Assessing the discrepancies between recorded and commonly assumed journey times in London

T. Hillel*[1], P. Guthrie[1] M Elshafie and Y. Jin[1]

[1] *University of Cambridge, UK*
* *Corresponding Author*

ABSTRACT Transport models for infrastructure investment and operations planning make use of generalised trip cost to predict travel choice decisions. In cities, the most important factors in the generalised cost is trip duration. When calibrating such models to achieve simulation fidelity, observed data such as the choice of destination and means of travel recorded in travel surveys are used in estimating model parameters. Ideally, observed travel durations should also be used in the model estimation. However, in the past it was infeasible to record the actual trip durations to any degree of accuracy in travel surveys. Trip durations derived from a transport network model were commonly assumed to be sufficiently representative. Increasing availability of better recorded trip durations from travel surveys and better modelled trip durations from online mapping present the promise of significant improvements in the fidelity of transport models. As a preamble to adopting such data, we investigate how the best developed recording of actual trip durations from the London Travel Demand Survey compares with the most advanced trip duration modelling from Google Map travel directions API. We find clear discrepancies between the two, with the discrepancies varying systematically for different means and purposes of travel. The magnitude of the discrepancies is greater than can be attributed to randomness or noise. The systematic nature of the discrepancies suggests that transport network modelling even in its advanced form still has a long way to go to represent the observed patterns of behaviour, particularly for non-commuting journeys which account for about 80% of all trips made in cities. Since the discrepancies may create a systematic bias in the model parameters, it is of critical importance to understand them better in future analysis.

1 INTRODUCTION

Transport models for infrastructure investment and operations planning use discrete choice models to predict travel on the transport network, based on generalised travel costs (Train 2009; Prato 2009; TfL 2014). In order to obtain sound predictions, it is essential to have good measurements of the generalised travel costs. In cities, the duration of travel is usually the greatest influence on generalised cost.

The *commonly assumed* definition of trip duration, used in both research and industry, is the time taken to complete the optimal route between two points on the network, as predicted by a transport network model. This includes timetable information on public transport services, and either observed or predicted congestion delays on roads. This common assumption has been made by convention because it is difficult to record systematically the actual trip durations in travel surveys, where all travel within a day or week need to be recorded.

Increasing availability of better recorded trip durations from travel surveys and better modelled trip durations from online mapping present the promise of significant improvements in both data sources. In this paper we investigate how the best developed recording of actual trip durations from the London Travel Demand Survey (LTDS) compares with the most advanced trip duration modelling from Google Maps API. The analysis is for passenger travel only.

2 DEFINITIONS OF TRIP DURATION

There are four alternative definitions of trip duration:

1. *Ideal duration (t_i)*: time to complete a trip as predicted by a transport model, free of effects of traffic congestion or delays.

2. *Commonly assumed duration (t_c)*: time to complete a trip as predicted by a transport model given predicted/observed traffic conditions, congestion and delays.
3. *Expected duration (t_e)*: time the passenger expects to take.
4. *Recorded duration (t_r)*: time the passenger records a trip as having taken.

Their characteristics are summarised in Table 1.

Table 1. Summary of trip duration definitions used in this study.

	Name	Includes congestion/ disruption	Modelled	Measured	Pertinence
t_i	Ideal	✗	✓	n/a	Low
t_c	Commonly assumed	✓	✓	n/a	Mid
t_e	Expected	?	✗	✗	High
t_r	Recorded	✓	✗	✓	High

Both the ideal and optimal durations are theoretical values which represent minimum journey times and are computed using a model. The expected and recorded durations are real world values that need to be observed or measured rather than computed.

Galotti & Bathelemy (2014) analyse the theoretical efficiency of the British public transport network by comparing the ideal route and duration for multiple journeys *(t_i)*. This value is independent of the conditions of the transport network, and so is of low relevance for real passenger journeys.

As discussed, trip assignment within state-of-the-art transport models uses the commonly assumed duration *(t_c)*. This is of greater relevance to real world journeys than the ideal duration *(t_i)*, as it is dependent on the network conditions. However, it still represents an idealised case, where the passenger takes the optimal route and travels as quickly as possible.

In reality, passengers make decisions based on their expected duration of a trip *(t_e)*. It is not possible to model the expected duration of a trip directly as it is highly dependent on a passenger's individual experience at that time. Instead this study investigates the recorded trip duration *(t_r)*, which is how long a passenger reports a trip to have taken. This is likely to have a strong relationship with the expected duration of repeating a similar trip in the future.

