# Mining Temporal Patterns of Transport Behaviour for Predicting Future Transport Usage

**Stefan Foell**
Computing Department
The Open University
Milton Keynes, UK
stefan.foell@open.ac.uk

**Gerd Kortuem**
Computing Department
The Open University
Milton Keynes, UK
gerd.kortuem@open.ac.uk

**Reza Rawassizadeh**
Computing Department
The Open University
Milton Keynes, UK
reza.rawassizadeh@open.ac.uk

**Santi Phithakkitnukoon**
Computing Department
The Open University
Milton Keynes, UK
santi.phi@open.ac.uk

**Marco Veloso**
Centro de Informtica e Sistemas
Universidade de Coimbra
Coimbra, Portugal
mveloso@dei.uc.pt

**Carlos Bento**
Centro de Informtica e Sistemas
Universidade de Coimbra
Coimbra, Portugal
bento@dei.uc.pt


*UbiComp '13 Adjunct*, Sept 8-12, 2013, Zurich, Switzerland.
ACM 978-1-4503-2139-6/13/09...$15.00.

## Abstract
There is huge potential in increasing the value of public transportation by creating novel travel information systems which are centred on the individual transport user. Especially, in dense urban cities where it is hard to oversee complex transport networks that are subject to frequent changes, maintenance and construction works, travellers want to be proactively notified about disruptions and traffic incidents relevant to their future behaviour. In this paper, we show how to mine characteristic patterns of the transport routines of urban bus riders for the design of novel travel information system that have the ability to understand forthcoming travel needs of individual users. We leverage on travel histories collected from automated fare collection system (AFC) to extract features of personal transport usage and study their predictive power to forecast whether people access public transport services on a future day or not.

## Author Keywords
Public Transport, Bus Rides, Transport Usage Prediction, Automated Fare Collection Data

## ACM Classification Keywords
H.2.8 [Database applications]: Data mining.

## Introduction

Innovation in public transport system is experiencing a remarkable paradigm shift in recent days. Historically, the effort in optimising transport experiences has mainly targeted the underlying infrastructure where the priority has been on providing reliable and cost-effective transport services [1]. However, recently the focus has shifted towards the richness of available information about public transport services which allow travellers to make more informed transport choices [4]. In particular, the design of public transport information systems that have the ability to incorporate the travel habits of citizens has great potential to increase the value of shared modes of transportations. While transport systems are physically designed for an anonymous mass of travellers, single transport users have individual information needs which are strongly related to their everyday mobility characteristics and routine behaviour. In this regard, information technology provides vast opportunities to optimise the usage and accessibility of public transport services by means of incorporating data about how people behave in an urban transport scenario.

Based on the increasing availability of digital footprints about the transport behaviour of urban citizens, a number of data-driven studies have been conducted to improve the design of public transport services in various ways. In order to inform people about crowded stations and raise awareness about expected transport conditions, it has been shown that typical periods of overfilling can be extracted from electronic ticketing data [2]. Further, in order to understand the flows of city residents in an urban transportation network, a transport demand model has been refined to predict the number of trips between pairs of intra-urban train stations [6]. Also, it has been demonstrated that official information provided by transport providers about the duration of trips is less accurate than predictions derived from transport usage traces of individual travellers [5]. Based on data about urban travel demand and official route planning information as maintained by transport providers, it could also be shown that popular routes in an urban transport network suffer from effective accessible transport services for disabled people using wheelchairs [3].

While in previous studies the focus has been very much on collective information of transport usage, in this paper we extend on these studies by mining personal transport patterns which provide useful insights for the design of personalized travel information systems for individual travellers. Based on large-scale ticketing data provided by automated fare collection (AFC) system about urban bus rides in Lisbon, we uncover travel regularity patterns that govern peoples' access to public transport system. In order to inform travel information systems with knowledge about forthcoming transport needs, we formulate a prediction problem which is to classify whether a traveller will access public transport services on a future day or not. In order to perform the prediction, we propose several temporal features of travel behaviour as input to a classification algorithm and identify the feature combination with the highest predictive power. In our evaluation, we show that our approach can achieve a high prediction accuracy of 77%, and outperforms two alternative baseline approaches by more than 49% due to the discriminative power of the temporal features we have derived. Hence, our work contributes important concepts to the development of personalised transport information systems that are able to understand travellers' transport routines. This in turn will contribute to making public transport more attractive in the future.

