ARTICLE IN PRESS

Pervasive and Mobile Computing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Individuals among commuters: Building personalised transport information services from fare collection systems

Neal Lathia [a,*], Chris Smith [b], Jon Froehlich [c], Licia Capra [b]

[a] *Computer Laboratory, University of Cambridge, UK*
[b] *Department of Computer Science, University College London, UK*
[c] *Department of Computer Science, University of Maryland, USA*

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Public transport
Personalisation
Urban data mining

ABSTRACT

This work investigates how data from public transport fare collection systems can be used to analyse travellers' behaviour, and transform travel information systems that urban residents use to navigate their city into personalised and dynamic systems that cater for each passenger's unique needs. In particular, we show how fare collection data can be used to identify behavioural differences between passengers: we thus advocate for a personalised approach to delivering transport related information to travellers. To demonstrate the potential for personalisation we compute trip time estimates that more accurately reflect the travel habits of each passenger. We propose a number of algorithms for personalised trip time estimations, and empirically demonstrate that these approaches outperform both a non-personalised baseline computed from the data, as well as published travel times as currently offered by the transport authority. Furthermore, we show how to easily scale the system by pre-clustering travellers. We close by outlining the wide variety of applications and services that may be fuelled by fare collection data.

1. Introduction

The proliferation of pervasive technology throughout urban environments, such as smart traffic meters, location-based bus systems, and cellular data, increasingly provides realtime information about city dwellers' location and mobility. This newfound availability of rich streams of urban data has led researchers to investigate its potential to aid urban planners, designers, and policy makers. Aggregate spatio-temporal analysis of urban data – often referred to as the "pulse of the city" – can reveal how people use and move between designed spaces and uncover otherwise hidden urban flows and structures. For example, geo-tagged collections of photographs have been used to identify tourist attractions [1], GPS traces from taxis have been used to inform future urban planning [2], and shared-bicycle station occupancy data has been used to aid the management of bicycle fleets [3]. In this work, we examine how a similar set of data can be used to provide *personalised* transit information services to individual city dwellers. We focus on one domain where data containing individuals' mobility patterns has yet to be leveraged to serve peoples' needs: public transport. Historically, public transport information systems have been centred on the system itself (e.g., location of the transit service, scheduled and estimated arrival times of trains or buses). In the following, we seek to explore how fare-collection data, which captures how individuals *use* public transport, can both uncover differences between individual travellers as well as be leveraged to provide more accurate information relating to their travels. The role of personalisation is particularly critical in this context: by providing a service that suits

* Corresponding author.
E-mail addresses: neal.lathia@cl.cam.ac.uk (N. Lathia), c.smith@cs.ucl.ac.uk (C. Smith), jonf@cs.umd.edu (J. Froehlich), l.capra@cs.ucl.ac.uk (L. Capra).

1574-1192/$ – see front matter
doi:10.1016/j.pmcj.2012.10.007

individual traveller needs and preferences, public transport satisfaction and usage will increase, which will invariably reduce congestion and pollution in city streets.

To date, two challenges have prevented the deployment of personalised information services in transport domains. First, user preference data has been considered notoriously difficult to collect without resorting to manual intervention [4]. Second, even when user preferences are available, leveraging them successfully remains an open challenge. Recent developments, however, offer new possibilities that promise to overcome these obstacles. On one hand, the introduction and widespread adoption of automated fare collection (AFC) systems throughout the world's urban public transport networks provide an opportunity to collect users' preferences and commuting habits (e.g., travel times, transport modalities) of millions of travellers in an implicit and highly scalable manner. On the other hand, personalisation has become a mainstream and highly active area of research in the context of Internet services, ranging from e-commerce, music, and news recommendation [5,6] to mobile location-based services [7]. Thus, there exists a vast toolkit of algorithms that have been successfully used on web-based preference data that can potentially be applied to urban transport networks. In this work, we address questions that relate to linking these two fields of research: what does AFC data tell us about *individual* traveller preferences and what sort of personalised travel services can be built from this data with personalisation algorithms?

1.1. Objectives and key results

In the following sections, we analyse a large, multi-modal, anonymised fare collection dataset of travellers using the Transport for London (TfL) public transport system. We do so from three perspectives: (i) the bird's-eye, system-level view, (ii) comparing individuals to each other, and (iii) comparing individuals' transit times to the published travel times from two selected stations. In doing so, a number of key results emerge: first, whilst an analysis of system-wide aggregate data would suggest all users follow the same pattern, individuals' usage of public transport is not uniform. Fare collection data can be used to uncover a series of emergent usage patterns that show a variety of travel habits. These more subtle usages emerge when clustering travellers based on their temporal patterns in the transit network.

Furthermore, we find that published travel times (which fail to take individual differences into account) substantially diverge from, and, in particular, underestimate, actual travel times even when accounting for transfers. Based on these insights, we propose and evaluate a set of algorithms that provide travellers with personalised travel time estimates, and perform an in-depth spatial, temporal, and user-analysis of where estimation errors emerge. We close by proposing and discussing a wide range of personalised services that can be built using this data. We also formalise a number of prediction and ranking problems that underlie these applications, and lay the foundations for future work by discussing the potential of augmenting AFC data with alternative repositories of traveller preferences.

1.2. Extensions to prior work

An earlier version of this work appeared in [8]. This paper extends our prior work in four main ways: first, we offer a more comprehensive analysis of travellers' behaviours to gain valuable insights into why individual's travel times vary. For example, we study trip distributions, temporal views of repeat trips, trip familiarity, as well as the travel patterns that emerge when clustering users based on travel time. We also conduct a novel enquiry that compares actual travel times (as measured by the fare collection data) with published inter-station travel times as they appear on posters affixed at stations, further highlighting the differences between static estimates and individual variations in transit times. Second, we compare our predictive techniques with static time-table based estimates, as offered by current travel planners; this new comparison was made possible after having obtained, via a public Freedom of Information request, detailed information about inter-station transit times for the entire London rail network. Third, we evaluate our prediction techniques on a complete dataset, that covers all users that travelled through the London rail network during a one-month period, as opposed to working with samples of 5% of users as done before; we thus avoid biases that may have resulted from the sampling technique used—something which was outside our control. Finally, we analyse prediction errors from new angles: whilst in [8] we analysed error in relation to simple grouping of users, time of travel, and familiarity with a given route, we now offer a detailed analysis of the temporal, spatial, structural, and user-trip characteristics of the experiments that we performed.

2. Smart card data from Transport for London

In this section, we describe the context of our study: measuring the differences that emerge between travellers who use the public transport infrastructure in London, England. We first give a brief overview of the geographical span and reach of the infrastructure (Section 2.1) and of the fare collection system in use (Section 2.2); we then describe the data that we obtained from the transport authority (Section 2.3).

2.1. Transport for London

The public transport system in London consists of several interconnected subsystems, incorporating multiple modes of transport. These include the London Underground, the Overground rail system, the Docklands Light Railway system (an automated train network operating in the east of the city), an extensive bus network, waterborne transport and portions of the UK National Rail network. The London Underground is made up of 11 lines totalling 402 km of track, connecting 260 stations[1] whose locations range from central London to as far as Chesham, Buckinghamshire, which is approximately 40 km from central London. Despite the name, only 45% of this network is actually underground. The transport network is separated into 9 fare zones, with Zone 1 encompassing central London and higher numbers representing regions further away from the centre of London, up to Zone 9, which contains a handful of outlier stations, including Chesham. The zoning system forms part of the fee structure for all rail travel in London, as well as approximating geographical distance from the centre of London.

2.2. The Oyster Fare Collection system

Automated Fare Collection (AFC) systems are in use in a number of cities throughout the world including Seoul, South Korea, Buenos Aires, Argentina, Rostov, Russia, and Chicago, USA. In 2003, TfL introduced its own: an AFC system which uses RFID-based smart card tickets called Oyster cards. The aim of this system is to replace all paper-based tickets with a single ticket that can operate across modalities: by 2008, TfL had issued over 16 million cards [9]; by 2009, this system accounted for approximately 80% of all public transport trips in the city [10]. Detailed information about each trip is captured when an Oyster card is used to enter or exit the public transport network. Beyond eliminating paper-based tickets, this automated system introduces an important benefit: sequences of individual travellers' trips can now be linked, allowing the transport operator to build a fine grained record of all people's movements within their network.

2.3. Oyster data overview

Our Oyster card dataset contains every single journey taken on the system's rail services using smart cards throughout the 31 days of March 2010 (which contains 23 weekdays and 8 weekend days). This amounts to roughly 89 million journeys, of which 70 million are tube journeys, with the rest made up of trips taken on National Rail, Overground and other rail systems. Each record details the day, anonymised user id, the origin and destination stations, entry time and exit time (measured as accurately as the minute of entry/exit), as well as departure and arrival zone. We note, however, that the dataset does not contain information regarding the actual route taken by individuals within the system but only their entry and exit points. Given the size and complexity of the London Underground network, there are often several distinct routes than can be taken between any two stations. To that end, when we discuss a "trip" or "journey", we mean a row of this data, which corresponds to an entry and exit of the system—we do not link or reason about adjacent entry/exit pairs of each user (which may indeed represent a multi-hop journey). Note also that in this paper we use "trip" and "journey" interchangeably.

We took two steps to clean the data. First, we removed any entries containing erroneous or inconsistent fields: entries were removed if the start time was earlier than the end time, or if the origin and destination were the same. Second, we found that certain stations were represented by more than one unique identifier: these were consolidated into a unique id number for each station. Approximately 2.5% of journeys were found to have end times earlier than the start time; a further 1.5% had the same origin and destination station and 0.5% of entries were repeated trips with the same start time but multiple end times. There were a number of entries that we did not remove because we could not determine whether these were genuine or erroneous. These include, for example, trips lasting less than two minutes or as long as twelve hours. In these cases, we opted to retain the data because we did not want to risk altering the representativeness of the dataset. After cleaning the data, we are left with 96.4% of the original data, amounting to 76.6 million trips by 5.1 million unique users—an average of 2.47 million journeys each day. We present a detailed analysis of this dataset next.

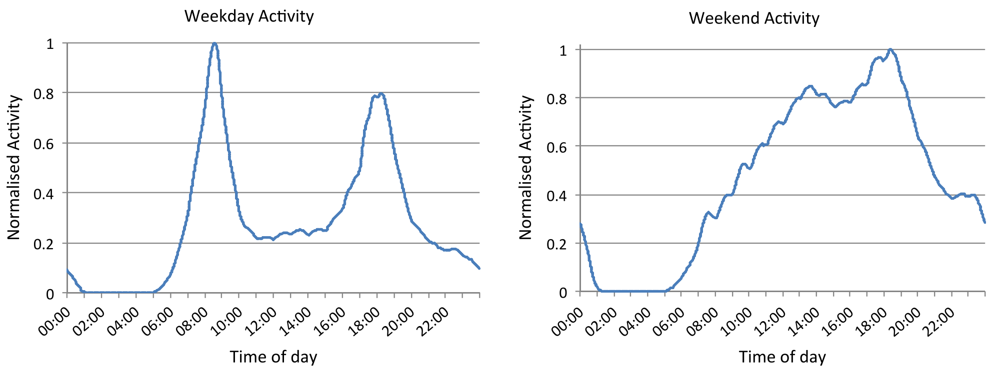
3. Analysing individual differences with Oyster data

There are three main hypotheses underpinning this work: (i) there exist fundamental differences between individuals' usage of public transport systems; (ii) these differences can be discovered by analysing AFC datasets; (iii) and, finally, that these individual differences can be used to build personalised travel services. In this section, we aim to validate the first two hypotheses, by first showing how marked individual differences exist in the Oyster card dataset, and secondly by demonstrating that these differences do matter when, for example, it comes to calculating travel times. We will then validate the third hypothesis in Section 4.

