Mining Mobility Data to Minimise Travellers' Spending on Public Transport

Neal Lathia, Licia Capra
Department of Computer Science
University College London
Gower Street, London, WC1E 6BT, UK
n.lathia, l.capra@cs.ucl.ac.uk

ABSTRACT

As the public transport infrastructure of large cities expands, transport operators are diversifying the range and prices of tickets that can be purchased for travel. However, selecting the best fare for each individual traveller's needs is a complex process that is left almost completely unaided. By examining the relation between urban mobility and fare purchasing habits in large datasets from London, England's public transport network, we estimate that travellers in the city cumulatively spend, per year, up to approximately GBP 200 million more than they need to, as a result of purchasing the incorrect fares.

We propose to address these incorrect purchases by leveraging the huge volumes of data that travellers create as they move about the city, by providing, to each of them, personalised ticket *recommendations* based on their estimated future travel patterns. In this work, we explore the viability of building a fare-recommendation system for public transport networks by (a) formalising the problem as two separate prediction problems and (b) evaluating a number of algorithms that aim to match travellers to the best fare. We find that applying data mining techniques to public transport data has the potential to provide travellers with substantial savings.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Information Filtering

General Terms: Algorithms

Keywords: Mobility, Public Transport, Filtering, Recommender Systems

1. INTRODUCTION

At face value, the problem of buying a ticket for travel on an urban public transport system seems simple and mundane. However, this is only the case for cities where there is only one fare or ticket type available for purchase. Increasingly, various metropolises around the world are diversifying the fares that they offer, in order to cater for the different needs of the city's residents: they may offer single, multi-trip, and tourist (or visitor) tickets, as well as tickets that restrict travellers to specific transport modalities (e.g., a bus or train-only ticket). Invariably, all of these tickets are offered at different prices. Travellers are now faced with the problem of having to purchase a ticket from a large list of candidate fares, by first estimating their own future travel needs, and then selecting the one that seems to cater for their needs best.

The public transport network in London, UK is a prime example of this scenario. The rich pricing scheme operated by Transport for London (TfL) introduces a variety of ticket options that range in price, transport modality, temporal validity *and* geographical boundaries. In fact, the optimal ticket for each of the city's travellers will depend on *who* they are (which determines which discounts they are eligible for) and three factors that directly affect cost: *where* they travel to and from (i.e., their geographic requirements), *when* they travel (e.g., rush-hour or day time) and *how frequently* they move between places, over time periods that span from single days to an entire year. While these multiple behavioural dimensions influence what the cheapest travel option will be, the current ticket sales process has no transparent link between usage and pricing: the decision of the best ticket is left unaided to each individual traveller. As a result, travellers are never informed as to whether they are indeed making the best decisions for themselves. TfL itself estimates that over GBP 300,000 is wasted *per day* by passengers buying paper tickets instead of opting for the electronic equivalent[1], and other investigations have revealed that approximately GBP 30 million of travel credit is sitting in the system, idle and unused[2]. These vast sums of wasted money all point to the fact that making the correct decision at the point of purchase is not only uninformed and lacking in transparency, but also incredibly difficult for travellers to reason about, in order to purchase the cheapest fare for themselves.

There are two significant obstacles confronting this complex decision: first, travellers need to *understand* the relationship between mobility and fares. To that end, we describe London's current pricing system in Section 2, which includes 7 temporal, 9 geographical, and 12 user categories, with the various conditions that influence the suitability of certain tickets for travellers. Second, each traveller needs to rely on her own memory and anticipated mobility patterns in order to decide what the best ticket will be. This is where the greatest opportunity for data mining lies: London's RFID-based Oyster card, much like Seattle's Orca

KDD'11, August 21–24, 2011, San Diego, California, USA.

[1] http://golondon.about.com/od/londontransport/qt/oyster_card.htm
[2] http://www.bbc.co.uk/news/10162991

card or the Tokyo Metro system's tickets (based on near-field mobile phone communication) is a uniform payment system that produces a digital record every time a trip is made and a ticket is purchased. Mining the travel data that is created as travellers enter and exit stations can be used to aide them when purchasing tickets. In this work, we propose and evaluate recommendation algorithms that aim to leverage each traveller's mobility patterns in order to determine what the best fare is for each unique individual. In particular, we make the following contributions:

- We present an extensive analysis of anonymised ticket purchasing behaviour and public transport usage datasets (Section 3); we link the two datasets in order to examine the relation between mobility and purchase habits and quantify the extent that London travellers waste money by buying the incorrect fares: our data shows that travellers overspend by approximately GBP 200 million per year by buying the incorrect fares.
- We compliment this analysis with the results of a survey (Section 3.3) that explores the heuristics and methods that travellers currently adopt when selecting the best fare for themselves.
- We design and evaluate (Section 4) algorithms that provide personalised ticket-purchase recommendations to travellers with an available travel history. We do so by splitting the problem into two: predicting future travel habits and matching travel habits to fares. We evaluate our proposals with accuracy metrics and by quantifying how much money travellers in our datasets could have saved (had they followed our recommendations).

We recognise that our contributions may reduce the gross income of travel operators. However, these systems can be used to encourage the adoption of public transport by promoting the cost-effectiveness of travel.

