Mining Public Transport Usage For Personalised Intelligent Transport Systems

Neal Lathia[1], Jon Froehlich[2], Licia Capra[1]
[1]*Department of Computer Science, University College London, United Kingdom*
[2]*Computer Science and Engineering, University of Washington*
n.lathia, l.capra@cs.ucl.ac.uk, jfroehli@cs.washington.edu

***Abstract*—Traveller information, route planning, and service updates have become essential components of public transport systems: they help people navigate built environments by providing access to information regarding delays and service disruptions. However, one aspect that these systems invariably lack is a way of tailoring the information they offer in order to provide personalised trip time estimates and relevant notifications to each traveller. Mining each user's travel history, collected by automated ticketing systems, has the potential to address this gap. In this work, we analyse one such dataset of travel history on the London underground. We then propose and evaluate methods to (a) predict personalised trip times for the system users and (b) rank stations based on future mobility patterns, in order to identify the subset of stations that are of greatest interest to each other and thus provide useful travel updates.**

I. Introduction

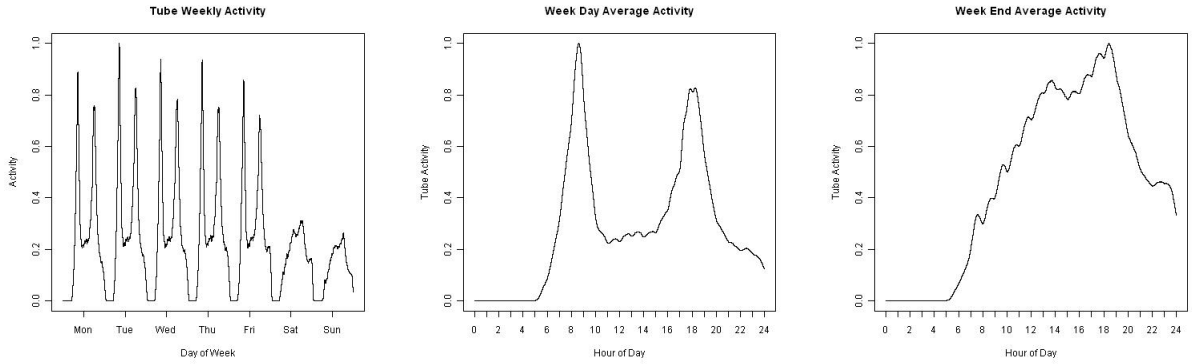
Interactive maps, route planners, and real-time service alerts have become essential components of public transport systems: they provide travellers with access to vital information[1] that reduces barriers to using public transit. A notable feature lacking in these systems, however, is the ability to dynamically tailor information to the individual needs of each traveller [1]. Most online transit tools, for example, have yet to incorporate an understanding of travellers' preferences or their mobility-related requirements—factors that can greatly impact the overall transit experience. As urban public transport systems continue to expand in size and complexity, so too does the amount of information that is available to travellers as they ride the transit system; for example, London's underground alone (excluding buses and train services) has 11 interconnected lines with hundreds of stations. Personalisation systems offer a rich opportunity to both tailor information to the individual traveler and reduce the complexity and need for manually searching for relevant transit information. In this paper, we explore automated methods to enable public transit personalisation; the goal here is to explore the viability of personalised travel information with little-to-no direct feedback from the travellers themselves.

A significant historical obstacle to personalising the public transport experience has been the lack of data about individual traveller preferences and routines. However, the introduction and widespread adoption of automated fare collection (AFC) systems offer a potential channel to this missing data. AFC transit systems forgo traditional fare media such as paper tickets or magnetic strip cards in favour of alternatives such as RFID-based smart cards (e.g., London's Oyster Card, Seattle's Orca Card) or near-field communication on mobile phones (e.g., the Tokyo Metro System). These new payment systems create a digital record every time a trip is made, which can be linked back to the individual traveller. Mining the travel data that is created as travellers enter and exit stations can give vast insight into the travellers themselves: their implicit preferences, travel times, and commuting habits.

The increasingly wide availability of AFC data has lead to an explosion of research primarily focused on how such data can be used to evaluate and study the performance of the transportation system itself: for example, through demand modelling [2], service reliability measurements [3], average route travel time estimation [3], and station transfer analysis [4]. In this work, we do not focus on what the AFC data reveals about the transportation system but instead what it can reveal about individual traveller behaviour. Indeed, recent work [3] states that, on average, only 46-62% of the time that users spend in the tube is actually spent riding the trains. The rest of the time is spent interchanging, walking, or waiting: differences between users (e.g., locals vs. tourists, youngsters vs. elderly) will strongly impact travel time and should be incorporated into transport route planning and notification services. In this work, we show how AFC systems can be used to uncover individual differences in travel patterns that, in turn, can be used to enable personalised transit services.

We focus on two facets of personalisation, both of which can be formalised as prediction problems: (a) predicting personalised travel times between any origin and destination pairs to provide users with accurate estimates of *their* transit time, and (b) predicting and ranking the interest that individual travellers will have in receiving alert notifications about particular stations based on their past travel histories. For this analysis, we use data collected from the London Underground (tube) system, which implements electronic ticketing in the form of RFID-based contact-less smart cards (Oyster cards). Unlike some AFC systems, Oyster cards must be used both when entering *and* exiting stations,

[1]For example, http://alerts.tfl.gov.uk/



(a) Week View (b) Week Day (c) Week End

Figure 1. An aggregate view of weekly tube activity (cumulative weekly, week day and week end journey frequency). Note how a two-spike temporal pattern corresponding to heavy commuting periods is visible on weekdays but not on weekends.

allowing us to track a traveller's origin, destination and travel time. In particular, we make the following contributions:

- First, we perform an extensive analysis of a large corpus of anonymised per-person usage of the London underground (Section II). Although aggregate summaries of the data point toward a consistent use of the system, we highlight measurable differences in transit usage between travellers that more concretely motivate the need for personalisation.
- Second, we propose and evaluate a set of simple algorithms to personalise travel time estimation (Sections III and IV). In doing so, we aim to implicitly capture aspects of underground usage that affects travel time such as route choice, the ability to physically move about a station, and route familiarity. We also evaluate a means of combining the different methods that takes into account the varying amount of data available for each prediction.
- Finally, we design and evaluate (Sections V and VI) a set of ranking algorithms that aim to identify which stations will be of interest to each traveller in their future mobility patterns.

