and lack of tools to process data [1] remain barriers to implementation. This thesis examines potential uses of large datasets and how they might be used to tackle this problem. This highlights a question of whether the data being used in this research qualifies as ‘Big data’, which requires a review of current definitions and a comparison to the characteristics of the datasets that will be examined.

The first academic paper to mention the term was published in 2003 [8, 9], but, Big data is a term that likely has its history in the late 1990s [8] with a series of talks from John Mashey of Silicon Graphics [8, 10] to which Diebold attests to in a reflective paper in 2012 [11]. Importantly at this stage the term did not fully encompass the current meaning that is attributed to Big data, but predicted the explosive rate of data generation that would occur in the two decades following the late 90s. It is quite clear then a key characteristic of Big data is the size of datasets. There is a consensus that Big refers to data that is too large to be adequately recorded, stored and processed for analysis by conventional means [12, 13, 14]. This means a datasets physical size it too big for single storage facilities to hold, i.e. a modern hard drive, and requires multiple (often thousands) of servers to record and store. As a result data can be comprised of matrices of billions of rows and columns [14].

This point is useful to compare to both datasets that are used in this analysis. The first dataset that is discussed was recorded during a vehicle and infrastructure instrumentation project conducted on the Merseyrail network that services Liverpool and the surrounding Merseyside area; results from this data analysis are discussed in chapter 3. A single vehicle was instrumented which recorded 16 data inputs at a data rate of 256 Hz. In this case the number of columns in question is significantly below the possible billion columns suggested by Qureshi [14], but due to the high data rate a single day of operation can produce datasets of 5 – 20 million rows, depending on how long the vehicle is in service. This results in an upper estimate of 1.8 billion rows of data that might require processing. Should the findings of this analysis be used for continued data monitoring on the Merseyrail network, it is expected that substantially more vehicles would require instrumentation; with the full fleet of 59 units this would produce an upper estimate of 107 billion rows of data for a three-month period, equivalent to approximately 1 billion rows of data per day.

The next dataset analysed, discussed in detail in chapter 6, is taken from energy meters installed on an entire DC EMU fleet. Data rates are significantly lower than the Merseyrail dataset recorded at 0.0167 Hz (once per minute). This results in only 86400 rows of data per meter in operation per day. The fleet has almost 300 units, as each unit has two DC meters an upper estimate of 51 million rows of data per day can be expected. Data for one year was used in this analysis which results in an upper estimate of 18 billion rows of data. Retrieval and processing of both datasets proved difficult and extremely time consuming, in both cases data had to be broken into accessible chunks and processed separately before output could be recombined to produce meaningful results. This plus the large amount of data described above suggests both datasets fit well into the initial definition of Big data.

The second note on how Big data is defined concerns the characteristics of the actual data and how the data are collected. Earlier literature suggests that Big data should have the following three characteristics [15, 16, 17]:

- Volume
- Velocity
- Variety

The so called 3 Vs were introduced by researchers at IBM and have been in use since the early 2000s [18]. Volume has already been discussed in the previous section, velocity is related to the speed at which the data are produced and recorded as well as the speed at which it can be processed then meaningful results gained from [15]. Variety is a somewhat understated term that refers to one of the most difficult aspects of Big data analytics [16], indeed Madden [19] suggests that this term is a catch all term meaning ‘too hard’ i.e. is difficult to perform analysis on. Variety is the term which is used to distinguish ‘small data’ which is typically stored in spreadsheets and databases between data that is available in any format and is normally unstructured or semi-structured. Data used in this research, although having substantial volume and velocity, are not varied compared to what is commonly associated with Big data. Table 1, reproduced from Tien [13], shows examples of data inputs that demonstrate large variety data which are distinctly different from the purely numeric or logic data used in this research.

Table 1 – Table showing Big data variety examples [13]

Scope	Example acquisitions	Example efforts
Data capture	Keystroke logger; clickstream; smart sensors; health monitors; drone sensors; samples	Monitoring software; website trackers; smart phone apps; RFID; ornithopters; memoto; compressed samples
Multisensory data	Visual detection; video cameras; light-field photography; beyond video and audio telepresence	Thermal imager; bug's eye; lytro; internet transmission of touch, smell & taste senses
Brain imaging	Magnetic response imaging; functional MRI (fMRI); Diffusion MRI (fMRI)	U.S.'s Human connectome project ($40M); E.U.'s Human brain project (Euro 1B); U.S.'s BRAIN initiative ($100M)
Real-time sensing	Real-time location data; real-time image display; real-time response	Smart phone-based, global positioning system (GPS); motion & image sensors; OLED TV; ocean observatories; smart grids; smart cities

The 3 Vs definition is still commonly quoted in literature, but there are also indications that current opinion is that these do not adequately define the characteristics of Big data [16, 18]. As of 2011 International Data Corporation (IDC) has introduced a fourth V term, 'Value' [18, 20]. A fifth term is also often described in literature, 'Veracity', however this term is closely related to value and can be considered to be a similar characteristic [18]. Veracity refers to the varying quality of data that is captured. Where data are purpose recorded there tends to be substantially more focus on maintaining its structure and quality. In the case of Big data, the data collected is simply everything that is available and as such it is inevitable that there will be differences in data quality [14]. Veracity is therefore often a hindering characteristic that is associated with Big data which in turn affect the value of data. Tien as early as 2003 [21] discusses the effects of data Rich information poor (DRIP) data that are a result of having the

original 3 Vs and high veracity. The result is Big data with low value as the information gained from it is poor. Veracity was known to be problematic in the datasets studied in this research in a number of areas:

- Merseyrail dataset
    - Frequent vehicle positioning data (GPS) errors and loss of signal.
    - Frequent periods of recorded data with no vehicle operation.
- Energy meter dataset
    - Frequent vehicle positioning data (GPS) errors and loss of signal.
    - Total energy recording not equally shared between meters on a single unit.
    - Meter failure – varied number of inputs not recorded, recorded with significant error or lack of all data.
    - Data in raw form ID tagged to individual vehicles only, unit number not available.
    - Data are not associated to relevant service codes or operational routes.

Data veracity issues are discussed in detail in relevant case study chapters. A significant amount of time was spent during the research period to develop a method to process data in such a way that removed or corrected data quality issues. Based on the definitions and concepts described in literature the datasets being analysed in this research mostly qualify as ‘Big data’. Big data analytics in single areas might take data inputs from many, often incomparable, types and sources suggesting that it is variety that defining term that has the greatest influence over the qualification as Big data. As a result of the relatively low variety of the data being examined in this research a similar term is given to the datasets to make this distinction: ‘Rich data’ high volume, velocity but relatively low variety datasets. The case for the fifth V, or these datasets value, is the subject of the case studies, results and conclusions of this research.

2.2 Data analytics in the rail industry

This section describes some examples of research where Big data analytics are used in rail industry specific areas. Optimised time tabling of services is one of the most important aspects of rail operation. A good timetable allows for high capacity, regular services, rapid recovery time in the event of delays and a high degree of interoperability with other modes of transport and customer satisfaction. Modern timetables are developed with computer modelling that attempts to design an optimal schedule based on a set of requirements which are usually frequency, unit headway, average and maximum unit speed and passenger needs. The desired result is a clock face timetable that is simple for passengers to use. With the availability of massive amounts of data more accurate timetabling is possible. Data that are described in literature relating to scheduling is commonly [1]:

- Accurate vehicle position data
- Passenger smart card data

A delay of more than a few minutes, in mainline operation, and often seconds in metro operation can cause disturbances across a whole rail network where eventual delays some distance from the original delay can be significantly longer [22]. In this event train arrival and departures must be re-scheduled to minimise the initial disturbance and quickly return the network to its normal operation. Frequent train position and speed data allow this to be done very accurately allowing dynamic re-scheduling and recovery [23]. As well as re-scheduling, position data has been used extensively to provide likely arrival time of trains to passengers via smart phone applications.

There is a significant amount of data available from ticket purchasing outlets. Specifically where smart cards and phone apps are used [24]. This type of data has been used to analyse and predict passenger flow in dense passenger use multimodal-transport networks [25, 26], as well as to contribute to increased security at transport hubs [27]. Depending on the available data that are recorded when a passenger makes a journey purchase there are a number of varying analyses that might be performed that provide a range of different benefits [24]. Where boarding time and data with location are recorded typical passenger travel behaviour can be identified as well as their seasonal variation [28, 29]. Where personal details are also connected targeted marketing strategies can be used as well as customer specific fare adjustment based on service loyalty [30]. Importantly better understanding of passenger flow and travel behaviour mean it can be reflected in timetable design [26]. Currently timetables are designed based on train performance and maximum capacity they can offer, rather than a direct link to data based passenger demand. Passenger demand has been attempted to be included in timetable design but this data has normally been gained via census data that uses only small and often non-representative samples of passengers [26]. As a result Big data use in timetabling is shown to have potentially significant impacts on passenger satisfaction.

There is a substantial lack of effective data use in performance and condition monitoring aspects of the rail industry. However, it is important to first establish what is meant by 'effective data use'. Condition monitoring and system performance are reliant on numerous sensors and collate substantial amounts of data for analysis. It is often the case that specific analyses are dependent on purpose gathered data. This means that data tailored to a specific aspect of rolling stock, be it a bearing fault, a wheel flat or electrical fault are monitored by a sensor that is intended solely for the detection of problems with that particular aspect only [31]. Although there are substantial amounts of data processing and analysis, this type of approach might be better described as sensor driven condition monitoring rather than data driven. Data driven condition monitoring is considered to be a more effective use of data than sensor driven condition monitoring. There are fewer examples of data driven condition monitoring. Some examples are [32], where vehicle acceleration data is also used to analyse track condition. In this case data can be re-used to discover ‘hidden’ information that are beyond the original use of the accelerometer data, providing added benefit [32]. The same is true of bespoke equipment that is designed to record track alignment for example. Laser scanning systems described in [33, 34] can effectively scan the track ahead of a moving train or from a static position. This requires a number of devices and fitment to numerous vehicles which incurs significant cost and requires a substantial recording period. Secondly, track condition maintenance using this equipment is often only done when regular service vehicles are off duty [32]. As a result only small parts of the network are able to be checked in a short period. It is likely that similarly effective track alignment data can be inferred from existing sensors that are fitted standardly to most rolling stock [32]. The advantage here is that all vehicles are able to take data more frequently over a greater degree of a network, providing a more data lead condition monitoring solution.

2.3 Driver advisory systems and data analytics

Compared to other land based transport rail is highly efficient [35, 36], this is largely due to the low rolling resistance of the steel wheel on steel rail contact, modern power electronics and electric motors in electric trains and greater efficiency of diesel rolling stock compared to road vehicles [37]. The inherent energy efficiency of rail transport is therefore an important aspect that can allow it to compete with other land-based transport modes [38, 36]. However, even with these benefits rail networks are some of the largest single consumers of electricity. Where electrification infrastructure exists, there are no carbon emissions at point of use, but there is an important carbon footprint associated with operation, for example in Great Britain some 326 ktCO$_2$e during 2014/15 [39] approximately 0.3% of the total national transport emissions [40]. British railway carbon emissions have risen almost 10% from 2013 to 2016 [41, 39] and so, there is a need to reduce both energy consumption and emissions from rail operations. In very general terms energy conservation during traction can be made in two ways:

- Capturing or returning energy

and

- Not using the energy in the first place.

Energy capture and storage refers to electrical energy generation during braking. This can be fed to other vehicles to use during acceleration, returned the electricity grid or stored on board a vehicle in batteries or other storage systems such as flywheels or supercapacitors [42]. Douglas [43] gives a detailed overview of a these energy reduction strategies and their interactions when implemented. Table 2 shows key factors identified in [43] for energy saving in rail.

Table 2 – Key factors for energy saving in rail (adapted from [43])

Route	**Service**	**Vehicle**
Line length	Interstation distance	Mass
Gradient	Driving style	Aerodynamics
		Power
		Efficiency
		Maximum speed

With access to detailed data regarding driver control (described in the following sections) there is an opportunity to assess its impact on energy consumption, as a result this research is focused on the effect of the key factor 'Driving style'.

Much of the literature concerning driving efficiency is based on models, commonly this is used in the development of Driver Advisory Systems (DAS) [44]. Although DAS will undoubtedly become widespread due to their potential energy savings of 15 to 25% [45, 46], there are problems with effective implementation. Mitchel [47] shows an example of energy consumption difference from driver trajectories, and a simulated optimal trajectory. In this case 90% of drivers could not follow the optimised trajectory, resulting in energy consumption increases of over 25% in most cases. There are a number of reasons that may be responsible for this difference ranging from trajectory complexity to driver priorities. Examples in [45] describe optimised trajectory implementation on a mining route in northern Scandinavia. Two drivers, a 'good' and a 'bad' are compared whilst attempting to follow the target trajectory. Although the 'good' driver achieved a significant reduction in energy consumption, they are not able to follow closely due to the difficult driver requirements of the complex DAS trajectory. The bad driver frequently exceeded the suggested speed limit and was forced to aggressively brake a number of times resulting in an energy consumption increase of 1500 kWh (for a 1.5 hour journey of approximately 80 km). In addition there is a significant amount of literature documenting driver error caused by advisory systems installed in cars. Arguably, the control of a car is substantially more complex as the driving environment is more dynamic and involves a greater number of vehicles in close proximity when compared to rail. However, the same message is conveyed in this literature: Advisory systems on the whole provide benefits to the driver, either to avoid accidents, reduce speed or to take caution providing that warning information is provided to a driver that is consistent with their current situation and unobtrusive [48]. Morris et al. [49] suggest that currently too little is known about driver interaction with new technology and that the majority increase the risk of driver distraction. Findings in Morris et al. [49] show that glance time at either advisory or journey planning equipment is often below thresholds for being dangerously distracted, but stress that increased complexity of in vehicle systems would undoubtedly cause dangerous distractions [49]. This is supported by other examples in automotive literature [50, 51] as well as the Rail Standards and Safety Board's (RSSB) field tests of a standalone DAS system [52] which showed overwhelmingly that drivers preferred a simple text based instruction system to keep workload at a minimum.

In [46] and [53] a more simplified DAS approach is presented. In this case, an optimal trajectory is described by well-defined driver regimes 1) Full power, 2) Cruise & coast 3) Braking. The trajectory profile is relayed to the driver one regime at a time in text form to reduce information processing and workload. Albrecht and Lupien [54] also suggest that a driver information system (DIS) can help with anticipation of conflicts, improve ride quality and reduce energy consumption based on track information, driver experience and professionalism. In this case specific driving instruction, as with a DAS, is not required. This can be seen as improving driver style using limited information and driver best practice.

Providing simple and non-distracting driver advisory systems have the potential to reduce energy consumption whilst maintaining driver competency and safety. Often, from a driver's perspective, there is less importance placed on energy consumption, the priority being instead on safety and keeping to the timetable. As a result there is a risk that driver advice systems could be ignored, particularly during busy time periods and in the event of delays. Chapters 3 and 5 present of observations taken from a data recorded on an instrumented EMU (electrical multiple unit) that aim to identify driver styles, and suggest optimal driver control from the best styles identified. Data from a vehicle in service is normally difficult to obtain. This is especially true for older rolling stock, which the Merseyrail network currently uses. Although modern rolling stock are now usually required to be fitted with energy monitoring instrumentation, this is often only stored at 1 to 5 minute intervals, which does not allow for detailed examination. Data rates from instrumentation that are discussed in these chapters are substantially higher and give a unique opportunity to examine multiple train parameters in high resolution. Rather than designing a real-time DAS, these observations provide detailed performance analytics from which efficient driving tactics can be identified and good practice passed on to all drivers. Such data analytics have been used in sport for many years [55] and are becoming more precise due to the availability of large amounts of data. The IMB Slamtracker [56] as well as research conducted during the 2014 soccer world cup [57] both highlight the competitive edge competitors can gain by identifying the elements that result in consistent successful play, the so called "keys to the match". Applied to the railway, performance indices can be identified that are suitable for use as driver instruction tools. Merseyrail rolling stock is to be upgraded by 2020, as an interim solution, there is potential to identify and instruct drivers on how to employ efficient driver styles, which if widely implemented could lead to energy savings.

3 Merseyrail Energy Monitoring Project (MEMP)

The Merseyrail network is centred in Liverpool in the UK, the electrified part of the railway services the surrounding towns in the Merseyside area and southern Chester area. Non-electrified sections of the railway provide services to Preston, Manchester and the North-East, these routes are not examined in this thesis. The electrified network consists of 120 km of track, has 67 stations and carries approximately 34 million passengers per year. The network consists of two major lines, the Northern line and the Wirral line. The Northern Line operates from Hunts Cross and Liverpool Central, providing services to Southport, Ormskirk and Kirkby. The Wirral Line operates from Chester and Liverpool Central and provides services to West Kirby and New Brighton. Merseyrail is operated by Merseyrail Electrics which is owned through a joint venture between Serco and Abellio with a turnover of approximately £144 million. Merseyrail receives the highest government subsidy of all the Train Operating Companies (TOCs), some 12.4p per passenger kilometre, this equates to £86 million per year, this subsidy has reduced by some 2.3% from previous years [58]. As well there is an annual demand to reduce costs whilst maintaining high quality service and increasing profit margin [59]. Many of the operational costs are essentially fixed or will potentially rise. This is the case for staff salaries, land and infrastructure costs paid to Network Rail and taxation. As a result there is an effort to realise savings in areas that do not have significant cost to implement and, even though small in comparison to total operational costs, can have an impact on profit margin. As a result Merseyrail have had an interest in investigating electrical losses and identifying solutions to minimise them as well as expressing interest in developing options to reduce traction energy consumption.

Even though Network Rail is the largest single user of electricity in the UK energy costs are still relatively cheap compared to many other aspects of operation. TOCs, the end user of electricity, are billed for their use by Network Rail and although costs are significant they only represent a small contribution to the total operational cost. Figure 1 shows a pie chart breakdown of Merseyrail's operational costs. The costs shown in Figure 1 sum to approximately £100 million which is short of the £144 million turnover discussed in [59]. The pie chart only shows the breakdown of the major operational costs only and does not include investment, donations, and various other non-operational costs. The author was informed that electricity costs are approximately £5 million [60] meaning that traction energy accounts for around 5% of the total operational costs. Merseyrail increased profit from £13.8 million in 2013 to £14.6 million in 2014, although only a small portion of the total operation costs, a saving of 10% on energy costs would equate to £500,000. This would have a significant impact on further profit margin; given that other operational costs a not flexible energy reduction research and projects have potential financial impacts that are worthwhile for Merseyrail. The Merseyrail Energy Monitoring project was undertaken to examine what options for energy saving were available to Merseyrail.

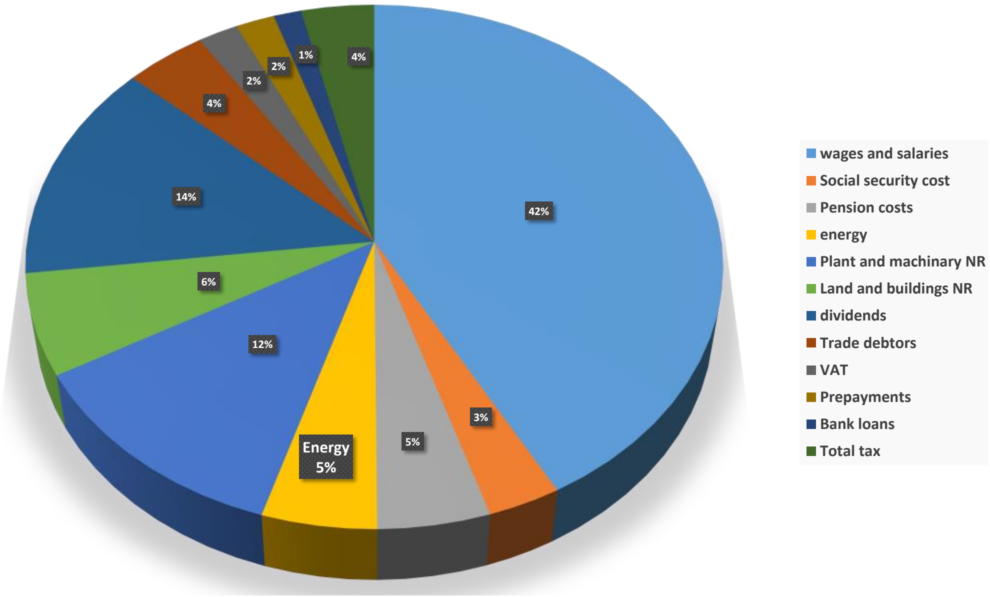


Figure 1 – A pie chart showing the breakdown of operational costs for Merseyrail for the year ending 4 January 2014 (52 weeks). Energy although a major contributing factor to the successful operation of the railway is only 5% of the operational cost.

3.1 Project Background

The MEMP was an electrical power network focused project undertaken by the University of Birmingham and Merseyrail. One of the aims of the project was to identify the origins of system losses in the network. Three substations and a single vehicle were instrumented with equipment designed and built by the Birmingham Centre for Rail Research and Education (BCRRE). The system was designed to show that high resolution data taken from different locations could be used to calculate instantaneous power flow to assess system losses [61]. Monitoring equipment installed allowed more detailed data acquisition than is standardly available from existing vehicle data logging equipment; one of the unique benefits was data that allowed examination between driver input and vehicle performance. This allowed a rare opportunity to examine the effect of driver style on energy consumption. The case study detailed in chapters 3, 4 and 5 discuss the development of a methodology which identifies different driver styles in real data taken from Merseyrail. Analysis of these driver styles in turn highlights where potential energy savings could be made.

Data were recorded when the instrumented vehicle was in operation. When data storage reached capacity new data would replace the oldest. Since retrieval of data was possible only when the vehicle was under maintenance datasets are not continuous over the three year project. As a result data used in the following chapters is comprised of continuous operation of the instrumented train over a period of three months, September to November, in 2011.

Findings detailed in this case study have been reported to Merseyrail and have resulted in increased interest for follow on projects and will likely have an impact on continued improvement of driver instruction. In the short term though this is extremely difficult to validate and is not considered in this thesis. Potential impacts on energy consumption as a result of driver style shifts based on these findings are presented in 5.6. Two further outputs from this case study are an academic journal [5] published in Proc. IMechE Part F, Journal of Rail and Rapid Transit, and a conference paper [6] presented at STech 2015 in Chiba, Japan. An extended version of this paper has been submitted to the International Journal of Railway Technology and is currently under review. These papers are included in Appendix 10.2, 10.3 and 10.4. Chapter 3.3 is reproduced from [5]; Chapter 5.5 contains the findings of [6] which is extended to include three more routes, detailed energy consumption estimations and optimised driver style suggestions. Conclusions from each paper are combined and presented at the conclusion of chapter 5.6.

3.2 Modelling and simulation

The University of Birmingham’s single train simulator (STS) examines the kinematic performance of rail vehicles and has been used extensively and described in the literature [62, 63, 64]. Prior to deeper analysis of the large datasets provided by the MEMP a vehicle modelling project was done. The modelling project had three aims:

- To generate a large dataset quickly to be used for initial testing.
- To vary driver style in a controlled and clear manner to produce extreme features.
- To use these varied driver styles to verify the effectiveness of a clustering algorithm to be used on real data.

3.2.1 The Single Train Simulator

This simulator has been used in a number of research projects where a significant number of simulation output has been presented in various published literature [62, 64]. Output is achieved by solving Lomonossoff's equation (equation 3) per distance iteration [65], which is essentially an extension of Newton's third law, equation 1, with consideration to various forces which a land vehicle encounters during movement. Distance can be set in the STS software, standardly each iteration is set to ten meters. Calculating over distance steps of ten meters allows a sufficient level of detail in the simulation output and keep the overall simulation time low. Simulations described in 3.2.3, 3.2.4 and 4 use a distance step of one meter for each iteration. This is due to, in some circumstances, camshaft simulation changes occurring in sub-ten meter distances meaning that these larger distance step missing these characteristics during simulation, the downside is that this significantly increases simulation time.

$$F = m_e a \tag{1}$$

$$Where\ m_e = m(1 + \lambda) \tag{2}$$

$$m_e \frac{\mathrm{d}^2 s}{\mathrm{d}t^2} = F - R_{\mathrm{total}} \tag{3}$$

$$Where\ R_{\mathrm{total}} = R_{\mathrm{motion}} + R_{\mathrm{gradient}} + R_{\mathrm{curviture}} \tag{4}$$

$$R_{\mathrm{motion}} = A + B\frac{\mathrm{d}s}{\mathrm{d}t} + C\left(\frac{\mathrm{d}s}{\mathrm{d}t}\right)^2 \tag{5}$$

$$R_{\mathrm{gradient}} = Mg \sin \alpha \tag{6}$$

$$R_{\mathrm{curvature}} = D\frac{mg}{r} \tag{7}$$

In the above equations m is the total mass of the vehicle, m_{e} is the inertial mass where rotary allowance λ increases the required force to begin movement.

Maximum force is limited by a resistive force, R_{motion}, which comprises of, A, resulting from the vehicle mass. B, resulting from bearing resistance and rolling resistance; it is speed dependant. C resulting from aerodynamic resistance; also speed dependant. Resistive forces due to changing gradient are considered in equation 6, when the vehicle is traveling uphill this adds to resistive force, when the vehicle travels downhill R_{gradient} becomes negative, instead assisting the vehicles movement. Curvature resistance is not active in the following simulation discussion, the D value is considered to be 0. Curvature resistance is problematic when track geometry has high curvature, i.e. there are tight bends in the route. The Merseyrail underground loop section has the tightest curve radius on the network, 210 m. There is a longstanding temporary speed restriction (TSR) put in place by network rail. This is due to the relative severity of the loop curve section as well as significant track degradation caused by water increased ingress and stray current [66]. Degradation lead to the derailing of a Class 508 three-car set in 2005 in the underground section beneath Liverpool Central Station [66]. Track curvature is likely to have a significant effect in this section however due to the TSR forces are reduced. Simulations that travel along the route are constrained by the TSR, as a result there are no differences in driver control in these section, driver control for this low speed can only be achieved in driver control position, notch 2. As a result, there is no vehicle performance difference between different driver styles. This means that whether curving resistance is included or not, the simulation output would be proportionally identical validating its setting at 0. As well the Hunts Cross to Southport line does not pass the loop section meaning that this high curvature does not need to be considered in this particular simulation.

The simulation requires route specific information and vehicle characteristics to solve the above equation. Route information contains a velocity profile, relaying the speed limit and gradient per distance interval. Station location and terminal location are also included which modifies stopping points along the route. Vehicle characteristics are included with a vehicle profile script, shown in Table 3.

*Table 3 – Vehicle characteristic profile - *Payload can be adjusted to better reflect passenger load, it the initial setting this is set to 0 indicating an unladen vehicle*

Physical Properties	**Value**
Mass – m	121.9 t
Rotary allowance – λ	0.1
Payload - m_{p}	0 t*
Inertial mass – m_{e}	$m_{\mathrm{e}} \times (1 + \lambda) = 134.09$ t
Power	1200 kW
Max Speed	120 kmh^{-1}

Davis Parameters	**Value**
A Parameter	1.6 kN
B Parameter	0.279 kN/ms^{-1}

C Parameter	0.00186 kN/m^{2}s^{-2}

Class 508 model only Properties	Value
Wheel Diameter	1.143 m
Wheel Radius – r_w	0.571 m
Gear Ratio – R_g	14 : 69
Number of Motors – n	8 per 12 axels
Equivalent Inertia – J	$\frac{m_\mathrm{e} \times {r_\mathrm{w}}^2 \times {R_\mathrm{g}}^2}{n} = 0.2254$

Performance Properties	Value
Gravity – g	9.81 ms^{-2}
Mass on driven axles – m_a	$m_\mathrm{e} \times \frac{8}{12} = 89.84$ t
Coefficient of friction – μ	0.1
Maximum tractive effort – TE_max	$(m_\mathrm{e} \times 1000) \times g \times \mu \times \left(\frac{m_\mathrm{a}}{m_\mathrm{e}}\middle/1000\right)$ = 96.947 kN
Maximum acceleration	${TE_\mathrm{max}}/{m_\mathrm{e}} = 0.72\ \mathrm{ms}^{-2}$

Electrical Properties	Value
Combined motor internal resistance – r	0.424 Ω
Motor control option	Set to 1 to calculate traction from Class 508 motor script.

STS simulations provide several useful outputs. Figure 2 presents default output from the simulation. For this demonstration, a four-point route has been created with a simple gradient profile and single station, inserted into the centre of the route. The Class 508 vehicle file has a switch selection option to enable and disable the motor model which is used when simulating varying driver inputs. In the following figures this option is disabled and output is generated with the default simulation options. Figure 3 shows a flow chart which describes the basic operation of the STS.

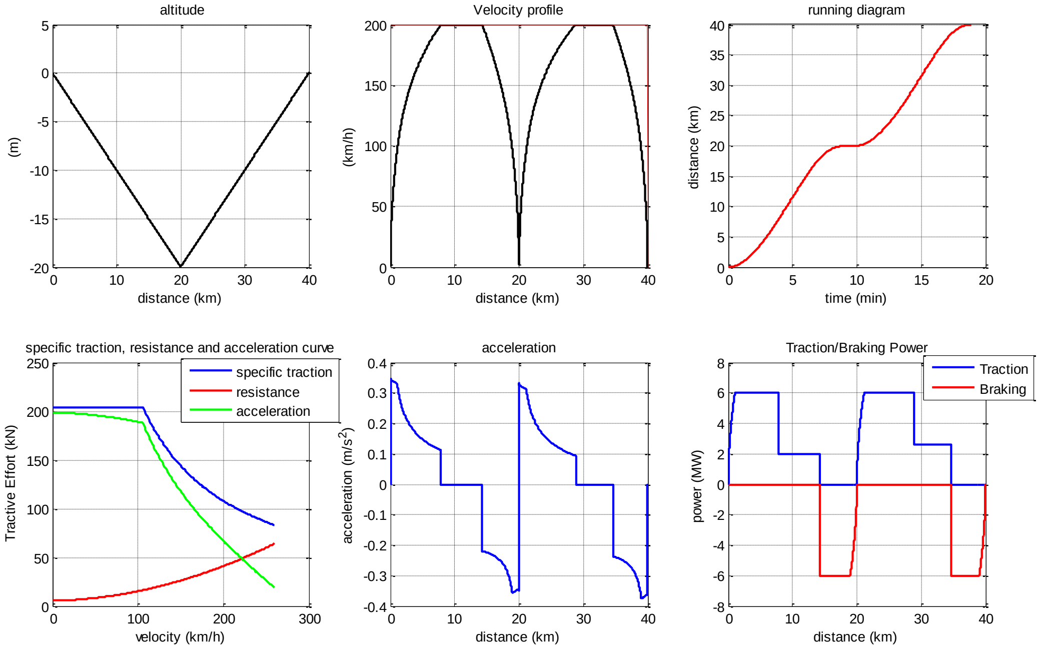


Figure 2 – Default output graphs from the STS for a simple 3 point route with single station

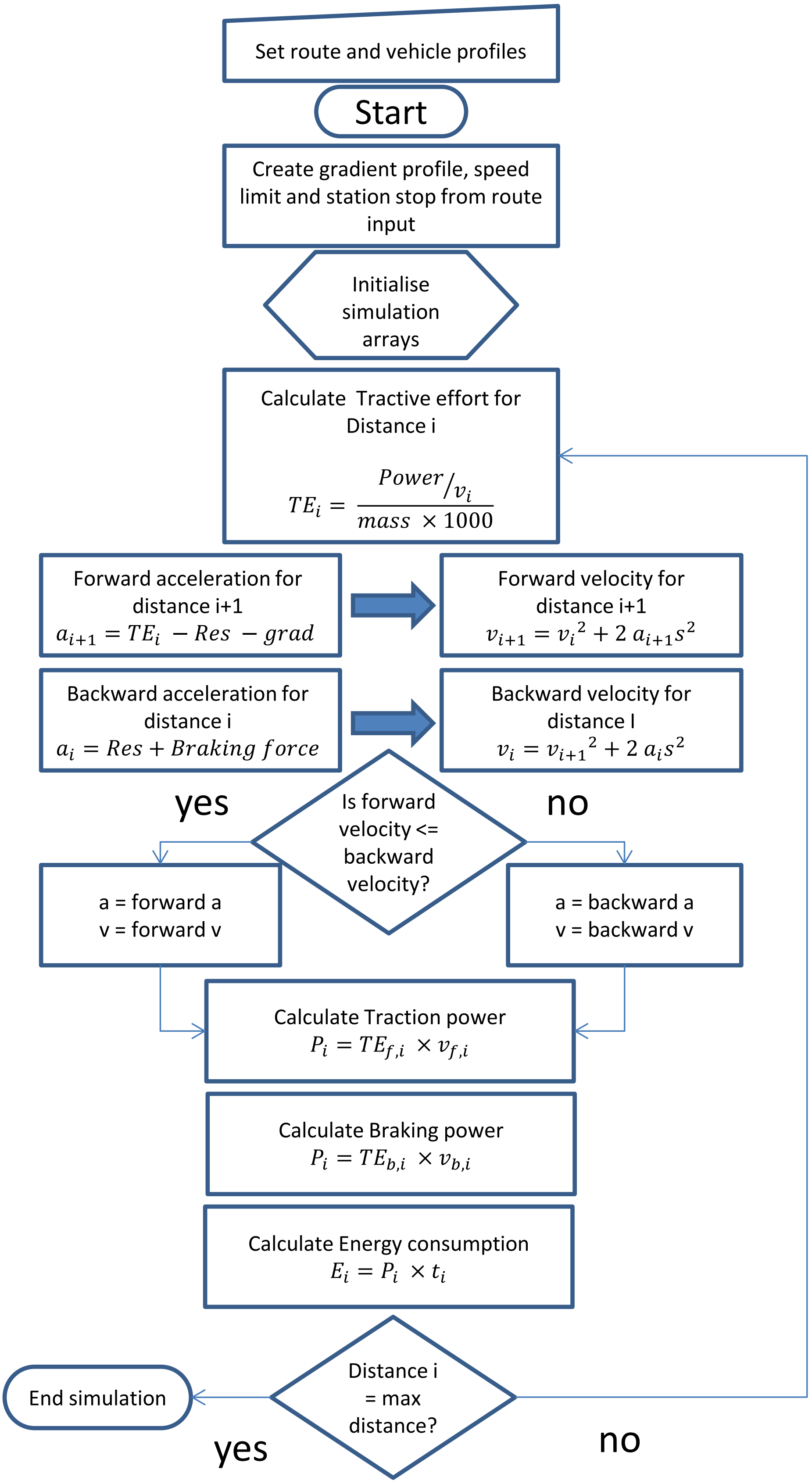

Set route and vehicle profiles
Start
Create gradient profile, speed limit and station stop from route input
Initialise simulation arrays
Calculate Tractive effort for Distance i
$TE_i = \frac{Power/v_i}{mass \times 1000}$
Forward acceleration for distance i+1
$a_{i+1} = TE_i - Res - grad$
Forward velocity for distance i+1
$v_{i+1} = {v_i}^2 + 2\, a_{i+1} s^2$
Backward acceleration for distance i
$a_i = Res + Braking\ force$
Backward velocity for distance I
$v_i = {v_{i+1}}^2 + 2\, a_i s^2$
yes
Is forward velocity <= backward velocity?
no
a = forward a
v = forward v
a = backward a
v = backward v
Calculate Traction power
$P_i = TE_{f,i} \times v_{f,i}$
Calculate Braking power
$P_i = TE_{b,i} \times v_{b,i}$
Calculate Energy consumption
$E_i = P_i \times t_i$
End simulation
Distance i = max distance?
yes
no

Figure 3 – A flow chart demonstrating the operational steps of the STS

3.2.2 The British Rail Class 507/8 EMU

This section introduces the British Rail Class 507/8 EMU with a detailed description of the operation of the camshaft rheostatic control traction system used by this vehicle. Part of this description is an expanded version of the background section of a published paper by the author [5].

The British Rail Class 508 is a camshaft controlled DC powered EMU that has been in service since 1978. The Class 508 consists of two motor cars (A and B) at either end of an un-motored (trailer) car, see Figure 4. All axles on A and B are driven by series wound DC motors. Although the driver control is common to both cars, the A and B end traction systems have independent control systems so can potentially act slightly differently. Driver control handle positions are commonly referred to as notches [67]. A mechanical camshaft automatically adjusts resistances and serial/parallel and field weakening configurations to maintain the desired traction current. The camshaft moves through 18 positions allowing four operational stages plus neutral. These stages are selected by the driver via the use of a five-notch handle designated 0 to 4. Operational stages correspond to handle position. In notch 0 the EMU receives no traction current. In notch 1, shunt operation, all camshaft resistors are connected and the motors are connected in series; this is used for manoeuvring only. Notch 2, is series only operation. Notch 3 is parallel operation, motors are connected in parallel to continue acceleration and trains can reach approximately half of the top speed. Notch 4 allows the final stage, weak field operation. Here field coils are shorted out of the motor circuit reducing magnetic field strength allowing the train to accelerate to its maximum speed. Selecting a higher notch position will result in a progression through all lower operational stages; a stage change is limited by the speed of the vehicle. These are controlled automatically by the camshaft current detector. A higher handle position setting will result in the camshaft moving through each of the lower settings in order. If the handle is moved from 0 – 4 all operational stages will be moved through as the vehicle speed increases until the demanded stage is reached. A change from a higher to a lower handle setting will reset the camshaft to position 1 and move to the requested position, for example, if the handle is moved from 4 – 3, stages will progress from weak field, to shunt, to series and finally to parallel. During braking, the power is disconnected and the motor circuit is connected to the braking rheostat. Figure 5 shows a typical acceleration phase of the Class 508, the notching current can be seen, showing a distinctive saw tooth pattern as the camshaft progresses and the train speeds up, all operational stages are demonstrated. Table 4 – A table summarising notch assignments and their function shows notch assignments.

Table 4 – A table summarising notch assignments and their function

Notch Assignment	Function	Operational effect
0	Traction system not engaged	Set to 0 when parked or when coasting
1	Shunting	Set 1 for manoeuvring, setting to 1 also resets camshaft from higher notches
2	Motors are connected in series	Begin acceleration, low speed operation
3	Motors are connected in parallel	Continue acceleration, mid-range speed operation
4	Weak field operation	Continue acceleration, maximum speed operation

Figure 4 – (Left) Instrumented Class 508 Birkenhead Depot, (Right) Class 507 (original three car variant) Meols station

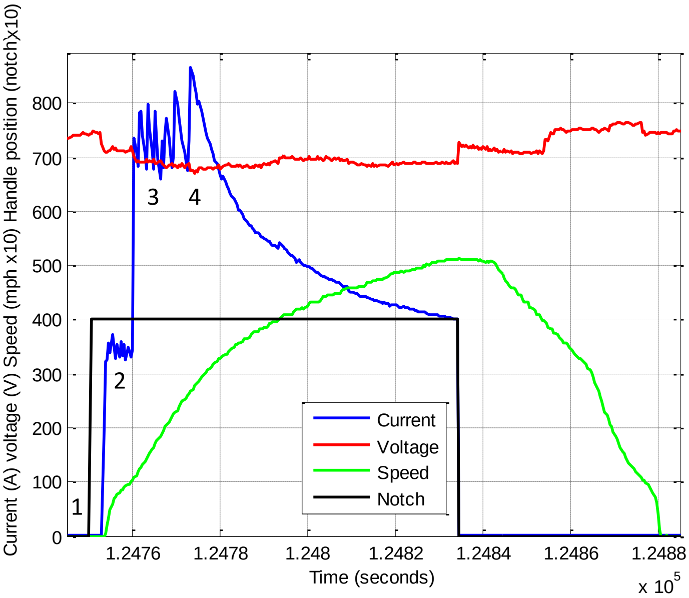


Figure 5 – Acceleration Phase of Class 508 in service. 1) Shunt operation, 2) Series operation, 3) Parallel operation and 4) Weak field operation.

3.2.3 Class 507/8 motor model

To accurately simulate driver interaction with the camshaft control of the class 508 a model of the traction system was designed. This model is then called as part of the traction and forward velocity calculation that the single train simulator provides. The traction system model produces a value for tractive effort based on current speed, traction current, camshaft resistor setting and line voltage. Figure 6 shows a flow diagram describing how the model functions. Initial values for camshaft resistors and line voltage are set at 1.95 Ω and 390 V. Although the nominal voltage of the network is 780 V the motor configuration at the start is series, as a result voltage across each motor is half of this resulting in the 390 V initial value.

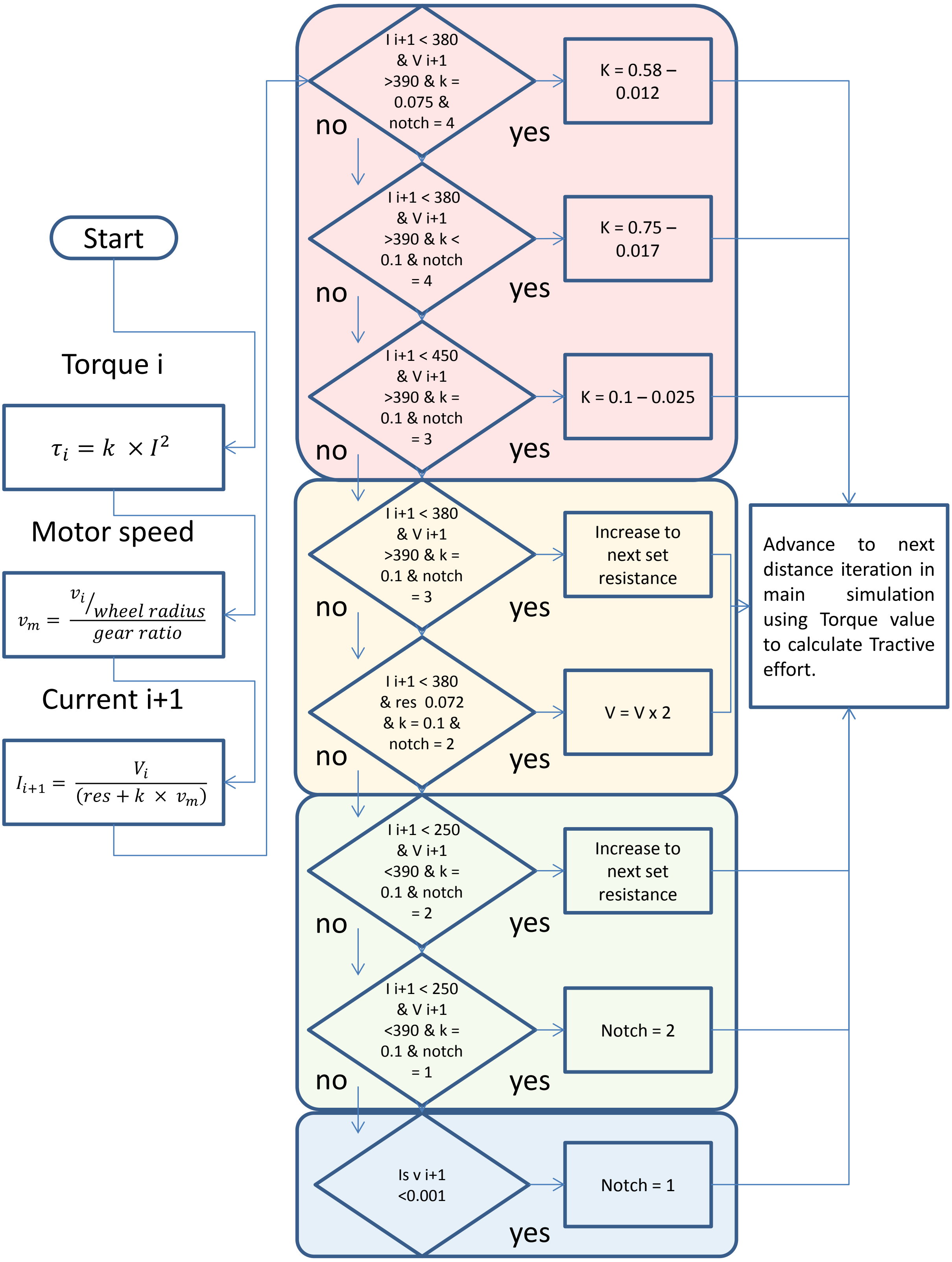

Start
Torque i
$\tau_i = k \times I^2$
Motor speed
$v_m = \frac{v_i / wheel\ radius}{gear\ ratio}$
Current i+1
$I_{i+1} = \frac{V_i}{(res + k \times v_m)}$
I i+1 < 380 & V i+1 >390 & k = 0.075 & notch = 4
yes
no
K = 0.58 – 0.012
I i+1 < 380 & V i+1 >390 & k < 0.1 & notch = 4
yes
no
K = 0.75 – 0.017
I i+1 < 450 & V i+1 >390 & k = 0.1 & notch = 3
yes
no
K = 0.1 – 0.025
I i+1 < 380 & V i+1 >390 & k = 0.1 & notch = 3
yes
no
Increase to next set resistance
I i+1 < 380 & res 0.072 & k = 0.1 & notch = 2
yes
no
V = V x 2
I i+1 < 250 & V i+1 <390 & k = 0.1 & notch = 2
yes
no
Increase to next set resistance
I i+1 < 250 & V i+1 <390 & k = 0.1 & notch = 1
yes
no
Notch = 2
Is v i+1 <0.001
yes
Notch = 1
Advance to next distance iteration in main simulation using Torque value to calculate Tractive effort.

Figure 6 – A flow chart demonstrating the Class 507/8 motor model. Blue – shunting, green – Series operation, orange – Parallel operation, red – Weak field operation.

Motor torque τ is calculated first, followed by tractive effort; this is done by considering gear ratio (R_g) and wheel radius (r_w) differences from motor to wheelset shown in equation (8). The tractive effort value F is then used by the main simulation to calculate vehicle speed over the next distance iteration v_{i+1}. Motor speed ω is obtained similarly and is shown in equation (9). Motor speed is essential as it is used to calculate the motor armature current I_{i+1} for the next ten meter distance step. Current I_{i+1} calculation is shown in equation 10, in this case r represents the total resistance provided by camshaft resistors in place.

$$\tau_i = k \times {I_i}^2$$

$$j = \frac{m_e \times {r_w}^2 \times {R_g}^2}{n}$$

$$F = \tau \times j \tag{8}$$

$$\omega = \frac{{}^{v}\!/_{r_w}}{R_g} \times j \tag{9}$$

$$I = \frac{V}{r + k\omega} \tag{10}$$

The existing current value I, i.e. the current at the distance interval being resolved by the simulation, is then passed through a series of if statements which determine the lowest permitted values for motor current. Higher current will result in greater torque and acceleration. Motor speed, and as a result vehicle speed are inversely proportional to armature current and as a result torque and acceleration. This means that torque and acceleration are greatest as the vehicle begins to move off and gradually decrease, speed increases over this time. When speed reaches its maximum and current its minimum the vehicle will continue to move at its balancing speed, this can be seen as the top speed for a certain operational stage. The four operational stages of the 508 described in section 3.2.2 result in continued acceleration. Within each operational stage there are several smaller stages that are a result of resistor changes. Without the resistor changes the vehicle would reach its balancing speed in the first 10 metres of operation. Before the balancing speed is reached a resistor is removed from the motor circuit automatically which increases voltage and current allowing a burst of acceleration, this produces a characteristic saw-tooth current profile and recognisable jerky motion during acceleration. Figure 7 Shows the first stage of operation (series) which relates to equation (10).

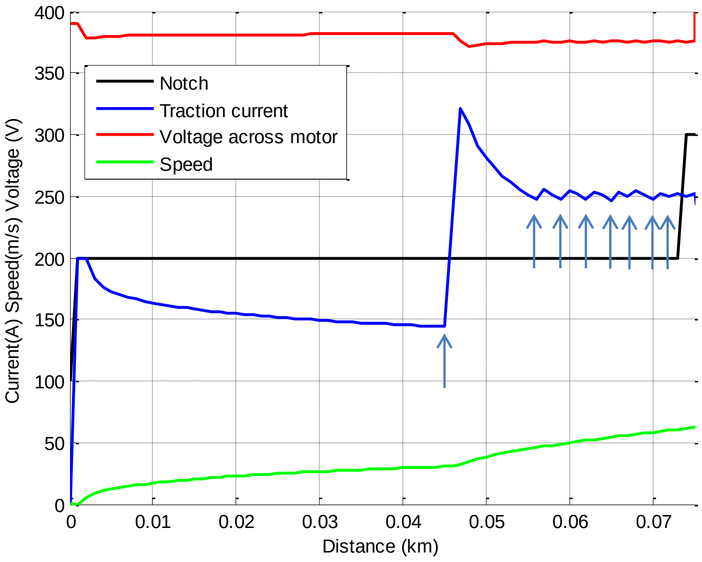


Figure 7 – Series operation (notch 2) of the Class 508 motor model, arrows denote camshaft resistor changes

The *if statements* approximate the automatic resistor changes a real electro-mechanical camshaft makes based on a low current threshold being reached, delaying these changes once the final resistor is retained has the effect of simulating different driver approaches this is discussed further in chapter 3.4. Once the first operational stage is completed, the traction system continues in parallel operation. This is approximated by doubling the voltage in the model; resistors are once again removed in a similar fashion to series mode. In the final operational stage, weak field operation, no more resistors can be removed from the motor circuit, in reality this requires the exciting magnetic field to be reduced to allow motor speed to increase; this is performed by shorting windings in the exciting magnetic field, k is reduced proportionally to the weakened field setting. Equation 11 shows the relationship between motor speed ω, field strength and flux.

$$\omega = \frac{V_m - R_m I_a}{k\Phi} \tag{11}$$

Where V_m is the voltage across a motor, R_m is the motor and camshaft resistance and I_m is the armature current. As k is proportional magnetic flux Φ, a reduction here approximates windings being removed from the exciting magnetic field. Figure 8 and Figure 9 show equivalent circuits demonstrating the motor resistor removal and exciting magnetic field removal process.