2. BACKGROUND

The services and pricing structures adopted by public transport authorities around the world are not uniform. However, in general, all systems will have an inherent relation between usage and pricing. In this work, we focus on London, England, due to the availability of data (although we note that our techniques could easily be adapted to other systems). 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 (about 8,000 buses serving 19,000 stops) as well as trams, river services, and other specialised services. At the broadest level, travellers must opt to either use a single, contact-less smart card (the Oyster card) to pay for their journeys or buy paper-based tickets. In order to encourage the use of the Oyster card, facilities are in place for travellers to automatically add credit to their card (*auto top-up*) and buy or renew travel passes online and using machines in each station; the ease of purchase of all fares on the Oyster card is roughly similar. However, use of an Oyster card does not determine what fares travellers should buy. In particular, there are a number of factors that must be considered when selecting tickets:

1. User Categories. There are 12 user categories, ranging from full fare-paying adults, to students, children of varying age ranges, disability/60+"freedom" passes (which entitle bearers to free travel), war veterans, bus discounts, groups and school parties. If any of these users also have one of many different additional cards (e.g., a National Rail membership card), they are entitled to further discounts.

2. Modality Restrictions. There are some tickets that are for exclusive use on the bus network, while others do not limit the transport modality that the traveller opts for.

3. Geographic and Routing Restrictions. The TfL rail network is subdivided into nine concentric zones. Zone 1 covers central London and higher-numbered zones are progressively further away from city centre. The cost of travel by train is influenced by the number of zones that are traversed; for example, a rush-hour single fare from Zone 6 to Zone 2 that goes via Zone 1 costs GBP 4.50, while the fare for a trip between the same zones, *without* going via Zone 1 costs GBP 2.50. There are particular cases where the actual route that is taken will change the cost. For example, a rush-hour single between Zones 1 and 2 (e.g., Goodge St to Archway) costs GBP 2.50, but if mobility is restricted to the Euston-Watford Junction train line (e.g., Euston to South Hampstead), the fare is GBP 2.00.

4. Travel Cards or Pay As You Go. Travellers can opt to pay for their trips on a per-trip basis (using *"pay as you go"*, PAYG, fares), or use 7-day, monthly, or annual passes, also called *travel cards*, which allow for unlimited travel within purchased zones. It is important to note that PAYG and *travel card* use is not mutually exclusive. Travellers can opt for combinations of the two; for example, they can buy a Zone 1 to 2 travel card and then use *pay as you go* for any travel that goes beyond these Zones (both tickets are stored within the same physical Oyster card). In this case, they will be charged a mixed fare, which corresponds to the PAYG cost for a single fare in the zones that are not covered by their travel card.

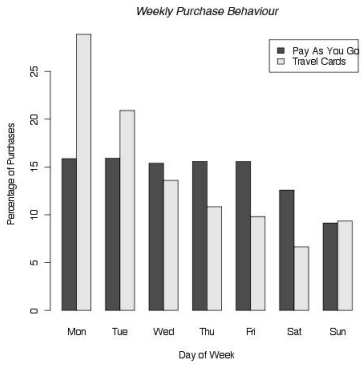
5. Temporal Categories. TfL has defined two distinct travel times: *peak* times, 6:30-9:30AM and 4:00-7:00PM Monday to Friday (reflecting the rush-hour commuting travellers) and *off-peak*, 9:30AM-4:00PM and after 7PM during week days and all of weekends and public holidays. PAYG as well as day-only travel card prices tend to, but do not always, differ in price between these times (note that the weekly/monthly/annual travel cards do not differentiate between peak and off-peak travel).

6. Additional Conditions. TfL has also implemented a *price capping* system for Oyster cards to limit the amount of PAYG credit that any one traveller can use in a single day. If a traveller has so many trips in one day such that it would have been cheaper to travel with a day pass, then the fare that will be charged will be the price-cap limit, which is the same as the day travel card ticket price (although there are separate peak and off-peak limits). Unfortunately, the capping does not transfer daily cards to further time periods (e.g., weekly, monthly), regardless of whether the latter would have been cheaper or not.

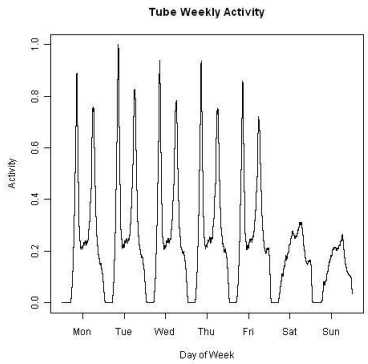
All the above conditions and related fare prices are subject to an annual review and continuous changes. Inextricably, a number of sub-optimal choices can easily be made: for example, using PAYG will be more expensive than travel cards for high frequency travellers. In this work, we will focus on full-fare paying adult tickets. How do travellers make purchases in this environment? In the following section, we examine the relation between mobility and purchase deci-

(%)	Type
Pay As You Go	
49.8	≤ GBP 5
24.2	GBP 5.01 - GBP 10
15.5	GBP 10.01 - GBP 20
7.7	GBP 20.01 - GBP 30
Travel Cards	
70.8	7-day travel card
15.8	1-month travel card
11.6	7-day bus/tram pass
1.9	1-month bus/tram pass

(a) Top-4 Purchase Types



(b) Weekly Purchase Distribution



(c) Week Ongoing Journeys

Figure 1: Fare Purchases: (a) the most frequently bought travel cards and top-up amounts, (b) *when* people purchase over a cumulative week, and (c) the cumulative ongoing (rail) journeys across a week.

Name	Users	Rail/Tube Trips	Bus Trips
D1	264,304	4,350,039	5,014,664
D2	267,357	4,315,821	4,734,435

Table 1: Two 83-day Travel history datasets of a 5% sample of TfL travellers. D1 is May-July 2009, D2 is October 2009-January 2010.

