Manuscript Title Page

Title: Railway System Capacity Planning based on Statistical Analysis, Machine Learning and Discrete-Event Simulation

Authors & Affiliations:

Mousavi, A. Systems Engineering Research Group, Electronic and Computer Engineering Department, Brunel University London, UK

ORCID: 0000-0001-7887-5286

Email: alireza.mousavi@brunel.ac.uk

Tel: ++44 1895 274000

Munro, H.J. Department for Transport, 67 Parliament Hill, London, NW3 2TB, UK

Email: Harry.J.Munro@gmail.com

Tel: 07805449960

Railway System Capacity Planning based on Statistical Analysis, Machine Learning and Discrete-Event Simulation

Mousavi, A. and Munro H.J.

ABSTRACT

The accurate forecasting of dwell times for modelling and predicting the capacity of underground transport systems is presented. Accurate forecasting of capacity helps to find the optimal allocation of investment into transport systems and deliver the right level of service. In order to identify the factors that influence the dwell time for through running stations, data was gathered from the London Underground. The most significant factor was shown to be the total number of passengers boarding and alighting, a random phenomenon. Some factors such as the degree of station automation, signalling and platform to train interface characteristics was not included in the study.

Given the parameters considered, evidence is presented to show that a K-nearest neighbour regression algorithm most accurately predicts the mean dwell time. New evidence has been presented to show that gamma distribution is the best-fit to estimate the dwell time. This distribution can be employed in discrete-event simulation models to ascertain resource capacity and schedules of a given railway system.

A case study of the London Underground's Victoria Line was modelled using discrete-event simulation implemented in Python. The achievable capacity of this line was predicted up to 2050 assuming that there is a linear relationship between the number of boarders, alighters and the fluctuation of population of London. Losses in achievable capacity were demonstrated as the number of boarders and alighters were increased. An important weakness of the study is the assumption that the number of passengers on the Victoria Line increases linearly with London's population.

KEYWORDS

Dwell Time, Railway, Capacity Planning, Modelling and Simulation, London Underground, Commuter

1. INTRODUCTION

According to the Greater London Authority (2010), in order to ensure economic productivity, London's transport capacity should match the increasing levels of crowding. The capacity of a railway is largely dependent on the frequency, or headway, at which trains can pass through bottlenecks within the system. The achievable headway is determined by three factors: the run-in run-out time (RORIT), the dwell time and the operational parameters of the site. The RORIT time is, in turn, determined by the signalling system, track and train parameters. The dwell time can depend on capacity limiting factors, human factors, and can be constrained by a number of factors including passenger behaviour and platform to train interfaces. Operational parameters, such as the routing of trains and the changing of drivers have an impact on capacity.

An average of one second saving of dwell time or run-out run-in time (RORIT) on London Underground's Victoria Line equates to economic savings of £917,000 per year (Goodwin, 2015). Transport for London budgeted £1.141billion on capital expenditure investment in the 2017/18 financial year, with two of the three largest capital spends (the Four Lines Modernisation and the Northern line extension) targeting increases in capacity (Transport for London, 2018). A significant portion of this spend is on decreasing the RORIT component of capacity through upgrading rolling stock and signalling system technology, which enables trains to run closer together safely. There is therefore significant benefit in understanding the dwell time and making efforts to increase capacity.

The ultimate capacity of a system is generally limited by a single bottleneck (Goodwin, 2015). The Theory of Constraints (TOC) states that there is always one constraint on a business' performance, and that that the key to unlocking progress is to eliminate the constraint (Ox & Goldratt, 1986).

The aim of this study is to identify the factors that influence dwell time on the London Underground and to produce a model which can be used to predict the dwell time. A probability distribution function for the dwell time needs to be developed to demonstrate how dynamic dwell time modelling can be applied to capacity planning. A discrete event simulation platform is then used to test the models and conduct *what-if-scenarios* to achieve optimal planning solutions.

In order to achieve the aim the main objectives of the study are to test the hypothesis that dwell times are random and tend to follow a specific probability distribution. Conduct data analytics methods on data from London Underground to identify the factors which influence dwell times. Conduct a correlation analysis and develop the model to estimate the distribution parameters of the dwell time. Finally, to visualise and test the models, develop a discrete event simulation model using the estimated dwell times for capacity analysis of the railway system.

The methods for realising the objectives of the study commences with the analysis of the distribution of dwell times using maximum likelihood estimation of the parameters followed by calculation of the coefficient of determination of the estimated distribution against the real world data as a means of assessing goodness of fit (Chambers et al 1983).

2. BACKGROUND LITERATURE

2.1 The Dwell Time

At the London Underground there are several definitions of dwell time. One definition is wheel stop to wheel start. However dwell time can also refer to door open to door close, and also the time available for passengers to board the train (Wong & Key, 2014). For the purposes of this study, dwell time will be defined as the time from wheel stop to wheel start where passengers are boarding and alighting as the authors have had access to extensive data on the London Underground estate. Holloway et al (2013) report that passenger boarding rates had some casual association with door width and stand-back distance (the distance passengers are allowed to stand from the edge of the platform). There are also a number of key factors which have been identified as under reported. These include the influence of platform edge doors (PEDs), platform layout (e.g. number of exits), platform width and the effectiveness of station assistant train services (SATS) on reducing dwell times (Wong & Key, 2014). SATS are members of staff that are positioned on the platforms and are meant to help maintain consistent passenger behaviour on platforms, in theory improving reliability and safety.