We first look at the data as a whole, which elicits the one-size-fits-all traveller's stereotype that emerges from an aggregate, system-level, perspective (Section 3.1). We then reveal how this bird's eye view of the data masks an underlying variety of traveller characteristics (Section 3.2). Indeed, by using agglomerative clustering on the data, we highlight how different travel habits emerge, each describing a different traveller's "persona" with their own travel pattern (Section 3.3). Finally, having demonstrated that individual differences in travel patterns do exist, we quantify their magnitude when calculating travel time: first, by comparing individuals' actual travel times with respect to those of the whole crowd

[1] http://www.tfl.gov.uk/corporate/modesoftransport/londonunderground/1608.aspx.

Please cite this article in press as: N. Lathia, et al., Individuals among commuters: Building personalised transport information services from fare collection systems, Pervasive and Mobile Computing (2012), doi:10.1016/j.pmcj.2012.10.007



(a) Week day. (b) Week end.

Fig. 1. Aggregate temporal views of (a) weekday and (b) weekend rail activity in London based on our AFC dataset. Scheduled services typically run between 5:00 and 1:30 AM. Note the dominant two-spike commuter pattern present on weekdays but absent on weekends. Furthermore, the scale of (a) and (b) is not the same: in general, there is much less activity during the weekends.

(computed by aggregating all records in the AFC dataset—Section 3.4); second, by comparing static travel time estimates, published in posters affixed in tube stations, with respect to actual travel times as recorded in the AFC dataset, for trips taking place within the Victoria tube line (Section 3.5).

3.1. *The pulse of London: an aggregate system-level view*

We begin our analysis by examining the aggregate usage patterns of the system. Later, we will contrast these aggregate results with clusters of individualised travel behaviour. To quantify the system usage over a given period of time (e.g., a day) we define a metric of *activity*. First, we divide the period into a vector of equal sized time bins. Next, for each trip that starts at time t_o and ends at time t_d we increment each bin b_i where $t_o \leq b_i < t_d$. Finally, we normalise the vector entries by dividing by the maximum observed value; this gives us a value between 0 and 1 representing the activity level at any one time. We show the results for weekday and weekend patterns in Fig. 1. Note that the normalisation was calculated independently for the weekday and weekend graph (i.e., the y-axes between the two are not interchangeable).

The pattern in Fig. 1(a) clearly reflects commuting behaviour. This behaviour disappears during weekends (Fig. 1(b)), when a gradual increase in activity throughout the day emerges, peaking at approximately 18:25 before trailing off into the evening. The lowest activity levels occur between 2:00 and 5:00 AM when rail services are closed. For most of the open hours during the weekday, the minimum activity level is just over 0.2—thus even during off peak periods the transit network is fairly well utilised. This mid-day lull is not apparent in the weekend graph.

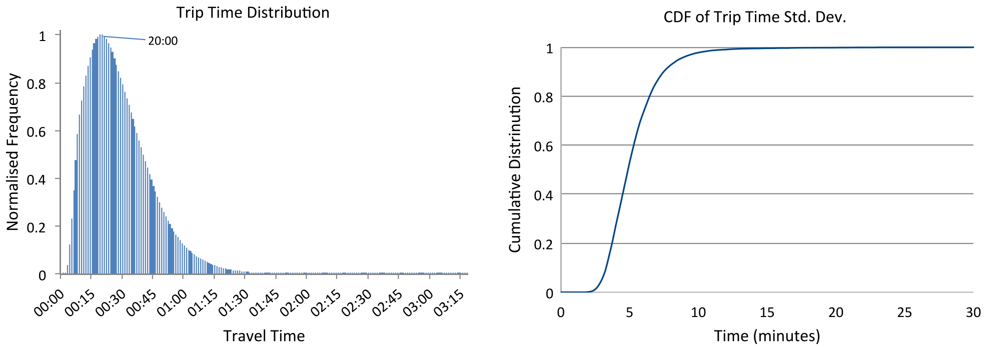
Given that our AFC data also provides trip entry/exit timestamps, we can also examine aggregate patterns in travel times in the London Underground network. From each trip tuple with start (entry) time t_o and end (exit) time t_d, the trip duration is simply $(t_d - t_o)$ minutes. The global mean trip time measured with the data is 28.41 min ($SD = 15.83$ min). Fig. 2(a) shows the distribution of trip times which roughly follows a normal distribution with a long tail. To further explore the level of consistency in trip times when the system is viewed as a whole, we also calculated the mean and standard deviation for trips between each possible pair of stations. Fig. 2(b) shows the cumulative distribution of standard deviations. Approximately 86% of trips have observations with a standard deviation of less than 10 min, and around 32% have a standard deviation of less than 5 min. This indicates that despite the complexity, size, and diversity of available routes on the rail network in London, trip times are surprisingly consistent.

A coarse view of the London rail network's usage thus reveals the following traveller model: a person who uses the network mainly to commute to and from work, with a journey duration that is consistent over time and that is roughly half an hour long. However, given the sheer size of the rail network and the diverse population of London, such a simple model would likely (or obviously) overlook other types of usage behaviours that don't fit the average case: indeed, it would seem counter-intuitive for a variety of travellers (e.g., students, night-workers, tourists, pensioners) to fit these patterns. Unfortunately, our data lacks any features that would allow us to categorise travellers into these well-known groups, thus seemingly limiting the applicability of using fare collection data to inform traveller information systems. In the following section, we demonstrate that the Oyster data *can* be used to gather implicit differences between users. We apply clustering techniques to merge them automatically into groups with similar behaviours: we then show how, with no explicit knowledge of who each traveller is (i.e., based solely on usage data), travel information systems can be augmented to provide tailored results to each traveller.

3.2. *Individuals among commuters*

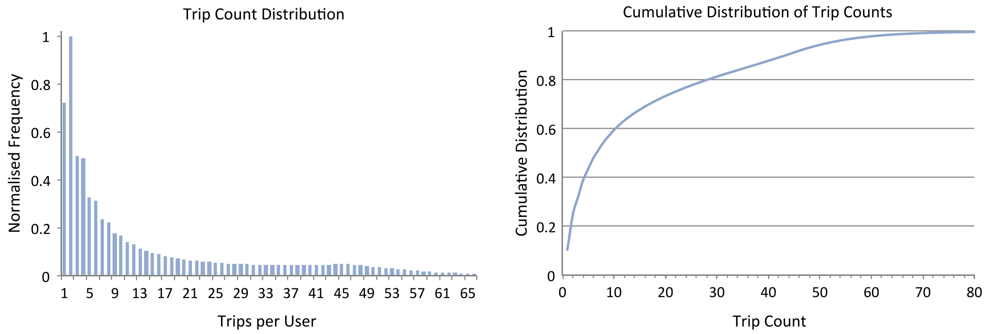
For the commuter stereotype to hold for all travellers, we would expect the following: (a) as the dominant travel pattern is a commute to and from work, the majority of travellers would have an average of 2 journeys (or more) per day;

Please cite this article in press as: N. Lathia, et al., Individuals among commuters: Building personalised transport information services from fare collection systems, Pervasive and Mobile Computing (2012), doi:10.1016/j.pmcj.2012.10.007



(a) Travel time distribution. (b) Standard deviation CDF.

Fig. 2. (a) The distribution of trip times showing that most trips are relatively short with a global average trip time of 28.41 ($SD = 15.83$) min. (b) CDF of trip time standard deviation, showing that 86% of trip times are within 10 min of the respective trip mean.



(a) Trip count histogram. (b) Trip count CDF.

Fig. 3. (a) The distribution of trip counts and (b) the cumulative distribution of trip counts across each user over the 31 days in our dataset. There is a substantial difference between users in terms of frequency of usage of public transport.

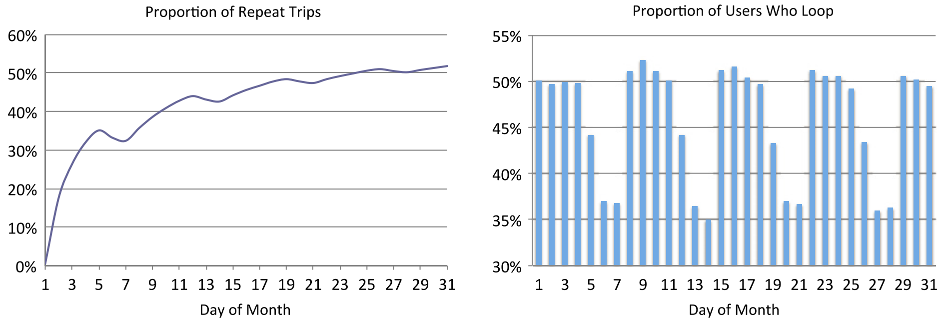
Table 1
Groups of users and average trips per group: a small proportion of users only travel on weekends, and nearly half of the users in the data set only appear during week days. Note, however, how those who travel on both week-days and week-ends achieve double the trips per user average.

	Users		Trips	
	Sum	Percentage	Sum	Avg/user
Week-day only	2,405,521	46.60	21,516,124	8.94
Week-end only	395,209	7.66	917,763	2.32
Both	2,361,787	45.75	56,407,420	23.88
Total	5,162,517	100	78,841,307	15.27

(b) furthermore, as the home and work place are not expected to vary much within a one month observation period, the same origin–destination trips should be repeated over and over again; (c) finally, for most travellers, the first origin and the last destination of the day should be the same, as this forms a closed loop with their commute. We now illustrate this is actually not the case, and that there exist many public transport passengers that do not fit this model. We then show how AFC data uncovers individual travellers' implicit profiles.

(a) *Average number of trips.* The mean total number of trips per user across the 31 days of data is 15.27, approximately one trip per user every two days: a significant proportion of travellers do not fit the commuting model. The differences go much further: Fig. 3(a) plots the distribution of user trip-count over the month, while Fig. 3(b) plots the same data as a cumulative distribution instead: as shown, the number of trips taken by each user ranges from 1 to well over 60, with the majority of users (66%) undertaking less than 1 trip every 2 days on average, and 2% (which still amounts to more than 100,000 users) taking more than 60 trips. Most notably, no more than 8% of users fit the expectation of 2 trips per weekday (that is, 46 or more trips overall).

We further examined these differences by grouping users into those who travel only during week days, only weekend days, and both: the sizes of each group (in terms of users, trips, and trips per user) are shown in Table 1. Nearly half of the travellers in our dataset use transit *only* during weekdays (46.6%) while 7.7% only use transit on weekends (the remaining



(a) Proportion of repeated journeys. (b) Proportion of users forming a loop.

Fig. 4. (a) The proportion of user-trip observations that have been seen before. (b) The proportion of users who form a loop with their daily travels. Both these results demonstrate the effect of the presence and absence of the commuting majority on weekdays and weekends respectively.