We believe that this paper not only highlights the potential value of AFC datasets to the data mining and personalisation research communities, but will also be of value to public transportation planners and operators.

II. The London Underground AFC Dataset

The London Underground consists of 11 interconnected underground lines, six fare regions, and over 250 tube stations. In this analysis, we use two datasets of London's tube usage from different 83-day periods (May-July 2009 and October 2009-January 2010). Each dataset is a 5% sub-sample of all users who were recorded during the two periods. A data point in our sample is a tuple in the form:

$$< u, (o, d), t_o, t_d >$$

Each one corresponds to a trip observation: the unique, persistent user id (u), the trip (with origin (o) and destination (d) stations), the time stamp t_o when u entered the origin station and the time t_d when u exited from the destination station. These time stamps allow us to compute trip time in minutes (t_d - t_o). We first filtered inconsistent entries from the data: we removed trips with invalid or missing origin or destination stations (caused by users who did not touch their ticket to the reader when starting or ending a trip), trips with the same origin and destination, and trips whose arrival timestamp was prior or equal to the departure time. Approximately 7% of the raw data was discarded; we were left with over 600,000 travellers and 12.7 million trips; details of each dataset is available in Table I.

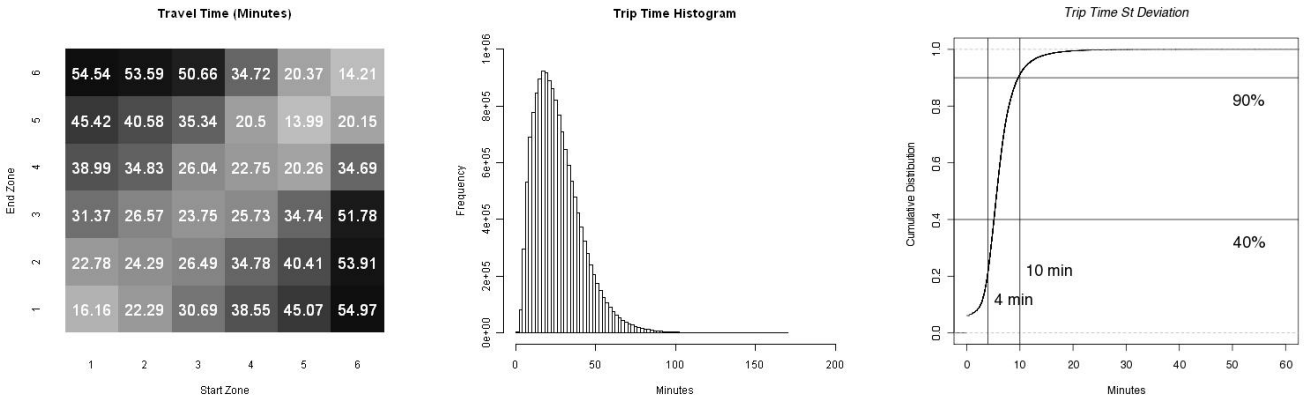
We begin by analysing the *aggregate* temporal usage patterns of the underground and the underlying differences that exist in *individual* traveller patterns. While the aggregate analysis may be used to reveal the large-scale geographic and sociocultural properties of the city, only a single type of usage pattern emerges: commuting to and from work. The focus on individual traveller patterns, instead, reveals a wide range of emergent user behaviours, which highlights the importance and potential benefits of augmenting intelligent transport systems with personalisation techniques.

A. Aggregate Behaviour

We first examine the overall system usage. The primary focus here is to highlight systemic patterns that (a) give a broad perspective of the usage of the system and (b)

Name	Date Range	Users	Trips
D1	03 May 2009 - 25 Jul 2009	298,294	7,534,700
D2	18 Oct 2009 - 09 Jan 2010	309,588	7,702,713

Table I
Tube Trip Datasets



(a) Zone O-D Travel Time Matrix (b) Trip Time Distribution (c) Trip Time Standard Deviation

Figure 2. (a) Zone O-D travel time matrix, showing the relation between the zonal structure of the London system and travel time, (b) Trip time distribution and (c) Standard Deviation CDF—both show that the majority of tube trips are short and tend to be very close to the trip's mean time.

may impact our ability to accurately predict travel times or stations of interest. This analysis also provides a necessary context within which to interpret our results.

*1) **Temporal Patterns**:* Figure 1 plots the cumulative number of ongoing trips over time, over the course of a week, weekday, and weekend day (Saturday *or* Sunday). The two distinct peaks in weekday activity (Figure 1(b)) reflect London's dependence on the tube as a means of commuting: the largest proportion of trips occur within the morning commute, 6.30 to 9.30am (22.95%), and the (longer spanning) evening commute, 4.30 to 8pm (29.19%). Unsurprisingly, these temporal patterns are not shared by weekends or national holidays where the number of ongoing journeys steadily increases during the course of the day, until approximately 7pm (Figure 1(c)).