A report prepared for London Underground Ltd identifies seven significant passenger, train and platform related variables, which have an effect on the dwell time. These factors are quantifies and a model to predict passenger boarding and alighting rates are proposed (Community of Metros and Imperial College London, 2013). The study had some success in validating its model for alighting rates, but was less successful validating the models predicted boarding rates against laboratory data. A key limitation of this study was the lack of validation against real-world data. To the best knowledge of the authors there are currently no recent published studies investigating the random behaviour of dwell times. Accurately estimating the probability distribution function of the dwell time is important from capacity modelling perspective.

2.2 Input Data Analysis

In this study Data Analytics techniques were chosen as the most relevant methodology to estimate the dwell times and the eventual capacity modelling. The maximum likelihood estimation (MLE) (Fisher, 1925) is one of the most popular methods for estimating parameters of distributions of data that appear stochastic in nature. The fit is computed by maximizing a log-likelihood function, with penalty applied for samples outside of range of the distribution. As the sample size grows larger the parameter estimates converge to an "optimal" value, the method is thus most suited to large datasets. The historical evolution of the MLE method can be viewed at (Meeker & Escobar, 1998). There are a number of different forms of regression analysis that can be used to model the relationship between a dependent variable and one or more independent variable(s). This study seeks to employ several different approaches and compare their effectiveness.

When dealing with multiple independent variables one popular technique tested in this study is multiple linear regression as described in Rencher and Christensen (2012). One advantage of this method is that the model is relatively simple and takes the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \cdots$$

K-nearest neighbour regression, as described in Altman (1992), is another regression method tested in this study. Decision tree regression is another non-parametric regression model that has been tested and is described in (Quinlan, 1986).

Feed forward neural network regression has also been tested. This is a model that is suitable for parametric and non-parametric regression. These neural network models are known as multi-layer perceptrons (MLP). The MLP is a series of logistic regression models, which result in forming a non-linear transfer function (Murphy, 2012). Open sourced Python based implementations of all the described regression models can be found online (scikit-learn developers, 2017).

2.3 Discrete Event Simulation and System Performance Engineering

Discrete event simulation has been continually developed since it first emerged in the 1950s. As simulation technology and computing power has grown, the complexity and power of such simulations has increased drastically. Many software packages have been developed to enable graphical modelling of discrete-event simulation. Recently, open-source simulation has emerged with the implementation of discrete-event simulation in Python (Simpy, 2016).

Systems engineering, an emerging discipline which takes an important place in a modern world of increasing complexity and interconnectivity, has been succinctly defined as "the creation and monitoring of requirements" (Tortorella 2015). Modelling and simulation is closely linked to systems engineering, and discrete-event simulation, which can now be employed with low cost through open source frameworks such as Simpy (Simpy, 2016), can be used to design and optimise system requirements at low cost.

On a railway, sites such as junctions, termini and depots, present a relatively high level of complexity to the systems engineer. The modelling of these sites involves a large number of variables, with interactions between independent entities, so non-linear behaviour presents itself. This is a characteristic of service systems, and it has been found that these systems can only properly be understood by using simulation (Longo, 2011) [Longo, F. (2011). Monte Carlo simulation, which discrete-event simulation employs, is used effectively to solve the problem of modelling complex systems (Zio, 2013).

Discrete-event simulation is powerful for modelling such systems. It has been shown to be an effective design tool for supply chain systems in the aerospace industry Visintin (2014). The non-linearity and complexity of such systems means that they share much in common with complex rail networks. It has been successfully employed to justify operational strategies at an airport runway (Mukkamala et al. 2008) and has been used to analyse urban freight on the Newcastle-Upon-Tyne (Motraghi and Marinov 2012).

Discrete-event simulation lends itself well to the integration with data analytics. This has been found to be particularly evident when integration DES with Python, since the data analytical aspect of Python can be integrated seamlessly with the simulation capability of DES; both the pre and post-processing of data can be integrated with a "sandwich" DES.

While not part of this study, there is clearly significant application for reinforcement learning in this framework. Output from simulations could be post-processed and fed back into a pre-processor and then into a new simulation, and so on and so forth. This could result in simulations that learn from their own experimentation, perhaps self-optimising, or being used to provide insights about optimal operation of the modelled system.

3. METHODOLOGY

Primary data was gathered from London Underground data sources. Dwell times by nature are stochastic and random; therefore the first step in the study was to investigate the patterns of dwell time occurrence. The distribution of data from the Victoria Station was plotted and a goodness-of-fit test (e.g. the Kolmogorov-Smirnov) conducted on the data to compare it to the estimated distribution. Multivariate analysis was applied to identify the significant factors with a p-value success criteria of less than 0.05. Regression analysis will be used to form a predictive model. Linear and non-linear regression models will be tested. A randomly selected 80% of the dataset will be used for developing the model and the remaining 20% of the dataset will be used for validation.