Name	Number of Purchases	
	PAYG	TravelCards
P1	1,646,987	134,721
P2	1,732,583	125,704

Table 2: Ticket purchase datasets from the same users and date ranges as Table 1.

sions by analysing two large datasets of London trips and fare purchases.

3. MOBILITY AND TICKET PURCHASES

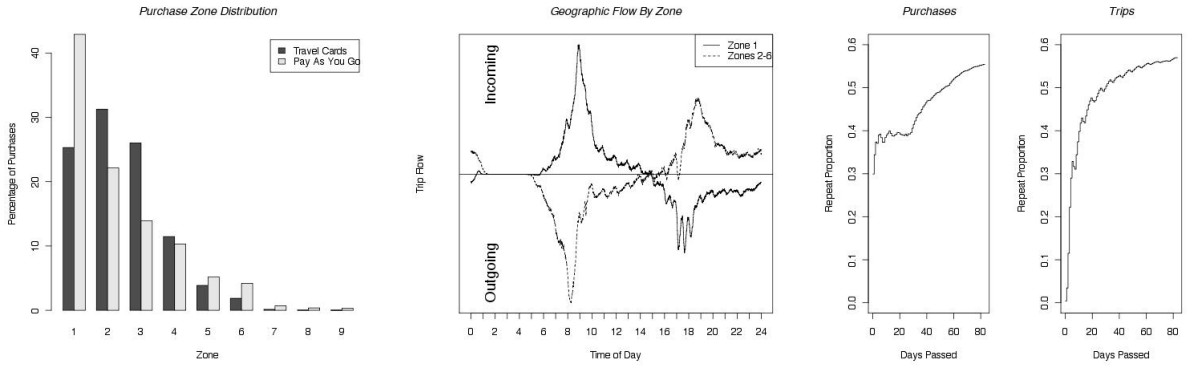
In this section, we first analyse anonymised purchase and travel behaviour data from TfL (Section 3.1), highlighting emerging trends and consistencies in both travellers' movements and purchases. We then quantify how much travellers overspend, by computing the cheapest fares for the trips they took (Section 3.2). Finally, we investigate the heuristics that travellers adopt to support purchase decisions, by reporting the results of an online survey (Section 3.3).

3.1 Dataset Analysis

In this analysis, we use two pairs of datasets of Oyster card usage from different 83-day periods (May-July 2009 and October 2009-January 2010). Each dataset is a 5% sub-sample of all users who were recorded during the two periods. The *mobility* datasets (D1 and D2), contain the travel history of the sampled users: an anonymised, unique user id, the modality, origin, destination, the journey's start and end times in minutes, and the travel ticket type as recorded by the user's Oyster card (note that, as travellers are not required to use their card to exit buses, we only have origin and start time for these trips; this does not affect pricing as bus fares do not depend on the trip destination). The purchase history datasets (P1 and P2) contain both PAYG credit and travel card purchases during the same time frame, along with where (which station), when (which day), and how much credit/what type of travel card (temporal and geographical boundaries) was purchased. Together, the two datasets sum to traveller spending of over GBP 500 million.

We took a number of steps to clean the data. First, the datasets suffer from the *edge effect*: some users will be making trips with travel cards that were purchased prior to the purchase history date range that we observe. Second, the purchase data has *missing entries*, potentially due to lost data from ticket sales done through authorised retailers. Note that, since PAYG fares are paid on a per-journey basis, these entries do not have this problem. To compensate for the edge effect, we pruned the profiles of those travellers who use travel cards until the date of the first (observed) purchase. If there was no observed purchase, or the traveller was using an annual card, we removed them from the dataset; we lost approximately 20% of the users. We then pruned any inconsistent or erroneous entries (trips with the same origin and destination, with end times prior to start times, or with unknown tickets), and then sampled all adult Oyster cards to produce the final datasets; an overview of these is provided in Tables 1 and 2. In this section, we analyse the combined datasets in order to examine the relation between purchase behaviour and mobility.

1. Ticket Purchases: travellers buy credit in small increments. Figure 1(a) shows the top-4 most popular travel card purchases (grouped by temporal validity) and PAYG credit amounts (grouped in GBP5 increments); these account for, respectively, 95.5% and 98.6% of all the purchases. This clearly shows that travellers tend to opt to *buy credit in small increments*: half of the credit purchases are less than GBP 5 and 74% are less than GBP 20, which indicates that credit is bought on an as-required basis. 7-day long passes are the most popular travel card purchase. We inspected this further by looking at the use of travel cards and PAYG by individual travellers. Overall, a majority of travellers do not use travel cards: 75.9% and 77% of the users in D1 and D2 respectively never make a trip with a travel card. However, travel cards are used more often than PAYG on bus trips (70% of trips are made with travel cards); having a travel card is an incentive to use the bus. This trend



(a) Zone Purchase Distribution (b) Geographic Flow by Zone (c) Repeat Trips and Purchases

Figure 2: *Pay As You Go* Credit and Travel Cards: (a) the geographic purchase distribution, by zone, (b) the cumulative number of entries and exits (flow) into Zone 1 and outside Zone 1, and (c) the average proportion of repeat trips and purchases in the datasets over time.

flips for rail trips (38.2% travel card, 59.4% PAYG), with the remaining fraction being mixed fares.