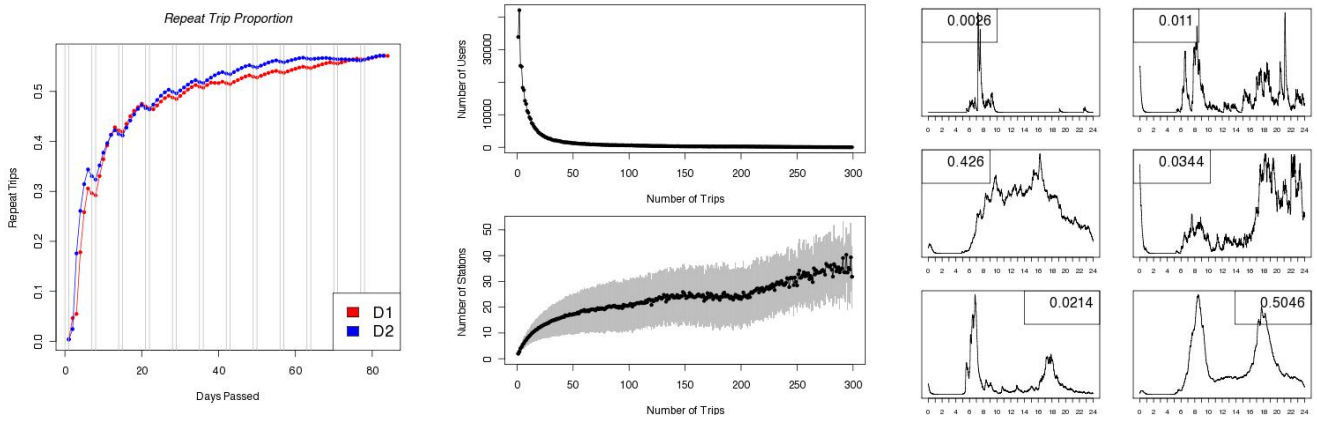
*2) **Station Visit Patterns**:* The mean number of stations visited per user per day is 2.54 (median 2, mode 2), while the mean number of trips per user per day is 1.87 (median 2, mode 2). A majority of users who appear daily in the system are travelling between just two stations (a "home" station to a "work" station and then back to a "home" station).

*3) **Travel Time**:* Travel time is calculated simply by subtracting the user's destination station exit time from when they entered the origin station ($t_d - t_o$). The global mean trip time for D1 is 26.81 ± 14.93 minutes, while for D2 it is 27.11 ± 15.12 minutes; the overall average trip time for the *entire system* is roughly half an hour. Finer grained travel time estimates can be obtained by incorporating zoning information. London tube zones are used to demarcate the city and its surrounding area into pricing tiers; Zone 1 is within central London, while stations in higher zones are progressively further away from the city centre. Although this structure is not strictly determined by geographic distance, inter-zonal average travel time increases proportionally to the number of traversed zones: Figure 2(a) shows the origin-destination matrix of average travel time between zones, shaded according to travel time (white = 0 minutes, black = 60 minutes). As expected, the longest trips tend to be those between Zones 1 and 6. Most notably, though, this matrix is not entirely symmetrical as we first expected. Lastly, we turned to the individual trips. Figure 2(b) is the overall trip time histogram: trip times in the tube network tend to follow a near-gaussian distribution with a long tail. From the data, we also computed trip time averages and standard deviations for each possible pair. Figure 2(c) shows the cumulative distribution of the trip time standard deviation. Approximately 90% of the trips are within ten minutes of the average; in fact, about 40% of the trips are within 5 minutes of the respective average trip time: this is an early indication that trip *mean time* is, in fact, a good predictor for how long it will take users to travel between a pair of stations.

A coarse view of London underground's usage patterns points to two main results: (a) the tube is used to commute to and from work, and (b) user travel times are very close to the mean travel times. This analysis would suggest the existence of a single *type* of traveller, with all users being similar to one another. In the following section, we demonstrate that this is not the case, by uncovering characteristics of the travellers that are indeed available in the data.

B. Traveller Characteristics

The above analysis limits the insights it provides to aggregate, system-wide behaviour. However, there are a number of differences that exist *between* users that remain unexplored. In this section, we highlight user-centric patterns of travel: repeat trips over time, usage similarities between different groups of users and relative travel times. The emerging differences in usage and travel time emphasise the potential that personalisation has to offer in this context.



(a) Repeat Trips Over Time (b) Trip Observations vs Stations Visited (c) Cluster Result Cumulative Start Times

Figure 3. (a) Proportion of Repeat Trips Over Time for each Dataset (vertical lines are weekends), (b) Number of users (top), and average number of stations visited (bottom), as trip observations increase, (c) User activity of different clusters.

1) ***Repeat Trips****:* In Figure 3(a) we show the temporal view of repeat trips: by the end of each dataset (83 days), approximately 60% of the user-trip pairs have been seen before (i.e., they are repeat trips). We also plot vertical lines in Figure 3(a) on weekend days. These correspond to the (small) dips in the measured proportion of repeated trips; users tend to be more regular in their movements during weekdays, as they commute to and from work. Commuting behaviour can be further illustrated by tracing the number of users per day whose last journey ends where their first journey began (i.e., their usage of the tube forms a round trip). For example, if a user travels from A to B, C to D, and lastly D to A, they form a full (albeit disconnected) circuit that begins and ends at A. We found that over 88% of the users per day form a loop with their travels. Note that this does not take into account multi-modal transport (e.g., if a user took a bus home), physical proximity of stations (e.g., if a user on their return trip exits at a station that is close to their original origin station but not the origin station itself), or loops that are formed over the course of multiple days. In fact, over 55% of the users who *did not* form a circuit with their trips took only 1 tube trip in that respective day.

While the global analysis pointed toward a system that is used in a highly regular way, users are not limited to commuting. Figure 3(b) plots the number of trips that a user takes against the average number of stations that the user has visited. As the number of trip observations increases, so does the breadth of stations that the user, on average, has visited. Interestingly, the graph appears to be segmented into two unique parts: it is composed of two point wise bounded parabolic growth functions with asymptotes corresponding roughly to 18 stations and 25 stations respectively. The pivot point between the two functions (at 110 trips) corresponds to roughly 2 trips/day on average in our dataset. Although we cannot be completely confident about why these three parts exist, it likely has to do with the primary reasons why travellers use the system (e.g., for commuting, for errands) and whether they have access to other forms of transit (e.g., a car, bicycle, or bus). Regardless, the key here is that the number of stations visited differs between travellers along with their frequency of travel.

2) ***User Activity Clustering****:* The overall average trips per user (in each of the full 83-day datasets) is 25.26 for D1, and 24.88 for D2 (both medians are 9), with approximately 9% of each set composed of users who we only see once. On the broadest level, users can be split into groups based on *when* they travel; whether they use the system throughout the entire week, or only during weekdays or weekends. As summarised in Table II, the smallest group is the users who *only* travel on weekends (approximately 8% of each dataset), while the largest group includes users who travel throughout the entire week including both weekday and weekend travel (56%). A noteworthy proportion of users (35%) travel only on weekdays.