Track circuits were used to collect the dwell times. Track circuits are simple electronic components which detect the presence of a train on a section of track, thus every train and every movement that occurs are captured in the central database. The dwell time was calculated using equation 1.

$$DwellTime = WheelStartTime - WheelStopTime \tag{1}$$

This track circuit occupation data is known in London Underground as NETMIS data. The month of November 2015 was used as a sample for this study, since this is the period of time for which the Rolling Origin and Destination Surveys (RODS) were carried out. It contained 527,332 rows of data.

Since the mean dwell time is of interest, this summary statistic needs to be calculated for each location and direction combination for each station on the network. A Python script was developed to process the raw NETMIS data in order to calculate the mean dwell at each location-direction combination. However, the mean for each location-direction combination cannot simply be calculated for the entire month. This would yield a dataset that is too small and could not be correlated with passenger number data for results that are statistically significant. To overcome this challenge a statistical approach called resampling was employed.

Resampling time series based data, such as NETMIS data, involves calculating a summary statistic (in this case the mean) for a period of time at fixed intervals in the data. Taking one data label, Victoria Northbound, as an example. Raw NETMIS data with the dwell time is calculated. If this data is resampled at a 2 minute frequency, with the arrival time as the resample reference point and the mean calculated, the data will reduce and look like table 1.

[**INSERT TABLE 1]**

3.1 Delays on the Tube from CUPID

London Underground also gathers data on delays. Delays are defined as service affecting failures (SAFs) which last two minutes or more. This includes escalator failures, lift failures, partial station closures, full station closures, platform closures, train delays and line suspensions. All SAFs of 2 minutes or more are recorded in a centralised database known as CUPID. In this study the SAF information used is from 2009 to present. Along with a description of the incident, the length of the delay, the delay duration, a basic root cause analysis and a calculation of lost customer hours is included. Each incident also contains data on location, time, train number (if applicable), line and direction (if applicable). This large database on reliability will be included as part of the dwell time study to investigate any correlation between the number of SAFs on a line and the dwell time. The primary disadvantage of CUPID data is that it does not contain data for service affecting failures less than two minutes. Additionally there is likely human error introduced in the recording of these incidents, which is done remotely at a control centre shortly after the incident occurs. Statistics based on this data can be calculated, such as the mean number of SAFs per year for each line and the mean duration of each SAF. The dataset is labelled by a combination of Line and direction for allocation of the SAF rate.

Processing the number of SAFs on each line direction combination yields the following summary statistics in table. Line availability was calculated by assuming that unavailability is defined by any train delay that affects the service. The SAFs in table are only train delay SAFs and do not include the number of service affecting failures for other categories, such as station closures or escalator failures. This is because train delay SAFs are directly relevant to the service. The SAF data can be seen in table 2.

[INSERT TABLE 2]

The "failure rate" instead of "line availability" data will be used in this study. It could be argued that Line Availability would be more appropriate, further analysis (scatter plot figure 1), data proves that the failure rate is strongly correlated with line availability, therefore failure data would be appropriate.

[INSERT FIGURE 1]

This study is concerned primarily with the "good service" as the performance indicator. Attempt is made to understand the factors which drive performance into the sub-2 minute region. The frequency of failures directly or indirectly affects the dwell time. Some failures at the system level such as station closures, platform closures and line suspensions results are excluded as they are not directly relevant to the train service and therefore dwell time. The overall train service affecting failure rates for each underground system are listed in table 3.

[INSERT TABLE 3]

3.2 Passenger Data from Rolling Origins and Destination Surveys (RODS)

In order to investigate the number of passengers on the network, and compare this data with dwell times, RODS[1] are used in this study. These surveys, which take place yearly each November across the network, involve a combination of counting of passengers boarding and alighting trains at each station along with an analysis of Oyster Card entry and exits count. The result is an estimate that can be provided network-wide for the number of boarders and alighters at each location. This data is broken down into 15 minute slots throughout the day to provide a high degree of granularity.

Looking at the data up close at a single location is revealing. Taking Holborn as an example, there are high fluctuations in the number of boarders and alighters throughout the weekday as can be seen in figure 2.

[1] https://data.london.gov.uk/dataset/tfl-rolling-origin-and-destination-survey

[INSERT FIGURE 2]

The mean dwell time for each 15 minute time period throughout the day is strongly dependent on the total number of boarders and alighters (figure 3).

[INSERT FIGURE 3]

3.3 Rolling Stock Technical Specifications

Primary data from internal London Underground rolling stock technical specifications provides information on the internal layout of each train. Data on the number of seating spaces in each train, along with the floor space capacity of each train was captured. This was done for all lines except the District Line. At the time of this investigation District Line runs two different rolling stock types, “S-stock” and “D-stock”, thus making it difficult to differentiate in the dwell time data which train is being represented. Therefore the District Line is excluded from this study. Calculations representing the number of trains on the line relative to the length of the line and the number of stations for each line were made. This variable will be referred to as *“train density”* and is calculated by equation 2.

$$TrainDensity = NTrainsOnLine/NStationsOnLine/LineLengthKM \quad (2)$$

Table 4 presents the train density.

