You are reminded that all work submitted as part of the requirements for any examination (including coursework) of Imperial College must be expressed in your own writing and incorporate your own ideas and judgements.

Plagiarism, that is the presentation of another person's thoughts or words as though they are your own, must be avoided with particular care in coursework, essays and reports written in your own time. Note that you are encouraged to read and criticise the work of others as much as possible. You are expected to incorporate this in your thinking and in your coursework and assessments. But you must acknowledge and label your sources.

Direct quotations from the published or unpublished work of others, from the internet, or from any other source must always be clearly identified as such. A full reference to their source must be provided in the proper form and quotation marks used. Remember that a series of short quotations from several different sources, if not clearly identified as such, constitutes plagiarism just as much as a single unacknowledged long quotation from a single source. Equally if you summarise another person's ideas, judgements, figures, diagrams or software, you must refer to that person in your text, and include the work referred to in your bibliography and/or reference list. Departments are able to give advice about the appropriate use and correct acknowledgement of other sources in your own work.

The direct and unacknowledged repetition of your own work, which has already been submitted for assessment, can constitute self-plagiarism. Where group work is submitted, this should be presented in a way approved by your department. You should therefore consult your tutor or course director if you are in any doubt about what is permissible. You should be aware that you have a collective responsibility for the integrity of group work submitted for assessment.

The use of the work of another student, past or present, constitutes plagiarism. Where work is used without the consent of that student, this will normally be regarded as a major offence of plagiarism.

Failure to observe any of these rules may result in an allegation of cheating. Cases of suspected plagiarism will be dealt with under the College's Cheating Offences Policy and Procedures and may result in a penalty being taken against any student found guilty of plagiarism.

Declaration of Authorship

I certify that I have read the definition of plagiarism given above, and that the work submitted for this coursework assignment is my own work, except where specifically indicated otherwise. In signing this document I agree that this work may be submitted to an electronic plagiarism test at any time and I will provide a further version of this work in an appropriate format when requested:

Signed: ____________________________________

Date: ____________________________________

Imperial College London

Master's Research Project

Analysis of Travel Time Reliability on London Underground

Author:
Fanny Heneine

Supervisor:
Prof. Daniel Graham

A project submitted in fulfilment of the requirements for the degree of Master of Engineering

in the

Railway and Transport Strategy Centre
Department of Civil and Environmental Engineering

June 2015

IMPERIAL COLLEGE LONDON

Abstract

Railway and Transport Strategy Centre
Department of Civil and Environmental Engineering

Master of Engineering

Analysis of Travel Time Reliability on London Underground

by Fanny Heneine

The implementation of Automatic Fare Collection(AFC) has induced a major leap in travel time analysis. By processing the data that has become available, travel time between any two stations can be calculated. Passenger expect from the public transit agencies a predictable service and a steady quality, and this invariability in the service attributes is defined as reliability. This research project focuses on quantifying reliability of journey time on London Underground from the user's perspective. Through nonparametric distribution fitting and regression modelling, the underlying factors affecting journey time are examined. The standard deviation of journey time between an origin station and its destination constitutes a direct indicator of its reliability, while reliability on a selected line at a designated time can be estimated according to semi-parametric regression models. These methods can be used by Transport for London as an effective tool to assess system performance and result in planning that reflects optimally the needs of travellers. They could also be developed and merged into a single reliability metric that would be made available to the users through journey planners.

Acknowledgements

I would like to thank Professor Graham for his supervision and guidance throughout every step of the process, and his help in structuring my research. I would also like to thank Ramandeep Singh for her involvement, and friendly advice, that were essential to my progress and finally, Kawthar Aljufairi for proof-reading my paper and for her constant support throughout this project and during my four years at Imperial College.

Contents

Declaration of Authorship **ii**

Abstract **i**

Acknowledgements **ii**

Contents **iii**

1 Introduction **1**

1.1 Introduction to the project 1

1.1.1 Automated Fare Collection System 1

1.1.2 Reliability 2

1.1.3 Application to the London Underground 3

1.2 Aims and Objectives 3

1.2.1 Research Question 3

1.2.2 Structure of thesis 5

2 Literature Review **6**

2.1 Travel Time Reliability 6

2.2 Reliability Metrics 7

2.3 Understanding reliability through regression models 8

3 Statistical Methods Used **9**

3.1 Parametric and Non parametric Distributions 9

3.1.1 The log-normal distribution 9

3.1.2 Nonparametric density estimation 10

3.1.3 Kernel Density Estimator 10

3.2 Regression 11

3.2.1 Polynomial Regression 11

3.2.2 Nonparametric Regression 11

3.2.3 Semiparametric Regression 12

3.3 Goodness of fit tests 12

3.3.1 Kolmogrov-Smirnov Test 12

3.3.2 Permutation Testing 13

3.3.3 Cross Validation and K-Folds 13

4 Journey Time Distribution Analysis **15**

4.1 Introduction to the Data 15

4.2 Data and Distributions Analysis . . . 15
4.2.1 Origin-Destination pairs along the same line . . . 15
4.2.2 Distribution Model Fitting . . . 21
4.2.3 O-D pair including an interchange . . . 25

5 Regression Analysis of Journey Time 29
5.1 Regression on the Victoria Line . . . 29
5.1.1 Regression of reliability with time . . . 30
5.1.2 Reliability with respect to standard deviation . . . 34
5.2 Regression on the Jubilee Line . . . 37
5.2.1 Regression of Reliability with Time . . . 37
5.2.2 Reliability with respect to Standard Deviation . . . 39
5.3 Findings . . . 42
5.4 Regression analysis on pairs with interchange . . . 43

6 Conclusion and Further Research 45

7 Bibliography 48

A R Code 50

B Figures and Underground Map 54

Chapter 1

Introduction

1.1 Introduction to the project

1.1.1 Automated Fare Collection System

In the main megalopolis of the world, public transport has become an essential tool in the day-to-day life of the urban population. It has progressively become an indicator of the level of growth of a city, contributing to its international reputation. Therefore, the transit agencies thrive to provide a service which quality and performance are satisfactory for the users, mainly through its efficiency and predictability. The performance of the service is constantly examined in order to highlight possible problems and implement new strategies that would enhance the passengers' experience. In the recent decade, the emergence of the automatic data collection system has given access to a new level of travel data analysis and has generated a broad range of applications, well beyond the previous manual data collection that restricted the extent of monitoring. Through the high accuracy of this ”big data”, the agencies can fully monitor the quality of their service and operate efficiently and cost-effectively (Chan, 2007).

The Automated Fare Collection system (AFC) is a type of automatic collection system. It has been increasingly adopted to automate the ticketing system of public transportation networks, since the introduction of the concept with the Octopus Card in Hong Kong in 1977. The AFC consists of smart cards that are tapped at a fare gate at the entry and exit stations, using sensors implanted in the cards, recording the time of entry and exit. They provide a detailed source of data about peoples transport habits as the duration of the trip can be calculated and assigned to peak or non-peak [1] travelling time, giving access to very detailed data that will quantify and measure the performance of the network.

[1] For London Underground, Transport for London defindes morning peak time as $6:30$am to $9:30$am and afternoon peak time as 4pm to 7 pm

Indeed, simple adherence to the operating plan of the system network has become insufficient to operate a transportation network. Efficiency has to be constantly improved and adapted to the users' need, while predictability of journey time appears as the main concern of users, and especially daily commuters. The detailed AFC data is an essential tool towards those objectives, as it can be used to conduct analysis which outcome will result in strategies to improve the performance of the network and optimise its running efficiency. The automated measurement of the journey times through entry and exit fare validation has become a mean of monitoring and detecting modifications in transit quality. The causes of these changes can be identified and the quentsubse passenger behaviour can be predicted. Moreover, Batty (2012) has demonstrated how AFC system data could reveal hidden aspects of travel behaviours.

Thus, the Automated Fare Collection system, by giving access to individual traveller information, is seen as the key to inform the design of future transport information services. By understanding mobility and travel behaviour data of individuals, the traditional transport information systems could evolve into personalised information spaces. (Foell et al., 2013)

1.1.2 Reliability

The users of the tube are more concerned with the predictability of a certain Origin-Dsetination (O-D) trip, rather than its mean journey time. The service provided is thus expected to be reliable. Abkowitz et al. (1978) defined reliability as "the invariability of service attributes which influences the decision of the travellers and transportation providers" and it can be defined from both the passenger and the supply-side perspective, with each having its own focus. Abkowitz et al. (1978) defines the operator reliability as the adherence to the operating plan in terms of schedule and capacity used in the vehicles. Reliability is central for transit operators as they are impacted in two ways: A variable service renders the efficient use of resources hard to achieve, increasing the cost of operation. Likewise, an unreliable service impacts the quality of the passenger experience, possibly causing changes in ridership rate and consequently, revenue (Uniman, 2009). With the emergence of technologies allowing vehicle location data, many studies focused on supply-side reliability by studying the relationship between operational quality and passenger experience, which was estimated through surveys. However, the recent emergence of the AFC systems, with very detailed smart card data, has yielded opportunities to look at reliability from the passengers' perspective without conducting manual surveys that only provided aggregate data. Indeed, users are most concerned about the predictability of journey times (Paine et al., 1976), which consitutes service reliability from their perspective. Any variability impacts their perceptions of service quality and accordingly, their travel behaviour (Polak et al., 2001).Therefore, transit operators aim to quantify passenger reliability in order to understand its underlying components and their impact on travel time. The development of Automated Fare Collection and the wealth of data it provides, represent essential information towards measuring this passenger reliability; an objective which this research paper investigates on the London Underground, using Oyster card data.

1.1.3 Application to the London Underground

The London Underground, serving a megalopolis of 8.3 million individuals and 270 stations over 402 kilometres (along 11 lines), is an impressive transport infrastructure that contributes to the international stature of the city of London. With 20 millions trips on a daily basis (week day), it is an extremely congested and complex network that requires constant monitoring and maintenance to sustain a satisfactory service.

Since 2000, Transport for London (TfL) is the organisation charged with managing Greater London road network and public transport modes. It is controlled by the mayor's office, who is responsible for the setting the budget and the long term vision for the city. When London Underground became part of TfL in 2003, the organisation introduced the Oyster smart-card system. With a steady decrease in the number of trips made by car in London, the public transports have been experiencing an increase of their share within transit. The Underground is the second most used transit mode after the bus: 27% of all trips as of 2007 (TfL, 2007). 85% of all trips are made using Oyster card and this proliferation of the smart-cards has given access to a wealth of data. In a business plan published in December 2007, Transport for London summarises its objectives towards achieving the mayor's vision: One of the key objectives to be implemented in the transport strategy is improving the regularity and reliability of door-to-door journey times on the Underground, an improvement on which this paper is centered.

1.2 Aims and Objectives

1.2.1 Research Question

The users of public transport expect from public agencies the provision of a predictable service. Commuters, in particular, through the daily repetition of the same trip demand a constant quality of service, exempt of any unexpected delay. With the recent emergence of tools as journey planners, users can evaluate the journey time of the intended trip, track live information about the transport services to be used and hence, predict the approximate journey time. However, a flaw of these systems is often overlooked, leading to unexpectedly longer trips; the reliability dimension. Indeed, these information tool only provide an indication of the mean journey time of the trip without mentioning how reliable this measure is. At a specific time of day, where congestion is as its uttermost, the journey time could be doubled and the passenger is subject to an unpredictably longer trip. By underestimating the travel time variability on the Underground, passenger bear longer journey times and endure the consequences of an unreliable service. Moreover, Uniman(2009) has demonstrated that the information provided by TfL on its passenger information system (the Journey Planner tool) seldom predicted an accurate journey time and passenger had to often include a time margin to allow for this inaccuracies. This research paper focuses on finding ways to quantify reliability and characterise it, to potentially implement a reliability metric on journey planners. This new indicator would allow a better estimation of the trip to be made, and become an effective tool towards choosing the optimal path, which maximises reliability on each segment of the trip. Reliability measures could also be used by TfL for estimation and prediction purposes, leading to more rigorous planning and scheduling.

This paper addresses reliability through the use of "big data". It focuses on the passenger perspective of reliability and explores the use of real journey time data obtained through AFC, to analyse the journey time distributions for certain Origin-Destination stations. Therefore, characteristics of travel time densities between O-D pairs are determined, in order to understand why they vary. For example, the compactness of travel time distributions provides a quantitative measurement of the service quality between selected pairs, and is hence a characteristic of reliability (Chan, 2007). This will be achieved through fitting nonparametric distributions to the journey time and providing visual analytics, which allows relatively abstract data transformations to be made intelligible to both data analysis specialists and domain experts at TfL (Beecham et al., 2013). The log-normal distribution is usually employed for travel time data type, and the nonparametric fit will be tested against it.

Secondly, examining and understanding the temporal variability on the London Underground is necessary to calibrate travel time estimation and prediction. Regression models will be fitted to understand this temporal variability of reliability on specific lines. In addition, regression models will be used to highlight any potential relationship between reliability of the line and the dispersion of the journey time distribution (and hence, the standard deviation). Through these models that will be fitted at a line level, the intent is to depict the fluctuations of reliability of travel time and the drivers that affect it.

The findings produced, primarily distribution fitting at an Origin-Destination level and regression models at a whole line level could potentially be used by service providers to establish reliability measures and their impact on travel times. These reliability measures could then be communicated to passengers, thus allowing them a more accurate prediction of journey time. They could also be a useful indicator of the most unreliable lines on which TfL could improve service quality and decrease the risks of unpredictable travel times.

1.2.2 Structure of project

In order to complete this research, the first step will be to review the literature on the previous studies concerning reliability on the London underground, to lay the foundation of the existing measures of reliability. The statistical methods used to complete the research will be detailed. Then, using SPSS Statistics and R, the data will be transformed into a usable format in order to select different O-D pairs and create journey time distributions. First, OD pairs separated by one route with no interchange, will be of interest, then OD pairs in between which there are several possibilities of optimal paths and potential interchanges. The comparison between the different types of O-D pairs will clarify how the characteristics of distributions vary and are a first attempt to understand the drivers or reliability. Secondly, reliability will be studied at line level through regression model fitting, as a mean of assessing service quality on these lines. The models will be used to understand the temporal variability of reliability and to establish any relationship with the standard deviation of travel time distributions along a line (implying no interchange between stations). Moreover, the relationship of standard deviation to reliability on O-D pairs including intermediate stations and more than one optimal route will be investigated to hold a comparison with the latter case, exempt of any interchange. Non-parametric models are to be used for the distributions and regressions fitting, and are to be analysed against parametric methods results. Findings of the research will be summarised and concluded and limitations of the methods will be discussed.

Chapter 2

Literature Review

2.1 Travel Time Reliability

As already mentioned, the journey travel time can be calculated using AFC data by subtracting the entry time from the exit time. It is however important to note that the journey travel time includes numerous components: the access[1], egress[2] and interchange walking time, the platform waiting time and the on-train time. Recently, many studies have focused on reliability of travel time as experienced by the passenger. Bates et al. (2001) as cited by Uniman (2009) suggests that one way to think about reliability from the perspective of the transit patron is as "an indicator of how much and how often outcomes deviate from the planned and/or expected outcome". As such, the study by Chan (2007) was one of the first to propose a measure of reliability centered on the passenger, describing the compactness of the travel time distributions as an indicator of reliability; a tight distribution revealing a more reliable service. The reliability factor metric, which will be detailed in the next section, found its origin in this principle. Using Oyster data of the London Underground, the study has demonstrated that the AFC can provide analysts with a better understanding of passenger demand and services, by developing an Origin-Destination flow matrix and service reliability metrics that interpret the variation in journey time between a given O-D pair. Uniman (2009) has proven that reliability is related to average performance and has established that the sole focus of transit agencies on average performance leads to an underestimation of the impacts of incidents on service quality and therefore, reliability. He highlights the importance of analysis that takes incidents and severe disruptions into consideration as they give rise to a non-negligible impact on service quality. He also develops a framework to break down performance into different categories that contribute to reliability and to quantify unreliability. He finally suggests that reliability information should become available to passengers through the existing information systems tools, like journey planners. Finally, Uniman et al. (2010) have developed a set of reliability

[1] Access time is the walking time between tapping in at the entry station and the midpoint of the platform

[2] Egress time is the walking time between the mid-point of the platform and tapping out at the exit station

measures that were used to monitor performance, gain insights into the cause of unreliability, and serve as an input to the evaluation of transit services.

2.2 Reliability Metrics

Journey Time Metric was the first mean introduced by London Underground in 1997 to assess service performance from the customer's perspective: it compares the estimated actual journey times experienced by passengers during a particular four-weeks period, obtained through manual surveys, with the scheduled travel time for that same period. Excess Journey Time (EJT) is the difference between the two measures and shows how well the network performed. However, the EJT was oriented towards average performance of the network and did not reveal enough about travel time reliability, due to its dependence on aggregate and limited data. With the emergence of AFC, new metrics have surfaced. Schrank et al. (2003) has determined three types of reliability measures: Measures of variability (compactness of travel time distribution), measures of additional budgeted time, also reconsidered by Uniman et al. (2010) as the reliability buffer time metric and measures of long delays. Chan (2007) introduced the Reliability Factor (RF) as the difference between the N$^{\text{th}}$ percentile and the median journey time, where the N$^{\text{th}}$ percentile is defined as an upper threshold beyond which the journey travel time is considered to be very unfavorable for the passenger. This tool is useful towards quantifying the spread of a distribution. The value of N should be chosen as an balance between the desired sensitivity of the measure and the point at which journey time becomes affected by individual behaviour. The proposed 95% percentile is a realistic level of significance and avoids any influence due to unusual behaviour.

$$RF_{OD} = JT_{95} - JT_{50} \qquad \text{if sample size} \geq 200 \tag{2.1}$$

The reliability factor hence represents the excess journey time of 1 out of 20 customers on a specific O-D link, and this reliability metric will be used along the paper. Uniman et al. (2010) used this Reliability Factor as a starting point for their research and introduced the Reliability Buffer Time metric (RBT), as an improved definition of this reliability factor: He characterised it as the amount of time the passenger must add to its usual travel time to guarantee, with a certain level of certainty that he will arrive on time.

$$RBT_{OD} = JT_{95} - JT_{50} \qquad \text{if sample size} \geq 200 \tag{2.2}$$

Moreover, Uniman (2009) presents another method of evaluating reliability on a line level by weighting the RBT between Origin-Destination pairs using the volume or flow of passenger on this trip Vol_{OD}.

$$RBT_{line} = \frac{\sum_{OD \in line}(Vol_{OD} * RBT_{OD})}{\sum_{OD \in line} Vol_{OD}} \tag{2.3}$$

The Excess Reliability Buffer Time Metric ($ERBT$) is an extension of the RBT, and defines quantitatively a level of reliability by separating performance into two categories: typical and incident-affected. Together, the two categories are used to estimate the $RBT_{overall}$, while incidents are excluded in the $RBT_{typical}$ measure. Uniman (2009) demonstrates how this metric it could become part of routine performance monitoring on the Underground lines.

$$ERBT = (RBT_{overall} - RBT_{typical})_{OD, within\, day\, time\, interval, n-days}$$

2.3 Understanding reliability through regression models

Uniman (2009) has developed ways of characterising reliability through metrics. He has developed a framework that "can be used to provide reliability information through existing information systems, helping to mitigate the effects of uncertainty on service quality". On the other hand, Fu et al.(2012) focuses on the journey time variability by times of day and between days. However, using regression to assess reliability and establish any relationship it might hold with time of day or dispersion of the distribution is an aspect that has not been examined yet. This project will use regression models and the data of the London Underground to reveal and quantify characteristics of reliability along lines, and between specific simple and more complex paths of Origin-Destination pairs.