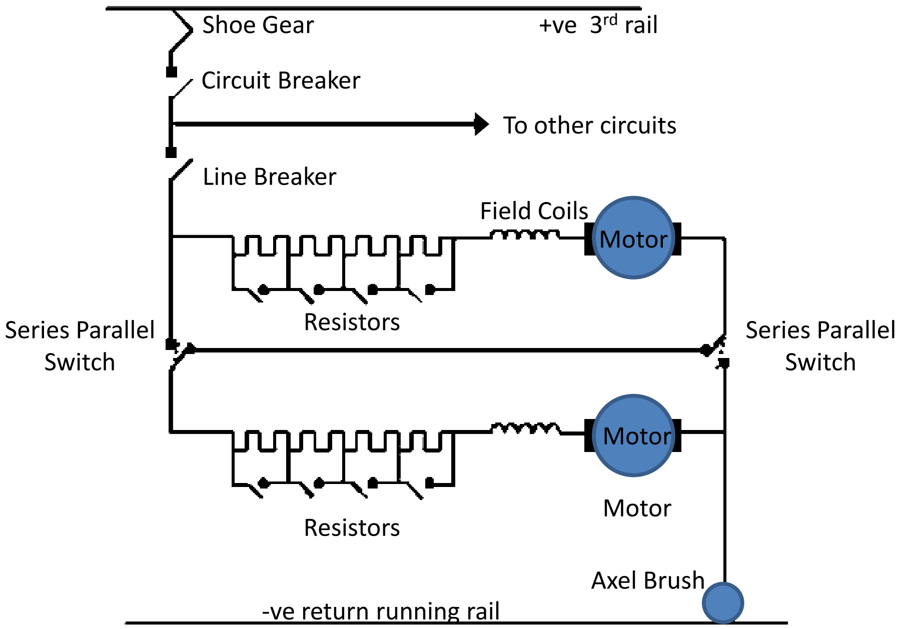


Figure 8 – DC camshaft controlled motor circuit adapted from [68]

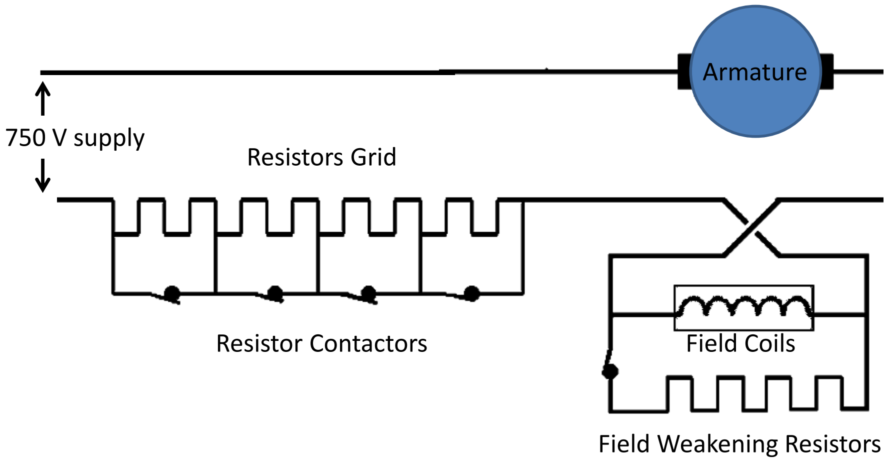


Figure 9 – DC motor Field weakening circuit
adapted from [68]

3.2.4 Two sample routes and simple power network simulations

Two routes were designed for use with the class 508 simulation. The instrumented vehicle was frequently in operation on these routes making them the most appropriate for simulation. Figure 10 and Figure 11 show the chosen routes, the Wirral line and Northern line of the Mersey rail network. Route design is a simple procedure providing route data is readily available. Both routes were used by the author and Merseyrail fleet manager Ian Jones. As part of a different project Ian provided a more accurate gradient and line speed profile using up to date information from Merseyrail. This was used to design the current routes used for simulation.

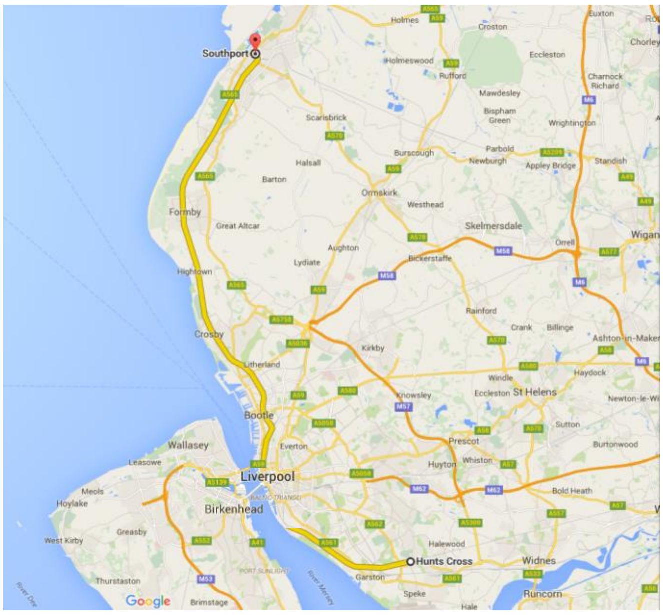


Figure 10 – Merseyrail network, route Hunts Cross – Southport

[69]

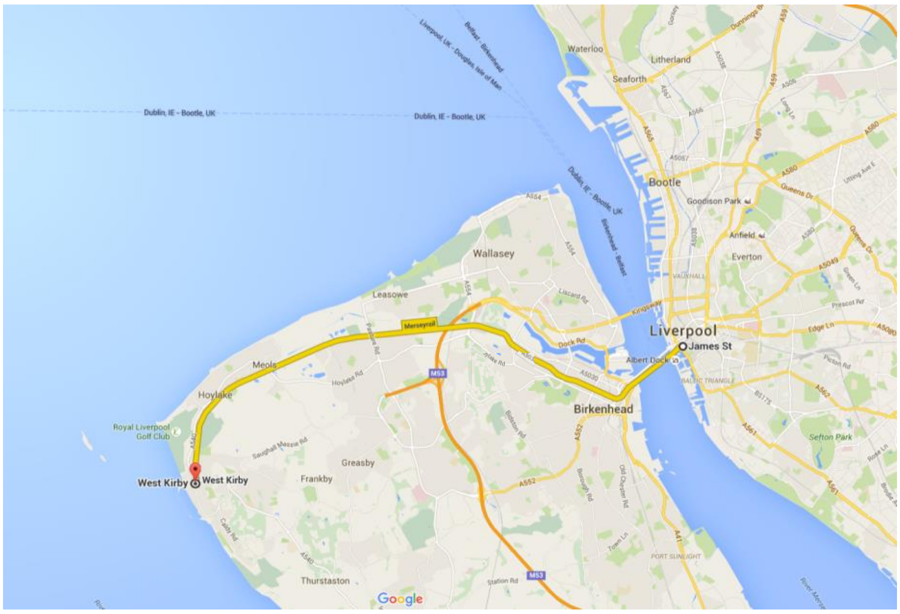


Figure 11 – Merseyrail network, Liverpool James St. – West Kirby

[70]

A route configuration file is prepared; for a basic variant, this includes four data arrays:

- speed profile
- gradient

- station location
- terminal location

Speed profile includes the various line speeds for the route and the distance into the route where each are implemented. Gradient includes the gradient between relevant points along the route, station location is a single array with distances into the route where stations are encountered, terminal location includes stations where a terminal time should be included, in this case, as only one journey is simulated a terminal is not necessary. A payload mass is included to represent a realistic passenger volume. 60% passenger loading was selected to compromise between a crowded morning or early evening vehicle and a less busy mid-afternoon or late evening vehicle. A standard 3 car set hold 192 seated passengers, a 60% passenger loading gives 115 passengers, additional passenger mass is therefore set to 8.9 tonnes (using the UK average weight of 77kg). Basic application of the motor model uses a constant value for line voltage, 390 V for the series operation stage and 780 during parallel and weak field operation. A mesh analysis was used to design an electrical network that provides voltage reaction to changing motor current and distance between feeding substations, this is added to the basic route profile variant and includes:

- substation distance
- substation resistance
- increased resistance per metre from each substation

Prior to vehicle movement the route gradient, stopping points and electrical network are calculated. The electrical network calculation gives an approximation of the impedance of the network providing varying line voltage as the vehicle approaches and moves away from each substation. This value is then taken by the motor model and a reactive voltage is calculated for use in successive current calculations, resulting in better electrical power simulation, this is used for electrical energy consumption comparison with real data. Figure 12 shows a final flow chart of the STS with all elements regarding the simulation of the class 508.

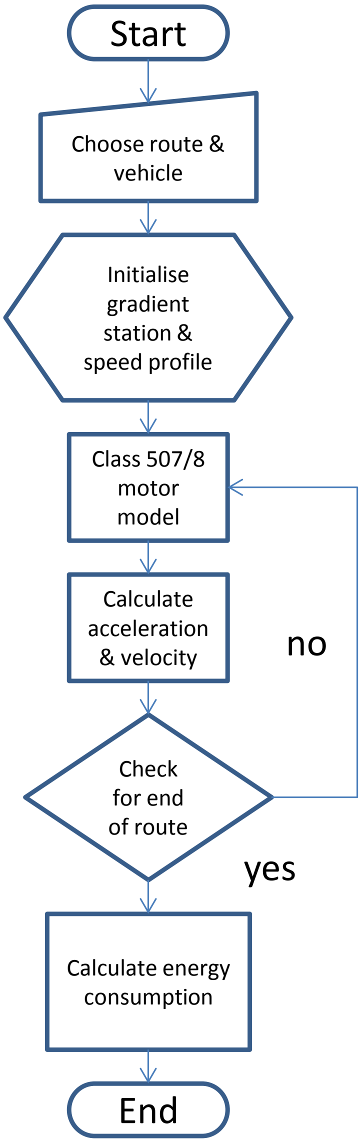

Start
Choose route & vehicle
Initialise gradient station & speed profile
Class 507/8 motor model
Calculate acceleration & velocity
no
Check for end of route
yes
Calculate energy consumption
End

Figure 12 – Flow chart showing an overview of the STS with motor model

3.2.5 Coasting control in the STS

Coasting is a driver technique where the vehicle accelerates to a desired speed, and allowed to decelerate by resistance forces alone; importantly it is unpowered during this period. An example of its use can be demonstrated with a two-vehicle example which is also shown in Figure 13. The no coasting vehicle accelerates to line speed and then cruises until approximately 37 km, the vehicle then brakes hard for the next station stop. The coasting vehicle accelerates to line speed and then cruises until approximately 33 km; this is followed by a period of resistive deceleration, the coasting, and a final braking section when the coasting speed of 10 ms^{-1} is reached. The coasting section demonstrated here is only tens of seconds, but as the traction equipment is unpowered has a significant effect on energy consumption.

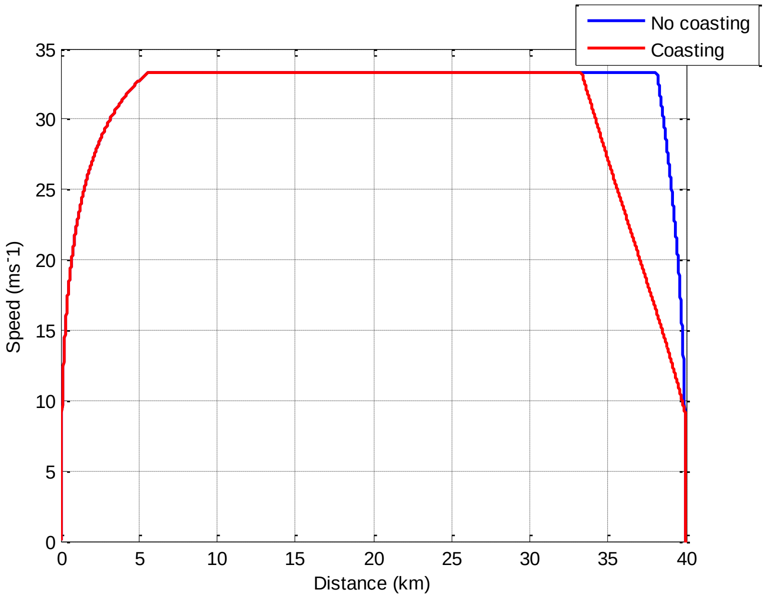


Figure 13 – A simulated trajectory showing coasting and no coasting settings

A coasting setting is built into the STS. Coasting requires the user to set a value of 1 or 0 to enable or disable. If 1 is selected a second choice of coasting speed must be specified. Deceleration via coasting or braking is back calculated from the vehicles next stopping point, as a result coasting speed is the target speed that must be achieved by the vehicle via resistive deceleration. Once the coasting speed is achieved additional deceleration by braking is required to bring the vehicle to a standstill at the next station stop.

3.2.6 Handle input

There are two very general control handle input ‘patterns’ which are used to simulate differing driver approaches on class 507/8 vehicles which are replicated in the STS. Different handle positions are often referred to as notches by drivers. During acceleration a driver may move the control handle from the 0 position (unpowered) straight into position 4 (highest handle position), this produces a rectangular driver input shown in Figure 14, and is described throughout this thesis as ‘full notching’.

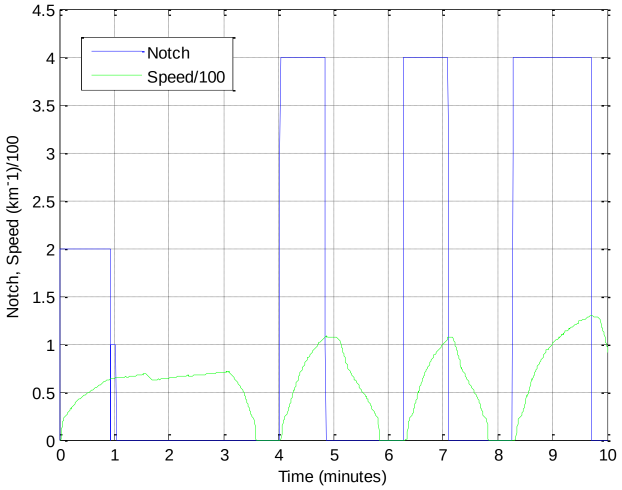


Figure 14 – An example of full notching from in service data

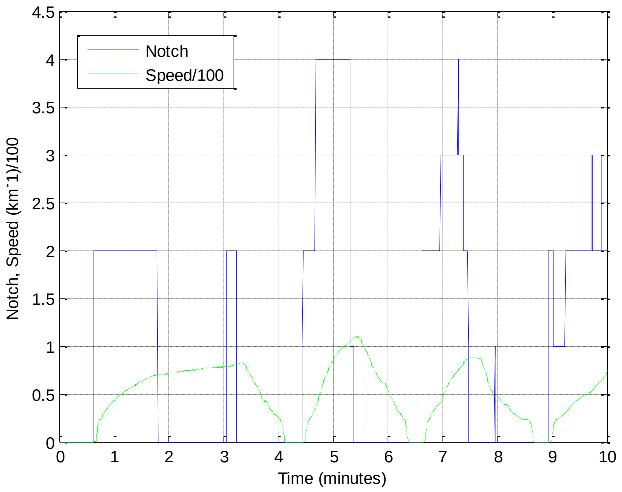


Figure 15 – An example of stepped notching from in service data

A driver may also move through each notch individually this produces a stepped like pattern shown in Figure 15. A detailed description of the operational effects of moving the control handle is presented in 3.2.2. The default operation of the class 508 motor model is to switch between each operational stage as quickly as possible; this produces performance similar to the full notch driver style (full notch model). Stepped notching requires additional functionality of the motor model. To approximate a driver holding the control handle in a particular notch for an extended period a lower current threshold motor variant was designed (stepped notch model). This variant is in effect in a number of points:

- after the final resistor of series operation
- before the change to parallel operation
- after the final resistor of parallel operation
- before the change between each field weakening stage

Figure 16 and Figure 17 show a comparison of resistor switch out current threshold and operational current threshold for both full notched and stepped notch models. The lower current threshold results in a delay in changing between operational stages. Full notching results in higher acceleration, the delay in changes between operation stages in stepped notching results in periods where the vehicle moves at constant speed for longer periods resulting in an overall lower acceleration. Varying acceleration has an effect on energy consumption, during acceleration, full notching will result in a higher energy consumption compared to acceleration via stepped notching. Figure 16, demonstrated the full notching model, however, in reality the model still has a stepped profile. In this case the model has reached stage 4 (weak field) in about 15 seconds, which approximates the actual camshaft resistor change of a full notch driver style closely. Shown in 3.3.2, Figure 25 is an in service data acceleration event showing a full notch and stepped notch input. The stepped notch changes are made at the same instance as the automatic resistor changes resulting in an acceleration performance almost identical to the full notching. This demonstrates that there is little difference between full notching and what could be termed ‘short stepping’. As a result the full notch motor model, although more closely resembling a short stepping control input, is thought to be adequate to demonstrate full notch control. This also suggested that the stepped notch model required longer delays between operational stages that were initially decided on; these were then implemented to demonstrate a difference in performance between models.

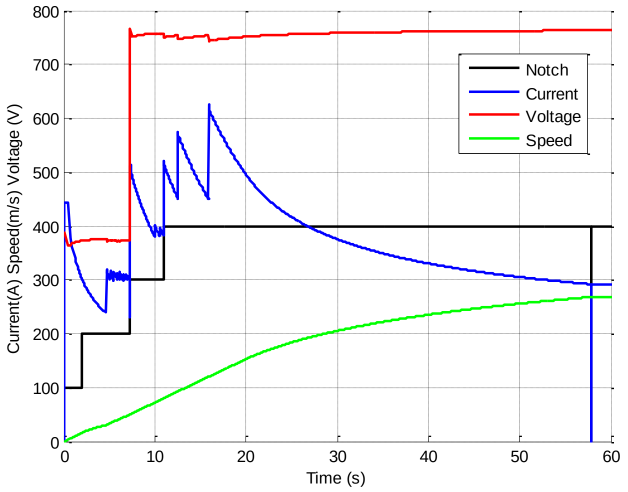


Figure 16 – Full notching in STS Class 507/8 motor model – note stepped notch profile

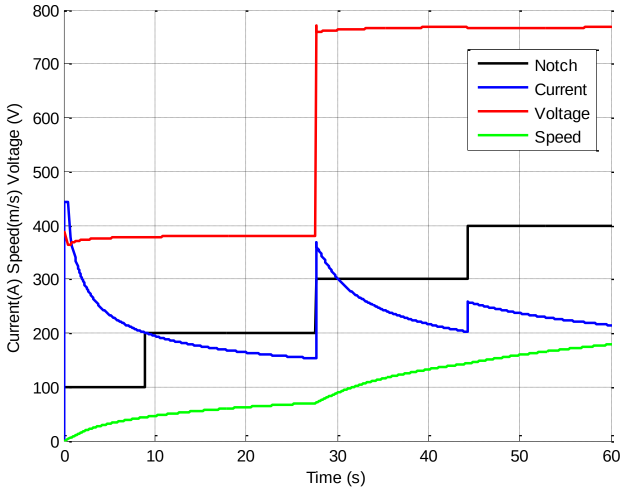


Figure 17 – Stepped notching in STS class 507/8 motor model – note low acceleration

3.3 Observations on a Camshaft controlled DC EMU

3.3.1 Background

The Merseyrail Energy Monitoring Project (MEMP) began in 2011 with monitoring instrumentation being fitted to a Class 508 camshaft operated EMU in service on the Merseyrail network. There are few examples of other similar projects in which detailed electrical data have been collected [71]. As part of the MEMP a British Rail Class 508 camshaft operated EMU was fitted with monitoring instrumentation. All instrumentation was designed and built by the University of Birmingham. The A and B cars have independent monitoring equipment. Instrumentation is housed in equipment bays on either side of the vehicle. Equipment is contained in boxes, named i – iv. Box i contains energy monitoring equipment and connects to voltage and current sensors. Boxes ii, iii and iv are responsible for tacho, driver control signals and bogie monitoring respectively. All boxes are connected using a controller area network (CAN) bus, boxes ii and iv are connected together using a fibre optic connection. Boxes ii and iv provide direction and speed data. Signals from Boxes i – iv are recorded onto solid state storage. Data required for processing are voltage, traction current, speed, handle/notch position, direction and GPS coordinates. Box iii monitors 9 channels and is used to record train control signal. Figure 18 shows a schematic of instrumentation connections.

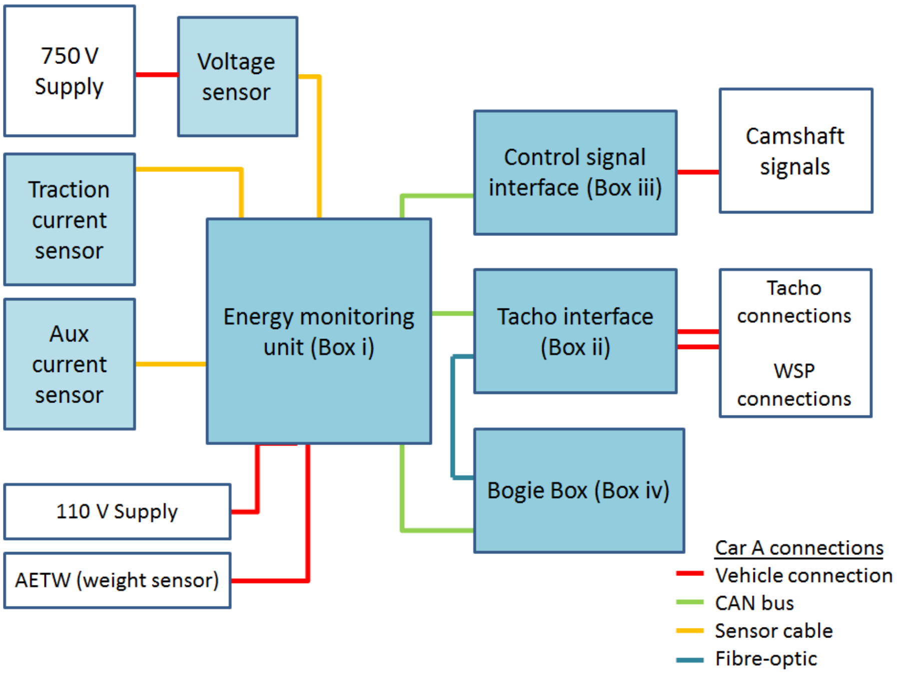


Figure 18 – Monitored train showing instrumentation box location, 'A' car

Table 5 shows recorded signals and sample rate which is the amount of samples taken by each instrument, the data rate is an average of samples per second. The data rate is the actual number of data per second used in the analysis. Although significantly lower than the sample rate the data rate still provides high resolution and is not required to be any higher for the purpose of this analysis.

Table 5 – Signals and data rates for instrumentation

Signal	**Sample rate (Hz)**	**Data rate (Hz)**
Traction current	8192	256
Auxiliary current	8192	256
Line voltage	8192	256
Notch	16	16
Camshaft position	16	16
Tacho	8	8
Passenger weight	1024	1
Temperature	1024	1
Body inertial measurement	1024	16
Bogie inertial measurement	8192	256
GPS position and time	1	1
Traction energy	8192	1
Auxiliary energy	8192	1

3.3.2 Methodology

Journey data from late 2011 were examined. Approximately 470 ‘power-up’ events were recorded, this includes full service runs, shunting and stabling. Significantly more data was available for the Northern Line and Southern Wirral line. The Northern line has three northbound routes, terminating in Southport, Ormskirk and Kirkby. Hunts Cross to Southport is the longest of the Northern line routes and so was chosen to be the focus route. Changes in the direction of travel were used to segment long datasets into shorter individual runs. Using GPS, station stop count and timetable information, 82 runs with total journey time of 61(± 2) minutes, from Hunts Cross to Southport were identified. Simulations discussed in the previous section were used to give comparisons between running time & energy consumption with real data. Dwell time in both simulation and real data was removed to show pure running time only. Figure 19 shows energy consumption vs time for simulated data and real data. A Pareto front or curve was generated using a pre-existing Matlab function. Pareto efficiency, denoted by the Pareto curve, in this case shows the optimal journey time for a specific energy consumption using simulation data output. This curve is used to identify examples of inefficient driving. For example real data points far from the Pareto curve demonstrate examples of wasteful driving, showing high energy consumption (120 – 130 kWh) with longer than expected running times. Initial observations were taken from 23 runs from September 2011. These runs were ranked low to high based on energy consumption and four exemplars were chosen; these being 0%, 33%, and 66% and 100% percentiles (Runs 1 – 4) these runs sit within the main body of data. Control and trajectory for each run are shown in Figure 20, Figure 21, Figure 22, and Figure 23. A further two runs were selected to describe variation in the data that is not consistent with simulations. The total energy consumption and journey time are given in Table 6. As the Merseyrail is a metro network the cruise period is very short meaning that train control during the acceleration and deceleration phases have the greatest influence on energy consumption. Specific driver inputs are then matched to particular high and low energy consumption runs.

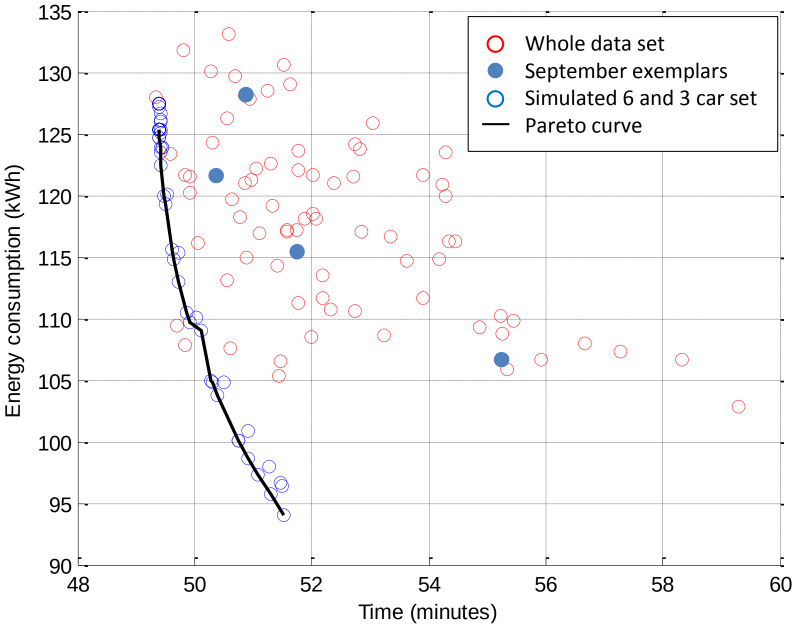


Figure 19 – Simulated and real energy vs time data

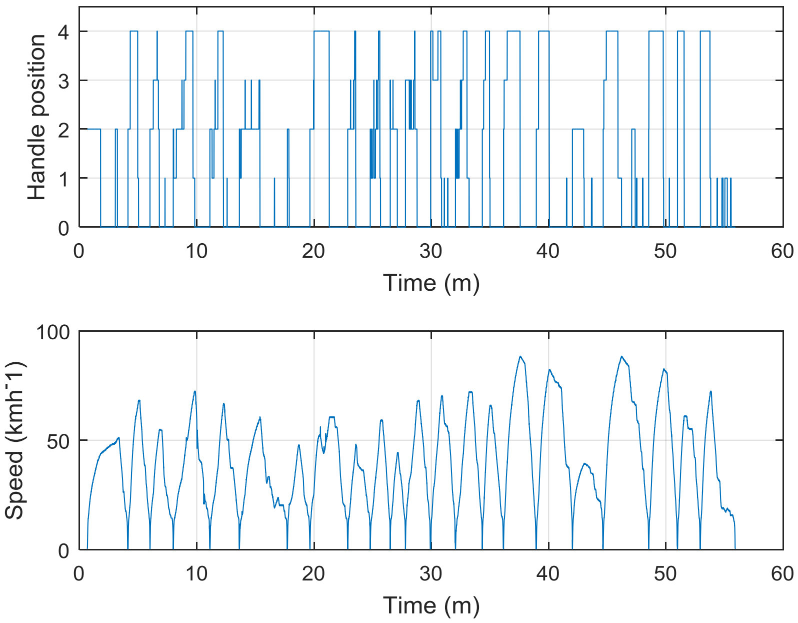


Figure 20 – Control and trajectory for run 1

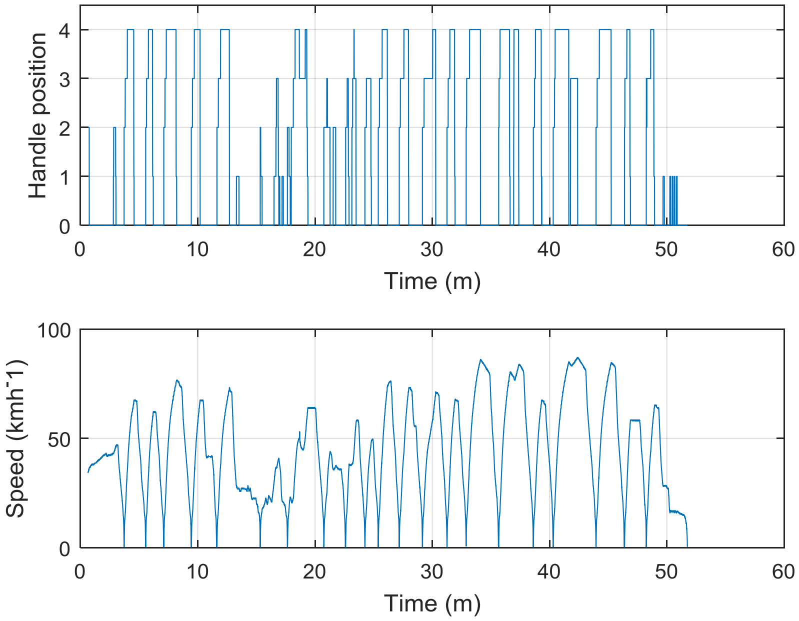


Figure 21 – Control and trajectory for run 2

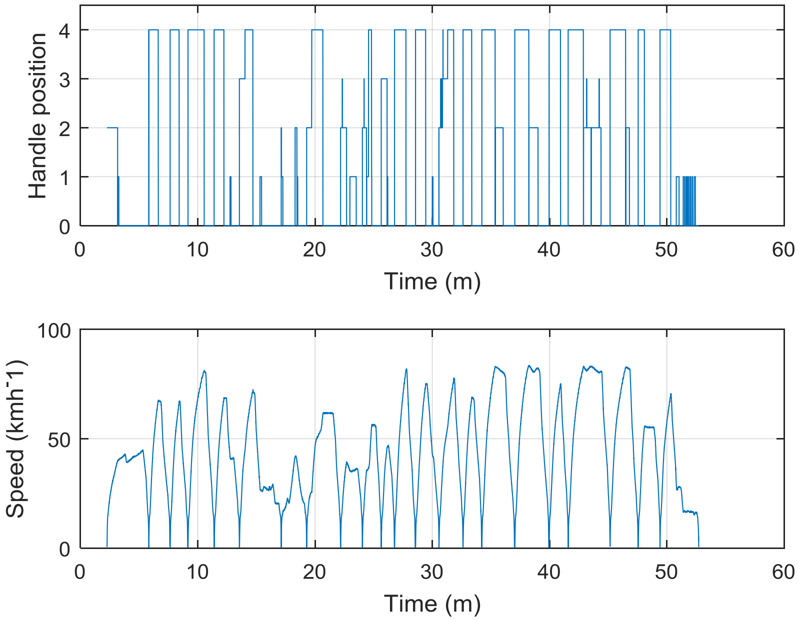


Figure 22 – Control and trajectory for run 3

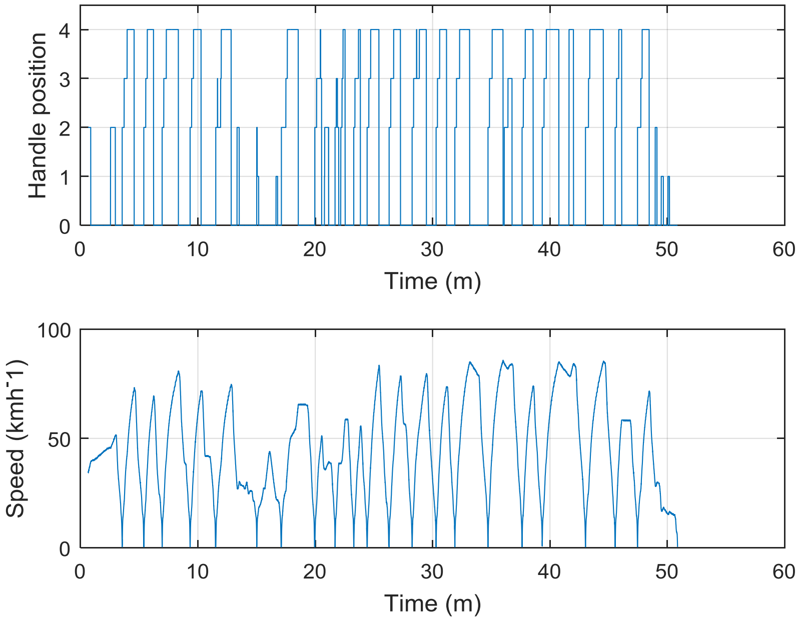


Figure 23 – Control and trajectory for run 4

Table 6 – Energy consumption and comparisons for each Run

Run	Energy (kWh)	% above least	Pure running Time
Run 1	106.6	-	55 m 20 s
Run 2	115.4	8%	51 m 40 s
Run 3	121.5	12%	50 m 30 s
Run 4	128.2	17%	50 m 50 s

Segmented runs were further divided into individual station hops, from one station to the next. The acceleration phase is defined as the first instant notch 1 is selected until the maximum speed is reached. A number of indices were tried using driver control data in the acceleration phase, these included examining the ratios between different notches and calculating the average acceleration for each station hop, and most usefully, identifying the percentage of notch 4 use.

The notching pattern in the acceleration phase signifies the degree of aggressiveness and moderation. An aggressive acceleration driver style has more time and distance spent in notch 4 during the acceleration period. A moderate acceleration has more notch 1 – 3 signals.

Due to safety requirements braking signals were not monitored. During the deceleration phase, whether braking or coasting, there is no traction energy consumption. As a result, a vehicle will use less traction energy if the deceleration phase begins with coasting (but the journey time increases). Coasting sections can be seen in the deceleration phase of each train hop, but they can be difficult to distinguish from braking. As well, during a number of runs, there is a tendency for drivers to employ sections of low-notch, powered deceleration before the braking period. This practice is non-standard and discouraged but common among experienced drivers [60]. Low-notching is seen to provide finer deceleration control as the braking configuration is not used until the latest point; this allows a degree of acceleration recovery if needed. The deceleration phase is characterised by identifying the total unpowered (where the vehicle is still moving) distance as a percentage of the total run distance. The total deceleration distance was calculated for each run by identifying sections of the trajectory which were unpowered (notch 0) with speed greater than 0. Increased unpowered distance denotes earlier deceleration, allowing longer coasting distance. Unpowered distance is reduced by later braking and low-notching.

3.3.3 Results

3.3.3.1 Acceleration

Figure 24 shows the percentage of acceleration phase distances spent in each notch. A greater distance in notch 4 during the acceleration phase indicates higher acceleration rates. Notch 4 use produces the same low to high ranking as energy consumption. As well as the lowest energy consumption, Run 1 has one of the lowest notch 4 use in the dataset. The driver frequently switches back and forth between notch 1 & 2 and 2 & 3 in the acceleration phase, shown in Figure 20, notch 4 use is kept to a minimum; as a result acceleration rate is low. This is a very cautious approach, and is not demonstrated in runs 2 – 4. This may be a response to environmental factors affecting adhesion. A notch 4 use percentage of less than 50% is not common and is considered very moderate. At 72% run 4 has 10% greater notch 4 use than the moderate Run 1 and more aggressive Runs 2 and 3.

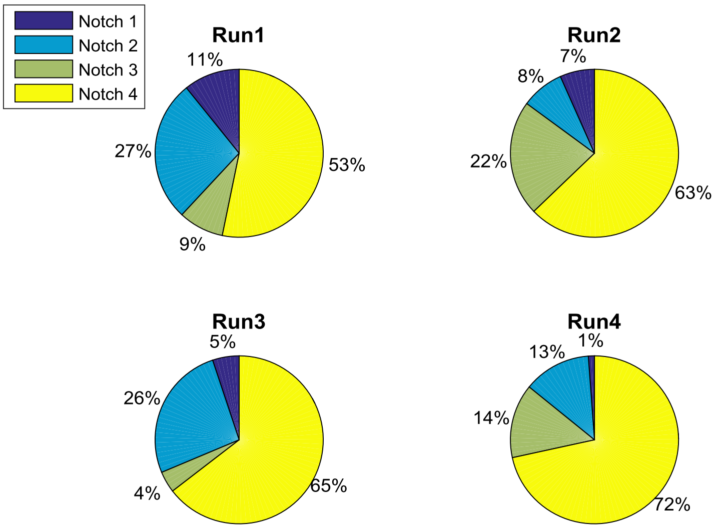


Figure 24 – Control handle position signal count for each notch per run

Examples of stepped and full notch control in the acceleration phase throughout the whole run can be seen in Figure 20, Figure 21, Figure 22 & Figure 23. Higher energy consumption runs coincide with a greater degree of full notch control.

Due to the nature of the camshaft-rheostatic control, varying vehicle performance and energy consumption can be achieved with different driver input. However, there are a number of considerations drivers would need to take into account to be able to benefit from different input. Figure 25 shows a hop between the same two stations from two different runs.

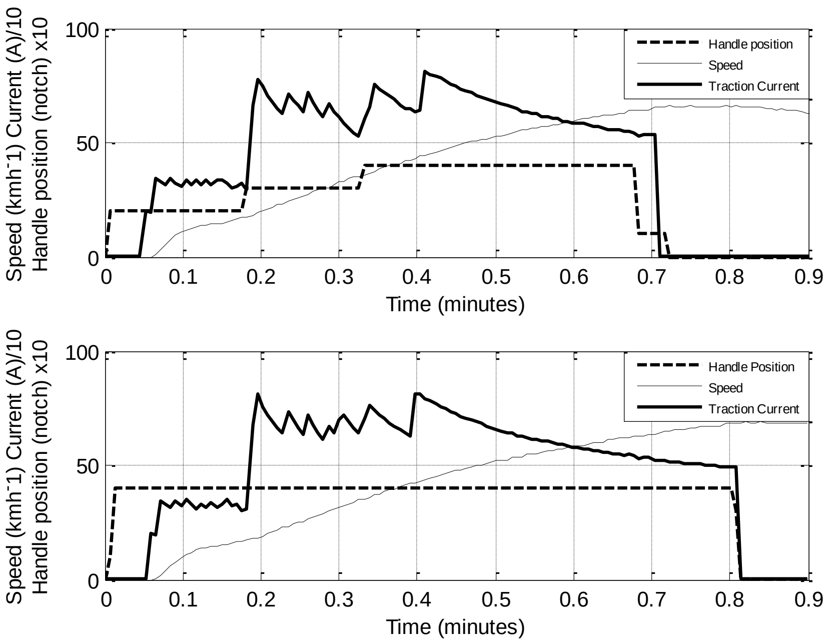


Figure 25 – Comparison of normal stepped and full control

The upper run shows a stepped input, the lower one a full input. The figure shows that although both inputs are different the vehicle performance is very similar. Figure 26 shows the energy consumption for this station hop. This can be explained by the electro-mechanical camshaft operation. If the driver selects notch 4 at the first instant, the camshaft will proceed through each stage automatically, if the driver steps through each notch, but selects the proceeding notch at the instant it would automatically change, as in the upper hop, there is very little difference in vehicle acceleration and energy consumption.

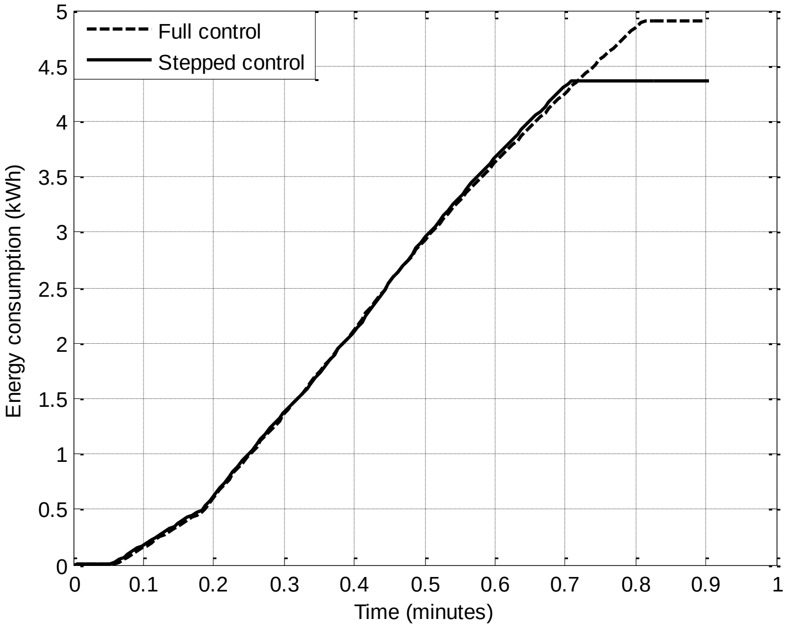


Figure 26 – Energy consumption between normal stepped and full control

Figure 27 shows two further comparisons of the same single station from the same two runs. The upper hop again shows a stepped control profile, in this case the driver holds in notch 2 for approximately one minute whereas the upper hop in Figure 25 shows notch 2 is used for 10 seconds only.

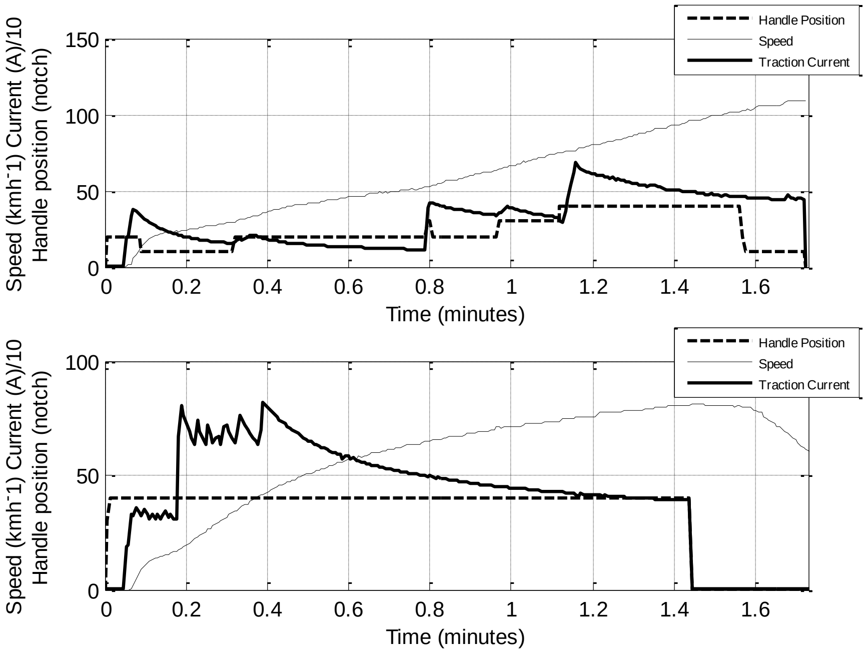


Figure 27 – Comparison of delayed stepped and full control

Resistive losses are reduced as camshaft resistors are removed from the motor circuit meaning the least electrical resistive losses are to be expected, in notch 2, where all resistors are switched out, at the same point in notch 3 and at the vehicles top speed. Moving from series to parallel operation, notch 2 to notch 3, replaces all the operational resistors back into the motor circuit, as a result efficiency drops. However, in the case of Figure 27 as the motor current has been allowed to drop during the notch 2, series operation, the required current for notch 3 is significantly lower, resulting in the camshaft rapidly switching out each resistor to reach the necessary operational current, such that that characteristic saw tooth pattern is no longer noticeable. This is repeated for notch 3 to 4. In each case the motor is allowed to operate in its most efficient condition for each operational stage for longer, as well current is reduced throughout the upper run resulting in a 2 kWh difference in energy consumption but reduced acceleration between 12 to 48 seconds, Figure 28 shows time vs energy consumption for the station hop shown in Figure 27. There is an 18 second difference in the time taken to reach the end of the acceleration phase which over the entire run would begin to affect the timetable.

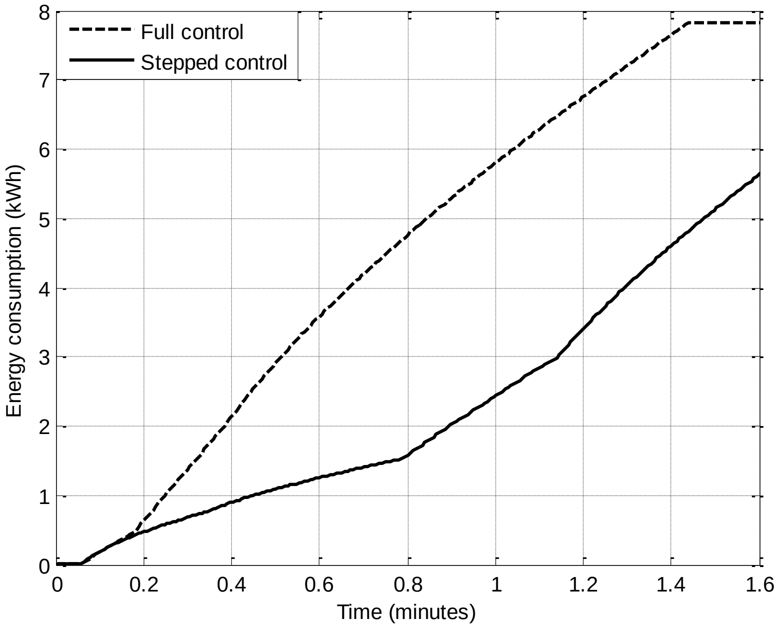


Figure 28 - Energy consumption between delayed stepped and full control

3.3.3.2 Deceleration

Figure 29 shows unpowered distance percentages – i.e. the percentage of the distance between two station stops used for deceleration. A higher percentage indicates a longer and more moderate deceleration phase. Runs 1, 2 and 4 have similar distances. Run 4 has the lowest total braking distance of the 4.

Figure 30 shows the unpowered distance trajectory curve for each run, these sections have no traction power use. Figure 31 isolates a single station hop from Hightown to Formby showing the deceleration phase and coasting distances for each run demonstrating differences in coasting strategy. In this case Run 1 has a relatively long coasting period (60 s), followed by braking. Run 1 briefly low notches after braking, before beginning a smaller second coasting period as this section is briefly powered, it registers as 0 in the braking phase trajectory, it does not represent the vehicle has stopped and restarted. Run 2 has two coasting phases (20 s and 30 s) with a period of acceleration between. This suggests that coasting began slightly early, compared to Run 1, resulting in excessive deceleration. Although a second coasting period is made, the acceleration section between contributes to an increased energy consumption, highlighting the importance of strategic coasting points. Runs 3 and 4 have a similar approach at this point, traction is briefly taken off, signified by the brief unpowered sections, and then returned for the following 40 s.

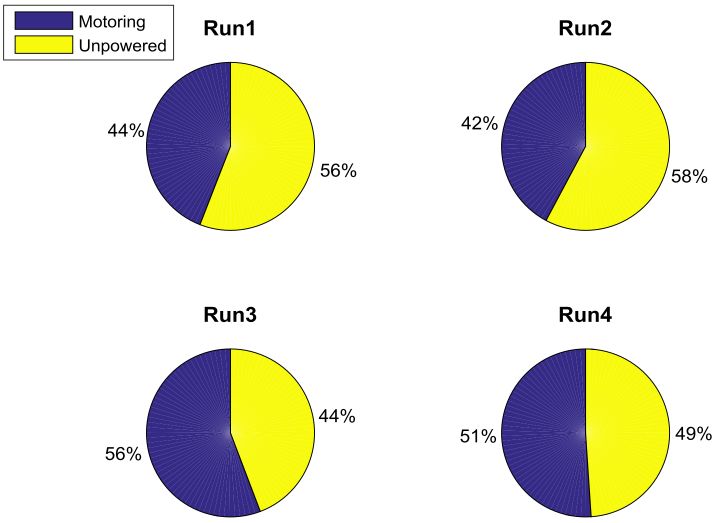


Figure 29 – Percentage of total distance spent braking and coasting

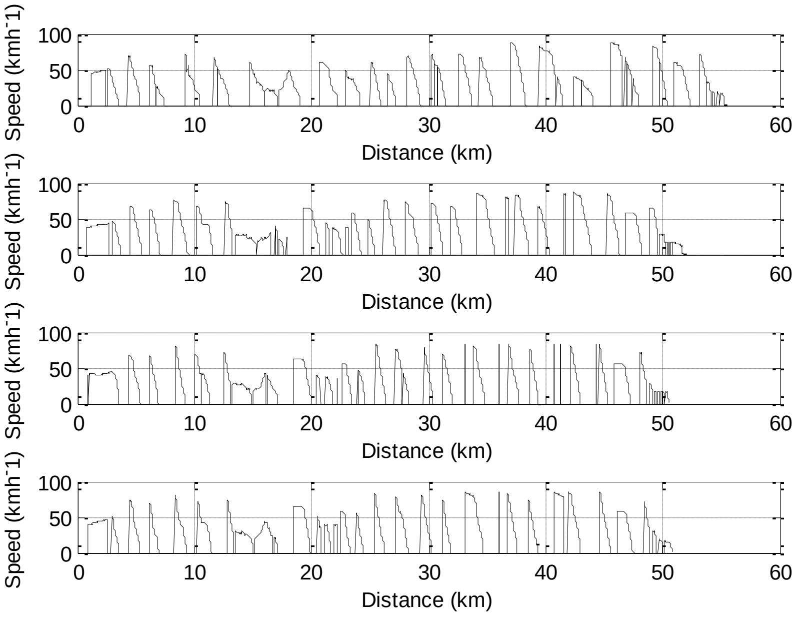


Figure 30 – Deceleration phase trajectory for each run

There are a number of points in each run where traction power is used during the deceleration phase causing a ‘chopped’ trajectory, Figure 30; these are examples of low-notching. As traction power is being used during the deceleration phase energy consumption is increased. It is not always appropriate to coast after the acceleration phase due to gradient effects, however there are numerous opportunities for coasting that are not always taken advantage of. This has been attributed to drivers preferring to reset the cam-shaft by switching to notch 1 during braking in preparation for moving off, as well as driver perception of the dynamic brake. Above 70 mph the dynamic brake is not designed to function, meaning that the friction brake is seen as more reliable during deceleration. At the end of the acceleration phase each run is set to notch 4, Run 1 begins coasting at approximately 80 s into the journey, there is adequate kinetic energy to allow the vehicle to coast a considerable distance, Run 2 begins coasting at approximately 60 s, the driver returns to notch 4, when it is likely that adequate kinetic energy would have been available to continue to coast. Both Runs 3 and 4 lower to notch

2 at the end of the acceleration phase, until they begin braking. Runs 2 – 4 demonstrate low notching. It would be unfair to label the deceleration phases of Runs 2 – 4 as poor examples of driver strategy, however, it is reasonable to suggest that energy savings could be improved by identifying areas where sustained coasting can be employed, as in Run 1, and to encourage drivers to use them. In the past coasting points on the Merseyrail had been used but due to lack of incentive to use them, and the low likelihood of drivers being accountable for not coasting, have become neglected. A particular useful part of this study is that there is now a good opportunity to present real performance data to drivers to aid in evaluation and instruction.

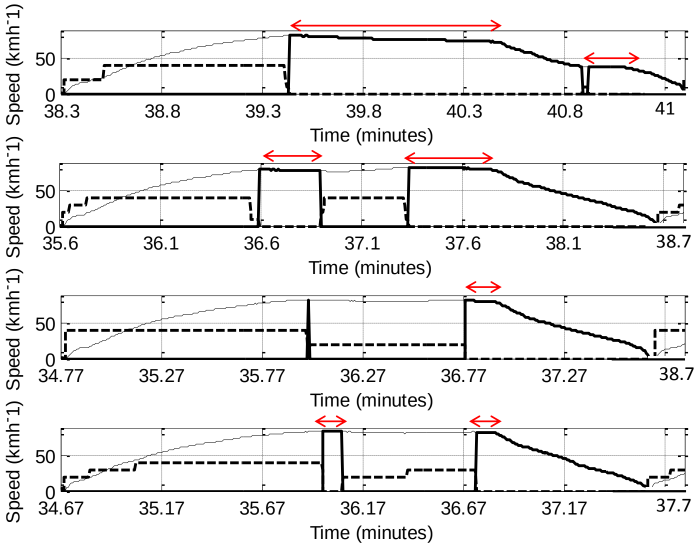


Figure 31 – Comparison of coasting distances for each run (1 – 4 top to bottom) for the station hop Hightown to Formby. Arrows show coasting, dashed line shows control handle position (x10)

3.3.3.3 Comparison runs

Figure 32 shows energy vs time with two additional runs (Runs 6 and 7) highlighted which are comparisons to Runs 1 and 4. Although Run 1 is low energy, it has a running time almost 5 minutes greater than the expected time shown by simulation. Figure 33 shows the control input and deceleration for Run 1 and its comparison, Run 6. Run 1 has a 10% increase in notch 2 use compared to Run 6, demonstrating greater use of delayed stepped control. Run 6's energy consumption is kept at 106 kWh by a slight increase in unpowered deceleration.