[INSERT TABLE 4]

3.4 Parameter Estimation

Goodness-of-fit tests were conducted to estimate the parameters of 84 (SciPy 2017b) different distributions against 30 different samples of dwell time data from across the London Underground Network. In this study the maximum likelihood estimation (MLE) method the will be used for estimating share, scale and location parameters for the data. The MLE method has the advantage of being quite suitable for large datasets. As the sample size grows larger the parameter estimates converge to the correct value. The goodness of fit is calculated by maximizing a log-likelihood function, with penalty applied for samples outside of range of the distribution. The parameters for each distribution were first estimated using maximum likelihood estimation (Fisher, 1925) (Meeker & Escobar, 1998). A probability plot was then generated comparing the sample data with quantiles of the estimated theoretical distribution, with a linear coefficient of determination (R-squared) value calculated to determine suitability of the estimated distribution (Chambers, Cleveland, Kleiner, & Tukey, 1983). The coefficient of determination provides an indication of how well the estimated distribution matches the sample data.

3.5 Data Selection and Experimental Design

The process of data selection and experimental design included 30 different scenarios. Since dwell time behaviour may be different at different locations and times of the day, these scenarios were selected in order to try and achieve a balanced set of samples. Eight lines were selected, two stations from each line were selected: one in “zone 1” and one in “zone 4” (with the exception of the Victoria Line for which the furthest from Central London station is in “zone 3”. For each station data was acquired in a single direction from the NETMIS database. Peak and off-peak data was acquired from the NETMIS database with peak data being defined as 07:00 to 09:00 and off-peak data being defined as 12:00 to 14:00.

A method for filtering out unusual dwell times is to filter out dwell times over a certain duration. Since it is known from previous studies that the number of boarders and alighters correlates with the dwell times, it can be hypothesised that the optimum cut-off point for filtering the dwell times can be found by calculating the correlation between the number of boarders and alighters and the dwell time for a range of cut-off points. The results of this exercise are below in figure 4. Note that this is with raw data that has not been resampled.

[INSERT FIGURE 4]

It was found that the Pearson correlation between the number of boarders and alighters and the dwell time peaks when the data is filtered for a cut-off point of 45 seconds. This suggests that boarding and alighting data does not generalise well to unusual dwell times, but does correlate well with normal-service dwell times of up to 45 seconds. Further filtering was applied by explicit selection of locations that would not yield data relevant to the goals of this study. For example, Termini (e.g. Uxbridge), Flat junctions (e.g. Baker Street), Reversing moves (e.g. Queen's Park and Rayners Lane), and stations with route in/out of depots (e.g. Northfields) were excluded in the study. At occasions there are irregular activities that causes non-parametric dwell time behaviour, such special cases were also filtered out to eliminate bias of the generalised model.

In order to find the optimal resampling period and with respect to maximising the signal while keeping the number of data points as high as possible, different resample periods and measure the correlation between boarders and alighters and dwell time were experimented. Since it is known that a relationship exists between these two variables, a stronger correlation should indicate a better dataset with an improved signal to noise ratio. The spearman correlation coefficient between the dwell time and the number of boarders and alighters was calculated for each resample period. The Spearman correlation coefficient was used since there is no evidence that the dwell time statistics linearly change with influencing factors (figure 5). Furthermore, subsequent modelling of the dwell time can take into account non-linear factors which the Pearson coefficient may discount. Based on the experiment 60min was chosen as the resampling time.

[INSERT FIGURE 5]

3.6 Data Analytics and Modelling

Based on the nature of the problem and the quality of the historical data regression analysis seems to be most promising method to link capacity planning and dwell time. The choice of analytical method, in this case regression techniques, will depend on a number of factors including complexity and linearity of the relationships. Ultimately the decision on which model to use will depend on the results of the validation.

Regression analysis is the process of estimating the relationship between different variables. Many forms of regression analysis exist and a number of methods have been tested in this study. For example, Multiple Linear Regression (MLR) is one of the most popular and simplest ways to project results in the form of:

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \ldots + c \tag{3}$$

Where y is the dependent variable in this instance mean dwell time, b_i is the constant associated with the factor x_j.

Decision trees are an another regression model described in (Quinlan, 1986). The method requires minimal data preparation such as converting the data into a standard normal distribution, however it is documented that decision trees are not as accurate as other approaches and they are not very robust, that is a small change to the data can yield big changes to the model (Gareth, Witten, Hastie, & Tibshirani, 2015). Since decision trees use Boolean logic to determine the structure of the tree, the method of modelling can be described as "white box", i.e. it is possible to explain what is happening. By contrast methods such as neural network regression are "black box" as studying its structure will not provide any insights into how the function is being generated (Dreiseitl & Ohno-Machado, 2002). There are also some disadvantages to using decision tree regression. The models created by decision tree learning algorithms can create overly complicated decision trees (Bramer, 2007). This may lead to a model that does not generalise very well. For example, it is possible to use a decision tree algorithm to fit the number of boarders and alighters to dwell time in the dataset (figure 6).

[INSERT FIGURE 6]

Figure 6 shows an R-squared value of 0.774. However it is clear that the algorithm is overfitting the model to the data, thus making it poor for generalization. This is a recognised problem with decision trees (SciKit-Learn 2016c).

The K-nearest neighbour algorithm is a commonly used method for regression analysis. This method works by determining a regression based on the values of the *k* nearest data points (Altman, 1992). The *k* nearest methods investigated are uniform weights and the inverse distance weights. Along with these sub-methods, the key determinant of the performance of this model is based on the choice of the number for *k*. Thus, this will also be varied.