Chapter 3

Statistical Methods Used

3.1 Parametric and Non parametric Distributions

3.1.1 The log-normal distribution

The most common parametric distribution method is the normal distribution. It is created through the addition of a large number of small random effects. Similarly, if a phenomenon is the consequence of the multiplicative effect of a large number of uncorrelated factors, the distribution becomes logarithmic normal. In other words, the logarithm of the data is normally distributed:

$$Y = ln(X) \tag{3.1}$$

$$f_X(x) = \frac{1}{x\sigma_{ln(X)}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{ln(x) - \mu_{ln(X)}}{\sigma_{ln(X)}}\right)^2\right] \tag{3.2}$$

In order to be able to fit the log-normal distribution to the data, the empirical parameters of $f(X)$ with X being the observed data are transformed, which means that they are expressed in terms of Y, the *logarithmic* form of the data.

$$\phi = \sqrt{\sigma_X^2 + \mu_X^2} \tag{3.3}$$

$$\mu_Y = \ln \frac{\mu_X^2}{\phi} \tag{3.4}$$

$$\sigma_Y = \sqrt{\ln \frac{\phi^2}{\mu_X^2}} \tag{3.5}$$

3.1.2 Nonparametric density estimation

Without a parametric assumption for a density $f(x)$, estimation of $f(x)$ over all points would involve estimation of an infinite numbers of parameters, know as a nonparametric estimation problem. The purpose of using nonparametric estimation is to reveal structural features of the data that parametric estimation might miss (Rosso et al., 2008). However, when using nonparametric theory to fit a data set, the resulting model is an approximation, and hence it includes an estimation bias. By increasing the complexity of the model (i.e. the size of the sample), the bias decreases while the estimation variance increases. The degree of complexity is hence controlled by a bandwidth, which is determined by data dependent rules. Powell (2009) explains that in order to determine the bandwidth sequence h_n, squared bias and the variance must both converge to zero at the same speed, thus ensuring consistency of the Mean Squared Error (MSE):

$$bias = E\Big[\hat{f}(x)\Big] - f(x) = \frac{F(x+h) - F(x)}{h} - f(x) \qquad \text{where } \hat{f}(x) \text{ is the biased density of f} \tag{3.6}$$

$$V(\hat{f}(x)) = \frac{f(x)}{nh} + 0\bigg(\frac{1}{n}\bigg) \tag{3.7}$$

$$MSE(\hat{f}(x), h) = \Big[E[\hat{f}](x)] - f(x)\Big]^2 + V\Big(\hat{f}(x)\Big) \tag{3.8}$$

Nonparametric methods involve approximation in the form of smoothing. The smoothing method used in this research project is the Kernel.

3.1.3 Kernel Density Estimator

The Kernel density estimator is the simplest smoothing method. Let X be a random variable with a continuous distribution $F(x)$ and density $f(x) = \frac{dF(x)}{dx}$. In order to estimate $f(x)$ from a random sample $[X_1, X_2, \ldots, X_n]$, the kernel density estimator is

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\frac{x - X_i}{h} \tag{3.9}$$

with kernel K and bandwidth h, which controls the degree of smoothing of a density estimate. The centre of the kernel is located right over each data point and the influence of the data point is spread about its neighborhood. The smoothing window is an interval $(x + h, x - h)$, where h controls how wide the probability mass is spread around a point. The Gaussian kernel and the

Epanechnikov kernel are to be used in the analysis.

$$K(x)_{\textbf{gaussian}} = (\sqrt{2\pi})^{-1} * \exp\Big(-\frac{x^2}{2}\Big) \tag{3.10}$$

$$K(x)_{\textbf{epanech}} = \max\left[\frac{3}{4}(1-x^2), 0\right] \tag{3.11}$$

3.2 Regression

When two or more variable are related, a way of exploiting this relationship to make it useful in engineering assessments and predictions is regression analysis. It involves the use of mathematical functions to model and investigate the dependence of one variable to one or more (multivariate analysis) other observable variables.

3.2.1 Polynomial Regression

Polynomial regression is a type of parametric regression model of n^{th} order. The relationship between the response variable Y and the explanatory variable X is expressed as follows:

$$y = \beta_0 + \beta_1 * x + \beta_2 * x^2 + ... + \beta_n * x^n + \epsilon \tag{3.12}$$

where β_0, $\beta_1 \ldots$ are unknown parameters also called the regression coefficients. The error residual term ϵ represents the difference between y and the deterministic component and follows a normal distribution $\epsilon \sim N(0, \sigma^2)$. The regression coefficients are determined using a weighted least squares method. Polynomial regression allows a nonlinear relationship between the explanatory and response variable and is used to approximate functions describing complex and nonlinear relationship. The higher the degree of the polynomial and the more sensitive the estimated y values are (Rosso et Kottegoda, 2008). In order to assess the regression, tests are carried out and diagnostic plots illustrate the results. These ensure that the initial assumption of the regression are not being challenged: The error term ϵ should have zero mean, and constant variance. The residual errors should be uncorrelated and normally distributed.

3.2.2 Nonparametric Regression

The main advantage of nonparametric models over parametric is their flexibility. In the nonparametric framework the shape of the functional relationship between covariates and the dependent variables is determined by the data, whereas in the parametric framework the shape is determined by the order of the model. Nonparametric methods rely on smoothing the scatterplot data. LOESS (locally scatterplot smoothing) and LOWESS (locally weighted scatterplot smoothing) are two nonparametric regression methods using locally weighted linear regression to smooth data: The smoothed value at each data point is determined by neighboring data points within the span, for which a regression weight function is defined. While LOWESS uses a linear polynomial model, LOESS relies on a quadratic one. These two methods are useful in fitting the data without specifying a parametric function and are assumed to well approximate the mean function locally. An alternative approach to be used is spline regression. A spline is a term for elastic rules that are bent to pass through predefined points and possess a sufficiently high degree of smoothness at a given number of points called knots. However, this method generates high variance at the boundaries, which is why it is combined with smoothing and parametric methods.

3.2.3 Semiparametric Regression

Semiparametric regression is a fusion between traditional parametric regression analysis and nonparametric methods, having a finite dimensional parameter of interest and an infinite dimensional nuisance parameter, as defined by Oakes(1981), as cited by Powell (1993). It is used in order to be able to work with parameters whilst the functional form of the data to be fitted is unknown. Since, it includes a parametric component, it relies on parametric assumptions and hence has to be checked against the initial assumptions for any inconsistency. The methods used in this paper rely on smoothing spline regression (nonparametric), using parametric polynomial effects: It represents the fit as piecewise polynomials, with the pieces connecting at points called knots, such as $k << n$ with n being the total number of observations in the fitted data. In R, the function used is *bigssp*, which allows to set the bandwidth of the smoothing process and the type of piecewise polynomial.

3.3 Goodness of fit tests

3.3.1 Kolmogrov-Smirnov Test

The Kolmogorov-Smirnov goodness of fit test is a nonparametric test that quantifies the distance between the empirical distribution function and the cumulative distribution function (cdf) of a continuous variable (Hogg, 2015). The test statistics represents the maximum absolute difference (usually, the vertical distance) between the empirical and hypothetical cdf. Hence, it tests the null hypothesis that follows:
Null Hypothesis H_0: The random variable follows the hypothetical distribution
Alternative Hypothesis H_1: The random variable has a different distribution

The test criterion D is the maximum absolute difference between $F_n(x)$ and $F_0(x)$ defined as:

$$D_n = \max\left|F_n(x) - F_0(x)\right| \tag{3.13}$$

The critical value of the test criterion $D_{n,\alpha}$, for a level of significance $\alpha = 0.01$ is inversely proportional to the size of the data sample:

$$D_{n,\alpha} = \frac{1.6276}{\sqrt{n}} \tag{3.14}$$

3.3.2 Permutation Testing

The Kolmogorov-Smirnov test would appear as the best suited to check whether a certain distribution could fit the data. However, due to the big data sample n of the data treated in the research, the critical values of D_n is too close to 0 and thus not realistic. Moreover, the parameters of the theoretical distribution are derived from the data set; Conover (1999) and Feigelson and Babu (2015) have shown that the KS probabilities are wrong if derived from the data set because the theory underlying the test requires independence between the curves under consideration. Hence, an alternative solution is to apply the permutation test, through bootstrapping[1]. The procedure is as follows: it is assumed that the null hypothesis is true. The number of bootstrap samples B is defined and generated under the null hypothesis. A KS statistic is calculated for each of these samples, giving B values of the test statistic and the D value from the complete initial data sample is compared to the distribution of the B number of D value. In addition, a two sided p-value is defined as the proportion of sampled permutations where the absolute difference (KS Statistic) generated by the random samples is greater than the KS statistic observed with the empirical data:

$$p = \frac{\sum D_{random} \geq D_{observed}}{B} \tag{3.15}$$

$$\text{if } p \geq \alpha \text{ the null hypothesis } H_0 \text{ cannot be rejected}$$

3.3.3 Cross Validation and K-Folds

Cross-validation is a statistical method useful towards validating a model. The data is divided into a training set and a validation set; the training set is used to fit the data and the resulting model can predict the validation set. Therefore, the model can be validated by comparing the actual validation set to the predicted one. In addition, the Mean Squared Error can be used to compare the goodness of fit of the model, through comparing the MSE of the training set to the MSE of the validation set, with:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y_i} - Y_i)^2 \tag{3.16}$$

[1] Bootstrapping refers to any test or metric that relies on random sampling replacement, or re-sampling

Therefore, through the MSE, an average of the quality of the predictions across the validation produces a measure of prediction accuracy. The K-fold cross validation is a ramification of the method where the data is divided into K subsets, and each of the subsets constitute in turn the validation set. This method is usually achieved through a for loop that allows selection of the most accurate model.

Chapter 4

Journey Time Distribution Analysis

4.1 Introduction to the Data

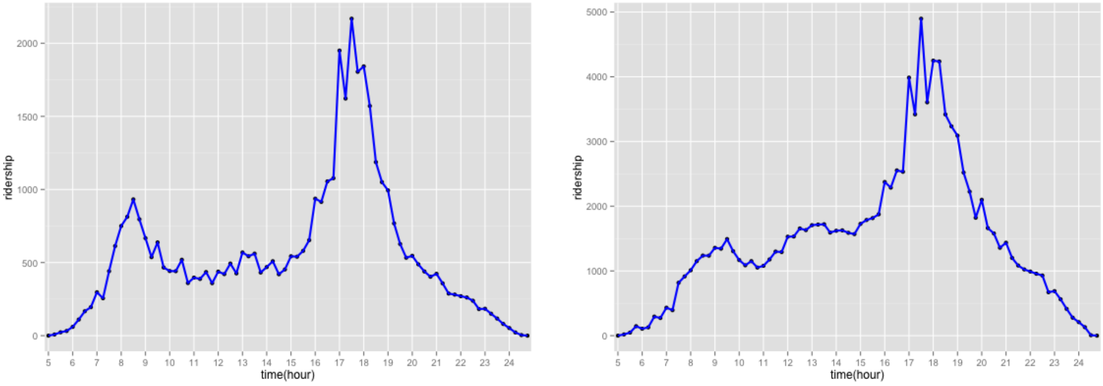
. In this chapter, a descriptive data analysis of some Origin-Destination pairs will be realized. The “big data”, provided by Transport for London, are data files that include all the Oyster card transaction on the tube on a designated day (Up to 20 millions transactions on a week day) between the 22nd of October and the 17th of November. Each recording represents one transaction and includes the information on the tapping-in, tapping-out and other related information to the Oyster card transaction. These transactions represent journey trips, which journey times will be calculated through subtracting entry time from exit time. It must be noted that only the Tuesday, Wednesday and Thursday data were used as the extreme congestion on Monday and Friday, as well as the lower ridership and the closures due to maintenance on weekends biased the data. Therefore, they have been excluded to conduct an analysis of the Underground on “normal” working conditions. Only the data with complete and valid entry and exit times will be considered. In addition, the extremely large data files are then reduced to the transactions pertaining to the Origin-Destination pair of interest. As the data files are to large to be opened with R, the SPSS statistics software is used to filter out the unnecessary recordings.

4.2 Data and Distributions Analysis

4.2.1 Origin-Destination pairs along the same line

First of all, different O-D pairs[1] along the Victoria and Jubilee Line are to be compared, within zone 1 and 2 of the Underground. A map of the tube is provided for reference in Appendix B

[1] It must be noted that when mentioning an Origin-Destination Pair, the data stored is for the two directions of the trip (the two stations are both an origin and a destination)



(A) Ridership between Warren Street and Victoria

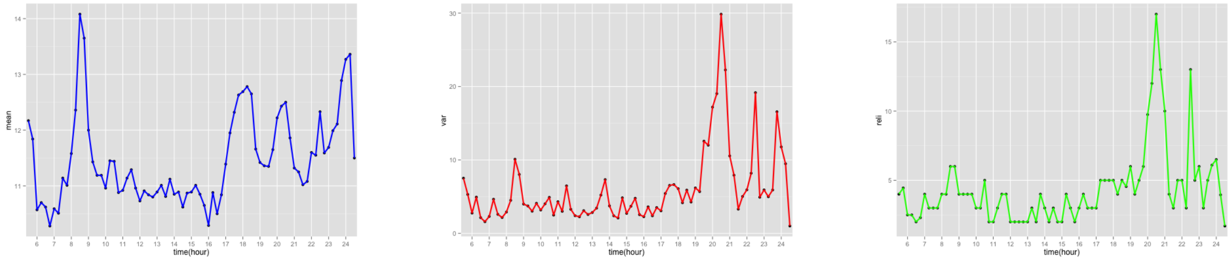
(B) Ridership between Victoria and Oxford Circus

FIGURE 4.1: Ridership Plots of O-D pairs along the Victoria line

The pairs were selected using the research by Chan (2007), pointing out the 50 largest O-D pairs in London Underground. The first pairs of interest are the Victoria-Warren Street pair, with 44,080 observations and the Oxford Circus-Victoria pair, with 117,431 observations, both over the 12 days of data. The ridership plots 4.1 depict the number of passenger between each O-D in 15 minutes time intervals (between 5am and 1am), and are categorised according to the time of entry. In figure A, ridership experiences a slow upward trend in the morning between 7am and 9am ($1,000$ passengers), hits a peak of over $2,100$ passengers in the evening between 4pm and 6pm, and sharply decreases between 6pm and 8pm. Even though the morning peak is lower than the evening peak, this trends corresponds to the peak time as defined by TfL and hence, reveals that it is a commuting trip. Passengers commute in the evening between Warren street, occupied by universities and offices to the more residential area of Victoria. On the other hand, the ridership between Victoria and Oxford Circus shows a different pattern. It increases progressively during the day and rockets up to 4500 passengers between 4pm and 8pm, putting into evidence the shopping characteristic of Oxford Circus, where people mainly go after work.

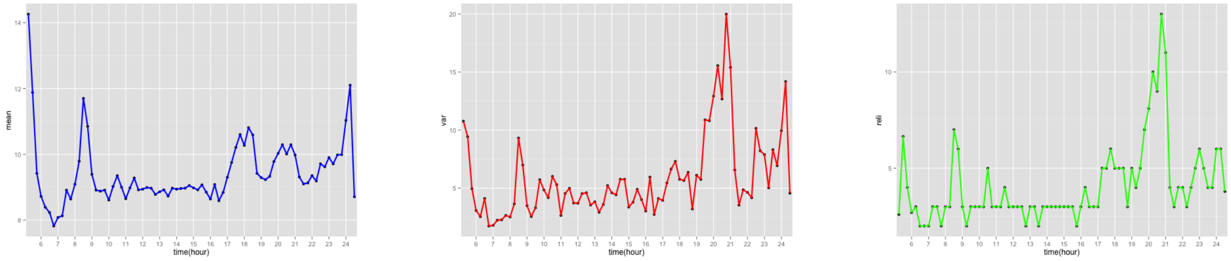
In order to discuss reliability, the mean and variance of journey time are calculated in 15 minutes spans on each O-D pairs. In addition, the reliability factor, as introduced by Chan(2007) and Uniman et al.(2010) in equations 2.1 and 2.2 comprises a significant value: It is a measure of how many minutes are to be It is an essential feature of the reliability as perceived by the passenger.

An interesting fact is the similarity of the fluctuations of the variables between the two O-D pairs, both part of the Victoria line. This suggests that the reliability measure on a segment of the Victoria line, within zone 1, at a specific time could be generalised to the whole line within the zone itself.The larger ridership in the AM peak times appear to strongly impact the mean travel time between Warren Street and Victoria, reaching a peak of 14 minutes. It could be explained by the inability of passengers to board due to congestion.The very slight fluctuations of the variance



(A) Mean Variability (B) Variance Variability (C) Reliability Factor

Figure 4.2: Plot of Parameters with respect to time of Day between Warren Street and Victoria (in minutes)



(A) Mean Variability (B) Variance Variability (C) Reliability Factor

Figure 4.3: Plot of Parameters with respect to time of Day between Victoria and Oxford Circus (in minutes)

and reliability factor during the day below 5 mins (except for the afternoon peak) show that these O-D trips are quite reliable. In the PM peak, the variance and reliability factor of both trip reach their maximum value, due to the very high ridership on both trips that creates congestion and hence longer access time, egress time and platform waiting time. The reliability factor of both trips, except for the PM peak, are quite stable fluctuating more or less around 5 minutes, which means most of the passengers on these trips would only need to add 5 minutes or less to their typical journey time to guarantee on-time arrival.

Having two very busy O-D trips on the Victoria Line, with respectively a noticeable commuting and leisure aspect, the travel time distribution of each O-D pairs is of interest. Due to the complexity of real travel time, empirical nonparametric distributions are fitted on the histograms of journey time. The kernel density function estimator is used to plot the nonparametric distributions; through smoothing over the neighboring data points, and its degree of smoothing, the bandwidth is determined to be $h = 1$. Indeed, because the transaction time is given in full minutes(seconds are not recorded by the AFC system), when using the widely used "Silverman's Rule of thumb" to determine the bandwidth:

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \tag{4.1}$$

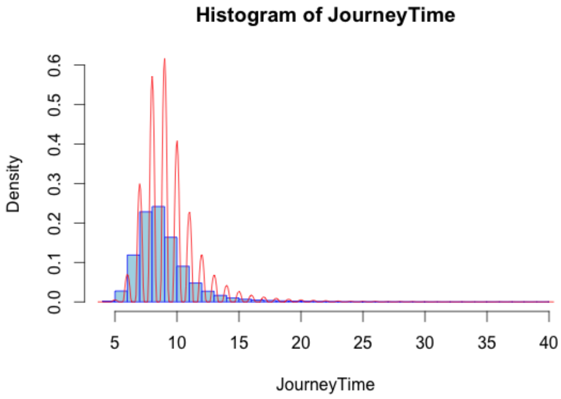
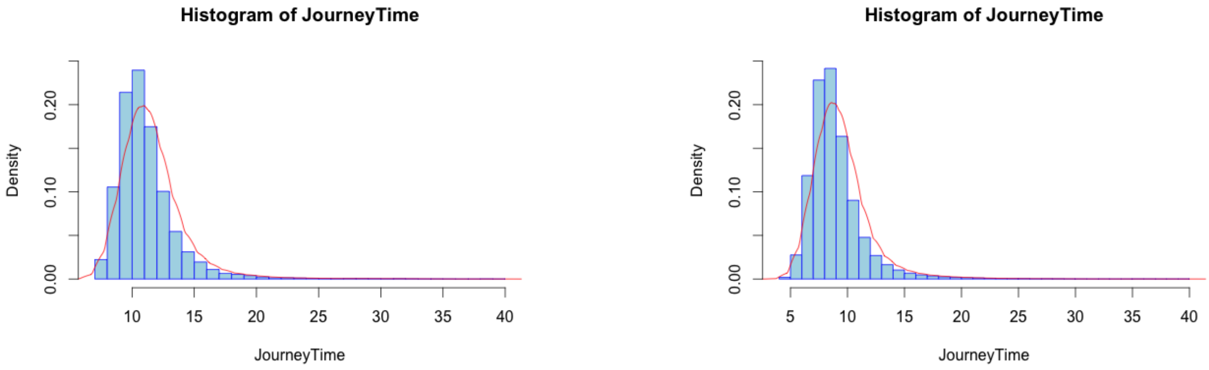


FIGURE 4.4: Kernel Density Estimation using Silverman's rule of thumb

an overestimation occurs, as illustrated in the figure 4.4 for Oxford Circus- Victoria. Therefore, the bandwidth was set to be $h = 1$, providing sensible distributions. The bandwidth is used as the standard deviation of the smoothing kernel. On R, the function *density* is used, which automatically determines the optimal bandwidth of the kernel.