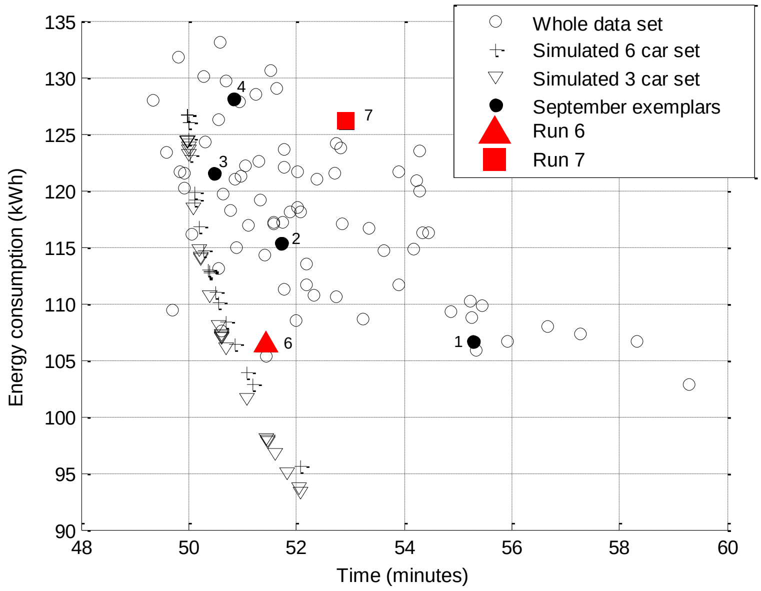


Figure 32 – Energy vs time plot with comparison plots for runs 1 (low energy run 6) and 4 (high energy run 7)

Figure 34 shows Run 4 with its comparison, Run 7. In this case Run 4 is closer to the pareto curve and Run 7 represents a less efficient driver style. As with Run 1 and Run 6, adhering to the pareto curve requires higher acceleration, this is seen with Run 4 which has high notch 4 use, a reduction in unpowered deceleration is also required, this results in Run 4 having one of the lowest unpowered deceleration scores in the whole dataset. As with the Run 1, the longer running time of Run 7 results from decreased notch 4 use and increased delayed stepped control. However, in this case there is a 20 kWh energy increase. This is related to remaining partly in notch 2 but mostly notch 3. As a result, acceleration remains low but energy consumption is significantly increased. Run 7 is a prime example of inefficient driving. It would be better to achieve a faster running time with similar high energy consumption, or a similar running time with lower energy consumption as shown by Runs 4 and 1.

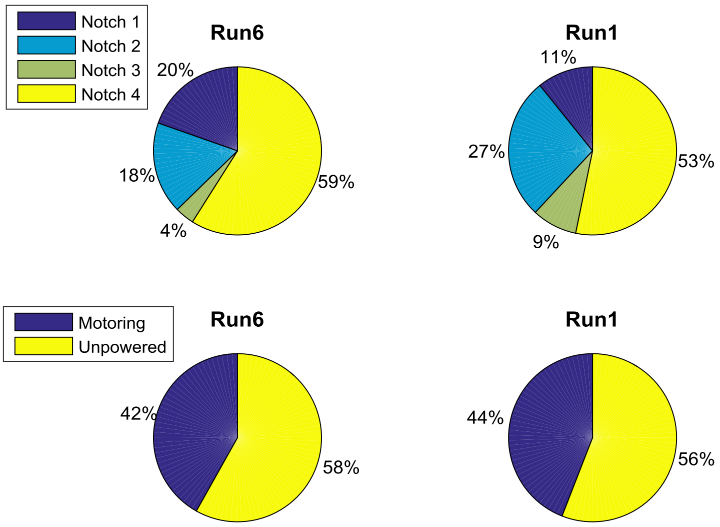


Figure 33 – Control input and unpowered deceleration between Runs 1 and 6

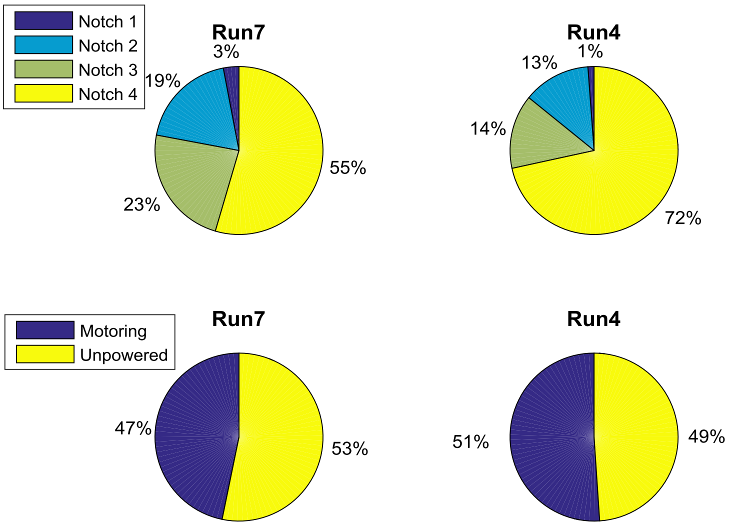


Figure 34 – Control input and unpowered deceleration between Runs 4 and 7

3.3.3.4 Characterisation of driver input

The acceleration and declaration percentage scores (Figure 24 and Figure 29) and energy consumption (Table 6) are able to adequately characterise a driver style in a simple way that can be used for evaluation of a particular journey on the Merseyrail network. Table 7 gives a brief summary of each score and how they relate to the observed driver style characteristics. These scores are used as input data for analysis in chapter 4.6.

Table 7 – A summary of acceleration and deceleration characteristics

Performance score	Description
Percentage of distance in notches	• Increased notch 1 – 3 will describe the effect of delayed notch 2 and 3 use. • Increased notch 2 indicates moderate acceleration with low energy, increase notch 3 indicates moderate acceleration with increased energy consumption. • Increased notch 4 use will describe entering notch 4 at the first instant, indicating a more aggressive acceleration. • Increased notch 4 use also describes longer use of the traction system.
Ratio of distance travelled unpowered : powered	• Indicates a longer distance without the traction system. • Indicates a deceleration phase where low notching is not employed.
Energy consumption	• Energy consumption provides a qualifier for the acceleration and deceleration scores. • Where drivers use normal stepped (10 s or less in notches 2 and 3), energy consumption can be used to demonstrate the similarity to full input (notch 4 at the first instant).

3.4 Conclusion

If drivers begin the acceleration phase from notch 4 or using a stepped approach, increased energy consumption in this phase can be offset by powering off at an earlier point and coasting. The early coasting point however, presents a gamble of sorts for the driver, if the vehicle begins to decelerate too quickly, the traction power will most probably have to be utilised again to allow the train to reach its destination on time, essentially undoing the savings made from early coasting.

Energy consumption reductions can be made in acceleration and deceleration phases. Both methods of energy reduction described here, delayed step control and increased unpowered deceleration (coasting) come at the cost of increasing running time. Observations demonstrate that when coasting is employed the data plot is closer to the expected energy vs running time curve, providing that the acceleration is more aggressive. In service operation where greater headway can be afforded, during off peak hours, means that low acceleration reduction of energy is appropriate. Run 1 begins at 13:20 and so is suggestive of this type of driver behaviour. During peak operation where there are more running time constraints it would be appropriate to use more aggressive acceleration and offset that energy consumption with coasting.

It is possible to characterise these behaviours by taking totals of each driving notch position, pure running time and energy consumption values as ‘scores’ for each particular run. As it is difficult to immediately compare a particular run from these 7 inputs, there is a need to separate and cluster them into groups that have similar performance. Chapter 4 discusses data processing and clustering steps, this is followed by their application to a large dataset in chapter 4.5. Driver styles used on the Merseyrail are then identified and descriptions of their effect on running time and energy consumption are given. A combination of driver styles over a long period of time reflects a sort of attitude or ‘culture’ of driving, this has been termed a ‘driver culture’. Optimisation of ‘driver cultures’ by identifying best performance from large datasets, such as these, have the potential to reduce energy consumption. This is discussed in detail in 5.5, 5.6 and 5.6.4.

4 Classifying simulated driver styles

Chapter 3 introduced preliminary findings taken from the Merseyrail dataset as well as describing aspects of the Northern line simulation. This chapter introduces data pre-processing steps and clustering techniques, the simulator discussed in the previous chapter is used to test the suitability of these.

Grouping synthetic data can validate the suitability of clustering techniques to be used with the real dataset. The findings described in section 3.3 have identified features, taken from a selection of real driver data, which can be used to characterise a particular style of driving. Feature extraction is computationally expensive and involves significant processing time. A large synthetic dataset generated by simulation can be made relatively quickly and allows testing of various analysis approaches. The best approach was selected and used to develop a methodology to find driver styles in the real dataset. The STS was used to generate synthetic data by altering acceleration and deceleration parameters; different parameters approximate varying driver styles. For a general appraisal of driver technique utilised by Merseyrail drivers the author discussed driver strategy with Merseyrail Fleet manager Ian Jones [60]. Figure 35 summarises the basic approaches that were discussed with Ian. These approaches form the basis of the driver styles that were designed for simulation.

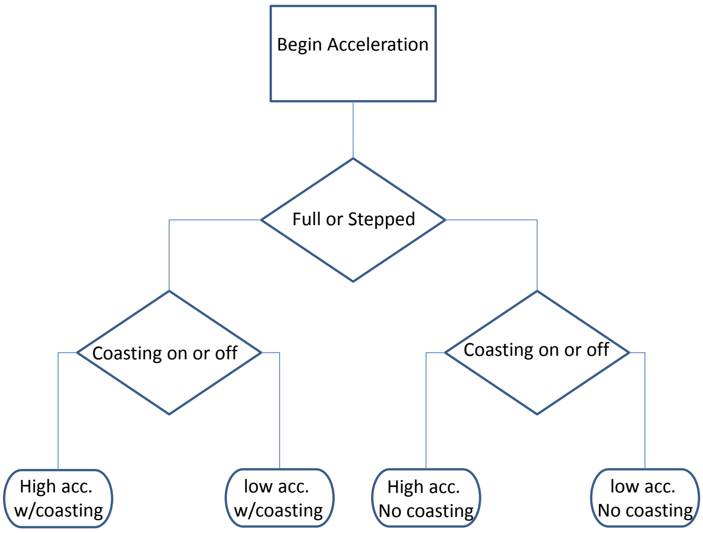


Figure 35 – Basic Simulated driver technique options

4.1 Feature Vectors

To reflect a long period of operation, simulations are repeated multiple times each having varied driver styles. There are approximately 70 services per day dependent on yearly timetable changes and varying weekend operation. The simulation is repeated 1000 times, equivalent to approximately two weeks of operation; driver style is selected randomly. This is performed by the following procedure:

- Coasting to be enabled or disabled selection
  - If enabled coasting speed is selected.
- The standard motor model (full notch) is selected

Or

- The lower current threshold model (stepped notch) is selected,

This produces one of the four driver styles demonstrated in Figure 35. Features identified in section 3.3 are extracted from each simulation and stored in an array to be processed. These features are summarised below:

- Time (s)
- Energy (kWh)
- Distance unpowered deceleration (m)
- Distance in notch 1 (m)
- Distance in notch 2 (m)
- Distance in notch 3 (m)
- Distance in notch 4 (m)

Dwell time differences have a significant effect on real data, for example, two journeys with similar running time could have very different energy consumption this could be a result of:

- A low acceleration low energy driver style where dwell time is reduced to adhere to the timetable.
- A high acceleration high energy driver style where dwell time is increased to adhere to the timetable.

This necessitated removal of dwell time and the use of ‘pure’ running time in real data observations in section 3.3. In the STS dwell time is fixed at 30 seconds meaning that a pure running time calculation is not required as varied dwell time cannot affect overall running time.

4.2 Randomised variation of driver styles

As real data was thought to be highly varied it was necessary to introduce a degree of variance between instances to provide a more rigorous test for data processing and clustering steps.

Braking rate and coasting speed are the first area to be varied. In all cases braking rate is affected, coasting speed is only varied if a driver style with coasting enabled is selected. Both braking rate and coasting have a normally distributed 10% error added to their initial value. Figure 36 shows a selection of trajectories with randomised coasting speeds.

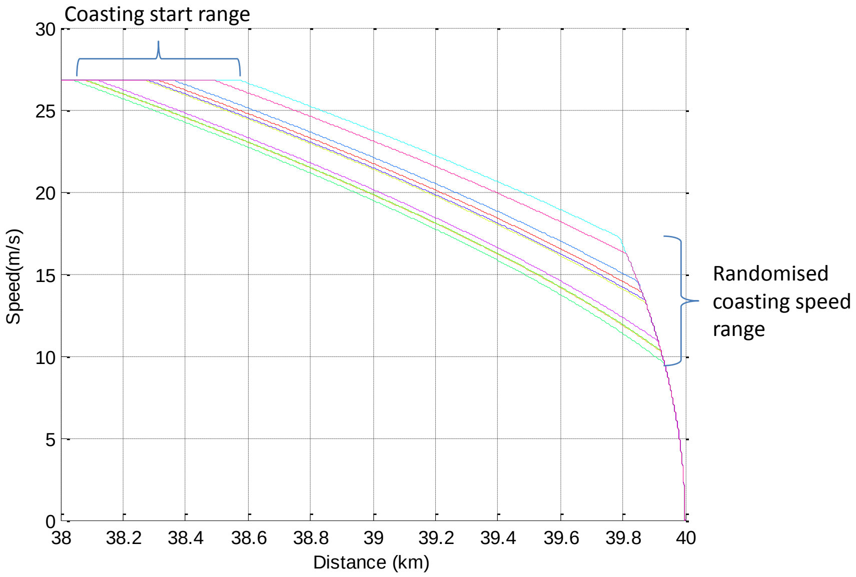


Figure 36 – Trajectories with random variation in coasting point

Current threshold values, determining the point at which operational stages change, are also varied. Again a normally distributed 10% error is added to their normal values. Figure 37 shows a selection of randomised stepped notch trajectories in the time domain which better emphasises the delay in notch change.

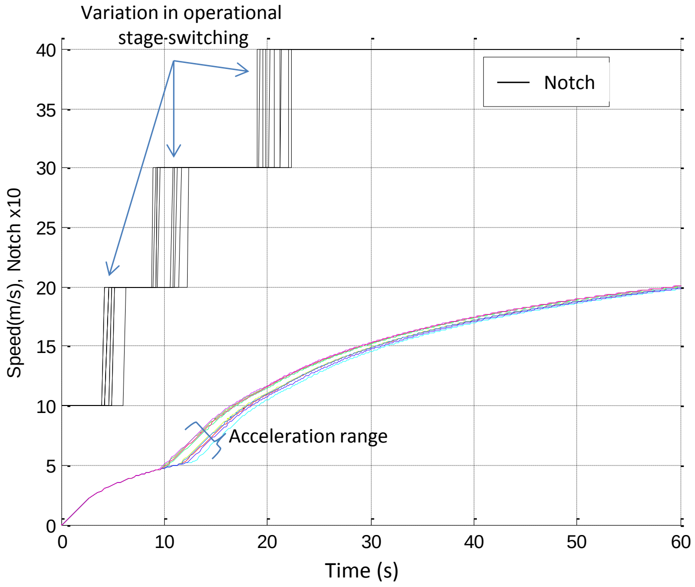


Figure 37 – Trajectories with random variation in operational stage switching in stepped motor model

The stepped motor model has the most opportunity for variation as drivers are expected to have a wide variance in the delay time between notch changes. However, drivers who full notch are expected to have very similar traction system performance meaning that variation in this model is significantly less, as a result it is difficult to discern and a figure and has therefore not been included. Output from simulations using randomised variation provides data which is closer to that which was expected from real data. Figure 38 shows a selection of randomised trajectories using both acceleration models (notched and stepped) as well as coasting selection enabled or disabled.

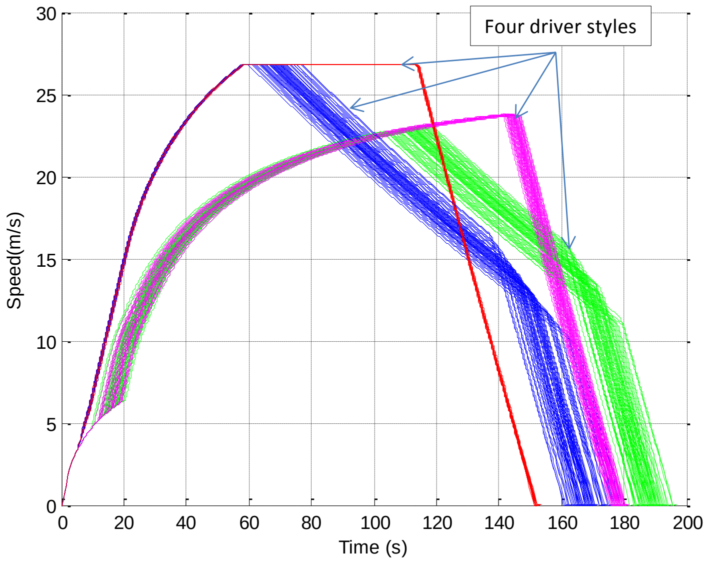


Figure 38 – Randomised trajectories showing four distinct simulated driver styles

The four driver styles are visible in Figure 38 can be described as:

- Blue; high acceleration with coasting.
- Red; high acceleration without coasting.

Both are full notch setting.

- Green; low acceleration with coasting
- Magenta; low acceleration without coasting.

Both are stepped notch setting. Prior to the discussion of simulation results an overview of clustering is given. Simulation clustering results are presented in 4.4.2.

4.3 Data clustering

Data analysis can be thought of as the process of sorting information and then making inferences. Analysis is the stage that makes information useful and as such is the key step in determining data value. Humans instinctively perform analysis on the environment around them, deciding, for example, fruits from vegetables or between colours. When faced with data that is not immediately familiar to the human senses, for example a matrix of numerical data, it becomes increasingly complex for humans to perform analysis and as such computer processing power is required, this is often a result of the speed at which analysis needs to be performed at. Computers, however, lack the instinctive ability to separate data into groups without human supervision. As such there has been an effort by all areas of the scientific and engineering community to develop such methods. Data clustering is a coverall term that describes the unsupervised organisation of data, of various input type, into groups of similar attributes [72]. Data analysis is frequently separated into confirmatory and exploratory; where the term clustering is used data analysis is exclusively considered exploratory [72]. Clustering is the separation of unclassified data meaning that there is often no prior knowledge of the structure of the data and the relationships within. In cases where clustering is used on datasets with known data relationships the discovery of otherwise hidden or unknown trends are normally the target of such analysis [73]. In cases where grouping techniques are used to confirm data relationships the term classification is used to distinguish [73]. Similarity between data clusters is often very difficult to determine, appropriate data preparation, insight into the number of clusters that should be identified and suitability of the clustered output are all the responsibility of the analysist; there is no magic algorithm that produces the perfect result. This means that there is a grey area associated with cluster analysis. This results in the risk of identifying clusters when there may actually be no existing structure to a dataset but also affords the analysist some freedom in interpretation [73].

4.3.1 The *k*-means algorithm

Many clustering algorithms exist but none are considered to be superior to others; algorithms are best suited to different problems. The *k*-means algorithm was chosen to be used in this research for a number of reasons. Firstly, it is not the goal of this research to design new clustering algorithms or to test a number of them to find the best suited. Secondly the k-means algorithm although over 50 years old continues to be one of the most frequently used clustering algorithms [73]. Also many other clustering algorithms are based on this and present only slight variations to either similarity metrics, i.e. data point distances or where to place the centres of individual clusters [72]. The main reasons for the popularity of k-means are its ease of implementation, low computational resource cost, speed of completion and documented empirical success [73, 72, 74]. The algorithm was designed independently by MacQueen [75] and Ball and Hall [76] and can be demonstrated with the following:

A dataset: $X = \{x_i\}$ Where $i = 1 \dots n$

can be separated into K number of clusters:

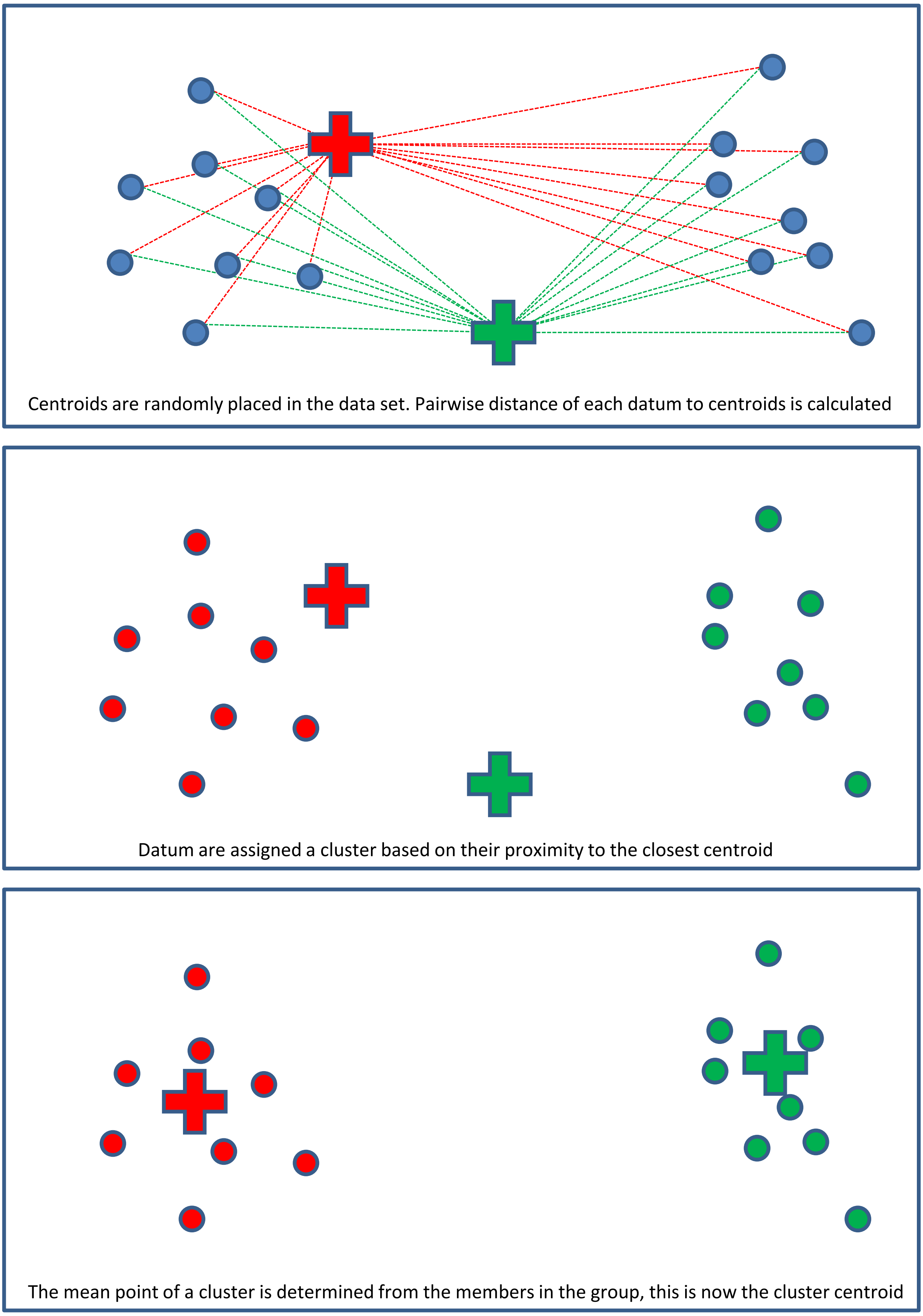
where $C = \{c_k, k = 1 \dots k\}$

When initiated the algorithm seeks to find minimum squared error between cluster mean of c_k (μ_k) and cluster members ($x_{1 \dots n}$) this is defined as:

$$J(c_k) = \sum_{x_i \in c_k} ||x_i - \mu_i||^2 \qquad (12)$$

Figure 39 shows a graphic implementation of a single iteration of the algorithm. Iterations continue until the squared error over all clusters is minimised:

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} ||x_i - \mu_i||^2 \qquad (13)$$



Centroids are randomly placed in the data set. Pairwise distance of each datum to centroids is calculated

Datum are assigned a cluster based on their proximity to the closest centroid

The mean point of a cluster is determined from the members in the group, this is now the cluster centroid

Figure 39 – Single iteration of the k-means algorithm

4.4 Problems with dimensionality

This section presents some of the common problems associated with visualising multivariate data and gives a tutorial on the Principal component analysis. If the reader is familiar with both they may wish to omit this section and continue at section 4.5.

Interpretation of data that can be useful is only possible if data is presented in a meaningful way. Without this the individual is normally reliant on computer software that is capable of assessing data and providing some sort of evaluation. Although there continues to be significant advances in machine learning there is still an important place for human interpretation of data. Where a dataset has large dimensionality it becomes difficult to interpret. To demonstrate, a dataset was synthesised using simple human body and dietary characteristics. These were selected as it is likely that the reader already has an understanding of the trends that are present in this dataset, i.e. taller people are heavier, or, heavier people eat more. A random selection of heights from the range of [1.5m – 1.9m] ± 1% was generated, mass was synthesised from each example using the relationship between height and a healthy BMI of 25, shown in equation 14.

$$BMI = \frac{mass}{height^2} \tag{14}$$

Random daily calorie intake was added from a range of [1000 – 3000] ± 1%, where calorie intake was greater the 1000 kcal but less than 2000 kcal two meals per day were selected, above 2000 kcal three meals per day were selected. Figure 40 shows the first two feature vectors, height and mass, where a linear trend is observable. Represented in two dimensions this trend is easy to interpret. The screen and page are limited to these two dimensions; above three dimensions it becomes increasingly more difficult to identify trends in the data. Figure 41 shows the inclusion of meals per day and demonstrates the increased difficulty in identifying exactly what the graph is showing.

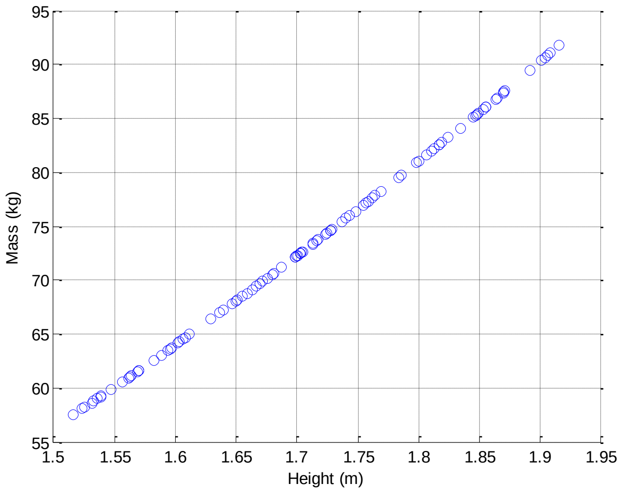


Figure 40 – Trend between height and mass in humans with a BMI of 25

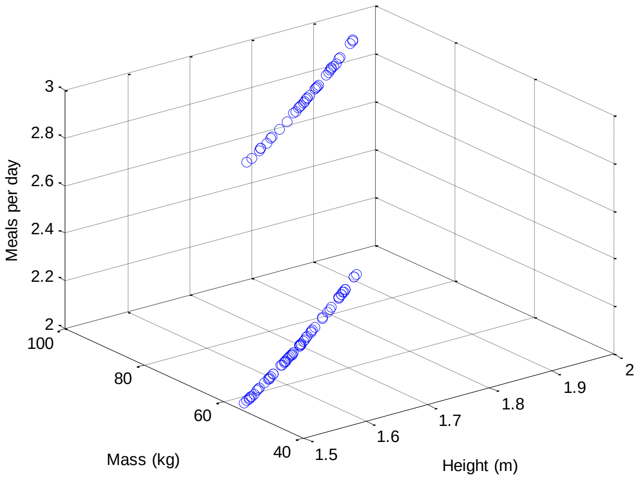


Figure 41 – Trend between height, mass and meals per day in humans with a BMI of 25

If interpretation of data trends were required with dimensionality greater than three the task for the interpreter becomes extremely difficult resulting in more creative visualisation to fully consider higher dimensions. This is shown with the inclusion of marker size and colour in Figure 42 to show journey time to work and daily calorie intake. Figure 42 is difficult to interpret; the relationships of each variable cannot be fully explored with this approach, which is with only five dimensions. If the Merseyrail dataset and STS simulation results are considered there are seven input variables, the feature vectors described in 3.3 and 4.1, that have to be interpreted to identify meaningful trends within the data. As a result the problem with dimensionality becomes apparent and can hinder data analysis.

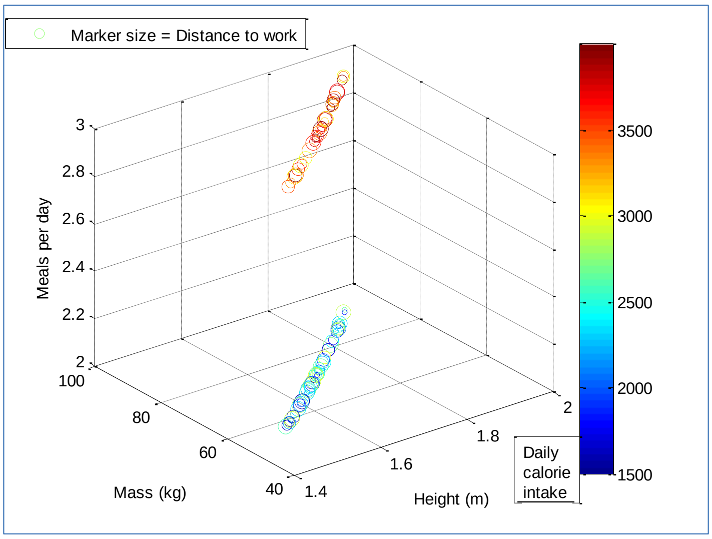


Figure 42 – Trend between height, mass, meals per day, distance travelled to work and daily calorie intake in humans with a BMI of 25

In many other fields where data clustering applications are required, in Big data and gene expression analysis for example, input feature vectors may number in the tens of thousands; this requires additional data processing steps to be taken to allow useful data interpretation. Although this particular dataset has only seven feature vectors this process is very important and relies on a dimensionality reduction step known as eigenvector decomposition and is conducted as a part of principal component analysis

4.4.1 Principal component analysis

Principal component analysis (PCA) is a very common and widely used technique that allows for dimensionality reduction of large datasets. It was developed by Pearson in 1901 [77, 78] and independently by Hotelling in 1933 [79, 78]. While the aim is to reduce dimensionality of multivariate data the technique preserves the structure or relevant information. Like k-means it is an unsupervised learning procedure that is dependent on input data only; PCA aims to maximise variance. Data reduction is very important in two main areas:

- Allows examination of multi-dimensional data in lower dimensions, i.e. a dataset with 20 variables can be plotted in two or three dimensions.
- Data is restructured in terms of its largest variation, i.e. the underlying trends of the data, or the structure are presented to the observer rather than separate inputs that may or may not interact.

PCA also allows for ‘original’ data to be restructured based on the largest variation within that particular dataset. Trends that do not contribute significantly to the overall variance of the dataset can be removed without substantial loss of information meaning that PCA is a form of lossy compression [80, 81]. The key factors of PCA are variance, covariance, eigenvectors and eigenvalues, a short overview of each is given next.

4.4.2 Variance

If one considers a dataset and wishes to identify features of that set, a good starting point is the mean value or the average, this is a very common and perhaps instinctive approach to data analysis, for example, when comparing exam grades the mean is a useful value to assess performance. However, in the case below, the means do not tell a sufficient amount of information about two very different datasets:

$$x = [2\ 2\ 10\ 25\ 5\ 8\ 2\ 9\ 14\ 23]\,, \bar{x} = 10$$

$$y = [10\ 10\ 10\ 10\ ], \bar{y} = 10$$

In this case variance is an appropriate tool to describe the data. Data variation is most commonly presented as standard deviation but variance is also frequently used [82, 81]. In both cases variation is described as the average difference or error to the mean value, in *x* a large variation from the mean and in *y* zero. This gives greater level of detail about the underlying structure of the two datasets. Standard deviation and variance differ only by one term and are shown in equations 15 and 16 [82, 81].

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x - \bar{x})^2}{(n)}} \tag{15}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x - \bar{x})^2}{(n)} \tag{16}$$

4.4.3 Covariance

If a dataset is one dimensional, a list of heights for example, mean and variance alone are adequate to make an interpretation about the data trend, when looking at a dataset with two or more dimensions covariance becomes particularly useful as a method of interpreting data. Covariance provides variation between all input variables, a list of heights compared to a list of masses for example. Covariance of an $m \times n$ matrix produces a square $m \times m$ matrix describing the variation between each input variable with each other input variable and is calculated via Equation 17. The diagonal of the covariance matrix will always compare two like input variables, as a result the variance is being calculated rather than covariance [82, 81].

$$cov = \frac{\sum_{i=1}^{n}(x - \bar{x})(y - \bar{y})}{n} \tag{17}$$

Figure 43 shows the covariance matrix calculated from input variables height and mass using synthetic data in 4.4.

$$\begin{pmatrix} 1.00 & 0.99 \\ 0.99 & 1.00 \end{pmatrix}$$

Figure 43 – Covariance matrix of height and mass

Understanding the covariance matrix is the first step towards making an informed decision on the underlying patterns in the dataset being examined. What is particularly important is the sign of the various entries in the covariance matrix [82, 81]. A positive value indicates a proportional relationship between input variables; a negative indicates an inverse relationship. A value of zero is indicative of no relationship between inputs. In Figure 43 there is a clear proportional relationship between height and mass based on the positive value. Figure 44 shows the covariance matrix in numerical and a colour map form. The colour map is particularly useful as relationships between various inputs can be identified very quickly based on the colour scheme. Variance, where comparison is between like input variables has been re-coloured white to aid visualisation and should be ignored. Strong proportional relationships can be seen between height and mass as well as meals per day and calorie intake. There are also inverse relationships between distance travelled to work and height, and distance travelled to work and weight. This trend is a product of the random selection of values for distance travelled to work. Most other values in the covariance matrix are close to zero and most likely represent noise.

$$\begin{pmatrix} 1.000 & 0.999 & 0.081 & -0.113 & 0.140 \\ 0.999 & 1.000 & 0.079 & -0.117 & 0.136 \\ 0.081 & 0.078 & 1.000 & 0.049 & 0.841 \\ -0.113 & -0.117 & 0.049 & 1.000 & 0.133 \\ 0.140 & 0.136 & 0.841 & 0.133 & 1.000 \end{pmatrix}$$

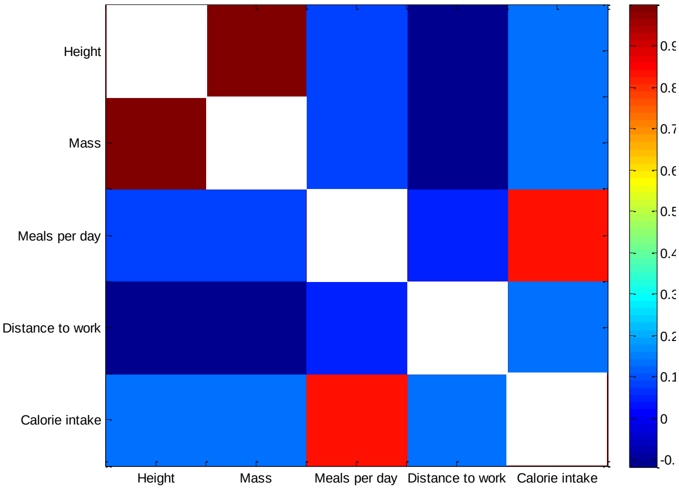


Figure 44 – Covariance matrix numeric (left) colour map (right) for synthetic data

The covariance matrix is useful as it explains the relationships, i.e. their variance, between different input variables. The next step is to be able to assess which particular variables show the greatest variance and which do not contribute to the overall underlying structure of the data. As the covariance matrix is a square matrix eigenvectors and eigenvalues for it can be extracted which allows the presentation of data in terms of principal components.

4.4.4 Eigenvectors and eigenvalues

Eigenvectors are a characteristic of square matrices where the eigenvector v will undergo a scalar transformation only when multiplied by the transformation matrix A, in this case the covariance matrix of the data. An eigenvector has the property as shown in equation 18 [81].

$$Av = \lambda v \qquad (18)$$

From equation 15 it is shown that the transformation of v by the covariance matrix A is equal to a scalar multiplication by λ which is the associated eigenvalue of the eigenvector v, this demonstrates that the vector v will no longer change direction, it will only get longer. The eigenvectors that represent the principal components of the matrix are those that have the largest associated eigenvalues. The eigenvector with the largest eigenvalue is known as the first principal component (PC1) and for a $d \times d$ matrix there will be d principal components. Eigenvectors describe the vector of the largest variance in the original dataset, the corresponding eigenvalue shows the variation across that particular eigenvector, as a result PC1 – PC*d,* demonstrate the most influential to least influential trends, or, the underlying structure which is present in the data. Eigenvectors are calculated by first identifying eigenvalues and solving the determinant of equation 18, shown in Equation 19 [83].

1
$$det(\Sigma - \lambda I) = 0 \tag{19}$$

2
$$det\begin{pmatrix}1.00 - \lambda \; 0.99 \\ 0.99 \; 1.00 - \lambda\end{pmatrix} = (1 - \lambda)(1 - \lambda) - (0.99)(0.99)$$
$$= \lambda^2 - 2\lambda + 0.02$$

3
$$\{\lambda_1 \lambda_2\} = \frac{1}{2}\left(2 \pm \sqrt{2^2 - 4 \times 0.02}\right)\{0.01 \; 1.99\}$$

4
$$\Sigma \, e_i = \lambda \, e_i$$

5
$$\begin{cases}1e_1 + 0.99e_2 = 1.99e_1 \\ 0.99e_1 + 1e_2 = 1.99e_2\end{cases} \; that \; is \; \begin{cases}0.99e_2 + 0.99e_1 = 0 \\ 0.99e_1 - 0.99e_2 = 0\end{cases} \; e_1 = e_2$$

6
$$e_i = \begin{bmatrix}1 \\ 1\end{bmatrix}$$
$$unit \; length \; l_i = \sqrt{1^2 + 1^2} = 1.41$$
$$e_i = \begin{bmatrix}1/1.41 \\ 1/1.41\end{bmatrix} = \begin{bmatrix}0.71 \\ 0.71\end{bmatrix}$$

7 Step 5 is repeated for the second eigenvalue 0.01

9
$$e_i = \begin{bmatrix}0.71 & -0.71 \\ 0.71 & 0.71\end{bmatrix} \; with \; eigenvalues \; \{1.99 \; 0.01\}$$

Where Σ is the 2 dimensional synthetic data covariance matrix shown in Figure 43.

The largest eigenvalue in the example above is 1.99; as eigenvalues describe the variance of the associated eigenvector it can be shown that by summing the eigenvalues and dividing each by the sum PC1 has 99.5% of the data variance and PC2 has only 0.5% of the variance. This indicates the underlying structure of the data is explained in the relationship of PC1 alone. PC2 can be removed and the original data transformed by the PC1 vector to provide a new one dimensional dataset, this is displayed in Figure 45.

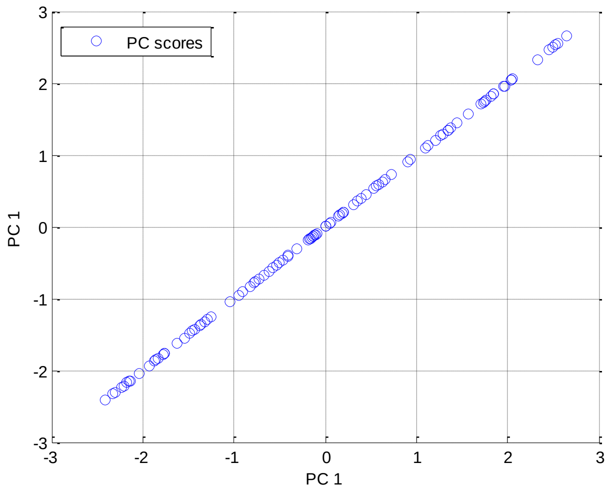


Figure 45 – PC scores generated from PC1 for height and weight data

In Figure 45 the eigenvector matrix of PC1 is multiplied by the original zero mean height and weight data (equation 18), this performs a linear transformation of the original two dimensional data to give a single dimensional array; individual data points are now in terms of their variation rather than their original value, these are termed principal component scores (PC scores).

$$PC\ scores = a_{data} \times \begin{bmatrix} 0.71 \\ 0.71 \end{bmatrix} \tag{20}$$

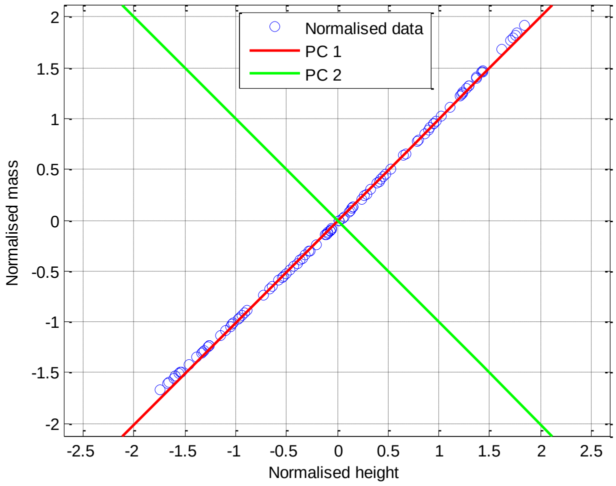


Figure 46 – Normalised data plot with PC1 and 2 vectors plotted

Figure 46 shows the normalised data plotted with eigenvectors that denote PC1 and 2. This is perhaps the best example showing how principal components clearly describe the underlying trend in datasets. PC2 is orthogonal to PC1, this is a feature that all successive principal components display, each new eigenvector is orthogonal to the last one, as a result successive PCs are no longer correlated and can be considered as unrelated to the previous data trend. In the case above the second trend, PC2, is extremely small and can be explained as a result of noise obtained during the generation of the dataset. The procedure shown in Equation 17 is applied to the whole covariance matrix shown in Figure 44. Eigenvalue, or variance, contributions from each principal component are shown as percentages in Figure 47. PC1 and 2 contain almost 80% of the variance in this particular dataset, as a result PCs 3 and 4 can be discarded and the PC scores can be generated based on PC1 and 2 alone.

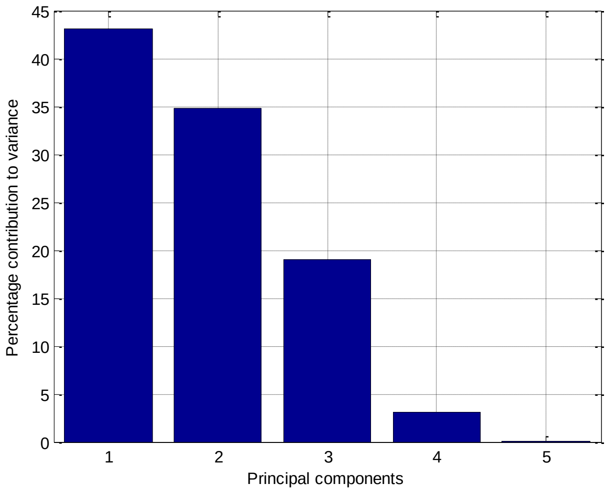


Figure 47 – Percentage contribution of each PC for synthetic data

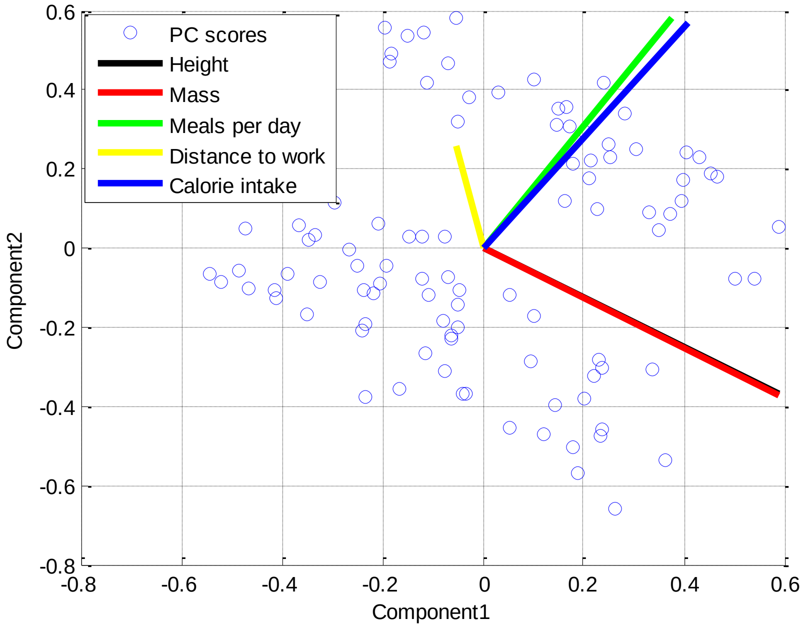


Figure 48 – A biplot of principal components and their generated PC scores

Figure 48 shows the PC scores plotted on PC1 and 2 in a biplot. This figure allows visualisation of the five inputs in two dimensions only. The figure shows a number of relationships in the data.

4.4.4.1 PC1

1. Height and mass for a BMI of 25 increases proportionally, this is the most significant trend on the horizontal axis.
2. Both meals per day and calorie intake also increase proportionally with height and mass although to a lesser degree.
3. Distance travelled to work has a very small horizontal score that is inversely proportional to the rest, suggesting that it has no significant relationship to the other inputs, this is due to it being randomly generated data, and would likely also reflect real data as this parameter does not generally affect body size and food consumption, unless, for example, all members of this dataset travelled by bicycle or walked.

4.4.4.2 PC2

1. Meals per day and calorie intake are the most significant trend in PC2, both increase proportionally on the vertical axis.
2. Once again height and mass increase proportionally but interestingly in this component both are inversely proportional to meals per day and calorie intake. This suggests there are two underlying structures to the data, the first accounting for 43% of the variance suggests that larger people tend to eat the most. The second accounting for 35% of the variance suggests that some smaller people are also eating a lot. In this case these trends are most likely a result of arbitrary selection of meals per day rather than any meaningful trend, but equally if this type of analysis were performed on a workforce or school population this result could suggest a hidden trend where food consumption is not totally based on size needs, this could then be used to identify problems such as overeating, spending habits, and other dynamics present in a large dataset.

3. Once again distance travelled has little to no impact on the dynamic of this dataset, we can now consider distance travelled as essentially redundant and it could possibly be removed from a future dataset reducing dimensionality before eigenvector decomposition is required.

Although the steps to identifying eigenvectors and eigenvalues are not difficult they become more complex and time consuming when calculated manually with square matrices greater than 2 × 2. As a result PCA is performed using existing Matlab functions.

4.5 Simulated data PCA and clustering results

A large dataset was generated using the STS. Driver style was varied as described in section 4.2; simulations use the Hunts Cross to Southport route file. Principal component analysis is applied to the dataset and followed by clustering algorithms. Figure 50 shows the covariance matrix, numeric and colour map, of the 1000 × 7 dataset, a sample taken from the dataset is shown in Table 8. From the covariance matrix it is possible to identify relationships between increased energy and notch 4 use, as well as increased time with increased notch 1 – 3 use, signifying reduced acceleration impacting on journey time and energy consumption. Interestingly the covariance matrix also suggests that increased coasting has only a small relationship with journey time, but does result in energy savings. At this point it could be suggested that even though the energy savings are less than those gained by lower acceleration, there is greater benefit to coasting due to the smaller impact on journey time.

$$\begin{bmatrix} 1.00 & -0.95 & 0.22 & 0.83 & 0.83 & 0.83 & -0.39 \\ -0.96 & 1.00 & -0.47 & -0.65 & -0.66 & -0.66 & 0.62 \\ 0.22 & -0.47 & 1.00 & -0.35 & -0.34 & -0.34 & -0.98 \\ 0.83 & -0.65 & -0.35 & 1.00 & 0.98 & 0.99 & 0.18 \\ 0.83 & -0.66 & -0.34 & 0.98 & 1.00 & 0.98 & 0.16 \\ 0.83 & -0.66 & -0.34 & 0.99 & 0.98 & 1.00 & 0.17 \\ -0.39 & 0.62 & -0.98 & 0.18 & 0.16 & 0.17 & 1.00 \end{bmatrix}$$

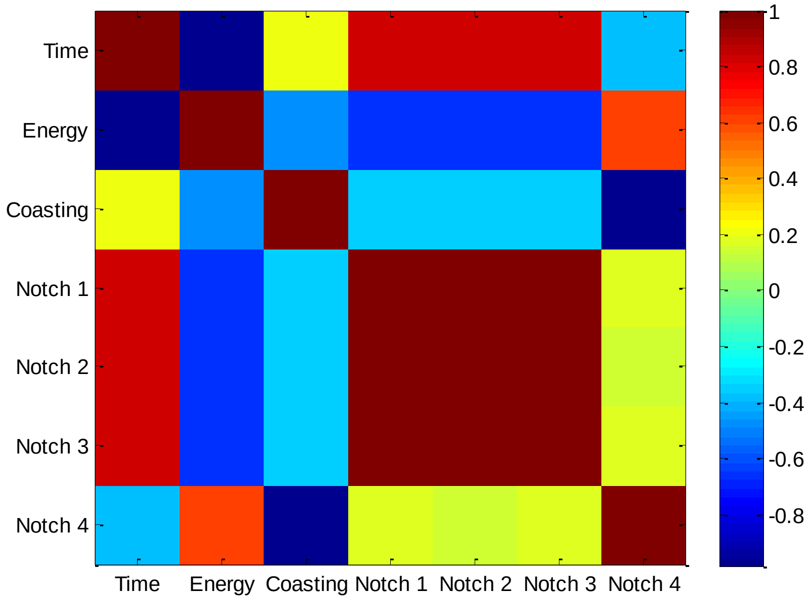


Figure 49 – Covariance matrix of normalised STS simulation data, numeric (above) colour map (below)

Table 8 – A sample of STS simulation output data

Time (min)	**Energy (kWh)**	**Coasting (m)**	**Notch 1 (m)**	**Notch 2 (m)**	**Notch 3 (m)**	**Notch 4 (m)**
56.0	116.7	6932	342	911	2171	31505
56.2	116.5	6942	365	969	2159	31426
57.6	99.9	15902	352	953	2139	22515
56.2	116.4	6944	327	1035	2201	31354

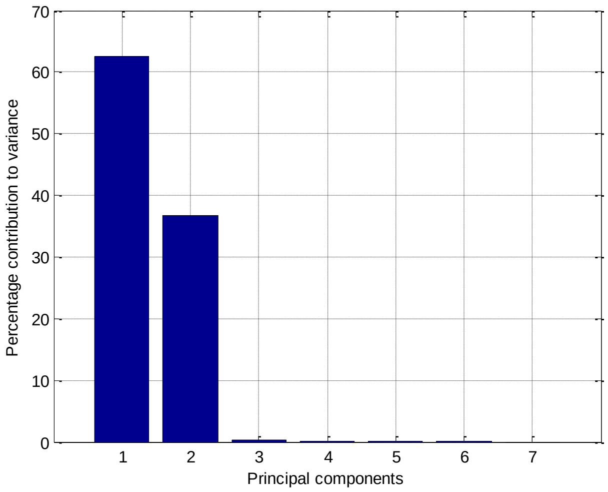


Figure 50 – Percentage contribution of each PC for STS simulation data

Eigenvalues are presented as percentages in Figure 50, showing the greatest variance is contained in the first two principal components, 62% and 36% respectively, as a result PCs 3 to 7 are discarded and PC scores are calculated using the PCs 1 and 2 only. Figure 51 shows the numeric values and colour maps of the remaining principal components. Matching colours, e.g. red and red or blue and blue show inputs have proportional relationships, opposing colours, e.g. red and blue, show inverse relationships, greens in this case show little or no importance in a particular data trend. As each PC is orthogonal to the previous PC and uncorrelated, inputs that have significance in one PC frequently have substantially less significance in the following PC, showing that their maximum variation has been 'used up' in an earlier trend.

$$\begin{bmatrix} -0.45 & -0.19 \\ 0.40 & 0.34 \\ 0.04 & -0.62 \\ -0.46 & 0.17 \\ -0.46 & 0.16 \\ -0.46 & 0.16 \\ 0.04 & 0.62 \end{bmatrix}$$

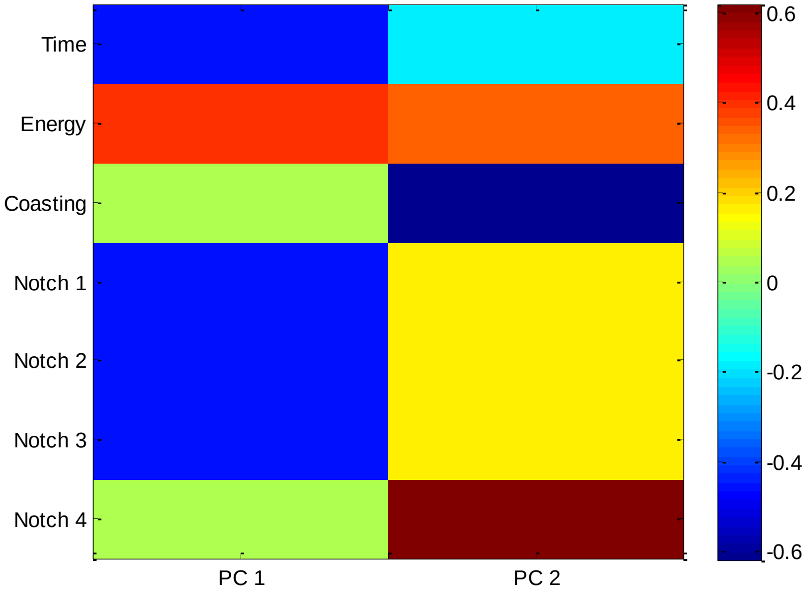


Figure 51 – Principal components, numeric eigenvectors (above), colour map (below)

The dynamic of the simulated data can be described using the colourmap as below.

