Practitioner Paper

Predictability of Public Transport Usage: A Study of Bus Rides in Lisbon, Portugal

Stefan Foell, Santi Phithakkitnukoon, Gerd Kortuem, Marco Veloso, and Carlos Bento

***Abstract*—This paper presents a study of the predictability of bus usage based on massive bus ride data collected from Lisbon, Portugal. An understanding of public bus usage behavior is important for future development of personalized transport information systems that are equipped with proactive capabilities such as predictive travel recommender systems. In this study, we show that there exists a regularity in the bus usage and that daily bus rides can be predicted with a high degree of accuracy. In addition, we show that there are spatial and temporal factors that influence bus usage predictability. These influential factors include bus usage frequency, number of different bus lines and stops used, and time of rides.**

***Index Terms*—Public transport, data mining, smart card data, urban computing, transport usage patterns, travel prediction.**

I. Introduction

PUBLIC transport plays an important role in sustainable development of cities as it copes with the rising demand for mobility and helps reduce carbon emissions [1]. However, from a passenger point of view, public transport systems such as buses can be complex and difficult to use, lacking in freedom and flexibility offered by privately owned vehicles [2]. Due to recent advances in information and communication technology, novel opportunities have emerged for improving public transport systems to be more user friendly and passenger centric [3]. In particular, the wide adoption of mobile devices has provided public transport providers new channels for information dissemination [4]. Being able to provide travelers instant access to public transport data, e.g., real-time information of arrival times, incidents, or delays [5], has shown to create a positive impact on experience and satisfaction with public transport services [6].

While public transport information systems have the potential to further encourage the use and adoption of public transport services, with current systems, the responsibility is on the side of the travelers to actively inquire and filter information about their journeys [7]. As transport systems are subject to frequent delays and failures (e.g., schedule changes, reroutings, station closures, and overfillings), there is a high risk that relevant transport updates remain unnoticed. To provide more direct support and guidance, there has been an emerging idea of personalized transport information systems that proactively provide useful personalized transport information updates with no (or minimal) user interaction [8]. As these updates are automatically prepared for upcoming journeys, personalized information systems can significantly reduce the effort needed to make effective travel decisions. To provide personalized information recommendations, consequently, it requires an understanding and recognition of the individual transport usage patterns such as a traveler's preferred stops, routes, and travel times, and an ability to predict future transport usage [9].

The premise that the users' transport behavior can be sufficiently understood to estimate future travel needs is therefore key to the feasibility of personalized travel information systems. The extent to which this premise is fulfilled and transport users exhibit predictable behaviors currently remains unanswered. A number of studies have previously demonstrated the utility of smart card data to study bus passengers' travel behavior [10], [11]. The use of smart card data has been highlighted as an emergent and important component of planning and management for public transport services, given that it can offer finer grained spatial–temporal information on travel behavior [12], [13]. However, existing work is predominantly focused on performance metrics of the transport system itself (e.g., service accessibility [14], travel times [15], travel demand [16], and network planning [17]) and not on how individual users rely on public transport systems as part of their daily routines.

To this end, this paper explores the predictability of using the public bus system from the angle on individual riders. We exploit the availability of massive trip records from electronic ticketing systems, consisting of millions of rides by hundreds of thousands of bus users in Lisbon, Portugal. In contrast to traditional paper tickets, electronic ticketing systems are based on smart cards, which are carried by passengers and swiped over on-board card readers installed on the buses [13]. Analogous to bank cards, smart cards are owned by single users, so that each time the traveler boards a bus, an entry is created in an electronic trip history that is associated with the card holder. Mining these data of bus transport usage allows us to analyze the extent to which the transport behavior of individual bus riders is predictable.

