Article:

Hickish, R., Fletcher, D. orcid.org/0000-0002-1562-4655 and Harrison, R. orcid.org/0000-0002-9323-8637 (2019) A rail network performance metric to capture passenger experience. Journal of Rail Transport Planning and Management. ISSN 2210-9706

https://doi.org/10.1016/j.jrtpm.2019.06.002

A rail network performance metric to capture passenger experience

Bob Hickish[1], David I Fletcher[1*] and Robert F Harrison[2]

[1]Department of Mechanical Engineering, The University of Sheffield, UK

[2]Department of Automatic Control and Systems Engineering, The University of Sheffield, UK

*Corresponding author: d.i.fletcher@sheffield.ac.uk, Department of Mechanical Engineering, The University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK

Declarations of interest: none

Abstract

For passenger rail operators worldwide a common concern is to better understand and improve passenger experience. Based on factors including train movement times and crowding, the Journey Time Metric and Disutility Metric can be used to quantitatively assess the journey experience of individual passengers. However an assessment of overall network performance is also desirable. This paper presents a whole-network assessment metric that captures passenger experience by aggregating and normalizing individual journey assessments. The newly developed metric is validated against customer satisfaction data measured in passenger surveys of the London Underground Limited Victoria Line with a statistically significant correlation ($P<0.005$) between the predictions and the measurements. It is found that there is a high degree of correlation ($\rho=1.00$, $P<0.005$) between the network scores calculated using the new whole-network assessment metric with either the Journey Time Metric or Disutility Metric despite their different formulations and countries of origin. Through development of the new metric it is identified that many commonly used network assessment metrics (e.g. Public Performance Measure and the end-to-end journey time of passengers) are insensitive to crucial aspects of passenger experience. The newly developed metric could be used by rail operators to better select strategies for improving passenger experience.

Keywords

Journey Time Metric, Disutility Metric, rail, network, passenger, assessment

Highlights

- A new whole-network assessment metric is developed to capture passenger experience by aggregating and normalizing individual journey assessments.
- Two different passenger journey assessment metrics of different international origin are compared.
- The new whole-network assessment metric is validated against measured data from the London Underground Limited Victoria Line.

Nomenclature

ψ – individual passenger journey score*
Ψ – distribution of passenger journey scores
ϕ – network score*
I – number of states in a passenger journey
i – counter for enumerating sequence of states
t_i – time passenger spends in their i^{th} state (seconds)
Ω – Value of Time weighting function*
α_i – passenger journey stage of i^{th} state*
β_i – vector describing conditions of passenger's i^{th} state*
ε – number of passenger train changes
ω – crowding penalty function*
δ – number of passengers on train
δ_{max} – train maximum capacity
γ –train crush capacity
μ – number of seats
η – crowding factor
c_1 to c_3 – constants
k_1 to k_7 – constants
R – number of passengers
d_r – distance travelled by r^{th} passenger
τ_B – Kendall Rank Correlation Coefficient B*

* Values specific to a metric are indicated with the superscript text: *JTM* or *DM*

1 Introduction

With demand for rail travel having doubled in the last 20 years (Davis, 2018) and 40% more passengers predicted by 2040 (Carne, 2018), rail travel has an increasingly important role to play in meeting the passenger journey needs of Great Britain (GB). To fulfil this role the rail industry Technical Leadership Group (2017) set targets for the GB network that included "improving [the] customer experience" of passengers. The experience of passengers and their satisfaction is also a network performance indicator for other rail networks internationally, e.g. across Europe (TNS Political & Social, 2013) and in Japan (Kunimatsu et al., 2012). Traditionally, however, rail networks have been assessed with train-focussed metrics. For example, the GB industry standard Public Performance Measure (PPM) describes the percentage of services that arrive at their *final* destination within five minutes (ten for long distance trains) of the timetabled time, this metric having no sensitivity to the effect on passengers if the train arrives late at intermediate stations, or to the comfort of their journey. In this paper a new method is developed which combines assessments of individual passenger journeys, i.e. *journey scores*, for all passengers in a network to give a *network score* that quantifies the experience of passengers. In a case study relating to the Victoria Line of the London Underground Limited (LUL) network, the whole-network assessment metric is validated against measured data from passenger surveys surmised by LUL (2018a). Furthermore, international comparison is made when the whole-network assessment metric is used with individual passenger journey assessment metrics from different countries of origin. The developed whole-network assessment metric will allow operators to provide a parameter summarising overall network performance from the passenger perspective, enabling this to be effectively optimised.

2 Metrics to assess networks

The aggregate of passenger end-to-end journey time has been used as a metric to assess network performance, for example by Vuchic and Newell (1968), Chang et al. (2000) and Cacchiani and Toth (2012). However, there is evidence that end-to-end journey time does not fully capture the passenger experience. For example, Susilo and Cats (2014) show that, for public transport travellers, factors such as station environment, ease of transfer, service frequency and safety are significant determinants of passenger satisfaction. Because Chen and Chen (2010) describe customer satisfaction as being affected by customer experience, in the current paper it is assumed that the satisfaction of a passenger is an indicator of their experience, and the effect of other factors such as ticket pricing is disregarded. Consequently, in the current paper, decreasing passenger dissatisfaction or disutility and increasing passenger satisfaction are considered to be equivalent to "improving passenger experience". The disconnect between passenger journey time and passenger satisfaction is evident in the results of a rail passenger survey by Transport Focus (2016) which showed that journey time has a smaller influence upon passenger satisfaction than punctuality of the service or cleanliness. Therefore, to better capture passenger satisfaction it is necessary to quantify a passenger journey in greater depth than journey time or punctuality alone.

2.1 Describing a passenger journey with stages

A passenger journey can be modelled as the combination and repetition of specific activities, i.e. *stages*. For example, Wang et al. (2015) state that a passenger journey can be well represented with the stages: walking into and out of a station, waiting on the platform, riding on a train and transferring between platforms. However, they do not take into account the relative impact of time in each stage upon the whole passenger experience. Vansteenwegen and Van Oudheusden (2007) and Sels et al. (2016) describe a passenger journey using two stages ("In Station" and "On Train") and capture the varying impact of time in different stages by weighting these times with a different Value of Time (VoT). The VoT concept has been developed in Transport Economics and describes, in monetary terms, the disutility experienced by a passenger over a time period. It can be thought of as the price a passenger would pay to reduce their travel time by one unit, hence a greater VoT indicates a worse experience for passengers. As well as being sensitive to the journey stage of a passenger, a VoT can be sensitive to the mode of transport, journey purpose and distance, for example having different values for travel by car, bus, train or other public transport (ARUP et al., 2015). Wardman (2004) showed that the VoT is sensitive to the activity of the passenger, and Vansteenwegen and Van Oudheusden provide values showing that passengers rate 1 minute of waiting in a station to be equivalent to 2.5 minutes on a moving train. By modelling the amount of time passengers spend in both of these stages and weighting it by the VoT for each stage, Vansteenwegen and Van Oudheusden create a network assessment metric which can capture the relative effect on passengers of time savings in either stage. However, their metric does not capture the effect of *crowding* (i.e. the number of passengers on a train relative to the number of seats and standing space) which can reduce the personal space and comfort of passengers, causing additional disutility and hence increasing the VoT.