4.5.1 PC1

1. Time and Notch 1 – 3 are proportionally related, energy is inversely related to these.
2. Coasting and Notch 4 are not significantly involved in PC1
3. These relationships suggest that the greatest effect on energy consumption is produced by increasing Notch 1 – 3 use, i.e. a low acceleration, however this has a significant impact on journey time.

4.5.2 PC2

1. Coasting and Notch 4 are strongly inversely related
2. Time and Energy are again inversely related although their significance in this PC is reduced as the majority of their variance is explained in PC1
3. Notch 2 and 3 have a weak proportional relationship in PC2 resulting from their high variance in PC1, this likely indicates that they are not significantly influential in PC2.
4. PC2 shows that energy consumption savings can be made by coasting, although these savings are less than those resulting from low acceleration; however the use of coasting has a much lower impact on journey time. This suggests that a high acceleration driver style might offset the energy consumption in the acceleration phase somewhat by using coasting. In doing so there is only a small journey time penalty.

4.5.3 Driver style clustering

Figure 52 shows a biplot of the PC scores and principal components, the biplot is shaded to give a general description of a data point present in a particular quadrant based on the relationships identified from the colour map. Colour shading of each quadrant is described below and is matched to the same colours used in described in 4.2. The horizontal axis represents the first principal component and shows the relationships between each input with the largest variance (62%). The vertical axis represents the second principal component and shows the relationships between each input with the second most variance 36%. The lines on the biplot represent the value of each input in the first two principal components. For example, Time (-0.45), notch 3 (-0.46) and energy (0.40) have the highest magnitude on the horizontal axis. This shows that the greatest variance, i.e. the greatest data trend is carried by the relationships of these three variables. On the vertical axis notch 4 (0.62), coasting (-0.62) and energy (0.34) have the highest magnitude. This shows that the greatest data trend is carried by the relationships of these three variables. It is possible to then evaluate the driver style of PC scores (scattered data plots). For example, closely clustered group on the upper right:

- High positive value on the horizontal axis, therefore
    - High energy
    - Low time
    - Low notch 3 use
- High positive value on the vertical axis
    - High energy
    - High notch 4 use
    - Low coasting use

This driver style then can then be matched to the red driver style generated 4.2, the highest energy 'Full notch. High acceleration, highest energy, no coasting, lowest journey time'. The other three clusters are similarly matched to the driver styles which generated them.

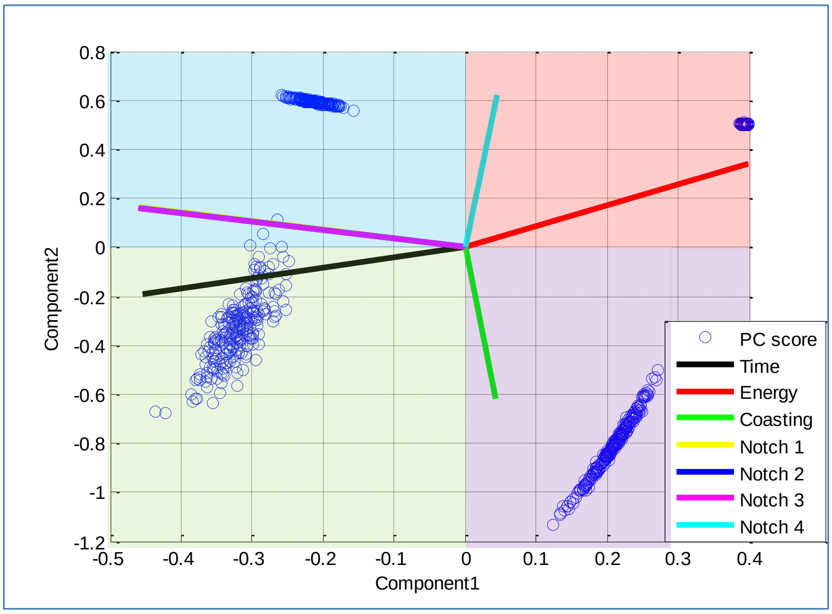


Figure 52 - Biplot showing STS simulation PC scores and principal components

Notch 1 and 2 are obscured beneath Notch 3 due to the similarity of PC1 and 2 inputs.

- Green – Stepped notch. Low acceleration, lowest energy, coasting used highest journey time
- Magenta – Stepped notch. Low acceleration, low energy, no coasting high journey time
- Blue – Full notch. High acceleration, mid energy, coasting used, mid journey time
- Red – Full notch. High acceleration, highest energy, no coasting, lowest journey time.

PC scores are then grouped using the k-means clustering algorithm, Figure 52 shows the result of classification by the k-means algorithm. Validation of the value for *k* (number of clusters) in this case is simple, the most important factor being that *k* is known *apriori i.e.* four driver styles were designed for the test. Also the analysist is able to visualise four distinct clusters with large separation between each, in this case it is quite simple to agree with the cluster results provided by the algorithm. At this stage it was not known whether the real data would also clearly cluster into as well defined groups, it was thought that a degree of drift between different driver styles would be present. This introduced the need for a range of cluster validation tools that would provide both an optimal value for *k* to use with real datasets. Colours represent the same driver styles that were designed during simulation, centroid (cross) numbers, are unrelated to a particular driver styles as clusters are randomly initialised successive runs of k-means would have a significant chance of providing the same clusters but cluster numbers are often different. This is especially evident with real data clustering and as such cluster numbers should not be compared between datasets, cluster 1 in a Hunts Cross to Southport may be very different to a cluster 1 from the return journey dataset; real data clustering is performed and considered separately for each dataset examined.

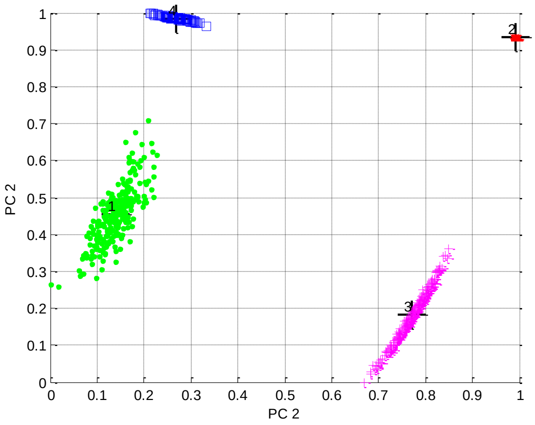


Figure 53 – K-means clustering of STS output for synthesised varied driver style, colours relate to the same driver styles shown in 4.2, Figure 38

4.6 Conclusion

Chapter 4 has shown the visualisation problems that are associated with multivariate data and introduced PCA as a data processing step that reduces dataset dimensionality. PCA produces a new dataset that expresses the greatest trends; dimensionality reduction allows these to be viewed on two or three axes which aids visualisation. PCA was applied to a synthetic dataset that was generated via simulation of the Northern Line on the Merseyrail network. The 7-dimensional dataset was decomposed into two principal components which express approximately 90% of the data variance (≈ 60% PC1 ≈ 30% PC2). PC 1 shows the inversely proportional relationship between energy consumption and time, there is also a strong relationship between increased notch 3 use and reduced energy consumption. In PC 2, Notch 4 is shown to have a strong relationship with increased energy consumption. Four distinct groups, in the PC scores of the synthetic dataset, are shown in Figure 52; these were clustered using the k-means algorithm. The quadrant where each group is situated represents a particular driving style which matches to the varied driver styles that were randomly selected during simulation. PCA and k-means have been shown to be able to independently separate and cluster data into the original varied driver style inputs validating their use in the driver style identification methodology. The following chapter details the use of the methodology on the real dataset.

5 Identification of a 'Driver Culture' and its effect on energy consumption on a DC rail network

The next stage in this project is to implement the methodology described in chapter 3.3 and chapter 4.5 on real datasets taken from multiple routes. This requires identification of driver styles that are in use followed by an extrapolation of their long term effect on energy consumption. A distribution of driver styles reflects the collective performance of a driver population and has been termed a 'driver culture'. Driver culture for each route analysed is presented.

5.1 Background

A long term goal of this project, as discussed with Merseyrail energy monitoring personal as well as driver and fleet managers, was to identify areas where energy savings could be implemented and to suggest steps to optimise driver strategy. The impacts of driver cultures are presented as potential energy savings that might be possible if implemented on the Merseyrail network.

5.2 Methodology

Six separate datasets were taken from the Merseyrail network for three routes where the instrumented vehicle was in operation. Table 9 shows each route with length, single journey duration, station stops and number of runs contained in each dataset. Data is separated using the same processing steps as described in 3.3.

Table 9 – Route information

Route	Length	Single journey duration	Number of stations stops	Number of services per day	Number of runs in the dataset
Hunts Cross to Southport	38 km	60	23	70	169
Liverpool to Ormskirk	18.5 km	30	12	60	319
Liverpool to West Kirby	28 km	34	15	60	14

A master map for each route is identified with its particular trajectory profile and used to match each run examined to the correct route. Station number is also used to confirm runs are matched to the correct route. Matched routes are also separated for direction of travel and the analysis is conducted separately for both directions. The instrumented vehicle is in operation for the majority of the time on the Northern Line in the three month period examined and as a result has substantially more data to use for analysis; data for the Wirral Line is limited.

For each route and direction PCA is conducted followed by driver style clustering. As discussed in section 4.4 visualization of 7 dimensional feature vectors are problematic, principal component analysis is used to reduce dimensionality as well as to identify the strongest variation within the data. Covariance matrix eigenvectors and eigenvalues of each dataset are calculated. A total variance of greater than 80% is usually considered enough to explain the majority of variance in a dataset, in a number of following cases three PCs are required to describe the variation in the datasets, however, it was decided that when trends do not have a significant influence on either journey time or energy consumption they relate to trends that have little use in the analysis and have been removed. This allows data to be presented in mostly in two or three dimensions but introduces information loss; approximately 30 – 35% when two dimensions are used and 20 – 25% with 3. As this data is not significantly related to areas of interest (energy and time) it is considered an acceptable loss of information. It is thought that with significantly larger datasets, such as the 1000 data points used in simulation, first and second PCS would show significantly greater variance (approximately 70 – 80%) but due to data focused on three months with only one instrumented train there is a limit on how many journeys are available for analysis.

5.3 Data decomposition

The following section shows eigenvector decomposition of datasets from each of the chosen routes and a description of the first two or three principal components shown in their various biplots and colour maps. There is a degree of similarity between the four large datasets; the small datasets, Liverpool to West Kirby and its return journey, are significantly different. This is thought to be largely due to the lack of data available on these lines. As a result in the following clustering validation and results sections both small datasets are no longer included.

The following sections frequently refer to notch position and the various effects on journey time and energy consumption, as a result a short recap on each notch and its function is given below. For a detailed description of notches and operation see chapter 3, section 3.2.2.

- Notch 0 – Traction system unpowered, resistors remain in position of preceding notch setting.
- Notch 1 – Camshaft reset – all resistors connected in series. Used for shunting
- Notch 2 – Series operation. Eight stages of resistor removal.
- Notch 3 – Parallel operation. Four stages of resistor removal.
- Notch 4 – Weak field operation. Three stages of field weakening.

5.3.1 Hunts Cross to Southport

This route was one of the first to be analysed, initial observations described in section 3.3 use this dataset. Features used for initial data analysis were taken from this route, it was also simulated and is used for synthetic data generation. This is largely due to the frequency of operation on this route and its ease of identification. Figure 54 shows a map view of the route, Figure 55 shows a profile of the gradient from Hunts Cross to Southport showing a general decline toward the Southport coastal area.

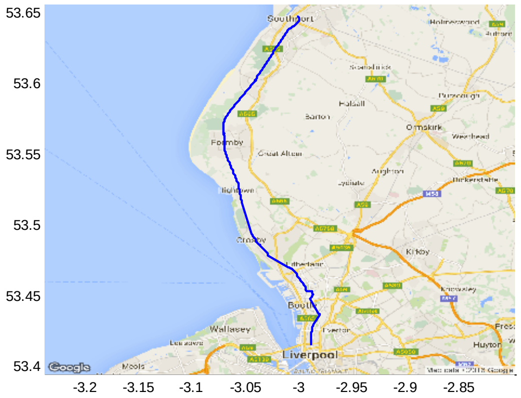


Figure 54 – GPS trace plotted to a map

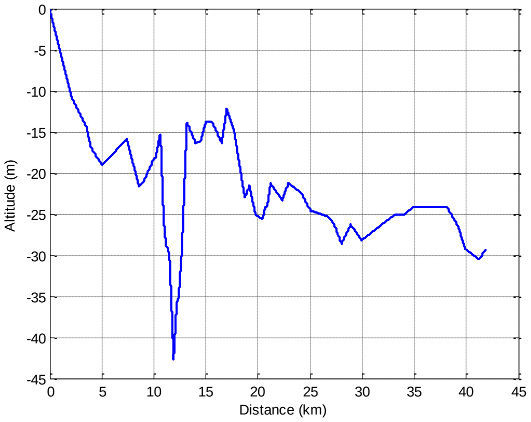


Figure 55 – Height profile of Hunts Cross to Southport route

Figure 56 shows eigenvalues as percentage contribution to overall variance, 36 % is contained in PC1, 24% contained in PC2 and 16% in PC3; an overall 76% of trend information is contained in the first three principal components.

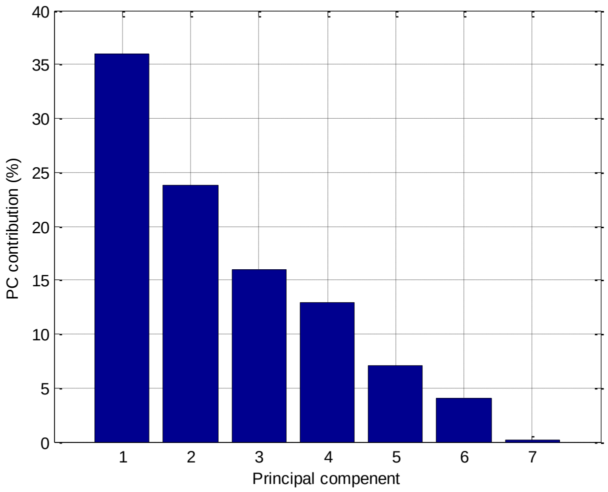


Figure 56 – Percentage contribution of each PC Hunts Cross to Southport

Figure 57 shows a colourmap of the first four principal components. As PCs 5, 6 and 7 contain less than 10% of the variance they are considered noise and automatically discounted, the trends shown in PC4 contribute little to energy. Although it contributes to approximately 13% of the variance it is also removed, this results in the PC1, 2 and 3 colourmap shown in Figure 58.

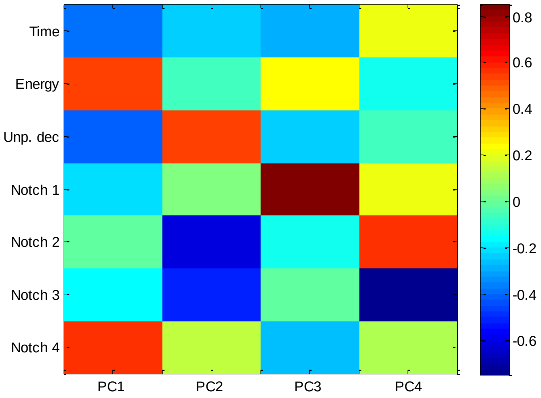


Figure 57 – 4 PC colourmap Hunts Cross to Southport

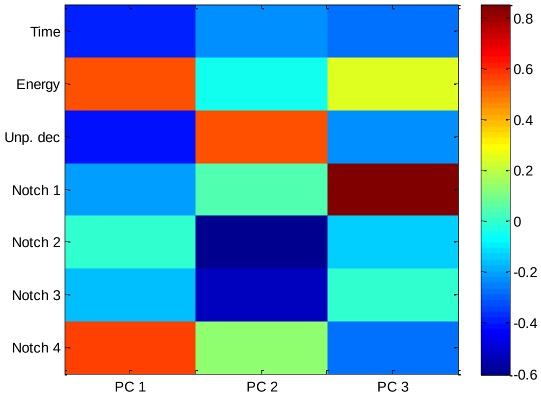


Figure 58 – 3 PC colourmap Hunts Cross to Southport

PCs 1, 2 and 3 are used to generate PC scores shown in Figure 59 and Figure 60. Based on the colourmap the following relationships are observable:

5.3.1.1 PC1

The reader is encouraged to read the PC summary following the bullet point descriptions of each trend to clarify what the trends mean in terms of the real driver styles.

- Energy consumption and notch 4 are strongly proportionally related, i.e. greater acceleration and a greater distance travelled while powering result in increased energy consumption.
- Time and unpowered deceleration are proportionally related, i.e. a greater distance coasted will increase journey time, both are inversely proportional to energy and notch 4 usage.
- Notch 1 and 2 usage are proportionally related and but have a lower influence on the variance in PC1 than those above. They are proportionally related to time and unpowered deceleration and inversely related to energy and notch 4 usage.

Summary: PC1 shows a split between two extremes, in this case we have examples of the lowest and slowest vs the highest and fastest (energy consumption & journey time) driver styles. As a result we can expect driver styles on the left of the horizontal axis of PC1 in Figure 59 to be low acceleration shown by increased notch 1 and 2 use, low energy, increased coasting and longer running time. Conversely on the right of the horizontal axis in Figure 59 we can expect to find driver styles with high acceleration, shown by increased notch 4 usage, high energy, little coasting but shorter running time. These driver styles are what are expected from the Pareto curve shown in 3.3, Figure 19, resulting from simulation data and the general trend of the real data shown in the same figure. These are examples of what can be thought of as 'efficient' driving. An efficient energy use driver style with a running time penalty (left) and an efficient time use driver style with an energy penalty (right).

5.3.1.2 PC2

The reader is encouraged to read the PC summary following the bullet point descriptions of each trend to clarify what the trends mean in terms of the real driver styles.

- The relation with the greatest influence in PC2 is between notch 3 and unpowered deceleration, this is an inverse relationship, where drivers are either coasting during the 'cruise' and deceleration periods or powering in notch 3.
- Notch 2 is proportional to notch 3 and inversely proportional to unpowered deceleration.
- In PC2 energy and time are proportional and both inversely proportional to notch 4 use and unpowered deceleration.

Summary: In PC2 a number of the relationships are opposite to those demonstrated in PC1, as each PC is orthogonal to the last it is not correlated. PC2 shows examples of what can be termed 'offset' driver styles and 'inefficient' driver styles. The top of Figure 59 shows increased coasting where drivers have used notch 4. This results in a slight energy and time saving and offsets the initial high energy consumption during the acceleration phase. The bottom of Figure 59 shows increased notch 2 and 3 use and reduced coasting resulting in a slight energy and time increase. This is typical of an inefficient driver technique. When energy consumption is high running time should be low, in this case there are examples of high running time *and* high energy consumption. This also suggests the use of 'low notching' (see 3.3.3.2 for a description of low notching) during potential coasting points.

5.3.1.3 PC3

The reader is encouraged to read the PC summary following the bullet point descriptions of each trend to clarify what the trends mean in terms of the real driver styles.

- Energy and notch 1 use increase proportionally in PC3 (shown in Figure 60) where all other input variables are inversely proportional to these or have very little influence in PC3.

Summary: Energy increase with notch 1 use is a further suggestion of the use of ‘low notching’ observed on this route (see 3.3), this is an example that shows energy increase with increased notch 1 use. Notch 1 is used almost exclusively during manoeuvring and briefly during acceleration. As a result a relationship between increased energy consumption and notch 1 use is likely a result of drivers selecting notch 1 during the deceleration phase where the camshaft is in position 1 and all resistors are connected in series, in effect causing ‘powered’ deceleration.

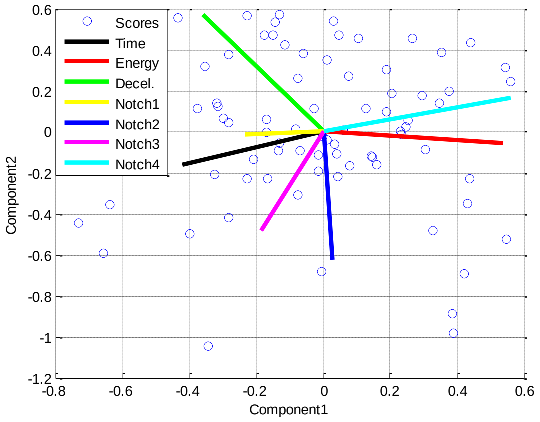


Figure 59 – Biplot showing Hunts Cross to Southport PC scores and principal components (1 & 2)

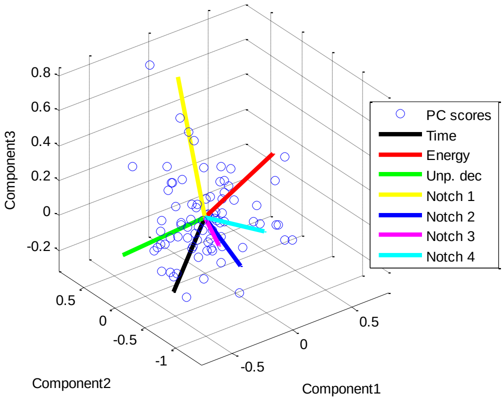


Figure 60 – Biplot showing Hunts Cross to Southport PC scores and Principal components (1 – 3)

5.3.2 Southport to Hunts Cross

On the return journey there is a general incline from the Southport coast back to Hunts Cross (see Figure 55), this appears to have influence on the effectiveness of coasting on this route related to the ‘offset’ driver styles. In this case it can be seen that coasting does not significantly improve energy consumption, this is discussed further below and is demonstrated in Figure 63 and Figure 65. Figure 61 is very similar to the percentage contribution bar chart for the Hunts Cross to Southport route. In this case PCs 1 and 2 contribute 61% to overall variance. Also PCs 3 and 4 have a very small effect on energy consumption and can be discarded along with ‘noise’ PCs 5, 6 and 7. PC scores are generated from PCs 1 and 2 only and are presented in Figure 64 and Figure 65. Although PC1 is very similar to the previous PC1, PC2 has a number of differences.

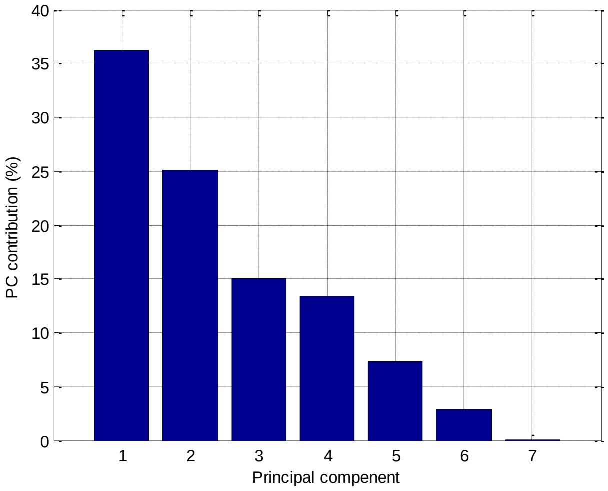


Figure 61 – Percentage contribution of each PC Southport to Hunts Cross

Figure 63 shows a colour map of Southport to Hunts Cross principal components, based on this the following observations below can be made

5.3.2.1 PC1

The reader is encouraged to read the PC summary following the bullet point descriptions of each trend to clarify what the trends mean in terms of the real driver styles.

- Energy consumption and notch 4 are strongly proportionally related, i.e. greater acceleration and a greater powered distance result in increased energy consumption.
- Unpowered deceleration is inversely related to energy consumption and notch 4 use, this shows that there are less examples of coasting where powering in notch 4 is prevalent which results in increased energy consumption.
- In this case time is proportionally related to unpowered deceleration and inversely proportional to energy and notch 4 use. Time has a lower influence (-0.2) in the first PC compared to the journey in the opposite direction (-0.4) suggesting it is not affected significantly by reduced acceleration or increased coasting.
- Notch 1 and 2 are inversely proportional in this route.

Summary: Although largely similar to Hunts Cross to Southport the return journey has an interesting difference in the first PC regarding journey time impact. Normally it is expected that for increased energy consumption journey time will be shortened, although that is the case in Southport to Hunts Cross the journey time is not substantially different compared to the outward journey. This suggests that a high acceleration, high energy driver style (right of the horizontal axis) is not gaining a time advantage compared to a low acceleration low energy driver style (left of the horizontal axis) meaning that driver styles further to the right of PC one are at risk of being inefficient as there is little benefit to the increased acceleration.

5.3.2.2 PC2

The reader is encouraged to read the PC summary following the bullet point descriptions of each trend to clarify what the trends mean in terms of the real driver styles.

- Notch 3 is inversely proportional to both notch 4 use and unpowered deceleration. Diver styles with more notch 4 use (upper vertical axis) also have increased coasting. The driver styles in the lower vertical axis display increased notch 3 use. This suggests a split between drivers who power in notch 4 then coast and those that cruise in notch 3.
- Notch 1 & 2 are both proportional to notch 3 and inversely proportional to notch 4 use & unpowered deceleration.
- Unlike the outward journey, PC2 shows energy consumption is *proportional* to unpowered deceleration and notch 4 use, driver styles in the upper area of Figure 65, although using increased coasting have increased energy consumption compared to the lower area that has reduced coasting.
- Time has little contribution to the dynamic of PC2.

Summary: Perhaps the most interesting conclusion from the return journey (Southport to Hunts Cross) is that their appears to be no ‘offset’ type driver style, in this case we see that increased notch 4 and unpowered deceleration in PC2 is contributing to increased energy consumption and that maintaining notch 1, 2 and 3 use results in a lower energy, lower running time driver style, with particularly increased amounts of notch 3 use. Prolonged notch 3 use has been observed to be a more inefficient way to drive in other examples [5] (see 3.3) but in this case is notable because this is not the case. This is likely a result of the return journey gradient which, in general, follows a slight incline from the Southport coast back to Liverpool.

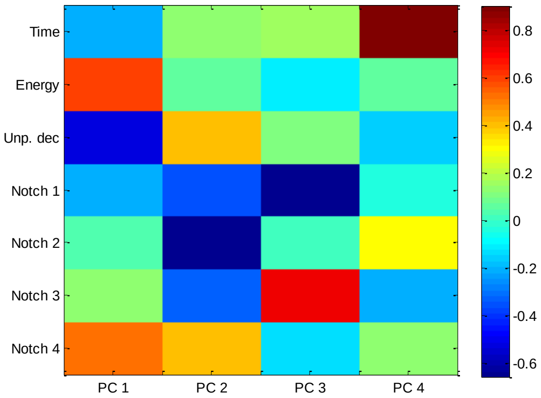


Figure 62 – 4 PC colourmap Hunts Cross to Southport

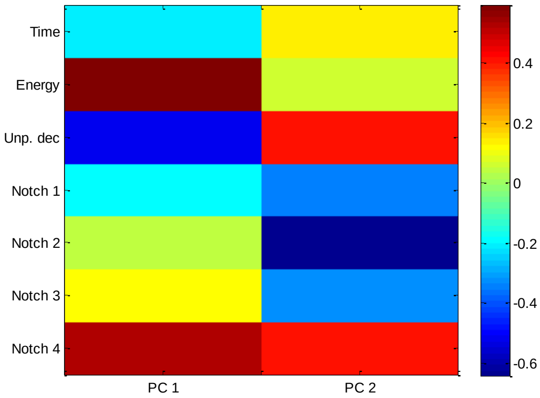


Figure 63 – 2 PC colourmap Southport to Hunts Cross

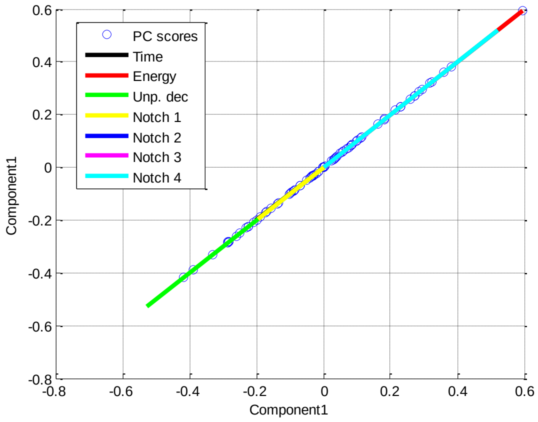


Figure 64 – Biplot showing Southport to Hunts Cross PC scores and principal components (1 only)

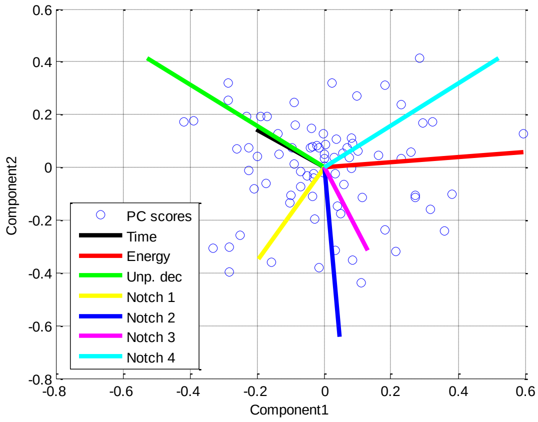


Figure 65 – Biplot showing Southport to Hunts Cross PC scores and principal components (1 & 2)

5.3.3 Liverpool to Ormskirk

This route is the central part of the Northern Line and is approximately half the length of the Hunts Cross to Southport route. The route separates from the main Northern Line at Sandhills and continues to Ormskirk. Once again there is a very similar pattern to the two previous routes analysed in terms of PC contributions to variance, 60% from the first two PCs and approximately 27% from PCs 3 and 4, PC s 5, 6 and 7 are again considered noise. As PCs 3 and 4 have little effect on energy consumption they are also discarded; PC scores are generated from PCs 1 and 2.

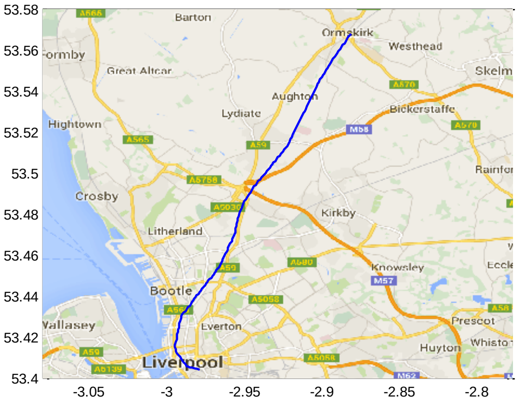


Figure 66 – GPS trace plotted to map

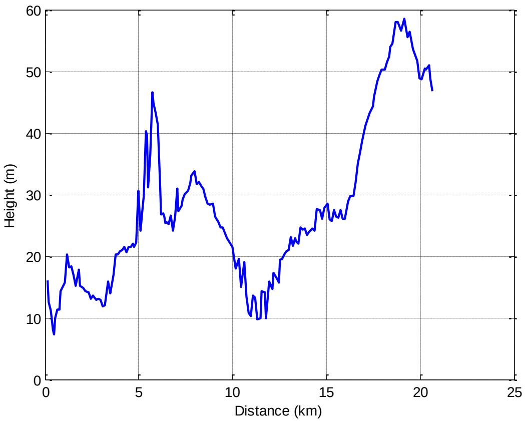


Figure 67 – Height profile of Liverpool to Ormskirk route

From the colourmap in Figure 70 and biplots in Figure 71 and Figure 72 the following can be observed:

5.3.3.1 PC1

The reader is encouraged to read the PC summary following the bullet point descriptions of each trend to clarify what the trends mean in terms of the real driver styles.

- Energy and notch 4 use are strongly proportionally related, unpowered deceleration is strongly inversely related to both. Drivers on the right of the horizontal axis are high acceleration and high energy consumption, on the left are low acceleration low energy driver styles which have increased coasting. In other routes time has been seen to have a strong inverse relationship to energy and notch 4 use, however in this case time has an almost zero value (≈ 0.05) indicating little to no involvement with the PC1 trends.

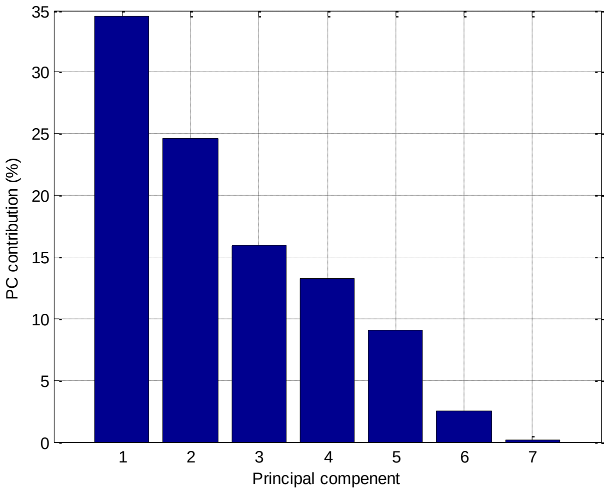


Figure 68 – Percentage contribution of each PC Liverpool to Ormskirk

- As with the Hunts Cross outward journey notch 3 use is inversely proportional to energy and notch 4 use, demonstrating a split between high and low acceleration driver styles on the right and left of the horizontal axis respectively (see Figure 71).

Summary: Interestingly PC1 in the Liverpool to Ormskirk route shows that time is not involved in the dynamic to a significant amount. This suggests that high energy driver styles do not gain a time advantage and low energy driver styles do not take a time penalty for the reduced energy consumption. As there has been a similar observation with the Southport to Hunts Cross route there is a suggestion that time variation from initial input data for certain routes does not vary significantly between individual runs to have a significant role in the underlying structure of the data, particularly in PC1. This is unlike simulation data where time had one of the largest dynamics in PC1 which was unexpected. This likely reflects the simulator having no upper limit on journey time allowing for a clear wide range of journey times compared to those found in Southport to Hunts Cross and, particularly, Liverpool to Ormskirk. This suggests that as a result of having a low range of journey time there is little benefit to using high acceleration, high energy driver styles on this route.

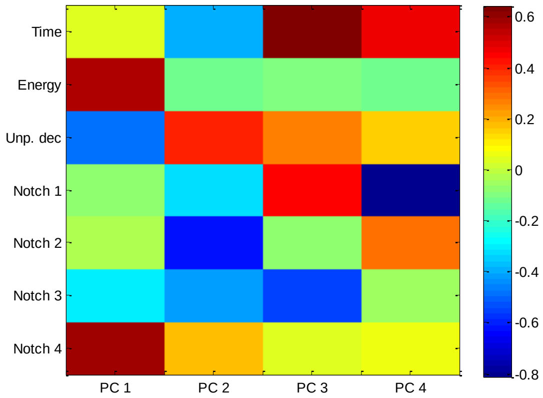


Figure 69 – 4 PC colourmap Liverpool to Ormskirk

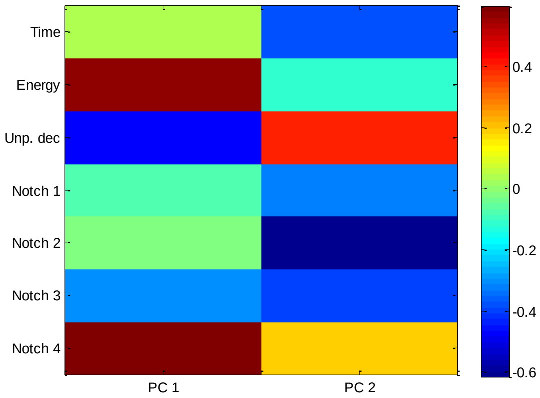


Figure 70 – 2 PC colourmap Liverpool to Ormskirk

5.3.3.2 PC2

The reader is encouraged to read the PC summary following the bullet point descriptions of each trend to clarify what the trends mean in terms of the real driver styles.

- Notch 1, 2, 3 and time are proportional, on the lower vertical axis are low acceleration slower run time driver styles, they are inversely proportional to unpowered deceleration, the upper vertical axis shows driver styles with increased coasting.
- Notch 4 use and energy are inversely proportional, driver styles with increased coasting on the upper vertical axis tend to be higher acceleration but have reduced energy consumption and vice versa on the lower vertical axis

Summary: PC2 is very similar to PC2 in the first route analysed, Hunts Cross to Southport. The area at the upper section of the vertical axis of Figure 72 shows that high acceleration i.e. increased notch 4, driver styles coincide with increased unpowered deceleration meaning that coasting is likely to be employed. This reduces energy consumption as well as journey time and is another example of an ‘offset’ driver style. The lower vertical axis shows a driver style that typifies ‘inefficiency’ where acceleration appears to be low and both journey time and energy consumption are increased. This is likely a result of prolonged use of notch 3.

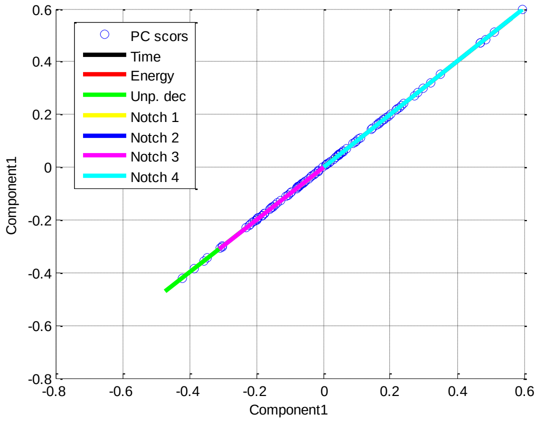


Figure 71 – Biplot showing Liverpool to Ormskirk PC scores and principal components (1 only)

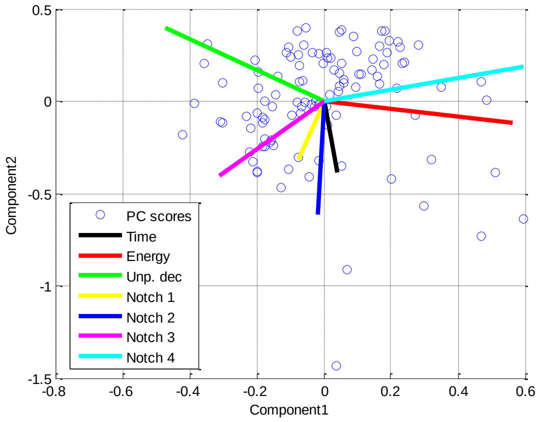


Figure 72 – Biplot showing Liverpool to Ormskirk PC scores and principal components (1 & 2)

5.3.4 Ormskirk to Liverpool

The return journey has numerous similarities to other routes analysed, however PCs 1 and 2 only contribute 57% to total variance which is the smallest of the large dataset routes. This is related to the relatively large 17% contribution of PC3 shown in Figure 73. Again PCs 5, 6 and 7 are removed based on their likely representation of noise and PC4 is also removed as its lower variance contribution does not have a substantial effect on energy consumption. PC scores are generated from PCs 1, 2 and 3 and are shown in Figure 76 and Figure 77. Based on colourmaps in Figure 74 & Figure 75 and biplots in Figure 76 & Figure 77 the following observations were made:

5.3.4.1 PC1

The reader is encouraged to read the PC summary following the bullet point descriptions of each trend to clarify what the trends mean in terms of the real driver styles.

- Energy and notch 4 use are proportionally related. As in all previous routes unpowered deceleration is inversely proportional to these, driver styles to the right are high energy, high acceleration and short running time, compared to those on the left which are lower energy, low acceleration and increase running time.
- Time is proportionally related to notch 1 – 3 and unpowered deceleration and as a result is also inversely proportional to unpowered deceleration and energy.

Summary: PC1 shows similar features to Hunts Cross to Southport demonstrating the split between high acceleration high energy vs low acceleration low energy. Unlike Hunts Cross to Southport the time advantage/penalty for high or low acceleration is substantially lower suggesting less advantage from high acceleration high energy driver styles, this is more in line with the Southport to Hunts Cross route, meaning that driver styles to the extreme right of PC1 in Figure 76 are good examples of inefficient high acceleration high energy driver styles. Interestingly there seems to be a clear split between the main body of PC scores and an isolated extreme group suggesting only limited use of these inefficient driver styles on this particular route.

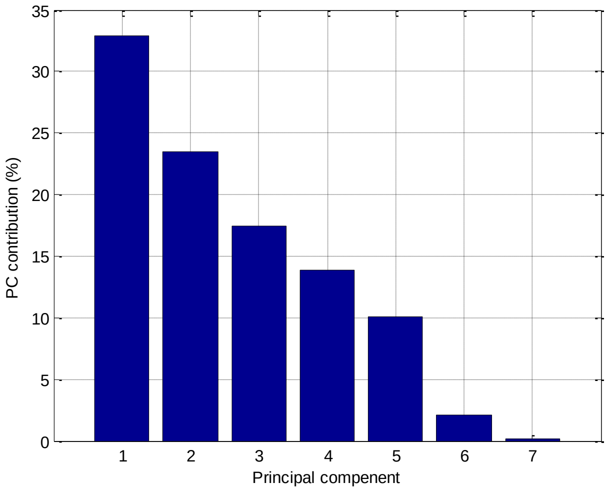


Figure 73 – Percentage contribution of each PC Ormskirk to Liverpool

5.3.4.2 PC2

The reader is encouraged to read the PC summary following the bullet point descriptions of each trend to clarify what the trends mean in terms of the real driver styles.

- Unpowered deceleration and notch 1 and 3 use have a strong inversely proportional relationship. Upper vertical axis driver styles are higher acceleration but employ coasting, lower driver styles are lower acceleration and have reduced coasting.
- Notch 4 use and energy consumption are also inversely proportional, meaning that Notch 4 and unpowered deceleration are proportional. Energy and notch 1 & 3 use are also proportional. Upper vertical axis high coasting driver styles have reduced energy consumption compared to the lower vertical axis low acceleration driver styles which have increased energy consumption

Summary: As in previous examples (except Southport to Hunts Cross) PC2 shows examples of high acceleration driver styles which are able to offset energy consumption by the use of coasting, as shown by the proportional relationship between notch 4 use and unpowered deceleration and their inverse relationship to energy consumption. Ormskirk to Liverpool shows the greatest energy savings in PC2. This, however, comes at the cost of a slight running time penalty as time is shown to increase with both notch 4 and unpowered deceleration shown in the top section of Figure 76. In the lower section of Figure 76 a further example of prolonged notch 3 use is shown to contribute to increased energy consumption although, this does not contribute to increased running time.

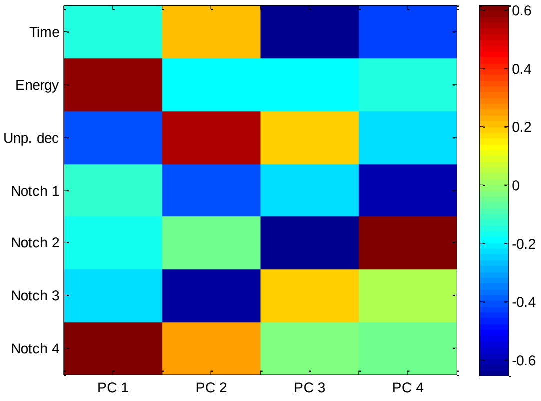


Figure 74 – 4 PC colourmap Ormskirk to Liverpool

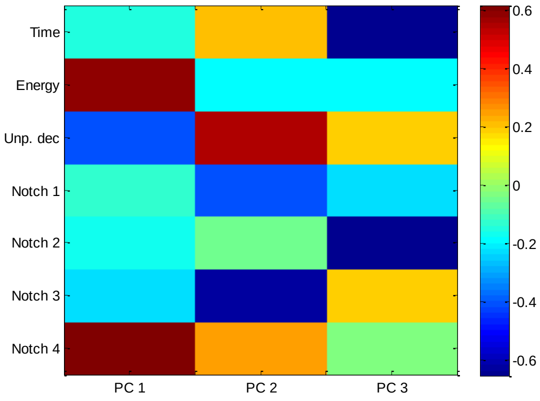


Figure 75 – 3 PC colourmap Ormskirk to Liverpool

5.3.4.3 PC3

The reader is encouraged to read the PC summary following the bullet point descriptions of each trend to clarify what the trends mean in terms of the real driver styles.

- There is a strong proportional relationship between notch 2 use and running time. Energy and notch 1 are also proportionally related to both although this is weaker than the above.

Summary: Figure 77 shows strong relationships between journey time and low notch control (notch 1 and 2) this coincides with energy consumption. This is a good example of ‘low notching’ driver strategy that has been observed in other examples above. In this case it can be seen that there is a significant time and energy penalty resulting from their use.

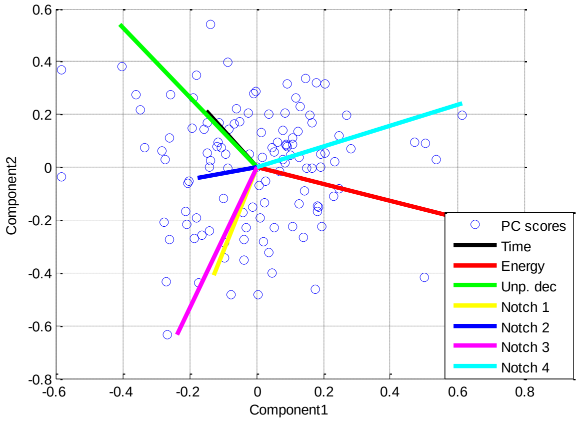


Figure 76 – Biplot showing Ormskirk to Liverpool PC scores and principal components (1 & 2)

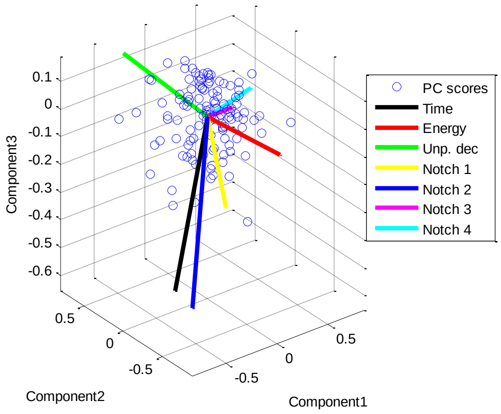


Figure 77– Biplot showing Ormskirk to Liverpool PC scores and principal components (1 – 3)

5.3.5 Small dataset routes

Small dataset routes cover outward and return Liverpool to West Kirby runs. From the large dataset analysed only 7 such journeys were identified indicating that the instrumented vehicle was mostly in operation on the Northern Line during this period. The small size of these sets would likely not have sufficient data to produce reliable driver style analysis results as it is certain that a large enough sample of the driver population was not captured in these journeys. Results shown in figures below demonstrate the differences between the larger dataset routes.

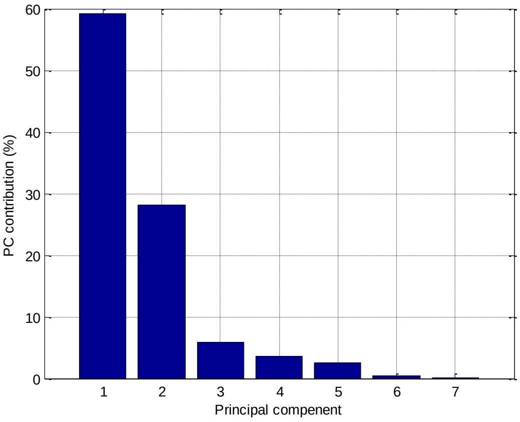


Figure 78 – Percentage contribution of each PC Liverpool to West Kirby

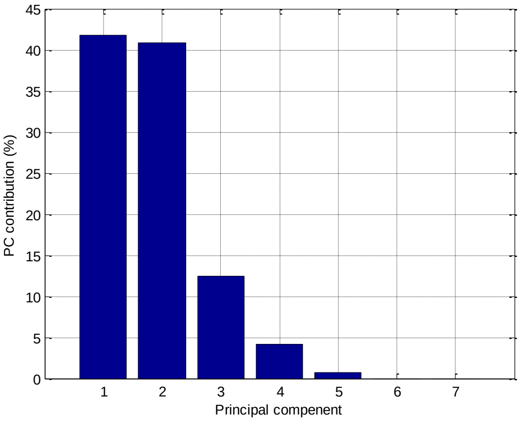


Figure 79 – Percentage contribution of each PC West Kirby to Liverpool

When compared to the small dataset routes there is little comparison to be made. Figure 82 shows the PC1 trend generally demonstrated in all other routes, i.e. energy consumption and notch 4 verses time and unpowered deceleration, this trend is not observable in the return journey shown in Figure 83 where energy is inversely proportional to high acceleration and low acceleration notches have no effect on energy consumption. PC2 shows similar inconsistencies, particularly in Figure 83 where a slight increase in unpowered deceleration has a very large effect on energy consumption which is not demonstrated to this extent in other routes observed. Small datasets are not included in further individual route analysis.