## Dataset

In this paper, we are relying on a sample of automated fare collection (AFC) data which contains records of millions of bus trips conducted by citizens in Lisbon, Portugal. In contrast to traditional paper tickets, AFC systems are based on smart cards (e.g. RFID based), which are carried by passengers and swiped over on-board card readers that are installed in buses. Analogous to bank cards, smart cards are typically owned by single users, so that each time a traveller boards a bus, an entry is created in an electronic trip history that is associated with the card holder. The data sample on which our study is based spans a period from 1st of April to 31th of May 2010, resulting in almost 9 weeks of bus usage traces (61 days). Even though the data is personalized, only anonymous information is provided and and no further attributes (e.g. name, address, etc.) are revealed which might identify the user. For the purpose of our study, we rely on information about the time of when bus services have been accessed as recorded in the AFC sample. Formally, this information is encoded as tuples $< u, t > \in H$, where $H$ represent the user trip history, $u \in U$ is the individual traveller (as identified by his/her travel card) and $t \in T$ indicates the bus boarding time. In total, we obtained $|H|$ =24,257,353 bus trips taken by $|U| = 809,758$ travellers.

## Mining of Travel Patterns

We have mined patterns which are hidden in the user's travel histories to gain important insights about the predictability of a user's travel behaviour. First, we look at different profiles of bus usage regularities. Then, we analyse typical bus usage periodicities which underlie the users' travel behaviour. Finally, we discuss the emerging characteristics of weekday and weekend travel found in the data.

*Weekly Travel Profiles*

We are studying a user's weekly travel profiles in terms of the average number of distinct days per week on which a passenger takes part in public transportation. Note that for this study we are not interested in the total frequencies of how many trips are conducted, which may involve several bus rides (e.g. interchanges or possible return trips) on the same day. In contrast, we capture a user's travel habits as a binary relation and say that a bus passenger has been active on a travel day if at least one bus trip has been taken. In Figure 1 we plot a histogram of the bus ride activities for binned ranges of average days per week on which bus services have been accessed. In order to gain statistical meaningful results for the purpose of our analysis, we have used those individual travel histories for which the time span between the first and the last trip recorded in the histories exceeds seven days. As the figure shows, the distribution is unevenly split across the spectrum of possible travel activities with a mean of $2.92$ days per week. The largest peak relates to travellers accessing bus services on a single day per week. Another peak can be observed for more frequent travellers who use bus services on approx. 4.5 days per week. We conclude that there exist a wide range of different mobility needs which should be incorporated by the next generation of public transport information system. While regular travellers more likely would welcome continuous travel updates, there is a high risk to obstruct travellers having reduced travel footprints. For this purpose, a deeper study of travel regularity is required to inform the design of travel information system with knowledge about a user's individual transport needs.

*Transport Usage Periodicities*

In order to extend our previous analysis, we mined temporal usage patterns that underpin the users' bus