2. Temporal Distribution: travel cards are bought on Mondays. Figure 1(b) shows the temporal distribution of purchases in a week: while credit purchases tend to remain constant throughout the week, travel cards are mostly bought in the early days of the week. Again, travellers are viewing the system on a weekly-basis, and may thus be discounting the potential benefit of month/year passes. Figure 1(c) shows the the cumulative number of ongoing trips over 7 days; the commuting majority dominate over the week day patterns.

3. Geographical Distribution: PAYG travellers do not plan ahead. Figure 2(a) plots the geographic distribution of purchases, split by fare zone (recall that Zone 1 is central London, with higher zones being concentric circles around it). The majority of PAYG purchases occur within Zone 1, while there are more travel cards bought in Zones 2 and 3 than central London. A potential explanation for this comes from the aggregate trip data. After clustering the stations by fare zone, we can count arrivals (+1) and departures (–1) from each zone and sum these two together. At any given time period, a zone's flow is positive if there are more arrivals than departures, and negative if there are more departures than arrivals. The resulting data is shown in Figure 2(b). The only zone to be positive in the morning (arrivals) and negative in the afternoon (departures) is Zone 1; all other zones, which have been grouped in Figure 2(b), each *individually* show departure trips in the morning and arrival trips in the afternoon/evening. Relating purchase behaviour to traveller flow indicates that PAYG credit, must be mainly purchased between commuter's outbound and return journey, while travel cards are more frequently bought before or after the day's travel.

4. Repeat Behaviours: travellers do the same thing again and again. To what extent are travellers making the same trips and buying the same fares repeatedly? We examined this question by defining a repeat trip made by an individual as a *train* trip with the same origin and destination as a previously taken trip (we do not include bus trips since we do not know where travellers alighted). Similarly, we defined a repeat purchase as either a PAYG credit purchase with *exactly* the same amount as has been purchased before, or a travel card that has been bought before. After the 83 days of each dataset, we find that over half of all the trips *and* purchases have been seen before, as shown in Figure 2(c): travellers are highly regular in both their movements and purchases.

Based on the above analysis, two facts emerge: (1) there exists high *regularity* (and thus predictability) of both travel behaviour (Figures 1(a) 1(b) 1(c)) and of purchase behaviour (Figures 1(a) 2(c)); (2) there exists a strong correlation between travel behaviour (e.g., what travel zones are crossed) and purchase behaviour (e.g,. what ticket type, PAYG versus travel card are purchased) (Figures 2(a) 2(c)). However, a question arises as to whether this *repeated correlation* between travel and purchase behaviours is optimal, or whether travellers are wasting their money instead. In the following section, we quantify how much people overspend by computing the optimal fares for each traveller.

3.2 Potential Savings

In this section, we look at what travellers *could have* saved, had they made the optimal purchase decisions for all their travel habits, that is, had they known, a priori, exactly what trips they were going to make and purchased the cheapest sequence of fares. We first encoded the tube network structure and implemented Floyd's algorithm [1] to find the shortest path between any origin and destination (by number of hops). This way, we can infer which zones each trip should traverse and thus which PAYG, travel card, and mixed fares are applicable. Although our data is from 2009, we used 2011 fares since historical fares are not currently available; this does not affect our results since (a) we are using the same fares to compute both the "actual" and optimal trip costs and (b) the relative benefit of one fare over another is the same.

We then compute, for each individual, a sequence of optimal fares by building a tree, where each node is a ticket and a chain of linked nodes denotes a sequence of purchase decisions. Tickets have two corresponding costs: the initial cost (which is zero for PAYG and non-zero for travel cards), and any additional charges that a traveller is subject to while using that ticket (e.g., mixed-fare trips). The nodes also have a geographic validity (i.e., zones) and expiry date: single fares expire immediately, while travel cards expire a

Spending	Cumulative	Trip Avg	User Avg
D1 Actual	13,797,168.10	1.47	52.20
D1 Optimal	11,395,244.90	1.22	43.11
Difference	**2,401,923.20**	**0.25**	**9.09**
D2 Actual	13,393,208.20	1.47	50.09
D2 Optimal	11,196,687.40	1.24	41.88
Difference	**2,196,520.80**	**0.23**	**8.21**

Table 3: Results (GBP) for D1 (May - July 2009) and D2 (Oct 2009 - Jan 2010) with the actual and optimal spending, quantified as cumulative, trip average, and user average spend.

number of days after purchase. For each trip that a traveller makes, the leaves of the tree that are expired or not valid are expanded by adding child nodes representing the tickets that the traveller could select from. The cheapest fare is computed with a depth-first search on the tree. Brute-force exhaustive search on this tree would have a space complexity of $O(b^d)$, where the branching factor b is the number of available fares and the depth d is the number of trips taken by the user. The resulting computational cost is prohibitive for all users who have taken more than a handful of trips in the 83-day periods covered by our datasets. We have thus adopted a number of heuristic expansion and pruning rules that reduce this search space [2], as we describe next.

Expansion Constraints. First, we implemented TfL's price "capping" system that prevents a traveller's daily PAYG cost from exceeding that of the relevant travel card. Second, we consider the *geographical* requirements of all the user's trips. For example, if a traveller will only commute between Zone 3 to Zone 4, then fares and travel cards that do not include these areas (e.g., a Zone 1-2 travel card) are not amongst the available options to expand the leaves. When expanding the tree, we further reduced the number of candidate child nodes by taking transport modality into consideration; for example, if we are expanding based on a rail trip, then bus-only fares are not included.