Horowitz (1978) showed that, as well as the journey stage, the "environmental conditions" that a passenger experiences during a stage (referred to as *conditions* in the current paper) affect the VoT. Horowitz considered weather conditions, that are not considered here, but also standing vs seated travel and crowding levels. Models to quantify the effect that crowding has upon the VoT have been developed for example by

Wardman and Murphy (2015) and Qin (2014). Two metrics developed in different international systems to assess individual passenger journeys across journey stages and crowding levels are the Journey Time Metric (JTM) and the Disutility Metric (DM).

The JTM has been developed by LUL and shared with the authors by private communication, the most informative accessible documentation being the investigations of Chan (2007) and Hickey (2011). It describes passenger journeys using five stages “Buying Ticket”, “Moving Through Station”, “On Platform”, “On Platform (Left Behind)” (where a passenger has not been able to board a suitable train because it is overly occupied) and “On Train”. The effect of crowding conditions are considered in the “On Train” stage by modifying the VoT with a *crowding penalty* that is dependent on the number of passengers, train capacity and seats. The DM has been developed in Japan and is documented in English by Kunimatsu et al. (2009, 2012). It takes a similar approach to the JTM, but resolves a journey using two stages (“On Train” and “In Station”) with weightings different to those used by the JTM. Similar to the JTM, the DM applies a crowding penalty for passengers in the “On Train” stage that is sensitive to the same factors as the JTM crowding penalty, however a different formula is used. The DM is used again by Kanai et al. (2011) to assess individual journeys as part of a network assessment metric used in a decision support tool for delay management. They discuss different methods of combining journey scores into a network score, however none of their methods normalize for the distance travelled by passengers, meaning that networks providing shorter journeys could compare favourably against networks providing longer journeys.

Moving from individual journey to network metrics, Ali et al. (2017) predict network performance by combining journey scores calculated using an individual journey metric with similarities to the JTM and DM. The network metric is demonstrated to predict observed simple qualitative relationships between timetable features and network performance, e.g. fewer train services result in worse network performance as determined by their metric.

The JTM, DM and the metric described by Ali et al. are the only metrics, found for this review, to capture the multi-stage nature of passenger journeys and weight the time spent in each stage *including* the effect of crowding. They therefore capture individual passenger journeys in more detail than the other metrics identified here which consider journey stages *or* crowding only. However, the parameter values used within the metric of Ali et al. could not be retrieved so this is excluded from further analysis. To the best of the authors' knowledge, no publicly available documents describe the validation or comparison of the JTM and DM, or network assessment metrics based upon them. This gap defines the targets of this paper, to make a comparison of the JTM and DM methods, and to develop a validated network metric based upon them.

3 Network assessment metrics that capture the passenger perspective

To assess a rail network we evaluate individual passenger journeys and examine the distribution of experiences. To evaluate modelled passenger journeys, we introduce the term *state* to describe a specific combination of journey stage and conditions. A passenger journey is decomposed into a sequence of states as shown in Figure 1, which illustrates an example four-state passenger journey. Shading is used to indicate which journey stage the passenger is in ("On Train" or "In Station"). Crowding is only considered in the "On Train" stage and text is used to indicate this. The markers t_0 to t_4 indicate the times at which the passenger changed state. At t_0, the passenger enters the origin station, at t_1 the passenger boards their train. At t_2 the train stops at an intermediate station where more passengers board making it crowded. The passenger journey *stage* does not change, but the *state* does. At t_3 the passenger reaches their destination station and exits at t_4. The number of states in a passenger journey, *I*, is variable dependant on the journey and we use the counter, *i*, to enumerate the sequence of states, $i = 1, 2, \ldots I$.

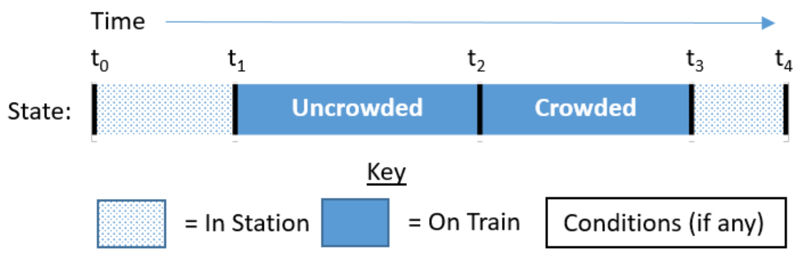

Figure 1 – An example passenger journey decomposed into four states. The journey is described with two stages: On Train and In Station. The shading of the state indicates the stage. Text is used to describe the conditions of the state. The markers t_0 to t_4 relate to the times when the passenger changed state.

The sum of VoT weightings across all states of a passenger journey can be used as an individual journey score. The following section describes how this is calculated when either the JTM or DM is used. The following section also compares how the JTM and DM calculate the crowding penalty. Section 3.2 then describes how the distribution of journey scores is evaluated to give a network score.

3.1 Calculating an individual journey score

A journey score calculated using the JTM is computed from the formula:

$$\psi^{JTM} = \sum_{i=1}^{i=I} t_i \Omega^{JTM}(\alpha_i^{JTM}, \beta_i^{JTM}, \omega^{JTM}) \tag{1}$$

Where ψ denotes the journey score, t_i, the time (in seconds) spent in the i^{th} state, Ω, the VoT weighting function, α_i and β_i, respectively the journey stage and conditions of the passenger's i^{th} state and ω the crowding penalty function. ψ^{DM} (given by (2)) is calculated similarly to ψ^{JTM}, but has an additional term to capture the relative disutility experienced by passengers changing train with a parameter for the number of times a passenger must change trains, ε, and a weighting factor, k_1. A value of 600 is used by Kunimatsu et al. for k_1, meaning that each train change has an associated disutility equivalent to 10 minutes (600 seconds) travelling on an otherwise unoccupied train. Table 1 provides the other parameter values for each metric.

$$\psi^{DM} = \sum_{i=1}^{i=I} t_i \Omega^{DM}(\alpha_i^{DM}, \beta_i^{DM}, \omega^{DM}) + k_1 \varepsilon \tag{2}$$

$\alpha_s^{JTM} =$	1	2	3	4	5
Description	On Train	On Platform	On Platform (Left Behind)	Moving Through Station	Buying Ticket
$\Omega^{JTM} =$	$1 + \omega^{JTM}(\beta_i^{JTM})$	2.5	3	2.7	2.5

$\alpha_i^{DM} =$	1	2
Description	On Train	In Station
$\Omega^{DM} =$	$1 + \omega^{DM}(\beta_i^{DM})$	3

Table 1 – The VoT weighting, Ω, for both metrics dependent on the journey stage, α, of a passenger's i^{th} state. A description of the journey stage relating to α is also shown. The VoT weighting for the On Train state is dependent on a crowding penalty function, ω, calculated using the conditions of the state, β. For the JTM, these values have been shared with the authors by personal communication and for the DM they are taken from Kunimatsu et al. (2012).