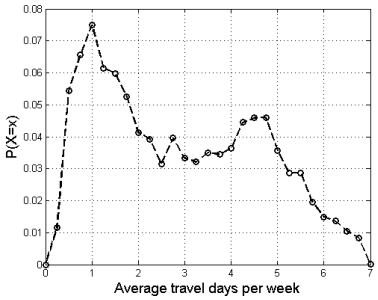


**Figure 1:** Pdf of weekly number of travel days per user

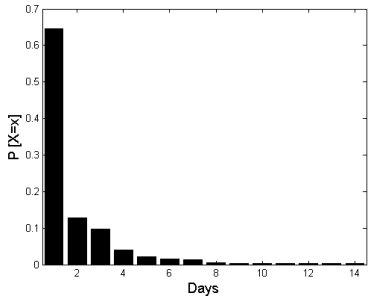


**Figure 2:** Pmf of travel periodicity across all users

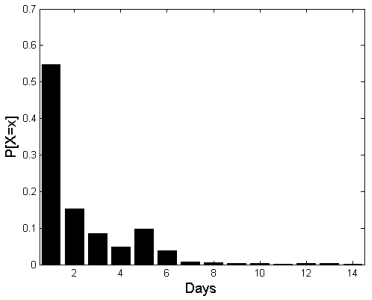


**Figure 3:** Pmf of travel stationarity across all users

transport behaviour. For this purpose, we studied typical transport usage periodicities, uncovering the typical durations which elapse between users accessed transport services. Figure 2 plots the observed periodicities in the data, measured as the temporal distance between consecutive days where a bus ride has been conducted. Across the user population, the usage periods are intermixed according to the shown occurrence probabilities. While the one day period makes up for the largest fraction, the probability of travelling at usage periods of increased lengths continually decreases. Note that the decomposition of the user's travel behaviour into various usage periods explains the variance in the travel activity profiles discussed before.

While the previous analysis reveals information about the occurrence of sequential travel decisions, it does not give insight about the extent of how long a travel pattern lasts once it occurs. In order to see if the bus usage is fragmented over the time, we analyse periods of continuous bus usage that can be observed in the data. We refer to this property as bus usage stationarity, since it sheds a light on the period of how long the user remains an active bus user over several days. It is determined by the maximum subsequence of consecutive travel days that can be extracted from the user's travel histories. The distribution of different stationarity periods is plotted in Figure 3. Frequently, people use bus services only on a single day, while on the following day no bus service is accessed. The probability of bus users taking trips over several consecutive days gradually decreases with a larger number of days. However, we can observe an outlier for the 5 day period which ranks third among all periods. This points to situations where people continuously use bus services on weekdays Monday to Friday to accommodate their mobility needs.

*Weekday/Weekend Travel Behaviour*
We analyse the extent to which the users' travel behaviour is concentrated on specific days of a week. For this purpose, we look at potential travel days in the users' travel histories embracing the days in between the first and the last recoded bus ride. Based on the information about the days when a bus ride was actually taken, we

then determine the probability that a bus service is accessed on a particular week day as shown in Figure 4. As can be seen, there is a clear difference in travel activities between weekdays (average probability of 0.5535) and weekends (average probability of 0.2399). Among weekdays, Mondays has the highest probability for a user taking a bus ride, followed by Thursday and then Wednesday. The lowest chance for a bus ride is associated with Tuesday, and then Friday follows as the day of the week with the second lowest travel probability. The likelihood for weekend travel activities is significantly lower, where traveller more often ride buses on Saturdays than Sundays as our analysis reveals.

## Transport Usage Prediction

The prediction problem which we address in the context of our work can be described as follows. Given is a user's $u$ history of past travel activities $H_u =< (t_1, b_1), (t_2, b_2), \ldots, (t_n, b_n) >$, where $b_i = \{travel, no\_travel\}$ encodes the users' travel choice in conducting a bus trip or not on date $t_i$. The history incrementally grows with each passed day, so that $(t_n, b_n)$ represents the travel activity of the most recent day $t_n$. Note that there is exactly one entry in a user's travel history for each day. Based on the historic travel information, we are interested in inferring whether it holds that $b_{n+1} = travel$ or $b_{n+1} = no\_travel$ for the upcoming day $t_{n+1}$.

Having accurate knowledge of a user's travel intent provides interesting opportunities for more proactive triggers of travel information which can make public transportation a more effective and enjoyable experience:

- People could receive alerts of time table changes before they would be affected by delays or unreliable services. Since such alerts are often experienced as disruptive and irritating if the user's intent is not matched, it is important to send notifications only when people would need them for the purpose of travel.
- In a similar fashion, users could be proactively be informed about their friend's transport patterns. Since revealing detailed mobility information might raise a privacy issue from the perspective of users, people might want only to share the information about their travel intents with their friends to allow for the exchange of social mobility predictions.
- In combination with 3rd party applications, predicted travel behaviour could provide useful input for value-added transport information services. For instance, in order to support travellers in handling ticket purchases more effectively, people could be asked whether credit should be added to their electronic travel card in case no money is left on it. This will allow people to travel with more certainty and avoid situations where they have to interrupt their trips in order to top-up their travel cards. In order to trigger this service in relevant stations, the users' requirements of accessing the transport system on an upcoming day needs to be predicted.