We delved further into the differences that emerge in temporal usage patterns by clustering the users' weekday travel. We applied *dendrogram* clustering [5], a form of agglomerative hierarchical clustering, on vectors of binned user trip start times. More formally, we split the 24-hour day

	Users (%)		Avg Trips Per User	
	D1	**D2**	**D1**	**D2**
Week-End Only	7.71	7.61	2.33	2.32
Week-Day Only	35.40	35.81	10.73	10.45
Both Week/Weekend	56.89	56.58	37.41	37.05
Total	100	100	25.26	24.88

Table II
GROUPS OF USERS AND AVERAGE TRIPS PER GROUP: A SMALL PROPORTION OF USERS ONLY TRAVEL ON WEEKENDS.

into 5 segments using Figure 1(b); each segment represents a particular time of day (i.e., early morning, morning commute, day time, evening commute, late evening). For each user who has made more than a single trip, we construct a frequency vector denoting the number of trips started within each time segment. The similarity between the travel patterns of user a and b, who have respectively made a total of A and B trips, can then be computed as follows:

$$d_{a,b} = \frac{1}{5} \sum_i |\frac{a_i}{A} - \frac{b_i}{B}|$$

In each iteration of the clustering algorithm, the two users who share the *smallest* value of $d_{a,b}$ are merged; the resulting vector is the sum of the two users. Due to the high volume of users that each dataset contains, we clustered 10 uniform-randomly selected subsamples of $1,000$ users and average the results. The iteration stops when a pre-defined number of clusters have been formed: we manually tuned and examined the results of varying cluster thresholds, and settled on 6 as this produced a variety of user profiles that are different to one another.

Cluster Results. Based on the clusters found above, we plot the cumulative start times per profile in Figure 3(c). Note that these images are *not* plots of the cluster centroids themselves, but rather are the cumulative start times of all members of the given cluster. The six profiles that are produced are (reading from left to right): morning-only travellers, irregular travellers (with, for example, a peak immediately *after* the end of the evening rush hour), day-time-only travellers, users who travel most frequently in the evening, early-morning commuters (who go to work *before* the normal rush hour), and, lastly, the commuting majority. The clusters each vary in size, ranging from very few average users (2.6%) for the morning-only travellers to an average of 50.46% for the largest group—the commuting majority. This is why Figure 1, which gives the overall cumulative views of the system, only reflects one of the groups that were found via clustering: it reflects the group that was the largest. However, a noteworthy point is that the second largest cluster (with, on average, 45.6% of the users) is that of the day-time-only travellers; a significant portion of the population does *not* fit the two-spiked commuting pattern.

3) ***Travel Time****:* As above, we also explored the extent that travel time differences emerge between users. We first looked at the relation between trip *familiarity*, or the number of times that a user has taken a trip, and their trip time relative to the overall mean. Given a user u who, on average, completes a trip (o, d) in $\overline{u}_{o,d}$ minutes (while the overall trip average time is $\overline{m}_{o,d}$), the normalised residual $r_{o,d}$ is:

$$r_{u,t} = \frac{\overline{u}_{u,t} - \overline{m}_{o,d}}{\overline{m}_{o,d}}$$

A positive residual indicates that the user tends to be *slower* than the overall mean, while a negative residual shows that the user's average travel time is faster than the mean. We computed the residuals for all trip observations and plot the averages against the trip frequency in Figure 4. The results show that trip residuals are positive, on average, when the trip frequency is small (less than 3); however, as trip frequency increases beyond 3, the residuals become, on average, negative. While we cannot be certain that our first observation of a user's trip corresponds to the first time the user takes the trip, the overall results point to the fact that as users become more familiar with a trip (e.g., locals vs. tourists), they also tend to be faster in completing it.

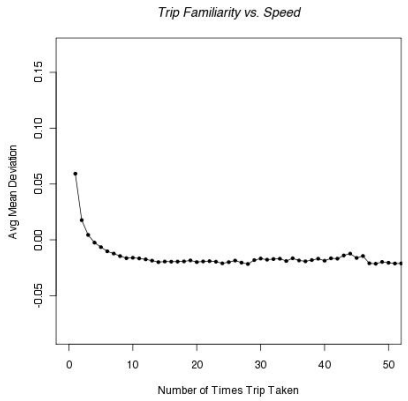


Figure 4. Trip familiarity vs. speed: comparing the trip frequency and trip time residual (notice how the residual becomes negative—users travel faster—as the familiarity with the trip increases)

In the following sections, we build models that estimate trip time by incorporating each user's travel history into the estimation algorithms. This perspective addresses many of the shortcomings of the above, such as accounting for trip familiarity and per-user transit speed differences, and aims at capturing the hidden variables that relate to the system users: their physical aptitude (i.e., their ability to move about the train station), their knowledge of the system, and their route choices.

III. Personalised Trip Time Estimation

In this section, we describe our proposals for computing personalised estimates of trip time. The goal here is to present travellers with more accurate estimates of their travel time than traditional aggregate estimates (e.g., the mean historical trip time between (o, d) pairs). Recall that our dataset contains a set of trip observations: the time it took a user to travel from an origin to a destination station when commencing the trip on a particular time and date. None of our proposed models below incorporate information about the London underground network topology (e.g., which stations are connected to one another via which trains), historical train arrival/departure data, service disruption histories, geographic distances between stations, route transfer data, station size, or train schedule information. Access to

any one of these additional datasets may very well improve personalised trip time estimation; however, in this work we investigate how well a personalisation system can perform with a simple, AFC-based dataset. We return to limitations of our dataset in the discussion.