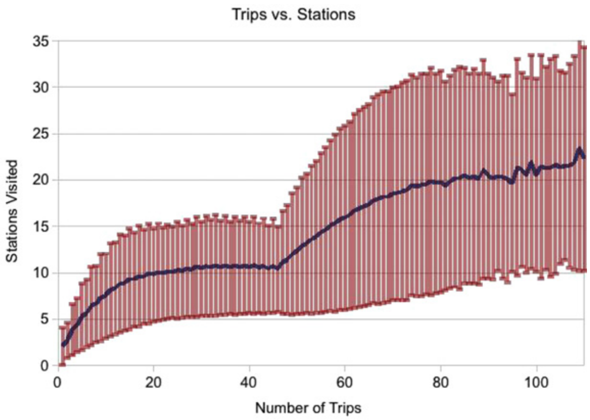


Fig. 5. Visualising the variance in the user's spatial exploration of the city, by plotting the number of trips taken vs. stations visited. The blue line is the mean, surrounded by the standard deviation in red. Note how the distribution changes near 50 trips in the month, which demarcates the difference between low-volume users and those who are using the system, on average, more than once a day. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

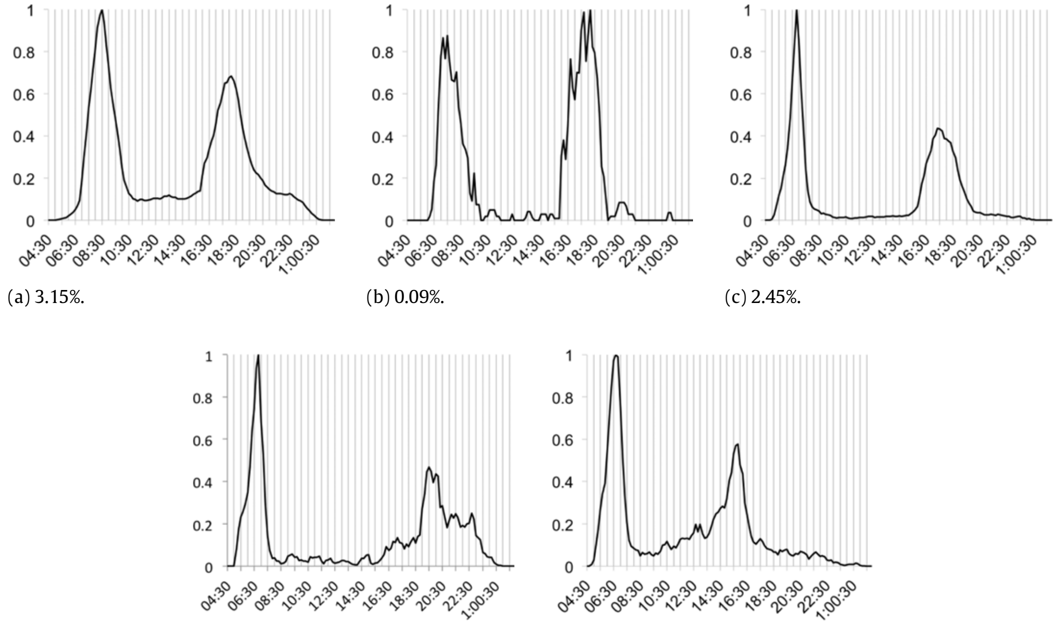
45.8% of travellers use transit on both weekdays and weekends). Unsurprisingly, travellers who travel on both weekdays and weekends average more trips overall in our dataset than the other two groups of travellers (AVG = 23.88, in the far right column in Table 1).

(b) *Repeated trips.* To examine repeat trip behaviour, we plot the cumulative number of user-trip pairs which have been observed previously as a proportion of the total number of trips observed thus far in Fig. 4(a). As time progresses, the proportion of repeated trips increases in a parabolic fashion, falling back slightly each weekend, and reaching 51.8% by day 31. In other words, about half of the trips observed throughout an entire month have been taken at least once before, with the other half being unique origin–destination tuples for that traveller during our period of observation (note: a different person may have taken that trip). Interestingly, in a study of over 250 car drivers, [11] found that the percentage of repeat trips reached 52.3% after 25 days of observation.

(c) *Closing loops.* Fig. 4(b) shows the proportion of users who form a loop each day. We count a user as having formed a loop if the origin station of their first recorded trip in a given day is the same as the destination of their last recorded trip that day. The proportion of users forming a loop appears to follow a regular pattern, with approximately 50% looping on Monday through Thursday, 44% on Friday, and 37% on weekends. Note, however, that one limitation in this analysis is that we do not account for the spatial proximity of adjacent stations. That is, a traveller may leave from one station in the morning and return to another adjacent station in the evening. In our analysis, this would not be counted as a loop. However, even with this limitation, approximately half of the travellers do close a loop in their daily trips.

(d) *Trips vs. stations explored.* Finally, we analysed the extent that usage of the public transport system (as measured by number of trips throughout the month) relates to spatial variance – in terms of the number of stations that the user's visit in the network (shown in Fig. 5). We find that, for low-usage users, the total number of stations visited remains relatively small – varying between 5 and 15 stations in the entire month. An interesting pivot point in the distribution appears after approximately 50 trips in the month. This inflection point indicates that those travellers who use the transit system more often also achieve a higher coverage of stations in the train network. More research is needed to uncover why the inflection point occurs near 50 trips.

Please cite this article in press as: N. Lathia, et al., Individuals among commuters: Building personalised transport information services from fare collection systems, Pervasive and Mobile Computing (2012), doi:10.1016/j.pmcj.2012.10.007



(a) 3.15%. (b) 0.09%. (c) 2.45%.

(d) 0.56%. (e) 1.17%.

Fig. 6. Clusters that portray varying kinds of 2-spike behaviour, including travellers who depart before the morning rush-hour and return before, during, and after the evening commute. The percentage under each cluster denotes the proportion of the total data that fell into the given cluster.

3.3. Clustering travellers to uncover emergent behaviours

In this subsection, we cluster travellers according to their average daily temporal usage of the train network. The goal here is to uncover patterns of behaviour that were not identified in the aggregate analysis above. To be able to compare travellers to one another, we represent each of them as a numerical vector which describes their travel habits. We then use an unsupervised clustering algorithm to discover and visualise emergent groups of similar people. Due to the scale of the data, we clustered random sub-samples of 2000 users, and repeated this process to generate a 10-fold cross validation of cluster results. In the following, we describe the steps taken in each run in detail.

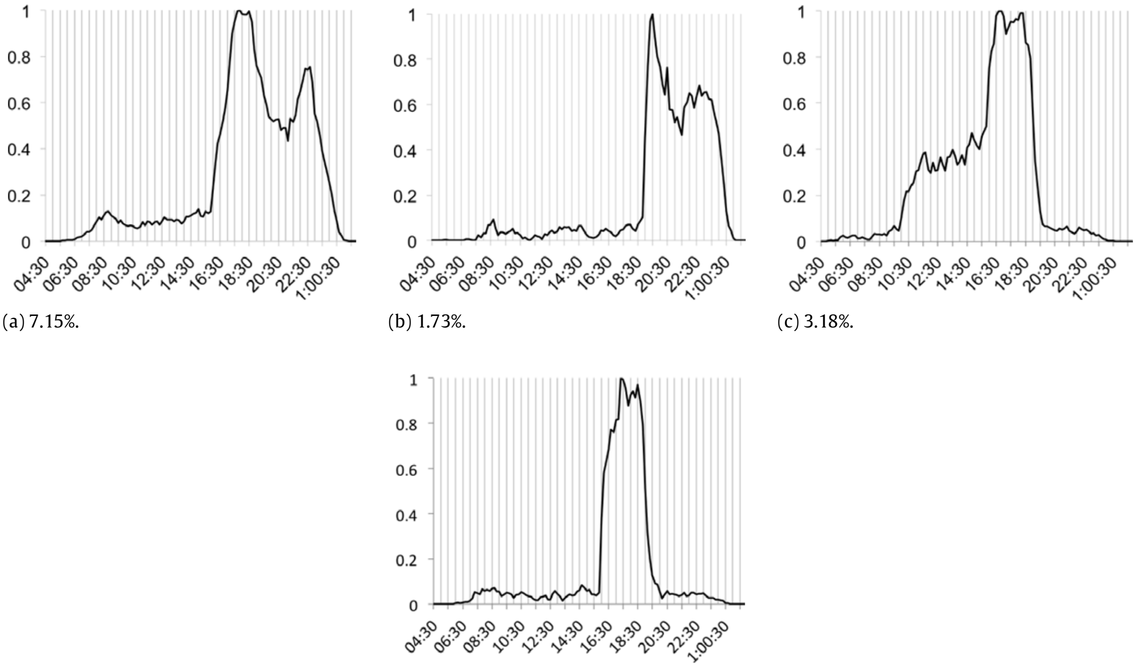
We first define travellers according to the *week day times* that they begin their trips; we split the 21.5 h day that the system is open for (from 4:30 till 2:00) into 5 segments, dividing the day around commuter hours (early morning 4:30–6:59, morning rush hour 7:00–9:59, day time 10:00–15:59, evening rush hour 16:00–18:59, and late evening 19:00–2:00). We then construct, for each traveller, a frequency vector of binned start times. That is, for each start time in the user's set of trips, the corresponding bin in the user's vector is incremented. We decided to use 5 segments because of the sparsity of the data in terms of trips per user (see Fig. 3(a)): finer grained partitions of the start times (e.g., 10-min bins) would result in many segments being empty, and thus presenting problems when measuring the similarity between vectors. We further pruned all users who had fewer than 2 trips in the entire data from this analysis to combat this issue.

Once we have a representation of each users' temporal travel habits as a numerical vector of values, we can measure the similarity between any two travellers. Given a pair of vectors, a and b, where user a has taken a total of A trips and user b has taken B trips, the similarity is computed as the absolute normalised difference between the vectors.

$$d_{a,b} = \frac{1}{5}\sum_{i=1}^{5}\left|\frac{a_i}{A} - \frac{b_i}{B}\right|. \qquad (1)$$

Under this metric, smaller values represent higher similarity. Finally, we implemented a form of agglomerative hierarchical clustering [12] to group travellers. This algorithm begins by assigning each user to an individual cluster, and then iteratively merges the two most similar clusters until a threshold similarity value is reached. In other words, we first construct a similarity matrix between all travellers, which will be populated with $d_{a,b}$ values for each possible pair of travellers. We then select the minimum value from the matrix, and merge the two vectors that share this similarity value to produce a new cluster. When merged, the two original clusters become the members of the new cluster, and the *centroid* is calculated as the mean of the member vectors. This process was continued until a maximum similarity value of 0.15 was reached; we found this value through a series of experiments. A similar clustering approach was used in [3] to automatically cluster shared temporal usage behaviours across communal bicycling stations in Barcelona, Spain.

This approach resulted in 15 individual clusters. By visually comparing the temporal behaviour in each cluster, four categories of behaviour emerged. The first set, shown in Fig. 6, represents a variety of different 2-spiked trends, including:



(a) 7.15%. (b) 1.73%. (c) 3.18%. (d) 2.58%.

Fig. 7. Clusters that portray varying kinds of evening-only behaviour.

(a) the typical commuting pattern, which appears in the aggregate data, (b, c) users who first travel before the morning rush-hour, but return with the crowd, although those in (c) proportionally travel less during the evening peak hour, (d) users who travel before the morning rush hour, peaking near 6:30 AM, and then again late in the evening, peaking at 8:00 PM, and lastly, those who again travel early and then again before the evening rush-hour, at about 4:00 PM. In other words, even the daily commuting habit is subject to a range of variants.

The other groups do not portray morning and evening travel habits. Instead, the travel patterns that emerge are condensed into particular times of the day. Fig. 7 depicts clusters of those travellers who tend to use transit only towards the end of the day, portraying evening-only behaviour. These range from: (a) a two-spiked pattern that only emerges in the evening, peaking near 6 PM and again near 11 PM; (b) a near two-spiked pattern that emerges even later during the day (7 PM–12 AM); (c) a set of travellers who begin moving after the morning peak hours and do so mostly throughout the afternoon rush-hour, and a similar set in (d) who exclusively travel throughout the evening peak-hour times.

The existence of travellers who exclusively use the system throughout the later hours of the day is counter-balanced by the third group (Fig. 8), with emergent patterns that centre on the morning hours only. Two sub-groups emerge: (a, b) morning travel with reduced usage throughout the day, and (c, d) morning only travel. Members of the latter sub-group also differ in terms of when they peak: (c) has members that travel during the "normal" commuting times, while (d) has members who move much earlier.

Finally, the last group, in Fig. 8, (e) and (f), exhibit the remaining behaviour: their travel is centred on the middle hours of the day. The difference between (e) and (f) is in terms of the variability in time-span: indeed, those in (f) produce – using their week day travels exclusively – a travel pattern that resembles the aggregate week-end travel activity.