Pruning Rules. We also defined a *pruning* function that, as the tree grows, removes the subtrees with ticket sets that are already more expensive than those in other subtrees. More formally, let us define $expire(X)$ as the expiry date of the cheapest fare sequence starting from a generic node X, and $cost(X)$ as the total cost of the cheapest fare sequence from node X. Given two sibling nodes X and Y, we prune X's branch if the following holds:

$$\big(expire(X) \leq expire(Y)\big) \wedge \big(cost(X) > cost(Y)\big) \qquad (1)$$

In other words, if a ticket has a shorter temporal validity than another and is already more expensive, then it can be pruned. Similarly, if a node with a smaller geographic range is already more expensive than another with a larger range, then this is due to the extra incurred cost of mixed-fare trips and the node can be pruned. For example, a ticket for Zone 3 to 4 is "smaller" than Zone 3 to 5, and any trips between Zone 3 and 5 using it will be charged at a mixed-fare rate.

The potential savings are shown in Table 3. Overall, we find that this 5% sample of travellers are spending just under GBP 2.5 million more than they need to. Overall, each user in D1, on average, overspends by GBP 10. Two points to note are: (a) using this sample of people to approximate the entire population would mean multiplying these figures by 20 and (b) both datasets only cover an 83 day period. In other words, we estimate that travellers cumulatively spend up to approximately GBP 200 million per year more than they need to, by simply buying the incorrect fares.

In the following section, we aim to answer this question by uncovering the heuristics used by travellers to guide their purchase choices; we will then propose algorithms that can provide travellers with cost-saving ticket recommendations.

3.3 Purchase Decision Survey

There are a number of aspects relating to purchase decisions that the datasets do not reveal; the datasets cannot tell us *why* people opt for the fares that they buy and any biases that they may be subject to. To explore this space, we designed an online survey. The survey was divided into three sections: (a) questions about *travel habits*: travel times, typical trip origins/destinations, and transport modes that are most often used; (b) questions about *purchasing habits*: typical top-up amounts and travel cards bought and why they are purchased; and lastly (c) open-ended questions about their impressions of cost-saving and fare selection. The survey was disseminated online (via twitter and mailing lists) and completed by 119 travellers (30% students).

The origin/destination pairs reported show that 92% of respondents travel into Zone 1 during a typical week day, much like the results we found in Figure 2(b). While self-reported travel behaviour broadly correlates to that observed in the two large datasets, we found significant differences in terms of habits: only 7% of respondents claimed to top-up by less than GBP 5, which does not match the proportion of these transactions (49.5%) we found in the datasets. Furthermore, a large proportion of respondents stated that they never buy travel cards (46%), although this falls short of the proportion of users in the larger datasets who also exhibited this behaviour. Only 18% of respondents claimed that they typically buy a 7-day travel card, although we had previously found this to be the most popular purchase choice.

These differences between self-reported and observed purchasing data can be explained in two ways: they may be due to the demographic biases in this sample of survey respondents (even though their self-reported travel behaviour did match that observed in our datasets), or they may be a consequence of the divergence between travellers's perceived versus actual purchase behaviour (e.g., they may be unaware of the high volume and cumulative effect of small transactions that they carry out). We explored the misalignment between travellers' perceived and actual behaviour further, by asking *why* they opt for the fares they buy. PAYG was preferred when users expected their travel to be irregular (38.8% of the total votes), and when they believed this was the cheapest option for them (with 32.2% of the votes). Travel cards were preferred for convenience (39.4%), and again because it was thought to be cost-saving (44.9%). However, despite money-saving being a top priority overall, our previous analysis demonstrates massive amounts of money is actually being wasted by the travellers. We also asked respondents about *when* they top up their PAYG credit versus when they purchase a travel-card, and the majority of them (77%) reported that they only top up when they have insufficient credit to enter the system, demonstrating no element of planning whatsoever in this behaviour.

In the following section, we show how data mining tech-

		Avg Trips/Day	Geography		Travel Time		Travel Modality	
Method	**Span**	**MAE**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Last Profile	1	0.72	0.9974	0.9976	0.8774	0.8313	0.9198	0.9161
	7	0.41	0.9939	0.9940	0.8532	0.8509	0.9049	0.9050
	30	0.30	0.9891	0.9899	0.8955	0.9172	0.9303	0.9298
Avg Profile	1	0.76	0.9996	0.9693	0.7267	0.7881	0.8239	0.8935
	7	0.38	0.9978	0.9759	0.8607	0.8214	0.9123	0.8722
	30	0.29	0.9916	0.9845	0.9156	0.9149	0.9444	0.9231
Moving Avg	1	0.69	0.9979	0.9963	0.5570	0.8285	0.7181	0.9443
	7	0.37	0.9956	0.9908	0.8096	0.8689	0.8866	0.9131
	30	0.29	0.9916	0.9844	0.9039	0.9192	0.9389	0.9314

Table 4: Averaged Results for D1 and D2 when predicting users' travel profiles: How many trips per day will they take? What zones will they travel between? Will they only travel off-peak? Lastly, will they only take bus trips?

niques can be applied to aide travellers by providing fare recommendations based on their trip history.