Table 1 shows the relative weighting both metrics put on each state (a lower value of Ω indicates a better passenger experience) and that the JTM describes a journey using five journey stages whereas the DM uses two. Both methods consider crowding only when passengers are in the "On Train" journey stage. The JTM crowding penalty, ω^{JTM}, is determined with the formula given by (3) using values given in Table 2.

$$\omega^{JTM} = \begin{cases} 0\,, & \delta \leq \mu \\ c_1 + c_2 \dfrac{\delta - \mu}{\gamma} - c_3 \dfrac{\delta\mu - \mu^2}{\gamma^2}\,, & \mu < \delta \leq \delta_{\max} \end{cases} \tag{3}$$

Where δ denotes the number of passengers, μ, the number of seats on the train, $\delta_{\max}$, the maximum passenger capacity, γ, the crush capacity and c_1 to c_3 constants. The crowding penalty formula given by (3) has been shared with the authors by personal communication from the Transport Planning department of LUL (Kelt, 2015). The second term of (3) captures the number of standing passengers relative to the crush capacity of the train and the third term captures the effect of seated passengers also. The value of γ describes the theoretical maximum number of people that can fit into the train assuming seven passengers per square meter of standing floor space. However, LUL have determined that the practical maximum capacity of a train is less than γ and under "normal operating conditions" the value of $\delta_{\max}$ is defined as 71% of γ. The DM crowding penalty, ω^{DM}, is determined with the formula given by (4) and requires computing the crowding factor, η, given by (5). The constants k_2 to k_7 and c_1 to c_3 are shown by Table 2.

$$\omega^{DM} = \begin{cases} k_2\eta, & \eta < 1 \\ k_3\eta - k_4, & 1 \leq \eta < 1.5 \\ k_5\eta - k_6, & 1.5 \leq \eta \leq 2 \end{cases} \tag{4}$$

$$\eta = \frac{k_7\delta}{\delta_{\max}} \tag{5}$$

Name	c_1	c_2	c_3	k_2	k_3	k_4	k_5	k_6	k_7
Value	0.85	1.915	1.03	0.027	0.0828	0.0558	0.179	0.2	2

Table 2 - Constant values used to calculate the crowding penalty, ω^{JTM} and ω^{DM}, in (3) and (4). For the JTM, these values have been shared with the authors by personal communication and the DM constants k_2 to k_6 are taken from Kunimatsu et al. (2012). The value of k_7 is informed by Nippon (2018).

The values of c_1 to c_3 have been derived by LUL and shared with the authors by personal communication (Kelt, 2015). The values of k_2 to k_6 are listed by Kunimatsu et al. (2012). Although Kunimatsu et al. do not explicitly define η, they describe it as the "congestion rate of the train", therefore it can be inferred as being proportional to $\delta/\delta_{\max}$. However because Nippon (2018) report the largest crowding factor (η) observed in Japan during 2017 as 2 (relating to when "bodies come into contact with each other and one feels considerable pressure"), the scaling factor k_7 is introduced into (5) and given a value of

2. The values of μ, $\delta_{\max}$ and γ are rolling stock specific and are defined by LUL for each fleet. For the LUL 2009 rolling stock (used on the Victoria Line and the subject of this investigation) their values are 288, 730 and 1028 respectively (Kelt, 2015)

Figure 2 compares ω^{JTM} and ω^{DM} on the y-axis for varying number of passengers (δ). The number of seats on the train is shown by a vertical dashed line and reflects that when $\delta \leq \mu$, the JTM does not apply a crowding penalty. A crowding penalty is applied by the DM even at this level of occupancy, but it is small in comparison to the minimum VoT weighting for passengers in the "On Train" journey stage (the dash-dot horizontal line). When $\delta > \mu$, the JTM applies a crowding penalty that is 4 to 8 times greater than the DM crowding penalty. For both metrics, the crowding penalty is always less than the minimum VoT weighting for the "On Train" stage. Both the JTM and DM models of crowding assume that passengers are homogenously distributed throughout the train and that passengers will always find and occupy a seat if one is available. Although this may not be realistic, it is the same for both models so the comparison is like-for-like.

The VoT weightings (in Table 1) and crowding penalty function for the JTM and the DM have been derived for the LUL network and Japanese railway respectively. It is therefore expected for these values to capture local preferences and expectations.

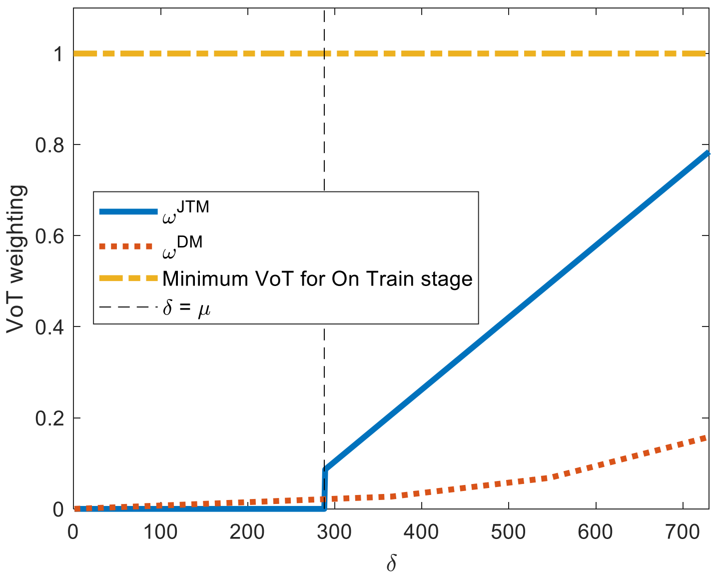

Figure 2 - The crowding penalty, ω, applied by the JTM and the DM for different numbers of passengers, δ, in LUL 2009 rolling stock up to its maximum capacity. The number of seats, μ, is shown by a vertical dash line. The minimum VoT weighting applied by both metrics to passengers that are in the "On Train" stage is shown by a horizontal dash-dot line.

3.2 Calculating a network score from journey scores

Networks provide journeys for multiple passengers so there is a distribution of journey scores. To improve the network assessment metric and ensure that journey scores only capture the quality of the service provided to the passenger by the network (and not the distance of the passenger journey which is a passenger choice), we normalize journey scores by the distance travelled. This allows like-for-like comparison of journey scores within the distance-normalized journey score distribution, Ψ, given by:

$$\Psi = [\frac{\psi_1}{d_1}, \frac{\psi_2}{d_2}, \dots, \frac{\psi_R}{d_R}] \tag{6}$$

Where ψ_r and d_r respectively denote the journey score and distance travelled relating to the r^{th} passenger and R the number of passengers. Different features of Ψ can be used to provide the network score, ϕ, for all R passengers conveyed. Although we wish to capture the effect of passenger numbers upon crowding, we also wish the network score to be independent of the number of journey scores within Ψ. Consequently, an additional passenger-number normalization step is included so ϕ^{JTM} and ϕ^{DM} are defined by:

$$\phi = \frac{1}{R}\sum_{r=1}^{r=R} \frac{\psi_r}{d_r} \tag{7}$$

Beyond this network score the characteristics of the distribution of Ψ can offer additional insight. For example, an operator wishing to examine the consistency of their service to passengers taking different journeys may evaluate the range of Ψ in addition to ϕ. In the current paper we focus primarily on ϕ to study quality of service provided to all passengers within the network.