To assess the feasibility of personalized transport information systems, we specifically present algorithms to predict bus usage of individual riders with respect to the bus lines and bus stops used in the next days. The results of our predictability analysis suggest that proactive personalized transport information systems are indeed feasible for a large population of bus riders, and thus, information needs can be predicted with decent accuracy. Finally, we discover characteristic features that describe predictable riders. In particular, we show that if riders travel close to peak times and travel scope is limited in relation to travel demand, high predictability is better guaranteed. Based on the insights into factors that determine predictable bus ride patterns, we seek to provide transport authorities useful information for increasing the travelers' information awareness and further improve their satisfaction with public transport systems.

Manuscript received March 12, 2014; revised July 28, 2014, January 31, 2015, and April 2, 2015; accepted April 9, 2015. Date of publication May 19, 2015; date of current version September 25, 2015. The Associate Editor for this paper was D.-H. Lee. *(Corresponding authors: Stefan Foell and Santi Phithakkitnukoon.)*

S. Foell and G. Kortuem are with the Department of Computing and Communications, The Open University, MK7 6AA Milton Keynes, U.K. (e-mail: stefan.foell@open.ac.uk; gerd.kortuem@open.ac.uk).

S. Phithakkitnukoon is with the Excellence Center in Infrastructure Technology and Transportation Engineering (ExCITE) and with the Department of Computer Engineering, Chiang Mai University, Chiang Mai 50200, Thailand (e-mail: santi@eng.cmu.ac.th).

M. Veloso and C. Bento are with the Center for Informatics and Systems, University of Coimbra, 3030-290 Coimbra, Portugal (e-mail: mveloso@dei.uc.pt; bento@dei.uc.pt).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2015.2425533

1524-9050 © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

II. Related Work

With the integration of pervasive sensors into public transport systems, an unprecedented amount of digital data has become available to analyze public transport systems as they are operated in the real world. In particular, smart card data, which provide direct access to histories of public transport journeys, have proven to be an invaluable source of information for optimizing public transport services [13]. For instance, Ceapa *et al.* [18] exploited smart card data to predict spatiotemporal events of overcrowding at London underground stations. Based on travel flows encoded in smart card records, Smith *et al.* [16] built a gravity model that explains the variance in travel demands between two underground stations in London. To study the accessibility of the London underground system for people with disabilities, Ferrari *et al.* [14] mined journey planning information and transport usage data.

Historically, data mining in the area of public transport systems has primarily focused on analyzing the travel demand of an aggregate mass of travelers. Recently, the focus of data-mining-based studies has expanded to improve the understanding of transport usage patterns associated with individual users. For instance, Lathia *et al.* [9] demonstrated that information from travel histories can be used to derive travel time estimates for individual riders that are more accurate than those provided by official schedules. Moreover, Lathia *et al.* [19] proposed a ticket recommendation system that helps travelers in choosing among various tickets, e.g., weekly or monthly travel cards, those that match best their travel needs. Foell *et al.* [20] developed a machine learning approach to predict travel intentions of riders. Based on features that characterize temporal usage patterns, prediction is made on whether or not the user will be an active rider on a future day.

As bus systems create vast route networks in cities and are among the public transport systems that are most difficult to maintain and use, the development of public transport information systems for bus riders has gained much attention over the recent years. For instance, Bejan *et al.* [15] developed an approach to exploit bus probe data for accurately analayzing journey times experienced by road users. Ma *et al.* proposed a trip chain model to identify and combine a series of bus rides into an end-to-end journey [21]. Mobile transport applications such as OneBusWay [6], Tiramisu [4], or PATH2GO [22] give smartphone users access to bus travel information from virtually anywhere. However, novel personalization concepts that are based on an understanding of transport usage patterns are not incorporated into state-of-the-art bus transport applications.

In this paper, we extend the previous studies of transport usage by analyzing specific aspects of individual bus rides. In particular, we investigate the predictability of daily bus usage of individual riders, considering both bus stop and bus line access patterns, and discuss the variation in riders' predictability governed by different characteristic features.