4. RECOMMENDING TICKETS

We formalise the ticket recommendation problem as follows. If we consider the factors that influence travel cost (as outlined in Section 2), we can conclude that knowing a traveller's *detailed* future trips, as origin-destination pairs, is not strictly necessary in order to recommend the best fare. This is due to the zone structure of the stations: the fact that a user went from Leicester Square to Heathrow Airport is not needed to determine cost (all that is required is that a trip was taken from Zone 1 to Zone 6). Instead, a number of features shared by these trips are more relevant. Determining the appropriate fare can be based on an estimate of the broad aspects of a traveller's habits, such as frequency and geographic areas of travel. More formally, given a sequence of T(u) trips by user u, a traveller's habits between time t and $(t + \Delta)$ can be summarised as:

$$P_u(t, t + \Delta) = \{d, f, b, r, pt, ot, G\} \quad (2)$$

where d is the number of trips, f is the average trips per day, b and r represent the proportion of trips that were taken with buses and by rail respectively, pt and op are the proportion of peak and off-peak trips, and G is an N×N matrix, with N = the number of zones in London, and each $G_{i,j}$ is the frequency count of trips between these two areas. We assume that, for other cities, a similar profile could be designed with a basic understanding of the public transport fare structure.

Given this set up, the task of recommending the next suitable fare(s) can be decomposed into two prediction problems. The first is predicting a user's *future mobility patterns* $P_u(t, t + \Delta)$, given a sequence of P_u from prior to t. The second problem is to fit the appropriate fare to a profile of mobility requirements (i.e., predicting a ticket given a travel profile). In the following sections, we look at and evaluate each of these prediction problems individually.

4.1 Predicting Future Travel Habits

The first step in determining the best fares for a traveller is forecasting their travel habits, in order to understand their needs. As above, each user's trips over a particular period can be condensed into a *profile* of values representing their geographical, temporal, and modality requirements. The objective here is therefore to predict these values, given a history of profiles. We applied a number of baseline algorithms to this problem.

1. **Last Profile:** assumes that travel habits are constant; a future profile between time t and $(t+\Delta)$ is predicted to be the same as the profile between $(t - \Delta)$ and t.

2. **Cumulative/Average Profile:** leverages more historical data than the last value; the modality, temporal and geographic features are simply sums of past behaviour (e.g., we predict that a user will travel in Zone 1-2 if that traveller has trips between these zones in the past). If we define H as the *set* of profiles, each of span Δ, that come before t:

$$H = [P(t_0, t_0 + \Delta), ..., P(t - \Delta, t)] \quad (3)$$

   Then the future profile trip frequency is defined the average of the past ones:

$$P_u(t, t + \Delta) = \frac{1}{|H|} \sum_{i \in H} H_i \quad (4)$$

3. **Moving Average Profile.** Lastly, we define a moving average profile, which gives additional weight to more recent profiles. At each time interval $(t, t + \Delta)$, the trip frequency f is defined as:

$$f_{t,t+\Delta} = \alpha \cdot f_{t-\Delta,t} + (1 - \alpha) \cdot f_{t-2\Delta,t-\Delta} \quad (5)$$

   All members of the traveller's profile are similarly updated, with the same scaling factor α.

Our experiments proceed as follows: given each user's trips, we can form a set of n profiles for a given time length (e.g., Δ =1-day profiles). We then use the historical profiles $[0, 1, ..., n - 1]$ in order to predict the nth profile: Δ remains fixed. We repeat this process for all users and varying time lengths. Our evaluation metrics focus on the aspects of these predictions that will directly affect cost: we first estimate the average trips per day for a given user, and measure error with the mean absolute error (MAE); we then measure precision and recall of classifying travel within each pair of geographic zones, whether a user will *solely* take bus trips, and whether a user will *only* travel during off-peak times.

All the results are shown in Table 4. We found that users' travel habits are easily estimated with these baselines: for example, precision and recall are consistently above 98% across all methods, when predicting what zones each user

will travel in. A number of interesting trends also appear: for example, estimating average trips per day becomes more accurate over longer time spans. Classifying whether travellers will solely travel in off-peak times and whether they will only take bus trips is estimated slightly less accurately, although precision and recall values are still very high (91-94%). To explain these highly accurate results, it is important to reinforce that we are predicting habits over very coarse features (zones rather than stations): at this granularity of geography, travellers are incredibly consistent.

Note that each of the algorithms we used views the components of travellers' profiles as independent from one another (which is not the case); there are a number of models that could leverage inter-dependencies between profile features in order to improve predictions. However, since we found that even these simple baselines produce highly accurate results, we leave an examination of more complex models [4] as future work.

Note that, in the above approach, each algorithm was computing predictions over a pre-defined prediction window Δ. In practice, in order to recommend tickets to travellers, multiple windows of different length will need to be considered simultaneously ($t + \Delta$, $t + \Delta'$, $t + \Delta''$, etc.), in order to be able to recommend what the best next purchase is, between tickets of different time validity (PAYG, week, and month travel cards), which we leave as future work.

4.2 Predicting the Next Cheapest Fare

The next problem we need to solve is matching travel habits to the cheapest fares; this setting heavily relies on an accurate forecast of each user's travel profile. In this work we assume that these were predicted accurately. In future work, we plan on investigating the effect of prediction errors in users' travel profile on the ticket recommendations they receive. We regard this as a generic classification problem: given a vector of values representing trip habits, we need to select from a set of categorical labels representing tickets. Classification has been widely addressed in the literature and there are a range of algorithms that are readily applicable to this scenario. Our evaluation proceeds as follows: given the trip data described above, we first compute the travel profiles and optimal fares for each user (as per Section 3.2). We thus produce a dataset of profile instances of *varying lengths* Δ, each with the corresponding label. We then randomly split this data into training (80% of the users) and test (20% of the users) sets. We split by user so that we could quantify how much these users would save with our algorithms; we also measure the accuracy of classifying the instances in the test set as the number of correctly classified instances over the total number of instances. The techniques we evaluated on our data include:

1. **Baseline**. This classifier simply returns the most frequent class in our training set (PAYG). Inherently, we expect this classifier to do relatively well; by (correctly) predicting the fares for those users who should travel on PAYG, it corrects the cases of users who bought travel cards without needing to do so.