Baselines. We have three available baseline estimates; two of these were discussed in the previous section. They include (a) the global mean trip time (roughly half an hour), (b) the inter-zone transit time $\overline{z}_{o,d}$, as pictured in Figure 2(a), and (c) the mean trip time between the station of origin and destination. More specifically, if we denote the set of N observed trip times between stations o and d as $T_{o,d}$, where each time is $x_{u,t}$ (the time it took user u when the trip was started at t), the predicted time for user u to travel from o to d is equal to the arithmetic mean $\overline{m}_{o,d}$ of observations to date:

$$\overline{m}_{o,d} = \frac{1}{N} \sum_{T_{o,d}} x_{u,t}$$

User Self-Similarity. The first assumption that we incorporate is that of user *self-similarity*: when users repeatedly make the same trip, they will tend to follow the same path within the system and therefore have similar travel time performance. We assume further that the speeds at which a traveler interchanges between lines and walks between the platform and station gates also tends to be consistent. We thus define $U_{o,d} \subset T_{o,d}$ as the set (of size M) of user u's trips between o and d, and $\overline{u}_{o,d}$ as the *user* mean time of these observations. In order to compensate for potential outliers in the user's set of trip times, we define $\overline{u}_{o,d}$ as the geometric mean of observed times:

$$\overline{u}_{o,d} = \left(\prod_{U_{o,d}} x_{u,t}, \right)^{1/M} = \exp\left(\frac{1}{M} \sum_{U_{o,d}} \ln x_{u,t} \right)$$

If the user has not taken the given trip before, then the overall trip mean is returned. More generally, the reliability of the mean computed by the moving average will be proportional to the number of trip observations M that are available. We use M as a weight: the personalised prediction $\hat{p}_{u,o,d}$ can thus be computed as a $\frac{1}{M}$-weighted combination of the baseline and the user mean:

$$\hat{p}_{u,o,d} = \left(\frac{1}{M} \times \overline{m}_{o,d} \right) + \left(\left(1 - \frac{1}{M} \right) \times \overline{u}_{o,d} \right)$$

Trip Familiarity Model. In the previous section, we discussed the relation between the number of times M we have observed a user taking a trip (which we use to quantify how *familiar* she is with it) and the average time it takes her to complete it. Figure 4 displays the relation between familiarity and trip time residual: on average, trip time is inversely proportional to familiarity. These observations translate into a predictive model as follows. Given a user u who has taken a trip between o and d M times, we define $F_{o,d}$ as the set of *user* mean times $\overline{u}_{o,d}$ of all users who have familiarity f_u that is at least M. The weighted average of all members of this set forms the personalised prediction:

$$\hat{p}_{u,o,d} = \frac{\sum_{F_{o,d}} (\overline{u}_{o,d} \times f_u)}{\sum_{F_{o,d}} f_u}$$

The intuition behind this model is to partition, for each traveller, all other users into two groups and use the most relevant one, in terms of familiarity, to compute trip time.

Trip Context Model. The third method we examine aims to uncover any similarities between users based on the temporal context of their travel. Context is a broad term that may refer to any of a number of characteristics: these include, for example, whether the user is commuting to or from work or travelling during non-peak hours, congestion, and the underlying average service availability. In this model, we do not explicitly formulate or quantify the precise context, but instead assume that users who begin travelling at the *same time* implicitly experience similar contexts. In other words, given a time interval of size w, we assume that all users who travel from o to d, starting their trip within a window of size $2w$ centred on t, are similar to the user u who travels from o to d at time t. For each prediction $\hat{p}(u, o, d, s)$, we define the set of trip times $W_{o,d} \subset T_{o,d}$, where each observation $x_{u,t}$ fits the condition:

$$(s - w) \leq t \leq (s + w)$$

The times in $W_{o,d}$ can be used to estimate the user's trip time by computing their (geometric) mean $\overline{w}_{o,d}$, as above. Note that this is the only method that considers a user's intended start time when formulating a prediction so that, for example, time-of-day factors such as commuting congestion and train arrival/departure frequency are implicitly incorporated. Broadly speaking, this method is a simple two sided moving average. However, given the sparsity of our underlying dataset, there may be a substantial number of travel windows for which we have little to no data for $\overline{w}_{o,d}$. To account for this, we implement a weighting scheme that is similar to that used in the self-similarity method above. Given a trip, for which we have M observations and N trips within the pre-defined window, personalised predictions are defined as a $\frac{N}{M}$-weighted linear combination of the baseline and window mean:

$$\hat{p}_{u,o,d,s} = \left(\frac{N}{M} \times \overline{m}_{o,d} \right) + \left(\left(1 - \frac{N}{M} \right) \times \overline{w}_{o,d} \right)$$

Combined Model. A pervading facet of personalisation techniques [6] is that *combined* models tend to predict preferences more accurately than any single model. However, in our case, traditional techniques such as linear regression failed to produce improved results; a potential cause for this is the varying sparsity in each personalised prediction method. Instead, we adopt a *chaining* approach to combine methods. Each of the proposals above relies on a *baseline* to resort to when there is insufficient data. For example, if

a user has taken a trip before, we can use the self-similarity method. However, if a user has not previously taken the trip, the mean time is returned; if the trip has never been taken before by any user, the zone transfer time is used. In our combined approach, we iteratively replace the baseline of one method with the output of another. For example, given a trip between o and d with N observations, the zone transfer and trip mean can be combined:

$$\overline{mz}_{o,d} = \left(\frac{1}{N} \times \overline{z}_{o,d}\right) + \left(\left(1 - \frac{1}{N}\right) \times \overline{m}_{o,d}\right)$$

This new baseline can then be used in the self-similarity approach:

$$\hat{p}_{u,o,d} = (\beta_u \times \overline{mz}_{o,d}) + ((1 - \beta_u) \times \overline{u}_{o,d})$$

In our experiments, we chain the methods as follows: the zone transfer is combined with the trip mean ($\overline{mz}_{o,d}$), which becomes the baseline for the trip context model, which is then fed into the user similarity approach.

IV. Trip Time Estimation Evaluation

In order to test the predictive power of the models above, we split our data into training and test sets using the trip time stamps. We used the last 9 days of data for testing (approximately a 90% training—10% test split). The sets contained a small proportion of both users and origin-destination pairs not been seen before; we pruned both of these from the test sets since our proposals focus on having some data regarding users and we assume that access to more data would have resolved this issue.