The multi-layer perceptron (MLP) neural networks is a powerful tool for regression analysis. The MLP is a series of regression models which are organised on top of one another. The neural network contains a number of hidden layers and hidden units within each layer. The purpose of these units is to model non-linear behaviour. Each of the units in the neural network has an activation function associated with it. There are a number of functions that can be used for the activation function such as the logistic sigmoid function and the rectified linear unit function. The sklearn algorithm for the MLP neural network attempts to optimise the squared loss in order to find the best model. A MLP neural network machine learning regression model was applied to the data. A stochastic gradient descent optimizer ('Adam') was utilised. This is a modification to the standard stochastic gradient descent optimizer and has been shown to work well on large datasets with improved model fitting speed and validation accuracy (Kingma and Ba 2014). The rectified linear unit function was chosen. The number of units per layer was varied from 2 to 50.

3.7 Model Validation and Verification

The R^2 value provides an indication of the model fit. However this simply reveals how well the model fits the sample dataset. As discussed, and particularly with more complex non-parametric methods, it is quite possible to create models that produce high R^2 scores and suggesting that they are good models when they are, actually, 'overfitting'. An overfitted model is unlikely to generalise well to other data. The data is therefore split into training and test sets. The training set is used for creating the regression model. The test set is used for testing the performance of the new regression model. The best performing model on the test set is then selected. For the purposes of this study 80% of the data will be partitioned for the training set and 20% is partitioned for the test set based on guidance by Murphy (Murphy 2012). The order of the dataset is randomised before being partitioned. In order to evaluate the performance of the model on the validation set, the Pearson correlation coefficient is calculated between the true mean dwell times in the test set and the predicted mean dwell times from the model. This method can also be visualised.

4. CAPACITY SIMULATION CASE STUDY

The Victoria Line makes a good candidate option for modelling capacity in this study due to its relative simplicity and recent uplift of performance to 36 trains per hour (Transport for London, 2017). A stochastic discrete-event simulation of the Victoria Line is constructed to assess and forecast capacity.

4.1 Simulation Model Logic

The model includes each station on the Victoria Line Northbound service. Each station will be modelled in a chain and will contain a full speed run-in time, a dwell time and a run-out time. The inter-station journey time will be excluded from this model since journey time is not of interest, only capacity. Run in and run out times were modelled using Railway Engineering Simulator.

Improvements in capacity can be found by utilising modern signalling systems that allow trains to run closer together in through running sections. However this only holds true if there are not bigger constraints that exist at stations. For single platforms in a series-like railway design, the constraint on the system will be found around the run in time, the dwell time and the run out time from platforms. Multiple platforms could in theory be one technical solution to removing constraints at stations. Simulating run-in, dwell and run out times seem to be suitable method of predicting capacity. Note that in this context capacity is independent from journey time. The bottlenecks which determine the maximum available capacity are found at the stations.

The dwell time mean is calculated per location from the multiple linear regression model. A gamma distribution was assumed a good fit for the dwell time with mean of 30s and standard deviation of 7.8s. It was shown to be the most generalisable distribution across the range of stations and times tested (see section 3). The dwell standard deviation σ_{dwell} is assumed to be a constant 7.8s and the variance 60.84s. The Gamma distribution takes two parameters: shape k and scale θ. Given the mean dwell μ_{dwell} and variance σ^2_{dwell} these parameters can be calculated as follows:

$$k = \mu^2_{dwell}/\sigma^2_{dwell} \tag{4}$$

$$\theta = \sigma^2_{dwell}/\mu_{dwell} \tag{5}$$

Thus, each location on the Victoria Line samples dwell times form a unique gamma distribution. Each station included in the model contains a RORIF (run-in and run-out time at full speed) and an associated number of passenger numbers (table 5).

[INSERT TABLE 5]

The simulation time has been set to 1,000,000 seconds. This is the equivalent of 277.78 hours of continuous service. At 34 trains per hour this over 9000 trains in the simulation. This ensures that the summary statistics in the results are statistically significant (95% confidence interval). The Service Affecting Failure rate was assumed a constant value of 6.844 service affecting failures per day. Figure 7 shows the simulation logic used to represent the run in and run out from the station.

[INSERT FIGURE 7]

The platform resource has a capacity of one and controls the movement of trains in and out of the station. Resources are seized on a first come first serve basis. If the resource is not available then the request for queues.

4.2 The Operational Recovery Time

Operational recovery can be described as the ability of the system to absorb delays - in other words the amount of *resilience* which the system has. In the London Underground, operational recovery is the amount of operational slack that exists in the process of running trains in and out from platforms. It is used to add resilience to a network and reduce the severity of the impact of service affecting failures. A 10 second recovery time at a station means that a train can take an extra 10 seconds to complete its run in, dwell and run out, without unduly affecting the timetable. Recovery can be thought of as scheduled extra dwell time, which is sometimes not used if the service needs to catch up in the case of a delay, or else can be used to absorb variations in dwell time or run in and run out time.