2. **Naïve Bayes**. This classifier, based on Bayes' theorem, assumes that each feature of the users' profiles is independent from the others. Given a profile with n features $P_u = \{F_1, F_2, ..., F_n\}$, the probability that the best fare is ticket C is estimated as:

$$p(C|P_u) = p(C) \prod_{i=1}^{n} p(F_i|C) \quad (6)$$

The best ticket is then selected as the class with the highest probability. For all features that do not relate to geographic requirements (i.e., all except feature G above), the posterior probability that feature F_i in class C has value v is estimated with a Gaussian distribution that is parameterised with the mean, μ_C, and variance, σ_C^2, of the feature F_i in class C:

$$P(F_i = v|C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} e^{-\frac{(v-\mu_C)^2}{2\sigma_C^2}} \quad (7)$$

The geographic features, instead, are transposed into binary variables, and the posterior is estimated as the proportion of instances in class C where F_i is non zero.

3. **k-Nearest Neighbours**. This technique operates by finding, for each test profile, the k most similar profiles; the predicted class is the most frequent class that appears in the neighbour set. We first defined similarity as the absolute difference between two profiles (thus, smaller values indicate higher similarity). We also introduced two small modifications: we gave higher weight to the average trips per day and days of travel (by using the squared difference) than to the proportion of bus trips (by ignoring the similarity between pairs of profiles where one only rides the bus and the other does not).

4. **Decision Trees**. The C4.5 algorithm [5] is a statistical classifier that generates a decision tree which can be used to classify test instances. It does so by recursively partitioning the data on a single attribute, according to the measured information gain of each split, where gain is defined relative to the *entropy* $E(S)$ of each group S:

$$E(S) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \ , \quad (8)$$

where n is the number of classes (i.e., tickets) and $p(x_i)$ is the proportion of S belonging to class i. In this work, we used the open-source implementation of C4.5 from the WEKA project [6].

	Accuracy (%)		Savings (GBP)	
Method	D1	D2	D1	D2
Baseline	74.99	76.91	326,447.95	306,145.85
Naïve Bayes	77.46	80.71	393,585.81	369,232.24
k-NN (5)	96.74	97.09	465,822.17	426,375.85
C4.5	98.01	98.29	473,918.38	434,082.81
Oracle	100.0	100.0	479,583.91	438,923.30

Table 5: Ten-fold cross validated classification accuracy for each dataset and average cumulative savings if travellers used each algorithm's recommendation.

We measure each classifier's performance with the proportion of correctly classified test instances (which we denote as accuracy); a summary of the ten-fold cross validated results

is shown in Table 5. The accuracy of the baseline is due to the large proportion of users in our datasets who have very few trips (thus making PAYG their cheapest fare). However, all the algorithms were able to outperform the baseline by varying amounts: the C4.5 decision tree produced the most accurate results, at over 96% and 98% for D1 and D2 respectively. We also quantified the savings that our travellers could have cumulatively achieved, had they followed each algorithm's recommendation, along with the maximum possible savings computed with the tree-based method described in Section 3.2 (denoted "oracle"). The total amount of possible savings, for both datasets, were over GBP 400 thousand. In both cases, the savings obtained by the C4.5 decision tree were less than 5 thousand pounds away from the optimal: over 99% of the potential savings were obtained by this algorithm's classifications. Interestingly, the baseline also provides modest savings: it seems from this that a large waste is generated by people buying travel cards that they then do not use.

5. RELATED WORK

By combining data from public transport and fare purchases, this study has a wide range of related research. At the broadest level, we categorise these into two different groups: understanding *mobility*, which has been studied using a variety of different data sources, and investigating and mining *decision making* contexts, by building recommender systems.

5.1 Mining and Modeling Mobility

Insight into human mobility is recognised as the key to understanding a variety complex phenomena, such as the spread of disease and traffic pattern forecasting [7], as well as giving an understanding into urban design and flow. It has been studied using mobile phone data [7, 8], travel data [3], and bank note dispersal patterns [9]. Fare collection data provides an alternative data source into people's commuting habits and mobility; the analysis in Section 3 highlighted a range of behaviours, ranging from aggregate, city-wide flows to individual modality choices.

These works examine mobility in isolation. Richer scenarios have also been researched: for example, Graham and Glaister [10] developed models to test the influence of traffic pricing on congestion and travel times. More generally, many transport-related projects are dedicated toward understanding mobility and building advanced traveller information systems (for example, [11, 12]). These systems all aim to help people travel, but rarely factor in the cost of doing so; we found that, in the context of public transport, there have been no efforts dedicated to helping commuters buy cost-saving fares.