III. Data Sets

In this study, we used data that contain bus usage information from one of the largest bus operators in Lisbon, Portugal. The data span a period from April 1 to May 31, 2010, consisting of nearly nine weeks of bus usage traces (61 days). There are two sets of data; the first data set (A) is the Automated Fare Collection (AFC) data, and the second data set (B) is the Automated Vehicle Location (AVL) data. Data set A provides the bus boarding history of passengers, which are identified by anonymous IDs of their smart cards. Data set B is a bus probe data that contain entries of bus arrival times at each bus stop along the bus routes, where a unique ID is assigned to each bus, bus stop, and bus line. To safeguard personal privacy, individual information was anonymized by the bus operator before leaving their storage facilities and was identified with an anonymous ID (hash code). Therefore, no personal information is exposed in this study.

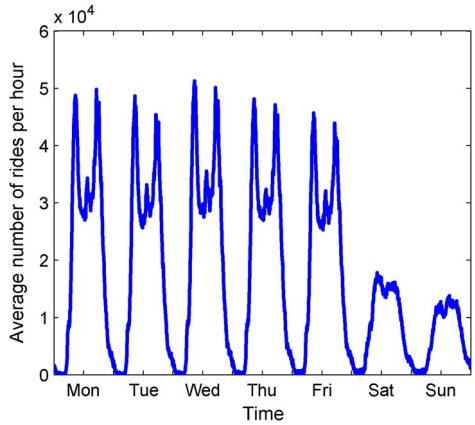

Fig. 1. Temporal distribution of average ridership demand (Mon–Sun) with significant spikes of high volume ride activity on weekdays as well as lower and more uniformly dsitributed ridership on weekends.

For the purpose of our study, we combined both data sets into bus usage histories that compose spatiotemporal information about a user's bus rides. Due to some inconsistency in timestamp recorded between the two data sets (as they were separately collected using different machines), we needed to clean the data by aligning the boarding times to the respective users and bus stops.

As a result, we obtained a cleaned data set of complete individual bus ride information. Formally, the data set consists of bus rides $\langle u, t, s, l\rangle \in H$, where H represents the entire ride history, $u \in U$ is the individual rider, $t \in T$ denotes the bus boarding time, $s \in S$ is the boarding bus stop, and $l \in L$ is the bus line taken by the user. In total, we obtained $|H| = 24{,}257{,}353$ bus rides taken by $|U| = 809{,}758$ users over the observation period. A total of $|S| = 2110$ distinct bus stops and $|L| = 93$ distinct lines were recorded. For each individual bus user u, H_u denotes the user's bus ride history, S_u is the set of visited bus stops, and L_u is the set of bus lines used by u. Fig. 1 shows a weekly distribution of the total average number of bus rides per hour. It is observed that buses are used mostly on weekdays. On weekdays, 21% of the usage are in the morning, between 7:30 A.M. and 10 A.M., whereas the other usage peak (also around 21%) is in the evening, between 4:30 P.M. and 7 P.M. Bus usage generally does not fluctuate on weekends.

In terms of bus travel demand, Fig. 2 shows the probability distribution of the number of rides per day. On average, 0.61 rides per day are taken by the users, which corresponds to 4.4 bus rides per week. It is notable that the majority of users (78%) rides a bus less than once per day on average. In addition, Fig. 2 features the distribution of the number of bus rides on active travel days (when bus usage is observed). We can see that, often, more than one bus trip is involved over the course of a travel day.

Similarly, Fig. 3 shows the probability distribution of the bus line and bus stop usage. As expected, bus journeys involve a higher number of distinct stops than lines. We observe that, on average, bus riders leave from 1.93 distinct stops per travel day, whereas they use 1.55 distinct lines. Intuitively, the same departure stop is rarely used twice in the same day, whereas trips with the same bus line are common, e.g., for commuting. On most days (52%), the bus riders take only one bus line, whereas, predominantly, two distinct stops (35%) are used to access a bus service followed by one stop (33%).