It is difficult to be precise when stating what operational recovery time needs to be built into the railway system. Increasing recovery time leads to a reduced frequency of service, but the resilience and recoverability of the service improves. To the best knowledge of the authors there is no study available that has successfully quantified the benefits and drawbacks. Instead, the amount of operational recovery built into the system is based on heuristic beliefs about the time. Generally a 10 second operational recovery time is built into the process of building London Underground timetables. The history of the 10 second heuristic and why it was chosen is not recorded in any literature. Based on the simulation results, the operational recovery time can be computed as:

$$OperationalRecovery = MeanMeasuredHeadway - MaxTargetHeadway$$

(6)

Similarly given a fixed operational recovery it is possible to predict the maximum capacity of the system.

$$MaxTargetHeadway = MeanMeasuredHeadway - OperationalRecovery \qquad (7)$$

Thus simulation presents the opportunity to frame recovery and the relationship to capacity in two lights: (1) the operational recovery can be varied to observe what TPH can be achieved, or (2) the target TPH can be varied to observe what operational recovery is needed to achieve the target TPH. For the purposes of this simulation, operational recovery is fixed at 8s and the mean headway achieved will be observed. The 8s was chosen as a heuristic operational recovery because with 2015 population the results produced a mean capacity of greater than 36 trains per hour – reflecting the true service rate produced today.

4.3 London Population Growth

As a case study this discrete event simulation will seek to model the capacity of the Victoria Line and investigate the effects of population growth on that capacity. The population forecasts in London up to 2050 are provided by the Greater London Authority, a projected gradual growth will increase London's population by 27.45% (i.e. reach 11,069,000) (see GLA 2015 for detailed annual rate). This simulation assumes that the number of boarders and alighters at each station on the Victoria Line will increase linearly with the population growth in London. Thus, the simulation will be able to test and provide a capacity forecast for each year up to 2050. A population multiplier is included in the simulation, which for a linear relationship is set to 1. This population multiplier variable can be used to limit the effect of the increase in population on boarders and alighters. This could be useful if finer predictions could be made about which stations on the Victoria line will experience the uplift in boarders and alighters. Since these data do not exist the relationship is assumed to be linear. The study also assumes a fixed recovery time of 8 seconds. The mean output headway will be calculated from the simulation results and a trains per hour figure will be based on this.

$$AchievableTrainsPerHour = MeanOutputHeadway/3600 \qquad (8)$$

5. ANALYSIS OF RESULTS AND FINDINGS

5.1 Data Distribution and Dependency Tests

The mean R^2 score across the 30 scenarios for each distribution was calculated. The Gamma distribution was found to be, on average, the best estimate for the dwell time (goodness of fit) with a mean R^2 value of 0.85. This means that by using the Gamma distribution, on average it can be expected that 85% of the variation in the data can be captured by the distribution. Prior to data analysis, dwell times of greater than 45 seconds were excluded from the analysis based on the peak correlation discovered at a cut-off point of 45 seconds. The dataset was resampled at a 60 minute frequency and the means for dwell time and passenger number data were calculated for each resampled period. Figure 8a shows the distribution of mean dwell times in the dataset. The mean of the mean dwell times is 25.28s with a standard deviation of 5.19s. The distribution of the mean dwell times appears to be normally distributed. The distribution of the number of boarders and alighters across the network is shown in figure 8b. The mean number of boarders and alighters in the dataset is 336.83 with a standard deviation of 477.21. The data appears to be exponentially distributed.

[INSERT FIGURE 8]

A dependency test of the resampled mean dwell times on the number of boarders and alighters demonstrates positive correlation where with increasing numbers of boarders and alighters dwell times increase. The calculated Spearman correlation was 0.6 with a p-value of 0 (Figure 9a).

Figure 9b shows the distribution of the Service Affecting Failures (SAF) rates. SAF rates are assigned at a line level, there are six values for the data. The mean SAF rate is 10.83 failures per day for a line with a standard deviation of 2.79, a goodness of fit test shows that SAF rate follows a uniform distribution *UNIF(6.39, 14.29)*. The dependency of mean dwell time on SAF rate shows positive correlation between the SAF rate and the dwell time with a Pearson correlation coefficient of 0.27 and a p-value of 0 (Figure 9b).

The distribution of standing capacity across the different lines was calculated as 133.4m2 and the standard deviation of 33.21. No particular statistical distribution could be observed, but the observation shows the standing capacity is heavily skewed towards the minimum. The dependency of dwell time on standing capacity shows a small but statistically significant negative correlation, with increased standing capacity correlating with a slight reduction in dwell time. The Pearson correlation coefficient was calculated at -0.027 with a p-value of 5.6e-7 (Figure 9c).

The mean seating capacity is 243.13 with a standard deviation of 33.21. The dependency of dwell time on seating capacity shows is a small but statistically significant negative correlation. With increased seating capacity correlating with a slight reduction in dwell time. The Pearson correlation coefficient was calculated at -0.19 with a p-value of 8.1e-273 (Figure 9d).

The dependency of dwell time on the train density statistic shows a small but statistically significant negative correlation. The increased train density correlates with a slight reduction in dwell time. The Pearson correlation coefficient was calculated at -0.0077 with a p-value of 7.7e-45 (Figure 9e).