Each of these definitions of trip duration is separate from the duration a passenger actually takes.

3 METHODOLOGY

This study assesses the discrepancies between recorded trip durations *(t_r)* taken from the London Transport Demand Survey (LTDS) and their corresponding commonly assumed trip duration *(t_c)*, generated using the Google Maps Directions Application Programming Interface (API).

3.1 LTDS

The data source for completed journeys for this study is the LTDS, a continuous survey carried out by TfL of a sample of households within London's orbital motorway, the M25 (TfL 2011).

Each household is surveyed on one day of the year, listing all of the members of the household, all of the vehicles that the household owns or has access to, and the estimated total household income. Each household member over 5 years of age then completes a trip diary, giving details of all of the trips made on the survey date. Details include the trip start-point, end-point, start time, trip duration *(t_R)*, means of travel and trip purpose.

This study uses data from the 2013/14 survey year, which contains 44,981 trips made by 18,877 individuals.

3.2 Google Maps API

The data source for generating optimal journey times is the Google Directions API. It generates more than one route for any origin-destination pair. In line with the modelling convention, we retrieve the optimal route as the commonly assumed trip duration *(t_c)*.

Google's representation of London's transport network is commonly considered fine grained and accurate. On the network, Google generates real-time traffic routeing using crowd-sourced movements data. Google also receives up-to-date public transport timetable and delay information from TfL and Network Rail. It is reasonable to consider this dataset to the most advanced estimation of trip durations.

Using this information, the Google Maps API can return an optimal route and the commonly assumed trip duration *(t_c)* calculated using a modified Dijkstra's algorithm (Dijkstra 1959; Casey et al. 2015).

3.3 *Processing the data*

The trips from the LTDS are sorted into the same trip classes as used in London's transport policy model LTS (Table 2). For each trip in the LTDS, an optimal route and duration is obtained from the Google Maps API. The trip requests to the Google Directions API are performed in time bracketed groups, according to their departure time and day of the week from the LTDS, for each means of travel:

- **Driving:** Trips sorted by weekday, Saturday, or Sunday departure. Within each day, trips sorted into groups of departure time within two hour intervals.
- **Transit (public transport):** Trips sorted into weekday or weekend departure. Within each day, trips sorted into day and night departure trips.
- **Cycling and walking:** No time bracketing used, as walking and cycling durations returned by Google are time independent.

Table 2. LTS model trip classes.

Time periods (weekday)	Morning peak:	07:00-10:00
	Inter-peak:	10:00-16:00
	Evening peak:	16:00-19:00
Means of travel	Walking	
	Cycling	
	Transit	
	Driving	
Trip purposes	Home-based work	
	Home-based education	
	Home-based other	
	Non-home-based work	
	Non-home-based-other	

4 RESULTS

4.1 *Scatter plots*

Figure 1 shows a scatter plot of t_c against t_r for all trips within the study. A bi-square linear regression, which is robust to outliers, is performed on the data. The regression line is well below the line $y = x$ which corresponds to $t_r = t_c$. This shows that the recorded trip durations tend to be substantially longer on average than the commonly assumed durations.

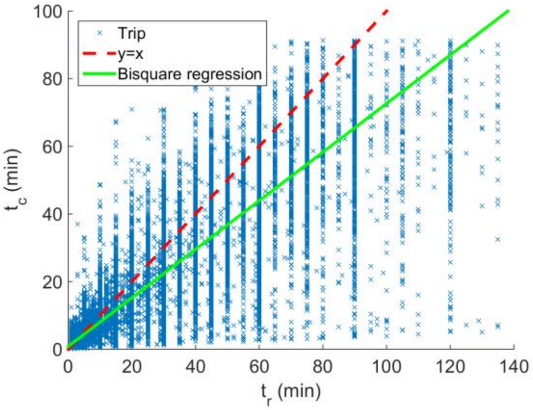

Figure 1. Scatter plot of all trips.