(A) Victoria-Warren Street (B) Victoria-Oxford Circus

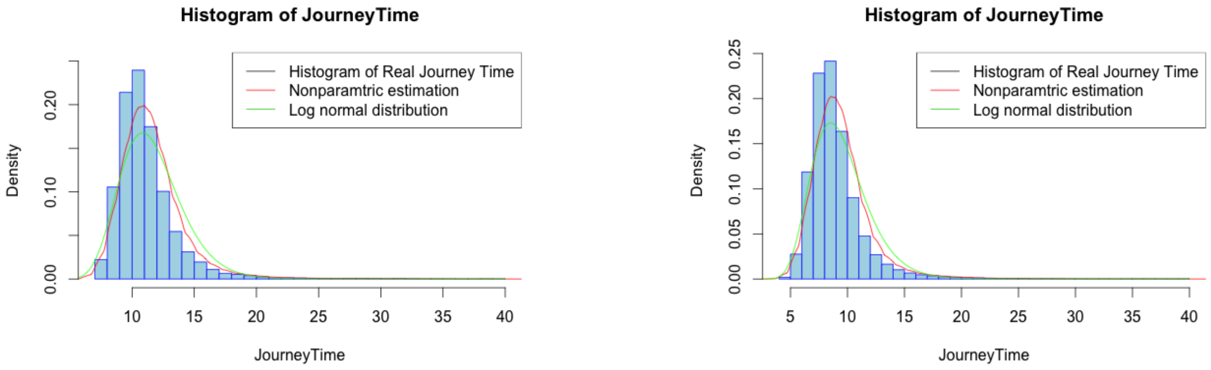
FIGURE 4.5: Non parametric distribution of Journey Time

The journey time non-parametric distribution in Figure 4.5 appears unsymmetrical and positively skewed, with a longer tail on the right. This shape, very similar to the log-normal distribution, is expected: The multiplicative effect of a large number of uncorrelated factor like the access time, egress time and the waiting platform time due to varying headways in the Underground lead to this heavy positive skewness. The congestion degree also constitutes a factor to be taken into account. On the other hand, the longer tail can be explained by cases of extreme congestion where travel time is longer, the diversity of passengers (for example, seniors would travel more slowly), irregularities on the lines that might be held at red lights etc .... The spectrum of passengers is large and includes everyone from commuters who are in a rush, to tourists who take their time or might get lost. In addition, the compactness of the travel time distribution is an indicator of the quality of the service between the O-D pairs. By comparing plot (A) and (B), it can be inferred that the Oxford Circus-Victoria pair provides a more reliable service than the Victoria-Warren Street pair due to its distribution being more compact. However, the reliability factor of both pairs is of 5 minutes, suggesting a constant quality of service along the Victoria Line, within zone 1 and 2. This is confirmed by Table 4.1, which reveals that the reliability factor for O-D pairs along the Victoria line oscillates between 4 and 6 minutes, provided the mean journey time does not exceed 20 minutes.

O-D pairs	**Mean Travel Time (min)**	**Reliability Factor (min)**
Oxford Circus-Victoria	9.46	5
Vauxhall-Oxford Circus	12.29	4
Victoria-Warren Street	11.62	5
Brixton-Oxford Circus	17.56	6
King's Cross-Green Park	12.80	6
Vauxhall-Warren Street	14.96	6

Table 4.1: Reliability metric of Origin-Destination pairs along the Victoria Line

Due to their familiar shape, the histograms are fitted with the log-normal distribution using the mean and variance of the empirical data, as shown in figure 4.6 However, it can be noted that the mode of journey time in the histograms are better fitted when using nonparametric than when using log-normal. The goodness of fit and comparison of two distributions will be elaborated in the next section.

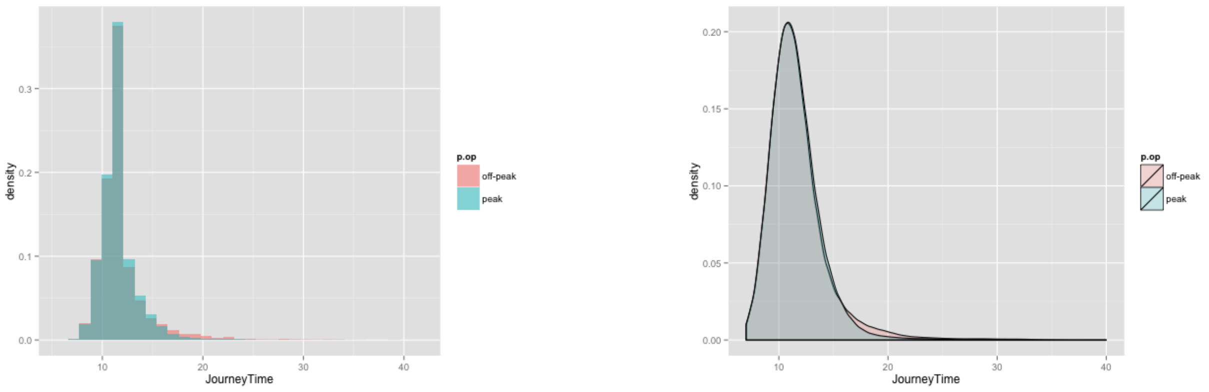


(A) Victoria-Warren Street

(B) Oxford Circus-Victoria

Figure 4.6: Fitted Lognormal Distribution

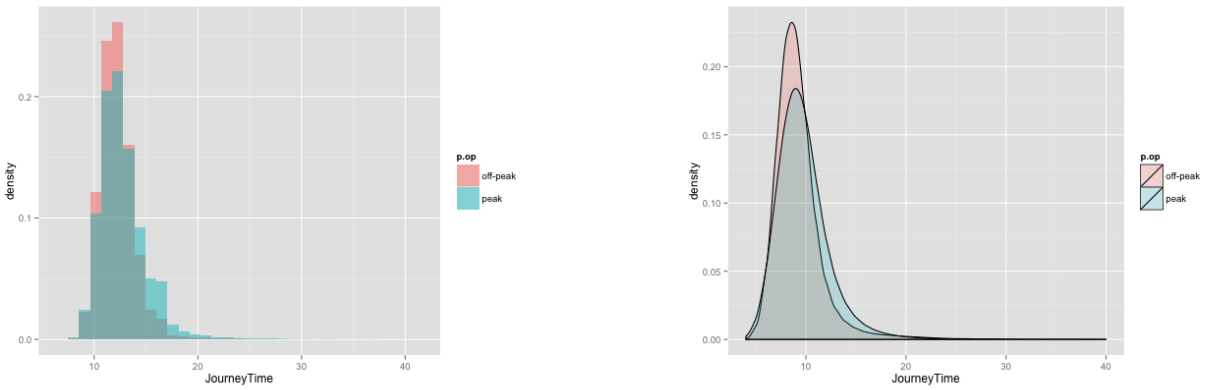
To check whether the peak and non-peak travelling time (as defined by TfL) significantly affect the journey time distributions, the travel times are classified according to this criterion and are plotted as shown in figure 4.7 and 4.8. The figures reveal that on the Victoria line, the travelling times are very similar in peak and off-peak, thus it is not necessary to segregate the data for its evaluation.



(A) Histogram

(B) Density

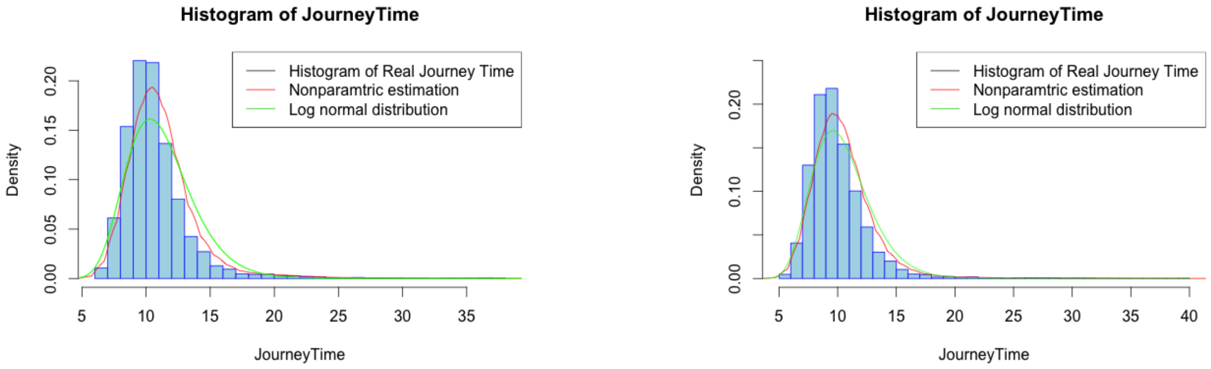
Figure 4.7: Journey Time Histograms and distributions in terms of peak and off-peak departure time between Victoria and Warren Street



(A) Histogram

(B) Density

Figure 4.8: Journey Time Histograms and distributions in terms of peak and off-peak departure time between Victoria and Oxford Circus



(A) Baker Street-Green Park

(B) Green Park-Waterloo

Figure 4.9: Journey Time Histograms and fitted parametric and nonparametric distributions

The above analysis has also been conducted on O-D pairs along the Jubilee line. The resulting graphs are not included to avoid a repetitive analysis. As an indicator, only the journey time histograms of two O-D pairs along the Jubilee Line are plotted: The Baker Street- Green Park pair and the Green-Park Waterloo pair (Figure 4.9). The compactness of both distributions are very similar, suggesting a stable service on the Jubilee line between these OD pairs, with reliability factors of respectively 4 and 5 minutes.

4.2.2 Distribution Model Fitting

In figure 4.6, it seems that the nonparametric distribution is a better fit for the journey time data on single O-D pairs. However, using a parametric log-normal distribution would be an advantage in order to model distributions pairs. The goodness of fit of the log-normal distribution will be tested in order to assess if it is suitable. Permutation testing and k-folds (cross validation) are to be used. The hypothesis to be tested are:
$\mathbf{H_0}$: The Journey Time random variable follows the log-normal distribution
$\mathbf{H_1}$: The random variable has a different distribution, typically nonparametric

For the Oxford Circus- Victoira pair, the assumed parameters of the log-normal are obtained through the transformation of the mean and the standard deviation of the empirical data (equations 3.5), as calculated in R:

$$\mu_X = 9.46min \qquad \sigma_X = 2.51min$$
$$\phi = \sqrt{\sigma_X^2 + \mu_X^2} = 9.79min$$
$$\mu_Y = \ln \frac{\mu_X^2}{\phi} = 2.213min$$
$$\sigma_Y = \sqrt{\ln \frac{\phi^2}{\mu_X^2}} = 0.261min$$

The number of bootstrap samples $B = 10,000$ is defined and data sets of $n = 100$, conform to this log-normal distribution, are randomly generated using R and the given parameters. Then for each one of the bootstrap samples, the D test statistic of the KS (under the Null Hypothesis H_0) is calculated and its distribution is as follow.

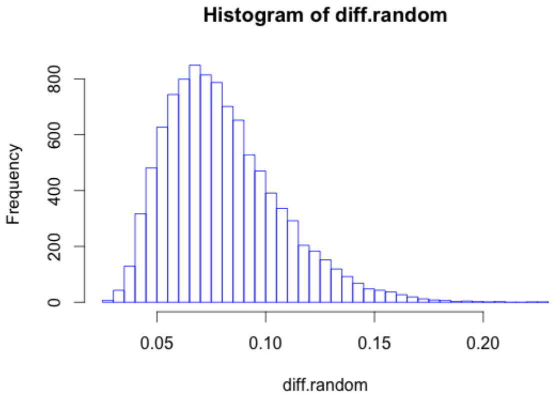


FIGURE 4.10: Histogram of D statistic of the KS test applied to the 10,000 samples in permutation testing

The observed D-value of the empirical data is calculated using equation 3.13

$$D_{observed} = \max\left|\text{ecdf}(JourneyTime) - \text{lognormal}(JourneyTime, \mu_Y, \sigma_Y)\right| = 0.148$$

This value lies within the right tail of the histogram, but is an accepted D-value. $p = 0.0184 \geq \alpha = 0.01$, showing that the Null Hypothesis cannot be rejected.

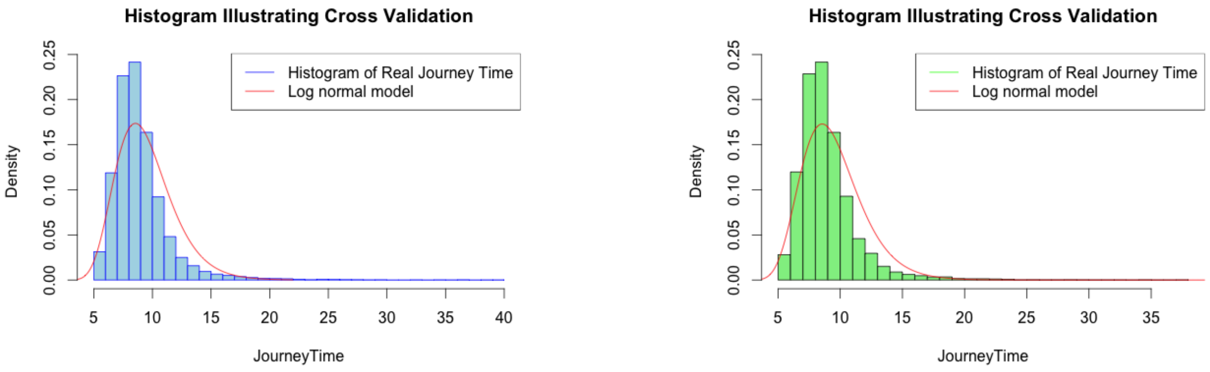
In addition, in order to strengthen the results, 10-folds cross validation is achieved on the data. The $117,430$ observation data is randomly split into 10 sets and using a for loop that matches all possible combinations, each of the set is successively assigned as the validation set while the 9 other sets are used as training sets to estimate a model and its parameters.
The results of one of the combinations, which appears as the optimal one, are presented. After having estimated the parameters of the log-normal distribution, which are equal to the parameters calculated for the whole set of data, the MSE of the model is calculated using only the 9 training sets, using equation 3.16.

$$\mu_t = \mu_{\text{training set}} = 2.213 \text{ min and } \sigma_t = \sigma_{\text{training set}} = 0.260 \text{ min}$$
$$MSE_{\text{training set}} = 9.750 \text{ min}$$

The estimated model is then fitted on the validation set, as presented in figure 4.11, and the simulated log-normal distribution values, randomly generated with R, are compared with the empirical journey time values:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(simulated_{\mu_t,\sigma_t} - observed_{\text{Journey Time}})^2 \text{ with } n = \text{length of validation set}$$

$$MSE_{\text{validation set}} = 9.768min$$

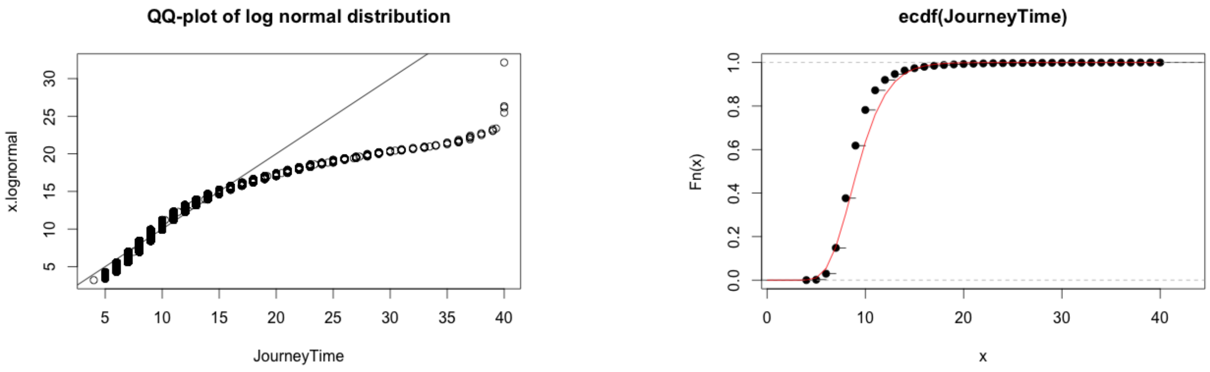


(A) Validation set fitted with estimated model (B) Different validation set model

FIGURE 4.11: Histogram to illustrate cross validation of parameters of log-normal model

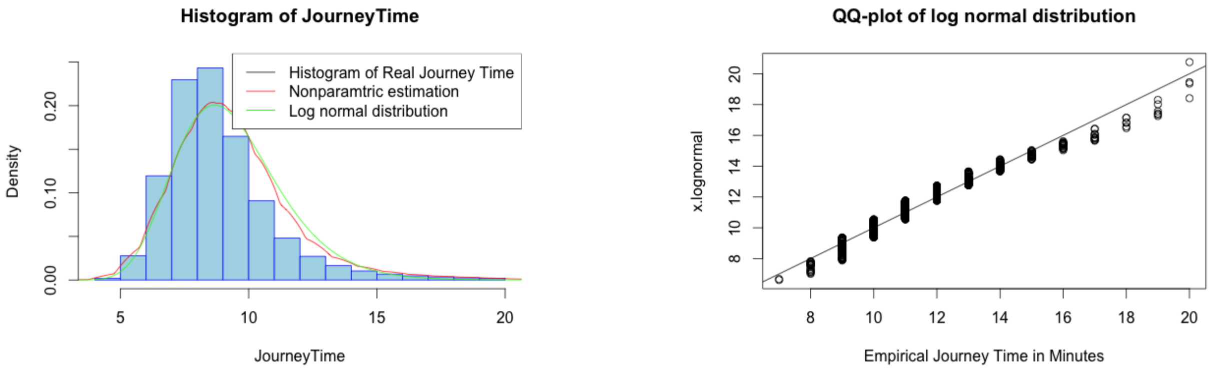
The relative error between the mean squared error of the validation set and the training set is close to zero (0.04%), thus proving that the log-normal distribution is suitable for this O-D pair along the Victoria Line.