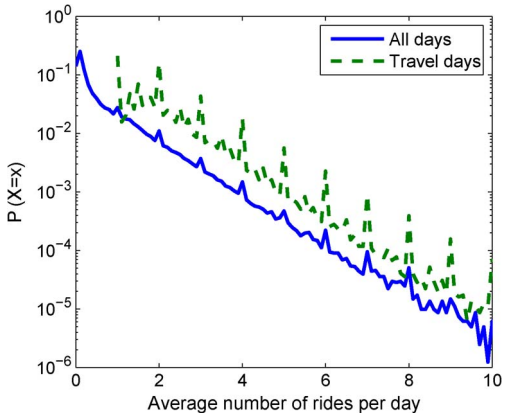

Fig. 2. Probability distribution of individual ridership demand. Two measures of ridership are shown: average number of rides per day (all days) and average number of rides per day when buses are actually used (travel days).

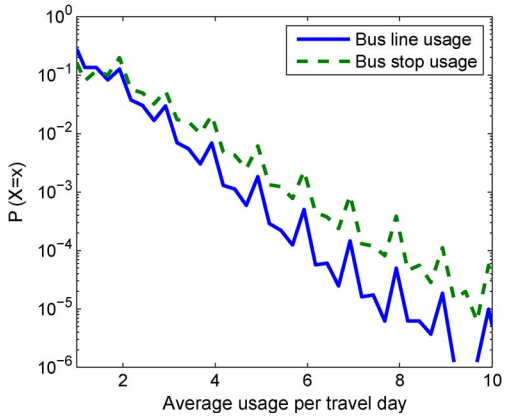

Fig. 3. Probability distribution of a rider's average daily usage, differentiating between stops and lines used. The distribution of bus stop usage is more skewed than bus line usage.

IV. Predictability of Bus Riders

To facilitate the design of future personalized transport information systems with predictive capabilities, we are interested in 1) the extent to which the user riders are predictable and 2) the classification of bus riders according to their predictability. In the following, we first present different prediction algorithms for next-day bus usage and measure corresponding prediction accuracy and then identify characteristic features in bus usage behavior, which are indicative of the rider's predictability.

A. Prediction Problem

In this work, we set out to predict bus ride behavior over the entire day of a week. Knowledge of riding patterns associated with single days gives transport information systems a useful horizon for planning, as potential interchanges or return trips by riders can be realized and recommended.

We formulate the prediction problem as follows: Given a particular user u and his/her bus ride history H_u on the day $d_t \in D$, the goal is to predict all bus lines $L_u(d_{t+1})$ and stops $S_u(d_{t+1})$ used in the next day $d_{t+1} \in D$. In the following, we present four different predictors that are designed to incorporate temporal features of travel decisions. The features exploited in our prediction are motivated by our previous work in characterizing bus usage patterns [20].

B. Prediction Algorithms

The following are prediction algorithms that take the rider's bus ride history H_u as input and make a prediction of bus lines $L_u(d_{t+1})$ and bus stops $S_u(d_{t+1})$. Subsequently, we give a description of bus line usage prediction of each algorithm. The described algorithms can be applied for bus stop usage prediction exactly the same way.

1) Continuation Predictor (CP): This predictor is based on the idea that travel behavior is characterized by a high degree of stationarity. Therefore, the assumption is that the user tends to behave similarly in the following day as before. Therefore, the predictor considers bus lines taken most recently. The degree of recency is determined by the parameter $r > 0$, which defines a sliding window centered on the current day $d_t \in D$. The prediction is thus defined as

$$CP_r(d_{t+1}, H_u) = \{l_i \in L| < t_i, s_i, l_i > \in H_u \wedge \\ \wedge t_i \in [d_{t-r-1}, d_t]\} \quad (1)$$

where $[d_{t-r-1}, d_t]$ denotes the interval spanning the previous r days prior to the current day d_t. By means of the recency parameter, the predictor can be configured with different time windows to incorporate different levels of recency of the past behavioral information.