4 Validation and comparison

To validate the network assessment metric, ϕ values are calculated using either the JTM or DM (ϕ^{JTM} or ϕ^{DM}) for the Victoria Line of the LUL network. For the same network, a network score is determined from measured Customer Satisfaction Survey (CSS) data, ϕ^{CSS}. The predictive values of ϕ^{JTM} and ϕ^{DM} are compared against the measured ϕ^{CSS} values and the correlation between their changes relative to a baseline year is quantified. The predictive values are then compared to each other to determine a relationship between the network assessment metric when either journey score metric is used. To calculate ϕ^{JTM} and ϕ^{DM} data describing the network operation was combined with data describing the passenger load and captures the effect of varying timetables and passenger loads over ten years. For the Victoria Line in the period investigated, the formation, length and interior layout of rolling stock remain constant, therefore the frequency of trains (determined by the timetable) has the greatest effect upon the passenger carrying capacity of the network. Decreasing the speed of trains on a line slows travel but also reduces headway with potential to decrease intervals between trains, so typically there is a trade-off between journey times and frequency. To meet increasing demand for travel, minimise crowding and generate more revenue, whilst maintaining competitive journey times against other transport modes, there is a pressure on LUL to balance this trade-off when updating their timetable.

4.1 Data sources

The data sources used in this investigation are: Victoria Line Working Timetable (WTT) numbers 31 to 41 (London Underground Limited, 2007, 2009, 2011, 2012a, 2012b, 2014, 2015a, 2015b, 2016b, 2016c, 2017), Access, Egress and Interchange (AEI) data provided by LUL (2016a), the Performance Data Almanac (PDA) (London Underground Limited, 2018a) and the Rolling Origin Destination Survey database (RODS) (London Underground Limited, 2018b). In the following section, the data is described in more detail.

4.2 Input data

The network operation data is taken from the WTTs and the AEI data. For each day, the WTTs provide the average train frequency and interstation run times for the three weekday operational periods on which our investigation concentrates: Morning Peak, Midday Off Peak and Evening Peak. Later operational periods are excluded because their timings are not consistent between the WTTs. The effect of this exclusion is unlikely to be significant because observing the RODS database indicates that this period is when the fewest passengers travel and so it has the least weighting on the network score. Weekends and holidays are not considered because they are more likely to be affected by events (e.g. sporting events or planned line closures for maintenance works) that affect passenger experience but are not captured in all the input data sources. The operational pattern described in the WTT is applied for every day the timetable was in effect (LUL update their timetable irregularly, but the date of introduction is provided be each WTT). The WTTs also provide the distance between adjacent station pairs. The AEI data describes the passenger travel time from station door to platform and vice versa, and platform to platform. The AEI data available relates to every four week period of the year beginning 2011 (the LUL reporting year begins on 1st April), over which the year mean is 2.23 minutes. Because data is only available for one year, this is applied for all years of the investigation, implicitly assuming that personal mobility within the station remains constant over this period.

The passenger load data is a combination of two data sources: the PDA and RODS. RODS provides the proportion of passengers included within the database that travel between adjacent station pairs in an operational period, i.e. *line section loadings*. However, this data does not describe whole passenger journeys (i.e. an origin and destination with any transfer stations). The PDA provides the total number of passengers travelling on the Victoria Line each year, and the quarterly CSS data. To collect the CSS data, LUL use questionnaires to ask approximately 2,500 passengers per quarter to rate, on a scale of 1 to 10, their satisfaction with their travel on the line of the last leg of their journey. The mean of the ratings is then multiplied by 10 and reported for each line by LUL.

4.3 Methodology

To calculate ϕ^{JTM} and ϕ^{DM}, the line section loading data was scaled by the yearly passenger numbers data and used to disaggregate the journeys of passengers who travelled further than the station adjacent to their origin, into a series of journeys between adjacent station pairs. For each operational period (Morning Peak, Midday Off Peak and Evening Peak) and line section, the number of passengers per train was calculated by dividing the number of passenger journeys in that period by the number of trains. Where demand for travel exceeded provision, the excess passengers were modelled as being “left behind” by one train before catching the next. The frequency of trains was used to determine the total passenger time spent in the “On Train”, “On Platform” and “On Platform (Left Behind)” stages. The journey score metrics were used to calculate the VoT weighting for these states. To avoid over-counting, the AEI time and weighting was only applied twice for each whole passenger journey defined by the PDA data rather than the RODS data. The “Buying Ticket” journey stage was disregarded because the use of pre-paid travel cards (“Oyster” cards) and contactless payment at ticket gates is common for this network. For example, in 2012 Oyster cards were used for over 80% of public transport travel in London (Transport for London, 2012). The inter-station distances were multiplied by the line section loadings so that the aggregate of the VoT weightings could be normalized by the total passenger distance travelled. This analysis was conducted for the Morning Peak, Midday Off Peak and Evening Peak operational periods of every weekday and was dependent on the daily timetable and yearly number of passenger journeys. To calculate the network score for that day, the values from the three operational periods of the day were summed. The year value was calculated as the mean of the year's day values. This process is illustrated by Figure 3 which shows ϕ^{JTM}, ϕ^{DM} and ϕ^{CSS} being calculated for corresponding years so that comparison is like-for-like. Because the CSS data is already normalized for passenger numbers and distance travelled, it is not relevant to normalize ϕ^{CSS} using (7).