The empirical cumulative distribution function reveals that the empirical data corresponds to the hypothetical log-normal.The QQ plots however reveals that the model is not perfectly fitted. For up to 18 minutes, the empirical data matches the theoretical log-normal distribution. However, it is heavily tailed, revealing that the empirical data contains some extreme cases that the log-normal does not account for. These are the travellers that either need a longer access and egress time, or face an issue during their trip, causing an unusually long journey time. Indeed, when looking at the journey time distribution in figure 4.6A, above 18 minutes, both the nonparametric and log-normal distribution tend towards zero and do not account for these cases. When looking at the data more closely, 99% of the passenger complete the trip in less than 20 minutes(only 482 cases out of 44080 need longer than that treshold). Therefore, the longer journey time are excluded from



(A) Quantile-Quantile Plot (B) Empirical cumulative distribution function

FIGURE 4.12: Probability plots of the log-normal distribution model between Victoria and Oxford Circus



(A) Histogram and Fitted distributions (B) QQ-Plot

Figure 4.13: Journey Time distribution fit excluding outliers

the data and as seen in figure 4.13, the log-normal distribution matches the nonparametric and the empirical data. As a conclusion, the log-normal appears as a relevant model for the Victoria and Oxford Circus pair.

Similarly, the process is repeated for others Origin-Destination pairs on the same line. The results are presented in the following table 4.2:

Permutation Test Results				
Line	Origin-Destination Pair	p-value	$D_{observed}$	$D_{random,max}$
Victoria	Oxford Circus-Victoria	0.0184	0.148	0.229
Victoria	Vauxhall-Oxford Circus	0.0122	0.156	0.229
Victoria	Warren Street-Victoria	0.0124	0.155	0.229
Victoria	Brixton-Oxford Circus	0.0215	0.145	0.229
Victoria/Picadilly	King's Cross-Green Park	0.0615	0.126	0.229
Jubilee	Green Park-Waterloo	0.0535	0.129	0.229
Jubilee	Baker Street-Green Park	0.0189	0.147	0.229
Jubilee	London Bridge- Bond Street	0.0136	0.157	0.229
Jubilee	Westminster-Baker Street	0.0151	0.154	0.229

Table 4.2: Results of Permutation Test on OD pairs with no interchange

From the table 4.2, the p value for all the Origin-Destination pairs presented reveals that the Null Hypothesis H_0 (The journey time follows the log-normal distribution) cannot be rejected with a level of significance of 99%. The 10-folds cross-validation testing confirm the results.

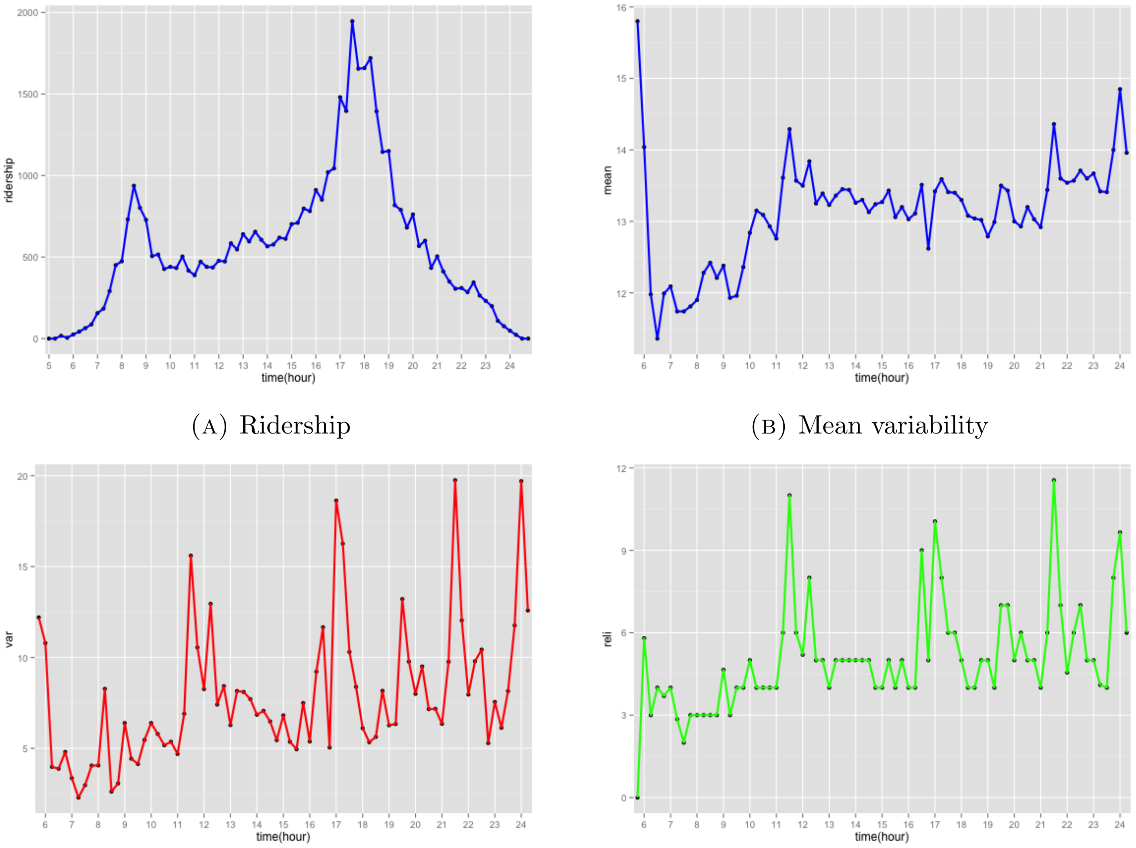
It can be concluded than for journeys between Origins and Destinations on the same line, the log-normal distribution was found to properly model travel time. Indeed, although the nonparametric distribution better captures the density of the mode of the data, as seen in the journey time histogram 4.6, the log-distribution has been proven to be acceptable. Moreover, the log-normal being a parametric model, it is interpretable and meaningful in terms of data analysis and can be used for predictions and forecasting purposes, and therefore preferable to use than the nonparametric. Finally, as seen in figure 4.13, when the 1% of extremely slow travellers are excluded from the data, the nonparametric and log-normal models are equivalent in fitting the data, hence reinforcing the preponderance of the use of log-normal.

4.2.3 O-D pair including an interchange

It has been concluded that log-normal distribution was found suitable for O-D pairs on the same line. Are the results reproducible for OD pairs when an interchange is completed during the journey or when different route choices are available? In order to verify that, the Waterloo-Oxford Circus pair, with $45,460$ observations over the 12 days, is tested (considering the different entry-exit station codes of Waterloo Station), as it comprises trips on the Victoria and Jubilee line, which were previously studied. Along this pair, the first segment of the trip is made on the Victoria line between Oxford Circus and Green Park (one-stop) and the second segment is made on the Jubilee line between Green Park and Waterloo (two stops); an interchange has to be made at Green Park station. Another option is to travel between this O-D pair using the Bakerloo line (four stops). The ridership, mean, variance and reliability factor are calculated in 15 minutes interval and plotted in Figure 4.14.

The ridership follows a "commuting" trend: An AM peak and a PM peak stand out, but the sharper increase to 1700 passengers per 15 minutes in the evening peak emphasises again the shopping destination that is Oxford Circus. Moreover, the important fluctuations of the mean, variance and reliability during the day can be pointed out: The variance varies between 3 minutes in the morning to up to 15 and 20 in the evening. Similarly, the reliability factor fluctuates around 3 minutes in the morning peak but reaches 9 minutes at several times during the day. When comparing this plots to the plots of O-D pairs on the same line, where the statistical vproperties are quite stable and only sharply increase at peak times, the extreme fluctuations of the statistical properties on this pair stand out. It can be concluded that the mean, variance and reliability factor of a journey time on this pair with an interchange and different possible paths are less reliable for journey time prediction than they are for simple O-D pairs.

As a consequence of the larger values of variance, the distribution is more disperse. Indeed, the possible factors affecting journey time are highly increased, such as the different path choices leading to the use of different lines (different speeds, headways and stops), congestion in the interchange stations, getting lost or going the wrong way at the interchange stations, explaining this continuous great fluctuation during the day.The mean journey time of the trip is a less reliable reference as the reliability factor fluctuates to high values.



(A) Ridership

(B) Mean variability

(C) Variance Variability

(D) Reliability factor Variability

FIGURE 4.14: Evolution of main statistics variable between Waterloo and Oxford Circus with respect to time of the day

The journey time histogram, fitted with a non-parametric and log-normal distribution and plotted in Figure 4.15, proves that the distribution is indeed more disperse. By comparing the nonparametric distribution of this pair with interchange to the nonparametric distribution of the O-D pairs along the same line, it can be pointed out that this distribution is less compact and hence has a larger standard deviation, revealing a more disperse data set. In other words, the service quality between two O-D pairs including an interchange is less reliable than between an O-D pair along the same line. Again, this is a direct repercussion of the many new factors that affect travel time when an interchange is to be made. Moreover, the time spent at the intermediate station represents a significant chunk of the total time and is independant of the service on the lines. As a consequence, the compactness of the distribution cannot be a direct indicator of the service on the lines when an interchange is completed. To check whether the log-normal distribution is suitable, permutation testing rates the Null Hypothesis.

The observed D-value of the empirical data is calculated using equation 3.13

$$D_{observed} = \max|\text{ecdf}(JourneyTime) - \text{lognormal}(JourneyTime, \mu_Y, \sigma_Y)| = 0.165$$

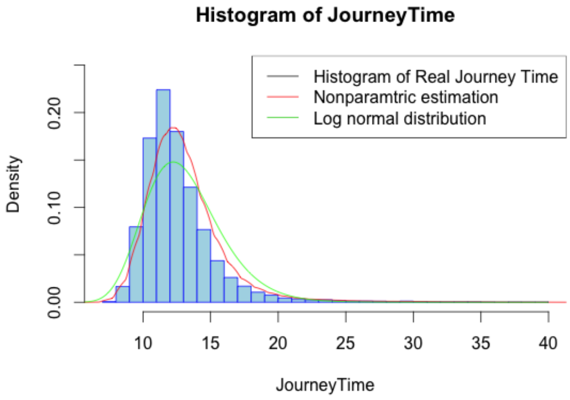


FIGURE 4.15: Histogram of Journey Time distribution between Oxford Circus and Waterloo with fitted distributions

This value lies within the extreme right tail of the D-values histogram 4.10. Similarly, the p value is calculated using equation 3.3.2 the $p = 0.0078 \leq \alpha = 0.01$, showing that the Null Hypothesis can be rejected. As a consequence, it can be inferred that, for an O-D pair which entails multiple optimal path, many new factors influence the journey time. The most basic of this factors is the route choice, that would heavily change the resulting journey time: Indeed, the different lines chosen might have different speeds, different headways of trains, longer distance to trace (and thus varying number of stops). The intermediate station itself could affect journey time through its congestion, the different transfer times for different interchange stations, varying waiting time on the paltform (varying headways) or the user going in the wrong direction. The combination of all these factors lead to the rejection of the log-normal distribution. In this case, the nonparametric is deemed more suitable to model the journey time, as the variance is larger and the distribution is more disperse.

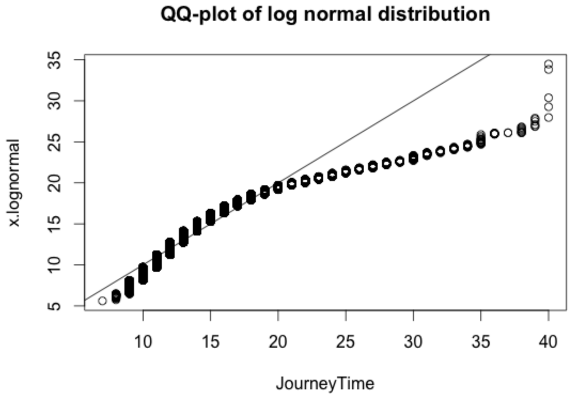


FIGURE 4.16: Quantile-Quantile plot of the Oxford Circus- Waterloo journey against the fitted log-distribution

The Quantile-Quantile plot 4.16 shown above supports the rejection of the log-normal distribution. It can be seen that it does not fit the data on both tails as the data is heavily tailed. Moreover, for this case, the trips above the 20 minutes limit cannot be filtered out as they represent a significant 5% of the data.

Other O-D pairs including one interchange on at least one of their optimal path are treated and tests are conducted to check whether the previous conclusion applies. Table 4.3 summarises the results.

Permutation Test Results				
Origin-Destination Pair	p-value	$D_{observed}$	$D_{random,max}$	number of obs.
Oxford Circus- Waterloo	0.0078	0.165	0.229	44,360
Baker Street-Victoria	0.0404	0.133	0.229	12,301
Bank/Monument-Green Park	0.0095	0.160	0.229	9,837
Green Park- Embankment	0.0785	0.155	0.121	975

Table 4.3: Results of Permutation Test on OD pairs with interchange or different route choices

From table 4.3, it can be seen that the conclusion drawn earlier cannot be generalised. Indeed, for both the Baker Street-Victoria pair and the Green Park-Embankment, which includes one interchange on its optimal path, the null hypothesis cannot be rejected and the log-normal model is still acceptable, as $p \geq \alpha = 0.01$. However, the Bank/Monument-Green Park pair, which includes one interchange on both of its optimal path[2], cannot be modeled using the log-normal distribution as the Null Hypothesis has been rejected. Again, the different path choices, as well as the interchange include in the journey time additional factors like walking time at the potential interchange station, headways of different lines at departure and intermediate stations, choice of path, which lead to a distribution that is best fitted when nonparametric.

As a conclusion, for pairs with interchanges along their route, no general conclusion can be drawn. It appears that when the pair has one interchange on its **only** optimal path, the log-normal model is suitable to model journey time. The new factors pertaining to journey time are similar for all travellers between this pair, as it is the same two lines utilised and the same interchange station subject to congestion. Consequently, the transfer time at the interchange station, the speed of lines and headways of trains constant on this path, resulting in the ability to characterise the distribution with parameters. On the other hand, when the pair has as interchange but at least **two** optimal paths, the number of factors inherent to journey times increases while these factors become too random for a parametric distribution to be fitted. As a consequence, reliability of the journey becomes harder to evaluate and cannot be parametrically characterised. It can be implied that in this case, only nonparametric distribution can capture the dispersion of the distribution. However, these conclusions must be treated with caution as the number of O-D cases tested is limited and the sample size for some O-D pairs over the 12 days of data is quite modest.

[2] One path along Victoria and District/Circle Lines and another path along Picadilly and Central Lines

Chapter 5

Regression Analysis of Journey Time

The users of the London Underground, and especially commuters, rely on Transport for London to provide a service with acceptable and predicted journey times, which are complied with in theory. Indeed, TfL guarantees a certain quality of service, with reasonable headways and fast transit and provides transport information through journey planners, which are quite accurate in predicting **mean** journey time. However, many factors that cannot be completely controlled by TfL impact the journey time of travellers. As they cannot bear unexpected delays journey times, reliability constitutes one of the main concerns of the users. Therefore, the variance and reliability factor [1] of the journey time, which represent the two main evaluation criteria of the service by the users, are to be investigated to highlight any relationship between these variables. Regression models are built to evaluate, at a line level, reliability as a function of time of day and to assess the relationship it sustains with the standard deviation of any journey on that line. The main purposes are to determine how reliable an O-D trip made at a specific time of day is and to assess whether reliability can be realistically characterised through regression models. As most of the O-D pairs considered were part of the Victoria and Jubilee line, the reliability of these two lines is to be analysed.

5.1 Regression on the Victoria Line

In order to investigate the reliability factor on the Victoria line, the OD pairs with no interchange along this line, examined in Chapter 4 are used. They are completed by the data of 6 other O-D pairs along this line and within zone 1 and 2, listed in Table 5.1.

[1] The smaller the reliability factor is, the more reliable the service. It is defined as the amount of time the passenger must add to his usual travel time to guarantee, with a 95% level of significance that he will arrive on time

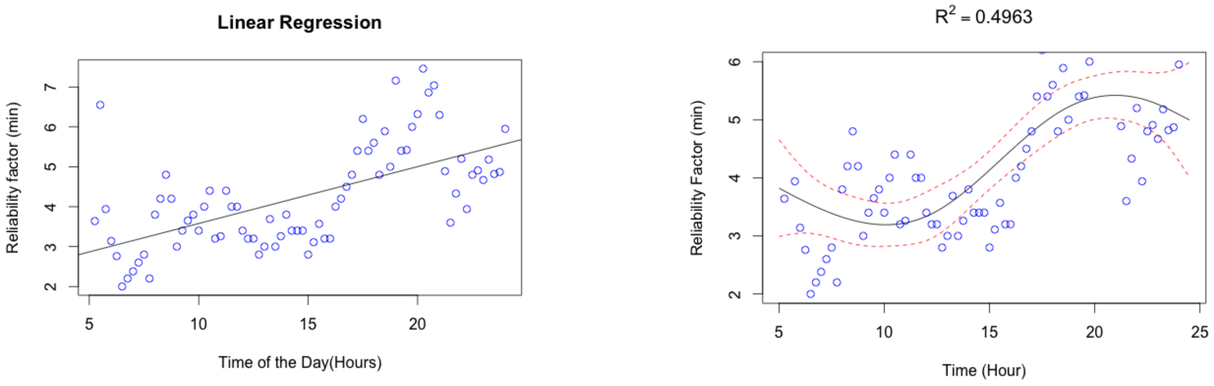
O-D pairs from Chapter 4	Additional O-D pairs
Oxford Circus-Victoria	Finsbury Park-Victoria
Vauxhall-Oxford Circus	Green Park-Oxford Circus
Victoria-Warren Street	Islington-Euston
Brixton-Oxford Circus	Oxford Circus-King's Cross
King's Cross-Green Park	Vauxhall-Euston
Vauxhall-Warren Street	

Table 5.1: Origin-Destination pairs along the Victoria Line used for analysis

5.1.1 Regression of reliability with time

The first aspect of interest is whether the time of day influences the reliability of a line and hence determining whether travelling at a specific time of day on the Victoria line would significantly impact the usual journey time of a user. As the reliability along the Victoria Line is to be investigated, the O-D pairs presented in Table 5.1 are treated to extract the reliability factor vector, which incorporates the reliability factor of the 12 days of data, over 15 minutes intervals. The mean reliability factor over all the 0-D pairs on the line are then obtained and several regression models are fitted to the data. This section summarises the results.

To start with, linear and smoothing spline regression models are used, as shown in Figure 5.1. R relies respectively on the functions *glm* and *bigspline* to perform these regressions. The cubic spline regression depicted uses a total of 4 knots with two interior knots and one at each extreme.