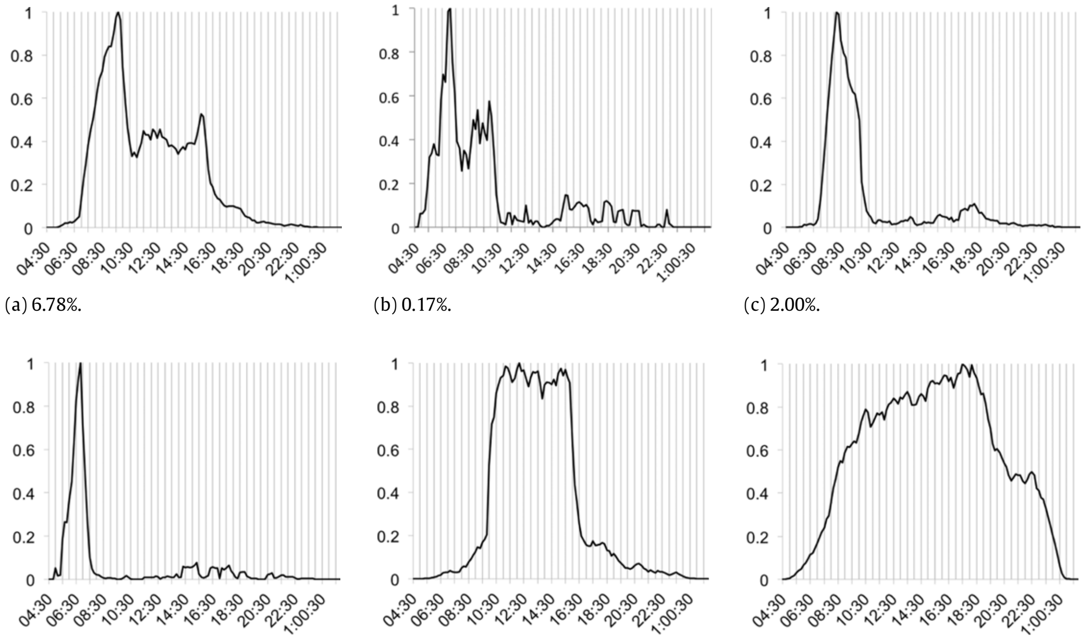
3.4. Travel time: individuals vs. the crowd

We now turn from exploring clusters of temporal usage patterns (i.e., when a traveller uses the transit system) to analyzing individual differences with regards to travel time (i.e., how long it takes a traveller to get from A to B). To do so, we calculate two values: the average trip time for a given origin–destination pair $\overline{m}_{o,d}$, and the mean travel time for a particular individual u's trips between the same stations $\overline{u}_{o,d}$. These two values can be used to compare each individual to the "crowd" of travellers who have travelled from o to d, by computing the normalised residual trip time $r_{u,o,d}$ for user u as follows:

$$r_{u,o,d} = \frac{\overline{u}_{o,d} - \overline{m}_{o,d}}{\overline{m}_{o,d}}. \tag{2}$$

The normalised residual represents the average excess journey time that an individual traveller experiences compared to the trip mean of all travellers. If $r_{u,o,d}$ is positive, this suggests that user u tends to be slower than average, whereas a negative

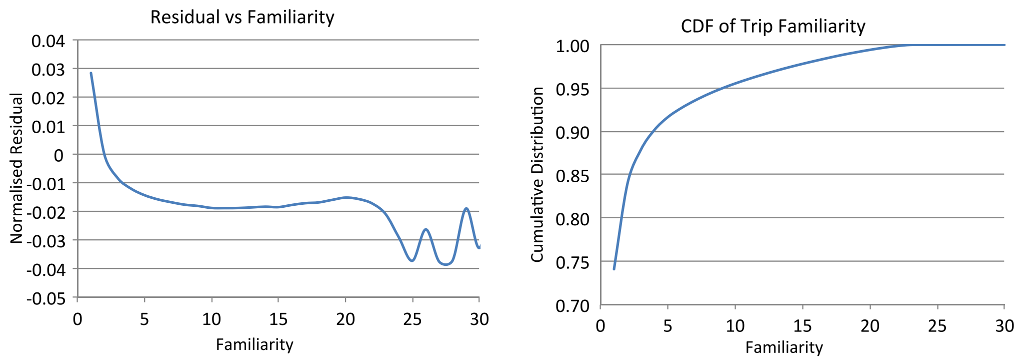
Please cite this article in press as: N. Lathia, et al., Individuals among commuters: Building personalised transport information services from fare collection systems, Pervasive and Mobile Computing (2012), doi:10.1016/j.pmcj.2012.10.007



(a) 6.78%. (b) 0.17%. (c) 2.00%.

(d) 0.20%. (e) 10.65%. (f) 31.14%.

Fig. 8. Clusters that portray varying kinds of morning and day-time only behaviour.



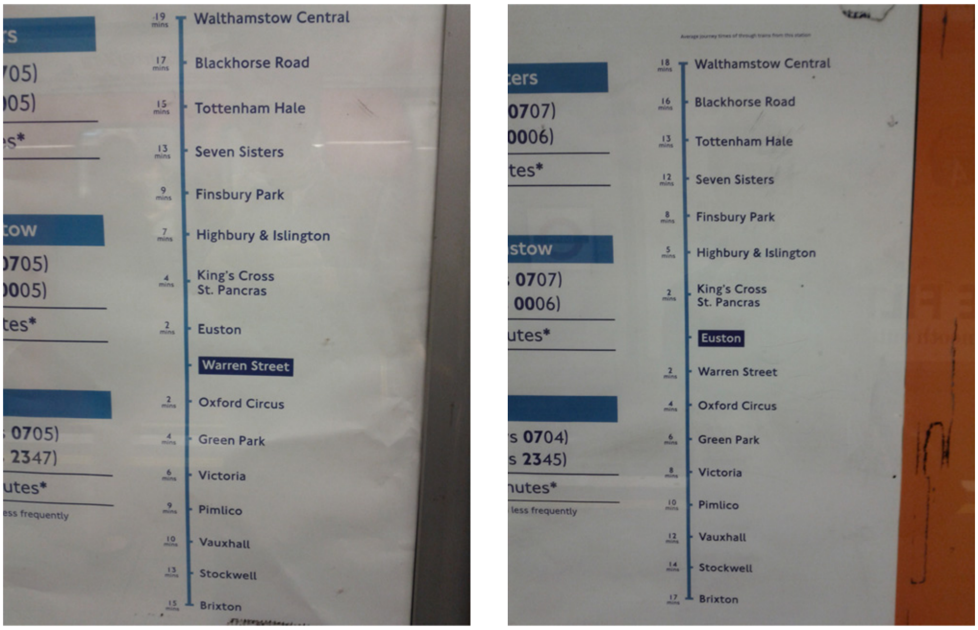
(a) Residual vs observed familiarity. (b) Observed familiarity cumulative distribution.

Fig. 9. (a) Residual trip time becomes negative as users become more familiar, (b) note that 74% of user-trip pairs have familiarity of 1.

residual indicates that the user is, on average, faster than the crowd. We then count the number of times that u has taken the trip between o and d as $f_{u,o,d}$ (which we denote as u's observed *familiarity* with the trip), and compare the resulting residual-familiarity pairs. The underlying intuition here is that, as users take a trip more frequently, they may no longer need to pause to look at signs or read maps and will generally be more adept at moving about the system: does the data confirm this intuition?

Fig. 9(a) shows the resulting relationship between trip familiarity and average residuals. User average residuals do, in fact, tend to become negative as the measured familiarity increases. This result has one limitation: the one-month view that we have of the data does not necessarily mean that, when we measure a traveller's familiarity with an (o, d) pair to be 1, the traveller has only ever taken the trip once. The user may very well have taken the trip previously, outside of the breadth of data available to us. That said, we remain sure that the user only takes this trip once during the month-long view that we possess; therefore, while the user may not necessarily be taking this trip for the first time, it is likely that this is not a high-frequency trip for that user. We examined the distribution of familiarity values closer in Fig. 9(b), which plots the cumulative distribution of familiarity among user trip means. Notice that for more than 74% of user-trip means, the familiarity value is 1, which suggests that the group of users who make a given trip once within the month contribute the most towards establishing the mean time.

Please cite this article in press as: N. Lathia, et al., Individuals among commuters: Building personalised transport information services from fare collection systems, Pervasive and Mobile Computing (2012), doi:10.1016/j.pmcj.2012.10.007



(a) Warren Street station. (b) Euston station.

Fig. 10. An example poster with travel times from (a) Warren Street's and (b) Euston's Victoria line platform.

3.5. *Travel time: individuals vs. static timetables*

The last step of our analysis focuses on a comparison between individuals' actual travel times, as recorded by our AFC dataset, and transit-time information provided by the London Underground. More precisely there are posters affixed in most platforms stating how long it should take to travel to any destination that is reachable from the given platform along the same tube line (i.e., requiring no interchanges). In particular, we selected two stations on the Victoria Line (Euston and Warren Street, with posters shown in Fig. 10). The advertised travel times range from 2 min to reach an adjacent station, to 19 min to reach the end of the line. We note that, although the stations are adjacent to one another, the listed travel times are not symmetrical: Euston to Walthamstow Central is advertised as an 18 min trip. Warren Street, which is 2 min (and 1 stop) from Euston, shows a 19 min journey to Walthamstow Central (rather than an $18 + 2 = 20$ min trip). We have no further details about these travel times: we do not know how they are computed or what they assume; a matter that we leave for future investigation. However, by being a source of information for travellers, the estimates they display should offer a relatively accurate picture of the amount of time necessary to travel between each pair of stations. How close are advertised transit estimates to the travel time measured by AFC datasets?

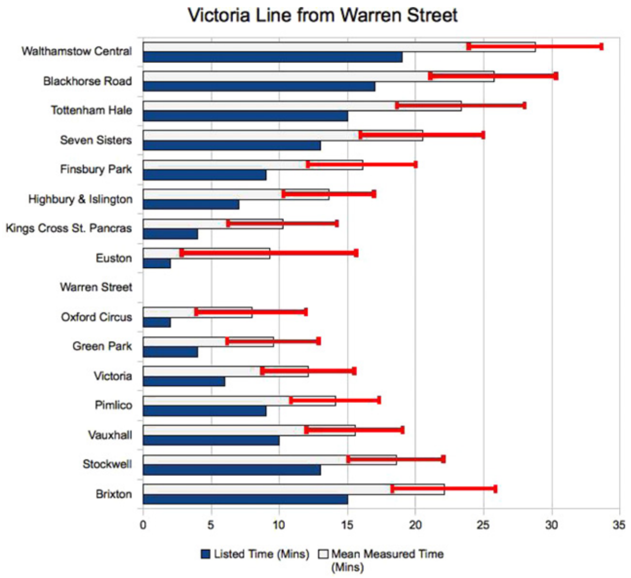
We measured how actual travel originating at these stations and completing at any other station on the same line, compares to the poster's expected trip duration. The results are shown in Fig. 11: blue bars represent the advertised number of minutes from that station captured on the posters, while grey bars represent the mean travel time for all observed trips between the given (o, d) pair, with red lines illustrating the variance. For example, the topmost bar in Fig. 11(a) is the mean travel time between Warren Street and Walthamstow Central.

The results show that actual mean trip times are much higher than listed times; indeed, the advertised times fall well under one standard deviation from each actual mean. The standard deviations are particularly large when travelling to adjacent stations. This may be explained by the fact that the transport authority is only advertising the *train travel time* between stations. Recent work [13] states that, on average, only 46%–62% of the time that users spend in the tube is actually spent riding the trains, while the rest is spent interchanging, walking, or waiting. These times, which are unaccounted for in static timetables, are still within the scope of the *travellers' perceived* travel time; furthermore, while 'time on train' is likely to be the same for all travellers, it is precisely these times spent elsewhere within the system that we expect to vary across individuals, and to account for the differences between individuals' travel times and the crowd already highlighted in Fig. 9(a).