We formulate the prediction task as a binary classification problem using statistical methods from the domain of machine learning. In this paper, we rely on a Bayesian classification method to determine a user's travel activity based on features $X_i$ that describe different temporal aspects of the user's transport behaviour. Since we assume the variables $X_i$ to be conditionally independent to simplify the prediction problem, we apply the Naive Bayes classifier so that the prediction is computed as:

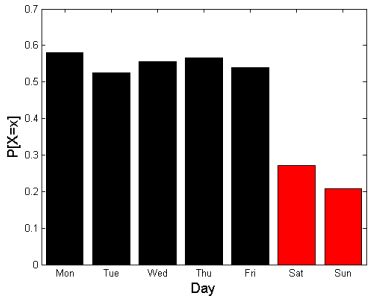


**Figure 4:** Pmf of travel activities over different week days

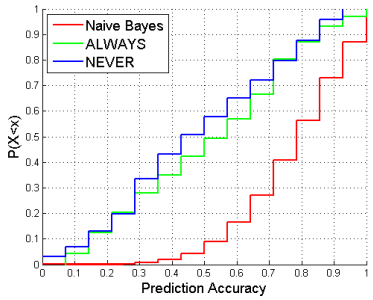


**Figure 5:** CDF of prediction accuracies for comparison with baseline predictors

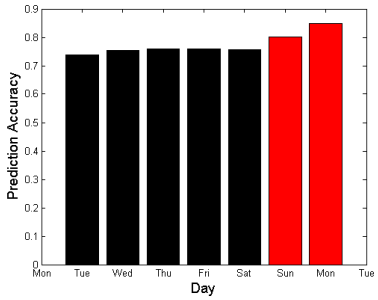


**Figure 6:** Average prediction accuracies for different week days

$$P(Y|X_1, \dots X_n) = \frac{\prod_{i=1}^{n} P(X_i|Y) \cdot P(Y)}{P(X_1, \dots X_n)}$$

The probabilities are maximum likelihood estimates based on the information contained in the user's travel histories which we use as training data for the classifier. The class variable $Y = \{travel, no\_travel\}$ has two distinct outcomes to differentiate among a user's travel and non-travel activity. By determining $P(Y = travel|X_1, \dots X_n)$ respectively $P(Y = no\_travel|X_1, \dots X_n)$, we choose the outcome with the higher probability as the prediction. In order to provide accurate predictions, the set of features $X_1, \dots X_n$ needs to be carefully chosen in in order to inform the classifier with relevant patterns about a user's travel intent.

*Features*

In the following, we propose and evaluate four different features $F = \{X_1, X_2, X_3, X_4\}$ as input to our classification algorithm which capture discriminative properties of a user's travel behaviour:

### Part of the Week

The first feature $X_1$ relates to coarse-grained information about the time when travelling might occur. We define it as

$$X_1 = week\_part(t_{n+1})$$

, where the feature can take values $week\_part(t_i) \in W = \{week\_day, week\_end\}$ to distinguish among different parts of the week. This is motivated by our previous finding, which showed cleared differences in transport usage on weekends and weekdays.

### Day of the Week

In addition, we also consider a more fine-grained resolution in information about the time of transport usage. The feature $X_2$ is given as

$$X_2 = day(t_{n+1})$$

, where day $day(t_{n+1}) \in D = \{Mon, Tue, \dots, Sun\}$ refers to the particular day of a week. This will allow us to

predict travel habits which are conditionally dependent on a specific week day.

### Travel Periodicity

Further, we include information about the typical time periods which underlie user's access to the transport system. For this purpose, we define the user's travel periodicity feature as

$$X_3 = t_{n+1} - t_r$$

, where $t_r = max_{1 \leq i \leq n}\{t_i | b_i = travel \wedge (t_i, b_i) \in H_u\}$ refers to the day where the most recent travel activity was observed. Note that the periods are measured in units of days $\in \{1, 2, \dots\}$ to account for the nature of our prediction problem. Consequently, we quantify the usage period as $t_{n+1} - t_r = n + 1 - r \in \mathbb{N}$ to identify the number of passed days since the most recent ride.