(A) Linear Regression

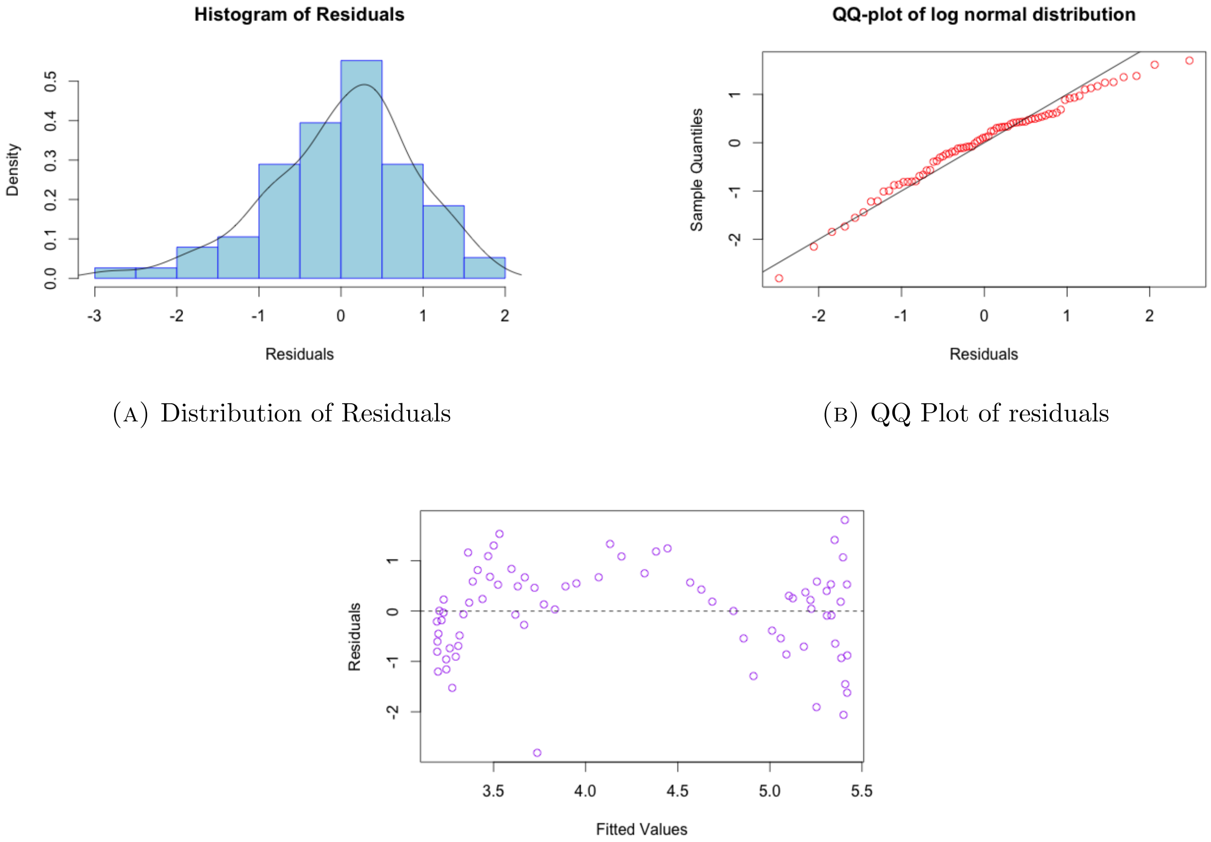
(B) Smoothing Spline Regression

Figure 5.1: Reliability factor according to time of day

Moreover, the red dotted line delimit the 95% Bayesian probability interval of the data. Also called a credible interval, they incorporate information from the prior distribution into the estimate, allowing more flexibility towards the estimation of parameters. Indeed, in Bayesian inference, a probability is a measure of the degree of certainty about values, meaning that parameters are random and data is fixed.

It is clear from the regression figures that the parametric linear regression does not fit the data properly and the cubic spline regression appears as a better fit. In order to investigate this semi-parametric model, the histogram of residuals is plotted, as well as the Quantile-Quantile plot to assess normality of the errors (Figure 5.2). The residual errors appear to have a Gaussian distribution, while the QQ plots confirms normality, although it suggests a skewed distribution. In plot C, the residuals disperse randomly around the $y = 0$ line suggesting constant variance, despite some outliers standing out of the pattern. The figures present evidence that the cubic smoothing spline method is suitable, and imply that no obvious problem can be pointed out, but some improvements could be made.

The rejection of the linearity model is confirmed by diagnostic plots that contradict the initial assumption of linearity[2]: The residual errors are not normally distributed and their variance is not constant.



(A) Distribution of Residuals

(B) QQ Plot of residuals

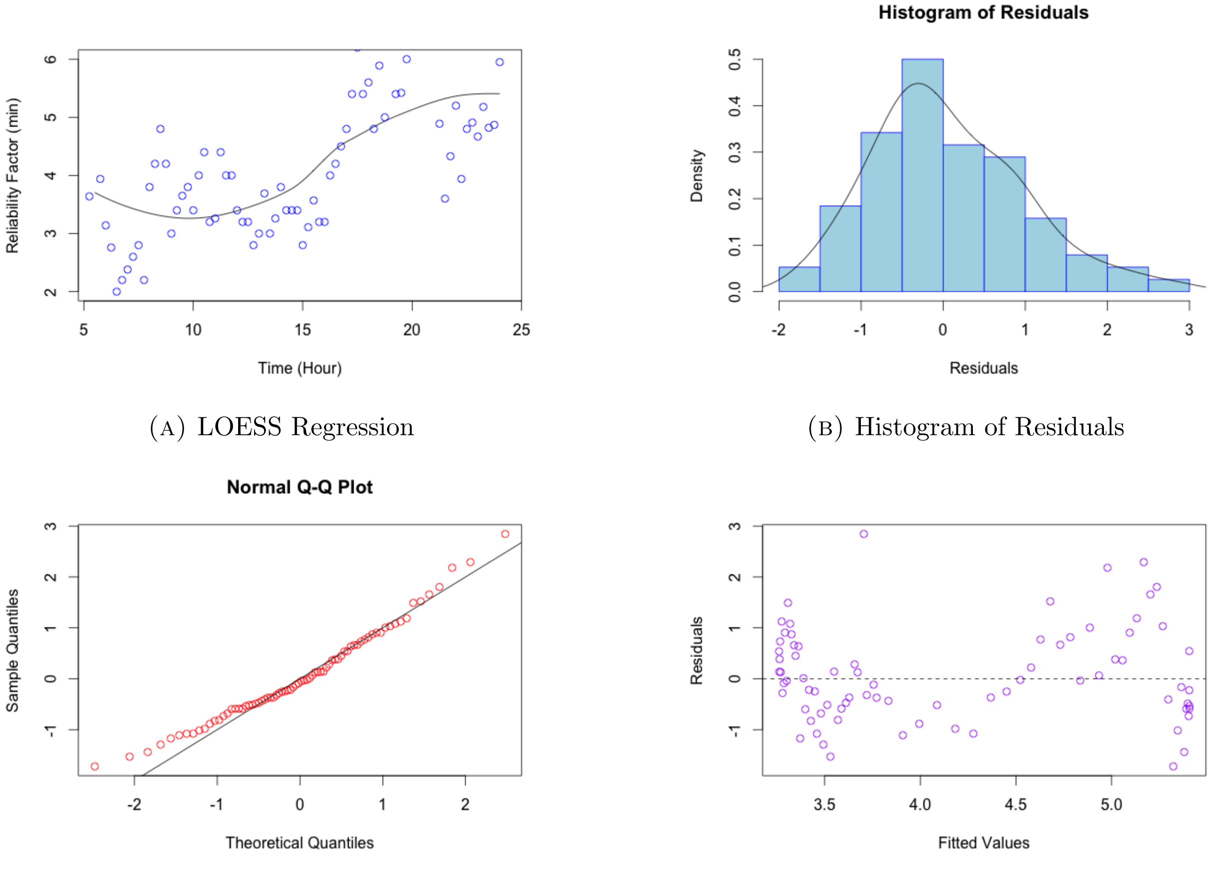
(C) Residuals Versus Fitted Values

Figure 5.2: Diagnostic Plots to assess Smoothing Spline regression

As the semiparametric method was found suitable but prone to improvements, a type of non-parametric regression is investigated, the Locally Scatterplot Smoothing method or LOESS: The smoothed value at each point is defined using neighboring points, within 0.5 minute span and a

[2] The plots are included in Appendix B

quadratic local fitting method. The span was determined by trial and error, as it appeared that with a lower span, the regression model over fitted the data. Figure 5.3 discloses the results. The two-plotted non-parametric model appear very similar when plotted on the scatterplot of the data. However, the diagnostic plots put into evidence a heavily tailed distribution of residual errors as well as a non-constant variance and numerous outliers on the fitted versus residual plot. The smoothing spline is deemed preferable to model the evolution of reliability during the day on the Victoria Line.



(A) LOESS Regression

(B) Histogram of Residuals

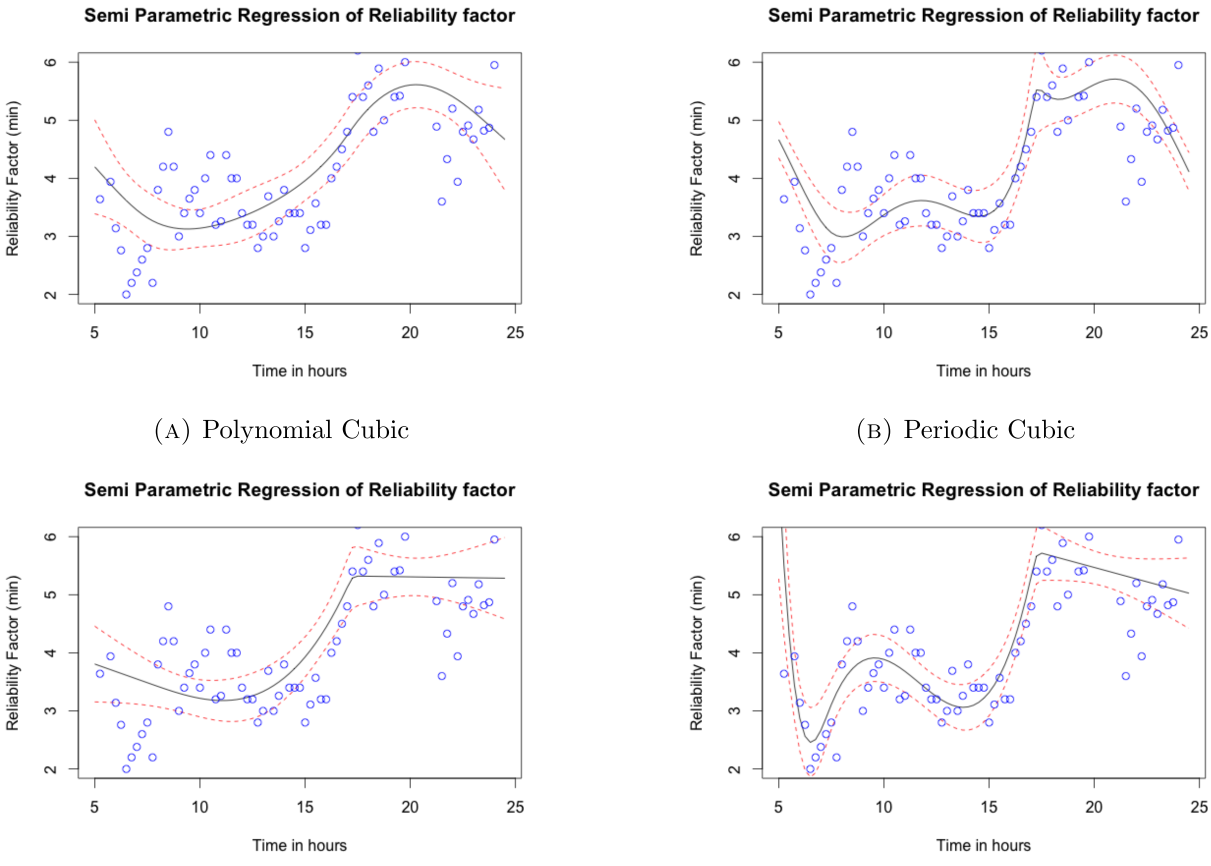
(C) QQ Plot

(D) Residuals Versus Fitted Values

Figure 5.3: LOESS Regression of reliability factor

From the numerous regression plots, the main trait that protrudes is the decrease of the reliability of the service along the day, as the reliability factor increases. The reliability factor is quite stable in the peak hours of the morning, around 3 minutes, and progressively increases to reach 5 minutes at 8pm. This could be the consequence of longer headways of the successive trains after that time, while the progressive increase in the afternoon is due to the PM peak time on the Underground that creates important congestion and thus increases the access, egress and often, the platform-waiting time due to being unable to board the first train that arrives to the platform.

Finally, as it was demonstrated that the semiparametric method could be improved, different types of semiparametric regression models are fitted to the reliability factor data frame using the *bigssp* function in the *bigspline* package in R. This tool also fits smoothing splines to the data but includes varying parametric effects. Figure 5.4 exhibits the different types of semiparametric regression: Plot (A), which is polynomial cubic is very similar to the previous cubic spline regression, but captures better the decrease of reliability factor after 8pm. Plot (B), (C) and (D) capture more accurately the evolution of the reliability factor on the Victoria line during the whole day but (C) and (D)are however hardly interpretable and hence unsuitable as a generic regression model for reliability. The periodic cubic regression of plot (B) on the other hand, though its periodicity, is interpretable and seems like a suitable option.



(A) Polynomial Cubic

(B) Periodic Cubic

(C) Cubic Thin Plate

(D) Cubic Type 2

Figure 5.4: Different Types of Smoothing Splines with Parametric Effect

Different types of regression models were fitted to the data. A boxplot 5.5 of the residual errors is used to compare the several selected models. The linear regression model boxplot is comparatively tall, revealing a more extended residual error. However, the three remaining models have similar residuals errors, which leads to the selection of a semiparametric method, the polynomial cubic smoothing spline. Indeed, through its parametric dimension, this model is easier to interpret and potentially predict.

As a conclusion, the variability of reliability with time of day can be well described with a cubic spline regression model, with higher reliability in the morning up to 3pm (lower reliability factor of 3 minutes), and a progressive increase in the reliability factor (5 minutes) between 3pm and 8pm corresponding to peak time on the tube. Finally, after 8pm, the Victoria line regains some reliability but the larger headways of trains do not allow a complete drop back to 3 minutes.

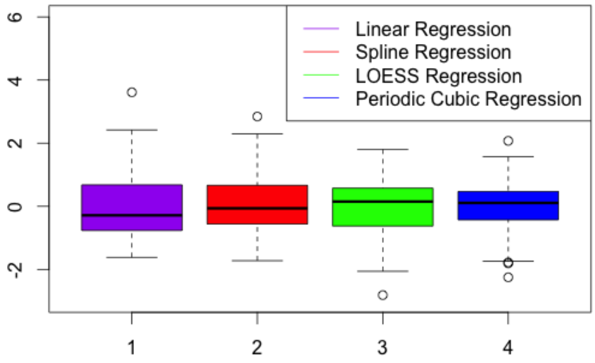


Figure 5.5: Boxplot to compare the residual error of different regressions

5.1.2 Reliability with respect to standard deviation

As already mentioned, the user of the Underground are not inclined to unexpectedly long journey. Indeed, the mean travel time does not represent a meaningful evaluation criteria as it is taken into account when planning a trip. Therefore, the main criterion to evaluate how reliable the service is the reliability factor, which states how much time the user must add to his usual travelling time to guarantee he will arrive on time. The relationship between this factor and the standard deviation of any journey along the Victoria line is investigated.

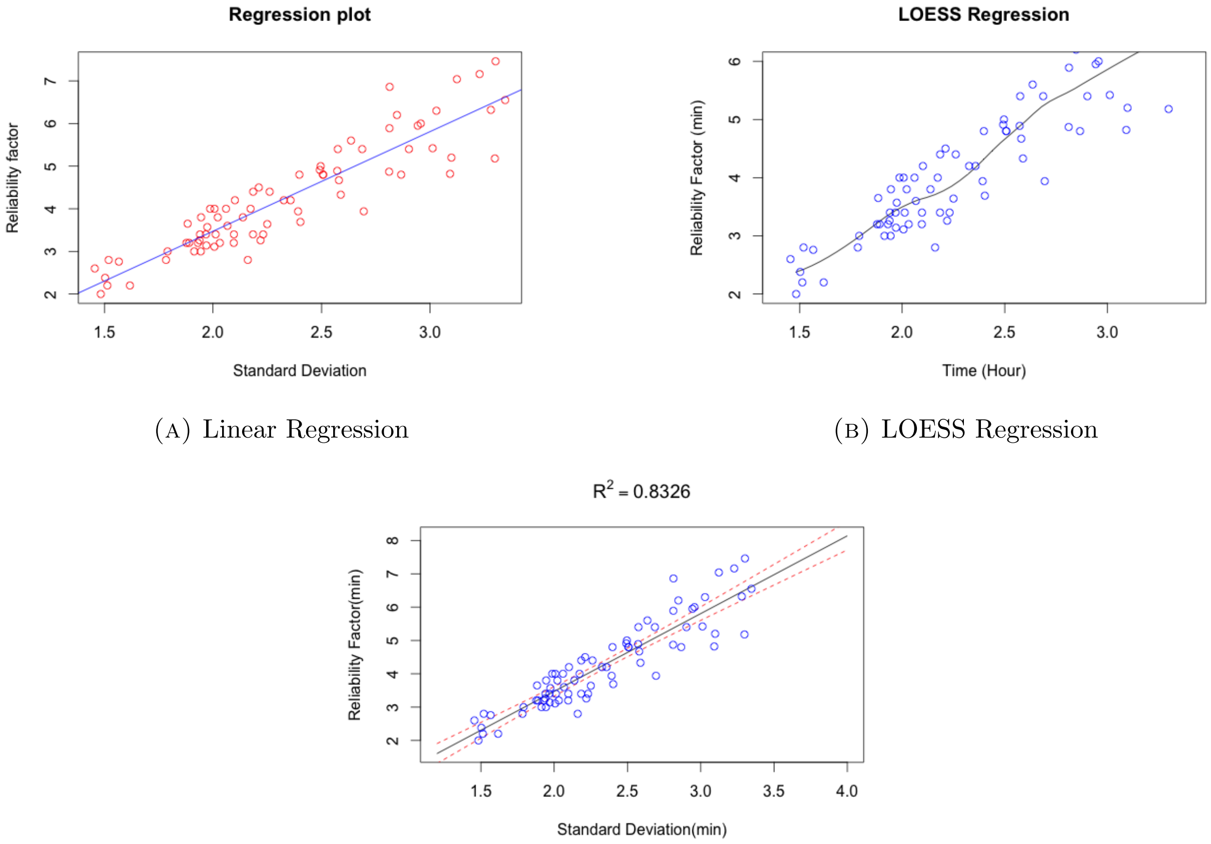
When estimating the linear, LOESS and Spline Regression of reliability factor as a function of standard deviation in R, the plots reveal a surprising fact: All regression curves are almost linear, hence proving that linear regression is acceptable in this case. This results is reinforced by the Mean Squared Error calculated for each case. The MSE is as its lowest when using Spline Regression, but the MSE_linear is within less than 10% of the MSE_spline, thus revealing that the relationship is almost linear.

$$\text{Mean Squared Error} = MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y_i} - Y_i)^2$$

$$MSE_{\textbf{linear}} = 0.519 \text{ min}, \; MSE_{\textbf{loess}} = 0.512 \text{ min and } MSE_{\textbf{spline}} = 0.473 \text{ min}$$

The squared roughness $R^2 = 0.8326$ of the spline regression, which shape appears almost perefctly linear, close to 1 remains a week indicator that the model fits the data well. The diagnostic plots shown in figure 5.7 reveal that the assumption of the linear method, as detailed in Chapter 3 are respected. The residuals are normally distributed as seen on the Quantile-Quantile plot and the residuals are dispersed randomly around the $y = 0$ line, suggesting that assuming the relationship is reasonable. They form a "horizontal band" around the line, suggesting that the variance of the error terms is constant. This suggests that the assumption that the relationship is linear is reasonable.Even though some residuals stand out from the pattern, suggesting the data includes some outliers.

Consequently, the reliability factor of any 0-D trip on the Victoria Line is linearly dependent on its standard deviation. This relationship highlights the fact that the standard deviation has a direct outcome on the reliability of travelling on the Victoria line an its significance should not be overlooked.