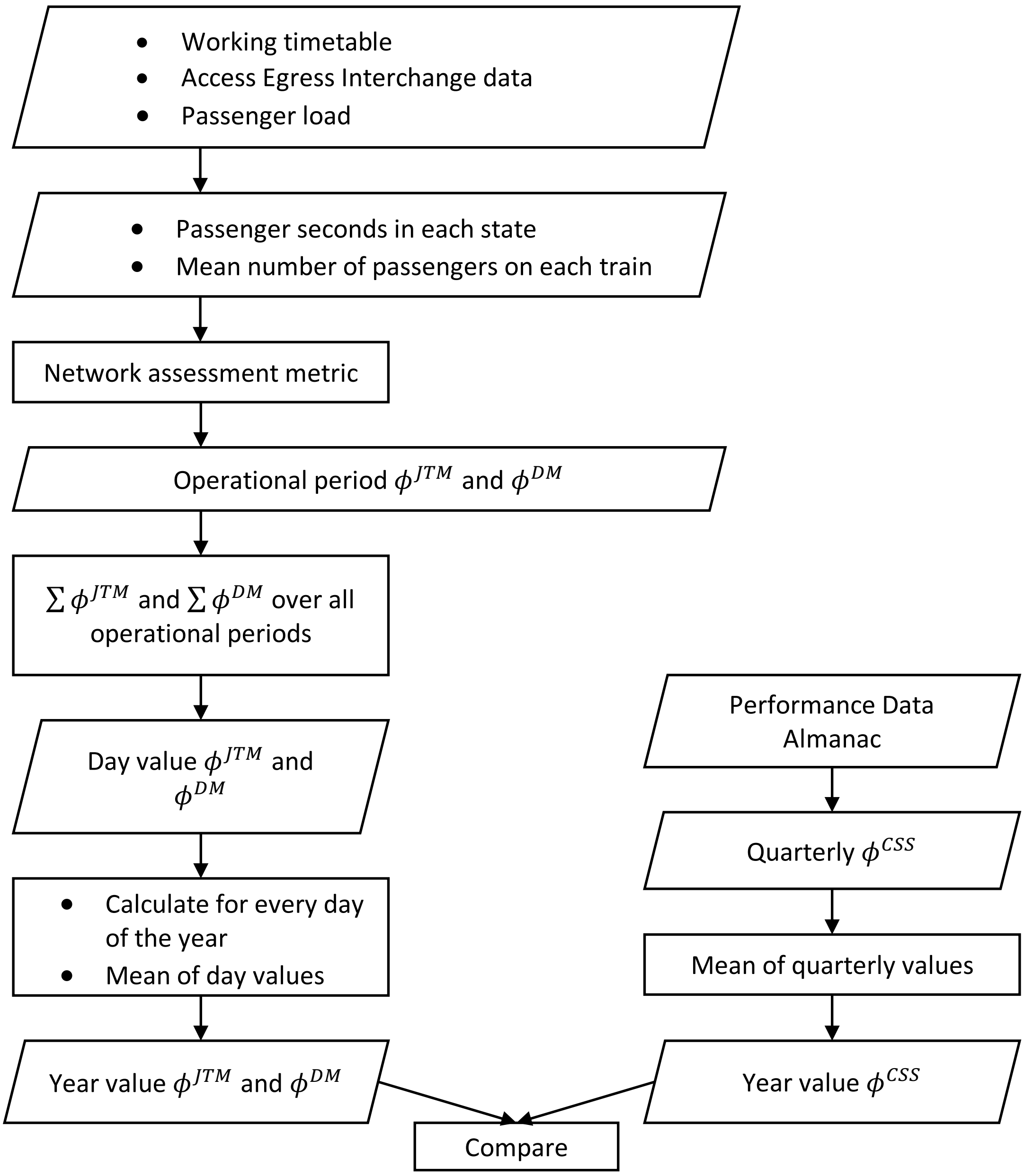

Figure 3 - The method for calculating the measured network score, ϕ^{CSS}, and predicted network score using the Journey Time Metric or Disutility Metric, ϕ^{JTM} and ϕ^{DM} respectively, from the Working Timetable (WTT), Access Egress and Interchange (AEI) data, passenger load data and Customer Satisfaction Survey (CSS) data.

4.4 Results

Figure 4 enables comparison of ϕ^{CSS} with ϕ^{JTM} and ϕ^{DM}, and also presents data where no distance or passenger normalization is applied, $\phi^{JTM\ (UN)}$ and $\phi^{DM\ (UN)}$, for the years 2008 to 2017. The number of passengers, R, is also included in the plot. Upward-pointing bars with values displayed on the left ordinate are used for ϕ^{CSS}, while $\phi^{JTM\ (UN)}$, $\phi^{DM\ (UN)}$, ϕ^{JTM} and ϕ^{DM} are represented by downward-pointing bars with values displayed on the right ordinate. Because the prediction metrics measure dissatisfaction and ϕ^{CSS} measures satisfaction, the right ordinate is inverted. A positive change in the vertical position of a bar-top for ϕ^{CSS} indicates a "better" performing network. R is also represented by markers with values displayed on the right ordinate. To allow comparison of relative changes on different scales and using different units, all series have been normalized against their 2008 value.

It can be seen that over time, in general, the measured network scores (ϕ^{CSS}) indicate improving network performance, with rising values relative to 2008. In general, this behaviour is successfully predicted by ϕ^{JTM} and ϕ^{DM}. However, $\phi^{JTM\ (UN)}$ and $\phi^{DM\ (UN)}$ predict deteriorating network performance and correlate with the increasing passenger numbers. It should be noted that, whilst the prediction metrics appear to give equal scores in 2008, this is because of the series normalization process. The importance of normalizing the predictive values by passenger numbers and distance travelled is clear if the metrics are to be compared over time.

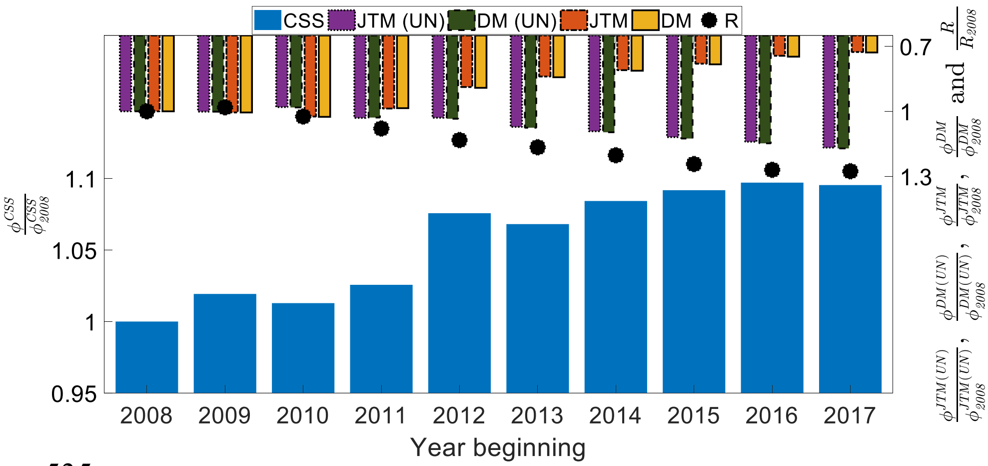

Figure 4 - Bar chart to compare predicted and measured network scores for different years and different prediction methods. Measured customer satisfaction scores, ϕ^{CSS}, are shown by the left ordinate. Predictions using the Journey Time Metric, ϕ^{JTM}, Journey Time Metric with no distance or passenger normalization, $\phi^{JTM\ (UN)}$, Disutility Metric, ϕ^{DM}, and Disutility Metric with no distance or passenger normalization, $\phi^{DM\ (UN)}$, are shown by the right ordinate which has been inverted. The right ordinate also displays the number of passengers, R. All values have been normalized against the corresponding 2008 value.

To investigate the importance of applying VoT weightings to different passenger states, Figure 5 enables comparison of ϕ^{CSS}, ϕ^{JTM}, ϕ^{DM} and a simple end-to-end journey time, ϕ^{EE}. To ensure like-for-like comparison, ϕ^{EE} has been normalized for passenger numbers and distance. The ordinates are similar to Figure 4 with the right ordinate now displaying ϕ^{EE} normalized against the 2008 value. To quantify the level of agreement between predicted and measured performance, Kendall's rank correlation coefficient B, τ_B, is calculated between the series of ϕ^{CSS} with each series of: ϕ^{JTM}, ϕ^{DM} and ϕ^{EE}. For the series of ϕ^{CSS} with ϕ^{JTM} and ϕ^{CSS} with ϕ^{DM} a value of -0.82 (P<0.005) is found (-1.0 indicates perfect (negative) correlation between prediction and measurement and 0 indicates no correlation). For the series of ϕ^{CSS} with ϕ^{EE} a value of -0.73 (P<0.005) is found, indicating worse correlation and that network assessment metric is improved by representing a passenger journey as a series of states and applying weighting to these.