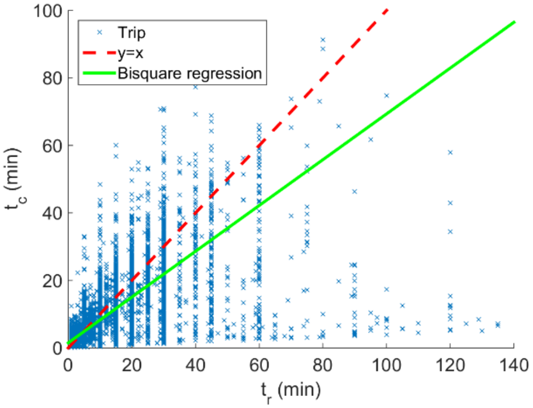

Figure 2. Scatter plot of walking trips.

Figures 2-5 show the scatter plots for each transport mode. Each plot contains trips for all trip purposes and trip departure periods. Each plot has different visual characteristics, which are shown numerically in Table 3. The bi-square regression gradient shows the average relationship between t_c and t_r and the Pearson correlation coefficient demonstrates the spread of the data. These values are also calculated for each journey purpose and departure period. There is wide variation in both the gradient of the linear regression and the value of the cross-correlation coefficient for each trip class.

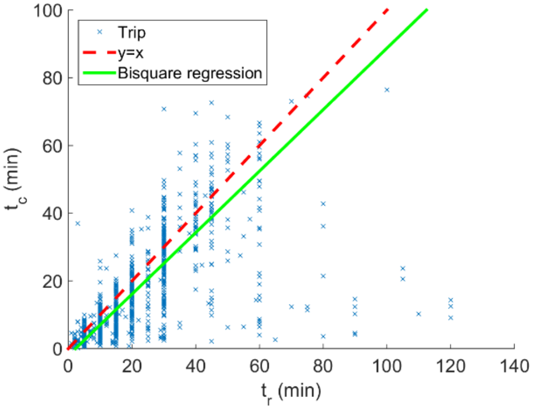

Figure 3. Scatter plot of cycling trips.

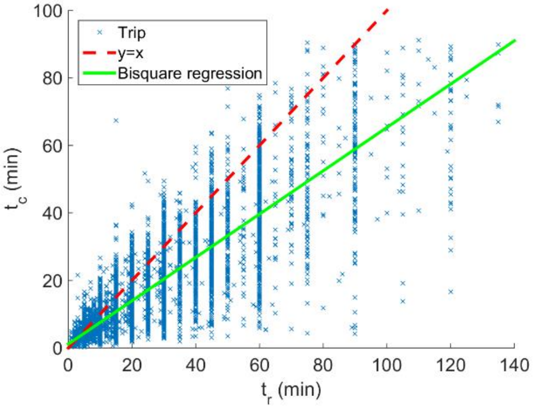

Figure 4. Scatter plot of transit trips.

All of the plots show strong banding of the recorded duration *(t_r)*. This relates to the fact that the recorded duration is a measure of how long a passenger perceives a journey to have taken. Below 60 minutes, the bands occur at 5 minute intervals, demonstrating that for short journeys the resolution of perceived duration is ±2.5 minutes, i.e. the trip durations are rounded to the nearest 5 minutes. For all of the plots, the strongest band above 30 minutes is at 60 minutes.

The bands at 55 minutes and 65 minutes are also much weaker than the other bands. This suggests that for the majority of the population, there is a tendency to round trip durations to 60 minutes. Above 60 minutes, the plot for all trips shows the strongest bands at 75 minutes, 90 minutes, and 120 minutes, showing the resolution for the majority of the population reduces to 15 minute and then 30 minute intervals.

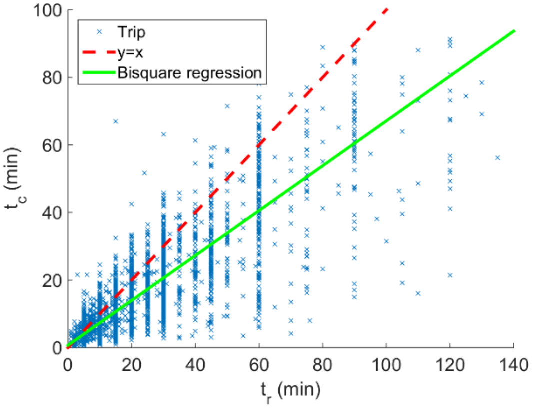

Figure 5. Scatter plot of driving trips

Table 3. Linear regression gradient, and correlation coefficient for each trip mode, trip purpose, and departure time period.