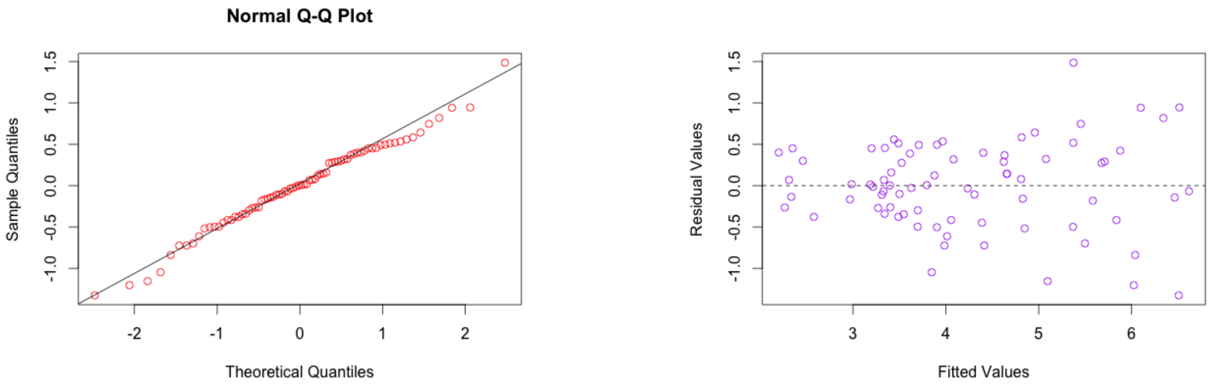


(A) Linear Regression

(B) LOESS Regression

(C) Spline Regression

FIGURE 5.6: Regression of Reliability Factor with Standard Deviation



(A) QQ plot

(B) Fitted VS Residuals plot

FIGURE 5.7: Diagnostic plots of linear Regression of Reliability Factor with respect to Standard Deviation

$$RF = 2.337 \times SD - 1.202 + \epsilon_i \text{ with } \epsilon_i \text{ the residual error of point } i \tag{5.1}$$

In order to confirm the results, a cross validation linear regression is conducted. Similarly as in Chapter 4, the data is divided into 10 folds and a for loop is used to find the regression model that minimizes the Mean Squared Error of both the training and validation set, which is illustrated in Figure 5.8.

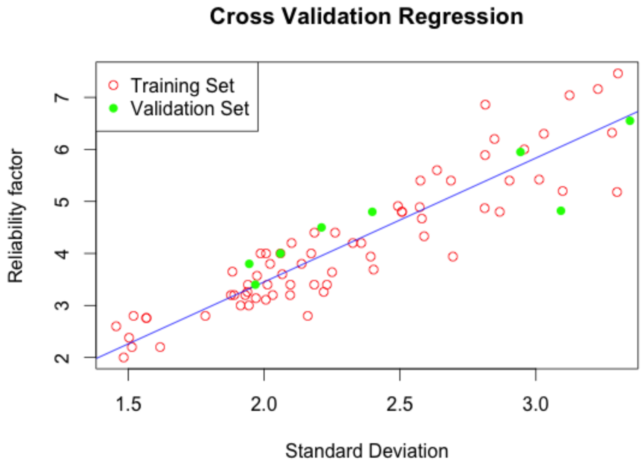


FIGURE 5.8: Regression Model Created Through 10-Folds Cross Validation loop

The regression model generated through this process and which combines the minimum Mean Squared Error of both the training set and validation set is shown below. It can be inferred that this model is as valid as the non-parametric LOESS model as their MSE are equal. The linear regression model linking the reliability to the standard deviation on the Victoria Line is thus validated.

$$RF = 2.381 \times SD - 1.310 + \epsilon_i \text{ with } \epsilon_i \text{ the residual error of point } i \tag{5.2}$$

$$MSE_{\textbf{training}} = 0.512 \text{ min and } MSE_{\textbf{validation}} = 0.545 \text{ min} \tag{5.3}$$

5.2 Regression on the Jubilee Line

Similarly, in order to investigate the reliability factor on the Jubilee line, the OD pairs with no interchange along this line, already examined for the log-normal distribution in chapter 4 are used. They are completed by the data of 5 other O-D pairs along this line and within zone 1 and 2, listed in Table 5.2.

O-D pairs from Chapter 4	**Additional O-D pairs**
Green Park-Waterloo	Bond Street-Waterloo
Baker Stree-Green Park	Westminster-West Hampstead
London Bridge-Bond Street	Canary Wharf-Westminster
Westminster-Baker Street	Canary Wharf-Green Park
Baker Street-Bond Street	Bond Street-Southwark

Table 5.2: Origin-Destination pairs along the Jubilee Line used for analysis

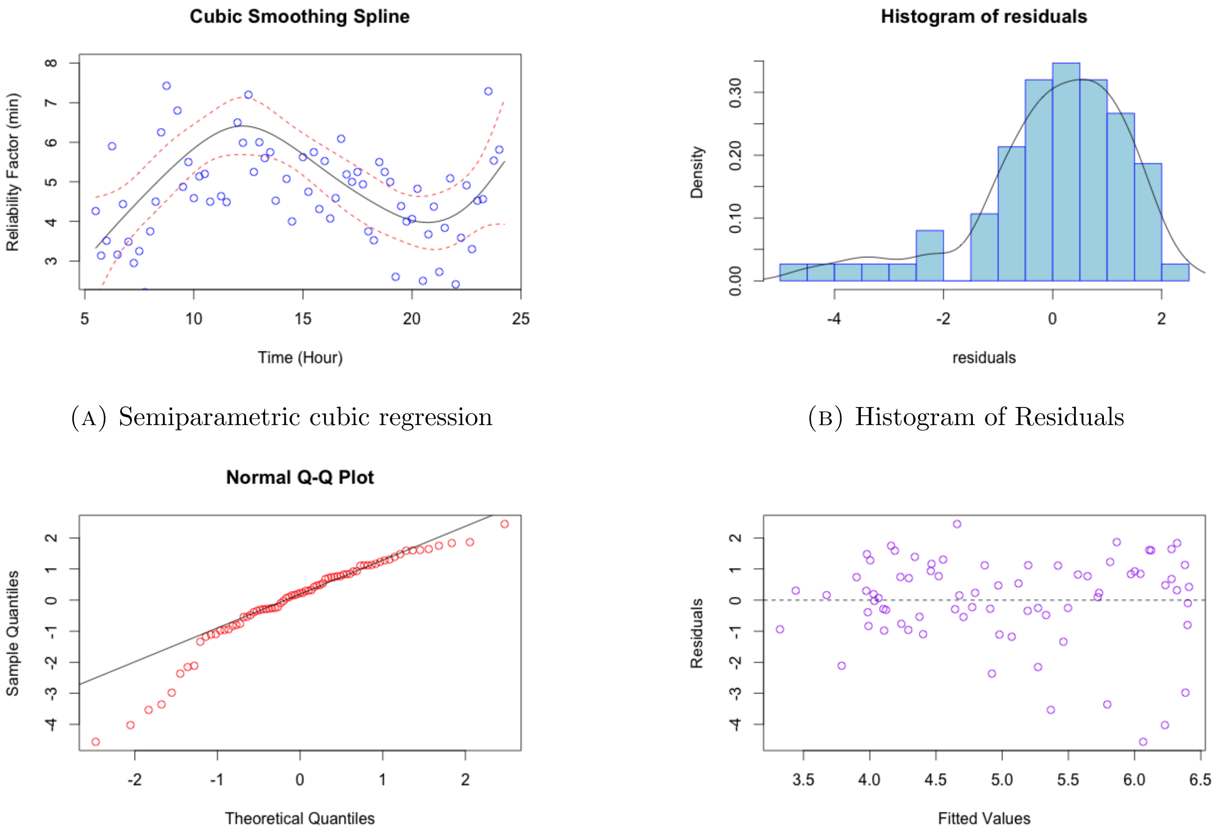
5.2.1 Regression of Reliability with Time

As it has been demonstrated on the Victoria Line that the polynomial cubic spline captured substantially the variability of the reliability along the day, this model has been applied to the reliability on the Jubilee line in Figure 5.9. This semiparametric model is also checked for any mis-specification or inconsistencies with diagnostic plots and it appears that many outliers lead to residual error that stand out from the horizontal random pattern in the Fitted versus Residual plot and to a heavy tailed normal distribution in the QQ-plot.

As a consequence, Cook's distance, which is used to estimate the influence of data point when performing regression analysis, permits elimination of outliers:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - y\hat{(i)}_j)^2}{MSE * p} \text{ with } i = 1, 2, ...n \tag{5.4}$$

p is the number of fitted parameters, in the linear case $p = 1$, $\hat{y}_j$ is the prediction from the regression model and $\hat{y_j}(i)$ is the prediction from the refitted regression model in which observation i has been removed. Applying this method through a for loop in R, the outliers are excluded and the regression model is refitted (Figure 5.10) The residual error satisfy the normality assumption, with constant variance, but its distribution still appear lightly tailed.



(A) Semiparametric cubic regression

(B) Histogram of Residuals

(C) QQ Plot

(D) Fitted Values Versus Residuals

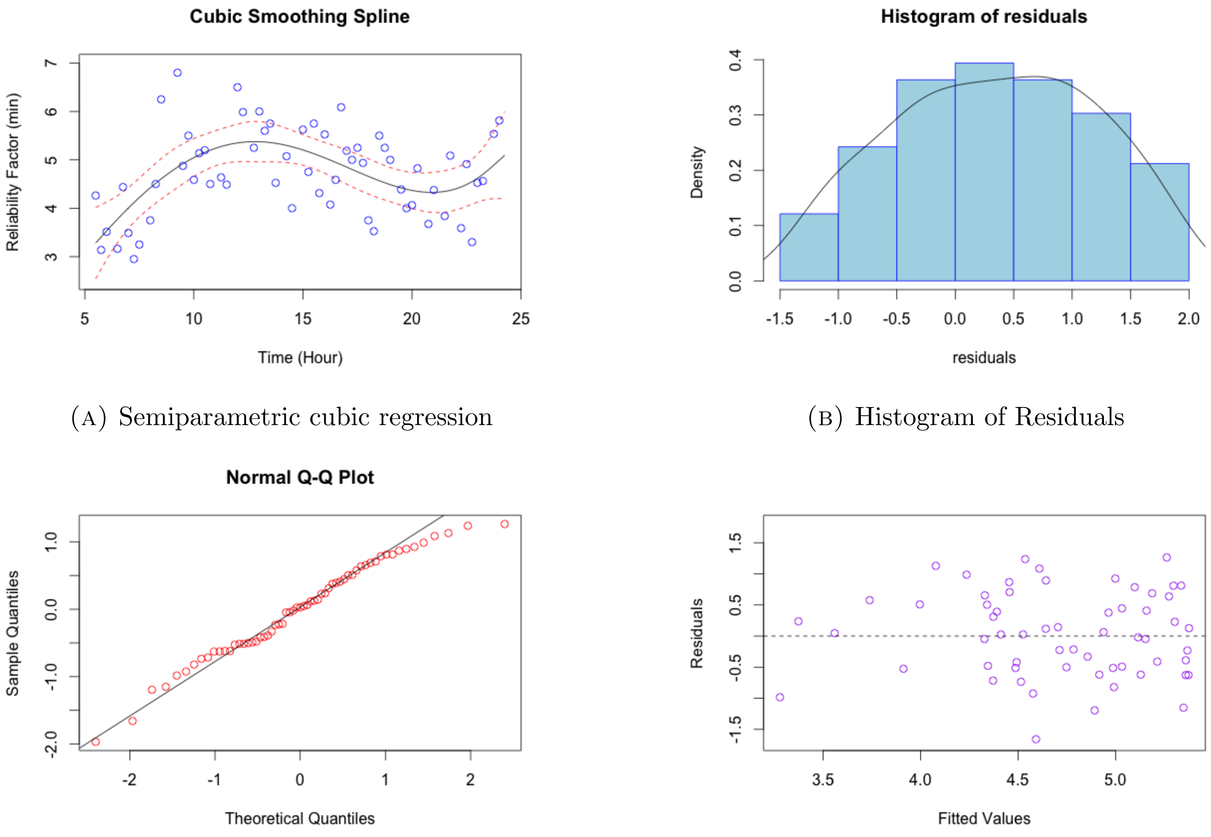
FIGURE 5.9: Linear Regression of Reliability factors with standard deviation and its diagnostic plots

The MSE before and after omission of the outliers is calculated, strengthening the suitability of the resulting cubic spline regression model, as the MSE is reduced by half.

$$MSE = 1.396 \text{ min } \textbf{with outliers}$$

$$MSE = 0.715 \text{ min } \textbf{without outliers}$$

As a conclusion, cubic smoothing spline regression has been found to be a suitable regression model for the variability of reliability with time of day on the Jubilee line as it satisfactorily captures its fluctuations along the day. In the AM peak, the reliability on the Jubilee line progressively decreases as the reliability factors rises from 3 to 5 minutes. The reliability on the line seems to slowly regain importance during the day, reaching a stable value of 4 minutes around 7pm.



(A) Semiparametric cubic regression

(B) Histogram of Residuals

(C) QQ Plot

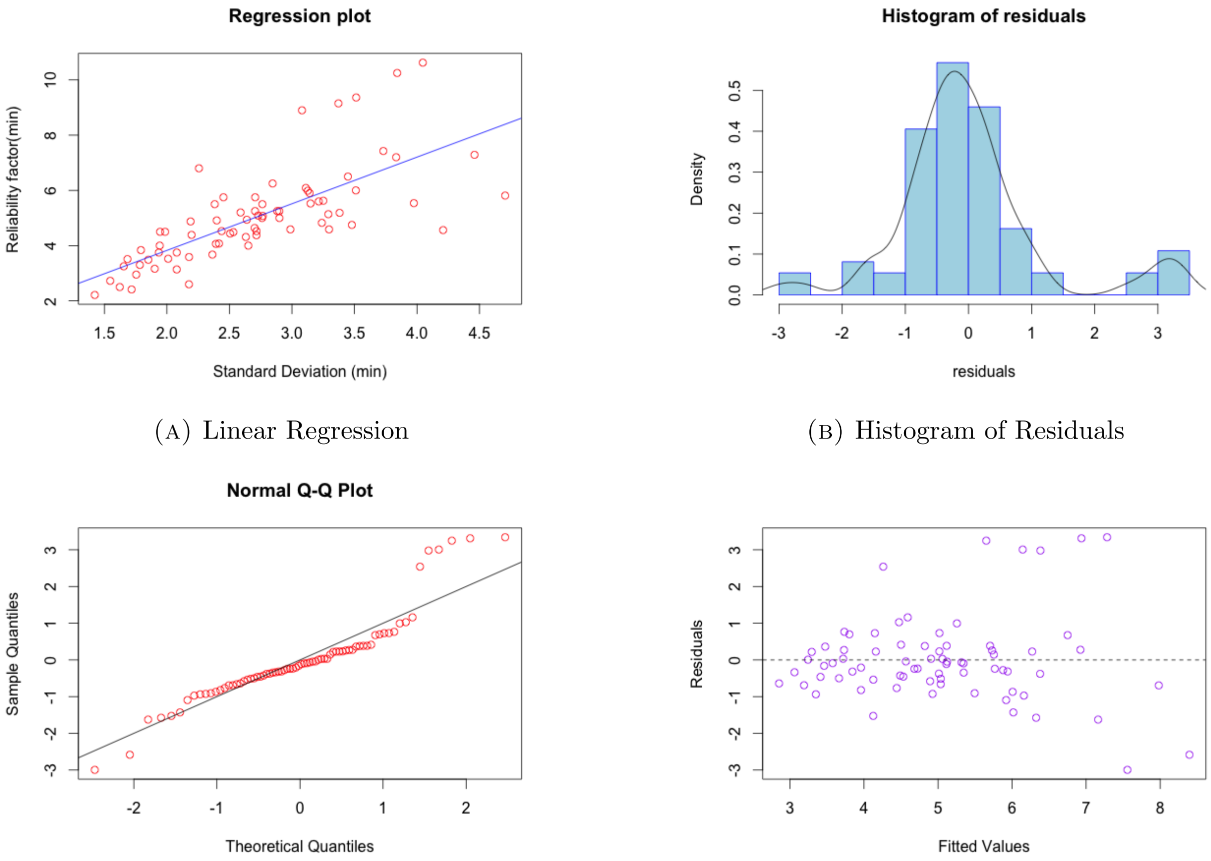
(D) Fitted Values Versus Residuals

FIGURE 5.10: Linear Regression of Reliability factors with standard deviation and its diagnostic plots

5.2.2 Reliability with respect to Standard Deviation

Building on the conclusion of a linear regression between the reliability factor and the standard deviation of a trip on a certain line, the linear regression model is generated for these two variables on the Jubilee Line. In order to come up with the plots 5.11, the reliability and standard deviation of all the listed O-D pairs (Table 5.2) is calculated in 15 minute intervals and stored into data frames.

In Figure 5.11, histogram of residuals is Gaussian, with the exception of both its tails. This result is confirmed by the Quantile-Quantile plot, which highlights heavy tails for the distribution of the residuals, as well as the fitted Versus residuals values plot, which reveals some extreme residuals. The bulk of the residuals is randomly scattered around the $y = 0$ line, forming a horizontal band and suggesting constant variance. Indeed, it can be implied that some values in the data set are outliers. Cook's distance is thus applied to the data to eliminate the outliers and their effect on regression, similarly to the previous section, and the linear model is re-fitted.



(A) Linear Regression

(B) Histogram of Residuals

(C) QQ Plot

(D) Fitted Values Versus Residuals

FIGURE 5.11: Linear Regression of Reliability factors with standard deviation and its diagnostic plots

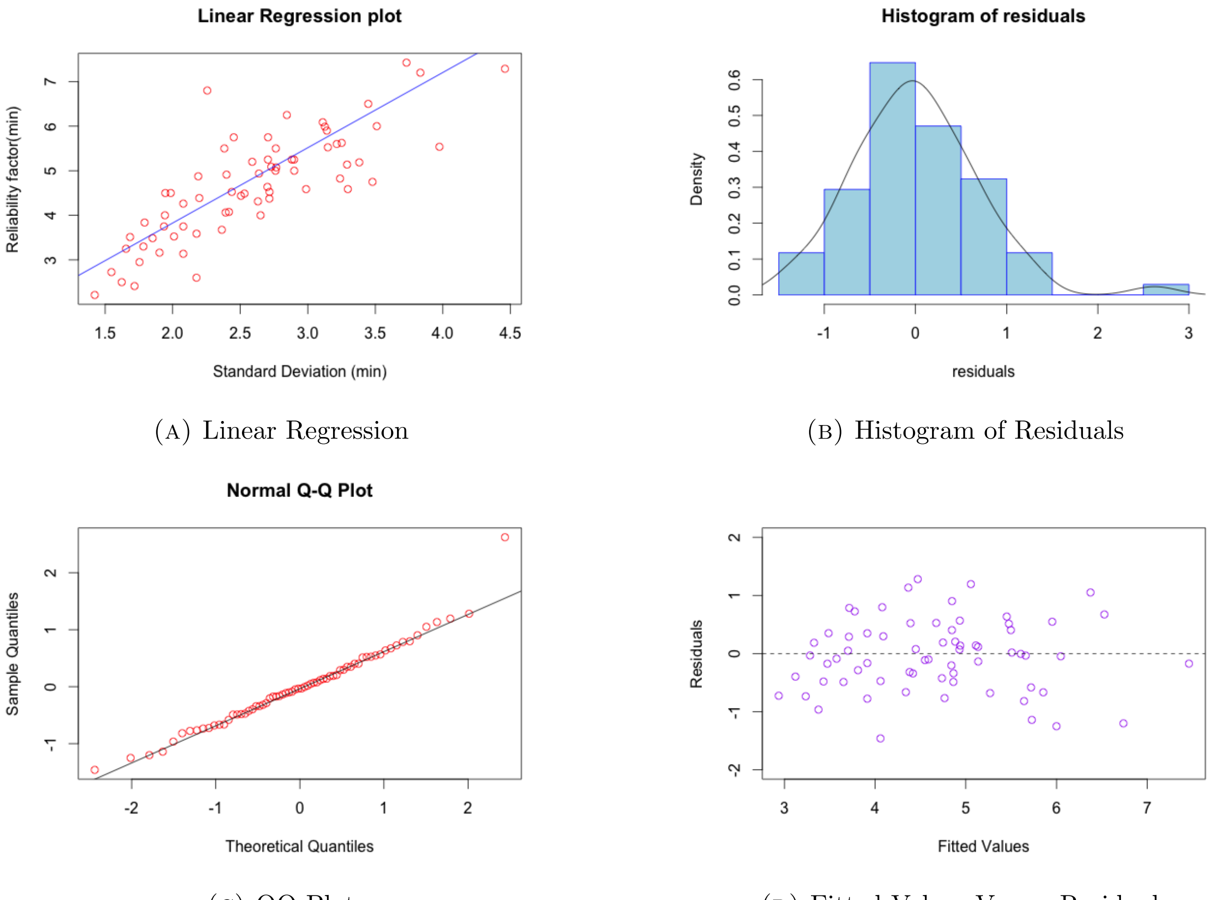
The Mean Squared Error is calculated before and after removing the outliers.

$$MSE_{\textbf{linear}} = 1.166 \text{ min } \textbf{with outliers}$$

$$MSE_{\textbf{linear}} = 0.686 \text{ min } \textbf{without outliers}$$

The MSE is reduced by almost half when removing the $m = 7$ outliers values using Cook's distance, showing the improved performance of the new linear model. The new regression model as well as its diagnostic plots are shown in Figure 5.12.