2) Weekday/Weekend Predictor (WP): An alternative approach is to construct a predictor that is able to deal with discontinuity in travel. Motivated by our previous work [20] that identifies differences in the weekday/weekend travel behaviors, the WP only considers either weekday or weekend histories for a prediction for a weekday and weekend, respectively. For instance, if a prediction is made for a Saturday, only the weekend travel history is considered. The prediction is defined as

$$WP_r(d_{t+1}, H_u) = \{l_i \in L| < t_i, s_i, l_i > \in H_u \wedge w_{end}(t_i) \\ = w_{end}(d_{t+1}) \wedge t_i \in [d_{t-7*r}, d_t]\} \quad (2)$$

where $w_{end}(t)$ is a determiner if a given date t is a weekend. The degree of recency of history data is determined by the variable r. For example, WP_1 makes a prediction based on the bus rides from the last weekday/weekend, whereas WP_∞ considers the entire history.

3) Same-Day Predictor (SP): We have previously shown that not only between weekdays and weekends but also among different days of a week travel habits tend to differ [20]. Therefore, the SP makes a prediction based on a travel history of a specific day of the week according to the predicting day. The SP is defined as

$$SP_r(d_{t+1}, H_u) = \{l_i \in L| < t_i, s_i, l_i > \in H_u \wedge day(t_i) \\ = d_{t+1} \wedge t_i \in [d_{t-7*r}, d_t]\} \quad (3)$$

where $day(t)$ is history data of a day of the week. In comparison with the WP, the SP is more selective in terms of the considered bus ride information.

4) Periodicity Predictor (PP): This predictor introduces the ability to learn and adapt to different periodicities of travel behavior. The CP assumes a constant daily periodicity, whereas both the WP and SP predictors base their predictions on fixed weekly periods. On an individual basis, the range of underlying periodicity varies considerably.

The PP is designed to incorporate the time period in which a specific bus line is taken. The rider's usage period of the bus line $l \in L$ on day $d_{t+1} \in D$ is defined as

$$p_{u,d_{t+1}}(l) = median\left(\{\Delta_{int_{u,l}}\}\right) \quad (4)$$

where $\{\Delta_{int_{u,l},l}\}$ denotes the set of intertrip times from all past rides with l. To ensure that the selected median is not biased by repeated daily rides on the same bus line (e.g., for return trips), only intertrip times greater than one day are included in this set. Based on these periods, we determine those lines that may reoccur on the predicting day $d_{t+1} \in D$. Therefore, let $lr(l) = \max\{t_i \in T | < t_i, s_i, l > \in H_u\}$ be the time of the user's u last ride with the bus line $l \in L$. Then, the prediction can be defined as

$$PP(d_{t+1}, H_u) = \{l_i \in L | < t_i, s_i, l_i > \in H_u \wedge \exists k \in N : \left(lr(l_i) + k \cdot p_{u,d_{t+1}}(l_i)\right) \in d_{t+1}\} \quad (5)$$

where $lr(l_i)$ denotes the time of the last ride with bus line $l_i \in L$ and the associated usage period $p_{u,r}(l_i)$ to anticipate if the next ride is about to take place on $d_{t+1} \in D$. If the projected periodic occurrence of the next boarding falls within the predicting day, then the bus line is added to $PP(d_{t+1})$. Consequently, instead of predicting the same bus lines with a fixed period (i.e., from the last week), the prediction is based on learned trip periodicities. Please note that for determining the usage period, a sliding window (r) could also be applied, and we leave this for future investigation.

C. Analysis of the Predictors

To evaluate the performance of each predictor, we separate our data into training and test sets. The test set contains the last two weeks of the bus usage, whereas the training set includes the rest. Predictions are made for each user $u \in U$ for bus lines $L_u(d_{t+1})$ and bus stops $S_u(d_{t+1})$ potentially taken in the next day $d_{t+1} \in D$. A prediction is made every day in the test set period, and the training set (H_u) consequently grows as a new prediction is continuously being made. Note that only bus riders with at least one ride in both training and test sets are considered in our analysis. As a result, we are left with a total of 380 197 riders for the predictability analysis.