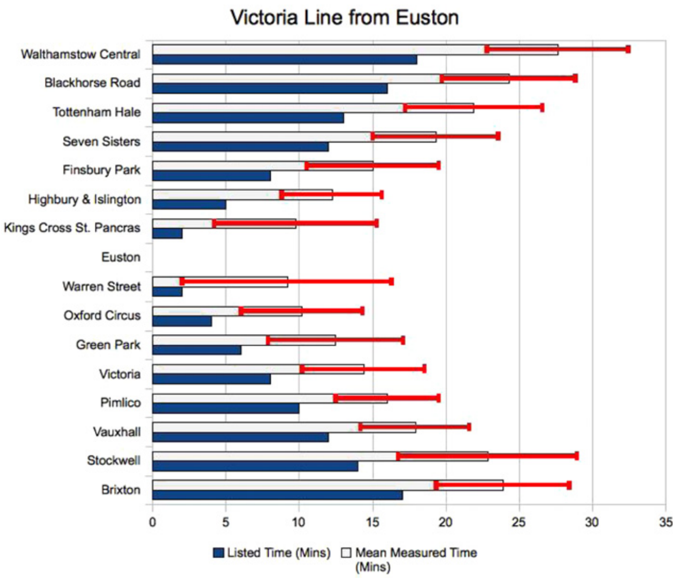
3.6. *AFC data analysis conclusions*

Based on the analysis conducted above, we draw the following three main conclusions:

- *More than one traveller type exists.* An aggregate analysis of the whole AFC data produces a single view of the underlying travellers, which is dominated by the commuting habits of a large subset of users. By using hierarchical clustering, we



(a) From Warren Street.



(b) From Euston station.

Fig. 11. Travel time from Euston and Warren Street: the blue bars represent the advertised travel time from posters at each station, while the grey bars are the mean travel time as measured with the Oyster data. Note that, even when plotting the variance as well (in red), most values are underestimated by the poster's times. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

showed that a number of different travel patterns emerge, thus discounting the value of modelling travellers using a single perspective.

- *The crowd's travel time significantly differs from that of the individual travellers.* The aggregate system view fails not only to describe the travel patterns of the majority of travellers, but also to capture the time it takes for individual travellers to complete their journeys. For example, we showed that as a user becomes more familiar with a trip, the time it takes to complete the trip decreases. We thus cannot use information about the crowd as a whole to accurately describe individual travel times.
- *Published travel times fail to capture substantial differences in travellers' transit times.* We cannot compute individual's travel times using aggregate AFC data, and nor can we do so by using published travel time estimates. Indeed, as the last part of our analysis demonstrated, the advertised inter-link times consistently (and grossly) underestimate the time that it takes people to travel between an origin and destination.

The analysis above thus demonstrates that there exist substantial differences among individual travellers, and that such differences are indeed *measurable* via the implicit behaviours captured by the Oyster AFC system. So what can this data be used for? We posit that AFC data can be used to build more *personalised* travel information services, that is, services that are geared towards the *individual* needs and characteristics of each single traveller. As a quantifiable example of such systems, in the next section we delve into the construction of a personalised trip time estimator. In particular, we propose a variety of prediction algorithms that leverage AFC data to implicitly capture individual travellers' characteristics, and thus accurately estimate personalised travel times.

4. Trip time prediction algorithms

In this section, we explore a number of methods that take advantage of a user's AFC-based travel history in order to provide personalised trip time estimations. Recall that each data point gives us the origin, destination, and start and end times of each trip. As such, the following prediction models do not include information regarding the route taken between stations or service disruptions, both of which are likely to affect travel times. Instead, the models described here explore the patterns and relationships highlighted in Sections 3.3 and 3.4, but with a specific focus on the *differences* between users.

4.1. Baselines

We first describe two baseline predictors against which the quality of our models will be assessed. The first baseline is based on the static travel time between adjacent stations. This data was obtained using a Freedom of Information (FOI) request[2] to Transport for London; it details the distance and travel time between adjacent stations in the London Underground network. This data does not include scheduled train departure or arrival information, only link traversal times. For example, for a train line with stations $s_1 \leftrightarrow s_2 \leftrightarrow s_3$, the FOI data includes tuples with the travel time (t) between each interconnected station: $\langle s_1, s_2, t_1\rangle$, $\langle s_2, s_3, t_2\rangle$. Note that the FOI request returned travel time data from September, 2008 whereas our AFC dataset is from March, 2010 (18 months later). We assume that travel times have not changed significantly in those 18 months. The time between non-adjacent stations was computed using Dijkstra's algorithm on a graph where each station is represented as a node, and the connections between each station are edges weighted by the link traversal time: this method therefore produces the shortest path between any pair of stations based on hop-count (instead of, for example, minimising number of transfers between lines); in the following sections, we refer to this as the *static* prediction method. Where an interchange occurs, we had a choice about how many minutes should be added to the total trip time: we experimentally determined this "transfer time penalty" to be 8 min (see Section 5.2.1).

Although the Oyster data contains 584 stations, the FOI request provided us with link traversal times that included only 355 (those covered by the London Underground), including all but one tube station. We therefore pruned trips between stations not present in the FOI request when testing all of the prediction algorithms.

The second baseline we use is derived from the Oyster data itself: the historical trip time average. This approach will not take into account any differences between travellers, and simply returns the mean travel time for all trips between any given two stations. More formally, for the set $T_{o,d}$, which contains N observations of a journey between stations o and d, the baseline estimated travel time is simply the arithmetic mean of all $x_t \in T_{o,d}$, where x_t denotes the travel time:

$$\overline{m}_{o,d} = \frac{1}{N}\sum_{T_{o,d}} x_t. \tag{3}$$

4.2. Personalised approaches

Given the two baselines above, we now set out to design and evaluate algorithms that account for the differences between users when making their predictions: such techniques could be used to provide personalised trip time estimates to travellers using the system. In particular, the techniques that are presented in the following sections build off the latter baseline by using different dynamic techniques to subsample the training data that is used to predict each user-trip pair. In particular, three types of information are leveraged: (a) the similarity between the current user and those who have historically travelled at the same time, (b) the similarity between the current user and those who have historically made the trip with similar frequencies, and (c) the extent that each traveller will repeat past behaviours.

There are two main challenges that we faced when designing personalised trip estimate algorithms using the Oyster data. First, the data is incredibly sparse and subject to each users' habits; we thus do not hold a complete view of how travel time over all (o, d) pairs varies with, for example, time of day. We address this below by weighting personalised predictions with the mean travel time, based on how much data is available. Second, the data is subject to noisy entries, such as when users encounter delays or other events whose occurrence significantly changes their normal travel time. We determined empirically that computing a *geometric* mean provided a more robust estimate when taking small samples of

2 www.whatdotheyknow.com/request/distance_between_adjacent_underg.

data for personalised predictions. Given a set $S_{o,d}$ of M observations of a given travel time between o and d, the geometric mean is computed as the log-sum of each member t:

$$\overline{g}_{o,d} = \exp\left(\frac{1}{M}\sum_{S_{o,d}} \ln t\right). \tag{4}$$

In the following sections, all references to a *mean* of data sampled using a given method refer to the usage of the geometric mean.

4.2.1. Trip context model

The first model aims to derive a trip time estimate by incorporating historical contextual information of each trip that users are taking. Ideally, such an approach would take into account each trips' details, including purpose of travel (e.g., business or leisure), service status, whether the user is travelling with luggage or children, etc.. However, we are limited by the information contained in the dataset and are therefore constrained to that which implicitly represents travellers' context: the historical average travel time for a journey at a particular time of day. We assume that, for a given start time s and a time interval w, all users who start a journey from station o to station d within a window of size $2w$ centred on s, experience a similar context. Thus we can define $\overline{w}_{o,d}$ as the geometric mean of the set of trip times $W_{o,d} \subset T_{o,d}$ in which each observation $x_{o,d,t}$ meets the condition:

$$(t - w) \leq s \leq (t + w). \tag{5}$$

The observation window defined above operates at the day level: factors which vary above this level, such as holiday periods and busy local events are not taken into consideration. Due to the sparsity of the data, the possibility exists that, for certain (o, d) pairs, the $\overline{w}_{o,d}$ value will be undefined. When computing the personalised prediction $\hat{p}_{u,o,d,s}$, we compensate this by returning a weighted combination of the window mean and the trip mean. We weight each component by the value $\frac{M}{N}$, where we have a total of N observations of trip o, d, and M observations within the predefined window; this weighting allows us to compensate for data sparsity by using the aggregate data where necessary.

$$\hat{p}_{u,o,d,s} = \left(\left(1 - \frac{M}{N}\right) \times \overline{m}_{o,d}\right) + \left(\frac{M}{N} \times \overline{w}_{o,d}\right). \tag{6}$$

4.2.2. Estimating with trip familiarity

While the model above only considered the implicit similarity shared by travellers based on *when* they travel, the next model incorporates notions of how *familiar* those users are with a given trip. In Section 3.4 we introduced the notion of user's *familiarity* with a trip, which is simply the number of times M we have observed the user-origin–destination tuple. We also found an aggregate inverse relationship between the familiarity and the deviation of the user mean from the trip mean: those who had taken the trip with high frequency tended to travel faster. We capture this relationship when computing predictions with the following definition: for a user u who has taken a trip between o and d a total of M times in the past, we define a neighbourhood of user means $F_{M,o,d} \subset T_{o,d}$ from users whose familiarity $f_{o,d}$ with the trip is such that:

$$\frac{M}{2} \leq f_{o,d} \leq 2M. \tag{7}$$

In other words, we compute a personalised prediction for a user by identifying, using a sliding window, those users who have historically taken the trip a similar number of times. To compensate for the sparsity of observations, we return a $\frac{1}{M}$-weighted combination of the trip mean and the familiarity mean:

$$\hat{p}_{u,o,d,f} = \left(\frac{1}{M} \times \overline{m}_{o,d}\right) + \left(\left(1 - \frac{1}{M}\right) \times \overline{f}_{o,d}\right). \tag{8}$$

4.2.3. User self-similarity predictions

The final model relies on the assumption that users may tend to exhibit consistent habits when navigating the system, including their choice of route and walking pace; in effect, it assumes that travellers' time spent in the system is likely to be more similar to their own historical travel times than, unlike above, those of others. More formally, let $U_{o,d} \subset T_{o,d}$ denote the set of user u's trip times between stations o and d. We compute the geometric mean $\overline{u}_{o,d}$, for each trip $u_{o,d} \in U_{o,d}$.

As above, due to the sparsity of the data, the user mean will be undefined for many (o, d) pairs (where the user has not previously made that trip within the scope of the training data), or it may be the result of very few observations. To account for this, the personalised prediction $\hat{p}_{u,o,d}$ returned by the user self-similarity model, is a $\frac{1}{M}$-weighted combination of the trip mean and user mean:

$$\hat{p}_{u,o,d} = \left(\frac{1}{M} \times \overline{m}_{o,d}\right) + \left(\left(1 - \frac{1}{M}\right) \times \overline{u}_{o,d}\right). \tag{9}$$

Table 2
Mean Absolute Error prediction results. The results are presented for all predictions (aggregate) as well as decomposed into groups of users, who travel on week days only, weekend days only, and both week day and weekend days. Estimating users' travel time based on how they travelled previously is, on aggregate, the most accurate method. However, travel time (context) becomes more accurate for those users who only travel on weekends.

Method	Mean Absolute Error (minutes)			
	Aggregate	Week day	Week end	Both
Static	9.82	9.91	10.71	9.72
Trip mean	3.30	3.32	3.69	3.28
Context	3.28	3.29	**3.68**	3.26
Familiarity	3.17	3.18	3.69	3.16
Self-similarity	**3.13**	**3.11**	3.81	**3.14**

5. Evaluation

In the above, we have defined three perspectives of historical travel time data (context, familiarity, self-similarity) that can be used to forecast how long it will take a traveller to transit between an origin and destination. In this section, we use the Oyster data in order to evaluate the accuracy of these forecasts, compared to the trip times that are estimated by both baselines. Recall that our dataset spans an entire month and includes 76.6 million trips. We test our prediction models by repeatedly splitting the Oyster data into two partitions: a 1-day test set taken from the final week of the data, with all days prior to it as a training. We repeated this procedure for each of the 7 final days in the dataset in order to provide 7-fold cross validated results, each of which has an approximate split of 95% training data to 5% test data. We next describe the accuracy metrics we used and the results we obtained (Section 5.1), and a detailed analysis of the user, spatial, and temporal aspects of the accuracy achieved (Section 5.2).