Category	Class	Gradient	Correlation
All	All	0.718	0.830
Transport mode	Walking	0.679	0.548
	Cycling	0.908	0.588
	Transit	0.679	0.751
	Driving	0.647	0.837
Purpose	Home-based work	0.752	0.840
	Home-based education	0.629	0.849
	Home-based other	0.681	0.822
	Non-home-based work	0.573	0.760
	Non-home-based other	0.655	0.712
Period	AM Peak	0.766	0.866
	Inter peak	0.669	0.797
	PM peak	0.694	0.848
	Other	0.736	0.820

4.2 Probability distributions

In order to create the probability distributions, a dimensionless ratio of recorded duration *(t_r)* to commonly assumed duration *(t_c)* is defined:

$$d = t_r/t_c \tag{1}$$

The ratio of two values is not a symmetrical operation, and as such the distribution of the ratios show heavy positive skew. This is shown in Figure 6, which plots smoothed kernel distributions of the ratio for each transport mode. The line $d = 1$ corresponding to $t_r = t_c$ is given for reference. To deal with the heavy skew, the natural logarithm of the ratio is taken to provide a symmetrical operation. This gives the following formula for the log-ratio (*r*):

$$r = \ln(d) = \ln(t_r/t_c) \tag{2}$$

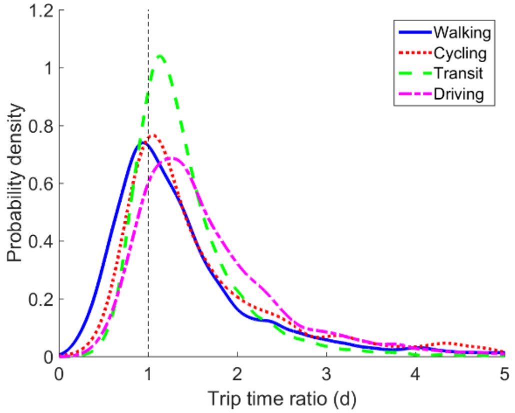

Figure 6. Skewed probability distributions by means of travel.

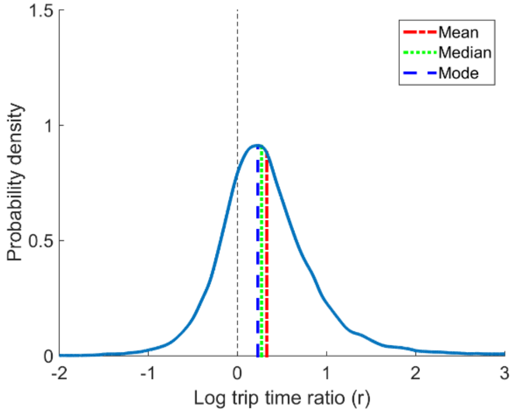

Figure 7. Probability distribution for all trips.

Figure 7 shows the smoothed kernel distribution plot of all trips combined. Here the line $r = 0$ corresponds to $t_r = t_c$. The mean, median, and mode are all to the right of this line, once again showing that the recorded durations *(t_r)* are on average significantly higher than the commonly assumed durations *(t_c)*.

Smoothed kernel distributions of the log-ratio *(r)* are generated for each trip class. The sample geometric mean and standard deviation of the ratios *(d)* can be calculated directly from the log-ratio *(r)*, using the following formulae:

$$\mu_g = \left(\prod_{i=1}^{n} d_i\right)^{1/n} = \exp\left[\frac{1}{n}\sum_{i=1}^{n}\ln d_i\right] \tag{3}$$

$$s_g = \exp\sqrt{\frac{\sum_{i=1}^{n}\left(\ln\frac{d_i}{\mu_g}\right)^2}{n-1}} \tag{4}$$

where:

$$r_i = \ln d_i$$

These values are given in Table 4 for all of the primary trip classes, alongside a calculation of the Pearson's moment coefficient of skewness of the log-ratios.

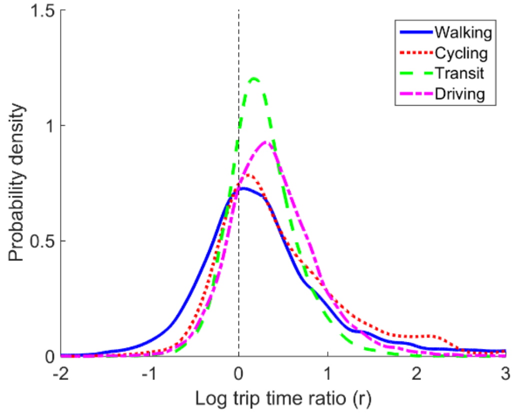

Figure 8. Probability distributions for each means of travel.