We used two metrics to quantify trip prediction error: the Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). These metrics are commonly used when evaluating trip prediction estimation [7], and are defined as follows. Given N predictions $\hat{p}_{u,o,d}$, each of a user u's trip time between origin o and destination d, with actual time $x_{u,o,d}$, the MAE is the average of the absolute deviations from the actual trip time:

$$MAE = \frac{1}{N}\left(\sum_N |\hat{p}_{u,o,d} - x_{u,o,d}|\right)$$

The MAE is in the same units as the predictions (minutes). The MAPE, instead, is the average relative error of each prediction, measured as a percentage:

$$MAPE = \frac{100}{N} \times \sum_N \left(\frac{|\hat{p}_{u,o,d} - x_{u,o,d}|}{x_{u,o,d}}\right)$$

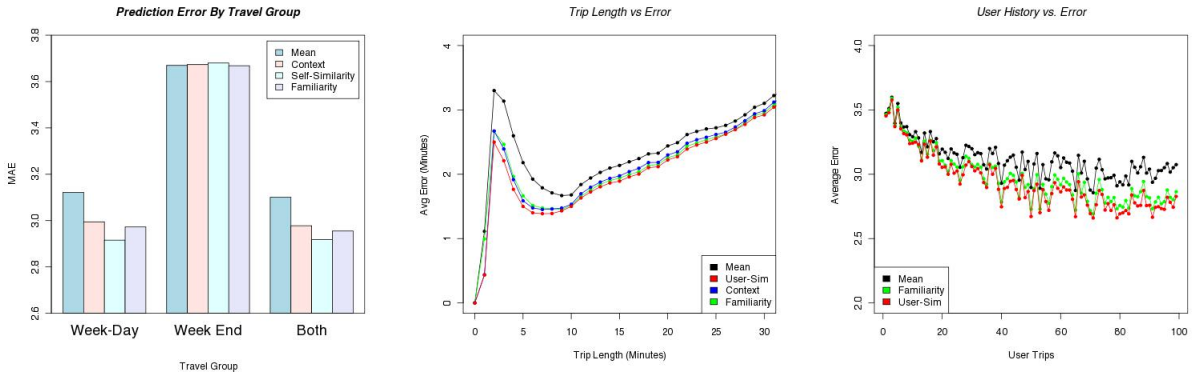
Results. The results for each model over the two datasets is shown in Table III. The global and zone mean transfer baselines perform the worst: MAE values range between 8 and 12 minutes. However, the mean trip baseline provides relatively accurate results: over both datasets, it consistently produces the lowest MAE and MAPE values. All of the proposed personalisation methods, while producing MAE results that lie between 2 and 3 minutes, outpredict the best baseline; the self-similarity approach, when used alone, is the most accurate. The combined model produces the most accurate results overall but there are diminishing returns: the accuracy gain when the methods are chained together is very low. Overall, these simple methods outpredict the baseline estimates by, on average, 11.22 seconds for D1 and 5.64 seconds for D2. However, the underlying predictions account for a wide range of users and trips. We therefore also analysed how our proposed models compared to baseline estimates for different traveler groups, trip lengths, and data amounts. Recall that travellers can be segmented into three different groups based on weekday-only travel, weekend-only travel, or both. We compare the predictive accuracy of each of these groups in Figure 5(a). The error is not distributed evenly between the groups; those who only travel on weekends show consistently high MAE values. In fact, the personalised methods do not outstrip the baseline for these users.

	MAE (Mins)		MAPE (%)	
Method	**D1**	**D2**	**D1**	**D2**
Baselines				
Global Mean	11.454	11.981	65.56	63.49
Zone Mean	8.561	9.215	46.878	47.36
Trip Mean	**3.109**	**3.650**	**13.306**	**14.02**
Personalised Models				
Trip Context	2.986	3.601	12.33	13.42
Familiarity Model	2.989	3.599	12.28	13.37
Self-Similarity	**2.924**	**3.556**	**11.97**	**13.17**
Combined Model	2.922	3.556	11.95	13.17

Table III
Prediction Mean Average Error and Mean Average Percentile Error Results

With regards to how prediction error varies with trip length, we hypothesised that individual differences in off-train time (e.g., time spent moving about a station) would be more pronounced for shorter trips (as off-train time would account for a more significant proportion of overall travel time). Figure 5(b) plots the relation between the actual time the trip took $(t_d - t_o)$ and the prediction error. It shows that the simple proposals that we put forward improve predictions from the baseline in all trips; moreover, the improvements are indeed greater for shorter trip lengths.

Lastly, we examined the extent that predictions improve as users continue travelling on the tube. Two of the methods that we proposed (the user-self similarity and familiarity methods) rely on augmenting predictions with the number of trips a user has taken. In Figure 5(c), we plot the average error against the trip frequency: these results clearly establish the potential benefits of personalisation, since the divergence between the baseline and the personalised approaches grows as travellers continue to use the system.



(a) Prediction Error By Group (b) Prediction Error By Trip Time (c) Prediction Error As Users Take More Trips

Figure 5. Decompositions of prediction error, by group (week-day only, week-end only, week-long travellers), by trip time, and by travel history size. Notice that week-end only travellers are consistently receiving the highest error.

The goal of this evaluation was to verify that personalised trip estimation methods that take into account facets of the *user* who is travelling from an origin to a destination can outperform a mean trip time predictor. We found that, while mean trip times already provide good estimates of travel time, simple personalisation techniques can produce more accurate results. However, these results do not hold over the entire population: there are some users who are consistently given bad predictions. This point does not detract from our objective; in fact, it enhances the need to personalise traveller's transit experience further.