5.2 Decision Making

Decision making is also at the forefront of research and popular literature [13]. In the field of data mining, there are two questions that arise: the first is *when* to make a decision in a volatile environment. For example, Etzioni *et al.* [14] mine airfare data in order to predict when the best time to buy a plane ticket is; in this case, airfare prices vary over time according to hidden variables (such as seat availability). Much like algorithmic trading, the basis of the decision is to predict the minimum of a time-varying price. The methods developed by these researchers do not apply to the problem of public transport fare selection, as prices will not vary at such heightened frequencies. Instead, our context relates to decision making in the face of an abundance of options. In general, the solution for the information overload problem [15] in web settings (e-commerce, movie rentals and music) is building online recommender systems [16]. As the web becomes evermore mobile, the breadth of contexts addressed by recommender systems is also growing to include urban navigation [17]. While personalisation and recommender systems have been widely adopted in online environments, there still exists a broad variety of contexts where users need to make complex decisions without the aide of these technologies. One area where recommender systems have little to no presence is the domain of the *individual, off-line* financial transactions and purchase decisions that people need to regularly make. This paper considers one such example: the problem of selecting which public transport ticket is best suited to a particular traveller.

6. CONCLUSION

In this paper, we investigated the financial benefits that can be offered to travellers after mining their mobility data. At face value, our proposals may be seen as detrimental to business, since they lower the gross income of the transport authority. However, by offering a service that enables travellers to understand and improve their spending habits, the adoption and use of public transport may actually rise, as travellers know that they are being given the best fare.

The details of this study are location-specific: we focus on the fare and public transport network structure in London, and also on the data that is available from the AFC system adopted by TfL. The applicability of these techniques to other cities will rely on the data collection and fare structures in place; for example, the fare scheme in Seoul, Korea, also factors in distance [18]. However, an overriding conclusion is that the *availability* and *mining* of AFC data presents many opportunities for personalised, dynamic services that cater for individual travellers [19].

We first evaluated baseline algorithms for estimating traveller mobility patterns, and found them to be highly consistent. We then evaluated algorithms that, given a travel profile, can predict what the best fare will be. Both prediction problems remain considerate of each traveller's privacy: they do not require the full trip history of each user to be stored, but only a less fine grained summary profile.

The evaluation metrics that we used focused on *accuracy*, *precision*, *recall*, and *potential savings*. In practice, rank-based metrics may also be helpful, since a deployed system would not want to recommend a single ticket for purchase, but rather present the most appropriate set of fares to each traveller, allowing them to make their own decisions. However, a useful side-effect of the classification algorithms that we evaluated is that they also pave the way for *explanations* to be given to travellers (e.g., we recommend the week pass since you tend to travel x times per week but only y times per month). In fact, a key role of this system would be to simplify the ticket purchase decision process, by decreasing the number of fare options offered to each traveller, so that they are aligned with their travel behaviour; as a consequence of the *informed* choices being made, users may then adapt their behaviour based on the fares they have purchased.

We conclude by noting how similar techniques could be applied to a myriad of scenarios: wherever traces of hu-

man behaviour can be collected, data mining can be used to complement the daily decisions that people make. Examples include recommending telephone contracts (or call plans) based on mobile phone usage or recommending investment and saving plans based on bank account data.

Acknowledgements. The authors would like to thank Angela Sasse and Philip Inglesant for help with the survey and Tamas Jambor, Joachim Neumann and Xavier Amatriain for their comments. 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.

7. REFERENCES

[1] R. W. Floyd. Algorithm 97: Shortest Path. *Communications of the ACM*, 5(6), June 1962.

[2] R. Zhou and E. A. Hansen. Breadth-First Heuristic Search. *Artificial Intelligence*, 170, April 2006.

[3] C. Roth, S.M. Kang, M. Batty, and M. Barthelemy. Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. *PLOS ONE*, 6.

[4] T. W. Anderson. *The Statistical Analysis of Time Series*. Wiley-Interscience, 1994.

[5] J. R. Quinlan. Improved Use of Continuous Attributes in C4.5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996.

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, and P. Reutemann. The WEKA Data Mining Software. *SIGKDD Explorations*, 11, 2009.

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

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

[9] D. Brockmann, L. Hufnagel, and T. Geisel. The Scaling Laws of Human Travel. *Nature*, 439, January 2006.

[10] D. J. Graham and S. Glaister. Spatial Implications of Transport Pricing. *Journal of Transport Economics and Policy (JTEP)*, 40(2), May 2006.

[11] B. Ferris, K. Watkins, and A. Borning. OneBusAway: Results from Providing Real-Time Arrival Information for Public Transit. In *ACM CHI*, Atlanta, GA, 2010.

[12] J.L. Ambite, G. Barish, C.A. Knoblock, M. Muslea, J. Oh, and S. Minton. Getting from Here to There: Interactive Planning and Agent Execution for Optimizing Travel. In *14^{th} Conference on Innovative Applications of Artificial Intelligence (IAAI)*, pages 862–869, Menlo Park, California, 2002.

[13] B. Schwartz. *The Paradox of Choice: Why More is Less*. Harper Perennial, 2004.

[14] O. Etzioni, C. A. Knoblock, R. Tuchinda, and A. Yates. To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price. In *ACM KDD*, Washington DC, USA, 2003.

[15] A. Borchers, J. Herlocker, J. Konstan, and J. Riedl. Ganging up on Information Overload. *IEEE Computer*, 31:106–108, 1998.

[16] F. Ricci, L. Rokach, B. Shapira, and P. Kantor, editors. *Recommender System Handbook*. Springer, 2010.

[17] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. Recommending Social Events from Mobile Phone Location Data. In *IEEE ICDM*, Sydney, Australia, December 2010.

[18] W. Jang. Travel Time and Transfer Analysis Using Transit Smart Card Data. *Transportation Research Board*, 2010.

[19] H. Bryan and P. Blythe. Understanding Behaviour Through Smartcard Data Analysis. *Proceedings of the Institution of Civil Engineers: Transport*, 160(4), 2007.