5.1. Metrics and aggregate results

We use two metrics to evaluate the prediction accuracy: the Mean Absolute Error, MAE, and Mean Absolute Percentage Error, or MAPE [14]. The MAE is measured in the same units as the trip times (minutes) and is simply the average absolute difference between the actual trip time and the corresponding predicted trip time. Let N be the size of the test set, then for each of N predictions, $\hat{p}_{u,o,d}$ of user u's time to travel from o to d, and the actual trip time $x_{u,o,d}$:

$$\text{MAE} = \frac{1}{N}\sum |x_{u,o,d} - \hat{p}_{u,o,d}|. \tag{10}$$

Table 2 compares the results from each model to the baseline predictions. As expected, the static link traversal-based data provides the least accurate forecast of travel time; on average, its estimates have an average error of 9–11 min. The trip mean that is computed using the Oyster data, instead, already provides a much stronger estimate. Yet it is outperformed by each of the personalised models, with the best performance provided by the *self-similarity* model, which tends to achieve an error just above 3 min—roughly comparable to the average waiting time between trains. Self-similarity, however, is not as accurate as using the time of travel (or context) for those users who travel exclusively on weekends. A potential explanation for the loss in performance is the fact that this is the group of users who have, on average, the smallest number of trips per person; prediction accuracy using the self-similarity method will thus begin to revert back to the trip mean. Fig. 12 displays these results visually for direct comparison.

An alternative view of the error can be obtained by normalising the absolute difference between the predicted and actual travel time using travellers' actual transit time, allowing us to compare errors for trips of varying length. The Mean Absolute Percentage Error (MAPE) does just this: it is the error measured as a percentage of the actual trip time and produces values in the range 0%–100%.

$$\text{MAPE} = \frac{100}{N}\sum \left(\frac{|x_{u,o,d} - \hat{p}_{u,o,d}|}{x_{u,o,d}}\right). \tag{11}$$

Table 3 shows the MAPE results, grouped as before: they highlight that up to 40% of travel time may be unaccounted for when using static estimates. This result echoes previous research by Jang [13] described above, which claimed that 46%–62% of travel time remains unaccounted for when exclusively using train time. Moreover, while the self-similarity method once again appears to be the most accurate on aggregate and the context method overtakes all others for weekend only users, the MAPE metric gives a higher priority to the familiarity technique for those travellers who take trips on both kinds of days.

5.2. Error analysis

In the previous section, we presented empirical results that demonstrate the improved accuracy in travel time estimation when using fare collection data. In this section, we delve further into the accuracy of the above methods, in order to fully

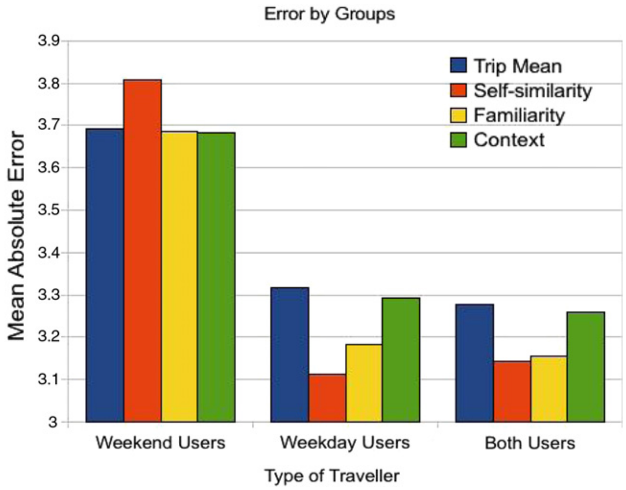


Fig. 12. MAE results, when users are grouped by when they travel (weekdays only, weekends only, both). The results show varying behaviours from the prediction methods: for example, users who travel during both weekdays and weekends are the most accurate group when using the trip mean, but the least accurate group when using self-similarity values.

Table 3
Mean Absolute Percentage Error prediction results. These results portray a similar trend to those reflected in the MAE results; however, the familiarity method overtakes the self-similarity method for those travellers who travel during both weekdays and weekends.

Method	Mean Absolute Percentage Error			
	Aggregate	Week day	Week end	Both
Static	37.67	37.26	40.62	38.03
Trip mean	13.42	13.20	14.39	13.62
Context	13.35	13.13	**14.37**	13.57
Familiarity	12.61	12.37	14.41	**12.83**
Self-similarity	**12.52**	**12.13**	15.19	12.88

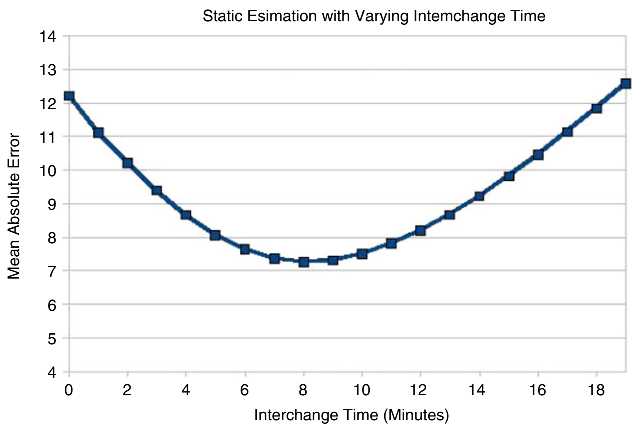
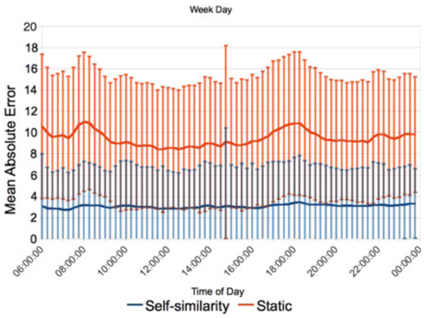
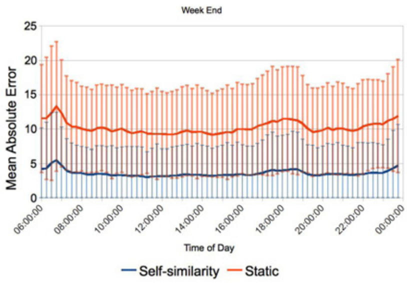


Fig. 13. The relationship between static transfer times and trip estimation mean absolute error. Lower values imply more accurate trip time estimates: in this work, we compare our proposals to the most accurate results found here (8 min).

understand the effect of static transfer times as well as how and where personalised approaches succeed. We investigate the extent that error is distributed over a number of attributes, including the time and day of the trip, the trips' length, the users' familiarity, and the spatial and transport network/structural properties of the journey. For the sake of clarity, we compare the static link traversal-based journey length estimates to the personalised self-similarity based predictions, which was found to be the most accurate method.



(a) Weekday.



(b) Weekend.

Fig. 14. The difference in accuracy over an average weekday and weekend: during the week day, the static data suffers from a rush-hour effect, while the data based approach remains flat.

5.2.1. *Static transfer time*

We first examine the impact of changing the inter-line transfer time penalty in the link traversal-based prediction method. We ran a series of experiments to determine which transfer time produces the most accurate results on the Oyster data we have. Fig. 13 plots the results: a transfer time of 8 min was the most accurate; lower values were likely to be under-estimating travel time, while higher values begin to unnecessarily apply too much time to transfers. Note, however, that this value remains constant across all users.

5.2.2. *Error by time of day*

We first investigate the variation in error values as the temporal attributes of the data vary. Fig. 14 shows the difference in error between the static and self-similarity predictions according to the time of day for weekday (Fig. 14(a)) and weekend (Fig. 14(b)) trips. The upper bound of the standard deviation of data-based predictions is (with the exception of one point) consistently below the mean of the prediction errors when using the static method. Moreover, static predictions seem to suffer from a peak-hour effect: the mean of weekday predictions increases during the rush hour times. On the other hand, the data-based predictions, which will capture how individuals tend to transit throughout the system, remains flat throughout the day and does not show this effect: in essence, personalised approaches overcome the challenge and need to dynamically modify trip estimation algorithms when the transport system is congested.

5.2.3. *Accounting total travel time*

Although personalised travel time estimation does not vary with the time of day, it may vary with the length of the trip that the user is taking: are the techniques able to better predict short or long trips? To answer this question, Fig. 15 illustrates the relationship between prediction error and the actual journey time. These results show how prediction errors tend to be *proportionally* higher for shorter trips. In fact, as trip lengths become longer, although the error increases, the proportional error becomes smaller; this points towards the fact that the less time travellers spend on trains, the more their individual differences will both emerge and determine their total travel time. Furthermore, when taking the standard deviations into account, the static and self-similarity distributions only begin to overlap after approximately 50-min long trips; recall that the analysis above indicated that majority of trips taken in the system have a travel time far below this threshold.

5.2.4. *Personalised approaches are not spatial*

In the trip time estimation methods presented above, we did not take any spatial features of the city into account: we now turn to this view of the data in order to visualise any potential spatial effects that emerge from the user-trip data. We hypothesise that the diversity of users entering stations in central London will be higher than, for example, stations in residential areas, and that most of the time estimate errors will thus emanate from the city centre. In other words, we examine the extent that trip time estimate errors vary according to the origin station of users' trips. To do so, we compute the average error of all trips beginning at each station, computing one MAE value for each origin in our dataset. We then plot the distribution of the results in Fig. 16(a) and examine the spatial distribution of error in Fig. 16(b). The former shows that there is indeed a non-uniform distribution of error (i.e., the origin station relates to the amount of error in the trip forecast). However, we did not find any direct relationship between error and the location of origin stations: stations of varying error are dispersed throughout the entire map of London, both within central London and outside of the confines of the city centre (16(b)). An analysis of the trip destinations' reveals similar results. A likely explanation for this result stems from the fact that our prediction algorithms do not use any spatial or structural information when computing trip estimates: the distribution of error may thus be more related to the origin–destination *pair*, as well as the amount of data available for the given pair and user who is taking the trip.

Please cite this article in press as: N. Lathia, et al., Individuals among commuters: Building personalised transport information services from fare collection systems, Pervasive and Mobile Computing (2012), doi:10.1016/j.pmcj.2012.10.007

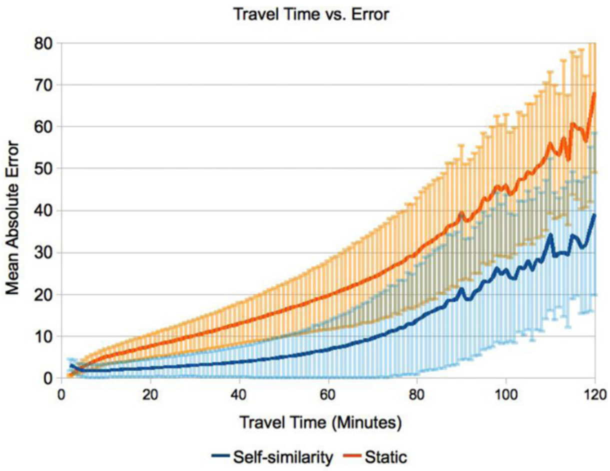
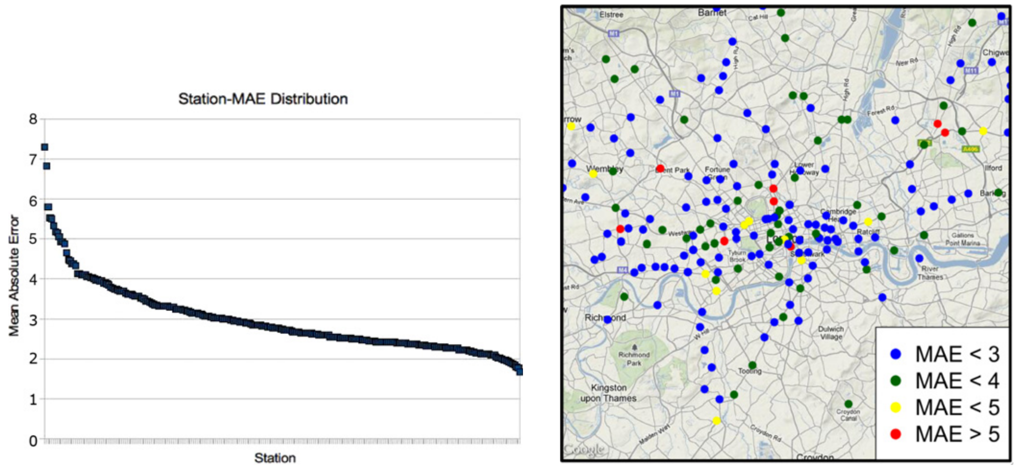


Fig. 15. Prediction accuracy compared to how long the trip actually took the user: transit time of longer trips are estimated less accurately.