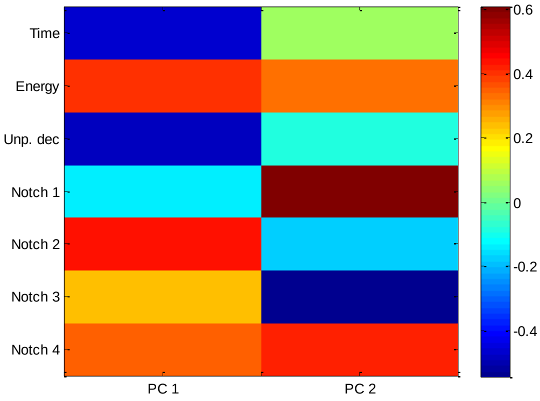


Figure 80 – 2 PC colourmap Liverpool to West Kirby

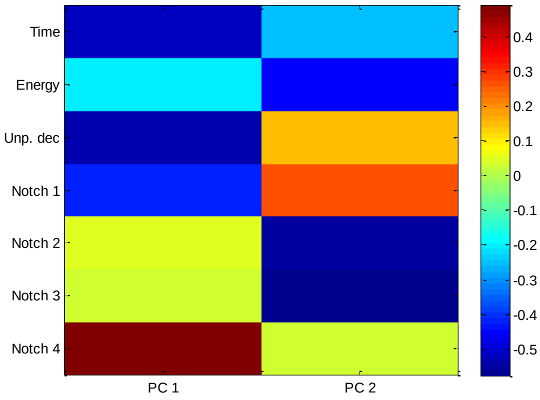


Figure 81 – 2 PC colourmap West Kirby to Liverpool

For both outward and return journeys two principal components were used to produce PC scores, percentage contribution for these are shown in Figure 78 and Figure 79, colourmaps are also shown in Figure 80 and Figure 81.

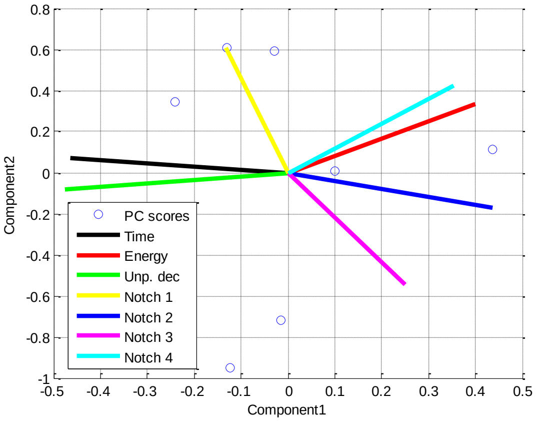


Figure 82– Biplot showing Liverpool to West Kirby PC scores and principal components (1 & 2)

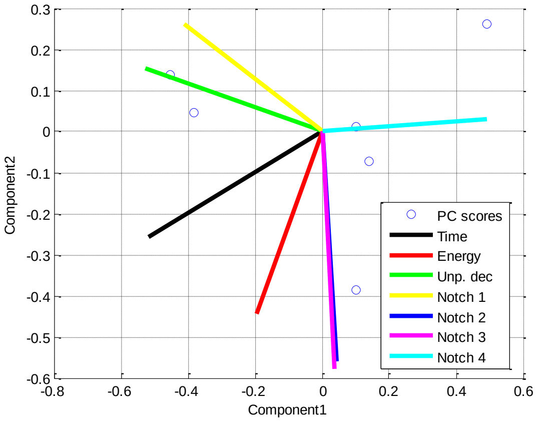


Figure 83 – Biplot showing West Kirby to Liverpool PC scores and principal components (1 & 2)

5.4 Cluster validation

Before clustering a suitable *k* (number of clusters) value must be decided upon. In the simulation clustering shown in 4.5 the number of driver styles, and therefore number of clusters, was known *apriori*. This allowed for accurate selection of *k*, however, with real data routes it is not known how many driver styles are present. In each case it was thought that well defined and isolated clusters would not be present as shown in the clustering of simulated driver styles. This means that a simple assumption of 4 clusters existing in each of the real data routes is not an appropriate way to proceed. In this case it is likely there is a degree of drift between the characteristics of driver styles across principal components. This means that clustering is less likely to be a process that identifies unique driver styles but is rather a technique to impose boundaries based on the limits of similarity between data which is closely grouped.

Choosing an appropriate *k* value represents an ongoing problem in the field of data clustering [73], accurate separation of data into its most likely groups can still not be done with precision [73]. A number of cluster validation tools exist which have been developed to identify an optimal value of *k*, tools used in this analysis are the Dunn index [84], Davies-Bouldin index [85] and Silhouette analysis [86]. The optimal choice does not automatically indicate the ‘right’ choice; rather it is a selection of *k* that satisfies the rule of the particular algorithm. As a result the final choice for *k* for each route is also dependant on the analyst’s discretion [73, 86] the main proviso being that the clustering results must be useful to the analyst [73]. For example, clustering suitability will tend to increase with the number of clusters until each data point is in its own cluster [86]. Obviously this defeats the point of clustering and so limitations must be introduced. Therefore an upper limit of 8 clusters is imposed on the optimal cluster validations, and clustering validation is repeated for each route so that number of clusters for each differs by only ± 1 cluster. Algorithms for three cluster validation indexes are shown below followed by their application to the Hunts Cross to Southport route. Each algorithm shown attempts to identify optimal clustering based on comparison of intracluster distance and intercluster distance, which is an assessment of cluster compactness and cluster separation. As *k*-means is randomly initiated and results can be significantly different between each initiation of the algorithm cluster validation is based on the *k* value with the highest frequency. This is taken from 100 runs of the *k*-means algorithm for *k* iterations from 1 to 8. Analysis is repeated until validation scores from each algorithm are the same.

The Dunn index predates both Silhouette and Davies-Bouldin indexes (presented below) but all are largely similar in their approach [84]. This algorithm attempts to identify intracluster similarity, i.e. low variance between members of the same cluster compared to the mean of each cluster. This is also termed the cluster separation. The Dunn index is determined by:

$$Dunn = min_{1 \leq i \leq} \left\{ min \left\{ \frac{d(n_i, n_j)}{max_{1 \leq k \leq n}(d(X_k))} \right\} \right\} \tag{21}$$

d(n_i,n_j) is the intercluster distance and determines cluster separation between clusters X_i and X_j, *n* is the number of clusters in a particular k-means iteration and so represents *k*. Large Dunn index values indicate more appropriate clustering [84, 87].

The Davies-Bouldin index aims to identify compactness of individual clusters and the separation between them [87]. Cluster appropriateness $R_{i,j}$ is determined by the ratio of S_i cluster scatter, determined by Euclidian distance to the cluster centroid, to $M_{i,j}$ cluster separation, determined by the Euclidian distance between clusters *i and j.* A lower *R* value indicates a tighter cluster scatter and greater inter-cluster distance [85]. Indexes for iterations of *k* are determined by:

$$DB = \frac{1}{N}\sum_{i=1}^{N} Max_{i\neq j}\left\{\frac{d(X_i) + d(X_j)}{d(n_i, n_j)}\right\} \tag{22}$$

N is the number of clusters in the current k-means iteration, *i* and *j* are clusters *d(*X_i*)* and *d(*X_j*)* are data points in clusters *i* and *j*, finally *d(*n_i*,* n_j*)* the separation between clusters [87]. This provides a ratio of the intracluster scatter to the cluster separation meaning that a lower DB provides a more appropriate value for *k* [85].

Silhouette analysis provides a comparison of mean intra-cluster distance *a(i)* to mean inter-cluster distance *b(i)* with the closest neighbour cluster for each datum [86]. *S(i)*, the silhouette score is determined by:

$$s(i) = \frac{b(i) - a(i)}{Max\{a\ (i),\ b(i)\}} \tag{23}$$

A low *a(i)* value indicates low intra-cluster dissimilarity, a large *b(i)* value indicates a large inter-cluster dissimilarity, therefore a mean *s(i)* value close to 1 indicates well-fitting clusters. The algorithm is repeated for *k* values up to 8, omitting $k = 1$. Figure 84 shows the cluster validation solution for the Hunts Cross to Southport route. The k-means algorithm ran for *k* values of 2 to 8 with 100 iterations for each, the best score for each index is selected. Best Scores for each value of *k* are compared and the 'best of the best' is selected as the most appropriate number of clusters for each validation algorithm. In this case '6' clusters are selected for the Hunts Cross to Southport route. Cluster validation is repeated for the remaining three routes Table 10 summarises the most appropriate amount of clusters for each dataset, these values are used as *k* for k-means clustering shown in the following section.

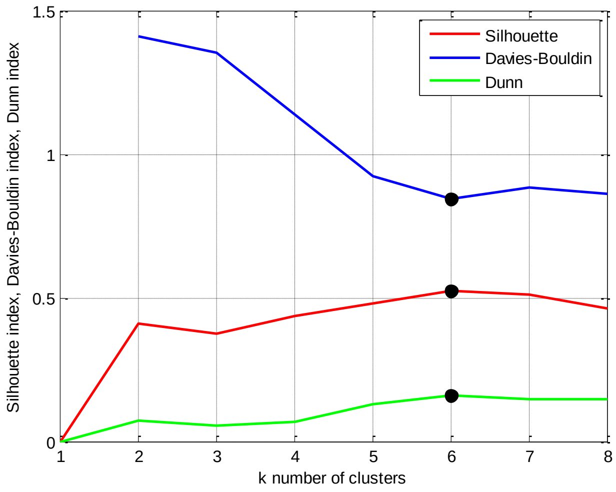


Figure 84 – Cluster validation for 100 hundred iterations of the k-means algorithm for k values 2 – 8

Table 10 – Routes and cluster number used in k-means clustering

Route	Most appropriate number of clusters
Hunts Cross to Southport	6
Southport to Hunts Cross	6
Liverpool to Ormskirk	7
Ormskirk to Liverpool	6

Other than Liverpool to Ormskirk the most appropriate number of cluster for each route was found to be 6 when the algorithm was allowed to run continuously until a consensus between each validation tool was found. It was found that the Liverpool to Ormskirk run would never find a consensus between each validation tool that resulted in a 6. After many cycles of the algorithm 7 clusters were found to be the most appropriate grouping. It is likely that this is a result of the disparate data plots in the right and lower right of the data plot in Figure 92 in section 5.3.3. This demonstrates the limits of the *k*-means algorithm when clustering non-circular or not tightly grouped data. It is likely that a higher value for *k* would provide better clustering results, however, this would become cumbersome and un-suitable for the calculation of energy consumption for each driver style and provide over detailed information, and as a result the decision was made to keep 7 clusters.

5.5 Results

This section presents clustering observations for each route, cluster centroids are then used to represent the characteristics of a particular driver style. Energy consumption, journey time and distribution of membership to each driver style are then shown. The distribution of differing driver styles has been termed a ‘driver culture’ and can be used to assess the status of driver technique on a particular route. By altering the driver style distribution, increasing the percentage of low energy style journeys for example, an improved energy consumption driver culture can be designed. A matrix is generated which contains percentile contributions for all possible driver style distribution permutations in 10% steps. Energy consumption and time for each permutation is added to the permutation matrix. Table 12 shows a small sample of driver style distribution permutations for the route Hunts Cross to Southport. Principal Component Analysis is then used to identify the relationship of driver styles to energy consumption and time.

By analysing multiple permutations of driver styles an optimised driver culture can be identified, this involves selecting a distribution of driver styles that have both reduced energy consumption and less impact on running time. The distribution of the optimal driver culture is based on service frequency and peak and off-peak times; the faster of the two optimal styles used to cover peak operation, the slower for off-peak. Services operate at 15 minute intervals from 06:00 to 23:00, services after 21:00 on the Liverpool to Ormskirk and return route operate at intervals of 30 minutes. Daily there are 70 services on the Hunts Cross to Southport route and 60 services on the Liverpool to Ormskirk route. Merseyrail peak operation is 06:31 – 09:29 [88], there is no specified peak evening time based on ticket price change, however, Merseyrail suggest 15:30 – 19:10 as the evening peak operating times [89]. The number of services in the peak and off peak range are calculated and divided between the two optimal driver styles for each route. Service peak and off peak times are shown in Table 11.

Table 11 – Service peak and off peak times

	Off peak	**Peak**	**Off peak**	**Peak**	**Off peak**
Time	06:06 – 06:31	06:31 – 09:29	09:29 – 15:30	15:30 – 19:10	19:10 – 23:21
No. of services	2	12	24	15	17
No. of services	0	13	24	15	8

Three daily energy consumption estimates are then made:

- A daily energy consumption estimate based on the real data distribution of driver styles
- An estimate based on a distribution taken from the optimised driver styles
- An estimate based on the lowest energy consumption driver style.

In a number of cases the lowest energy consumption driver style is very close or exceeds the scheduled journey time for its particular route. This suggests that lowest energy consumption driver styles are not a simple solution to energy reduction as the timetable cannot always be adhered to. This does, however, provide a lower boundary of possible energy consumption.

Table 12 – Sample of driver style distribution permutations for the Hunts Cross to Southport route

Driver style 1 %	**Driver style 2 %**	**Driver style 3 %**	**Driver style 4 %**	**Driver style 5 %**	**Driver style 6 %**	**Energy consumption**	**Time**
70	0	0	0	20	10	114.36	53.44
60	10	0	0	20	10	113.96	53.39
50	20	0	0	20	10	113.57	53.35
40	30	0	0	20	10	113.18	53.30
30	40	0	0	20	10	112.78	53.26
20	50	0	0	20	10	112.39	53.21

It is very important to point out that cluster designations have no relationship to each other between routes, cluster 1 in Hunts Cross to Southport is not to be compared to cluster 1 in Southport to Hunts Cross or in any other route for example. Each cluster should be considered unique to each route. This is because of two important points, a) clusters are randomly initialised and do not always represent the same area of data for each analysis, b) routes are not homogenous enough to be compared like with like; even if a cluster is present in a similar location it might represent a very different selection of variance due to the different PCA result between each route. This is especially true for examples where PC2 shows energy increase with increased unpowered deceleration in Southport to Hunts Cross for example. As a result general driver approach will be considered across routes but driver cultures are not compared to each other, rather specific recommendations are given on a route by route basis.

Daily Energy consumption estimates that result from cluster analysis are taken from real data from the A car of the instrumented vehicle. This means that approximately half of the true energy consumption is presented. As there is no indication of vehicle configuration in the datasets and information relating to what configuration was in operation on the particular date is no longer available it is not possible to determine whether data represents 3 or 6 car variants.

Simulations in [5] Show there is little difference between energy consumption of a single power car for 3 and 6 car variants. It has been assumed that configuration is not affecting the analysis. Therefore daily energy estimations are for single power car journeys only and likely underestimate total daily energy consumption on a particular route insignificantly. This being said, percentage savings likely reflect actual savings and are more important than the numerical energy consumption values.

Table 13 – Route and Driver style options for energy reductions

Route	Peak option	Off peak option
Hunts Cross to Southport	5	6
Southport to Hunts Cross	2	5
Liverpool to Ormskirk	4	3
Ormskirk to Liverpool	1	2

5.5.1 Hunts Cross to Southport

Cluster validation identified 6 clusters for this route; these are shown with PC scores in two and three dimensions in Figure 85 and Figure 86. Most of the clusters are focused around the central (mid acceleration, mid energy mid running time) and central upper regions (increased acceleration, increased coasting and reduced energy). Cluster three is quite distinct from the other clusters and represents the highest energy driver style, approximately 130 kWh a 17% increase over the lowest energy consumption driver style cluster 2, with a low running time of 51 minutes. The most populous styles are cluster 1 and 4, which are the second lowest (113kWh) and second highest (124kWh) energy consumption driver styles; see Figure 87 for energy time and cluster membership values.

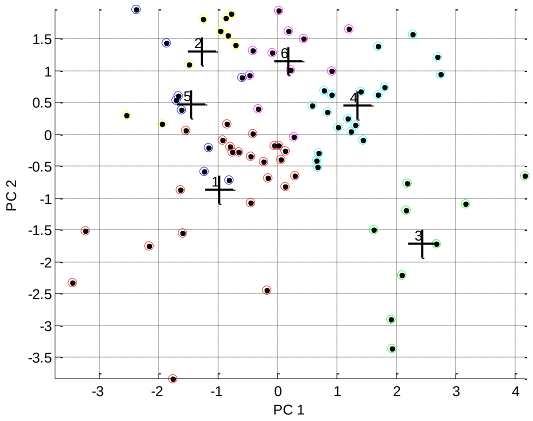


Figure 85 – k-means clustering Hunts Cross to Southport (PCs 1 & 2)

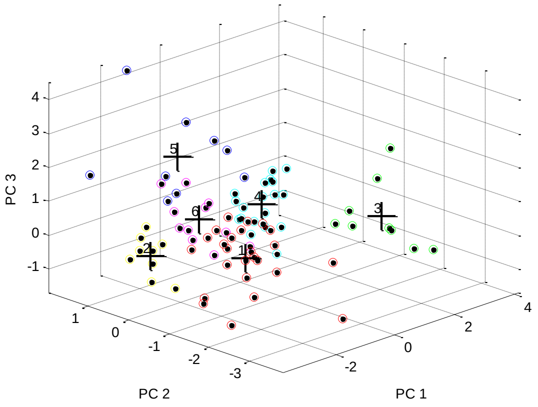


Figure 86 – k-means clustering Hunts Cross to Southport (PCs 1 – 3)

PCA results are shown in Figure 88; PC scores in this case are the transposed 8 dimensional permutation matrix, a sample of which is shown in Table 12. For the greatest impact on running time driver styles 3 and 4 should be maximised, but this comes at the cost of a significant energy consumption increase, especially with the case of driver style 3 which increases energy consumption in both PC1 and 2. Low energy consumption driver styles 1 and 2 increase running time, in PC2 driver style 1 also increases with energy consumption, suggesting that although the lowest energy consumption style overall its time penalty does not come with the greatest saving. Driver styles 5 and 6 show little relationship to energy consumption and running time, 5 shows a slight increase in time for a small energy saving and vice versa for 6. However, In PC2 it is shown that they provide both energy and time savings when increased. This suggests a distribution that maximised driver styles 5 and 6 would produce an overall energy reduction with little effect on running time. This is also suggested in Figure 87, both driver styles 5 and 6 have comparable running times (52 minutes, & 53 minutes) to the highest energy, driver style 3 (51 minutes). Although styles 5 and 6 are not the lowest energy consumption runs overall their running time suggests a more efficient driver approach that is more suited to adhering to the timetable. They are then selected as the ‘optimal’ driver styles to provide an energy saving estimate for the Hunts Cross to Southport route.

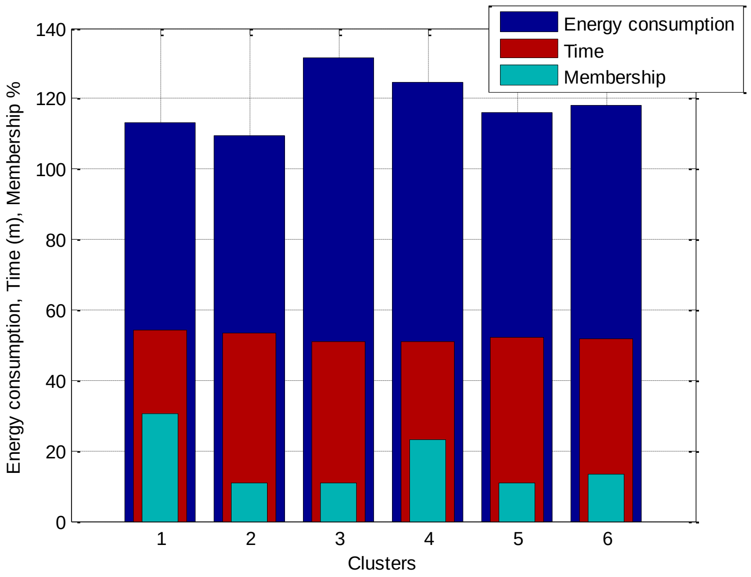


Figure 87 – Time and energy distribution of clusters Hunts Cross to Southport

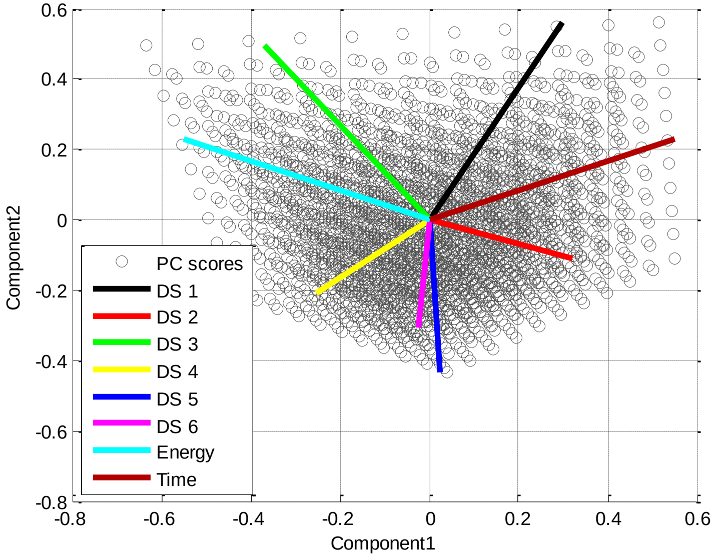


Figure 88 – Principal component biplot of Driver style relationships to energy and time Hunts Cross to Southport

5.5.1.1 Daily energy consumption estimate

Based on the number of data points in each cluster a driver culture can be calculated which is then adjusted to represent a single day of operation with 70 services. Table 14 summarises energy consumption, distribution and number of runs for each driver style for a single day on the Hunts Cross to Southport route. Number of runs for each driver style is calculated and multiplied by that particular styles energy consumption giving daily energy consumption (blue section) a total of 8294.8 kWh. The second estimate is made using only optimal driver styles which are highlighted in green; this estimation provides an energy estimate of 8173.5 kWh (red section) approximately a 1.5% reduction. In the lower dark green section the third estimate is given using the lowest energy consumption only, this provides a daily energy consumption of 7660.2 kWh, which is a saving of 8%.

Table 14 – Energy consumption estimations Hunts Cross to Southport

	DS 1	**DS 2**	**DS 3**	**DS 4**	**DS 5**	**DS 6**	**Total**
Energy single run (kWh)	113.4	109.4	131.5	124.8	116.0	118.0	
No. in cluster	25.0	9.0	9.0	19.0	9.0	11.0	82.0
Distribution %	0.3	0.1	0.1	0.2	0.1	0.1	1.0
No. of runs	21.0	8.0	8.0	16.0	8.0	9.0	70.0
Total energy (kWh)	2380.9	875.5	1052.2	1996.6	928.0	1061.6	8294.8
No. in opt. %					0.61	0.39	1.0
No. of opt. runs					43	27	70.0
Total energy opt. (kWh)					4953.3	3220.2	8173.5
No. in opt. single %		1.0					
No. of opt. single runs		70.0					
total energy opt. single (kWh)		7660.2					7660.2

5.5.2 Southport to Hunts Cross

K-means clustering identified 6 driver styles on this route 40% of which are in the central cluster 1 driver style shown in Figure 89. Cluster 5 is close to cluster 1 suggesting a gradual drift between the two rather than a well-defined boundary. Clusters 2, 3, 4 and 6 all have similar membership, approximately 15% of the runs in each and are situated as more distinct groups around the central cluster 1 and 5 group. Cluster 3 is the highest energy (136 kWh), lowest running time (51.4 min) driver style and is characterised by high acceleration by increased notch 4 use as well as prolonged notch 3 use. Cluster 4 is the lowest energy (112 kWh), highest running time (60 min) driver style which is characterised by low acceleration followed by notch 4 use and coasting. As time is shown as ‘pure’ running time i.e. with dwell time removed driver style 4 would not be able to adhere to the timetable as the entire journey including dwell time is 60 minutes (from the Merseyrail timetable) .

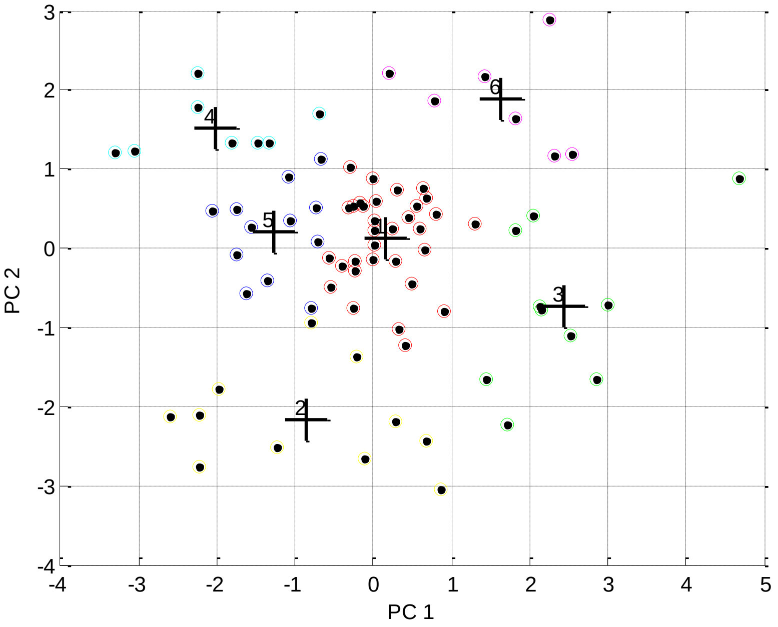


Figure 89 k-means clustering Southport to Hunts Cross (PCs 1 & 2)

PCA analysis was repeated for the Southport to Hunts Cross driver style permutation matrix; Figure 91 shows the PC scores and is plotted on the first two principal components. Driver distributions with increased styles 1, 3 and 6 show increased energy consumption where only driver style 3 provides a significant time advantage. Increased driver styles 4 and 5 reduce energy consumption, 4 significantly, however with substantial time penalties, this is less so with increased driver style 5 in PC1. In PC2 driver style 5 contributes to both reduced running time as well as energy consumption. Driver style 2 shows a similar relationship to energy and time. As a result both are selected for the optimal driver styles for this route. Driver style 2 has shorter running time (54.7 minutes) for very similar energy consumption 117 kWh compared to style 5’s 116 kWh, resulting in style 2’s selection for peak operation and style 5 for off peak.

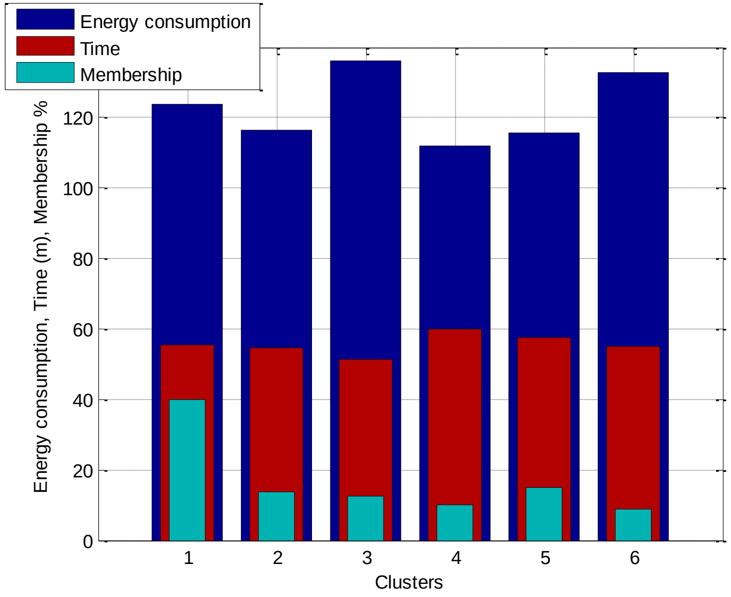


Figure 90 – Time and energy distribution of clusters Southport to Hunts cross

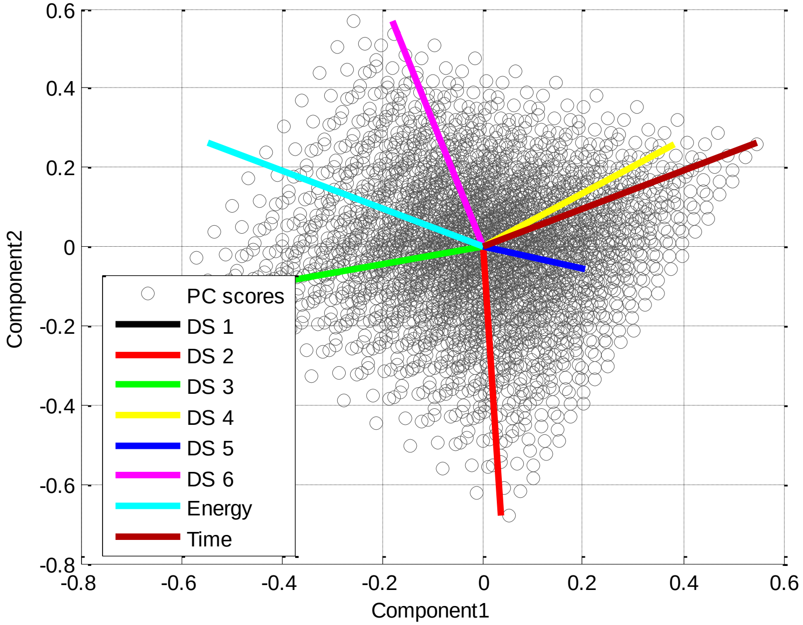


Figure 91 – Principal component biplot of Driver style relationships to energy and time Southport to Hunts Cross

5.5.2.1 *Daily energy consumption estimate*

The Southport to Hunts Cross route is the return journey for the first route analysed and so the same peak and off peak service are used. Table 15 summarises the energy consumption, distribution and number of runs for each driver style, again three estimates are made. The return journey based on the distribution of driver styles in the data produces an energy consumption of 8587.8 kWh shown in the blue section. The optimal peak and off peak driver styles shown in the red section produce an energy consumption estimate of 8134.2 kWh a reduction of 5.4 %. The final estimate uses driver style 4 only, the lowest energy consumption style, in this case 7844.9 kWh a reduction of 9%.

Table 15 – Energy consumption estimations Southport to Hunts Cross

	DS 1	DS 2	DS 3	DS 4	DS 5	DS 6	Total
Energy single run (kWh)	123.7	116.6	136	112.07	115.6	132.8	
No. in cluster	32.0	11.0	10.0	8.0	12.0	7.0	80.0
Distribution %	0.4	0.1	0.1	0.1	0.2	0.1	1.0
No. of runs	28.0	9.6	8.8	7.0	10.5	6.1	70.0
Total energy (kWh)	3462.5	1122.6	1190.3	784.5	1214.3	813.6	8587.8
No. in opt. %		0.39			0.61		1.0
No. of opt. runs		27			43		70.0
Total energy opt. (kWh)		3149.2			4972.9		8122.1
No. in opt. single %				1.0			
No. of opt. single runs				70.0			
total energy opt. single (kWh)				7844.9			7844.9

5.5.3 Liverpool to Ormskirk

Clustering for the Liverpool to Ormskirk route produced 7 driver styles which are shown in Figure 92. Most clusters appear tightly grouped on the upper left of the figure; this is likely a result of the two distinct and largely dispersed clusters 2 and 7. Clusters 1, 3, 4 and 6 have very similar membership with approximately 17% in each, and cluster 5 slightly less with around 14%. The main body of clusters are characterised by varying degrees of low to mid acceleration and energy on PC1 and increased coasting reduced energy across PC2, energy consumption varied from 50 – 60 kWh. Clusters 2 and 7 are characterised by high acceleration, low coasting and prolonged notch 3 use, energy consumption is over 65 kWh. Interestingly driver style 2 has the longest journey time. Style 7 does not show reduced journey time for its increased energy consumption suggesting inefficient driver technique in both.

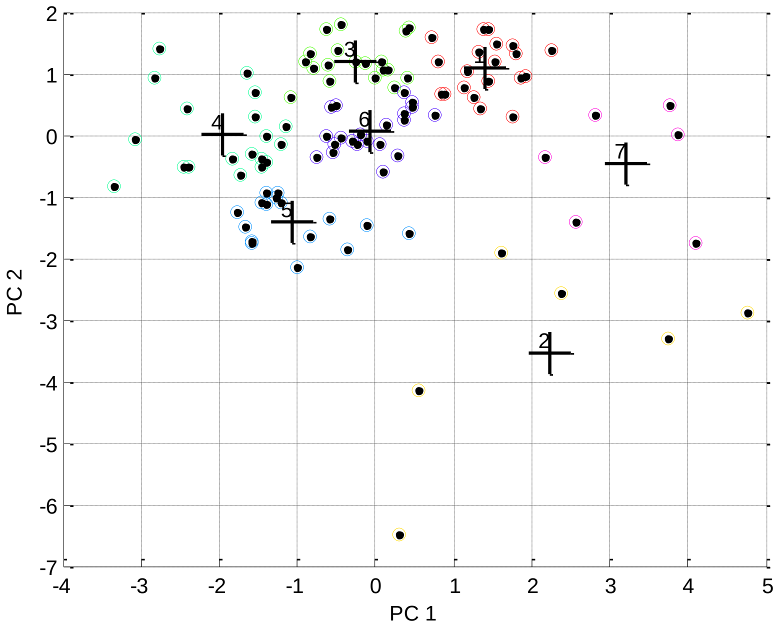


Figure 92 – k-means clustering Liverpool to Ormskirk (PCs 1 & 2)

PCA analysis of the Liverpool to Ormskirk permutation matrix in Figure 98 shows the lowest three energy consumption driver styles, 3, 4 and 5 also have reduced journey time. Driver style 4 should be selected for the optimal peak time driver style as it has the lowest energy consumption as well as a good journey time. With the closely grouped clusters journey time is almost equal between each driver style (22 – 23 minutes), 5 being slightly increased at 26 minutes. Driver style 3 should be selected for off peak; although style 5 has slightly lower energy consumption 3 has better running time.

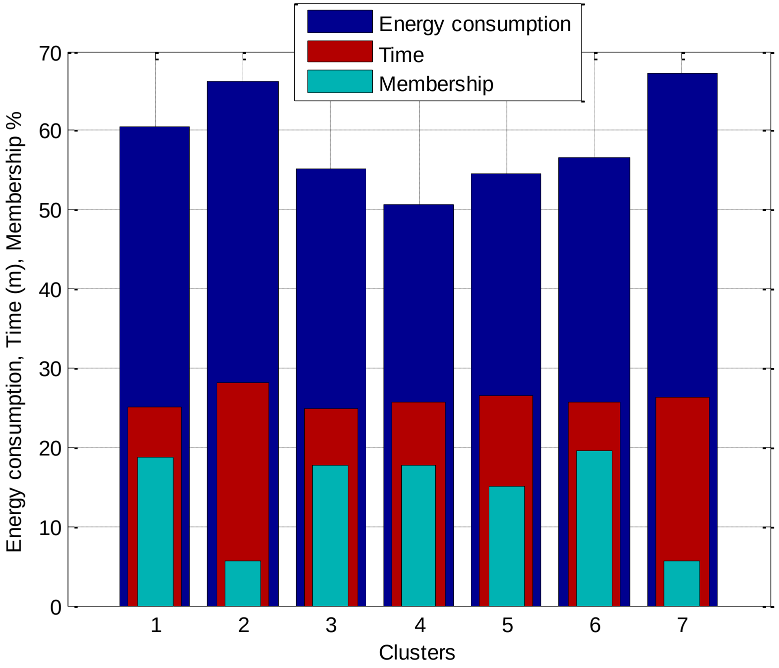


Figure 93 – Time and energy distribution of clusters Liverpool to Ormskirk

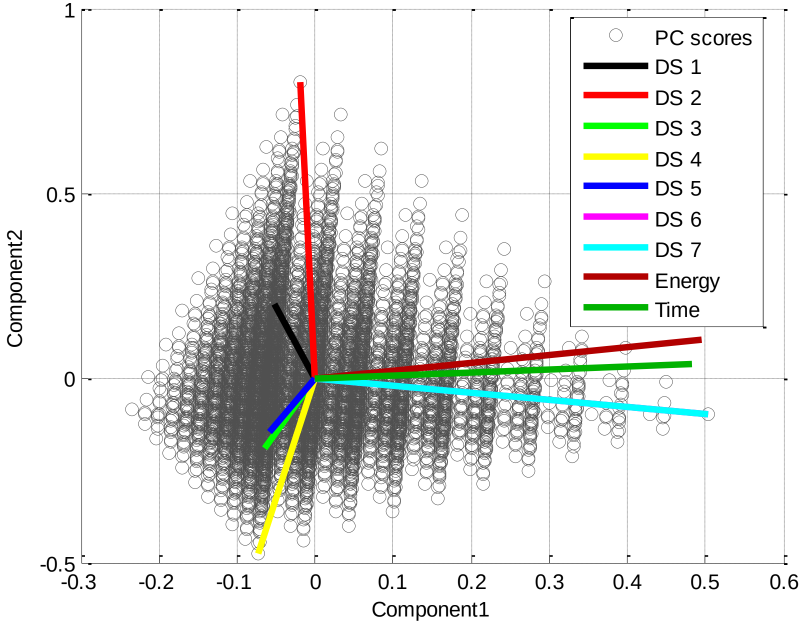
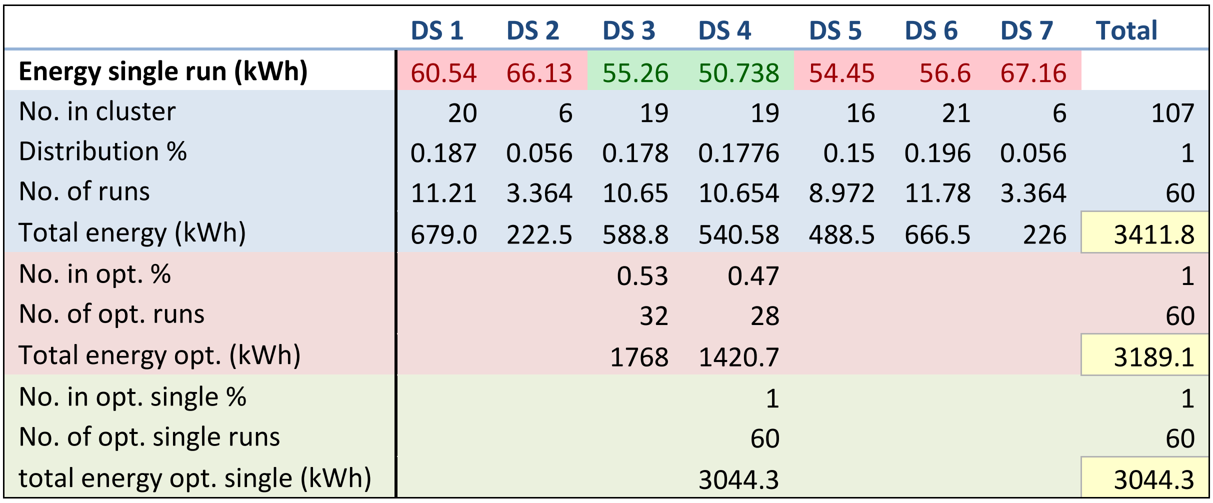


Figure 94 – Principal component biplot of Driver style relationships to energy and time Liverpool to Ormskirk

5.5.3.1 Daily energy consumption estimation

The three energy consumption estimations are made again for Liverpool to Ormskirk, in this case the shorter route follows a daily 60 service timetable. Based on the driver culture taken from real data an energy estimate of 3411.8 kWh is given, shown in Table 16 (blue section), optimal driver styles 3 and 4 produce 3189.1 kWh, a 6.5% saving (red section) and the lowest energy driver style only estimate is 3044.3 kWh an 11% reduction.

Table 16– Energy consumption estimations Liverpool to Ormskirk



	DS 1	DS 2	DS 3	DS 4	DS 5	DS 6	DS 7	Total
Energy single run (kWh)	60.54	66.13	55.26	50.738	54.45	56.6	67.16	
No. in cluster	20	6	19	19	16	21	6	107
Distribution %	0.187	0.056	0.178	0.1776	0.15	0.196	0.056	1
No. of runs	11.21	3.364	10.65	10.654	8.972	11.78	3.364	60
Total energy (kWh)	679.0	222.5	588.8	540.58	488.5	666.5	226	3411.8
No. in opt. %			0.53	0.47				1
No. of opt. runs			32	28				60
Total energy opt. (kWh)			1768	1420.7				3189.1
No. in opt. single %				1				1
No. of opt. single runs				60				60
total energy opt. single (kWh)				3044.3				3044.3

5.5.4 Ormskirk to Liverpool

The return journey on the Ormskirk section of the Northern line, like Hunts Cross to Southport and its return journey, has 6 driver styles identified by optimal clustering. Unlike its outward journey the driver styles are more centrally grouped suggesting a greater degree of drift between similar driver styles rather than well-defined boundaries which can be seen on the PC cluster plots in Figure 95 and Figure 96. Cluster 3 appears as the least compact cluster, data is grouped centrally and to the right of PC1 which suggests a large variance. Style 3 can be characterised as mid to high energy and acceleration, it also has the highest energy consumption but not the shortest journey time. Style 1, 2 and 4 all have similar or shorter running time to 3 and in each case have approximately 10% less energy consumption. 25% of the runs fall into the driver style 2 which is a mid-energy, high acceleration with coasting driver style.

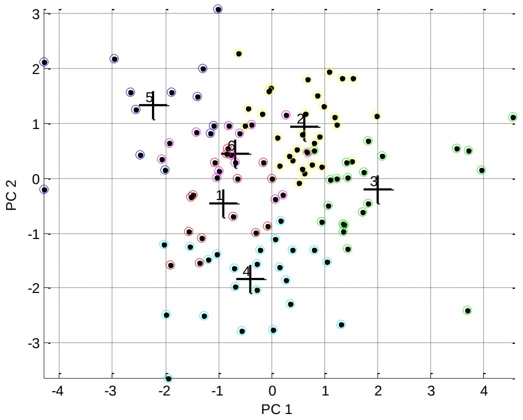


Figure 95 – k-means clustering Ormskirk to Liverpool (PCs 1 & 2)

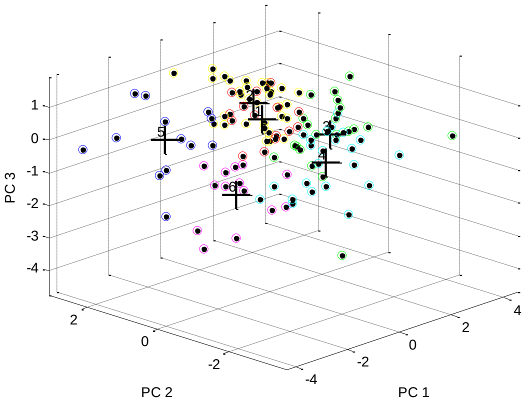


Figure 96 – k-means clustering Ormskirk to Liverpool (PCs 1 – 3)

Based on the PCA of the permutation matrix of this route, Figure 98, driver styles 1 and 2 are selected for peak and off peak optimal styles respectively. Both driver styles show a slight proportional relationship with energy consumption in PC1 but have a significant inverse relationship with both energy and time in PC2. This suggests that more journeys made with these driver styles should produce an energy consumption estimate that is lower than the existing driver culture; this will also have very little impact on journey time.

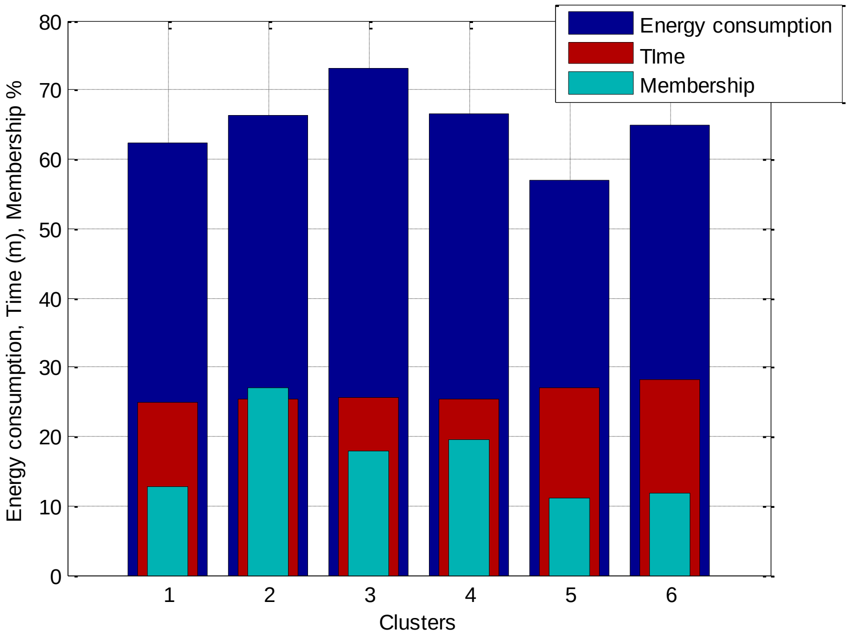


Figure 97 – Time and energy distribution of clusters Ormskirk to Liverpool

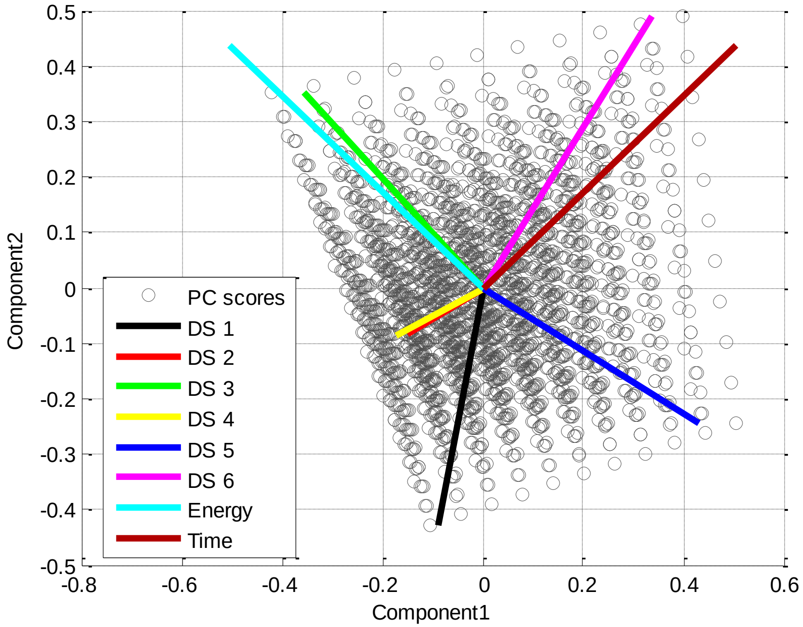
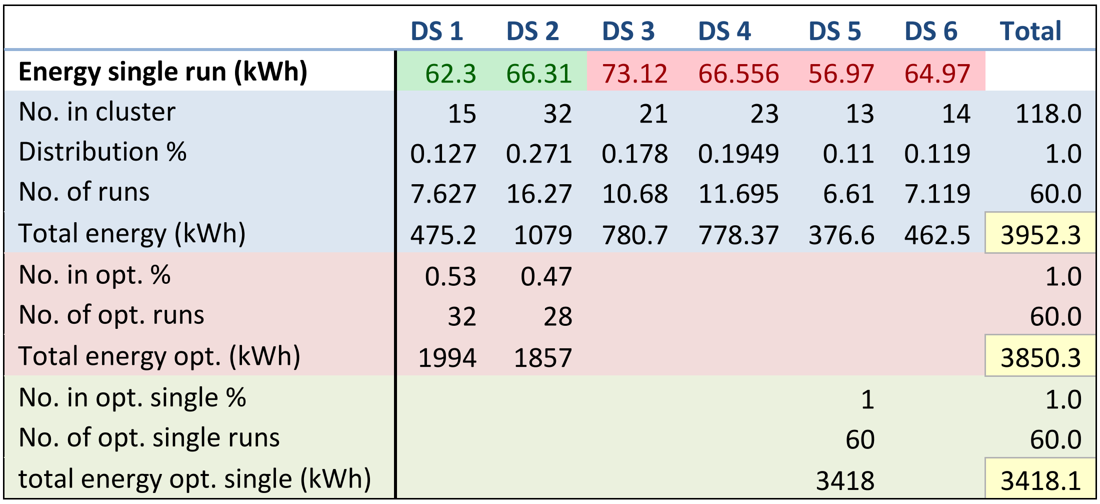


Figure 98 – Principal component biplot of Driver style relationships to energy and time Ormskirk to Liverpool

5.5.4.1 *Daily energy consumption estimation*

Table 17 shows the final energy consumption estimates made for the Ormskirk to Liverpool route, the driver culture from real data results in a daily energy consumption of 3952.3 kWh and is shown in the blue section. Using the optimised driver culture of driver styles 1 and 2 the energy estimate is 3850.3 kWh (red section) a 2.6% reduction over the original driver culture. The green section, lowest energy driver style only, gives an estimate of 3418.1 kWh which is a substantial 14% energy consumption reduction.

Table 17 – Energy consumption estimations Ormskirk to Liverpool



	DS 1	DS 2	DS 3	DS 4	DS 5	DS 6	Total
Energy single run (kWh)	62.3	66.31	73.12	66.556	56.97	64.97	
No. in cluster	15	32	21	23	13	14	118.0
Distribution %	0.127	0.271	0.178	0.1949	0.11	0.119	1.0
No. of runs	7.627	16.27	10.68	11.695	6.61	7.119	60.0
Total energy (kWh)	475.2	1079	780.7	778.37	376.6	462.5	3952.3
No. in opt. %	0.53	0.47					1.0
No. of opt. runs	32	28					60.0
Total energy opt. (kWh)	1994	1857					3850.3
No. in opt. single %					1		1.0
No. of opt. single runs					60		60.0
total energy opt. single (kWh)					3418		3418.1

5.5.5 Total energy savings

Total energy savings for the routes analysed are presented in this section, each routes optimal and maximum energy savings are considered and two total savings estimates are provided. In each case the savings are a 'percentage of a percentage' of the total energy consumption which is the combined energy consumption for each route. Table 18 below summarises the energy consumption of each routes existing driver style, its contribution to the total energy consumption and their savings.