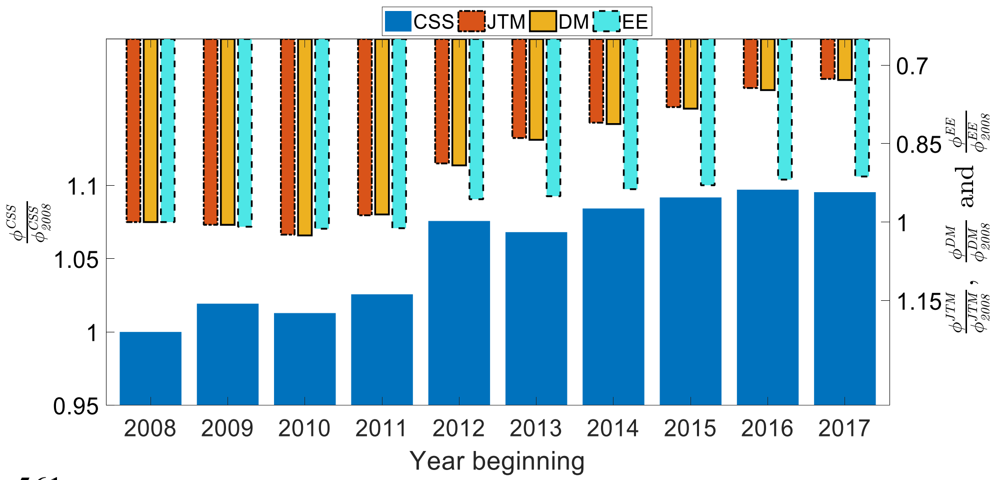

Figure 5 - Bar chart to compare predicted and measured network scores for different years and different prediction methods. Measured customer satisfaction scores, ϕ^{CSS}, are shown by the left ordinate. Predictions using the Journey Time Metric, ϕ^{JTM}, Disutility Metric, ϕ^{DM}, and end-to-end journey time, ϕ^{EE}, are shown by the right ordinate which has been inverted. All year scores have normalized against the 2008 value for the corresponding metric.

To explore the importance of the crowding penalty Figure 6 enables comparison of ϕ^{JTM} and ϕ^{DM} against the case where no crowding penalty has been applied in the calculation, $\phi^{JTM\,(NC)}$ and $\phi^{DM\,(NC)}$. The y-axis displays the raw values of ϕ, i.e. they are not normalized against the 2008 value. To determine what proportion of the network score is contributed by factors other than the crowding penalty, the value of $\phi^{(NC)}/\phi$ is calculated. For the JTM and DM series respectively, a mean value of 0.91and 0.99 is found both with a standard deviation less than or equal to 0.002. This behaviour is discussed in Section 5.

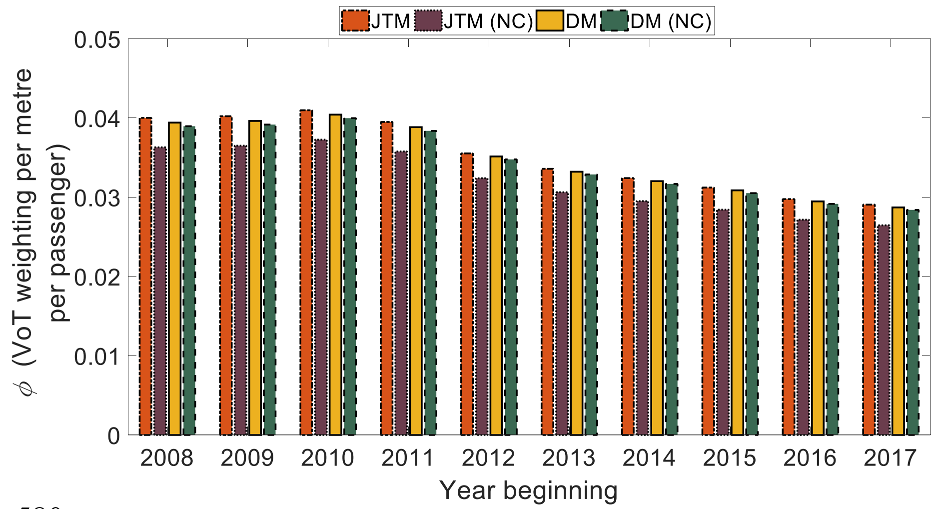

Figure 6 – Bar chart to compare the predicted network scores, ϕ, for different years and different prediction methods. Predictions using the Journey Time Metric, ϕ^{JTM}, and the Disutility Metric, ϕ^{DM}, are compared against the case where no crowding penalty is applied, $\phi^{JTM\,(NC)}$ and $\phi^{DM\,(NC)}$ respectively.

Figure 7 plots ϕ^{DM} against ϕ^{JTM} for the data from the years 2008 to 2017. The strong linear relationship of the data (ρ=1.00, P<0.005) suggests that, in general, similar changes in network performance are predicted by the JTM and the DM. A linear fit to this data shows a gradient of 1.013 (95% confidence bounds of 1.012 and 1.015). The intercept has been forced to the origin because both metrics are zero under the same condition: when no passenger time is spent in the network. The gradient implies that ϕ^{JTM} is consistently approximately 1.3% greater than ϕ^{DM}, but both are reacting consistently to external change over the period investigated.

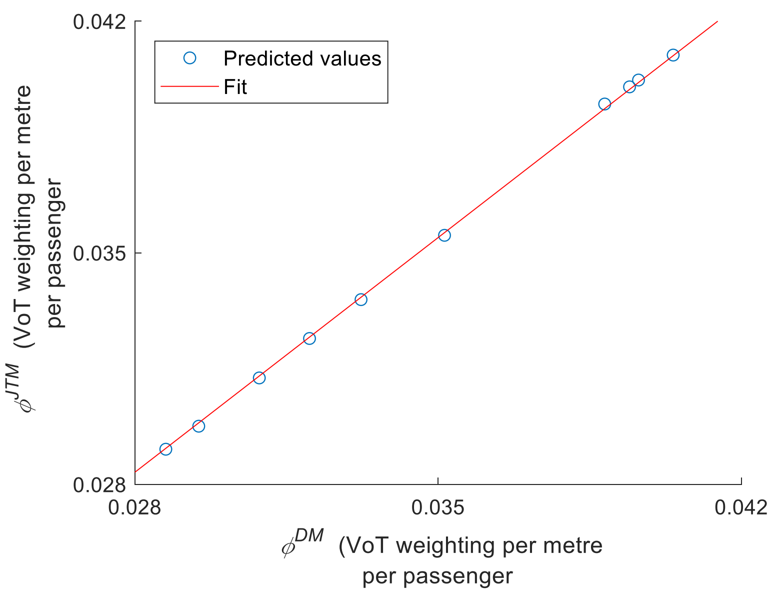

Figure 7- The relationship between the ten network score predictions for the Victoria Line from 2008 to 2017. The fit has an intercept forced to the origin and a gradient of 1.013.

5 Discussion

The results in Figure 4 indicate that, to successfully predict behaviour of ϕ^{CSS}, it is necessary to normalize the network assessment metric by the number of passengers and the distance they travel. In this investigation, the ratio between different line section loadings remains constant for all years therefore the value of R plotted in Figure 4 represents changes to passenger numbers and distance travelled. Consequently, the results in Figure 4 show that without passenger numbers and distance normalization, the predicted network scores become sensitive to both. This effect is unwanted therefore including passenger number and distance normalization within our network assessment metric is supported.