[INSERT FIGURE 9]

Both linear correlation metrics were calculated between each possible factor and the mean dwell times. All results are statistically significant to 99% confidence levels due to the sample sized used.

Table 6 presents Spearman and Pearson correlation scores between tested factors and mean dwell time.

[INSERT TABLE 6]

5.2 Regression Analysis Results

The R^2 coefficient of determination is calculated for each model. This value provides a score for the proportion of variance in the dwell time following the trend of input variables and the extracted model. A number of regression models were tested.

A. Multiple Linear Regression

Equation 9 represents the multiple linear regression model derived:

$$y = 0.0063x_1 + 0.4124x_2 - 0.0189x_3 + 0.0008x_4 - 0.0064x_5 + 23.646 \tag{9}$$

Where: y = Dwell Mean, x_1 = Total Boarders and Alighters, x_2 = SAF Rate, x_3 = Seating Capacity, x_4 = Standing Capacity, and x_5 = Train Density.

The R^2 value for this multiple linear regression fit was calculated to be 0.434 with the validation correlation score found to be 0.658.

B. Decision Tree Regression

With varying maximum depth, one can observe the plotted dependency of R^2 and validation correlation on the decision Tree Depth (Figure 10).

[INSERT FIGURE 10]

The R^2 value for the decision tree regression model was maximised at 0.89 with a tree depth of 30. However, this resulted in a model that was overfitted as the validation score was lowest at this point with a correlation of 0.63. The best model had a validation correlation coefficient of 0.73 between the test dwell means and the predicted dwell means. The decision tree depth for this model was 7. The R^2 value of this model was found to be 0.52.

C. The K-nearest Neighbours Regression

The default parameters for the model were used from SciKit learn library (SciKit-Learn 2016b). This resulted in a leaf size of 30 using the minkowski method with a power of 2. Uniform weights were tested as well as inverse distance weights. The number of neighbours was varied in order to find the optimal R^2 value and the results shown in figure 11.

[INSERT FIGURE 11]

The R^2 value for the decision tree regression model was maximised at 0.92 using the inverse distance weights method with 6 neighbours. However this has clearly produced a model that is overfitted to the data since the validation score is 0.68. This validation score improves with a greater number of neighbours and by changing the method to uniform weights. The uniform weights model produces the best validation correlation score of 0.71 with 18 neighbours in the model and with the R^2 value of 0.54.

D. Neural Network Regression

The number of hidden layers was varied from 10 to 200 and the results plotted below in figure 12.

[INSERT FIGURE 12]

The neural network model validation correlation score is maximised at 0.7 at several points. The smallest hidden layer size where this score is maximised is found with 30 hidden layers. The R^2 value with this model was found to be 0.48.

In summary, the R^2 tests and the corresponding scores provide a measure of how well the model fits the training data. The validation correlation score provides a measure of how generalisable and thus valid the model is. A summary of regression models tested is presented in table 7.

[INSERT TABLE 7]

E. Victoria Line Simulation Case Study Results

The mean headway at each station was calculated and an equivalent trains per hour (TPH) value was extracted from this. Since each year up to 2050 was tested, an equivalent forecast for the capacity of the line was made. The results can be seen in figure 13.

[INSERT FIGURE 13]

Figure 13 - Predicted Victoria Line achievable capacity as London population increases

The TPH results show a steady decrease in achievable capacity on the line as the population increases based on the assumption of a linear increase in boarding and alighting numbers with population. With 2015 population levels and therefore 2015 numbers of boarders and alighters the breakdown of capacity at each station can be seen in figure 14.

[INSERT FIGURE 14]

The maximum RORIF + recovery margin + mean dwell time is 98.5s at Oxford Circus, 98.4s at Stockwell and 96.6 at Victoria. This indicates that Oxford Circus is the bottleneck site to limiting capacity to 98.5 second headways or 36.55 TPH. With the 2050 population forecast and assuming that this translates linearly to an increase in passenger numbers at each station, the individual station results is presented in figure 15.

[INSERT FIGURE 15]

Oxford Circus had a total limiting time of 105.5s, Stockwell had 104.4s and Victoria 102.6. This indicates that Oxford Circus provides the bottleneck limiting capacity to 105.5 second headways or 34.12 TPH.

6. CONCLUSIONS

The consideration of dwell times is important for the planning and optimisation of railway transport systems. This study has presented new evidence on the distribution which a dwell time follows for through running stations. This estimated distribution can be utilised in discrete-event simulation, for the purpose of capacity planning and troubleshooting. This model and methodology is generalisable to other similar railway transport systems. Transport systems which differ significantly from the London Underground may require modelling of the dwell using data from those systems, however the methodology can remain consistent.

The study presented reveals that the dwell times on the London Underground network most commonly follow a gamma distribution. It is recommended that the gamma distribution is employed in simulations where the dwell time needs to be sampled from a continuous probability distribution.

A number of factors were tested and found to have a statistically significant correlation with the mean dwell time. The total number of boarders and alighters was found to have the strongest correlation with mean dwell time. The service affecting failure rate for the line positively correlates with the mean dwell time, likely due to the knock on effect of delays on the line causing mean dwell times to increase while the service recovers. The number of trains on the line relative to the number of stations and the line length, was also found to negatively correlate with mean dwell times. This is hypothesised to be due to the psychological effect of having more frequent train services on dwell time. The seating capacity and standing capacity of the train negatively correlates with mean dwell times, with seating capacity being a stronger factor. This may help in future design of trains.