The improved regression and diagnostic plots reveal that the linear regression model is suitable to characterise the relationship between the reliability factor and the standard deviation. The cubic smoothing spline and the LOESS nonparametric models are plotted to strengthens this conclusion (Figure 5.13) and the Mean Squared Errors are calculated. Indeed, the nonparametric plot are close in shape to the linear regression plots, while the MSE are smaller but not significantly, as the relative difference is smaller than 5%. Therefore, the linear model being the simplest and most interpretable is recommended.



(A) Linear Regression

(B) Histogram of Residuals

(C) QQ Plot

(D) Fitted Values Versus Residuals

FIGURE 5.12: Linear Regression of reliability factors with standard deviation excluding outliers

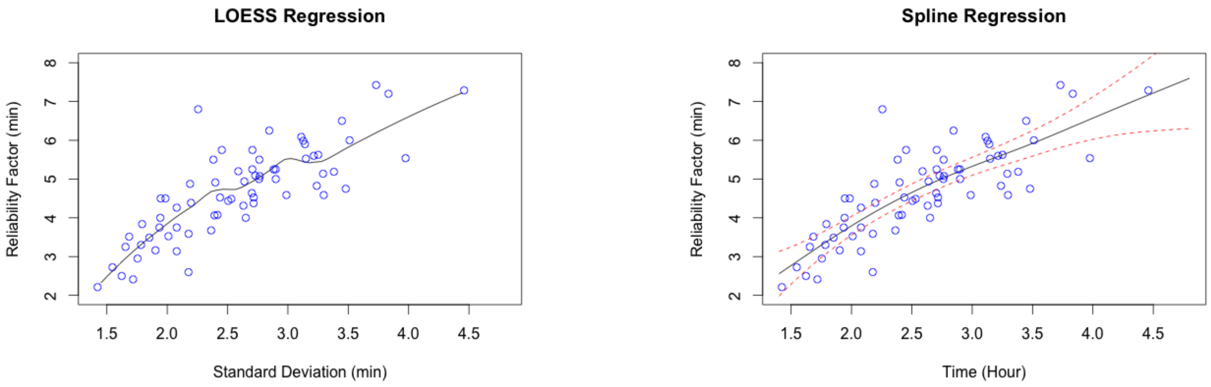
$$MSE_{\textbf{LOESS}} = 0.651\text{min}$$

$$MSE_{\textbf{Spline}} = 0.668\text{min}$$

The boxplot 5.14 of residuals errors of the different models confirms the previous conclusion, as the range of the residual error is equal for the three models. As the linear model is found to be very similar to the nonparametric and semiparametric models, it is hence selected as the preferred one due to its parametric nature.

Consequently, the reliability factor of any 0-D trip on the Jubilee Line is linearly dependent on its standard deviation. This relationship highlights the fact that the standard deviation has a direct impact on the reliability of travelling on the Jubilee line and its significance should not be overlooked.

$$RF = 1.489 \times SD - 0.818 + \epsilon_i \text{ with } \epsilon_i \text{ the residual error of point } i \tag{5.5}$$



(A) LOESS Regression (B) Spline Regression

Figure 5.13: Nonparametric Regression Models for the Jubilee Line Reliability Factor

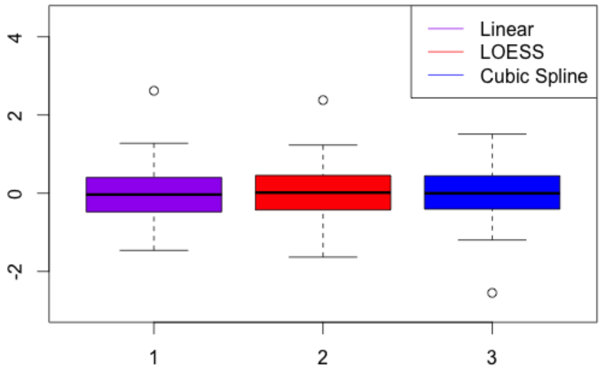


Figure 5.14: Boxplot of residuals errors to compare regression models

5.3 Findings

It can be highlighted that, by looking at both the regression models of the reliability factor with time along the Victoria and Jubilee Line (within zone 1 and 2), the cubic smoothing spline regression model appears very promising in capturing the evolution of reliability throughout the day, as it depicts well the variability of reliability engendered through peak and off-peak times. This model could become a useful tool towards predicting the reliability of a selected line at a specific time of day. Similarly, for both the Jubilee Line and the Victoria Line, a linear regression model was fitted to model the relationship between standard deviation and reliability of O-D trips. The reliability of a selected trip could thus be estimated through the dispersion of the journey time distribution. As a conclusion, the reliability factor on the Victoria and Jubilee Lines is a dependent variable that responds to two variables, which are time of day and standard deviation. The first one is treated semi-parametrically while the latter one is handled with a linear model. Therefore, semiparametric and linear regression appear as effective tools towards characterising and quantifying reliability on a specific line. It would be interesting to investigate whether this linear relationship between reliability factor and standard deviation or time of day could be extended to all lines within zone 1 and 2 of the Underground, in order to provide users with reliability measurements that would prevent them from underestimating travel time.

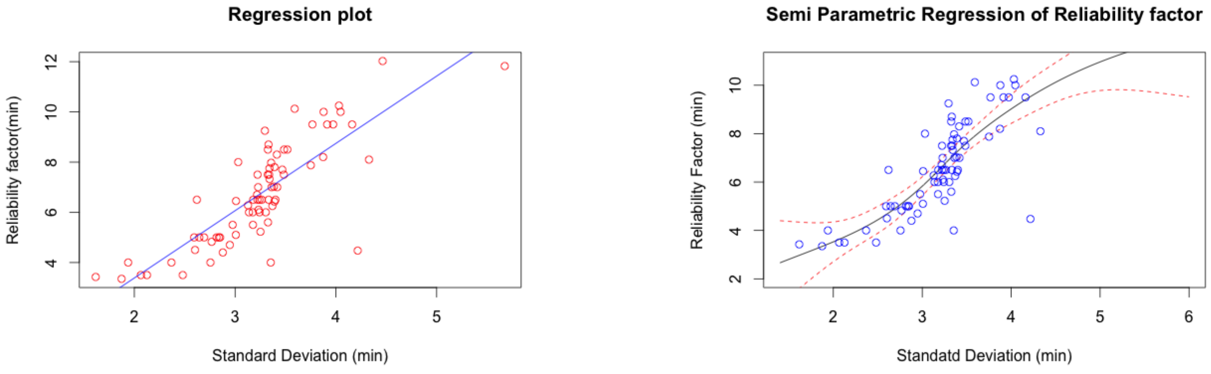
5.4 Regression analysis on pairs with interchange

In the light of the previous results, the reliability of O-D pairs along the same line is directly related to the standard deviation of the journey time. An interesting aspect would be to check whether these results are consistent throughout all types of O-D pairs within zone 1 and 2. Therefore, the reliability and standard deviation of O-D journeys including an interchange and/or two optimal paths, listed in Table 5.3, were extracted.

Complex O-D pairs data	
O-D pairs	**Number of obs.**
South Kensington- Oxford Circus	51,763
Oxford Circus-Waterloo	44,360
Baker Street-Victoria	12,301
Bank/Monument-Green Park	9.837
Green Park- Embankment	975

TABLE 5.3: Origin-Destination pairs including an interchange and/or two optimal paths

An average reliability vector and standard deviation vector were calculated for the 12 days of data and both a linear and semiparametric regression model are fitted to the data, as seen in Figure 5.15.



(A) Linear Regression

(B) Semi Parametric Regression

FIGURE 5.15: Regression models for Origin-Destination pair with interchange

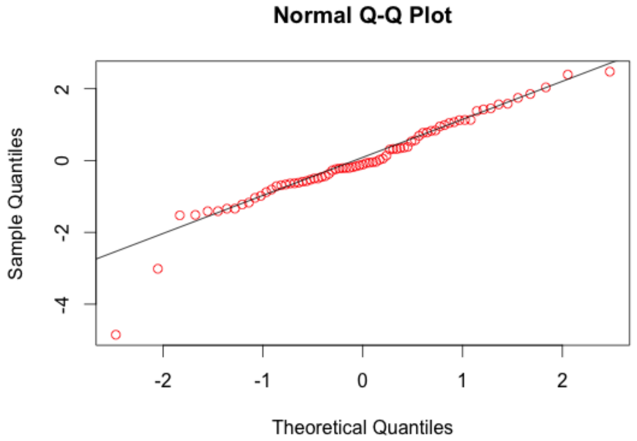


Figure 5.16: QQ plot of Linear Regression

The relationship between reliability and standard deviations seems to remain linear, even when fitted with a semi-parametric model as the latter produces a very similar fit. This is confirmed by the similarity of the MSE of both models, suggesting that linearity can be validated.

$$MSE_{\textbf{Linear}} = 1.16\text{min}$$
$$MSE_{\textbf{SemiPar}} = 1.11\text{min}$$

Finally, the Quantile-Quantile plot 5.16 of the residual errors confirms that the initial assumptions are respected, as the error appears normally distributed.

From the above analysis, it can be inferred that the linear regression relationship between standard deviation and reliability is not restricted to O-D pairs without an interchange (on the same line) as the results were extended to a set of O-D pairs within zone 1 and 2, which all included an interchange on their route. Reliability factor is linearly related to standard deviation, strengthening Chan's (2007) statement that reliability is linked to the compactness of travel time distribution.

Chapter 6

Conclusion and Further Research

One of the main objectives of Transport for London is improving the regularity and reliability of journey times on the Underground. Indeed, commuters and users require a constant quality of the service, exempt of any unexpected irregularities. With the emergence of journey planners and live tracking information, users can evaluate and predict the travel time of the intended journey. However, these tools are based on the calculation of journey time through the mean and have been proven to often underestimate the true journey time of travellers, leading to unpredicted delays for the users. Being able to characterise and estimate reliability in the Underground has arisen as a priority.

In order to quantify reliability, this paper investigates different approaches. For O-D pairs along the same line, reliability of journey time has been found to respect a pattern. The factors contributing to the total travel time, including access, egress, platform-waiting and on-train time have a multiplicative effect that is well captured by the log-normal distribution. However, when the optimal path between two O-D is more complex and includes interchanges or several optimal paths, this statement does not hold. The journey time distribution is more disperse and nonparametric distributions have to be used. In other words, the journey time is inherently less reliable through the intervention of numerous new factors from the choice of path, which raises many new variables like the speed of lines chosen, their respective headways and number of stops, to the state of the interchange station that might be subject to congestion or to different length of walk for platform change. Therefore, in the latter case, only nonparametric method can be used towards quantifying reliability and journey time dispersion.

Secondly, this paper offers new insight into the variability of reliability through regression models. Reliability along the Victoria and Jubilee line, within zone 1 and 2, has been found to be related to the dispersion of journey time as well as the time of day. Reliability of a trip on a selected line and at a specific time of day can hence be approximated. Indeed, reliability obeys a linear relationship to standard deviation (regardless of the fact that the O-D includes or not an interchange), and a cubic spline relationship to time of day. This builds on the results of Chan (2007), which described the compactness of the journey time distribution as an indicator of the level of reliability, by creating a direct relationship between the dispersion of the distribution of journey time and reliability. However, the fluctuations along time are more complex and seem to follow the pattern of peak and off-peak time, which explains the cubic aspect of the fitted spline. By using these regression models to fit the chosen reliability metric, reliability on the lines of the London Underground could be further investigated. The figures produced, with straightforward interpretation, would expose the lines on which improvements have to be made by Transport for London. The regression models would become effective tools to assess system performance and would result in better planning and scheduling, for example reducing headways of trains when reliability appears to suffer from the consequences of congestion, or towards a more efficient use of train capacity.

In the light of these findings, it would be interesting to research further into the regression models to build a multivariate analysis that could adjoin the two independent variable (time of day and standard deviation) to produce one single reliability measurement. This metric would capture the fluctuations of reliability along a line as a function of both time of day and dispersion of the journey time distribution between an Origin-Destination pair and could be used as an essential metric in journey planners tools, providing a plus or minus value of travel time with the estimated journey time to ensure on-time arrival for the user. Similarly, with that extra information, passengers would be more likely to select an optimal path between the several options available to complete a trip.

Validity and Limitations of the findings

This paper faces many limitations. To start with, the size of the data appeared as the first hindrance. Each day data file, which exceeds 4 GB, had to be run through SPSS for every Origin-Destination pair studied, leading to an extremely slow process of acquiring ”clean data” for every O-D pair investigated. Therefore, although the data of 24 working days was available for analysis, only 12 were processed through in order to increase the number of Origin-Destination to be treated. Working with only 12 days reduces the sample size and might influence the accuracy and credibility of the statistics produced during analysis. Moreover, the choice of excluding week ends of the data was made, due to an extremely different dynamic of the Underground during this period. Therefore, the paper only treats reliability during the week days and further research should be completed to quantify this measure during week-ends, comparing it with the results for week days. Likewise, another interesting aspect would be to measure reliability as a function of day of the week to infer whether reliability varies depending on this variable.

The reliability of only two lines out of the eleven has been examined, and although the results on both of them were similar, the conclusions cannot be generalised without rigorous checks. The analysis is also bound to Central London (Zone 1 and 2), and does not consider the Greater London area.

Moreover, the access and egress time of each user are included in the journey times and consequently, are variables that bias the result of the performance of the lines. Therefore, a further research could focus on this matter in order to rigorously study the performance. The on-train time could be isolated by analysing station flows to estimate walking times and deduce platform-waiting time, as suggested by Chan (2007).

Finally, interchanges appear as a major limitation to reliability measurements as the users, when shifting line on the Underground, do not have to tap their Oyster card and therefore, no data is recorded. The path choice of the passenger remain assumed and travel behaviour is hardly apprehensible, except through surveys and flow analysis. Indeed, as the path used remains unknown and the factors to consider numerous, reliability measurements are scarcely interpretable. Another interesting prospect would then be to further investigate journey time performance for journeys including interchange. Indeed, they represent almost half of the trip undertaken on the Undergound and although it has been demonstrated that for these trips, reliability linearly depends on standard deviation, their travel time distribution remains complex to characterise. By focusing on the path choices, these problems could be overcome and accurate reliability measurements would become achievable.

Bibliography

Abkowitz, M., Slavin, H., Waksman, R., Wilson, Nigel et al., 1978. *Transit Service Reliability.* Cambridge, Massachussets: U.S. Department of Transportation Systems Center.

Batty, M., 2012. Smart Cities, big data. *Environment and Planning B*, 39, pp.191-93.

Beecham, R., Wood, J. & Bowerman, A., 2013. Studying commuting behaviours using collaborative visual analytics. *Computers, Environment and Urban Systems* , (47), pp.5-15.

Chan, J., 2007. *Estimation and Journey Time Reliability Metrics Using Automated Fare Data.* MS Thesis. Cambridge: Massachussets Institute of Technology.

Conover, W.J., 1999. *Practical Nonparametric Statistics.* 3rd ed.

Feigelson, E. & Babu, G.J., 2015. *Beware the Kolmogrov-Smirnov Test!* Paper. Pennsylvania: The Pennsylvania State University.

Foell, S., Rawassizadeh, R. & Gerd, K., 2013. Informing the design of future transport information services with travel behaviour data. In *Workshop on SenCity: Uncovering the hidden pulse of a city.* Zurich, 2013. The Open University.

Fu, Q., Liu, R. & Hess, S., 2012. *On considering journey time variability and passengers' path choice behaviours: An empirical study with the use of Oyster card data on London Underground.* Paper. Leeds: Institute for Transport Studies, University of Leeds.

Hogg, V.R., Tanis, A.E. & Zimmerman, L.D., 2015. Test of Statistical Hypotheses. In *Probability and Statistical Inference.* Ninth Edition ed. Pearson. pp.363-414.

Lathia, N. & Capra , L., 2011. *How smart is your Smartcard? Measuring Travel Behaviours, Perceptions and Incentives.* Paper. London: Department of Computer Science, University College London.

Paine, F.T., Nash, A.N., Hille, S.T. & Brunner, G.A., 1976. Consumer Attitudes towards Auto versus Public Transport Alternatives. *Journal of Applied Psychology*, 53(6), pp.472-80.

Paine, F.T., Nash, A.N., Hille, S.T. & Brunner, G.A., 1976. Consumer Attitudes towards Auto versus Public Transport Alternatives. *Journal of Applied Psychology*, 53(6), pp.472-80.

Polak, J., Bates, J., Jones, P. & Cook , A., 2001. The Valuation of Reliability for Personal Traveler. *Transportation Research, Part E*, 37(2), pp.191-229.

Powell, J.L., 1994. Estimation of Semi-parametric models. In *Handbook of Econometrics*. New Jersey: Elsevier Science. pp.2444-65.

Powell, J.L., 2009. *Notes on Nonparametric Density Estimation*. Notes. Berkeley: Department of Economics, University of Califronia.

Rosso, R. & Kottegoda, N.T., 2008.*Methods of Regression and Multivariate Analysis*. In Applied Statistics for Civil and Environmental Engineers. 2nd ed. Blackwell Publishing. pp.326-397.

Rosso, R. & Kottegoda, N.T., 2008.*Model Estimation and Testing*. In Applied Statistics for Civil and Environmental Engineers. 2nd ed. Blackwell Publishing. pp.230-316.

Schrank, D., Turner, S., Margiotta, R. & Lomax, T., 2003. *Selecting Travel Time Reliability Measures*. Texas Transportation Institue and Cambridge Systematics.

Transport for London TfL, 2007. *Transport Planning Business Operations*. London Travel Report. London: TfL.

Uniman, D., 2009. *Service Reliability Measurement Using Smart Card Data: Application to the London Underground*. Master's Thesis. Cambridge: Massachussets Institue of Technology.