Table 18 – Summary of Driver culture energy consumption and energy savings

	Energy (kWh)	% of total energy	Opt saving %	Min saving %	Saving % of total energy (opt)	saving % of total energy (min)
Hunts Cross to Southport	8294.8	34.2	1.5	8	0.5	2.7
Southport to Hunts Cross	8587.8	35.4	5.4	9	2	3.2
Liverpool to Ormskirk	3411.8	14.1	6.4	11	1	1.5
Ormskirk to Liverpool	3952.3	16.3	2.4	14	0.4	2.3
Total	24246.8	100			4.0	10.0

The totals for the optimal driver culture and minimum energy consumption driver cultures in Table 18 show that there is a potential for a 4% to 10% energy reduction possible over all of the routes examined. These routes comprise of approximately half of Merseyrail and so do no fully describe the possible energy saving potential of altering driver styles across the entire network, however, a 4 – 10% energy saving would provide substantial operational cost reduction where upgraded traction packages or in cab driver advisory systems are unlikely to be fitted or are financially restrictive.

5.5.6 Optimised routes and control input

As well as potential energy consumption reductions a second output for this project was to produce recommendations to assist Merseyrail driver managers in developing energy saving driver strategy. This can be achieved by identifying optimised trajectory and control input from the driver culture analysis. However, as cluster centroids represent the mean distance between members of each cluster they do not represent real data. They can be converted back into seven dimensional input data (representing time, energy, unpowered deceleration and notch 1 – 4) by the following steps:

- Multiplying their PC score by the transpose of the retained principal components matrix
- Adding mean and standard deviation

This is the reverse process of generating the PC scores, but these results are not indexed to original recorded data as they do not represent individual data points. To identify an optimal trajectory and the control input that produces it, the closest PC score data point is used to represent the cluster centroid. Figure 99 to Figure 106 show trajectory and control input for peak and off peak driver styles and the minimum energy consumption style for each route. As a general approach driver cultures provide an overall strategy for a particular route, for example on the Hunts Cross to Southport route a combination of styles 5 and 6 for the optimal and 2 for the energy saving cultures are identified. This suggests that maximising low to mid acceleration with increased amounts of coasting is the best way to reduce energy consumption whilst adhering to the timetable on this route. Driver cultures are however, too vague to provide accurate advice or recommendations to drivers, they are perhaps best used as a checking system to identify whether drivers are following driver strategies recommended by instructors. As a result actual trajectories and control input for both optimal and energy saving driver cultures for each route are identified and demonstrated here. In most cases optimal driver strategy is identified via complex simulation providing trajectories that are likely to be substantially improved over those presented here. However, these are based on real data, are connected to actual control input (often not the case with simulated optimised trajectory) and most importantly are within the capability range of drivers; they show optimisation by driver best practice rather than mathematically. This is considered to be a novel method of driver style optimisation and addresses the documented problem of how to efficiently convey the trajectory in terms of control to the driver.

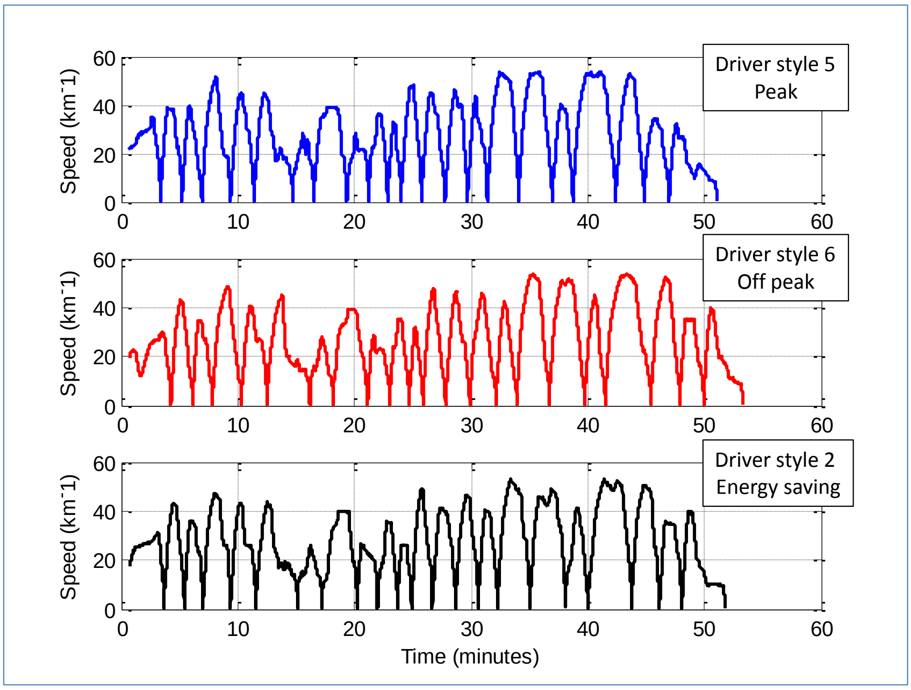


Figure 99 – Hunts Cross to Southport optimised and energy saving trajectories

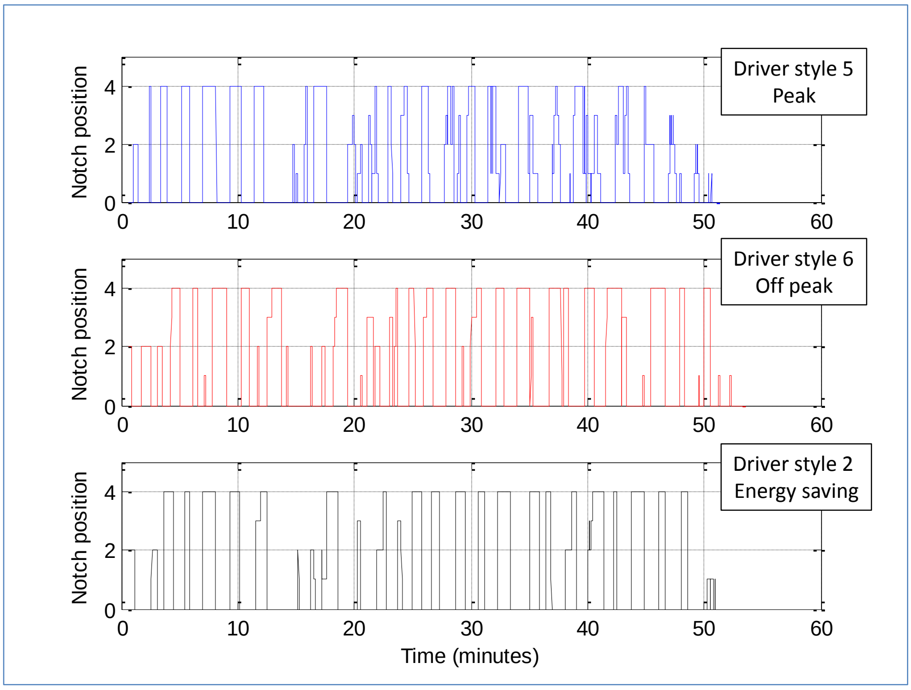


Figure 100 – Hunts Cross to Southport optimised and energy saving control input

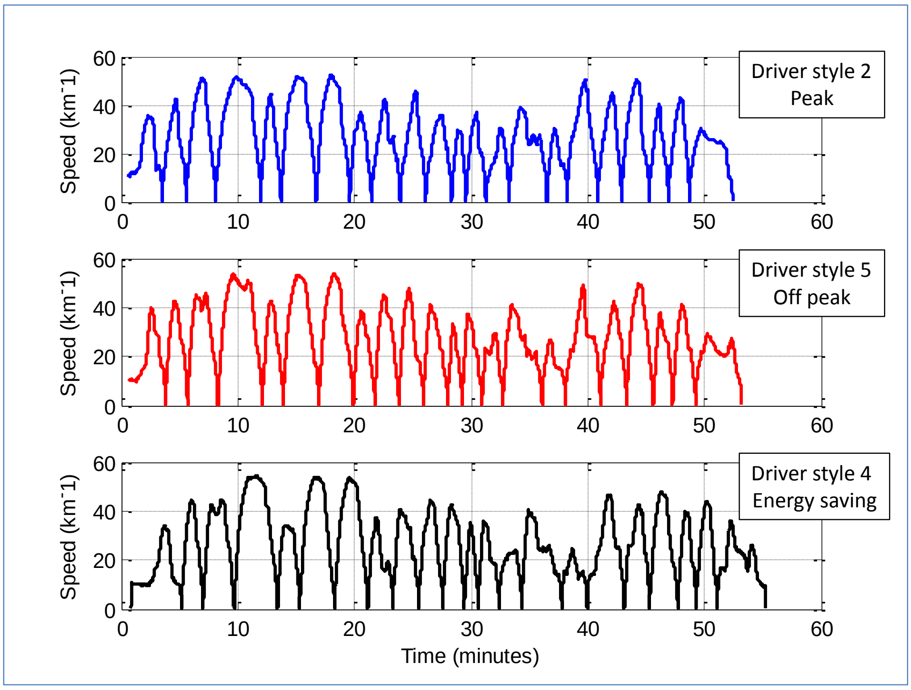


Figure 101 – Southport to Hunts Cross optimised and energy saving trajectories

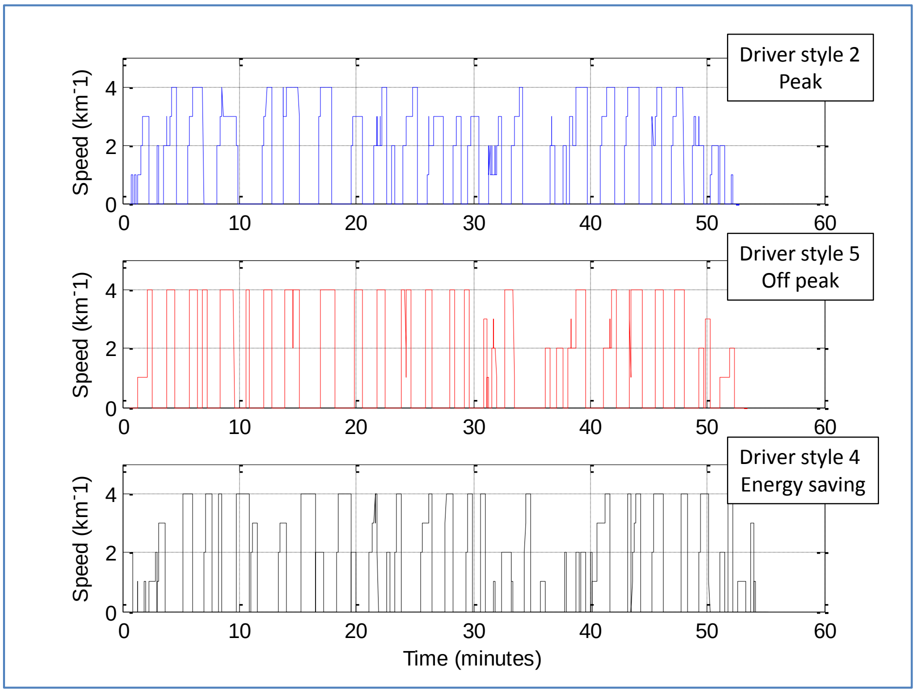


Figure 102 – Southport to Hunts Cross optimised and energy saving control input

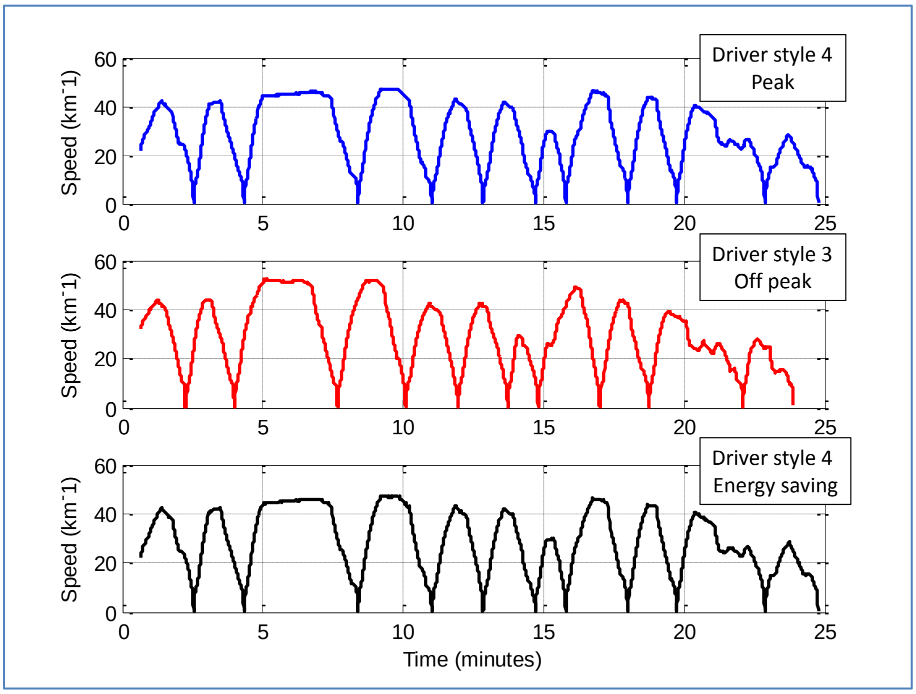


Figure 103 – Liverpool to Ormskirk optimised and energy saving trajectories

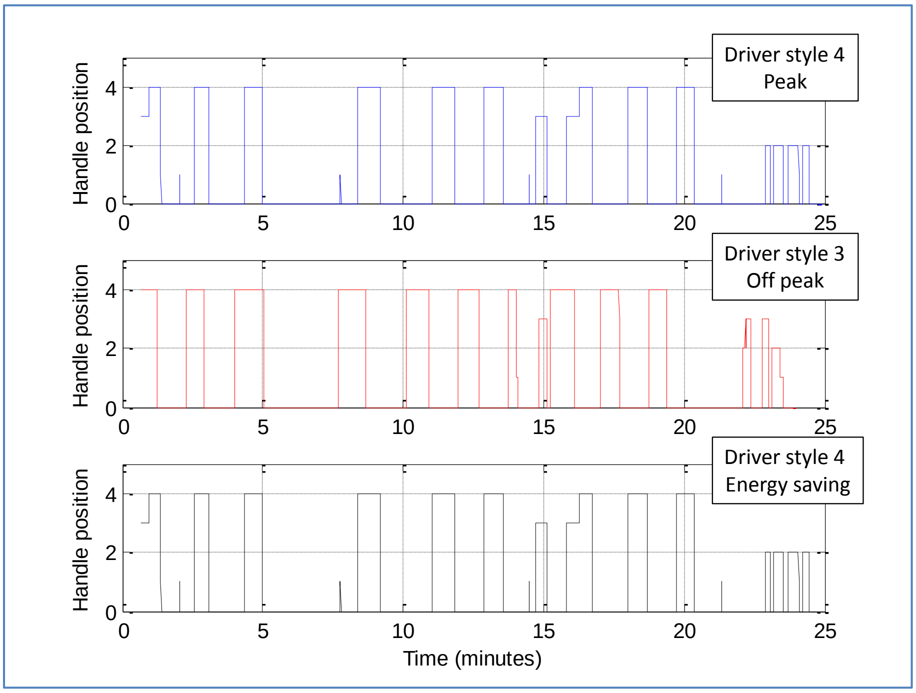


Figure 104 – Liverpool to Ormskirk optimised and energy saving control input

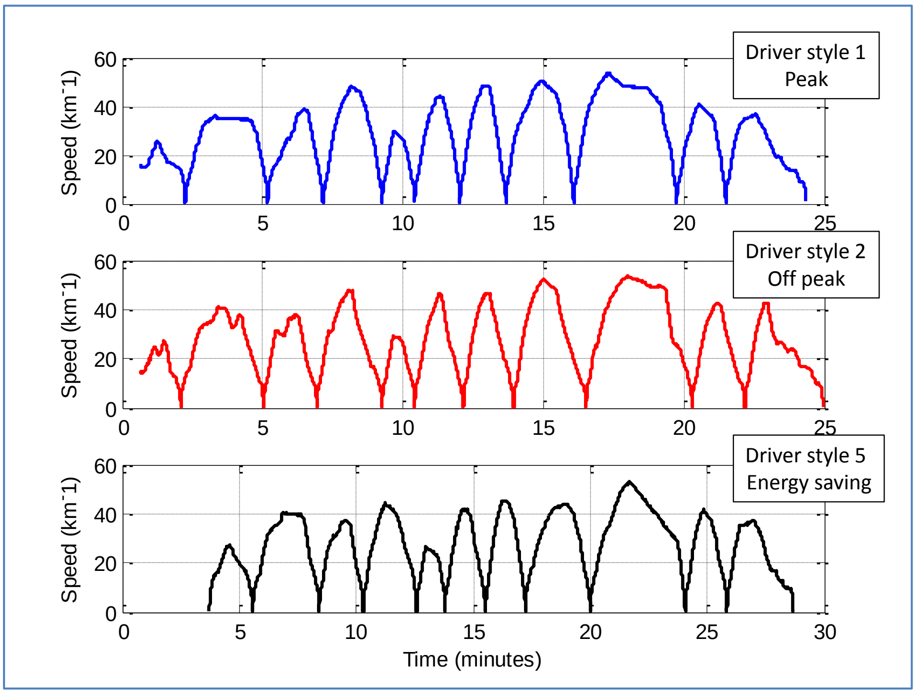


Figure 105 – Ormskirk and Liverpool optimised and energy saving trajectories

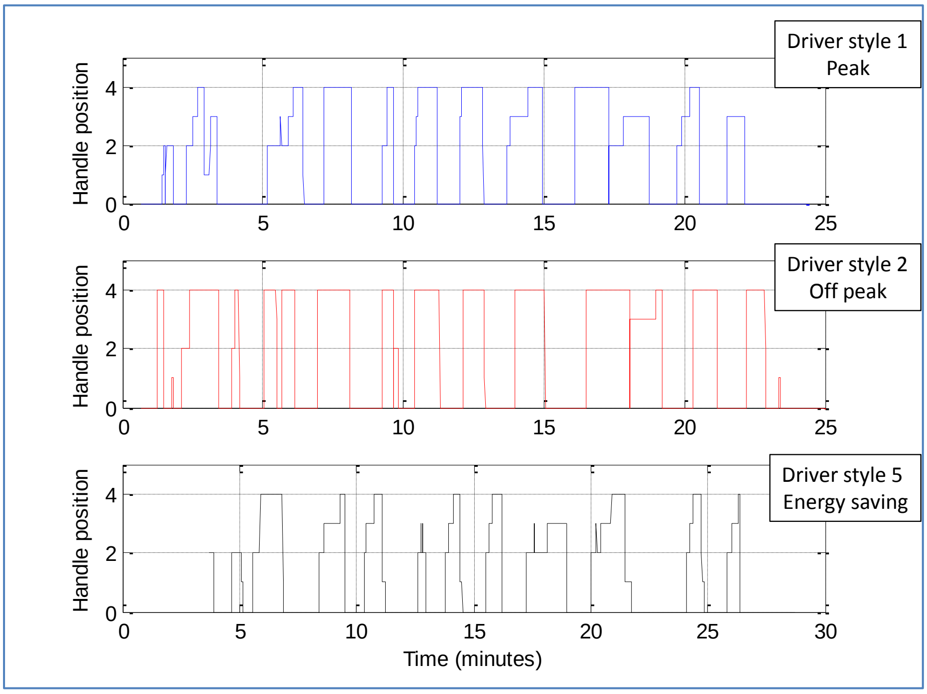


Figure 106 –Ormskirk to Liverpool optimised and energy saving control input

5.6 Discussion

Results show that there are potential energy savings that can be made with the introduction of optimised or energy saving driver cultures on the Merseyrail network. Based on the optimised driver styles selected for each route savings of 4% are likely, with the minimum energy driver styles 10% savings are suggested. In a number of cases the journey times that are required for energy saving driver styles to be implemented cannot follow the timetable meaning they might be inappropriate for use, particularly during peak operation. For example during off peak operation where passenger traffic is very low services might be removed and the slower, energy saving driver styles implemented. A system like this is already used on the late services on the Liverpool to Ormskirk route where there are only two services per hour due to reduced passenger need. It is likely that the removal of 1 service (three services per hour) during the lowest passenger number periods coupled with the lowest energy consumption driver style would result in further savings. This is a possible solution to allow the use of slower but greater energy saving driver styles.

Identification of driver cultures and altering them to select optimal driver styles or the lowest energy consumption style has been shown above to provide energy savings that are significant. 5 – 25% energy savings have been shown to be possible from a number of other rail projects that are in research stage or have been put into operation that are directly related to affecting driver control [90, 91, 92]. In most cases these projects involve the development of sophisticated in cab systems and require substantial financial investment from a rail operator to be a realistic prospect. As mentioned in the results section the type of analysis presented in chapters 3 – 5.5 has the potential to be substantially cheaper and does not rely on installation and development of new technology, as a result may be an appealing option to rail network with similar financial constraints as Merseyrail. This being said the analysis carried out here is not without problems; the following discussion will highlight a number of these and propose some solutions. With the problems considered the value of this analysis has to be decided, and how might this analysis be used best in future applications.

5.6.1 Personnel cooperation

Altering driver styles to give lower energy consumption distributions is reliant on driver instruction and co-operation. Currently there is no system in place that incentivises improved driver performance with regards to reducing energy consumption. Drivers are currently instructed to drive with maximum acceleration, unless poor adhesion conditions exist and brake at the latest point to adhere to the timetable [60]. Data where coasting is prevalent has been attributed to ‘legacy training’ driving through discussions with Merseyrail personnel. For example, in the past coasting points were used across the network but have become neglected and are no longer significant parts of driver instruction [60]. Much of the discrepancy between differing driver styles is thought to reflect overlapping training regimes as well as driver competency and willingness to provide high performance. In a practical application without any kind of incentive it is unlikely that implementation of an energy saving driver culture would have 100% participation.

5.6.2 Accurate selection of optimal control

Secondly although optimal driver cultures were selected using the data analysis methodology described above it is difficult to return to the exact original driver input which provided it. This is a result of a data processing step that sum the collective use of individual notch settings and unpowered deceleration (coasting) meaning that the original driver input is not preserved. As a result optimal driver cultures are very general for a given route and do not give specific instruction on how to tackle each part of the route; they show the total use of each control input but not the 'where and when'. A solution to this is to preserve all driver input data and to match PC score points to these original points. This is possible as PCA generation of PC scores is completely reversible. If PCs were removed as they contained little relevance to data trends, such as all examples in 5.3, there is a degree of information loss but this is not problematic as input data indexing is not affected and matches to original input data can be made. The greater problem is related to the location of the cluster centroids that are used to define the individual driver styles and cultures. The centroid is the mean of all points within the particular cluster and as a result does not have an original data input that can be retrieved. Instead a point that is close to the centroid can be selected and used to give an example of a likely driver input that originated from the centroid position. However, in a number of instances the centroid is the mean of a disparate cluster where the lack of compactness means that a driver input that is close enough to the centroid cannot be accurately selected. Figure 107 shows selection of a good centroid representative data points and a cluster with low compactness where a good representative data point cannot be accurately selected.

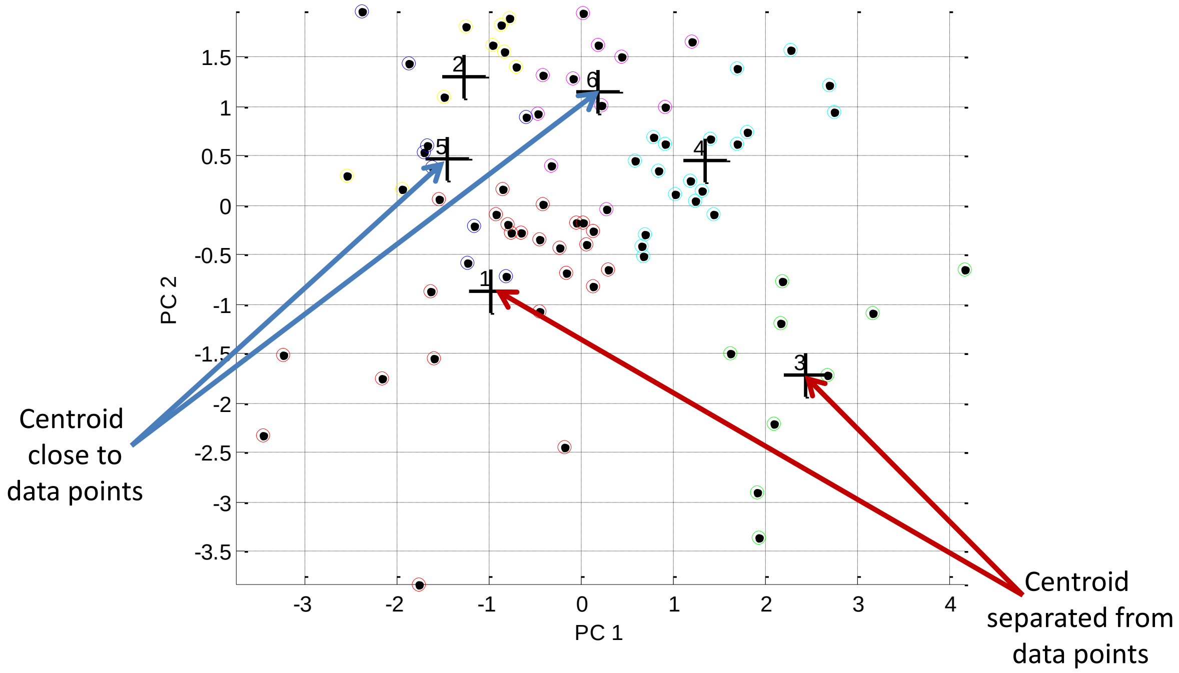


Figure 107 – PCA clustering showing centroids that can be represented by close data points and uncompact clusters where centroids cannot be represented well

Figure 108 shows control input data for three data points close to centroid 5 in Figure 107, all are valid selections for representations of the optimal control that the centroid represents, but only two of these selections have close similarities. The bracket on the lower two runs show sections which are very similar, after this the lower run begins to differ and eventually finishes the journey almost one minute later. The upper run, although having a generally similar stepped pattern, is distinct from the lower two runs. All three however fit closely to the same driver style as their combined control input, running time and energy consumption are very similar, the vehicle trajectory for all three is also very similar.

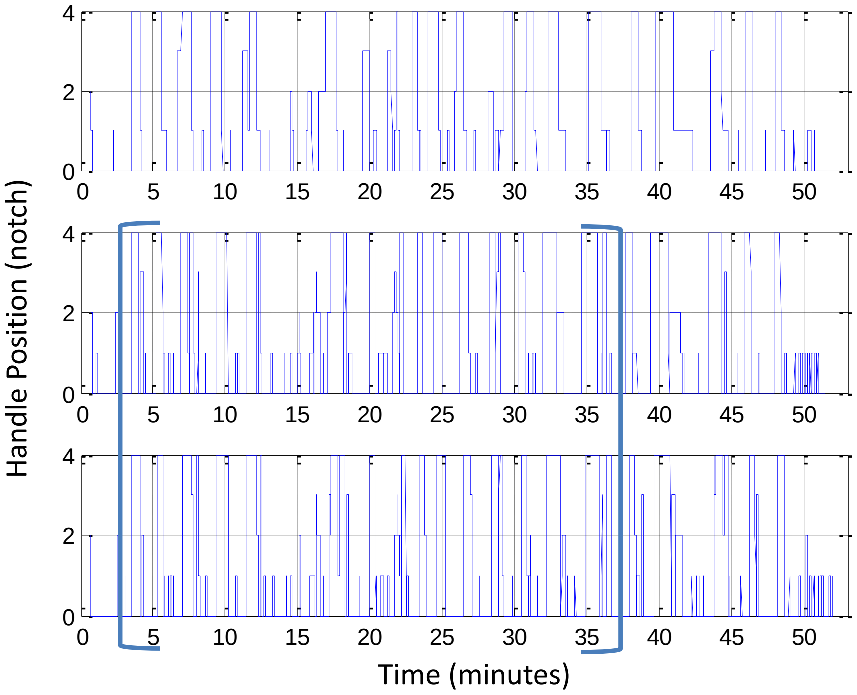


Figure 108 – Three driver inputs belonging to the same cluster group, brackets highlight areas of similarity

The main issue is that the selection of control input that represents the optimal driver style might not actually be well matched as a result the control input that would be recommended to a driver might not necessarily be the optimal approach that is required. A possible solution to this would be to take an average control input from nearest data points to the cluster centroid. However, as data points which were close would undoubtedly be varied in number and distance between each route examined it would be difficult to pinpoint the exact number that should be used to provide this average control input so that it was consistent.

5.6.3 Snapshot of driver culture

A further consideration is that this analysis is only a ‘snapshot’ of driver activity on a single vehicle over a three month period. It is likely that a large portion of the driver population on the Northern line had the opportunity to drive the instrumented vehicle but this is by no means certain, as a result it is not possible know if examples of all drivers are captured by the data. As well, there is no information available to determine whether drivers who were recorded have an equal amount of time in the instrumented vehicle as each other. In the worst case the data could represent a single driver producing the entire driver culture or, in the best case, there are at least two or three examples of each driver. The result of this uncertainty prevents these results, particularly the optimised trajectories, being presented as final driver instruction recommendations to Merseyrail. But this project does serves to demonstrate the potential of high resolution data capture and analysis.

5.6.4 Environmental Factors

Environmental factors such as varied weather, temperature, rail adhesion caused by ice or snow or leaf fall and high winds undoubtedly influence driver strategy to maintain safe operation. One of the initial goals of this research was to tie the rich energy data sets from the instrumented vehicle to weather data, this is also described in Appendix 10.1.2. Unfortunately, weather data does not have high geographic resolution i.e. there is only data from two or three weather stations separated by tens of kilometres, and only sampling once per hour. This is sufficiently high resolution for weather forecasting but not to establish driver behaviour in the event of an environmental change. Secondly, information regarding adhesion quality is kept only short term and was never available for comparison during this research, which leads to a possibility that driver style characteristics might be the result of nothing more than adaption to changing environmental characteristics. Even with this possibility it is not thought to be the case. The data used in this research is taken from the autumn winter period where a general more defensive driving approach is recommended, as a result there is an expectation that a large portion of the driver population would apply this to maintain safety, in effect somewhat removing the effects of poor adhesion and poor weather conditions. Had the data used been taken from a seasonal transition period, summer to autumn for example, there might be a more significant effect on driver style caused by environmental factors. Establishing a seasonal change would be an interesting finding from a follow up research project. It is also thought that only extreme environmental factors have enough of an effect to change to overall style of a driver's behaviour. As these factors are rare they are unlikely to be well represented in the dataset. A follow up research project that had access to driver control, weather data and specific driver would be able to make comparisons of 'normal' driver behaviour to that which is affected by an extreme condition, a storm for example. As this data is not available it is not possible to validate at present. This uncertainty as well as those mentioned above reinforces that final recommendations cannot be given to Merseyrail based on these findings alone, however the interesting findings and unanswered questions could develop into a potentially rewarding research area that can be supported by a number of follow up research projects.

5.7 Conclusion

It is very unlikely that technologies such as regenerative braking, hybrid traction or energy storage systems will be implemented on the Merseyrail network while Class 508 rolling stock is in operation. It is likely that a DAS system will be present in new rolling stock that will be introduced in the next five years. As an interim solution to reducing energy consumption the identification of performance indices and their use in driver training is an achievable and relatively short term solution that has the potential to provide energy savings on the Merseyrail network. Preliminary observations suggested that metrics in acceleration and deceleration can provide simple characterisations of driver performance which can be evaluated in terms of energy consumption. These observations were used to develop a methodology that identified different driver styles in operation on four routes and how their use affects both energy consumption and running time; this was termed a driver culture. Potential energy savings of 4 – 10% would be possible if driver cultures comprised of optimal and energy saving driver styles were introduced. This would require continuous monitoring of driver performance, data analysis and driver instruction based on this. Maximum savings are also dependant on driver cooperation. It is reasonable to assume that if driver instruction could target the adoption of distinct driver approaches that were suitable for the timetable demand in operation that there is potential to achieve significant energy savings.

This might be seen as a very simplistic driver advisory system. Rather than real time driver commands being relayed to the driver, flexible but monitored driver styles were taught to drivers. With access to their own performance data over short time periods drivers could evaluate how they have performed and plan how to improve energy consumption. This is especially applicable to rail networks like Merseyrail where driver mileage is relatively low and the network is small, both in route miles and financial capability. As a result the main recommendation to give to Merseyrail should be that instrumenting their vehicles long term, across a greater degree of the fleet would be the next step in continuing this analysis. A second aspect of a possible second round of instrumentation would be frequent short term analysis. This would necessitate the introduction of an automated data retrieval system and a standalone analysis system that could provide output results without the need of manual pre-processing and a dedicated analysist. Ideally data would be presented to Merseyrail personal much like performance data that is common in the sports world [93]. Strava, a performance analytic application, is commonly used by cyclists to monitor continuous performance over routes cycled, examples are shown in Figure 109. The individual has access to various data, in a graphic format, which shows how a particular journey was made, and where improvements might be made, there is also the option to compare to other cyclists on the same or similar routes.

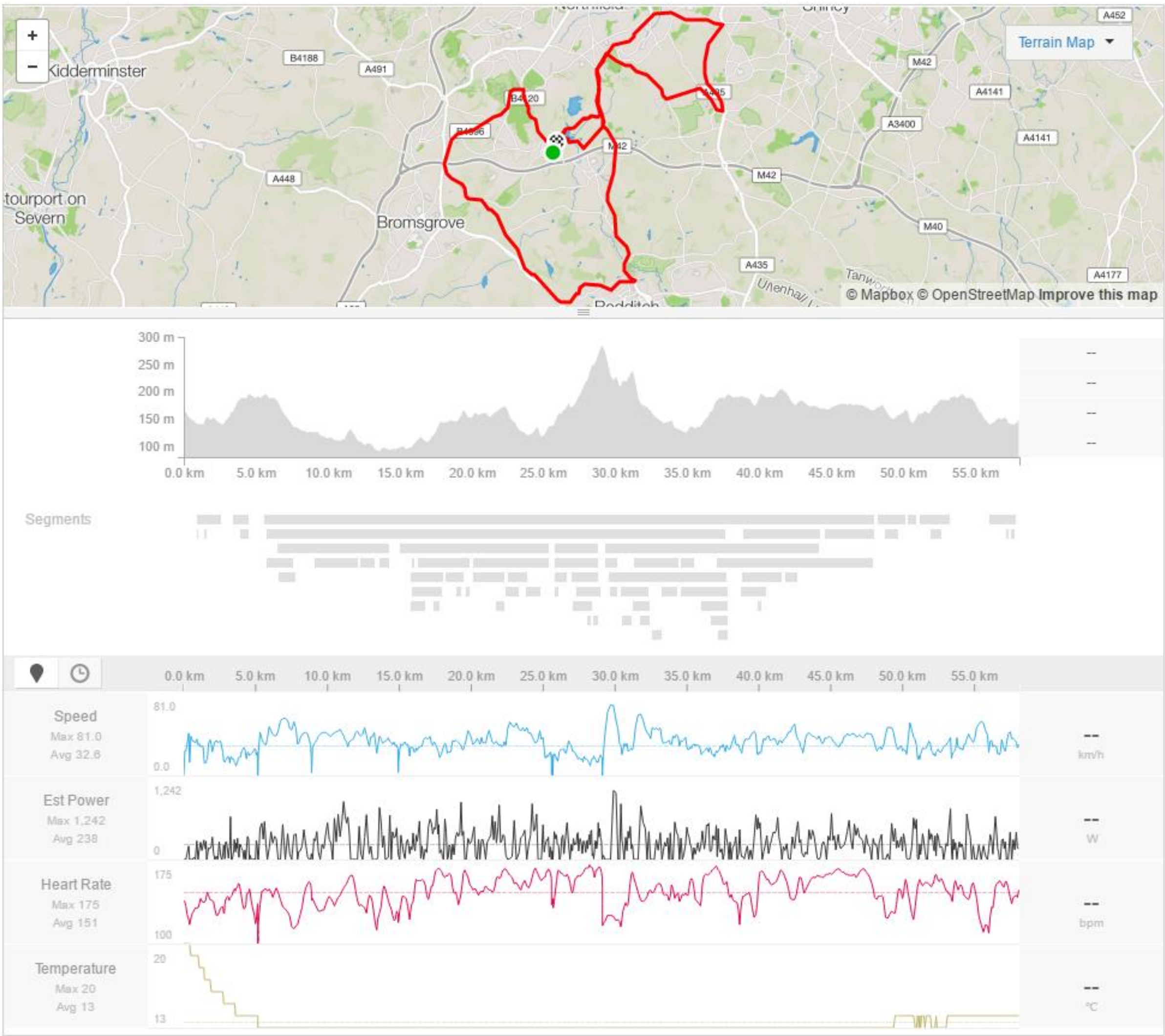


Figure 109 – Strava output graphs showing cyclist performance over an established cycling route

The final goal of the analysis and results provided by the Merseyrail energy monitoring project would be the development of a similar system where Merseyrail personal have access to energy consumption data, tied to driver control data, and an indication of how drivers are performing, in a very similar vein to the Strava application.

5.8 Summary

Chapters 3 to 4.6 have discussed various aspects of the Merseyrail energy monitoring project which are summarised shortly below:

5.8.1 Instrumentation

In 2010 three substations and a British Rail Class 508 were instrumented and data were collected over a three year period. These data were used to identify electrical system losses in the Merseyrail Network. An advantage of the instrumentation equipment was that driver control was logged and was able to be matched to energy consumption and route data. This allowed for an investigation into the effects of varying driver control on energy consumption and the development of a methodology to identify different driver styles from large datasets taken from the instrumented vehicle.

5.8.2 Initial Observations from datasets

Observations of six journeys were made from data taken from operation on the Hunts Cross to Southport route of the Merseyrail Northern line. These observations showed significant variation between driver technique and produced a range of energy consumption. Approximately 75% of journeys were shown to have ‘efficient’ driver technique, where high energy consumption journeys resulted in low running time and vice versa. This was expected and demonstrated to be similar in simulation. High energy consumption high journey time examples were observed in approximately 25% of journeys examined. It was shown that these were likely a result of drivers not using coasting, cruising in less efficient control notches and 'low notching', where drivers decelerate with low traction power setting engaged. These observations were used to design metrics for assessing driver style by using energy, pure run time (journey time minus dwell time) and distance travelled in each control notch. These metrics were used as input data in simulation and data clustering. Information in this chapter culminated in the publication of a journal paper published in the Journal of Rail and Rapid Transit [5].

5.8.3 Simulation

A model of the instrumented vehicle was designed with a modified version of the University of Birmingham's single train simulator (STS). Simulations were used to generate a large synthetic data base; this was used to test different analysis approaches and assess the suitability of a clustering algorithm to be used on real data.

5.8.4 Data decomposition and clustering

Due to problems associated with high dimensional data the input datasets were processed using principal component analysis to produce two and three dimensional datasets. These expressed the underlying structure of the original input. This allowed for analysis of individual driver styles used on four different routes on the Merseyrail network. K-means clustering was used to group the new datasets into 6 (for three routes) and 7 (for 1 route) distinct driver styles.

5.8.5 Driver culture

The distributions of each driver style on a particular route are termed driver cultures, energy consumption from each driver culture was calculated as well as an optimal driver culture (reduced energy consumption with low impact on running time) and an energy saving driver culture (use of lowest energy consumption driver style only, possible affecting journey time).

Energy consumption savings were made by comparing the existing consumption to that resulting from optimised and energy saving cultures. Savings ranged from approximately 2 – 16 % between different driver cultures and particular routes. Total energy consumption savings for the combination of routes (approximately half of the network) were estimated to be between 4 and 10% depending on whether optimal or energy saving driver styles were in use.

5.8.6 Optimal control and trajectory

Each route had and optimised and energy saving trajectory and control input for each route identified which if followed by drivers could result in energy savings similar to those identified. This is not able to be validated for this project as there are a number of constraints that prevent short term adoption of new driver instruction, as a result the implementation of identified trajectories and control is an area that is intended for a possible future continuation of this project.

6 ATOC Partial Fleet Metering

Responsible energy use is one of the key challenges of the early 21st century, with the prospects of dwindling fossil fuel reserved, population increase and the energy cost of modern high energy life styles requiring society to be able to accurately monitor energy use and account for inefficient use. The rail industry is the UK’s largest consumer of electricity, is dependent on accurate energy monitoring to provide fair energy use billing to operating companies as well as to demonstrate its role in carbon emission reduction. The vast majority of modern rolling stock comes fitted with energy monitoring systems as standard, however, as rolling stock tends to have a full service life of 20 – 30 years there remain a significant number amount of vehicles in operation in the UK that were put into service prior to the mass adoption of individual vehicle energy monitoring. For these vehicles energy meter retro-fitment is the most appropriate solution to this problem, The capital costs of retro-fitment are however, very high across large fleets and it takes a significant period of time to recover the initial costs via improved-accuracy billing. As a result full fleet metering of older rolling stock uptake is slow; instead operating companies might choose to only partially meter their fleets [94]. This allows for a portion of their fleet to be accurately billed. The portion that remains unmetered is billed based on dated energy consumption models, which likely overestimate energy consumption and are no longer valid for use. Thus, the present situation is essentially a trade-off between significant upfront costs or the possibility of annual over billing.

A billing model that uses partial fleet metering (PFM) data to provide accurate energy consumption estimates for a whole fleet is therefore desirable for train operating companies (TOCs) to reduce the impact of initial metering costs and to provide more accurate energy billing.

6.1 Background

The methodology presented here has been developed over the course of three research projects:

- A simulation project
- A small AC EMU fleet project
- A whole DC EMU fleet project

The main aim of this work is to determine the uncertainty associated with extrapolating the total energy consumption of an entire fleet from a subset of metered trains and identifying the minimum size a subset should be to produce an accurate billing estimation. The methodology section outlines a method to characterise the statistical error involved in estimating full fleet energy consumption from a subset of trains. Two approaches were investigated, these are termed:

- Whole network approach
- Route specific approach

Specific data processing requirements for each approach are given before the general methodology is presented. This is followed by case studies detailing the application of each approach to a large dataset. The following section gives a short introduction to two research projects [95] that were used to develop the methodology.

6.1.1 Simulated data project

A project to develop a methodology to allow partial fleet metered energy billing was conducted using simulated data [96]. This project used the University of Birmingham's single train simulator (STS) and Monte Carlo analysis to develop and validate the methodology before its use on real data. Synthetic data were generated to approximate the operation of a DC network in operation over two routes (100 miles and 30 miles) for 18 hour operational periods. A dataset representing a fleet in operation for a year was required. Energy consumption of a fleet was assumed to be normally distributed; a range of energy consumptions to be were simulated. This involved defining a range of operational parameters including:

- Driving style changes
- A range of passengers per train based on a national travel survey [97]

A normally distributed selection of energy consumption values were generated from these simulations allowing random selection of energy consumption rather than individual simulations for each journey. Findings showed standard deviation of estimation error (σ_{ε}) from partial fleets of greater than 5% of the original fleet size was 1%. This suggests that with this partial fleet size energy consumption estimation would be accurate to ±1% of the actual consumption with a probability of approximately 68%. These initial results indicated the potential for applying the methodology to operational data.

6.1.2 First in-service data project

In March 2012 energy data from two commuter routes was collected and used analysed using the same methodology as the simulation dataset. Raw data included GPS longitude and latitude which allowed energy data to be matched to relevant service codes. Approximately 1000 journeys were analysed for fleet sizes of approximately 30 and approximately 70 units each units of differing rolling stock. In this case σ_{ε} of 1% was obtained with a partial fleet size of greater than 40%. This difference was attributed to a wider distribution of energy data input indicating that real data is significantly less homogenous than that developed for the synthetic data. Normalisation of energy input data by distance and time significantly improved the analysis producing a σ_{ε} of 1% with a partial fleet size of 15 – 20% of the original fleet.

6.2 Methodology

6.2.1 Recorded data

The consistency of the error characteristic from a partial fleet is observed using two different approaches, whole network and route specific. Data input used for the analysis consisted of the following variables:

- Time
- Vehicle ID (single power car)
- Latitude
- Longitude
- Active energy
- Regenerated energy

Each vehicle is fitted with an electricity meter that logs the above data once every minute. For the whole network approach, route variation is not considered, instead daily operational activity of all units are compared. The route specific approach requires extra data processing steps, this is used to isolate individual journeys and group them with matching routes before comparison.

The two approaches were designed to test differing data quality situations. In the event that location data is not available or of low quality the whole network approach is suitable, this approach has lower complexity but comes with a penalty to the accuracy of the analysis; the result is an increased partial fleet size requirement. The route specific approach involves increased complexity due to additional data processing steps, but results show that partial fleet size requirement is reduced. As the results and data are commercially sensitive steps to censor the operating company and various aspects of the project have been made. Censoring covers the following areas:

- Operating company name.
- Operating company location.
- Area of operation.
- Route name and location.
- Rolling stock class.
- Rolling stock unit configuration.
    - This details the number of vehicles in a particular unit; three configurations were examined and have been titled configurations A, B and C.
- Fleet size by number.
- Description of characteristic operating features that could be used to identify the particular network.

Datasets for the year 2014 were obtained for individual electrical meters for an entire DC EMU fleet. During this period a number of vehicles were not in operation, and a subset operated on infrastructure that was not considered in this research. As a result the ‘total fleet’ that was examined consisted of approximately 80% of the actual fleet size that exists.

Raw data are poorly sorted and contained in .csv files which are too large to be handled by spreadsheet software such as Microsoft excel. Memory constraints also made analysis not possible. As a result a significant amount of time was spent dividing raw data into smaller files. These were sorted based on vehicle number, day and time. Raw data was obtained as separate files for each month; a single month takes approximately 2 – 3 days to separate, sort and then recombine. After this pre-processing step vehicles data are matched to their relevant unit number to allow daily totals for both energy meters to be calculated. The last stage of processing is to prepare input data for analysis. For each day following totals are recorded in a row vector:

- Active energy
- Regenerated energy
- Distance travelled
- Time in operation

As well as

- day,
- month and
- unit number

This results in a structure array of elements (representing each unit) containing an array of daily operation over the year. Active energy, regenerated energy, distance travelled and time in operation are stored cumulatively, representing each unit’s yearly total. These values are taken as the input for analysis.

6.2.2 Data normalisation

As data input for the whole year varies significantly it must be normalised to allow comparison. Normalisation is made by considering:

- Energy consumption rate
- Specific energy consumption

Energy consumption rate is defined as:

$$E_r = \frac{E_{\mathrm{t}}}{T_{\mathrm{t}}} \tag{24}$$

Where E_{t} is the cumulative total net energy consumption (active - regenerated) and T_{t} is the cumulative total time in service.

Specific energy consumption is defined as:

$$E_{\mathrm{e}} = \frac{E_{\mathrm{t}}}{S_{\mathrm{t}}} \tag{25}$$

Where E_{t} is the cumulative total net energy consumption (active - regenerated) and S_{t} is the cumulative total distance travelled. Normalisation using time and distance have shown similar results in previous projects [95, 96]. Energy meters do not provide a distance measurement; distance is calculated using latitude and longitude GPS data, using the Haversine formula [98]. In this analysis normalisation by distance provides the best results, this is a result of time data not accurately reflecting a unit's actual operational time. Units are stood down for extended periods for maintenance and are frequently powered during these periods. This gives very low energy consumption recordings for long periods of operation which do not reflect normal in-service use. As a result all analysis uses normalisation by distance i.e. kilowatt hours per kilometre which does not suffer from the same issue as negligible distance is recorded during the stood down period.

6.2.3 Vehicle configuration separation

6.2.3.1 Whole network approach

Data input should be homogenous, i.e. daily kilowatt hours per kilometre inputs should not significantly differ, as energy consumption estimates are made by extrapolating energy consumption from a subset of vehicles. A large variation in specific energy consumption means that individuals in a particular subset will have a greater calculation weighting leading to larger errors in the estimation.

Figure 110 - Cumulative specific active and regenerated energy for all units. Red braces highlight sub distributions that are a result of car configuration energy consumption differences.

Figure 110 shows cumulative specific energy consumption of each unit. Values reach a steady state toward the end of the year; the distribution of these final values can be used to assess the homogeneity of input data. Net active energy shows a substantially wide variation of input data, approximately 3 – 8 kWh km^{-1}. Regenerated energy shows a range of approximately 0.2 – 2.5 kWh km^{-1}. Most importantly both graphs show smaller groupings in the ranges, these were thought to be sub distributions within the larger distribution.

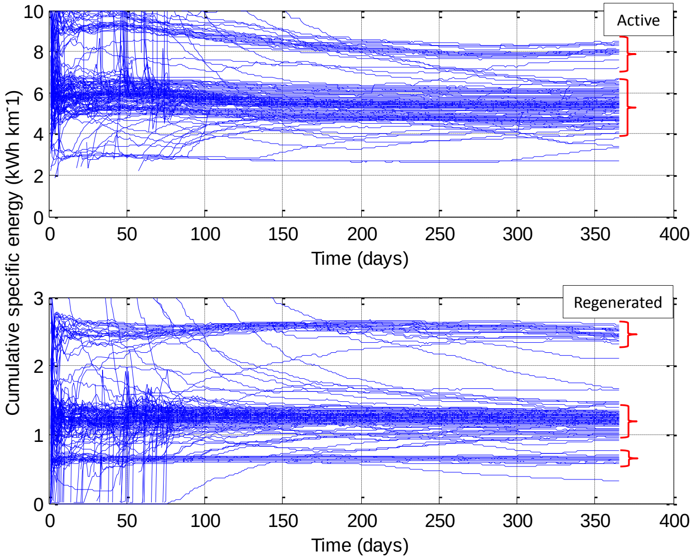


Figure 110 - Cumulative specific active and regenerated energy for all units. Red braces highlight sub distributions that are a result of car configuration energy consumption differences.

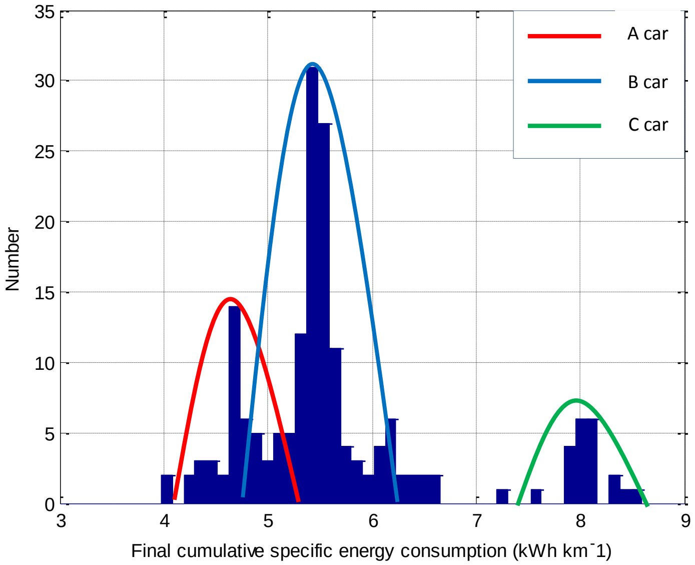


Figure 111 – Histogram of final cumulative specific energy consumptions showing car configuration sub distributions

Figure 111 shows these distributions with sub distributions highlighted. These were initially problematic as analysis would not be accurate using this input data. During a consultation with Neil Ovenden, engineering supply chain lead at ATOC/RDG, [99] it was suggested that this could have be the result of varying unit car configuration. For example, two units with identical distance travelled but different masses would have varied specific energy values. Also it was made known that the various configurations, A, B and C car, are in operation with different stopping patterns. A high frequency stopping pattern results in an average lower top speed over a particular route. As a result although distance travelled will be identical energy consumption per kilometre would be reduced. Final data input is then separated based on car configuration and each considered separately, this improves data homogeneity. Separating data into car configuration results in three separate fleets and as a result datasets of varying size. The largest of which is the B car configuration, with A and C car configurations being approximately equal. The A and C car fleets and datasets are 20% of the B car fleet and dataset size. For the whole network approach, data analysis can proceed with this format of data. For the route specific approach further steps are necessary to produce suitable input data.