### Travel Stationarity

The last feature gives insight about the user's travel stationarity measure. We define it as

$$X_4 = t_r - t_s$$

, where $t_r$ is again the most recent day of travel, and $t_s$ is first day before $t_r$ in the history where no transport usage was observed. Formally, $t_s$ is given as $t_s = max_{1 \leq i \leq n}\{t_i | t_r > t_i \wedge b_i = no\_travel \wedge (t_i, b_i) \in H_u\}$. The time difference $t_r - t_s = r - s + 1$ hence gives insight into the number of consecutive days of accessed transport services in order to capture the extent to which the user's travel behaviour persists over time.

## Evaluation Results

In order to evaluate the performance of our predictor, we computed forecasts of future transport usage for the last 14 days of the data sample, and provided the preceding trip histories $H_u$ as initial training data to our classifier (approx. 80% - 20% split in time). With each additional day that passed, we also let the history $H_u$ grow and updated our predictor with the most recent travel information. In order to avoid any bias from possible cold-start problems, we pruned those users from our data who have taken no trips during the test period and $< 10$ times during the training period.

As our prediction task is an instance of a binary classification problem, the predictions result in true positive (TP), false positives (FP), false negative (FN) and true negatives (TN). Based on the number of occurrences of each class, we then measure prediction accuracy as

$$accuracy = \frac{TP + FN}{TP + FP + TN + FN}$$

which evaluates the fraction of correct predictions over all predictions made. The accuracy is measured independently on a per user basis, and the average is computed over all users to report on the prediction accuracy for the entire population.

**Feature Sets and Accuracy** In order to identify the most discriminative feature sub-set $F' \subseteq F$ with the highest prediction accuracy, we have evaluated our predictor over the power set $\mathcal{P}(F)$ of all features. Table 1 lists the achieved prediction accuracies for each possible feature combination $F' \in \mathcal{P}(F) \setminus \emptyset$. Among the feature sets that contain only a single feature item, $X_2$ (Day of the Week)

ranks best. For features sets of size two, $X_3$ (Travel Periodicity) and $X_2$ denotes the best combination, which further increases the prediction accuracy. For feature sets of size three, we can observe the best accuracy for the combination of $X_1$ (Part of the Week), $X2$ and $X_3$. This also denotes the global maximum, since no further accuracy gain can be achieved when using the combination of all features. We can conclude that there is the following ranking in discriminative power for the purpose of predicting travel activities when being combined in a single feature set: $X_2, X_3, X_1, X_4$ (ordered by high to low discriminative power)

**Comparison with Baseline Approaches** For the analysis of the effectiveness of our predictor, we compare it with two baseline approaches which cover extreme ends of a spectrum of optimistic and pessimistic forecasts. The most optimistic approach assumes that a user rides buses on every single day (ALWAYS). This approach is tailored towards the large fraction of active bus riders, who access bus services on many days per week. On the other side of the spectrum, we include a pessimistic approach which assumes that the user literary never rides buses (NEVER). This allows us to judge the extent to which travellers would have missed travel updates in case of potential bus rides. In Figure 5, we plot the CDF of the prediction accuracies achieved by the two baseline approaches as well as our predictor. Among the baseline approaches, ALWAYS performs better (average accuracy of 52%) than NEVER (average accuracy of 48%) since more travel days could be observed in the validation set than non-travel days. On average, our predictor outperforms ALWAYS by 49% and NEVER even by 61% in prediction accuracy. This demonstrates that knowledge about the user's temporal travel patterns can significantly improve the predictions.