The F-score is an effective metric for set-based prediction problems [23], and it is used to measure the prediction accuracy here. For each rider, we compute the average F-score achieved over all days where bus rides have been observed. Formally, the F-score is defined as

$$F\text{-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \quad (6)$$

It denotes the harmonic mean of precision and recall. Hence, the F-score reflects the tradeoff between false positives and false negatives incurred by a prediction. A higher F-score implies better prediction. In a case where both precision and recall are zero, the F-score is assigned to be zero here to alleviate the division-by-zero problem.

Fig. 4 shows the prediction accuracies achieved by the predictors. In our evaluation, we have included a baseline approach (ALL), which basically bases its prediction on all history data. Hence, the ALL predictor is essentially CP_∞. For both prediction scenarios (i.e., bus line and bus stop), we can observe similarity across the different predictors. The SP_∞ predictor achieves highest prediction accuracy among other approaches. This suggests that day of the week is a relevant discriminator for predicting bus usage. As the PP does not perform well, we can say that knowledge of bus usage periodicities may not be relevant. We speculate that at least two rides need to be observed to build up a usage periodicity that could reduce fault learning. Other predictors that work better than the baseline approach are CP_7, WP_1, and WP_∞. However, there is a notable gap in the prediction accuracy compared with the best predictor SP_∞.

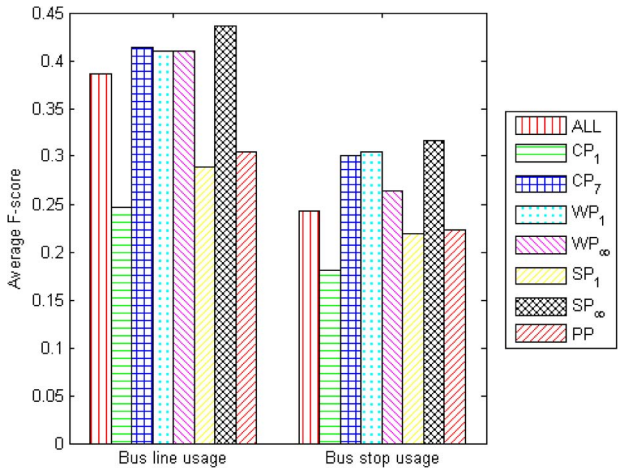

Fig. 4. Bar plot of the F-score achieved by all predictors. The left bars show the average F-scores obtained for bus line usage forecasts, whereas the right bars feature the average F-scores for bus stop usage prediction.

Compared with bus line prediction, bus stop prediction appears to be less accurate. This is in line with our previous study [20] that revealed a higher degree of variability in bus stop usage.

Although the ALL predictor uses the entire ride history for prediction, the recall (84% for bus line usage and 72% for stop usage) demonstrates that not all relevant transport decisions are captured in the user's ride history, and occasionally, new bus lines/stops (never observed before) are used. To forecast rides of new bus routes, different approaches in prediction need to be explored, which is part of our future work.

D. Predictability of Bus Users

Predictability of the bus users is important as it provides a preliminary indication for potential deployment of predictive capabilities for the next generation of intelligent public transport systems. The overall cumulative distribution function (cdf) of the bus users' F-scores is shown in Fig. 5. It reflects on the potential impact of predictive capabilities, e.g., proactive transport notifications and recommendations, both in terms of precision and recall. We first examine the top 33% users according to the F-score, which constitute a targeted rider group that may benefit from future predictive travel information systems. We find that based on these riders, daily bus stop usage can be accurately predicted with 68% precision and 81% recall. This means that two out of three predictions are correct, on average, whereas the predictions cover a large fraction of all stops visited. Even better predictable is the bus line usage with the precision of 84% and the recall of 91%. When all riders are considered, the predictability drops expectedly with the precision of 52% and the recall of 72% for the bus line usage and the precision of 39% and the recall of 56% for the bus stop prediction.