V. Station Interest Ranking

We now turn to the second prediction problem. This relates to identifying the subset of the transport system that each user is interested in. In this case, we define interest according to visit frequency: a user who continuously frequents particular stations (i.e., where trips originate or end) is likely to be interested in any disruptions affecting those stations. This problem can be formulated in many different ways. For example, it could be viewed as a binary classification problem: given a user u, will u travel to or from station s at time t? However, due to the broad nature of disruptions that can occur (e.g., planned outages, unexpected delays, passenger incidents) coupled with the wide availability of different routes available to travel between stations, we are not just interested in knowing whether a user will travel between two stations at a given time, but whether information affecting those stations is, more generally, relevant to that user. The alternative is, therefore, to view this context as a ranking problem: given a user u, what are the stations that u is most interested in? More formally, given the same dataset of trip observations (i.e., user u taking trip with origin o and destination d at time t with trip length x), we would like to create, for each user, a personalised *ranking* of the stations $s \in S$ that will reflect their future mobility patterns. We describe three such methods:

Baseline. The baseline generates the same ranking for each user, by sorting the stations according to *popularity*: stations that are frequented the most are ranked higher than those that are visited less. In other words, the score $\hat{r}_{u,s}$ given to station s (for user u) is strictly determined by the number of trips originating at s (o_s), and ending at s (e_s), normalised by the total number of station visits N:

$$\hat{r}_{u,s} = b_s = \frac{o_s + e_s}{N}$$

User History. The user history method augments the baseline to include higher weights for stations that the user u has visited in the past. The preference that a user u has for station s is the proportion of u's trips that originate or end at s:

$$h_{u,s} = \frac{o_{u,s} + e_{u,s}}{N_u}$$

We produce a final score for each station as a sum of the baseline and proportion of u's station visits that were to s:

$$\hat{r}_{u,s} = b_s + h_{u,s}$$

A natural extension to this ranking method would be to incorporate the *recency* of trips into the scoring function, allowing stations that have been visited frequently *and* more recently have a higher score. However, due to the limited time span of our data, we leave a thorough investigation of this point as future work.

Station Similarity Model. The final model we propose is a further refinement of the above. Given our dataset, we can create a co-occurrence station matrix C, where each entry $c_{i,j}$ is the frequency count of trips that have stations i and j as their endpoints (regardless of direction). Higher values of $c_{i,j}$ thus denote stations that are similar to each other, in that users frequently travel between them (note that this definition revolves around usage similarity rather than geographic proximity). C is symmetrical along the diagonal, and $c_{i,j} = 0$ if $i = j$ or no user has ever travelled between

i and j. The $c_{i,j}$ values are normalised with each row sum, to produce normalised matrix W, with entries $w_{i,j}$.

This matrix can then be used to increase the score of stations that are similar to those that the user has previously visited. Given a station s in the set of stations S_u visited by user u, we increase the score of all of s's neighbours n in N_s by $w_{s,n}$. The final weighting for each station is produced as follows:

$$\hat{r}_{u,s} = b_s + \beta h_{u,s} + \sum_{s \in S_u} \left(\sum_{n \in N_s} w_{s,n} \right)$$

We note that since a particular subset of stations (mainly in central London) are popular destinations, then these neighbours may appear in more than one of u's station's neighbourhoods: its weights may be increased more than once. This feature had a pronounced effect on the results, by detracting from the gain achieved with the user history weights $h_{u,s}$. In order to accommodate for this, we give a higher weight β to the history weights $h_{u,s}$ when computing the final score.

This model is, in effect, a similarity thresholded nearest neighbour approach, where distance is based on usage similarity. The choice of a station neighbourhood model, rather than a user neighbourhood model, has the benefit of being highly scalable: the data contains less than 300 stations, with hundreds of thousands of users; the station-oriented model thus has significantly less computations to perform when measuring the similarity between any two stations.

VI. Interest Ranking Evaluation

In this section, we evaluate the extent that our proposals correctly identify stations that are of interest to each user. However, our data does not include the full picture: we do not know the *actual* stations that each user is interested in and, thus, we can only infer interest from each user's travel history. In fact, since we also do not know the particular routes that users take, we cannot be sure about relevant station notifications *along* a route; due to this, and since these stations may be inferred by inspecting the tube graph, we omit them from this evaluation. In order to accommodate for this, we adopt a measure that has previously been used to evaluate personalised content ranking on the web: the average percentile ranking [6].

Given the same pre-defined test period as above, we define the *interest* that a user u has in station s as the proportion of times that the user starts or ends a journey at the station throughout the test period. In other words, if a user takes only 1 trip in the test period (and thus visits two stations, the origin and destination), the measured interest in the each station will be 0.5; interest values lie in the range $[0, 1]$. We assume that interest, defined in this way, is correlated to each user's actual station preferences. Note that, unlike above, in this section we restrict our test set to those users who have appeared at least once in the training period.

	Percentile-Ranking	
Method	**D1**	**D2**
Baseline	0.2467	0.2561
User History	0.0642	0.0611
Station Similarity	**0.0591**	**0.0555**

Table IV
Station Interest Ranking Results

We define $rank_{u,s}$ as the percentile ranking of station s for user u in the ranked list of stations; if $rank_{u,s} = 0$, then the station appears first in the list, while $rank_{u,s} = 1$ implies that the station was the last in the list. We combine these with each user's interest in the station $interest_{u,s}$ and average the results:

$$\overline{rank} = \frac{\sum_{u,s} interest_{u,s} \times rank_{u,s}}{\sum_{u,s} interest_{u,s}}$$

This measure is independent of the actual size of the ranked list, and produces values between 0 and 1. Lower values are inherently better; they reflect the case where stations that will be frequently visited (high interest) are ranked higher.

Results. The percentile-ranking results, for each dataset, are shown in Table IV. The baseline produces lists with a percentile ranking of approximately 0.25. Most notably, both personalised methods provide a large improvement over the baseline by reducing the percentile ranking to less than 0.07. The station similarity method ultimately produces the best results; percentile ranking is just below 0.06. These figures tell us that using the above techniques to rank station notifications would significantly improve the relevance of information residing in the top-ranked places: by using AFC data, we can give users important notification updates without any further input from them.

VII. Related Work and Discussion

There is a broad literature on predicting travel time, ranging from bicycle rides [8] to car trip [9] duration. A simple yet important point differentiates our work from many in this domain: current solutions either do not have access to per-user data [8] or explicitly focus on the aggregate usage [3]. We take a personalised perspective instead: we mine public transport usage data to uncover individual characteristics of travel behaviour, and then leverage it to build user-tailored travel time estimates.