Figure 8 shows the smoothed kernel distribution plots for each means of travel. Walking has the modal value closest to the $r = 0$ line. However, it has high positive skewness and variance. Cycling trips show a very similar distribution to walking. Transit trips have the lowest variance and skewness, reflecting their constrained nature (transit trips are constrained to train lines/bus routes, which generally run to a fixed schedule). The value of transit trips is also lower than that for driving trips.

Table 4 also gives the statistical properties of the distributions for each trip purpose and departure time. As with choice of the means of travel, the distributions for trip purpose are distinct with clear differences. Home-based work (commuting) trips have the geometric mean closest to the origin, as well as the lowest variance and skew. These trips are repeated regularly, and as such there is a high incentive for passengers to research and select the quickest route. Home-based trips tend to show lower variance and skewness to non-home-based trips.

The distributions for each departure period are relatively closely matched compared to those for different means of travel and trip purpose, as shown by their similar geometric mean, standard deviation and skewness.

Overall the distribution of the log ratios varies significantly for each trip class. This is indicated with the properties shown in Table 4.

Table 4. Geometric mean, standard deviation and skewness of the log-ratio.

Category	Class	Geometric mean	Geometric S.D.	Skewness
All	All	1.390	0.568	0.918
Means of travel	Walking	1.328	0.741	1.002
	Cycling	1.608	0.719	1.047
	Transit	1.311	0.389	0.575
	Driving	1.472	0.494	0.648
Purpose	H.B.W.	1.275	0.396	0.527
	H.B.E	1.380	0.519	0.477
	H.B.O.	1.384	0.527	0.666
	N.H.B.W	1.473	0.601	0.716
	N.H.B.O.	1.538	0.801	0.864
Period	AM Peak	1.363	0.502	0.560
	Inter peak	1.431	0.623	0.966
	PM peak	1.421	0.534	0.689
	Other	1.355	0.562	1.012

5 CONCLUSIONS

There are clear discrepancies between the commonly assumed trip durations such as used in transport models and trip durations as recorded by the passenger as reflected in the survey data. Crucially, as is shown by the geometric mean and skewness of the data, the discrepancies are non-uniform across the modes of travel and trip classes. The patterns of variation in the duration of actual trips compared to the commonly assumed duration for different classes of trip is not captured in the generalised costs calculated by current transport models, which may have significant implications regarding the assumptions made for model calibration, validation and predictions.

The analysis carried out in this study is subject to imprecisions inherent in both the recording by the surveyed travellers and in the derivations of the Google based travel times, but the discrepancies are both greater in magnitude and more systematic than can be attributed to randomness or noise. This would appear to warrant more in-depth analysis. Emerging availability of more directly sensored travel data would make this increasingly feasible in future work.

ACKNOWLEDGEMENTS

This research is an update from Tim Hillel’s MRes thesis which was undertaken as part of the Future Infrastructure and Built Environment Centre for Doctoral Training at the University of Cambridge, which is funded by the UK Engineering and Physical Sciences Research Council (EPSRC).

We thank Transport of London for their provision of the LTDS data, and Gerard Casey from the Centre for Sustainable Development at the University of Cambridge for assistance in data retrieval from the Google Maps API. However, the authors are solely responsible for the views or any remaining errors.

REFERENCES

Casey, G. H., Silva, E. A., Soga, K. et al. 2015. An ABM supported by real-time big data: The case of HSR vs. Aviation, *Proceedings of the ICE - Transport* Under Review.

Dijkstra, E. W. 1959. A note on two problems in connexion with graphs, *Numerische mathematik* **1.1**, 269-271.

Gallotti, R. & Barthelemy, M. 2014. Anatomy and efficiency of urban multimodal mobility, Nature *Scientific reports* **4**.

Prato, C. G. 2009. Route choice modeling: past, present and future research directions, *Journal of Choice Modelling* **2.1**, 65–100.

Train, K. E. 2009. *Discrete choice methods with simulation.* Cambridge university press, Cambridge.

Transport for London 2014. *The London Transportation Studies Model (LTS)*. Transport for London, London.

Transport for London 2011. *Travel in London, Supplementary Report: London Travel Demand Survey (LTDS)*. Transport for London, London.