E. Influential Factors on Predictability

The last section explores the predictability of the bus riders, i.e., how predictable bus usage is. Here, we extend our analysis to the factors that tend to influence the predictability of the bus riders. Fig. 6 shows the variation in predictability captured by the F-score with respect to the user's ridership demand f_u (defined as the average number of rides taken per day). The predictability rises and peaks at 1.35 rides per day, representing the maximum predictability across different demand levels. In other words, users who ride a bus 1.35 times a day, on average, are the most predictable. This seems to suggest that people

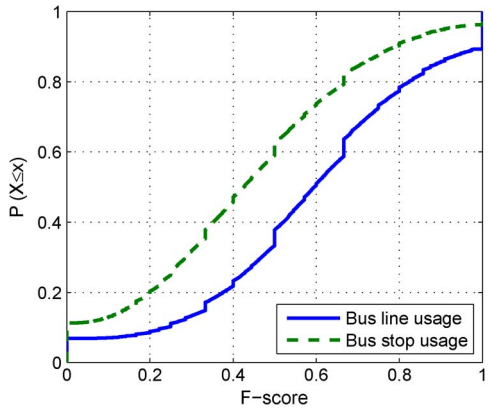

Fig. 5. CDF of the F-score values among all riders for bus line and bus stop usage predictability. Each rider is associated with the highest F-score that has been achieved among all predictors.

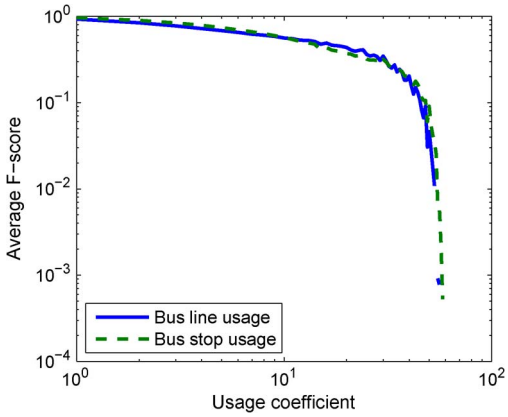

Fig. 7. Predictability of riders' daily bus usage as a function of usage concentration, i.e., the ratio of distinct stops/lines used and average number of rides per day.

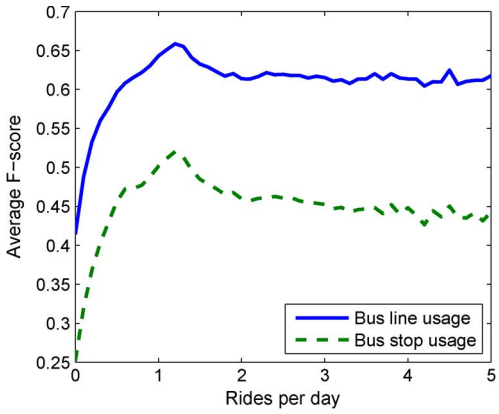

Fig. 6. Predictability of riders' daily bus usage as a function of ridership demand.

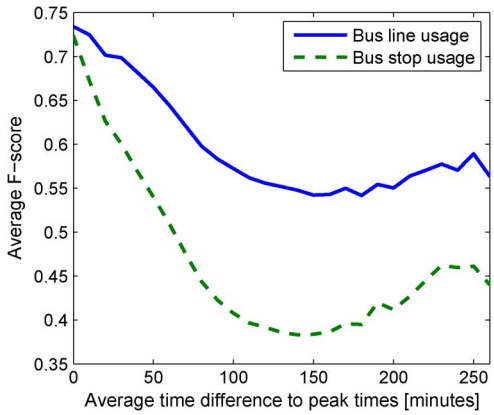

Fig. 8. Predictability of riders' daily bus usage as a function of the average minimum time difference from the morning and evening travel peaks.

who ride buses to commute to work on weekdays (twice a day, one for each direction, i.e., home–workplace and workplace–home) are the majority of these predictable users. Moreover, it can also be observed that the predictability is more or less constant for those with at least one ride a day, on average. Furthermore, the bus line is more predictable than bus stop.