(a) Station MAE distribution.

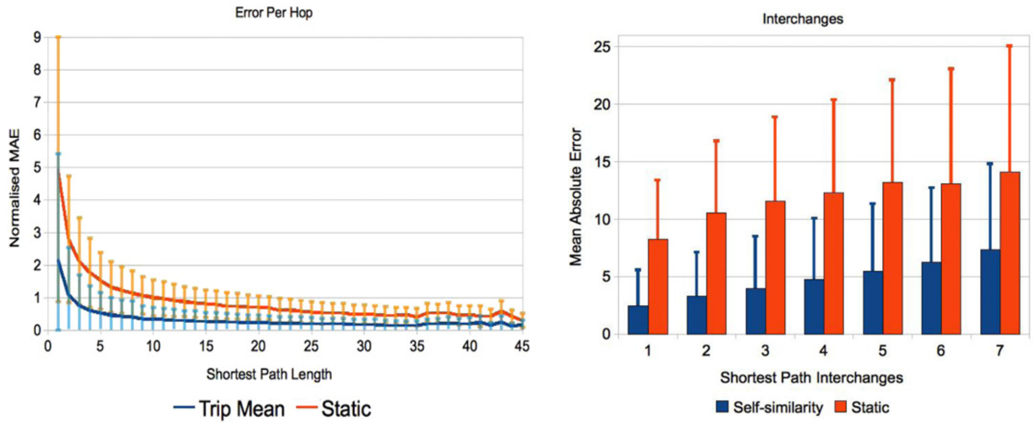
(b) Spatial errors.

Fig. 16. The spatial distribution of trip time estimation error, based on the station of origin. Note how there does not seem to be an obvious relation between the location of stations and estimation error: those origin stations that produce the highest error are both inside and outside of central London.

5.2.5. Comparing to structural properties of the network

Finally, given the graph-structure of the rail network in London, we can compute a variety of other features of users' trips. These include, for example, the minimum hop count between two stations. From these shortest paths, we can compute the number of hops that trips may take, as well as the potential number of interchanges that users will be faced with. Of course, these values do not take into account any heuristics that travellers may use when navigating the network, such as opting for a longer trip (in terms of hop count) in order to reduce the number of interchanges, or making interchanges to other tube lines which may lower the overall travel time. However, the key idea is that we can use these pre-computed values to examine the potential effect that users' choices as well as characteristics have on travel time across a variety of trips with similar characteristics. For example, if we know that the shortest path on the journey between a particular origin and destination has 2 interchanges, observing a high variance in the trip time may highlight (a) users who are fast (or slow) at interchanging, or (b) where users diverge from one another by making different route choices.

To compare trip length (by hop count) to error, we normalised the MAE values by dividing them by the number of hops, producing a error-per-hop metric. The results comparing trip hop count to normalised accuracy are shown in Fig. 17(a). Without normalising, the results appear to be much like those in Fig. 15 (comparing travel time to accuracy): longer trips generate worse accuracy results. Once we normalise, we produce values that represent the amount of error produced for each segment of (the shortest path for) the trip. We observe that, as trips become longer, the error per hop quickly diminishes; in fact, the highest error (as well as standard deviation) is for trips with a hop count below five, or short trips



(a) Trips' hop count.

(b) Number of transfers.

Fig. 17. Comparing the prediction error to computed characteristics of trips: (a) the trip's shortest path hop count and (b) the number of line changes required by the shortest path.

where (as we hypothesised) individual differences may become more prominent. The relationship between interchanges and forecasting error, instead, is found in Fig. 17(b). Indeed, we find that as the number of interchanges on the shortest path for the trip increases, so does the mean error for those trips. Reverting from static predictions, which assumes that all user transfers will take a constant amount of time, to the fare-data substantially reduces the error in trip time prediction.

5.2.6. *Error analysis conclusions*

The sections above explored the different forces that may be at play when making predictions about travellers' journey times using static link-traversal times and Oyster card data. A notable limitation of the analysis above is that each group (e.g., trips taken in a certain time bin, trips with a certain length, etc.) contains a varying number of trips, thus limiting our ability to directly compare groups to each other. However, we found that accuracy varies along a number of dimensions:

- *By user group*: Those users who only travel during week ends (and thus have, per user, much less data associated to them) are more difficult to predict. In fact, the most accurate predictor for this subset of travellers was the one that leverages trips taken during the same time of the day.
- *Total travel time*: The accuracy of our techniques was inversely proportional to the length of the actual trip. As above, this may be due to the fact that we have much less data for long trips (since the majority of trips taken in the system are short, and the global average trip time is below 30 min).
- *Structural properties of the trip*: We found that accuracy waned as the number of interchanges on the shortest path between an origin and destination increased. Furthermore, shorter trips (in terms of shortest path hop count) produced a higher variance per hop; both point to scenarios that are more difficult to predict when user differences may play a bigger role in determining travel time.

We also uncovered two dimensions that were not affected by personalised trip estimates. Both are properties of trips that inherently do not take into account individual differences; we found that accuracy did not seem to change with:

- *Time of day*: When binned into time segments, the personalised approaches produced consistent results throughout the day for both weekdays and weekends. The static data, on the other hand, shows a small loss in mean accuracy during peak-hours of the week days.
- *Spatial layout*: While the distribution of error across origin stations is not uniform (reflecting also the fact that the number of trips originating from each station is not the same), there was no apparent pattern between *where* the station is and *how much* error is associated with it.

In conclusion, we find that fare collection data is an immediately available, accurate source for travel time predictions that implicitly captures a variety of behaviours expressed by the city's residents. In the following sections, we place this empirical analysis into the context of broader personalisation and smart cities research; we then close by discussing future lines of research and the applicability of using fare collection data for a variety of other (including non-transport related) systems.

6. Related work

The work that relates to our task above – designing personalised information services for urban public transport systems – spans multiple fields. In this section, we present an overview of work that seeks to build *personalised* services, leverage mobility data from *smart cities*, and improve transport *information systems*.

6.1. Personalisation

Personalisation has been a key component of web-based systems; the most prominent example is its use for recommendation in e-commerce [5]. Such systems often rely on collaborative filtering algorithms, which can operate both on data from how users rate content (e.g., on a 1–5 Likert scale) as well as on implicit behavioural data [15]. These systems 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.

The lack of personalisation in transport information systems has been noted in the past [4]. At the same time, the notion of augmenting the performance of personalised systems by including a variety of contextual features is starting to be explored [6]. Beyond the work above, there seems to be a great opportunity for the overlap between the two; recent examples include incorporating preferences into route planning [16] as well as using mobility data to recommend money-saving public transport fares [17].

One of the key aspects of successful recommender systems is that they tailor information in a 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.

6.2. Smart cities and urban informatics

Broadly speaking, understanding the mobility patterns of large groups of humans has been a focal point of scientific research; researchers have investigated this theme using mobile phones [18] as well as the exchange of bank notes [19]; the AFC dataset we use here may very well contribute these insights as well. Within urban contexts, there is a growing body of research centred around the emerging field of urban informatics or "smart cities" [20], which is the study of human behaviours and urban infrastructures made possible by the increasingly digitised and networked city. Such research often incorporates advances in data mining, signal processing, sensing, and databases to process, analyse, and store massive quantities of data about human behaviour and urban infrastructure performance.

The bulk of research in this domain emanates from using embedded or mobile sensors in order to sense how people interact, navigate in and use urban spaces. For example, [21] used bluetooth and specialised mobile phone software to track human-to-human interactions. Further examples include measuring the spatio-temporal signature that emerges from Barcelona's shared bicycle scheme's stations [3]; visualising the flow of pedestrians, public transit, and vehicular traffic using mobile communication patterns [22,23]; characterising land usage from mobile phones [24], and using GPS-sensors embedded in taxis for urban planning [2]. Kostakos et al. [25] and Sadabadi et al. [26] rely on distributed Bluetooth receivers to track and predict travel speeds based on the Bluetooth MAC identifiers of passing devices; [27] use GPS for transport and traffic/congestion monitoring. Outside of the transport domain, [28] have used AFC data in order to evaluate models that would allow mobile phone users to share digital media while they may be riding the same train.

One of the drawbacks of raw sensor readings is that they lack qualitative descriptions of peoples' behaviours (e.g., reasons for behaving as they are, contextual features of their actions); this certainly applies to our data as well—we do not know *why* people are travelling or *what* individual traits they may have that affects their journey time. Researchers have thus been turning to social media data to fill this void: recent examples include measuring community well-being from tweets [29], using explicit "check-ins" to understand the relation between urban space and social events [30], and tracking where people take photos in urban environments to gain an insight into tourist dynamics [1].

Most of the cited work, however, continues to focus on aggregate analysis—which, as we have demonstrated above, will invariably not include the individual differences between the people who, together, form the aggregate. In doing so, they are more useful for addressing topics related to urban design and planning, rather than attempting to uncover opportunities for personalised information services.

6.3. Transport information systems

There is a broad literature on predicting temporal features in the context of transport systems, ranging from bicycle stations' capacity [3] to car trip duration [31]. A simple yet important point differentiates our work from many in this domain: current solutions either do not have access to per-user data or explicitly focus on the aggregate usage. 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 prevents us from building trajectory or sub-route-based models (e.g., [14] or [11]) 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

Please cite this article in press as: N. Lathia, et al., Individuals among commuters: Building personalised transport information services from fare collection systems, Pervasive and Mobile Computing (2012), doi:10.1016/j.pmcj.2012.10.007

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). Indeed, [32] found that many users choose a longer and slower route through the underground system, having been influenced by the schematic design of the tube map; similar effects were also found in other public transport systems. Knowledge of the routes taken by travellers would certainly aid in the discovery of variability. 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.

In the context of transport systems, AFC datasets have been studied extensively as a means to assessing service quality and demand. Previous work includes using Oyster data for origin-destination matrix estimation and journey time reliability [13], and transit demand modeling [33]. For example, [34] uses another TfL dataset to estimate an origin-destination travel time matrix, and shows how this can be used to aid the service planning process. Similarly, [35], focusing on the London Overground subsystem, develops a set of service reliability metrics, and [13] uses an AFC dataset from the city of Seoul, South Korea, to reveal transfer and travel time patterns, also as an aid to service planning. The pervading theme of these research efforts is to demonstrate the advantages AFC data analysis offers over traditional survey methods. What is lacking from these studies, however, is detailed analysis of individual traveller patterns.

Recently, a number of mobile applications have emerged that support urban residents using public transport. These include Tiramisu [36] and OneBusAway [37]. Each application differs in how it collects information for the travellers; in fact, Tiramisu relies mostly on crowd-sourced contributions from bus-riding participants. The addition of further data, such as AFC data, can only extend the range of potential applications that can be built, and amplify the potential influence of personalisation of these services on travellers' decisions. Data such as travellers' actual GPS locations, measured by their mobile phones, social connections, leveraged from available online social networks, areas of interest, stored in the databases of location-based services, as well as preferences elicited via in-situ, automatic surveys are just a sample of the myriad available data sources that can contribute to building better ATIS services. The availability of such data would allow travellers to, for example, coordinate their travels with their friends and be notified about disruptions that will affect their journeys to events they intend to attend [7].