Uniman, D.L., Attanuci , J., Mishalani, R.G. & Wilson, N.H.M., 2010. *Service Reliability Measurement Using Automated Fare Card Data*. Transportation Research Record: Journal of the Transportation Research Board, (2134), pp.92-99.

Appendix A

R Code

Travel Time Distribution for an O-D pair

```
## Merge all data sets between an O-D pair
mergeddata<-rbind(set1,set2,set3,set4,set5,set6,set7,set8,set9,set10,set11,set12,use.names=FALSE)

mergeddata1<-subset(mergeddata,V16!=V58)

JourneyTime<-mergeddata1$V62-mergeddata1$V42
mergeddata2<-data.frame(mergeddata1,JourneyTime)

## Non parametric travel time distributions of Specific OD pairs:
d <- density(JourneyTime, bw=1, kernel="epanechnikov") # returns the density data
plot(d, main="Kernel Density of Journey Time") # plots the results
hist(JourneyTime,breaks=40, ylim=c(0, 0.25), col="lightblue", border="blue", freq=FALSE)
lines(d, col="red")

h <- 1   # smoothing factor
testk <- bkde(JourneyTime, bandwidth=h, kernel="epanech")
lines(testk,type="l",lwd=2, col="green", main=paste("KDE, h =",h),xlab="journeytime",ylab="testk")
legend("topright",c("Histogram of real journey time","Non parametric estimation 1","Estimation 2"),lty=

## Plots of riderships to highlight peak travel times
t <- cut(mergeddata3$V42,breaks=seq(300,1500,length=81),labels=c(seq(5,24.75,by=0.25)),
include.lowest=TRUE,right=FALSE,dig.lab=4,ordered_result = FALSE)

dat <- data.frame(mergeddata3,JourneyTime,t)
C <- table(t)
C<- data.frame(C)
colnames(C) <- c("time","ridership")

ggplot(data=C,aes(x=time, y=ridership,group=1)) +geom_point() +
geom_line(group=1,size=1,color="blue")+
scale_x_discrete(breaks=c(5:25),"time(hour)")

## Calculate Statistical Values
vol<-length(JourneyTime)
```

```
mean<-round(mean(JourneyTime),2)
var<-round(var(JourneyTime),2)
skewness<-round(skewness(JourneyTime),2)
reli<-quantile(JourneyTime ,0.95 ,na.rm=T)-median(JourneyTime)

## Log normal distribution estimation
phi = sqrt(var + mean*mean)
mu    = log(mean*mean/phi)
sigma = sqrt(log(phi*phi/(mean*mean)))

x = rlnorm(800,meanlog=mu,sdlog=sigma)
grid = seq(0,40,.1)
lines(grid,dlnorm(grid,meanlog=mu,sdlog=sigma),type="l",col="green")
legend("topright", c("Histogram of Real Journey Time","Nonparamtric estimation","Log normal distributio

## Calculate mean, variance and reliability factor for intervals of 15 minutes and plot variability dur
calc <- ddply(dat, .(t) , summarize, mean = round(mean(JourneyTime),2),  var = round(var(JourneyTime),2

ggplot(calc,aes(x=t, y=mean,group=1)) +geom_point() +
geom_line(group=1,size=1,color="blue")+
scale_x_discrete(breaks=c(5:25),"time(hour)")

ggplot(calc,aes(x=t, y=var,group=1)) +geom_point() +
geom_line(group=1,size=1,color="red")+
scale_x_discrete(breaks=c(5:25),"time(hour)")

ggplot(calc,aes(x=t, y=reli,group=1)) +geom_point() +
geom_line(group=1,size=1,color="green")+
scale_x_discrete(breaks=c(5:25),"time(hour)")

## Store reliability vector and standard deviation vector
relivector1 <- calc$reli
sdvector1 <- sqrt(calc$var)
t4 <- seq(from=5.5, to=24.25, by=0.25)

reliframe <- data.frame(calc$reli,t4)
write.csv(reliframe, file="Reli Victoria- Warren street")

sdframe <- data.frame(sdvector1,t4)
write.csv(sdframe, file="SD Victoria-Warren Street")

## Peak and off peak data historgrams
p.op <- ifelse((mergeddata3$V42>=960 & mergeddata3$V42<=1200), "peak", "off-peak")
table(p.op)

finaltable <- data.frame(mergeddata3,JourneyTime,p.op)

peakdata1 <- subset(finaltable,p.op=="peak",select=c(journeytime,p.op),row.names= NULL)
nonpeakdata1 <- subset(finaltable,p.op=="off-peak",select=c(journeytime,p.op),row.names=NULL)

dataforhist <- rbind(peakdata1, nonpeakdata1)
row.names(dataforhist) <- seq(nrow(dataforhist)) # To cancel the row names!

ggplot(dataforhist, aes(JourneyTime, fill = p.op), col=c("blue","red"), border="black") + geom_histogra
ggplot(dataforhist, aes(JourneyTime, fill = p.op)) + geom_density(kernel="biweight", adjust=5, alpha =
```

Testing Procedures

```
## Permutation Testing with bootsrapping
numofpermutations=10000
diff.random=NULL

for (i in 1 : numofpermutations) {
set.seed(i)    ## to ensure the results are reproducible
pseudoData = rlnorm(n=100, meanlog=mu, sdlog=sigma) # this generated n samples from the NULL hypothesis
# we then get the empirical CDF for this data
pseudocdf = ecdf(pseudoData)
# calculate our KS statistic based on pseudocdf & plnorm,  we store this value of the KS statistic
diff.random[i]=max(abs(pseudocdf(unique(pseudoData))-plnorm(unique(pseudoData),meanlog=mu, sdlog= sigma
}

a=ecdf(JourneyTime)
diff.observed <- max(abs(a(unique(JourneyTime)) - plnorm(unique(JourneyTime), meanlog = mu, sdlog = sig
pvalue= sum(abs(diff.random) >= abs(diff.observed)) / numofpermutations
print (pvalue)
extreme <- max(diff.random)

# pvalue > alfa=0.01 : Fail to reject Null Hypothesis
hist(diff.random,border="blue")


## 10-FOLDS CROSS VALIDATION

## Divide teh data into 10 folds
library(caret)
k<-10
folds <- createFolds(JourneyTime, k , list = TRUE, returnTrain = FALSE)

for(i in 1:k){
set.seed(i)
train <- dataset[folds$subsets[folds$which != i], ] #Set the training set
validation <- dataset[folds$subsets[folds$which == i], ] #Set the validation set

mtrain <- mean(JourneyTime[folds$train])
vtrain <-var(JourneyTime[folds$train])

phi2 = sqrt(vtrain + mtrain*mtrain)
mu2   = log(mtrain*mtrain/phi2)
sigma2 = sqrt(log(phi2*phi2/(mtrain*mtrain)))

## Calculate MSE of fitted data (training set) and compare it with MSE of validation set, to choose the
l1=length(JourneyTime[folds$validation])
grid2=seq(0,40,length=l1)
sim=dlnorm(grid2,meanlog=mu2,sdlog=sigma2)
diff <- sim-JourneyTime[folds$validation]
MSEval <- sqrt(sum(diff^2)/l1)

l3<- length(JourneyTime[folds$train])
grid3=seq(0,40,length=l3)
```

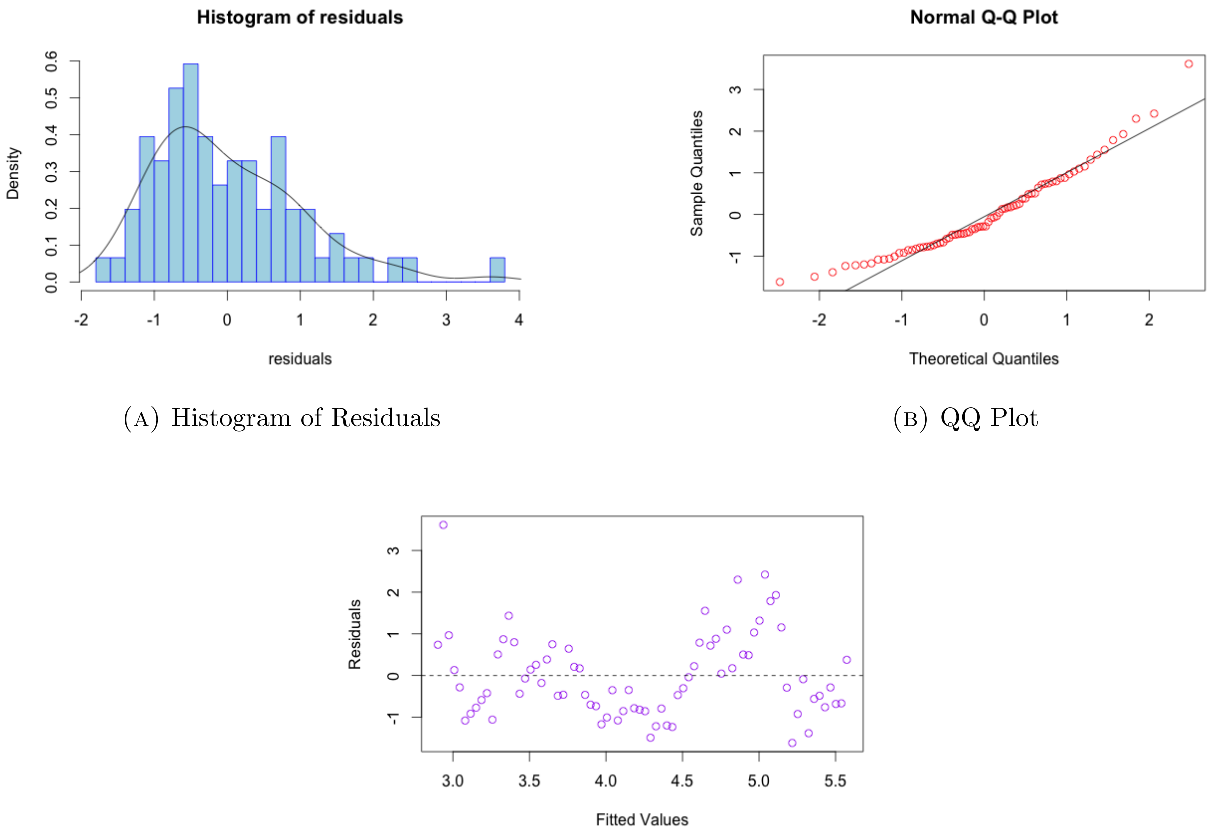
```
sim3=dlnorm(grid3,meanlog=mu2,sdlog=sigma2)
diff3 <- sim3-JourneyTime[folds$train]
MSEtrain <- sqrt(sum(diff3^2)/13)
```

Regression modelling

```
## Linear Regression of Reliability Factor with Standard Deviation
LinearFit <- glm (relitotal ~ sdtotal)
summary(LinearFit)
plot(sdtotal,relitotal, col="red", xlab="Standard Deviation", ylab="Reliability factor", main="Regressi
abline(LinearFit, col="black")

## Diagnostic Plots
d <- density(LinearFit$residuals)
hist(LinearFit$residuals, breaks=20, col="lightblue", border="blue", xlab="residuals", main="Histogram
lines(d)

qqnorm(LinearFit$residuals,col="red", xlab="Residuals",main="QQ-plot of log normal distribution")
qqline(LinearFit$residuals)

plot(LinearFit$fitted.values, LinearFit$residulas, ylim=c(-2,2),
xlab = "Fitted Values", ylab = "Residuals",col="purple")
abline(h=0, lty=2)

## LOESS Regression
model1=loess(relitotal4~sdtotal4, model=FALSE,span=0.5,degree=2,parametric=FALSE,family="gaussian",meth
newdata4=data.frame(sdtotal4=seq(1.4,4.8,length=68))
spred=predict(model1,newdata4,se.fit=TRUE)
plot(newdata4$sdtotal4,spred,type="l", ylim=c(2,8),xlab="Standard Deviation (min)", ylab="Reliability F
points(sdtotal4,relitotal4, col="blue")

## Spline Regression
model2=bigspline(t3,relitotal3, nknots=4)
newdata=data.frame(t3=seq(5.5,24.25,by=0.25))
spred=predict(model2,newdata,se.fit=TRUE)
yhat=spred[[1]]
yhatse=spred[[2]]
plot(newdata$t3,yhat,type="l", ylim=c(2.5,7),xlab="Time (Hour)", ylab="Reliability Factor (min)",main="
lines(newdata$t3,yhat+2*yhatse,lty=2,col="red")
lines(newdata$t3,yhat-2*yhatse,lty=2,col="red")
points(t3,relitotal3, col="blue")

## Semi Parametric (Cubic Smoothing Spline Regression)
model3=bigssp(relitotal~t,type="cub",nknots=4,skip.iter=FALSE)
newdata=data.frame(t=seq(5.5,24.25,by=0.25))
spred=predict(model3,newdata,se.fit=TRUE,include="t")
yhat=spred[[1]]
yhatse=spred[[2]]
plot(newdata$t,yhat,type="l",ylim=c(2,6),xlab="Time in hours", ylab="Reliability Factor (min)",main="Se
lines(newdata$t,yhat+2*yhatse,lty=2,col="red")
lines(newdata$t,yhat-2*yhatse,lty=2,col="red")
points(t,relitotal,col="blue")
```

Appendix B

Figures and Underground Map



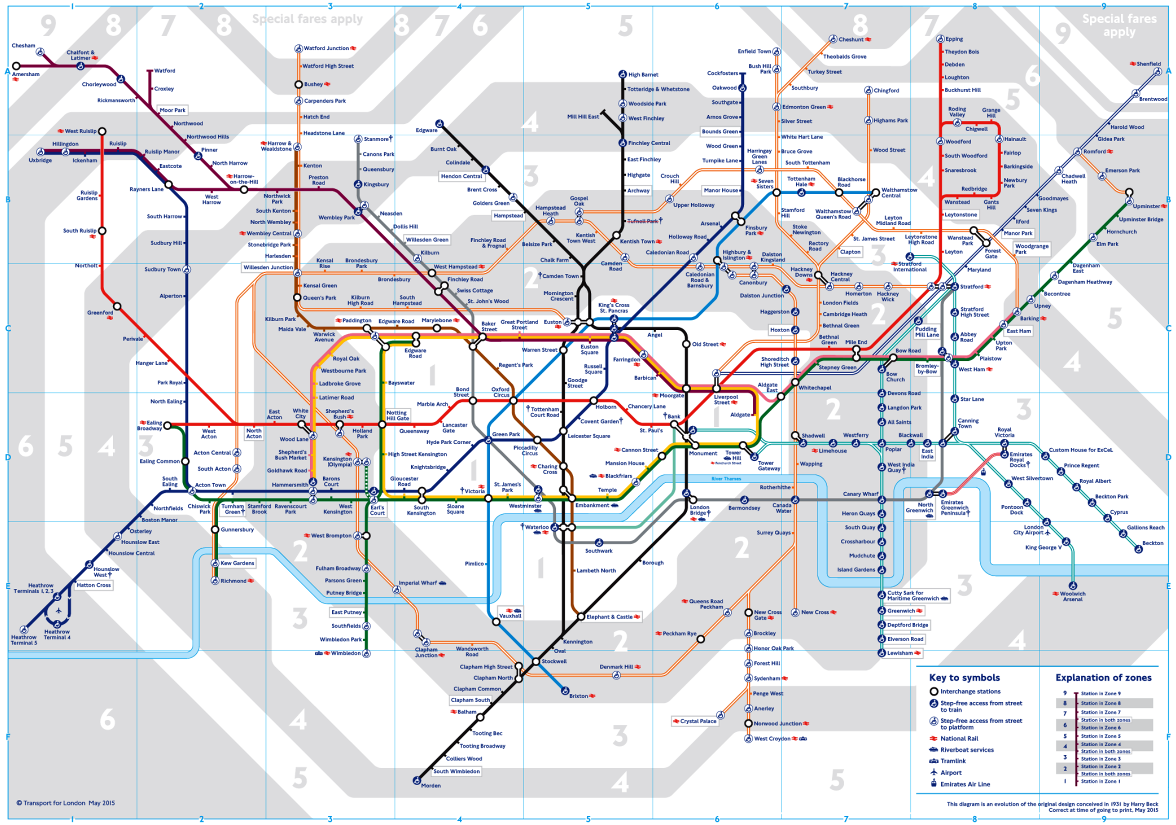
(A) Histogram of Residuals

(B) QQ Plot

(C) Fitted Values Versus Residuals

FIGURE B.1: Diagnostic Plots for Linear Regression of Reliability with time on Victoria Line

Tube map



Key to symbols

- ○ Interchange stations
- Step-free access from street to train
- Step-free access from street to platform
- National Rail
- Riverboat services
- Tramlink
- Airport
- Emirates Air Line

Explanation of zones

9	Station in Zone 9
8	Station in Zone 8
7	Station in Zone 7
	Station in both zones
6	Station in Zone 6
5	Station in Zone 5
4	Station in Zone 4
	Station in both zones
3	Station in Zone 3
2	Station in Zone 2
	Station in both zones
1	Station in Zone 1

This diagram is an evolution of the original design conceived in 1931 by Harry Beck
Correct at time of going to print, May 2015

† Check before you travel

Camden Town
Sundays 1300-1730 open for interchange and exit only.

Canary Wharf
Step-free interchange between Underground, Canary Wharf DLR and Heron Quays DLR stations at street level.

Covent Garden
Exit only until early November 2015.
Also on Saturdays and Sundays westbound trains will not stop. Please use Leicester Square instead.

Emirates Air Line
Emirates Greenwich Peninsula and Emirates Royal Docks
For full information about operating times and fares please visit *tfl.gov.uk/emiratesairline*

Heron Quays
Step-free interchange between Heron Quays and Canary Wharf Underground station at street level.

Hounslow West
Step-free access for manual wheelchairs only.

London Bridge
There is no step-free access to/from the Jubilee line until mid-August 2015.

Stanmore
Step-free access via a steep ramp.

Tottenham Court Road
Central line trains will not stop at this station until early December 2015.

Tufnell Park
Station closed from Monday 8 June until mid-March 2016.

Turnham Green
Served by Piccadilly line trains until 0650 Mondays to Saturdays, 0745 Sundays and after 2230 every evening. At other times use District line.

Victoria
During station improvement works there will be reduced interchange capacity between Circle/District and Victoria lines. Avoid queues at busy times by changing trains elsewhere.

Waterloo & City line
The Waterloo & City line is open between Bank and Waterloo 0621-0026 Mondays to Fridays and 0803-0026 Saturdays.
Between Waterloo and Bank 0615-0030 Mondays to Fridays and 0800-0030 Saturdays.
Closed Sundays and public holidays.

West India Quay
Not served by DLR trains from Bank towards Lewisham at certain times.

Night Tube Service From the early hours of Saturday 12 September 2015, Londoners and visitors will be able to travel on the Jubilee, Victoria and most of the Piccadilly, Central and Northern Tube lines all night on Fridays and Saturdays

Key to lines

- Bakerloo
- Central
- Circle
- District
- District open weekends, public holidays and some Olympia events
- Hammersmith & City
- Jubilee
- Metropolitan
- Northern
- Piccadilly
- Victoria
- Waterloo & City
- DLR
- Emirates Air Line
- London Overground
- TfL Rail

MAYOR OF LONDON

tfl.gov.uk

24 hour travel information
0343 222 1234*

Sign up for email updates
tfl.gov.uk/emailupdates

@TfLTravelAlerts

*Service and network charges may apply. See tfl.gov.uk/terms for details.

UNDERGROUND

TRANSPORT FOR LONDON
EVERY JOURNEY MATTERS