6.2.3.2 Route specific approach

Individual routes travelled on Network Rail’s South of the River Thames DC network were identified and their longitude and latitude data stored. These were then interpolated to give equally spaced GPS points at 100 m intervals. The interpolated route data were then stored and are referred to as master routes; there are 46 master routes that are used in this analysis. The aim of the route specific approach is to separate all unit data into like routes and only compare energy data from these. As individual routes have varying gradient, curve and speed limit characteristics it is likely that this has a further effect on energy per kilowatt in the same way as unit configuration. By comparing data from like routes the homogeneity of the input data should increase meaning that the energy analysis is more accurate.

During a single day a unit will not remain on the same route, as a result daily operation should first be separated into individual journeys. This is done by the following procedure shown in Figure 112. Speed data for a single day and unit are parsed through. If a speed of 0 ms^{-1} is recorded a counter is increased by 1, the counter represents time the vehicle is stationary. If the vehicle begins to move the counter is reset to 0. If the counter reaches 10, indicating a stationary period of ten minutes, the index of the stop is recorded. At the end of the single day instances of data between stopping points are recorded as separate journeys. The process is repeated for every operational day for each unit.

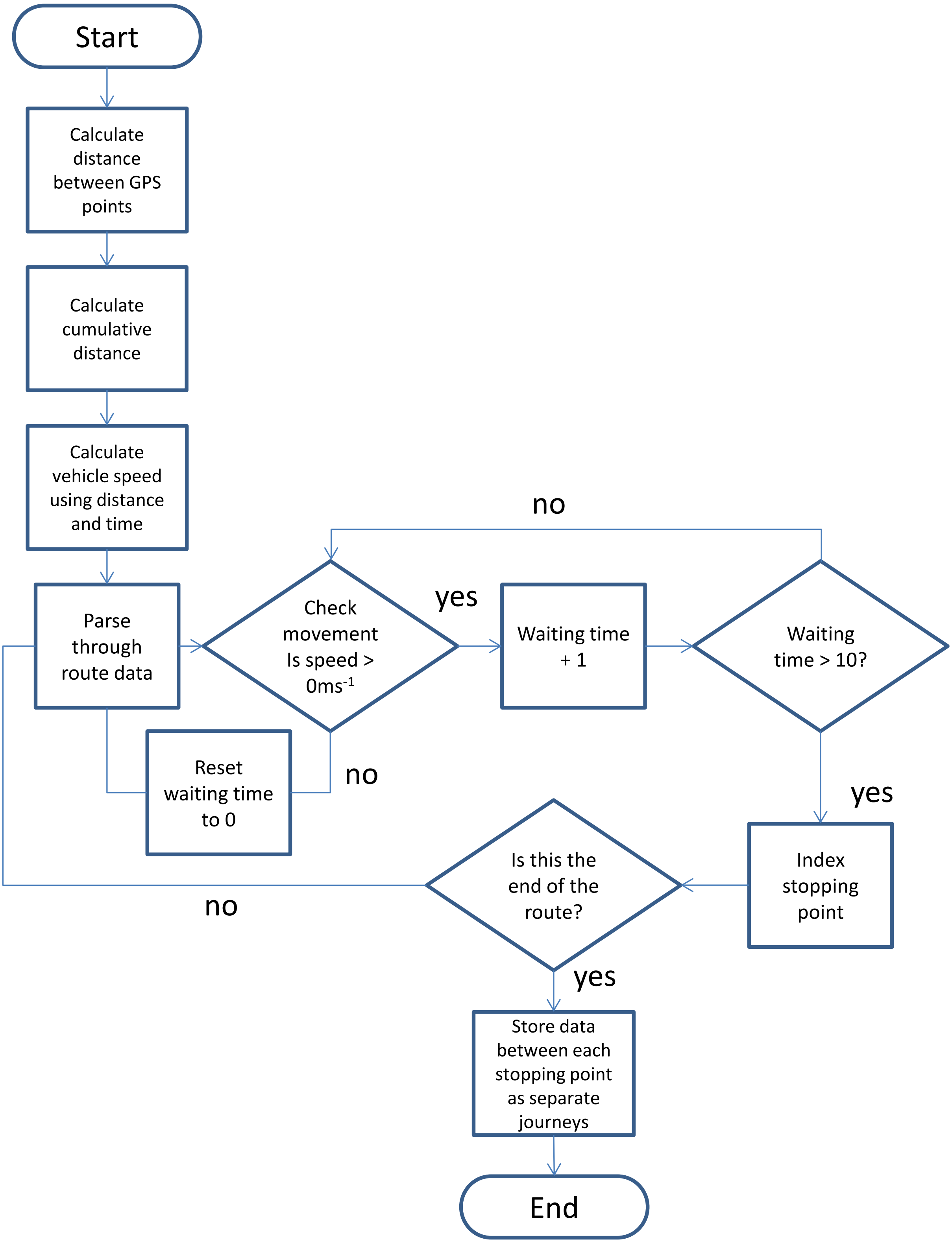
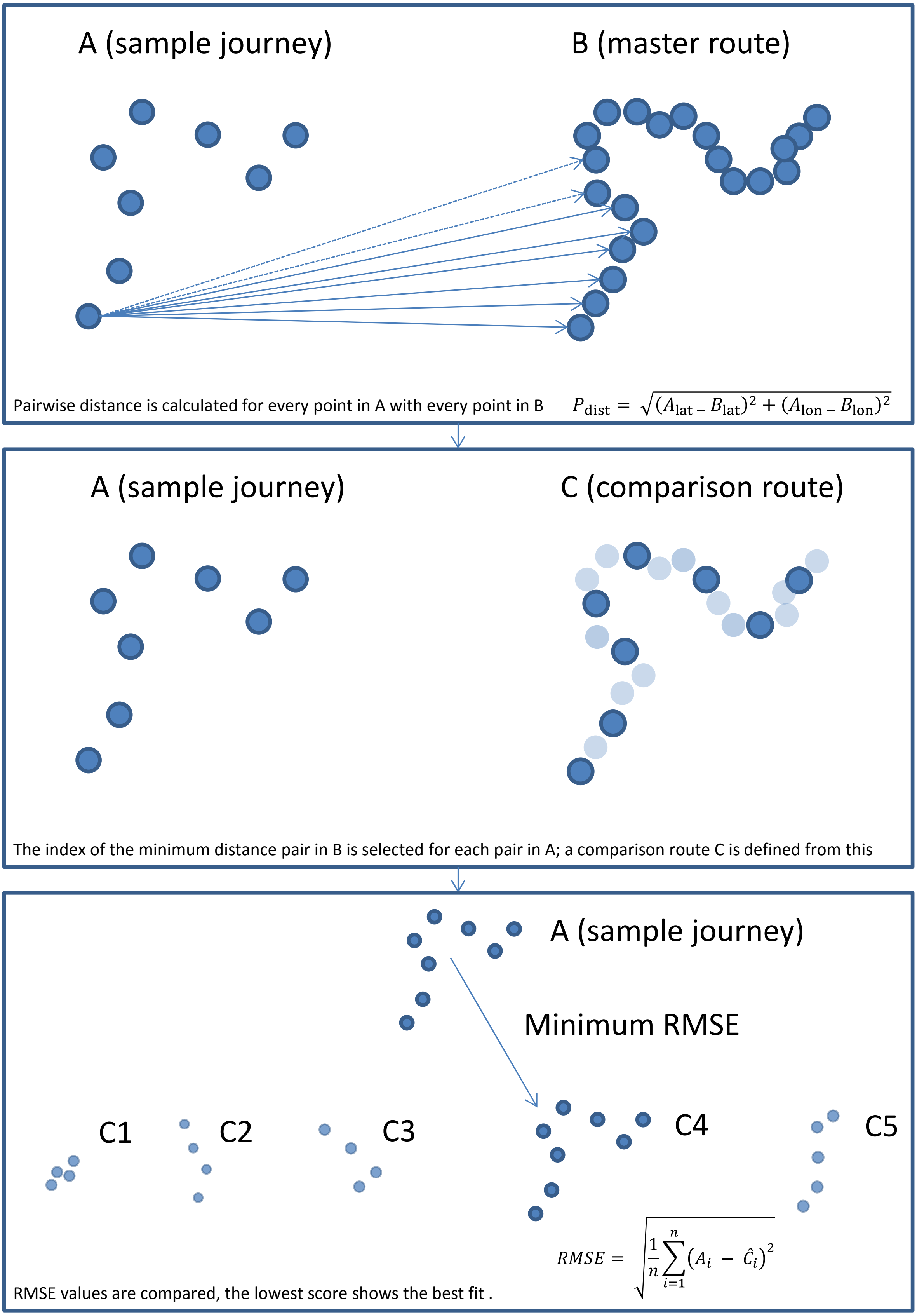

Start
Calculate distance between GPS points
Calculate cumulative distance
Calculate vehicle speed using distance and time
Parse through route data
Check movement Is speed > 0ms-1
yes
Waiting time + 1
Waiting time > 10?
no
no
Reset waiting time to 0
yes
Index stopping point
Is this the end of the route?
no
yes
Store data between each stopping point as separate journeys
End

Figure 112 – Flow chart demonstrating the route separation process

6.2.3.3 Route identification

Journeys should now be sorted by route, this involves matching the latitude and longitude of each journey stored in the previous process to one of the pre-defined master routes. For each GPS pair in a sample journey the closest pair of points in a master route is identified and indexed. A third 'comparison' route is defined from this, then the root mean squared error (RMSE) of the comparison route and sample journey is calculated. The RMSE gives a score of similarity between sample journey and master routes. When RMSE values for each route are calculated the minimum, i.e. the most similar, is chosen and the sample journey stored under that particular route, Figure 113 shows this process. Routes are then separated into car configuration as with the whole network approach. Analysis is conducted separately for each route dataset. Route data is processed in the same way as the whole network approach and data input used for analysis is in identical format.



Pairwise distance is calculated for every point in A with every point in B

$$P_{\text{dist}} = \sqrt{(A_{\text{lat}} - B_{\text{lat}})^2 + (A_{\text{lon}} - B_{\text{lon}})^2}$$

The index of the minimum distance pair in B is selected for each pair in A; a comparison route C is defined from this

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - \hat{C}_i)^2}$$

RMSE values are compared, the lowest score shows the best fit .

Figure 113 – Route matching process

6.2.4 Analysis

With all data processing steps complete the analysis is made, depending on the approach selected there will be a varying amount of input datasets, 3 for the whole network approach, one for each car configuration, for the route specific approach 40 datasets for the A car configuration, and 46 each for the B and C car configurations. In each case, however, the analysis follows the same steps:

- Energy estimations are made from a subset of size *n*
- The estimations are compared to the known energy consumption
- The errors are recorded
- The standard deviation of errors for the subset is calculated
- The subset size is reduced
- The process is repeated until there is only 1 vehicle in the subset

A subset is first sampled from the whole input dataset. This subset represents the part of the fleet that would be metered for PFM. The first subset sampled contains 100% of the units in the fleet and so is identical to full fleet metering. As specific energy consumption is used subset and total energy consumption are comparable. The energy estimation then is the total specific energy consumption of every unit in the subset. The known energy consumption is the total specific energy consumption of every unit in the dataset being analysed.

6.2.4.1 Error calculation

The error between the subset total energy and known energy consumption is calculated by using equation 26.

$$\varepsilon_{\mathrm{i}} = \frac{E_{\mathrm{k}} - E_{\mathrm{s}}}{E_{\mathrm{k}}} \tag{26}$$

Where ε_{i} is the error for an estimation of a subset, E_{k} is the known energy consumption and E_{s} is the subset energy consumption. The subset is randomly sampled 10,000 times and each error recorded. The standard deviation of this error is then calculated. With a subset that is the same size as the full dataset fleet there will be an error of 0 as there is no difference between estimated and known energy consumption values. As a result the SE will also be 0. The next step in the analysis is to decrease the subset size by a single unit and repeat the random sampling. The reduced subset size will introduce a small error between estimation and known energy values. As the subset size decreases there will be a greater variation in the range of estimations for a particular subset size, i.e. the σ_{ε} will increase. For each dataset the point of interest is the fleet size that is required to give a σ_{ε} of 1%. This point is the minimum number of units that need to be metered to give an energy estimation that is at the lower limit of accuracy. Figure 114 shows the analysis process steps and demonstrates the larger variation of error with smaller subsets.

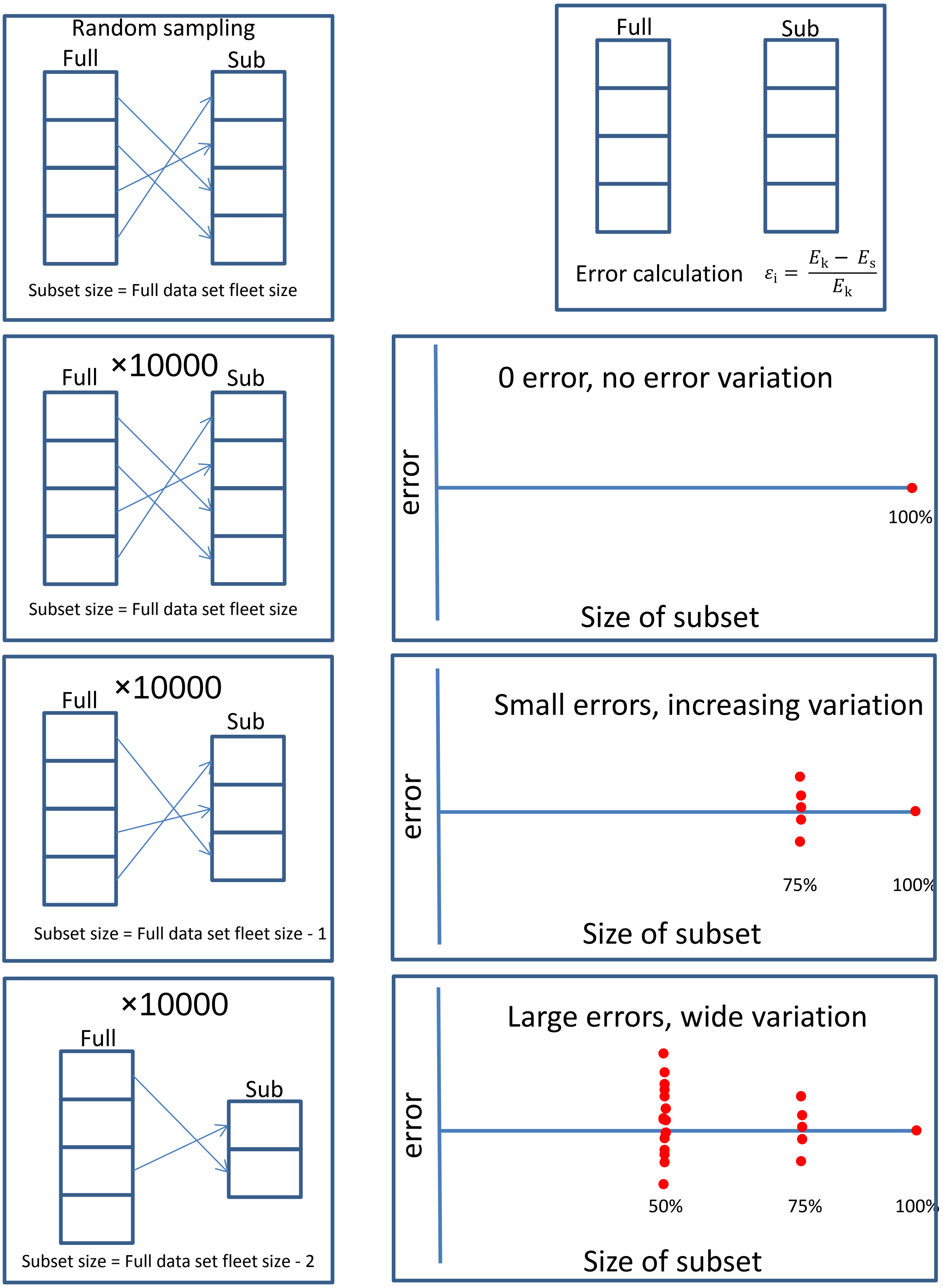

Random sampling
Full
Sub
Subset size = Full data set fleet size
Full
Sub
Error calculation $\varepsilon_{\mathrm{i}} = \frac{E_{\mathrm{k}} - E_{\mathrm{s}}}{E_{\mathrm{k}}}$
Full
×10000
Sub
Subset size = Full data set fleet size
0 error, no error variation
error
100%
Size of subset
Full
×10000
Sub
Subset size = Full data set fleet size - 1
Small errors, increasing variation
error
75%
100%
Size of subset
×10000
Full
Sub
Subset size = Full data set fleet size - 2
Large errors, wide variation
error
50%
75%
100%
Size of subset

Figure 114 – Analysis process showing increased error in energy consumption estimation with smaller fleet subsets.

6.2.4.2 Problematic data

Two case studies detailing the application of both whole network and route specific approaches were conducted, however, it became apparent that there were a number of issues regarding data quality. The need to process datasets automatically at high speed makes its very difficult for human observation of all journey data. This adds the possibility that some of the data is incorrect and would likely be included into the full datasets without a human analyst being able to spot these problems.

Outliers in datasets are often unavoidable. Steps to make data as homogenous as possible have been described above, however, even when implemented significant variation in specific energy consumption remained. This was bought to the attention of the author when looking at the range of input data. Occurrences of energy values with negative and less than 1 were noticed in final datasets. These anomalies were traced back to the raw data to eliminate the possibility that bad processing code had caused them. As an extra precaution raw data was also checked on the data holding companies' database for the same units, meters and dates. This confirmed that data errors originated at the electricity meter and was not a result of data processing. Raw data showed numerous points where a meter on a unit periodically stopped recording. In some cases the second meter would continue to work. As total cumulative specific energy for a unit is calculated from the sum of both energy meters this causes a significant discrepancy and causes points where energy per kilometre is extremely low. Negative values were produced where cases of both meters on a particular unit discontinued data recording but regenerated energy continued to be logged. The problem was reported to ATOC and rectified for continuing energy monitoring. This data was removed from the analysis.

6.3 Results

6.3.1 Whole network approach

The influence of the unit configuration is the focus of this approach. Units can be in A, B and C car configurations. Data is organised and processed as described in the methodology section above, total specific energy consumption for every unit is used to provide energy consumption estimates from fleet subsets which decrease in size until only one unit remains. The analysis shows that the characteristics of each configuration are similar but minimum fleet size required for a 1% σ_ε varies. Figure 115 (left) shows the σ_ε curve for the A car configuration where a partial fleet size consisting of 45% of the total A car fleet is required for a σ_ε of 1%. It was initially though that the B car configuration would give the best results as it has the largest fleet which would provide in a more homogenised distribution of energy input values. The A car configuration, however, provides the best results for the whole network approach. As the B and C car configurations are both largest and smallest fleet sizes this suggests that accuracy is less dependent on this and more centred on unit homogeneity; this reflects the duty cycles of each configuration. Analysis complexity is significantly reduced with the whole network approach as it does not need identification of single journeys and route matching. Where input data is for units operating on largely similar routes the whole network approach provides good results as shown with the A car configuration (45%). However, due to route specific nuances, i.e. varying gradient, speed limits and stopping patterns, input data, total operation over the whole network of each unit has limited homogeneity with the B and C car configurations, where 1% σ_ε occurs at 60 and 96% metered partial fleet, shown in Figure 115 (centre and right), making this approach unsuitable for use with this particular network.

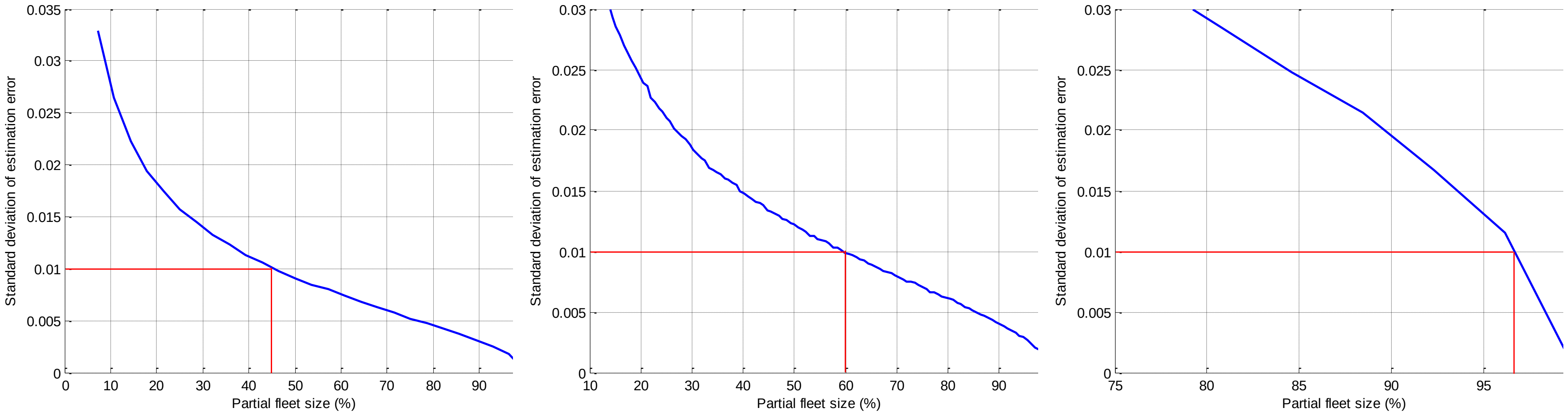


Figure 115 – Standard deviation of estimation error curve for A, B and C car configurations (left to right) using whole network approach

6.3.2 Route specific approach

Here estimation error is focused on the influence of the route taken by trains. Therefore, data coming from the multi-route service of units in operation across all routes were identified and analysed separately as described in 6.2. Journeys in both directions were not separated.

There are significant improvements to error standard deviation indicating greater homogeneity at the route level compared to the whole network level. σ_ε curves for each configuration operating on all routes are calculated and a mean σ_ε curve generated from these, SE curves for each route are shown in the following figures in light grey. Figure 116 (left) shows the A car fleet; the mean required fleet percentage is 35% for a σ_ε of 1%. The larger B car fleet has a wider distribution to the A car fleet but a lower mean required fleet percentage of 22%, shown in Figure 116 (centre). The wider distribution of σ_ε curves is a result of the wider range of routes taken by this configuration. The C car configuration has the largest distribution with required fleet percentages for its routes ranging from 10% – 97% and a mean of 78% shown in Figure 116 (right). Although this is an improvement over the whole network approach the 5 car configuration still results in a large partial fleet suggesting that the C car fleet data is not suitable for determining partial fleet size.

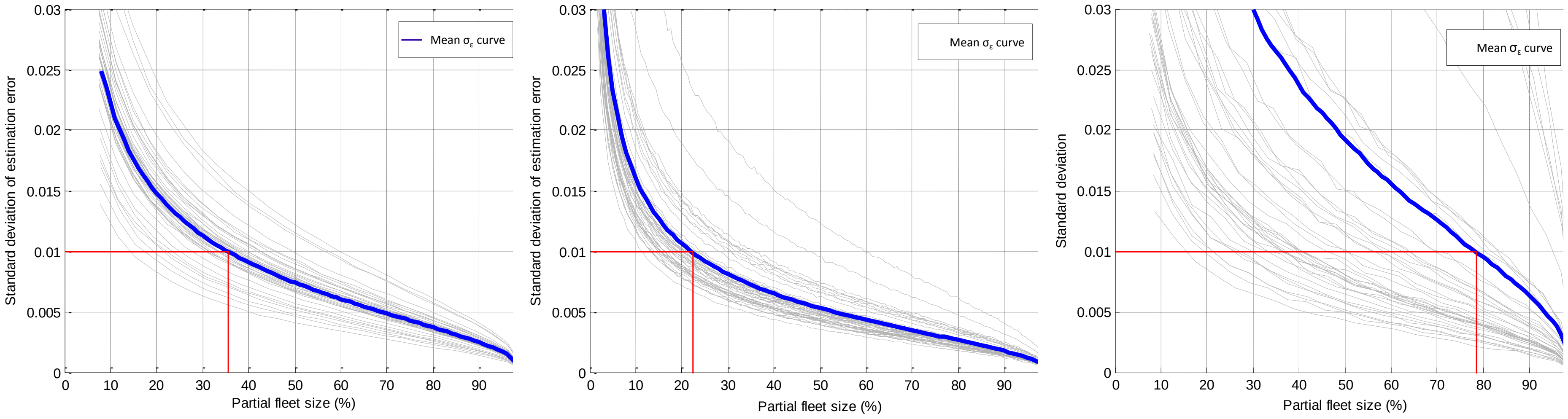


Figure 116 – Mean standard deviation of estimation error curve for A, B and C car configurations (left to right) using route specific approach

It should be noted that this does not necessarily suggest that the C car fleet cannot be partially metered. The poor results for this configuration only serve to show that the input data is not sufficiently homogenised, meaning that more normalisation data processing is required before this particular fleet can be used in the analysis. Partial fleet metering of the C car fleet can be chosen adequately by the A and B car results. This is supported by findings in the first in-service data project. Here σ_ε curves for differing and mixed rolling stock produced similar error characteristics with a 15% – 20% metering requirement. Input data homogeneity was increased due to the small route sample examined. The relatively similar error characteristic between varying rolling stock in the previous project and the DC fleet analysis suggest analysis suggest that C car partial metering is possible based on the results of the A and B car fleets.

The A and B car configurations in the route specific approach provide lower partial fleet sizes compared to the whole network approach. The mean results from the A and B car configurations were used to make fleet percentage metering requirement recommendations to ATOC and have since been used to advise future partial fleet metering policy [99].

6.3.3 Analysis Duration

Results in both analyses are taken from datasets which were recorded over a 12 month period. During discussions between the author and Neil Ovenden (engineering supply chain lead at ATOC/RDG) it was established that a key aspect of the research was to prove that PFM on a partial fleet size of 30% would be possible. Results show that this is the case providing that data for a significant period of time is used in the analysis. Billing periods tend to be much shorter, 29 – 30 days [96], which prompted an investigation into the minimum duration a dataset should be recorded over to produce a σ_ε of 1% with a fleet size of approximately 30% or less. The statistical analysis was repeated with increasing data duration sizes. Figure 117 shows this process.

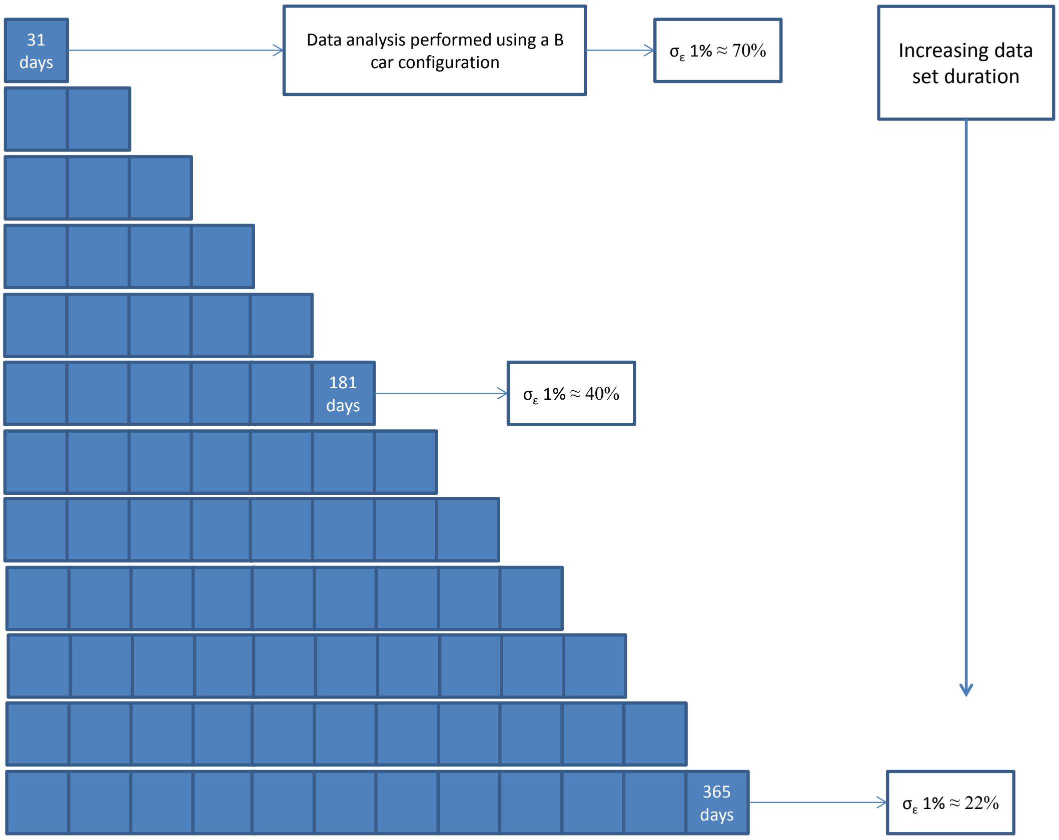

31 days
Data analysis performed using a B car configuration
σε 1% ≈ 70%
Increasing data set duration
181 days
σε 1% ≈ 40%
365 days
σε 1% ≈ 22%

Figure 117 – Cumulative dataset duration analysis procedure

This analysis uses the B car configuration route specific approach data only as it yields the best results overall. The dataset is further separated into cumulative month samples as shown in Figure 117. For example, the first analysis uses only the first 31 days of 2014. This is followed by the dataset from the 1st of January to the 28th of February, this continues until the analysis for a whole year is completed. In this case, route specific σ_ε curves are not considered; only mean curves are calculated and recorded, for each time period. Figure 118 shows that as dataset duration increases the metered fleet requirement decreases. After a single billing period metered fleet requirement is at approximately 83%, to be within the desired 30% range recorded data duration should be greater than 212 days or after 7 months. After this point, continued reduction of required partial fleet size decreases.

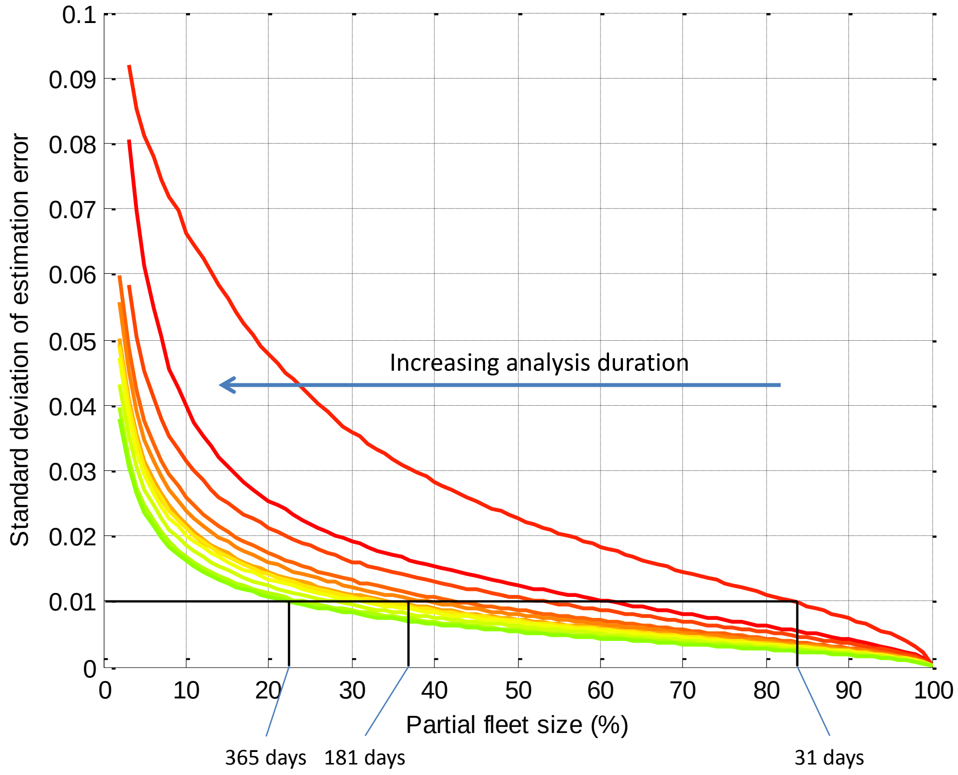


Figure 118 – Mean standard deviation of estimation error curves of increasing analysis duration

6.4 Discussion

Data extracted from this rail network energy database have been used to derive energy consumption estimates from partial datasets. This simulates fleets which may have only a proportion of their fleets fitted with energy meters. The distribution of energy data does in some cases follow a normal distribution which means that statistical analysis may be used to derive consumption estimates with known mathematical accuracy.

Normalised estimation inputs (specific consumption and consumption rate) produced lower estimation errors compared to the non-normalised input in previous studies, but this is only demonstrated in the specific energy consumption case. This is a result of differences in the pre-processing of data, whereby time is not always consistent with the actual operational time. In this case, operational distance is much more reliable and results in the lower estimation errors. In all cases, specific consumption has produced the best results and is recommended as the most suitable way to continue this analysis. This was observed in both the whole network approach and the route specific approach.

The wide distribution of energy consumption between different configurations means that collective data, i.e. all three-unit length configurations (for the class of vehicle analysed) in a single dataset cannot be used together. When separated into individual configurations results improve significantly, however this results in fleet sizes which are very different. It was thought that small configuration specific fleets would produce poor results as but this was shown to not be the case; in the whole network approach the A car configuration produces the best results, which are a significant improvement over the large B car configuration fleet. Poor results obtained from the C car configuration are therefore thought not to relate to the fleet size but rather the very varied duty cycle. A number of steps to normalise within car configurations were made prior to final data processing which had little effect on results. This involved an attempt to normalise against stopping pattern by dividing input data by average speed. Two different stopping patterns on the same route would produce a higher and lower average speed, however, when implemented no improvements were made. Identifying points where the vehicle is stopped is problematic, due to the relatively low resolution of the data. The primary focus of this dataset is to calculate energy billing and as a result there is no connection to service code and diagram data. This means that to match a particular journey to a known stopping pattern requires further dataset acquisition. This data is available but is only kept short term (approximately 4 weeks) meaning that accurate matching of known stopping pattern is not possible for this dataset. Therefore it is unlikely that successful normalisation of car configuration input data is possible.

Although separation by car configuration improves results the whole network methodology should only be used when vehicle position data is unavailable, this is demonstrated by the significant results improvement with the implementation of the route specific approach. Results from A and B car configurations suggest that the minimum fleet size that should be metered could be in the range of 22% – 35% of the full fleet based on a dataset covering twelve months. As the DC fleet concerned is already fully metered it does not make sense to remove meters and continue with a PFM. However there are a number of midlife fleets in operation in the UK that do not have energy metering. In this case PFM based on these results are a valid option. For a 200 unit fleet full metering costs would be approximately £1.5 – £2m [94, 99]. PFM at 30% fitted could provide savings of up to £1 – £1.4m on initial fitment costs. Secondly continued use of PFM on such a fleet would mean that billing could be based on an improved estimate system rather than the existing model. 30% PFM results in an σ_{ε} of approximately 1 standard deviation. Energy consumption data from the 377 fleet has been shown to closely approximate a normal distribution. A progressively larger dataset would likely show further tending towards this distribution. As a result approximately 68% of billing estimations (1 standard deviation) taken from the metered 30% would result in a ±1% error.

Based on the results there is a valid statistical basis to permit partial fleet metering to estimate traction energy bills. The fully metered fleet data can be used as a base case to compare partial fleet energy consumption to; this can be used to check whether there are significant changes in energy consumption after each billing period. Statistical analysis demonstrating this is discussed below. Partial fleet data can then be added to the base case to develop a larger base case sample. With the prospects of PFM continuing over a long period of time it is crucial that significant changes to the fleet operation are factored into the analysis and billing calculation. Significant changes encompass any alteration to the normal operation of the fleet, for example if 30% of the 4 car fleet had a rapid or non-stop service introduced, or speed limit changes were made to large sections of a route. Changes such as these would undoubtedly affect energy consumption per kilometre. This would mean vehicle homogeneity would be affected and energy data no longer representative of an entire fleet. Also any significant improvement of a vehicles' efficiency is likely to have similar substantial affects as operational changes; re-tractioning or introducing energy saving lighting systems would constitute efficiency improvement changes. In these cases steps to normalise data from more efficient vehicles should be taken, or a whole fleet analysis should be repeated to re-establish a base dataset that can be used for comparison to future partial fleet samples.

In late 2016 new billing rules including PFM will be suggested to TOCs by Network Rail and the ATOC. Based on TOC responses these rules will be finalised in February 2017 [100]. Draft rules discussed with the author and ATOC suggested that they would set the partial fleet requirement to cover at least 20% of the total mileage of the fleet i.e. that 20% of the units in a fleet should have energy meters installed [100], very close to the findings of this research. At 20% billing estimate standard deviation of error increases slightly, from the mean standard deviation of error curve for the B car fleet, Figure 116, a 20% metered fleet would result in a 1.3% estimation error. Based on the known approximated energy bill of Southeastern, £6.5 million, PFM with a 1% error (22% of the fleet) would result in a billing estimate with ± £65,000, a 1.3% error ± £84,500, depending on the size of the fleet a reduction of meter fitment of 2% is approximately equal to the potential loss resulting from a billing error increase of 1% - 1.3%. For example, assuming that energy costs between similar sized networks are similar, Arriva Trains Wales operate a fleet of approximately 100 units, at £7,500 - £10,000 per meter installation a partial fleet of 22% would cost £165,000 – £220,000, a 20% fleet £150,000 – £200,000. Initial savings of £15,000 – £20,000 would be achieved with PFM reduction of 2%, the potential over-billing estimate would be £19,000. This suggests that reduced PFM (to 20%) carries the risk of losing initial installation cost savings after only 1 year of billing assuming the worst-case scenario where billing estimation has a 1.3% error.

Similarly, a more cautious TOC that opted to install metering on 35% of a fleet could expect billing estimation error to reduce to 0.75%, using the same examples as above a ± £37,500 difference from the known energy cost can be expected. The additional costs of installing 13% more of the fleet would result in initial costs of £262,500 – £350,000 (based on the 100-unit fleet of Arriva Trains Wales). Compared to the potential over-billing of a 22% PFM (£65,000) there would be annual savings of £27,500 in the event of a potential over-bill with a 35% PFM. The difference in initial installation costs between a 22% PFM and 35% PFM are £97,500 – £130,000, meaning that initial costs could be recovered in 4 – 5 years. Once again this assumes the worst-case scenario where a TOC is annually over-billed at the maximum error for a particular percentage fleet. In the case where electricity is billed significantly closer to the actual energy use or lower than, recovery costs for a larger PFM percentage could take longer to recover.

The data analysis conducted in this project is intended to be a 'one-off' research project, findings can be used to suggest minimum fleet numbers without requiring the analysis to be performed again. In truth this is both unlikely to be the case and increases financial risk for energy suppliers and TOCs intending to used PFM. The following section describes hypothesis testing that can determine the similarity of successive data sets and demonstrates the importance of maintaining fleet energy consumption homogeneity. As well as continued data set homogeneity testing repeat partial fleet minimum numbers should be calculated annually using the historic (i.e. original) dataset that continually adds successive year's partial fleet datasets. This is dependent on the willingness of TOCs paying for data to be held and for analysis to be conducted. Based on the costs involved with the analysis conducted by the author and BCRRE, plus likely extra costs of formalising the analysis technique and providing a standardised report and data output, the financial saving provided by introducing billing from PFM is still worthwhile. To give an example total energy estimations for Merseyrail and Southeastern are show in Table 19. Merseyrail's estimation was obtained from [60] and Southeastern from [101] and [102].

Table 19 – A table showing additional charges and under billing for energy consumption based on potential error increases resulting from PFM

	Known Operator Traction Energy Bill	**Energy bill estimate +1% error**	**Energy bill estimate - 1% error**	**Energy bill estimate +3% error**	**Energy bill estimate - 3% error**
Merseyrail	£5,000,000	£5,050,000	£4,950,000	£5,150,000	£4,850,000
Southeastern	£6,500,000	£6,565,000	£6,435,000	£6,695,000	£6,305,000

PFM energy billing aims to give an estimation with a ±1% error of the mean energy consumption. If successive energy metering analysis is not carried out and billing estimates begin to vary beyond this 1%, additional costs to the TOC or energy provider are significant, almost £200,000 pounds in the case of a ±3% error with Southeastern. This is still a small cost when compared with the total operational costs of the railway, however, when compared to the relatively low cost of data analysis (potentially a few 10s of thousands of pounds) there is likely a good business case for continued data analysis that complements PFM energy billing; it is better to pay for annual data analysis rather than risk increased errors in PFM billing estimations that might arise without continued analysis.

6.5 Conclusion

Chapter 6 has shown how the collection and analysis of a Rich dataset has been used to develop a method to statistically validate the accuracy of partial fleet metering for electricity billing of rail vehicle fleets. Two approaches to analysis were designed, a whole network approach and a route specific approach. Results show that the whole network approach provides the best results where fleet size required for energy metering is 22 – 35% of the entire fleet. The route specific approach requires the availability of vehicle position data and increased data processing time. Where these are not available the whole network approach can be used, however, minimum fleet requirement is increased to 45 – 60% of the fleet. The difference between both approaches is related to the increased homogeneity of input energy data when considering individual routes compared to input energy from combined routes. Homogeneity is increased by separating units into car configuration specific inputs, in this case A, B and C car variants and performing separate analysis on each. The fleet percentages shown above are the minimum values from A and B car configurations which provided the best results. The C car configuration results were poor in the whole network approach, requiring approximately 90% of the fleet to be monitored to provide accurate billing estimations. In the route specific approach the C car configuration results vary from 10 – 97%, this suggests that energy input data from the C car fleet is unreliable and currently unsuitable for use in the analysis. A second factor that reduces the fleet size requirement is the duration of the analysis. Results taken from a B car fleet analysis with varying duration time shows that required metered fleet size reduces from 84% after 31 days to 22% after 365 days and reaches the necessary 30% fleet size at approximately 200 days.

The following chapter is a continuation of this analysis. It is likely that partial fleet metering will become a viable option for operators that do not currently have access to whole fleet energy metering. In the event that an operator opts for a partial fleet metering billing model, it is essential that there is very little difference in a successive year’s specific energy consumption (kWh/km). This would rely on little variation in the vehicles’ performance. This can be tested using null hypothesis and alternate hypothesis testing of data samples to identify significant differences in datasets from different years. This can be used to show that vehicle homogeneity has not changed, or can be used to show that any substantial change to rolling stock efficiency or service pattern has resulted in a significant difference between successive datasets.

7 Successive dataset validation

7.1 Introduction

This chapter describes a statistical analysis using a single route specific approach input dataset and a synthetic input dataset. The single route dataset represents historic data, this is used to compare against future datasets to determine whether there is significant difference between the two. This analysis can be used three areas:

- To determine whether successive datasets from the fully metered fleet are consistently similar.
- To determine a partial fleet dataset is comparable to the historic dataset.
- To investigate the effects of duty cycle changes and efficiency improvements on the analysis.

Successive fully metered datasets should be combined to continually build an accurate base case that can be used to determine whether partially metered fleet energy consumption is accurate. This is dependent on a successive fully metered data being similar to historic data. As only a single year of data is available this is not testable with real data. As a result a synthetic successive year's dataset is generated by sampling from the existing set and introducing random variation. Electricity meters installed on the DC fleet have a precision error of 1 – 2% which is added to the synthetic dataset. Three tests are performed using both datasets.

- A comparison of two years of fully metered data
- A comparison of a fully metered dataset to a 30% partial fleet dataset

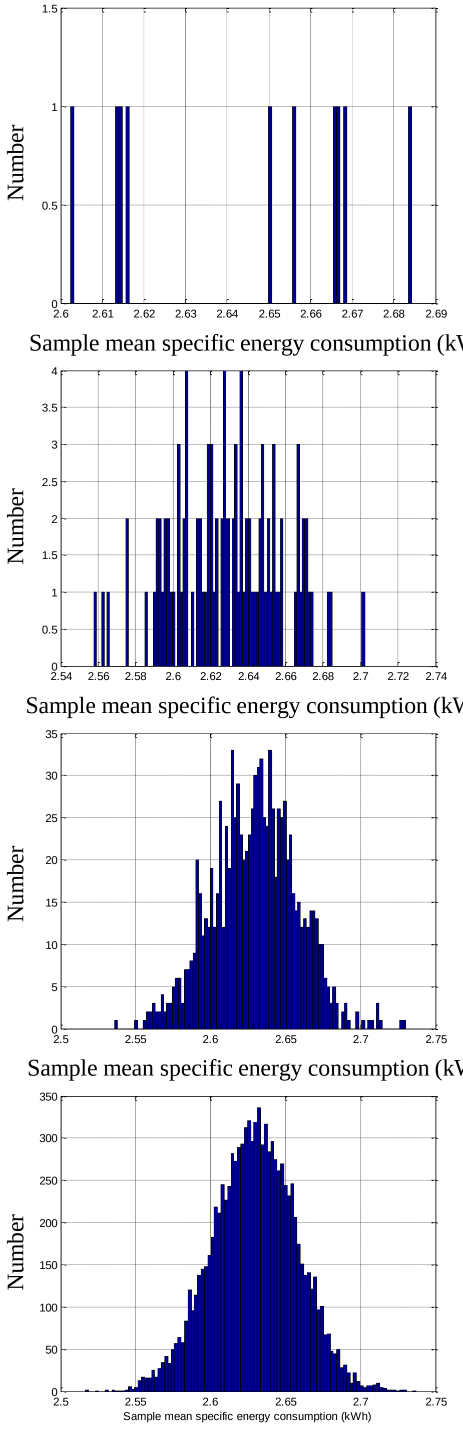
- A comparison of a fully metered dataset to a 30% partial fleet dataset where a percentage of the whole fleet has undergone re-tractioning providing a 15% efficiency improvement.

Each follows hypothesis testing and aims to identify the probability that an extreme result is produced when assuming the null hypothesis. This can then be used to determine whether a successive dataset is significantly different to historical data. The slight non-normal distribution of both datasets is problematic, accurate probability calculation relies on a normal distribution. As a result hypothesis testing uses the sampling distribution of both the single route dataset and synthetic successive year dataset.

7.2 Hypothesis testing

Due to the development of a key area of statistics and probability theory known as Central limit theorem (CLT) statistical testing using null and alternative hypothesis is possible, and is especially useful where total population data is not available and information relating to the mean and standard deviation is not known. It is considered to be one of the most important ideas in statistics due to its simplicity; its application allows the analysis of complex problems by approximation [103].

CLT was developed by de Moivre and improved by Laplace, but wasn't known in its classical form until the 1930s where it became widely used in many statistical applications [103, 104]. CLT shows that when random samples are taken from a non-normal dataset and their means calculated, the distribution of the means will tend to be normal [104], this is demonstrated in Figure 119. This is known as '*the sampling distribution of the sample means*' (SDM); its mean is denoted by $\mu_{\bar{x}}$. This should not be confused with sample and population mean which are different values. With a small dataset sample means will tend to have increased positive or negative skew and increased kurtosis [104, 105].



30 random samples are taken from the Historic data set, the mean of the 30 samples is calculated and their frequency plotted. After 10 iterations the normal distribution is not clear.

After 100 iterations there is a general normal distribution starting to take shape.

After 1000 iterations the normal distribution is much more clear but there is still a degree of negative skew.

After 10,000 iterations a normal distribution curve can be easily approximated. A perfect curve would only be apparent with infinite iterations.

Figure 119 – Sampling distributions showing the tendency toward a normal distribution over 10,000 iterations.

When the number of samples taken is greater than 30 the distribution of the sample means becomes increasingly normal [105]. Secondly the mean of the SDM $\mu_{\bar{x}}$, is the same as the mean of the original dataset (the original population mean μ) [105]. This allows the use of $\mu_{\bar{x}}$ to represent the population mean. As sample size increases, the standard deviation of the sampling distribution, denoted by $\sigma_{\bar{x}}$, reduces, this is shown in Figure 120, distribution curves become much tighter as sample size (n) increases.

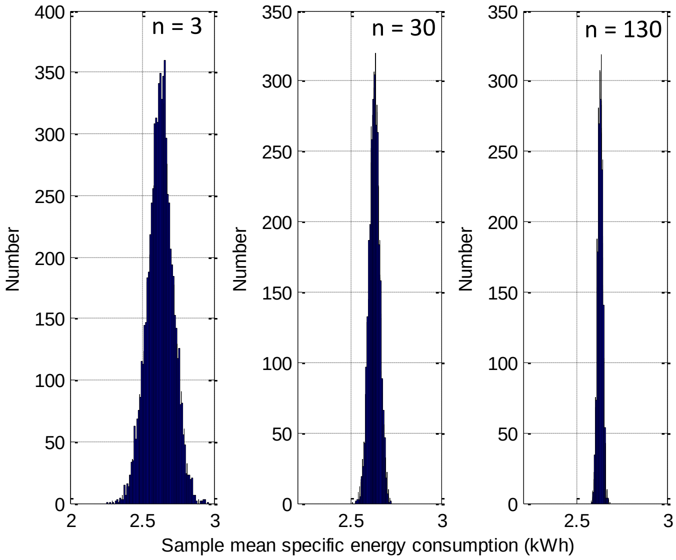


Figure 120 – Sampling distributions of 10,000 iterations with increasing sample sizes (n); standard deviation decreases as n increases

It is possible to calculate the variance of a SDM, when using a single sample, by dividing the original dataset (population) variance σ^2 by the sample size [105] shown in equation 27:

$$\sigma_{\bar{x}}{}^2 = \frac{\sigma^2}{n} \tag{27}$$

By square rooting n and σ the standard deviation of the SDM can be obtained. Where the population variance or standard deviation is not known the standard deviation of a sample with sample size greater than 30 [105] can be used to approximate the SDM's standard deviation. This is also known as the standard error of the mean (SE), shown in equation 28.

$$SE_{\bar{\mathrm{x}}} = \frac{s}{\sqrt{n}} \tag{28}$$

To identify significant difference between data sets a third SDM is required. This is the SDM of the difference between SDMs of each data set. The SDM of difference's mean is calculated simply by subtracting both datasets SDM mean, $u_{\bar{x}_1} - u_{\bar{x}_2}$, this can also be stated as $u_{\bar{x}_1 - \bar{x}_2}$; the absolute value is used. The difference in variance is the sum of a SDM's variance. Equation 29 is used to calculate the SDM's standard deviation of difference, denoted by $\sigma_{\bar{x}_1 - \bar{x}_2}$.

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}} \tag{29}$$

As this test investigates the probability that $\mu_{\bar{x}}$ of the follow on dataset is greater than or less than $\mu_{\bar{x}}$ of the historic dataset a two tailed test is required. A *z* score is the number of standard deviations from the mean that is associated with a particular probability and is calculated by:

$$z = \frac{x - \mu}{\sigma} \tag{30}$$

Each test is designed to find a 95% confidence interval, where α = 0.05.This means that a *z* score that coincides with a probability of 97.5% in a two tailed test should be identified rather than 95% as probabilities both sides of the mean should be summed; this is shown in Figure 121. Using a *z* table α = 0.05 can be found with a z score of 1.96.