7. Conclusion

Transport operators collect large quantities of coarse-grained mobility data from their AFC systems. In this work, we have investigated how these data can be used to augment the quality of traveller information systems, with a particular focus on travel time estimates. We have showed how individual travel habits, when uncovered, differ from aggregate commuting patterns, and have identified a number of features (contextual, familiarity, self-similarity) that improve the accuracy of travel time estimates based on fare collection data. Our work above has centred on trip time prediction; however, the potential use of AFC data for building personalised information services extends well beyond this particular use case. In this section, we enumerate a number of other applications, related to both the transport domain and broader urban scenarios.

7.1. Future transport information systems

There are a number of studies that AFC data may be used for to further improve travel time estimates. First, our work has centred on uncovering individual differences, and thus we limited the scope of the forecasting techniques to explainable sample (e.g., user, context) means of the data. A clear direction of future work is to investigate the wide range of statistical machine learning techniques that may have more accurate estimating capabilities. The data may also be re-examined, to include facets such as: blending it with the network structure (e.g., quantifying the number of route choices between an origin and destination), examining cross-user similarity in order to predict "cold-start" trips (i.e., trips that a user has never taken in the past), as well investigating inter-trip similarity in order to, for example, estimate a trip from A to C based on trip segments (A, B), (B, C). A common source of estimated trip times is the TfL Journey Planner, which combines scheduled in-vehicle travel times with estimated interchange walking times, as well as real time service information such as delays and line closures. A definitive assessment of the predictive models we proposed could be gained by comparing the results to the estimates provided by TfL. Unfortunately, this was not possible for the study above as no historical data from the Journey Planner was available, not least from the period covered by the dataset.

While we have used the Oyster data as a means of measuring the time it takes travellers to move between an origin and destination, the entry/exit data could equally be treated as a unary "vote" of interest for that particular location. Combined with spatial information about the layout of the system, such a view of the data could be used to easily and quickly identify which subset of travellers will be affected by particular service disruptions [8]. This aspect of travel information becomes particularly critical as the scale of the transport system increases: recall that the TfL public transport infrastructure is a vast, multi-modal network of underground trains (11 interconnected lines with 270 stations), overground trains (5 lines with 78 stations) and buses (8000 buses serving 19,000 stops) as well as trams, river services, and other specialised services. This implies that, for any particular disruption or service update, a varying subset of the travellers using the system will find that being informed about the event is relevant to their mobility. Furthermore, disruptions are not necessarily geographically isolated events; they range in scale (e.g., at a particular platform, station, on an entire line), and will have unforeseen consequences as travellers route around them. We believe that AFC data can be used to examine, at a large-scale, not only *who* should be informed about disruptions, but *how* masses of travellers react to varying disruptions.

Similarly, the timeliness of notifications could be improved by leveraging AFC data. Currently, TfL provides travel alerts[3]; travellers who are interested in notifications relating to particular stations and routes must manually input the origin and destination stations as well as the time and frequency of travel (e.g., travel from Archway to Goodge St. at 8:00 AM). Automated emails (and, in the event of major disruptions, text messages) are sent to travellers around the travel times that they specified. By not being linked to AFC data, these alerts are invariably not adapted to the current location or travel habits of the user, and may thus be less relevant than a real-time, AFC-linked, location-based service could be. This problem is further highlighted by the fact that travellers' perceptions of their needs and habits often does not match their actual usage of the public transport system [38].

An alternative direction for future research is to investigate the clustering potential in the data—from both the user and spatial perspective. Being able to pre-process the AFC data by grouping similar users would lower the dimensionality of the data and allow for accurate predictions to be served to travellers using highly scalable approaches. Spatial clustering, instead, could be used to uncover how the transport network operates from a system-perspective, based on its usage; similar approaches have already been used to identify how the city self-organises into a polycentric, hierarchical structure based on mobility patterns [39].

7.2. AFC data: beyond information systems

In closing, we posit that the potential for AFC data may very well extend beyond transport information systems. For example, the rise of location-based social networks like Foursquare allow for detailed analyses of the structure and usage of the city [30]. Oyster data is, essentially, a less granular form of such data: it gives information about the whereabouts of individuals at coarse locations over different periods of time. To that end, this data could be used to both plan and measure the efficacy of outdoor and within-station advertising [40]; provide social and location based recommendations; and link travellers to local offers (e.g., offered by GroupOn). In fact, the repository of data held by the transport authority extends well beyond the data that we analyse here: they further hold data related to customers' fare spending habits [17] as well as fine-grained details about the type of card that the traveller is using (which holds a sizeable amount of demographic data on it; e.g., child, adult, pensioner): a vast array of opportunities to build, understand, measure, and engage with urban residents, via this data, exists and is waiting to be harvested.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7-SST-2008-RTD-1) under Grant Agreement n. 234239.

References

[1] F. Girardin, F. Calabrese, F.D. Fiore, C. Ratti, J. Blat, Digital footprint in: uncovering tourists with user-generated content, IEEE Pervasive Computing 7 (2008).
[2] Y. Zheng, Y. Liu, J. Yuan, X. Xie, Urban computing with taxicabs, in: ACM Ubicomp, September 2011.
[3] J. Froehlich, J. Neumann, N. Oliver, Sensing and predicting the pulse of the city through shared bicycling, in: 21st International Joint Conference on Artificial Intelligence, Pasadena, California, 2009.
[4] N. Wilson, The role of information technology in improving transit systems, in: Transportation@MIT Seminar, Boston, USA, 2009.
[5] G. Linden, B. Smith, J. York, Amazon.com recommendations: item-to-item collaborative filtering, IEE Internet Computing 7 (2003) 76–80.
[6] G. Adomavicius, A. Tuzhilin, Context-aware recommender systems, in: Proceedings of the 2008 ACM Conference on Recommender Systems RecSys 08 16, 335, 2008.
[7] D. Quercia, N. Lathia, F. Calabrese, G.D. Lorenzo, J. Crowcroft, Recommending social events from mobile phone location data, in: IEEE International Conference on Data Mining, December 2010.
[8] N. Lathia, J. Froehlich, L. Capra, Mining public transport usage for personalised intelligent transport systems, Data Mining, in: IEEE International Conference on, pp. 887–892, 2010.
[9] L.S. Weinstein, Introduction to transport for London, in: Online Presentation, London, UK, February 2008.
[10] L.S. Weinstein, TfL's contactless ticketing: Oyster and beyond, in: Transport for London, London, UK, September 2009.
[11] J. Froehlich, J. Krumm, Route prediction from trip observations, in: Intelligent Vehicle Initiative, SAE World Congress, Detroit, Michigan, 2008.
[12] W.H.E. Day, H. Edelsbrunner, Efficient algorithms for agglomerative hierarchical clustering methods, Journal of Classification 1 (1984) 7–24. 10.1007/BF01890115.
[13] W. Jang, Travel time and transfer analysis using transit smart card data, Journal of the Transportation Research Board (3859) (2010).
[14] H.V. Lint, Empirical evaluation of new robust travel time estimation algorithms, in: 89th Annual Transport Research Board, January 2010.
[15] Y. Hu, Y. Koren, C. Volinsky, Collaborative filtering for implicit feedback datasets, in: IEEE ICDM, 2008.
[16] J. Letchner, J. Krumm, Trip router with individualized preferences: incorporating personalization into route planning, in: 18th Conference on Innovative Applications of Artificial Intelligence, Boston, Massachusetts, 2006.

[3] see: http://alerts.tfl.gov.uk/.

[17] N. Lathia, L. Capra, Mining mobility data to minimise travellers' spending on public transport, in: ACM Conference on Knowledge Discovery and Data Mining, San Diego, California, USA, August 2011.
[18] M. Gonzalez, C. Hidalgo, A.-L. Barabasi, Understanding individual human mobility patterns, Nature 453 (7196) (2008) 779–782.
[19] D. Brockman, L. Hufnagel, T. Geisel, The scaling laws of human travel, Nature 439 (2006) 462–465.
[20] M. Foth (Ed.) Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City, 2009.
[21] N. Eagle, S. Pentland, Reality mining: sensing complex social systems, Pers Ubiquitous Computing 10 (2006) 255–268.
[22] F. Calabrese, C. Ratti, Real time rome, Networks and Communications Studies 20 (2006) 247–258.
[23] C. Ratti, R.M. Pulselli, S. Williams, D. Frenchman, Mobile landscapes: using location data from cell phones for urban analysis, Environment and Planning B: Planning and Design 33 (5) (2006) 727–748.
[24] V. Soto, E. Frias-Martinez, Robust land use characterization of urban landscapes using cell phone data, in: Workshop on Pervasive and Urban Applications, San Francisco, USA, 2011.
[25] V. Kostakos, T. Kindberg, et al., Instrumenting the city: developing methods for observing and understanding the digital cityscape, in: Proceedings of the 8th International Conference on Ubiquitous Computing, Springer, 2006.
[26] K. Sadabadi, M. Hamedia, A. Haghani, Evaluating moving average techniques in short-term travel time prediction using an AVI data set, in: Transportation Research Board 89th Annual Meeting, Washington DC, USA, January 2010.
[27] B. Storey, R. Holtom, The use of historic GPS data in transport and traffic monitoring, TEC, 2006.
[28] L. McNamara, C. Mascolo, L. Capra, Media sharing based on colocation prediction in urban transport, in: ACM International Conference on Mobile Computing and Networking, MobiCom, San Francisco, USA, pp. 58–69, September 2008.
[29] D. Quercia, J. Ellis, L. Capra, J. Crowcroft, Tracking "gross community happiness" from tweets, in: ACM Computer Supported Cooperative Work, Seattle, Washington, February 2012.
[30] A. Bawa-Cavia, Sensing the urban: using location-based social network data in urban analysis, in: Workshop on Pervasive and Urban Applications, San Francisco, USA, 2011.
[31] S. Handley, P. Langley, F. Rauscher, Learning to predict the duration of an automobile trip, in: 4th International Conference on Knowledge Discovery and Data Mining, New York, 1998.
[32] Z. Guo, Mind the map! The impact of transit maps on path choice in public transit, Transportation Research Part A: Policy and Practice 45 (7) (2011) 625–639.
[33] K. Chu, R. Chapleau, Enriching archived smart card transaction data for transit demand modeling, Journal of the Transportation Research Board (2063) (2008).
[34] F. Gordillo, The value of automated fare collection data for transit planning: an example of rail transit od matrix estimation, Master's Thesis, MIT, Department of Civil and Environmental Engineering, September 2006.
[35] 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.
[36] J. Zimmerman, A. Tomasic, C. Garrod, D. Yoo, C. Hiruncharoenvate, R. Aziz, N.R. Thiruvengadam, Y. Huang, A. Steinfeld, Field trial of Tiramisu: crowd-sourcing bus arrival times to spur co-design, May 2011.
[37] B. Ferris, K. Watkins, A. Borning, OneBusAway: results from providing real-time arrival information for public transit, in: Proceedings of CHI, Atlanta, GA, April 2010.
[38] N. Lathia, L. Capra, How smart is your smartcard? measuring travel behaviours, perceptions, and incentives, in: Proceedings of ACM International Conference on Ubiquitous Computing, Beijing, China, September 2011.
[39] C. Roth, S. Kang, M. Batty, M. Barthelemy, Structure of urban movements: polycentric activity and entangled hierarchical flows, PLoS ONE 6 (2010).
[40] D. Quercia, G.D. Lorenzo, F. Calabrese, C. Ratti, Mobile phones and outdoor advertising: measurable advertising, IEEE Pervasive Computing Magazine 10 (2011) 28–36.

Please cite this article in press as: N. Lathia, et al., Individuals among commuters: Building personalised transport information services from fare collection systems, Pervasive and Mobile Computing (2012), doi:10.1016/j.pmcj.2012.10.007