Our data also prevents us from building trajectory or sub-route-based models (e.g., [7] or [5]) since the *actual* route that a user undertakes between any origin and destination is unknown to us; in many cases, there are a wide variety of candidate routes. Implementing heuristics to derive route choices (for example, minimising the number of interchanges or minimising the hop-count on the tube graph) does not resolve cases where two routes seem equal on the

applied heuristic (e.g., they both have one interchange) or when the heuristic derives results where travel time may increase (e.g., in cases where changing line would have reduced travel time). An important area of future work will be to incorporate additional contextual information such as the London underground network topology, train scheduling information, and service disruption history in order to further bound our estimates and assist in identifying anomalies.

This paper also adds to the growing body of work on smart cities or urban informatics [10], which is the study of human behaviours and urban infrastructures made possible by the increasingly digitised and networked city. For example, Gonzalez et al., 2008 [11] and Ratti et al., 2008 [12] use mobile phone-based location data to study human mobility patterns; Kostakos et al., 2006 [13] and Sadabadi et al., 2010 [14] rely on distributed Bluetooth receivers to track and predict travel speeds based on the Bluetooth MAC identifiers of passing devices. Most of the cited work, however, continues to focus on aggregate analysis rather than attempting to uncover opportunities for personalisation services.

Personalisation has been a key component of web-based systems; the most prominent example is its use for recommendation in e-commerce [15]. Such systems often rely on *collaborative filtering* algorithms [6], which automatically compute personalised rankings of e-commerce items based on the predicted interest a user will have for each one of them. A noteworthy point of these methods is that they use measured similarities across items (in our case: between two *different* trips), in order to formulate predictions. Our methods, instead, focus on inter-user similarity within a single trip. In fact, preliminary analysis of the data shows that relative transit speed is *not* consistent: just because a user travels quickly between an origin and destination, does not mean that s/he will continue to be faster than others on a different trip.

One of the key aspects of successful recommender systems is that they tailor information in transparent way; users should be able to infer *why* they are being recommended what they receive. Our trip estimation proposals above come with the same benefit: they not only allow for more accurate predictions, but also *reasons* why those predictions may be correct. For example, the self-similarity model justifies any prediction it makes based on the average time that it previously took the same user—these points can be used to directly enhance the experience that travellers have with personalised route planning systems. Such transparency can also be used to inform travellers about how their data is being used, to cater for those who may be wary or have privacy concerns.

VIII. Conclusion

This paper is the first to explore the potential opportunities that AFC datasets provide for personalisation services. Our in-depth analysis of two large datasets of London's tube usage demonstrated that there are substantial differences between travellers that emerge from this usage data. Although users remain highly regular, and often form loops in their daily travels, there are a variety of different usage groups, ranging from people who only travel during week days to clusters of users who commute *before* the morning rush hour. Based on these insights, we proposed three simple personalised prediction methods (trip familiarity, user self-similarity, and trip context). We showed that not only do these methods outpredict the baseline, but tend to improve over time as travellers continue to use the system. We also proposed a means of *ranking* stations to match traveller's interests; we showed that by tracking where users have been, systems that anticipate where they will go can be built. Ultimately, the key prediction—whether it be trip time or station interest—will be application-specific; further prediction problems can be defined based on the information needs and goals of each traveller. However, regardless of the prediction at hand, the key conclusion remains: to incorporate data about system users to build personalised intelligent transport systems.

References

[1] N. Wilson. The Role of Information Technology in Improving Transit Systems. In *Transportation@MIT Seminar*, Boston, USA, 2009.

[2] K. Chu and R. Chapleau. Enriching Archived Smart Card Transaction Data for Transit Demand Modeling. *Journal of the Transportation Research Board*, (2063), 2008.

[3] J. Chan. Rail Transit OD Matrix Estimation and Journey Time Reliability Metrics Using Automated Fare Data. Master's thesis, MIT, Department of Civil and Environmental Engineering, June 2007.

[4] W. Jang. Travel Time and Transfer Analysis Using Transit Smart Card Data. *Journal of the Transportation Research Board*, (3859), 2010.

[5] J. Froehlich and J. Krumm. Route Prediction From Trip Observations. In *Intelligent Vehicle Initiative, SAE World Congress*, Detroit, Michigan, 2008.

[6] Y. Hu, Y. Koren, and C. Volinsky. Collaborative Filtering for Implicit Feedback Datasets. In *IEEE ICDM*, 2008.

[7] H. Van Lint. Empirical Evaluation of New Robust Travel Time Estimation Algorithms. In *89th Annual Transport Research Board*, January 2010.

[8] J. Froehlich, J. Neumann, and N. Oliver. Sensing and Predicting the Pulse of the City through Shared Bicycling. In *21st International Joint Conference on Artificial Intelligence*, Pasadena, California, 2009.

[9] S. Handley, P. Langley, and F.A. Rauscher. Learning to Predict the Duration of An Automobile Trip. In *4th International Conference on Knowledge Discovery and Data Mining*, New York, New York, 1998.

[10] M. Foth, editor. *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*. 2009.

[11] M.C. Gonzalez, C.A. Hidalgo, and A-L. Barabasi. Understanding Individual Human Mobility Patterns. *Nature*, 453(7196):779–782, 2008.

[12] C. Ratti, R. M. Pulselli, S. Williams, and D. Frenchman. Mobile Landscapes: Using Location Data From Cell Phones for Urban Analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748, 2006.

[13] V. Kostakos, T. Kindberg, and et al. Instrumenting the City: Developing Methods for Observing and Understanding the Digital Cityscape. In *In Proc. of the 8th International Conference on Ubiquitous Computing*. Springer, 2006.

[14] K. Sadabadi, M. Hamedia, and A. Haghani. Evaulting Moving Average Techniques in Short-Term Travel Time Prediction Using an AVI Data Set. In *Transportation Research Board 89th Annual Meeting*.

[15] G. Linden, B. Smith, and J. York. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEE Internet Computing*, 7:76–80, 2003.