Choosing a typical significance level of 0.005, the results shown in Figure 5 are statistically significant evidence that the null hypothesis (that predicted and measured data are uncorrelated) can be rejected. Although the choice of significance level is arbitrary (Wasserstein and Lazar, 2016), considering the JTM and DM have been developed from empirical studies of passenger preferences and there is evidence that end-to-end journey time influences passenger experience (Transport Focus, 2016), we choose to accept the alternate hypothesis that there is correlation between CSS data and predictions with our network assessment metric when using the JTM, DM or end-to-end journey time. Because τ_B^{JTM} and τ_B^{DM} are closer to -1 than τ_B^{EE}, these results suggest that using our network performance metric with the JTM or DM better predicts relative changes to the CSS data than using end-to-end journey time. However, observing tables calculated by Walker (2016) indicate that even the 80% confidence intervals of τ_B^{JTM}, τ_B^{DM} and τ_B^{EE} are too large to determine a statistically significant difference between the values of τ_B^{JTM}, τ_B^{DM} and τ_B^{EE}. To determine a statistically significant difference by reducing the confidence interval without altering the significance level, more years of data for comparison are needed in the series of ϕ. It is unsurprising that τ_B^{JTM} and τ_B^{DM} do not equal -1.0 because, in this study, ϕ^{JTM} and ϕ^{DM} do not capture the effect of some factors, beyond the timetable and passenger load, which may affect ϕ^{CSS}, e.g. delayed trains. Our network assessment metric using the JTM or DM can capture the effect of some of these other factors, but the limitation of data available to this study means that they are not well captured by the model of network operation used. Similarly, because of factors such as survey design and implementation, the CSS data may not fully capture influencers to passenger experience that distinguish ϕ^{JTM}, ϕ^{DM} and ϕ^{EE}, e.g. if the surveys were not conducted during times of high travel demand the effect of crowding will not be well captured. Consequently, not being able to determine a statistically significant difference in the accuracy of ϕ^{JTM}, ϕ^{DM} and ϕ^{EE} might also be a limitation of the measured CSS data.

Section 3.1 describes that for low passenger numbers, ϕ^{JTM} is insensitive to crowding (because no crowding penalty is applied), whereas ϕ^{DM} is. However when some passengers are standing (the normal operating regime for many GB services, e.g. 70% of services into London St. Pancras during the morning peak (Peluffo, 2018)), ϕ^{JTM} will be more sensitive to crowding than ϕ^{DM} because it applies a crowding penalty four to eight times greater. This is confirmed by the results of Figure 6 which demonstrate that the contribution of the crowding penalty to the network score is on average 9% and 1% for the ϕ^{JTM} and ϕ^{DM} respectively. Section 3.1 also describes that the DM applies a greater VoT weighting than the JTM to passengers who are "In Station". Because the VoT weightings of the JTM and DM have been derived from surveying passengers, this may reflect local differences in passenger expectations where the metric was developed. For example, when used in our network assessment metric the JTM (developed in London) penalises crowding more and delay on the platform, less, than the DM (developed in Japan). This suggests that when considering a specific network, it is important to ensure the use of VoT weightings relevant to the passengers of that network. However, the similarity of the ϕ^{JTM} and ϕ^{DM} values in the results indicate that the difference in weightings placed on different passenger journey states approximately cancel out (for the study network in the years investigated). The results in Figure 7 show a high degree of correlation (ρ=1.00, P<0.005) between network scores calculated using the JTM and network scores calculated using the DM, despite their different formulations and countries of origin.

Considering all the results together suggests that using our newly developed network performance metric with the JTM or DM can be used to predict network performance from the passenger perspective, and successfully aggregates across passenger states to capture effects such as crowding and different journey stages. There is evidence that the network assessment metric, using either the JTM or DM, better predicts changes to customer satisfaction than end-to-end journey time. Because the JTM, CSS data and network operation data are all related to LUL, this result might be considered special to this case where there is a "closed-loop" between metric and validation. However, the DM has no connection to the LUL data but is demonstrated here to achieve similar outcomes. This indicates the result is not special to the "closed-loop" case.

6 Conclusions

Passenger journeys are multi-stage and the conditions of a journey stage, e.g. crowding when on a train, can vary. We have introduced the term "state" to describe a specific combination of stage and conditions. A passenger journey can be described as a series of states and the literature has shown that the relative time spent in each of these will have different effect on the overall experience of the passenger. Measuring the passenger end-to-end journey time alone, or the train punctuality at final destination (as used in the common UK performance measure, PPM) will not capture this. The JTM and DM are journey assessment metrics that can capture individual journey experience by applying a VoT weighting to time spent in each state. Both metrics sum the weighted time spent in each state, but they use different weightings, journey stages and the DM applies an additional penalty for train changes. Both apply a crowding penalty to capture the additional disutility caused to a passenger when traveling on a train with other passengers. For networks operating in the regime where some passengers cannot find a seat, the crowding penalty applied by the JTM is four to eight times greater than the DM. In this regime, the assessment of network performance using the JTM is more sensitive to crowding than when using the DM. Both the JTM and the DM can be used as part of a network assessment metric we introduce where the network score is taken to be the aggregate of journey scores normalized by the distance travelled and the number of passengers. It is found that, for the Victoria Line of the LUL network from 2008 to 2017, there is a high degree of correlation (ρ=1.00, $P<0.005$) between the network scores calculated with the JTM and network scores calculated with the DM, despite their different formulations and countries of origin. Extending the number of different networks in this comparison is an area for future work, to determine if this result is network-specific or general.

When comparing network scores against measured values of customer satisfaction for the same network (obtained from surveys) there is statistically significant evidence ($P<0.005$) to reject the null hypothesis that predicted and measured changes do not correlate. Considering other evidence from the literature, we therefore accept the hypothesis that predicted and measured changes are correlated which means our network assessment metric can be applied to predict the relative performance of different networks from the passenger perspective. For the data available, our network assessment metric using the JTM or the DM better predicted relative changes to customer satisfaction than end-to-end journey time. However, to determine a statistically significant difference more data for comparison is required. Therefore future work is to investigate networks where more than ten measurements of network performance can be collected and corresponding predictions computed (in the case of our experiment each measurement corresponds to a year over which passenger satisfaction data is available corresponding to the timetable operated that year, but any timescale in which a system change and its effect can be measured may be considered in future experiments). This might be achieved by re-investigating the Victoria Line in the future as additional years of customer satisfaction data become available. Further future work is to investigate networks where a more detailed description of the passenger route is available so that the effect of train transfer on passenger experience can be captured. The network assessment metric could then be validated for journeys which include this activity and might also allow a statistically significant difference with end-to-end journey time to be discerned. Updating the network assessment metric with new VoT weightings to capture other factors which influence passenger experience (e.g. cleanliness and journey purpose) is also an area for future work.

7 Acknowledgements

Funding was received from Network Rail and EPSRC grant number EP/M508135/1. The authors would like to thank Nigel Kelt and David Winslett at London Underground for discussion and access to data, and Dr Chikara Hirai and Dr Taketoshi Kunimatsu at Railway Technical Research Institute in Tokyo for discussions and documentation regarding use of network metrics in Japan.

8 References

ALI, A. A., ELIASSON, J. & WARG, J. 2017. Measuring the Socio-economic Benefits of Train Timetables Application to Commuter Train Services in Stockholm. *Transportation Research Procedia,* 27**,** 849-856.

ARUP, INSTITUTE FOR TRANSPORT STUDIES LEEDS & ACCENT. 2015. *Provision of market reserch for value of travel time savings and reliability: Non-Technical Summary Report* [Online]. London, UK: Department for Transport. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/470229/vtts-phase-2-report-non-technical-summary-issue-august-2015.pdf [Accessed 20/03/2019].

CACCHIANI, V. & TOTH, P. 2012. Nominal and robust train timetabling problems. *European Journal of Operational Research,* 219**,** 727-737.

CARNE, M. 2018. *Strategic Buisness Plan 2019 - 2024* [Online]. London, UK: Network Rail,. Available: https://cdn.networkrail.co.uk/wp-content/uploads/2018/02/Strategic-business-plan-high-level-summary.pdf [Accessed 15/11/2018].