**Accuracy on Different Week Days** We further analysed if the predictability of travel activities depends on particular periods of a week. For this purpose, we have determined the average accuracy of all predictions made on a specific week day. The results of our analysis are plotted in Figure 6.

| **Feature Set** | **Accuracy** |
|---|---|
| Part of the Week | 0.757 |
| Day of the Week | 0.763 |
| Travel Periodicity | 0.708 |
| Travel Stationarity | 0.655 |
| Day of the Week, Part of the Week | 0.763 |
| Day of the Week, Travel Periodicity | 0.769 |
| Day of the Week, Travel Stationarity | 0.723 |
| Part of the Week, Travel Periodicity | 0.766 |
| Part of the Week, Travel Stationarity | 0.719 |
| Travel Periodicity, Travel Stationarity | 0.675 |
| Day of the Week, Part of the Week, Travel Periodicity | 0.774 |
| Day of the Week, Part of the Week, Travel Stationarity | 0.730 |
| Day of the Week, Travel Periodicity, Travel Stationarity | 0.729 |
| Part of the Week, Travel Stationarity, Travel Stationarity | 0.728 |
| Day of the Week, Part of the Week, Travel Periodicity, Travel Stationarity | 0.738 |

**Table 1:** Prediction accuracies for different feature subsets

As can be seen, there is clear trend of improved predictability of weekend days compared to the prediction accuracy achieved for weekdays. Among all days, Sunday represents the most predictable day with an average prediction accuracy of 85%. Then, Saturdays follows as

the second best predictable day given an average prediction accuracy of 80%. Since less active public transport usage can be observed on weekends given our data sample, predicting non-travel activity in this time period results in highly accurate forecasts. For weekdays, there is a constant trend in predictability given an average prediction accuracy of 75% which holds approx. for all days. Due to the high variability in transport usage under the week where most of the travel activities are concentrated, a straightforward pattern of travel behaviour is more difficult to capture so that future transport usage becomes less predictable on average.

## Discussion and Future Work

Public transportation is deeply interwoven into the fast-paced lives of urban citizens to support their high demands for mobility. Digital footprints of urban citizens' usage of public transport systems provides the basis for transport information services which can increase the relevance of travel information delivered to individual riders by exploiting patterns residing in personal transport data. In this paper, we have shown that longitudinal data from automated fare collection (AFC) systems can be mined to uncover characteristic patterns of temporal regularities in accessing transport system (e.g. travel times, travel periodicities, travel stationarity). We then have devised a predictor which addresses the problem of classifying the user's behaviour into travel and non-travel days to sense a user's forthcoming travel intent. Based on our analysis, we could determine the best combination of predictive features which enables forecasts of future transport usage with a high accuracy of 77%.

In this paper, we have focused on a subset of the information only which is available in a typical AFC dataset to identify temporal aspects of transport usage. In future work, we will explore how further data about the usage of the transport systems in terms of accessed stops, services, and routes can be mined and predicted to unveil more complex patterns with finer-grained information about the transport habits of urban citizens. Even though we specifically evaluated bus riderhsip data, we believe that our pattern mining and prediction approach is generic enough to also predict mobility behaviours in relation to other transport means (e.g. trains, subways).

## References


[1] Camacho, T., Foth, M., and Rakotonirainy, A. Pervasive Technology and Public Transport: Opportunities Beyond Telematics. *IEEE Pervaisive* (2013), 18–25.

[2] Ceapa, I., Smith, C., and Capra, L. Avoiding the Crowds: Understanding Tube Station Congestion Patterns from Trip Data. In *Proc. of the ACM SIGKDD Intl. Workshop on Urban Computing* (2012).

[3] Ferrari, L., Berlingerio, M., Calabrese, F., and Curtis-Davidson, B. Measuring public-transport accessibility using pervasive mobility data. *IEEE Pervasive Computing 12* (2013), 26–33.

[4] Ferris, B., Watkins, K., and Borning, A. OneBusAway: Results from Providing Real-Time Arrival Information for Public Transit. In *Proc. of the 28th Intl. Conf. on Human Factors in Computing Systems (CHI '10)* (2010).

[5] Lathia, N., Smith, C., Froehlich, J., and Capra, L. Individuals among commuters: Building personalised transport information services from fare collection systems. *Elsevier Pervasive and Mobile Computing* (2012).

[6] Smith, C., Quercia, D., and Lici. Anti-gravity underground? In *Proc. of the Second Workshop on Pervasive Urban Applications (PURBA)* (2012).