One significant limitation to this study is that the relationship, if any, between the factors and the variance of the dwell time has not been investigated. It could be the case that the factors identified and investigated positively or negatively correlate with the dwell time variance. For the purposes of this study, in particularly evident in the discrete event simulation of the Victoria Line, it has been assumed that variance of the dwell time remains constant (stationary covariance). Future studies could investigate the effect of factors on dwell time variance. This would be an important investigation from the perspective of reliability, as a greater variance in dwell times is equivalent to a less dependable service for customers.

This study has demonstrated the method of testing several regression models and using the Pearson correlation coefficient to test predicted values against a partitioned validation dataset. The results indicate that the K-nearest neighbours regression model is the best performing algorithm for forecasting mean dwell times. Neural network decision tree models were also tested. The decision tree models performed the worst and the neural network models performed well but not as well as the K-nearest neighbours model. A simple linear regression can also be employed with less power than K-nearest neighbours and neural networks models but with more power than the decision tree network. The advantage of using the simple linear regression approach is found in the simplicity of the model to communicate with practitioners.

The capacity of the Victoria Line was assessed assuming a linear relationship between the number of boarders and alighters and the population change of London forecast up to 2050. The multiple linear regression model established in this study was employed. Given these assumptions the peak capacity constraint on the Victoria Line was shown to decrease from 36 to 34 trains per hour with an 8 second recovery margin built into the RORIF time. Mitigating this loss in capacity could involve reducing the dwell time or improving the RORIFs at the bottleneck locations. The primary limitation of this study is the assumption of the linear increase in boarders and alighters with the London population.

NOTE

Compliance with ethical standards

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

REFERENCES

Altman, N.. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician , 46* (3), 175–185 (1992).

Bramer, M. *Principles of Data Mining.* London: Springer. (2007)

Chambers, J., Cleveland, W., Kleiner, B., & Tukey, P. *Graphical Methods for Data Analysis.* Wadsworth (1983).

Community of Metros and Imperial College London. *Dwell Time Recalibration.* London: Imperial College London (2013).

Dreiseitl, S., & Ohno-Machado, L. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics , 35* (5-6), 352-359 (2002).

Fisher, R. A. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society , 22*, 700-725 (1925).

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. *An Introduction to Statistical Learning.* New York: Springer (2015).

Goodwin, T. *Changing Behaviours on the London Underground: A Journey into Dwells.* London: University College London (2015).

Greater London Authority. *2015 Round Population Projections.* Retrieved 2017 from https://files.datapress.com/london/dataset/2015-round-population-projections/2016-10-21T14:23:54/long_term_trend_2015_round.xlsx (2015).

Greater London Authority.. *Economic Evidence Base - Summary Version.* Retrieved January 22, 2018 from https://www.london.gov.uk/sites/default/files/gla_migrate_files_destination/2evidence-base-2010-summary.pdf (2010, May).

Holloway, C., Roan, T., & Tyler, N. *New Deep Tube Train: Design Features Affecting Boarding and Alighting of Passengers.* London: University College London (2013).

Kruskal, W. H. Ordinal Measures of Association. *Journal of the American Statistical Association , 53* (284), 814–861(1958).

Meeker, W., & Escobar, L. A. *Statistical Methods for Reliability Data* (1st ed.). John Wiley & Sons (1998).

Murphy, K. P. *Machine Learning: A Probabilistic Perspective.* London, England: The MIT Press (2012).

Ox, J., & Goldratt, E. *The Goal: A Process of Ongoing Improvement.* Croton-on-Hudson: North River Press (1986).

Quinlan, J. R. *Induction of Decision Trees. Machine Learning* (1st Edition ed.). Kluwer Academic Publishers (1986).

Rencher, A. C., & Christensen, W. F. *Methods of Multivariate Analysis* (3rd Edition ed., Vol. Wiley Series in Probability and Statistics). John Wiley & Sons (2012).

scikit-learn developers. *Decision Tree Regression*. Retrieved December 13, 2017 from SciKit Learn: http://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html (2017).

scikit-learn developers. *sklearn.linear_model.LinearRegression*. Retrieved December 13, 2017 from SciKit Learn: http://scikit-Learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression (2017).

scikit-learn developers. *sklearn.neighbors.KNeighborsRegressor*. Retrieved December 13, 2017 from SciKit Learn: http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html#sklearn.neighbors.KNeighborsRegressor (2017).

Team SimPy. *Overview of SimPy*. From SimPy Website: https://simpy.readthedocs.io/en/latest/

Transport for London. (2018). *Budgfet 2017/18.* London (2017).

Transport for London. *TfL.gov.uk.* Retrieved December 13, 2017 from London Underground World Class Capacity Sub Programme Review (corrected version): http://content.tfl.gov.uk/pic-20170628-item07-lu-world-class-capacity.pdf (2017).

Wong, H., & Key, A. *Tapir Phase 1 Report.* London: Transport for London (2014).

Zio, E. *The Monte Carlo Simulation Method for System Reliability and Risk Analysis.* London: Springer (2013).