CHAN, J. 2007. *Rail transit OD matrix estimation and journey time reliability metrics using automated fare data.* Master of Science in Transportation, Massachusetts Institute of Technology.

CHANG, Y. H., YEH, C. H. & SHEN, C. C. 2000. A multiobjective model for passenger train services planning: application to Taiwan's high-speed rail line. *Transportation Research Part B-Methodological,* 34**,** 91-106.

CHEN, C. & CHEN, F. 2010. Experience quality, perceived value, satisfaction and behavioral intentions for heritage tourists. *Tourism Management,* 31**,** 29-35.

DAVIS, R. 2018. *Rail Factsheet* [Online]. London, UK: Department for Transport. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/761352/rail-factsheet-2018.pdf [Accessed 13/12/2018].

HICKEY, S. 2011. *Improving the estimation of platform wait time at the London Underground.* Master of Science in Transportation, Massachusetts Institute of Technology.

HOROWITZ, A. J. 1978. The subjective value of the time spent in travel. *Transportation Research,* 12**,** 385-393.

KANAI, S., SHIINA, K., HARADA, S. & TOMII, N. 2011. An optimal delay management algorithm from passengers' viewpoints considering the whole railway network. *Journal of Rail Transport Planning & Management,* 1**,** 25-37.

KELT, N. Personal communication. "RE: LU Crowding and Left Behind". 2015, 3rd Decemeber.

KUNIMATSU, T., HIRAI, C. & TOMII, N. 2012. Train timetable evaluation from the viewpoint of passengers by microsimulation of train operation and passenger flow. *Electrical Engineering in Japan,* 181**,** 51-62.

KUNIMATSU, T., HIRAI, C., TOMII, N. & TAKABA, M. 2009. Evaluation of timetables by estimating passengers' personal disutility using micro-simulation. *In:* HANSEN, I. (ed.) *RailZuirch2009 - Third International Seminar on Railway Operations Modelling and Analysis.* Zürich, Switzerland: ETH Zürich.

LONDON UNDERGROUND LIMITED 2007. Victoria Line Working Timetable No. 31. London.

LONDON UNDERGROUND LIMITED 2009. Victoria Line Working Timetable No. 32. London.

LONDON UNDERGROUND LIMITED 2011. Victoria Line Working Timetable No. 33. London.

LONDON UNDERGROUND LIMITED 2012a. Victoria Line Working Timetable No. 34. London.

LONDON UNDERGROUND LIMITED 2012b. Victoria Line Working Timetable No. 35. London.

LONDON UNDERGROUND LIMITED 2014. Victoria Line Working Timetable No. 36. London.

LONDON UNDERGROUND LIMITED 2015a. Victoria Line Working Timetable No. 37. London.

LONDON UNDERGROUND LIMITED 2015b. Victoria Line Working Timetable No. 38. London.

LONDON UNDERGROUND LIMITED. 2016a. *FOI 0121 1617.xlsx* [Online]. London, UK: What Do They Know. Available: https://www.whatdotheyknow.com/request/i_need_to_understand_the_definit [Accessed 20/12/2018].

LONDON UNDERGROUND LIMITED 2016b. Victoria Line Working Timetable No. 39. London.

LONDON UNDERGROUND LIMITED 2016c. Victoria Line Working Timetable No. 40. London.

LONDON UNDERGROUND LIMITED 2017. Victoria Line Working Timetable No. 41. London.

LONDON UNDERGROUND LIMITED. 2018a. *Performance Data Almanac* [Online]. London, UK: Transport for London. Available: https://tfl.gov.uk/corporate/publications-and-reports/underground-services-performance [Accessed 29/11/2018].

LONDON UNDERGROUND LIMITED. 2018b. *Rolling Origin Destination Survey* [Online]. London, UK: Transport for London. Available: https://tfl.gov.uk/info-for/open-data-users/our-open-data [Accessed 20/12/2018].

NIPPON. 2018. *Japan's Most Congested Train Lines* [Online]. Tokyo, Japan: The Nippon Communications Foundation. Available: https://www.nippon.com/en/features/h00255 [Accessed 23/10/2018].

PELUFFO, N. 2018. *Rail passenger numbers and crowding on weekdays in major cities in England and Wales: 2017* [Online]. London, UK: Depatment for Transport. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/728526/rail-passengers-crowding-2017.pdf [Accessed 29/12/2018].

QIN, F. 2014. Investigating the In-Vehicle Crowding Cost Functions for Public Transit Modes. *Mathematical Problems in Engineering,* 2014**,** 1-13.

SELS, P., DEWILDE, T., CATTRYSSE, D. & VANSTEENWEGEN, P. 2016. Reducing the passenger travel time in practice by the automated construction of a robust railway timetable. *Transportation Research Part B-Methodological,* 84**,** 124-156.

SUSILO, Y. O. & CATS, O. 2014. Exploring key determinants of travel satisfaction for multi-modal trips by different

traveler groups. *Transportation Research Part a-Policy and Practice,* 67**,** 366-380.
TECHNICAL LEADERSHIP GROUP. 2017. *Rail Technical Strategy Capability Delivery Plan* [Online]. London, UK: Rail Safety and Standards Board. Available: https://www.rssb.co.uk/rts/Documents/2017-01-27-rail-technical-strategy-capability-delivery-plan-brochure.pdf [Accessed 7/11/2018].
TNS POLITICAL & SOCIAL. 2013. *Europeans' satisfaction with rail services* [Online]. The European Commission,. Available: http://ec.europa.eu/commfrontoffice/publicopinion/flash/fl_382a_en.pdf [Accessed 08/02/2019].
TRANSPORT FOCUS 2016. National Rail Passenger Survey - Spring 2016 Main Report. London, UK.
TRANSPORT FOR LONDON. 2012. *Join in the celebrations across the capital this summer with a limited edition Summer Oyster card* [Online]. London, UK: Transport for London. Available: https://tfl.gov.uk/info-for/media/press-releases/2012/june/join-in-the-celebrations-across-the-capital-this-summer-with-a-limited-edition-summer-oyster-card [Accessed 4/2/2019].
VANSTEENWEGEN, P. & VAN OUDHEUSDEN, D. 2007. Decreasing the passenger waiting time for an intercity rail network. *Transportation Research Part B-Methodological,* 41**,** 478-492.
VUCHIC, V. R. & NEWELL, G. F. 1968. Rapid Transit Interstation Spacings for Minimum Travel Time. *Transportation Science,* 2**,** 303-339.
WALKER, D. A. 2016. JMASM38: Confidence Intervals for Kendall's Tau with Small Samples (SPSS). *Journal of Modern Applied Statistical Methods,* 15**,** 868-883.
WANG, Y., TANG, T., NING, B., VAN DEN BOOM, T. J. J. & DE SCHUTTER, B. 2015. Passenger-demands-oriented train scheduling for an urban rail transit network. *Transportation Research Part C: Emerging Technologies,* 60**,** 1-23.
WARDMAN, M. 2004. Public transport values of time. *Transport Policy,* 11**,** 363-377.
WARDMAN, M. & MURPHY, P. 2015. Passengers' valuations of train seating layout, position and occupancy. *Transportation Research Part a-Policy and Practice,* 74**,** 222-238.
WASSERSTEIN, R. L. & LAZAR, N. A. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician,* 70**,** 129-133.