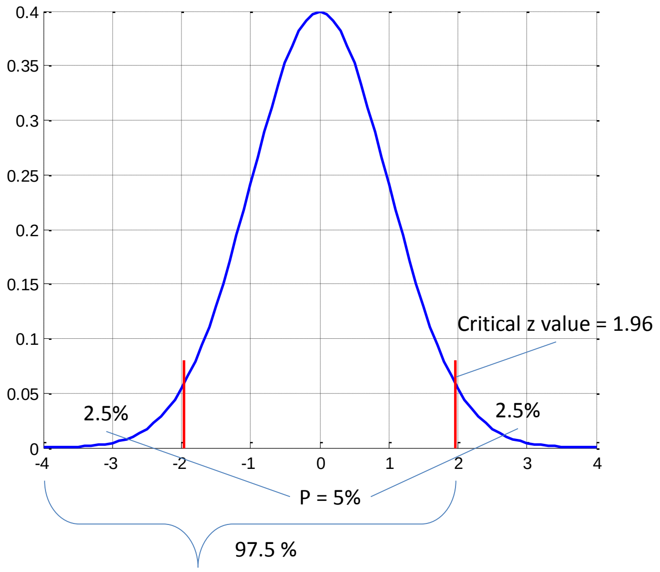


Figure 121 – Normal two tailed distribution curve showing critical z value where probability$p = 5\%$*, this value is at 97.5% of a one tailed curve as p both sides of the mean are summed*

Assuming the null hypothesis is true, a result more extreme than the z = 1.96 threshold would have a less than 5% chance of occurring. If this is the case the null hypothesis can be rejected, suggesting that there is significant difference between historic and follow on datasets.

In each of the following only one test is performed, this means that the probability of a rare event being repeated is not possible as would be with multiple hypothesis tests of the same type. This can lead to the rejection of the null hypothesis as a 'false positive' is more likely to occur, this is known as a type I error.

7.3 Test 1

A single route specific energy input dataset is chosen randomly, a second comparison route is randomly sampled from this with normally distributed 2% errors added; this is known as the follow on dataset. Both datasets have the same length i.e. energy consumption from 130 units are included in each array. Figure 122 and Figure 123 show the distribution of energy consumption for each dataset. Null and alternate hypotheses are determined, described below.

7.3.1 Null and alternative hypotheses

H_0 = Mean specific energy consumption shows little difference in successive years

$$\mu_1 - \mu_2 = 0 \tag{31}$$

The difference between the mean of both populations is equal to 0. This is equivalent to:

$$\mu_{\bar{x}_1} - \mu_{\bar{x}_2} = 0 \tag{32}$$

The difference between the SDMs' mean is equal to 0, as they are the same as the population means. This can ultimately be written as:

$$\mu_{\bar{x}_1 - \bar{x}_2} = 0 \tag{33}$$

H_1 = The mean energy consumption of the follow on dataset is significantly greater or less than the mean energy consumption of the historic dataset in successive years.

$$\mu_{\bar{x}_1 - \bar{x}_2} \neq 0 \tag{34}$$

The difference between the SDMs' means distribution is not equal to 0.

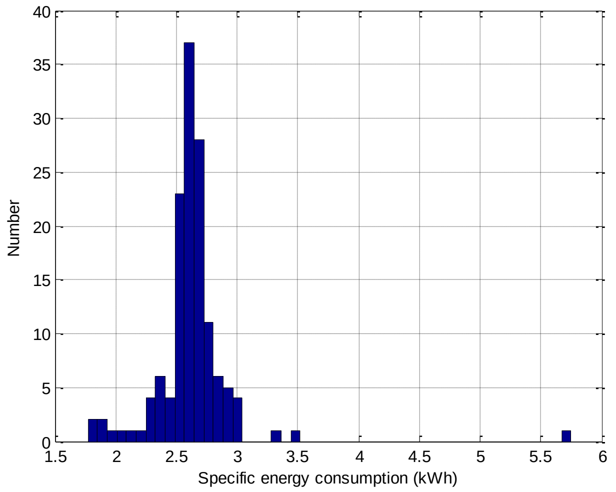


Figure 122 – Histogram showing distribution of energy consumption from the historic dataset

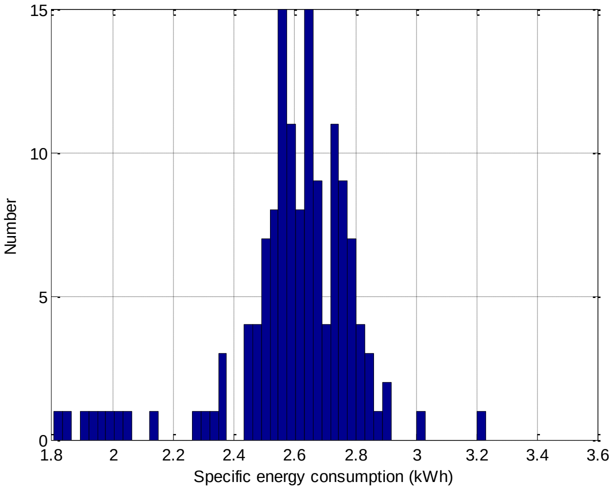


Figure 123 – Histogram showing distribution of energy consumption from the follow on dataset

7.3.2 Test 1 inputs and calculation

A large sample is taken from each dataset where:

Historic dataset SDM = population mean, $\mu_{\bar{x}_1} = \mu_1 = 2.68$

Historic dataset SDM standard error, $s_1 = 0.15$

Historic dataset sample size $n_1 = 30$

Follow on dataset SDM = population mean, $\mu_{\bar{\mathrm{x}}_2} = \mu_2 = 2.61$
Follow on dataset SDM standard error, $s_2 = 0.19$
Follow on dataset sample size $n_2 = 30$

Difference between SDM standard deviation:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}} = \sqrt{\frac{(0.15)^2}{30} + \frac{(0.19)^2}{30}} \tag{35}$$

$$= 0.043$$

Difference between SDM means if *z* score is 1.96:

$$= \sigma_{\bar{x}_1 - \bar{x}_2} \times 1.96 \tag{36}$$

$$= 0.084$$

Actual difference between SDM and population means:

$$\mu_{\bar{\mathrm{x}}_1} - \mu_{\bar{\mathrm{x}}_2} = 2.68 - 2.61 \tag{37}$$

$$= 0.070$$

7.3.3 Test 1 result

Difference between SDM means at z score of 1.96 (0.084) is greater than the difference between the actual SDM means (0.070) therefore the probability of producing this result is greater than 5%, this means that the null hypothesis can be accepted; the follow on dataset is not significantly different to the historical dataset; the probability of producing a difference between datasets such as this is:

$$z = \frac{\mu_{\bar{\mathrm{x}}_1} - \mu_{\bar{\mathrm{x}}_2}}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{0.070}{0.043} \tag{38}$$

$$= 1.63$$

From a standard normal z table $p = 1 - 0.95$ when z = 1.63

For a two tailed test $p = (1 - 0.95) \times 2$

$$p = 0.10$$

Assuming the null hypothesis is true there would be a 10% chance of producing a difference in sampling distribution sample means as extreme as this.

7.4 Test 2

A single route specific energy input dataset is chosen randomly, a second comparison route is randomly sampled from this with normally distributed 2% errors added then 30% of this dataset is sampled randomly and stored, this sample is known as the PFM dataset. The historic dataset contains 128 inputs, the PFM dataset 38. Figure 124 and Figure 125 show the distribution of energy consumption for each dataset. Null and alternate hypotheses are determined, described below

7.4.1 Null and alternative hypotheses

H_0 = Mean specific energy consumption shows little difference in successive years

$$\mu_{\bar{x}_1 - \bar{x}_2} = 0 \tag{39}$$

The difference between the SDMs' means distribution is equal to 0.

H_1 = The mean energy consumption of the PFM dataset is significantly greater or less than the mean energy consumption of the historic dataset in successive years.

$$\mu_{\bar{x}_1 - \bar{x}_2} \neq 0 \tag{40}$$

The difference between the SDMs' means distribution is not equal to 0.

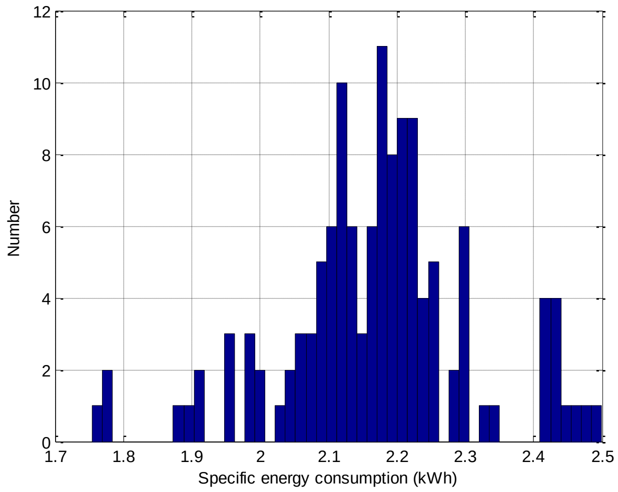


Figure 124 – Histogram showing distribution of energy consumption from the historic dataset

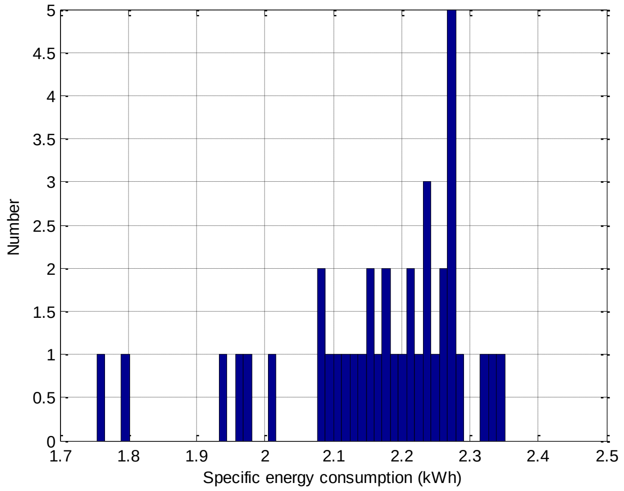


Figure 125 – Histogram showing distribution of energy consumption from the PFM dataset

7.4.2 Test 2 inputs and calculation

A large sample is taken from each dataset:

Historic dataset SDM = population mean, $\mu_{\bar{x}_1} = \mu_1 = 2.19$

Historic dataset SDM standard error, $s_1 = 0.124$

Historic dataset sample size $n_1 = 30$

PFM dataset SDM = population mean, $\mu_{\bar{x}_2} = \mu_2 = 2.17$

PFM dataset SDM standard error, $s_2 = 0.12$

PFM dataset sample size $n_2 = 30$

Difference between SDM standard deviation:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}} = \sqrt{\frac{(0.124)^2}{30} + \frac{(0.12)^2}{30}} \tag{41}$$

= 0.035

Difference between SDM means if *z* score is 1.96:

$$= \sigma_{\bar{x}_1 - \bar{x}_2} \times 1.96 \tag{42}$$

= 0.070

Actual difference between SDM and population means:

$$\mu_{\bar{\mathrm{x}}_1} - \mu_{\bar{\mathrm{x}}_2} = 2.68 - 2.61 \tag{43}$$

= 0.015

7.4.3 Test 2 result

Difference between SDM means with a z score of 1.96 (0.070) is greater than the difference between actual SDM means (0.015) therefore the probability of producing this result is greater than 5%, this means that the null hypothesis can be accepted; the PFM dataset is not significantly different to the historical dataset; the probability of producing a difference between datasets such as this is:

$$z = \frac{\mu_{\bar{x}_1} - \mu_{\bar{x}_2}}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{0.015}{0.035}$$

= 0.214

From a standard normal z table $p = 1 - 0.66$ when z = 1.63 (44)

For a two tailed test $p = (1 - 0.66) \times 2$

$$p = 0.67$$

Assuming the null hypothesis is true there would be a 67% chance of producing a difference in sampling means as extreme as this.

7.5 Test 3

A single route specific energy input dataset is chosen randomly, a second comparison route is randomly sampled from this with normally distributed 2% errors added. This test will assess the impact of a portion of the fleet being re-tractioned which would provide efficiency improvements. 50% of the fleet is randomly chosen; specific energy consumption for this selection is multiplied by 0.85 to introduce a 15% efficiency improvement. Then 30% of this dataset is sampled randomly and stored, this sample is known as the re-tractioned PFM dataset. The historic dataset contains 128 inputs, the re-tractioned PFM dataset 38. Figure 124 and Figure 125 show the distribution of energy consumption for each dataset. Null and alternate hypotheses are determined, described below

7.5.1 Null and alternative hypotheses

H_0 = Mean specific energy consumption shows little difference in successive years

$$\mu_{\bar{x}_1 - \bar{x}_2} = 0 \quad (45)$$

The difference between the SDMs' means distribution is equal to 0.

H_1 = The mean energy consumption of the re-tractioned PFM dataset is significantly less than the mean energy consumption of the historic dataset in successive years.

$$\mu_{\bar{x}_1 - \bar{x}_2} < 0 \tag{46}$$

The difference between the SDM's means distribution is less than 0. This is the only test which differs and does not require a two tailed result. This is due to the prior knowledge that re-tractioning will only reduce energy consumption, as a result the difference in sampling distributions and therefore population means can only be negative, and as such less than 0.

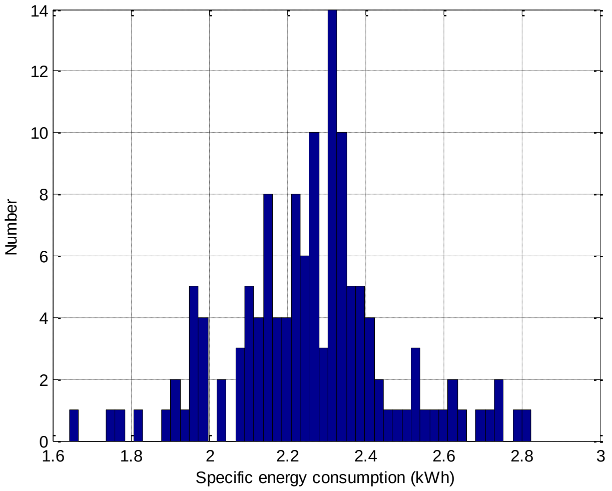


Figure 126 – Histogram showing distribution of energy consumption from the historic dataset

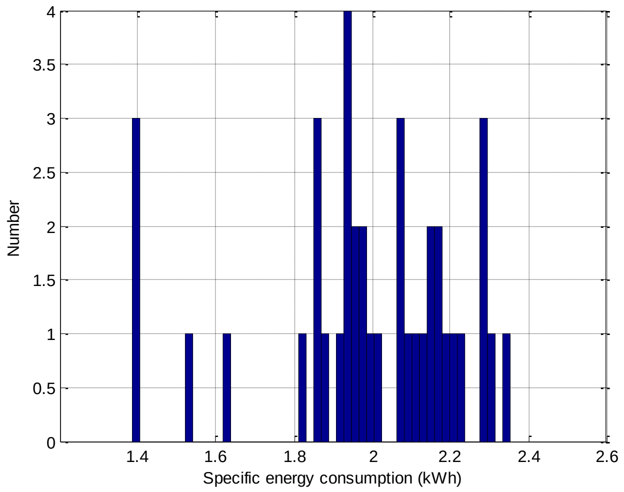


Figure 127 – Histogram showing distribution of energy consumption from the altered PFM dataset

7.5.2 Test 3 input and calculation

A large sample is taken from each dataset:

Historic dataset SDM = population mean, $\mu_{\bar{x}_1} = \mu_1 = 2.25$

Historic dataset SDM standard error, $s_1 = 0.21$

Historic dataset sample size $n_1 = 30$

Re-tractioned PFM dataset sample mean $\mu_{\bar{x}_2} = \mu_2 = 1.96$

Re-tractioned PFM dataset SDM standard error, $s_2 = 0.26$

Re-tractioned PFM dataset sample size $n_2 = 30$

Difference between SDM standard deviation:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}} = \sqrt{\frac{(0.21)^2}{30} + \frac{(0.25)^2}{30}} \quad (47)$$

$$= 0.059$$

Difference between SDM means if *z* score is 1.96:

$$= \sigma_{\bar{x}_1 - \bar{x}_2} \times 1.96$$

$$= 0.115 \tag{48}$$

Actual difference between SDM and population means:

$$\mu_{\bar{\mathrm{x}}_1} - \mu_{\bar{\mathrm{x}}_2} = 2.25 - 1.96$$

$$= 0.290 \tag{49}$$

7.5.3 Test 3 result

The difference between SDM means at *z* score of 1.96 (0.115) is less than the difference between actual SDM means (0.290) therefore the probability of producing this result is less than 5%, this means that the null hypothesis should be rejected. The re-tractioned PFM dataset is significantly different to the historical dataset; the probability of producing a difference between datasets such as this is:

$$z = \frac{\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{0.290}{0.059}$$

$$= 4.91 \tag{50}$$

$$p < 0.0001\%$$

As the z score is greater than 3 standard deviations the probability of producing this difference between dataset means, assuming the null hypothesis is true, is less than 0.0001%. This is extremely significant and demonstrated the large affect such a substantial change to rolling stock would produce. In this case if this data were added to historic data it would no longer be applicable to the billing estimation model. The analysis would have to be repeated on a fully metered fleet again to establish population mean and standard deviation and steps to normalise between variations in rolling stock.

7.6 Conclusion

A series of statistical tests were performed on real data and synthetic data generated to provide data for successive years of energy metering. The tests show that it is possible to identify significant variation ($\alpha = 0.05$) between historic and successive datasets. This allows for a successive year’s dataset to be added to the historic dataset. This was also shown to be possible with a full fleet historic dataset and a 30% PFM fleet. In the event of substantial changes to either the operational characteristics (i.e. changes in stopping pattern, speed limit changes or driver behaviour) or traction and energy characteristics (i.e. re-tractioning or the introduction of LED lighting) where energy efficiency is improved it is likely that there will be an effect on the historic dataset if data such as these were included. The third test found that where half a fleet had efficiency improvements of 15% there was significant difference between datasets. Where energy changing parameters are introduced the historic data should ‘reset’ and a new analysis made.

8 Conclusion

Data acquisition and analysis are of extreme importance to the rail industry. This research has set out to determine the added value of large datasets, demonstrating an aspect of the potential of Big or Rich data analysis for the rail industry. This has been done via two data analysis case studies. Data were taken from operational rail vehicles over a period of three years for one and a period of one year for the second. The first case study involved a single instrumented rail vehicle in operation on the Merseyrail network with high resolution data. The second case study involved electricity metering data from all DC EMU vehicles in operation on a UK rail network. This chapter will evaluate the effectiveness of each case study's response to the relevant hypothesis established in section 1.3.

8.1 Merseyrail energy monitoring project

Is it possible to identify driver characteristics, behaviour or possibly individual drivers from a Rich energy dataset?

The data analysis documented in chapters 3 – 4.6 show that driver characteristics and types of driver behaviour can be identified from energy data. Feature vectors describing the control input of a driver, running time and energy consumption were identified for approximately 400 journeys on 4 different routes on the Merseyrail network. These features were shown in section 3.3 and a journal paper by the author [5] as a method to classify driver style. Clustering analysis was used to group feature vectors into distinct groups; for three routes 6 such groups were identified, one route showed better clustering results with 7 groups. When separated into distinct groups the centre of each could be determined. The control input, time and energy of this ‘centroid’ were used to determine the characteristics of the driver style. The location of a centroid projected in terms of two or three principal components gives insight into the behaviour of a particular driver style. Chapter 4.6 section 5.3 gives a detailed description of the behaviours encountered for each route, Chapter 4.6 section 5.5 describes the clustering of each route. This case study shows that it is possible to use energy data, providing that control data is also available, to identify and categorise driver styles and examine the behaviour of those particular styles.

Individual drivers were not identified from the dataset. A methodology was developed to identify individual drivers from the Merseyrail dataset; however, lack of information matching specific drivers to journeys prevented conclusions to be made (see Appendix 10.1.1). Without this information algorithm training data could not be established meaning that driver identification could not be validated. It was decided that the identification of individual drivers should be set outside the scope of this research and be the subject of potential follow projects. This would allow for the collection of relevant driver identification data and reliable conclusions.

Could such characteristics be used to provide potential energy savings?

Clustering of driver styles was used to develop the concept of a ‘driver culture’. This is a distribution of driver styles in operation over a period of time; driver cultures were presented in chapter 5. Driver cultures show energy consumption per day resulting from the contribution, i.e. number of journeys, in a particular driver style. Three driver styles from each culture were selected:

- Peak operation

- Off peak operation
- Energy saving operation

The peak and off-peak operation driver styles were so called optimal driver styles, where energy consumption and running time are both reduced. As running time and energy consumption are largely inversely proportional, the optimal driver styles are a compromise between both. The shortest journey time of the two is selected for peak, the longer for off peak. Based on peak and off peak timetables a new driver culture is designed where peak and off peak driver styles comprise approximately 40% and 60% of journeys in a single day. This is used to demonstrate energy savings that would be possible if drivers adopted these driver styles only. Savings range from 1.5% – 6.4% for each route and equate to an estimated 4% saving combined. A second driver culture is designed that uses energy saving operation, i.e. the lowest energy driver style, in a particular driver culture for 100% of journeys. Savings range from 8% – 14% for each route and equate to an estimated 10% saving combined. The energy saving operation runs the risk of incurring a journey time penalty. In a number of cases journey time is very close to the allowed running time or slightly over. This means there would be little or no recovery time and the possibility of problems in the event of a delay. It was suggested that services could be removed to allow an energy saving driver culture to be used to minimise the risk of not meeting the timetable.

The 4% – 10% energy saving only accounts for half of the Merseyrail network, as detailed data is not available for the entire network it is not known how energy savings might change if Driver Culture analysis could be applied to its entirety. Energy savings based on current findings could provide a financial benefit to Merseyrail of approximately £200,000 – £500,000 per year based on an annual electricity bill of £5m [60], this would be an increase of 3.4% based on the stated £14.6 million Merseyrail profit. If the analysis were only applied to the Northern line only this would provide a 4% – 10% energy saving on a £2.5 million energy bill giving savings of £100,000 – £250,000 and profit increase of only 1.7%. The increase in profit from 2013 and 2014 was approximately 6%, if energy reduction using Driver Culture analysis as well as other operational cost reduction steps were made, further profit increases could be realised.

As well as energy reduction, carbon reduction is also one of the recent goals of Merseyrail. Carbon reductions of 400t per year have been estimated to be possible using Driver Culture analysis. To put this saving into perspective, a round trip from the UK to the far-east with a passenger capacity of 200 people would produce approximately 60t of CO_2. This means that the estimated carbon savings from Merseyrail would be the equivalent to only 6 or 7 long distance return flights, suggesting that this contribution will not be particularly effective in combatting rising carbon emissions and their climatic effects.

At this stage, it is not known whether the methodology developed in the MEMP is suitable for generic use, i.e. can it be applied to any rail network? The Merseyrail network has many commonalities between mainline rail networks but its standalone or isolated nature means that many of the findings shown here are possibly not repeatable outside of this network. For example, Merseyrail can operate with consistent headways and adhere to set timetables due to there only being Merseyrail Class 507/8s in operation. This potentially allows for an environment where driver behaviour (Cultures) are more apparent, i.e. drivers are able to readily repeat similar patterns on a daily basis, that on a more dynamic and larger mainline railway possibly do not happen. Secondly Merseyrail rolling stock is unique on GB railways. Although dated, the control system of the 507/8 EMU provides discreet output data allowing for the identification of feature vectors to use in driver style clustering. With modern rolling stock where driver control provides a continuous data output it is likely that there would be more complexity in identifying feature vectors for clustering. There are several other problems that would require consideration before implementing Driver Culture analysis to another rail network. An important conclusion to make is that from this research that potential energy savings of 4 – 10% are only possible for this network and that savings on other networks would require a repeat of the research stage rather than immediate implementation. The author is currently applying for funding to investigate whether Driver Cultures exist on larger rail networks, if this funding application is successful a follow up two-year project is planned to begin in the autumn of 2017.

8.2 ATOC partial fleet metering

Is it possible to use historic energy data from a rail network to validate the use of a partially metered fleet, where only a portion of vehicles have electricity metering, to provide accurate electricity bill estimations?

The case study documented in chapter 6 shows that it is possible to use an historic energy dataset to determine the size of a partially metered fleet. This required the consideration of individual unit configurations and normalisation of data by distance. From this two approaches were tested which produced varying results. The whole network approach proved to be faster and less complex to implement but required a minimum of 45% of the total fleet to be installed with metering equipment to provide satisfactory energy billing estimates (based on the A car configuration). The route specific approach data took significantly longer to pre-process and was more complex to implement, however, results were improved. Mean partial fleet requirement was a minimum of 22% of the total fleet (based on the B car configuration). This was below the 30% figure that ATOC required.

What would the minimum size of such a fleet be and what level of statistical accuracy would estimate bills have?

As described above the minimum size of a partially metered fleet could be as low as 22%. This figure uses results from the B car configuration only. Recommendations in chapter 6 consider results from both A and B car configurations, which suggest that a range of 22 – 35% of a fleet would be adequate to provide accurate energy estimates. The C car configuration provided poor results in both approaches. It is though that the C car configuration would provide better results if input data normalisation could be improved. At present it is not clear why this configuration does not yield similar results to A and B car configurations as well as results produced from earlier simulation and smaller fleet analyses [96]. In the event that the datasets can be kept long term and future datasets are made available for use, improving the C car configuration results would be an area where further work could be made.

The minimum partial fleet size required by ATOC was 30%, at this size energy estimation errors begin to increase. The criteria of suitability for a fleet of this size were based on the standard deviation of the energy estimation error (σ_{ε}). Under the empirical rule a $\sigma_{\varepsilon}$of 0.01 would have a 68% probability of producing a specific energy consumption estimate that is within ±1% of the recorded mean. This assumes that the estimation error is normally distributed, which larger B car configuration demonstrates.

If partial fleet metering is implemented on DC fleets that currently do not utilise full fleet metering or use dated billing models there is a potential for financial benefit. As there is a move to encourage the use of fleet metering, initially PFM could reduce the installation by 70% if only 30% of a fleet is metered. More importantly the analysis can be used to charge PFM fleets more accurately than the existing billing models are able, which would likely yield fairer energy prices over a continued period.

8.3 Key achievements

The key achievements of this research are listed below.

- Developed a methodology to determine driver styles in operation on a DC rail network.
- Identified driver cultures and their effect on energy consumption.
- Optimised driver cultures to provide energy savings of 4 – 10%

- Identified a series of optimised driver control required to adhere to lower energy driver cultures.
- Developed two methodologies to determine minimum required fleet percentage for Partial fleet metering of a DC EMU fleet
    - A whole network methodology suitable for a dataset that does not have vehicle position data available.
    - A route specific methodology to be applied to individual route where vehicle position data is available.
- Using the route specific approach it was shown that a minimum of 22% of a fleet in operation on a particular route could be used for partial fleet metering.
- Identified a minimum metering period of 7 months is needed to provide accurate PFM energy consumption estimates.
- Designed a methodology to test a successive year's meter data to establish significant similarity or dissimilarity to the original fully metered fleet data.

This research set out to establish whether there is additional value to be found in the use of Rich datasets for varied railway applications. The methodologies developed have demonstrated substantial energy and financial savings can be made if implemented on a rail network. These findings reuse existing data that were recorded for different original purposes and utilise either inexpensive or existing instrumentation. This allows the potential implementation of both methodologies to other rail networks at low cost with short term benefits that can provide an alternative to high cost driver advisory systems, in the case of the Merseyrail case study, and full fleet meter fitment, in the case of the ATOC partial fleet metering case study.

9 References

[1] A. Thaduri, D. Galar and U. Kumar, “Railway Assets: A Potential Domain For Big Data Analytics,” vol. 53, p. 457–467, 2015.

[2] N. Attoh-Okine, “Big Data Challenges In Railway Engineering,” in *IEEE International Conferen* D.C., 2014.

[3] A. M. Zarembski, “Some Examples Of Big Data In Railroad Engineering,” in *2014 IEEE Internatio* Washington D.C., 2014.

[4] Technical Strategy Leadership Group, “Rail Technical Strategy 2012,” RSSB, 2012.

[5] R. J. Ellis, S. Hillmansen, E. Stewart, P. Tricoli, P. Weston and C. Roberts, “Observations Of Trai Camshaft Operated DC Electrical Multiple Unit,” *Proceedings Of the Institution Of Mechanical E Rail And Rapid Transit,* pp. 1-18, DOI: 10.1177/0954409715589618, 2015.

[6] R. Ellis, P. Weston, S. Hillmansen and I. Jones, “Identification Of A 'Driver Culture' And Its Effect A DC Rail Network,” in *International Symposium On Speed-Up And Sustainable Technology For I* Chiba, 2015.

[7] R. Ellis, G. Nicholson, S. Lu, S. Hillmansen, P. Tricoli and N. Ovenden, “A Method to Derte Percentage On A Rail Network,” *Undetermined - Paper in preparation,* pp. 1 - 15, 2016.

[8] J. Dean, Big Data, Data Mining And Machine Learning; Value Creation For Buisiness Leaders And Jersey: John Wiley & Sons, 2014.

[9] F. X. Diebold, “"Big Sata" Dynamic Factor Models For Macroeconomic Measurement And Forecasti *And Econometrics, 8th World Congress Of The Econometric Society*, 2003.

[10] S. Lohr, “New York Times - Bits; The Origins Of ‘Big Data': An Etymological Detective Story, Available: http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detect April 2016].

[11] F. X. Diebold, *A Personal Perspective On The Origin(s) And Development Of \Big Data":The Phe Discipline*, Pennsylvania: University of Pennsylvania, 2012.

[12] M. Minelli, M. Chambers and A. Dhiraj, Big Data Big Analytics; Emerging Business Intelligence An Businesses, Hoboken, New Jersey: John Wiley & Sons, 2013.

[13] J. M. Tien, “Big Data: Unleashing Information,” *Journal Of Systems Science And Systems Engine* 151, 2013.

[14] S. R. Qureshi, “Towards Efficient Big Data And Data Analytics,” in *2014 Conference On Industry DOI: 10.1109/CSIBIG.2014.7056933*, Indore, 2014.

[15] S. Sagiroglu and D. Sinanc, “Big Data: A Review,” in *International Conference On Collaboratio (CTS),* DOI: 10.1109/CTS.2013.6567202, 2013.

[16] M. van Rijmenam, “Datafloq: Why 3V's Are Not Sufficient To Describe Big Data,” 7 Augus https://datafloq.com/read/3vs-sufficient-describe-big-data/166. [Accessed 6 April 2016].

[17] J. Hurwitz, A. Nugent, F. Halper and M. Kaufman, Big Data For Dummies, Hoboken, New Jersey:

[18] M. Chen, S. Mao and Y. Liu, “Big Data: A Survey,” *Mobile Networks And Applications,* vol. 19, no

[19] S. Madden, “From Databases To Big Data,” *IEEE Internet Computing,* pp. 4 - 6, May/June 2012.

[20] J. Gantz and D. Reinsel, “Extracting Value From Chaos,” IDC IView, Framingham, Massachusetts,

[21] J. M. Tien and D. Berg, “A Case For Service Systems Engineering,” *Journal Of Systems Science A* 12, no. 1, pp. 13-38, 2003.

[22] L. Chen, C. Roberts, F. Schmid and E. Stewart, “Modeling And Solving Real-Time Train Resch Bottleneck Sections,” *IEEE Transaction On Intelligent Transportation Systems,* vol. 16, no. 4, pp. 1

[23] M. Mazzarello and E. Ottaviani, “A Traffic Management System For Real-Time Traffic *Transportation Research Part B,* vol. 41, p. 246–274, 2005.

[24] M.-P. Pelletier, M. Trépanier and C. Morency, “Smart Card Data Use In Public Transit: A Literat *Research Part C,* vol. 19, p. 557–568, 2011.

[25] M. Faizrahnemoon, A. Schlote, L. Maggi, E. Crisostomi and R. Shorten, “A Big-Data Mo Transportation With Application To Macroscopic Control And Optimisation,” *International Journa* pp. 2354–2368, DOI: 10.1080/00207179.2015.1043582, 2015.

[26] P. Bouman, M. Lovric, T. Li, E. van der Hurk, L. Kroon and P. Vervest, “Recognizing Demand Pa For Agent-Based Micro-Simulation Of Public Transport,” in *Proceesings Of The Seventh Worksh Transportation*, 2012.

[27] I. Gokasar and S. Kevser, “Using “Big Data” For Analysis And Improvement Of Public Transport *ASE Big Data/Socialcom/Cybersecurity Conference,* Stanford, 2014.

[28] B. Agard, C. Morency and M. Trépanier, “Mining Public Transport User Behaviour From Sma
Symposium On Information Control Problems In Manufacturing - INCOM, Saint-Etienne, 2006.

[29] M. Bagchi and P. R. White, “The Potential Of Public Transport Smart Card Data,” *Transport Policy*

[30] M. Utsunomiya, J. Attanucci and N. Wilson, “Potential Uses Of Transit Smart Card Registration And
Transit Planning,” *Transport Research Record: Journal Of The Transportation Research Board,* no

[31] R. W. Ngigi, C. Pislaru, A. Ball and F. Gu, “Modern Techniques For Condition Monitoring Of
Journal Of Physics: Conference Series, vol. 364, no. 1, 2012.

[32] X. Wei, F. Liu and J. Limin, “Urban Rail Track Condition Monitoring Based On In-Service Vehicle
Measurement, vol. 80, pp. 217-228, 2016.

[33] A. Soni, S. Robson and B. Gleeson, “Extracting Rail Track Geometry From Static Terrestrial
Purposes,” in *The International archives Of Photogrammetry, Remote Sensing And Spatial Informa
Commission V Symposium,* Riva del Garda, 2014.

[34] Severn Partnership, “Mobile Mapping - Market Harborough Station,” 201
http://www.severnpartnership.com/case_study_item/case-study-market-harborough-station/. [Acces

[35] S. Hillmansen, “Sustainable Traction Drives,” in *5th IET Professional Development Course
Infrastructure And Systems (REIS 2011),*, London, 2011 doi:10.1049/ic.2011.0188.

[36] G. J. Hull, C. Roberts and S. Hillmansen, “Simulation Of Energy Efficiency Improvements On Com
Traction Systems (RTS 2010), IET Conference, 2010.

[37] Association Of American Railroads, *The Environmental Benefits of Moving Freight By Rail,* Wa
American Railroads, 2015.

[38] P. Johnson and S. Brown, “A Simple In-Cab Schedule Advisory System To Save Energy And Imp
in *IET Conference On Railway Traction Systems (RTS 2010),* Birmingham, 2010 doi:10.1049/ic.20

[39] Network Rail, “Delivering a Railway Fit For The Future,” Network Rail, Milton Keynes, 2014/15.

[40] Department of Energy and Climate Change, “2014 UK Carbon Emissions,” Department of Energy a
Statistics, London, 2014.

[41] Network Rail, “Sustainability Update 2013 Performance Summary,” Network Rail, Milton Keynes,

[42] T. Ratniyomchai, S. Hillmansen and P. Tricoli, “Recent Developments And Applications Of Energy
Railways,” *IET Electrical Systems In Transportation,* vol. 4, no. 1, pp. 9 - 20, 2014.

[43] H. Douglas, C. Roberts and S. Hillmansen, “Method To Evaluate Solutions For Complex Systems: *The Institution Of Civil Engineers - Transport,* p. DOI: http://dx.doi.org/10.1680/jtran.16.00017, 20

[44] R. Liu and I. M. Golovitcher, “Energy-Efficient Operation Of Rail Vehicles,” *Transportation Rese* 932 doi:10.1016/j.tra.2003.07.001, 2003.

[45] L. Yang, T. Liden and P. Leander, “Achieving Energy-Efficiency And On-Time Performance With *Intelligent Rail Transportation (ICIRT), 2013 IEEE International Conference* 10.1109/ICIRT.2013.6696260.

[46] T. Albrecht, C. Gasse, A. Binder and J. van Luipe, “Dealing with Operational Constraints In Ener *Conference on Rail Traction Systems (RTS),* Birmingham, 2010 doi:10.1049/ic.2010.0028.

[47] I. Mitchel, “The Sustainable Railway Use Of Advisory Systems For Energy Savings,” *IRSE News,* p

[48] A. Lindgren, A. Angelelli, P. A. Mendoza and F. Chen, “Driver Behaviour When Using An Integrat For Advanced Ariver Assistance Systems,” *IET Intelligent Transport Systems,* vol. 3, no. 4, p. 390–

[49] A. Morris, S. Reed, R. Welsh, L. Brown and S. Birrell, “Distraction Effects Of Navigation And Gre From Field Operational Tests (FOTs) In The UK,” *European Transport Research Review,* vol. 7, 015-0175-3, 2015.

[50] J. D. Lee, J. D. Hoffman, H. A. Stoner, B. D. Seppelt and M. D. Brown, “Application Of Ecologic Support Systems,” in *Proceedins To The IEA 16th World Congress On Ergonomics,* Maastricht, 20

[51] T. Brunetti Sayer, “Assessment Of A Driver Interface For Lateral Drift And Curve Speed Warning Auditory And Haptic Warnings,” in *Thid International Driving Symposium On Human Factors In And Vehicle Design,* Iowa, 2005.

[52] RSSB, “GB operational Concept Standalone Driver Advisory System (S-DAS),” RSSB, London, 2

[53] K. Rahn, C. Bode and T. Albrecht, “Energy-Efficient Driving In The Context Of A Communication (CBTC),” in *2013 IEEE International Conference On Intelligent Rail Transportatio* doi:10.1109/ICIRT.2013.6696261.

[54] T. Albrecht and J. van Lupien, “What Role Can A Driver Information System Play In Railwa *Symposium On Control In Transportation Systems*, 2006.

[55] A. Benjamin and M. Vijay, “Analyitics Informs Beyond ‘Moneyball’:The Rapidly Evolving World September 2011. [Online]. Available: http://www.analytics-magazine.org/special-articles/391-b evolving-world-of-sports-analytics-part-i.pdf. [Accessed 12 August 2014].

[56] P. Lucey, A. Bialkowski, P. Carr, E. Foote and I. Matthews, “Characterizing Multi-Agent Team Tracings: Evidence From The English Premier League,” Association For The Advancement Of Arti

[57] E. Misirlisoy and P. Haggard, “Asymmetric Predictability And Cognitive Competition In Football *Biology,* vol. 24, no. 16, pp. 1-5 doi: http://dx.doi.org/10.1016/j.cub.2014.07.013, 2014.

[58] Office of Rail Regulation, “Costs and Revenused of Franchised Passenger Tain Operators in the UK

[59] Merseyrail Electrics 2002 Limited, “Report and Fiancial STatements,” Merseyrail Electrics 2002 Li

[60] I. Jones, Interviewee, *Fleet Engineer, Merseyrail.* [Interview]. 12 August 2013.

[61] E. Stewart, P. Weston, C. Roberts and S. Hillmansen, “Monitoring The Energy Consumption In DC *On Railway Traction Systems (RTS 2010),* Birmingham, 2010.

[62] S. Hillmansen and C. Roberts, “Energy Storage Devices In hybrid Railway Vehicles: A Kinematic A *Institution Of Mechanical Engineers, Part F: Journal Of Rail And Rapid Transit,* vol. 221, pp. 135

[63] Y. V. Bocharnikov, A. M. Tobias, C. Roberts, S. Hillmansen and C. J. Goodman, “Optimal Driving Saving On DC Suburban Railways,” *The Institution Of Engineering And Technology, Power Applic* 682, 2007.

[64] S. Lu, S. Hillmansen, T. K. Ho and C. Roberts, “Single Train Trajectory Optimisation,” *IEEE Transportation Systems,* vol. 14, no. 2, pp. 743 - 750, 2013.

[65] D. Meegahawatte, S. Hillmansen, C. Roberts and M. Falco, “Analysis Of A Fuel Cell Hybrid Commu *Of Power Sources,* vol. 195, no. 23, p. 7829–7837, 2010.

[66] Rail Accident Investigation Branch, “Rail Accident Report: Derailment near Liverpool Central Un 2005,” Rail Accident Investigation Branch, Department for Transport, Derby, 2006.

[67] General Electric, “Class 507 Instrumentation Technical Documentation,” Undated.

[68] P. Connor, “Schematic Of A Simple Traction Motor Power Control Circuit,” Railway Technical We

[69] 2015 Google, “Google Maps Image Of Merseyrail Network, Hunts Cross To Southport,” 7 Octob https://www.google.co.uk/maps/@53.5021561,-3.1040094,11z. [Accessed 7 October 2015].

[70] 2015 Google, “Google Maps Image Of Merseyrail Network, Liverpool James St. To West Kirby Available: https://www.google.co.uk/maps/dir///@53.4141592,-3.1030997,12.25z/data=!4m2!4m1 2015].

[71] C. Landi, M. Luiso and N. Pasquino, “An On-Board Monitoring System For Electrical Railw *Transactions, Instrumentation And Measurment,* vol. 57, no. 10, pp. 2250-2256 doi:10.1109/TIM.2

[72] A. K. Jain, M. N. Murty and P. J. Flynn, “Data Clustering: A Review,” *ACM Computing Surveys* 1999.

[73] A. K. Jain, “Data Clustering: 50 Years Beyond K-Means,” *Pattern Recognition Letters,* vol. 31, no.

[74] P. S. Lloyd, “Least Squares Quantization In PCM,” *IEEE Transactions On information theory,* Vol 1982.

[75] J. MacQueen, Some Methods For Classification And Analysis Of Multivariate Observations, Univers pp. 281 - 297.

[76] G. Ball and D. Hall, “ISODATA, A Novel Method Of A Data Analysis And Pattern Classification 699616,” Stanford Research Institure, Stanford, CA, 1965.

[77] K. Pearson, “On Line And Plance Of Closest Fit To Systems Of Points In Space,” *Philosophical M* 11, pp. 559-572, 1901.

[78] M. Sewell, *Principal Component Analysis,* London: University College London, 2008.

[79] H. Hotelling, “Analysis Of A Complex Of Statistical Variables Into Principal Components,” *Journ* vol. 24, no. 6 & 7, pp. 417-441 & 498-520, 1933.

[80] S. G. Hoggar, Mathematics Of Digital Images: Creation, Compression, Restoration, Recognit University Press, 2006.

[81] I. T. Jolliffe, Principal Component Analysis Second Edition, New York: Springer, 2002.

[82] L. I. Smith, *A Tutorial On Principal Components Analysis,* Ithica: Cornell University, 2002.

[83] V. Lavrenko, *Lectures 18 and 19 In The Introductory Applied Machine Learning (IAML) Cour* Edinburgh: Victor Lavrenko, [online] https://www.youtube.com/playlist?list=PLBv09BD7ez_5_yapAg86Od6JeeypkS4YM, 2014.

[84] J. C. Dunn, “A Fuzzy Relative Of The ISODATA Process And Its Use In Detecting Compact Well *Of Cybernetics,* vol. 3, no. 3, pp. 32-57, 1973.

[85] D. L. Davis and D. W. Bouldin, “A Cluster Separation Measure,” *Pattern Analysis And Machine In On,* Vols. PAMI-1, no. 2, pp. 224 - 227 doi:10.1109/TPAMI.1979.4766909, 1979.

[86] P. J. Rousseeuw, “Silhouettes: A Graphical Aid To The Interpretation And Validation Of C *Computational And Applied Mathematics,* vol. 20, pp. 53 - 65 doi:10.1016/0377-0427(87)90125-7,

[87] E. Rendon, I. Abundez, A. Arizmendi and M. E. Quiroz, “Internal Versus External Cluster Valid *Journal Of Computers And Communications,* vol. 5, no. 1, pp. 27-34, 2011.

[88] Merseyrail, “Standard Tickets,” Merseyrail, 2016. [Online]. Available: http://www.merse travel/standard-tickets.aspx. [Accessed 2 February 2016].

[89] British Broadcasting Corperation, “BBC News - Peak Rail Times 'Are Confusing', Which? Suggest Available: http://www.bbc.co.uk/news/uk-11092037. [Accessed 2 February 2016].

[90] L. Yang, T. Liden and P. Leander, “Achieving Energy Efficient And On-Time Performance With *2013 IEEE International Conference On Intelligent Rail Transportation (ICIRT),*, Beijing, 2013.

[91] R. Kuwahara, T. Aoki, Y. Kamo and Toshiba Corporation, “Eco-Running Profile Trial Run In *International Symposium On Speed-Up And Sustainable Technology For Railway And Maglev Syste*

[92] T. Albrecht, A. Binder and C. Gassel, “Applications Of Real-Time Speed Control In Rail-Bound Pu *IET intelligent Transport Systems,* vol. 7, no. 3, pp. 305 - 314, 2012.

[93] N. Cummings, “Predictive Analytics Makes Its Mark On Rubgy,” *The Analytics Network,* pp. 1-2, A

[94] Network Rail, “PR13: ORR Consultation On Electricity For Traction Charges For CP5,” Network R

[95] S. Hillmansen, P. Tricoli and K. Taransenga, “Statistical Analysis Of The Efficiacy Of Partial Fleet F Follow On Work,” ATOC Report Reference - 120330/BCRRE/SH/ATOC02, Birmingham, 2012.

[96] G. Nicholson and S. Hillmansen, “Statistical Analysis Of The Efficiacy Of Partial Fleet Fitment Of Birmingham, 2011.

[97] Depatment for Transport, “National Rail Travel Survey Overview Report,” Department for Transpo

[98] G. R. van Brummelen, Heavenly Mathematics: The Forgotten Art Of Spherical Trigonometry, Princ

[99] N. Ovenden, Interviewee, *Engineering Manager.* [Interview]. 2 March 2015.

[100] Network Rail, “Traction Electricity Rules,” Network Rail, Milton Keynes, 2017.

[101] Souteastern, “Environment, Our Commitment To The Environment,” 201 https://www.southeasternrailway.co.uk/about-us/environment. [Accessed 26 March 2017].

[102] Southeastern, “Southeastern Environmental & Social Report 2007,” 2007. [Online]. ahead.com/content/dam/go-ahead/corporate/documents/Sustainability%20Reports/PerformanceAndReports/Archive/southeaste [Accessed 26 March 2017].

[103] S. Saifuddin, *Characteristic Functions And The Central Limit Theorem,* Waterloo: University Of W

[104] H. Fischer, A History Of The Central Limit Theorem; From Classical To Modern Probability Theor

[105] R. V. Hogg, E. A. Tanis and D. L. Zimmerman, Probability And Statistical Inference 9th Edition, N

[106] C. Bunks and D. McCarthy, “Condition-Based Maintenance Of Machine Using Hidden Markov N
And Signal Processing, vol. 14, no. 4, pp. 597-612, 2000.

[107] A. Kumar, F. Tseng, Y. Guo and R. B. Chinnam, “Hidden Markov Model Based Sequential
Diagnostics,” in *International Joint Conference On Neural Networks*, Hong Kong, 2008.

10 Appendix

10.1 Further work

This section gives a brief overview of two areas of research that were conducted during the second year of the PhD, due to limitations of the datasets that were in use findings could not be verified. As a result the author decided to conclude further experimental research with the aim to continue the work in the future. Also this area of research was discussed with supervising staff, where it was decided that potential research projects could be designed for future research students at PhD or MSc. level providing that improved datasets became available.

10.1.1 Dataset analysis using hidden Markov models

Hidden Markov models (HMMs) are commonly used in the field of speech recognition. A normal Markov model is a stochastic process that is used to give the probability of a system’s state changes based on the current state only; memory of previous states are not required to make predictions of future states. A HMM uses observations related to hidden states to predict state changes i.e. the likelihood of a certain observation can be used to determine a current state which then allows prediction of a future state. In the case of speech recognition a particular word, the state, is unknown. A sound, the observation, can be used to give likelihood of which part of a word is being spoken; this can be used to determine the probability of which word is uttered. Speech recognition is dependent on the predetermination of sounds and their relation to particular words; this is known as training data.

The concept has been used in a number of other engineering disciplines, [106] shows identification of faulty gearbox states and [107] describes the use of HMMs to pinpoint the characteristics of drill bit failure. It was thought that HMMs could be used to identify individual drivers in the Merseyrail dataset. An HMM was used on the Hunts Cross to Southport dataset. However, as driver identification i.e. which driver was driving on a particular day or time was not contained in the dataset and this information being deleted by Merseyrail by the time of analysis, it was impossible to determine a training dataset to make comparisons to. As a result any matching of data to a particular 'driver' could never be validated. As a proof of concept, selections of data were used as training input. These were then compared to the whole dataset where the training set was also included. The HMM matched various drivers to the training set, and importantly, matched training data to their equivalent inputs in the whole dataset, suggesting that HMMs would be successful in identifying like driver styles and potentially individual drivers. A continuation of this research would require driver identification to be included in the dataset, or available long term to be manually linked to each journey. This could be done simply with a number designation for each driver. Secondly there would need to be many examples of each driver in the dataset. As this data comes from a single vehicle there is no guarantee that there are examples of every driver recorded.

It was thought that HMMs might also be suitable for identifying track characteristics. An analysis was conducted to try to identify examples of rail wheel slip based on the effect that such an event would cause in electrical data. Slip would cause a very brief current spike followed by a cut-off to 0. This is a result of the wheel slip protection equipment intervening to reduce the slip. The author inquired with Merseyrail about areas that have common adhesion problems, in the hope that data from these areas could be used as training data as they ought to provide good examples of the current cut-off. Two problems became evident, 1) there were not enough examples of the expected current characteristic and 2) it was not possible to confirm that the characteristics that were identified were actually related to slip events. As a result this area of research was not continued, although it is thought that the concept would be successful if specific slip examples were known and could be used as training data.

10.1.2 Effects of weather changes on driver style

Of particular interest to the author were the possible effects of weather changes on driver style, for example:

- Do drivers become more defensive during heavy rain?
- Are there seasonal variations in driver strategy?

To approach this author obtained access to the MIDAS weather database. Weather data, particularly precipitation amount, was taken and matched to the relevant operational period in the Merseyrail database. Over the three month period, September to November, there was no discernible effect of changing levels of precipitation on driver behaviour. However, MIDAS weather data is only available for the central Liverpool region, meaning that weather data is at a relatively low resolution geographically. Low resolution is also a factor in the data rate, with most weather information being taken only hourly. The lack of a relationship between weather data and driver style is likely related to this. For a continuation of this project it would be beneficial to have humidity and temperature capability build into the instrumentation to allow for higher resolution recording of this data.

Seasonal variation regenerative braking use on the ATOC dataset was investigated at the request of ATOC. It was expected that drivers would use regenerative braking less during periods in the year where adhesion typically becomes lower, i.e. in the autumn and winter months where leaf mulch, precipitation and frost conditions are more frequent. This is a result of drivers being encouraged to drive more defensively, meaning that there is less opportunity to coast or use dynamic brakes. Analysis showed that there was actually increase in regenerated braking energy early in the year, the summer months have a slight dip followed by the lowest in late summer/autumn. However, the difference between seasonal variation was within electricity meter error and likely to be a result of noise rather than an actual observation. As a result in both Merseyrail and ATOC datasets a knowledge gap remains that if approached in a continuation of this research could potentially provide significant financial benefit to the rail industry.