In addition to bus ride demand, we introduce the notion of bus usage concentration in our analysis. For bus line usage, usage concentration is defined as the ratio of the number of distinct lines used and the ride demand ($|L_u|/f_u$). Similarly for bus stop usage, it is defined as the ratio of the number of distinct stops used and the user's ride demand ($|S_u|/f_u$). Fig. 7 shows the predictability that varies with usage concentration, which tends to follow the power law with exponential cutoff. This result intuitively suggests that frequent riders with smaller numbers of used bus lines and stops are much more predictable.

Finally, we examine the temporal bus usage patterns. Previously, it has been shown that bus usage has two peak times: one is in the morning centered around $t_m = 8:45$ am, and the other is in the evening centered around $t_e = 5:15$ pm. These usage peaks suggest that there is a salient rhythm or pattern in travel behavior (e.g., home–work commuting). A question arises from this context—if riders who consistently adhere to peak times are more predictable. Hence, for any ride taken at time t_r, we compute $\Delta t = \min(|t_r - t_m|, |t_r - t_e|)$ to determine the time difference between the boarding time of the ride and the closest peak time. Fig. 8 shows the predictability with respect to the users' average peak time difference. It can be observed that riders who travel closer to the peak times are more predictable. The predictability drops as the travel time becomes more distant from the peak time but starts to rise again toward the peak time difference of about 250 min, which is approximately near lunch time.

These results suggest that there are spatial and temporal patterns in bus usage behavior that play an important role as influential factors on predictability of bus usage.

V. Conclusion

In this paper, we have mined large-scale data collected by the AFC and AVL systems in Lisbon, Portugal to study the predictability of bus usage. In contrast to existing transport usage studies that are mostly concerned with aggregate travel characteristics, e.g., travel demand estimation, we have examined travel behavior patterns of individual bus riders. Understanding of bus user behavior is important for future development of the personalized transport information systems that can provide proactive assistance to the users. In this study, we have shown that daily bus usage can be predicted with a high degree of accuracy for a large proportion of the riders. In addition, we have uncovered that there are spatial and temporal factors that influence the predictability.

This work leverages on the availability of bus ride histories for predicting travelers' transport decisions. As part of our future work, we will continue to investigate on the predictability of bus usage as well as other public transport modes, e.g., train, taxi, and bike. In

particular, we will explore approaches to consider riders with limited travel histories such as tourists.

REFERENCES

[1] "Action plan on urban mobility, communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, and the Committee of the Region," Eur. Commiss., Brussels, Belgium, Sep. 2009.

[2] B. Gardner and C. Abraham, "What drives car use? A grounded theory analysis of commuters' reasons for driving," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 10, no. 3, pp. 187–200, May 2007.

[3] T. Camacho, M. Foth, and A. Rakotonirainy, "Pervasive technology and public transport: Opportunities beyond telematics," *IEEE Pervasive Comput.*, vol. 12, no. 1, pp. 18–25, Jan./Mar. 2013.

[4] J. Zimmerman *et al.*, "Field trial of tiramisu: Crowd-sourcing bus arrival times to spur co-design," in *Proc. CHI*, 2011, pp. 1677–1686.

[5] L. Weigang, W. Koendjbiharie, R. de M Juca, Y. Yamashita, and A. Maciver, "Algorithms for estimating bus arrival times using GPS data," in *Proc. IEEE 5th Int. Conf. Intell. Transp. Syst.*, 2002, pp. 868–873.

[6] B. Ferris, K. Watkins, and A. Borning, "OneBusAway: Results from providing real-time arrival information for public transit," in *Proc. 28th CHI*, 2010, pp. 1807–1816.

[7] N. Wilson, "The role of information technology in improving transit systems," in *Proc. Transportation@MIT Seminar*, 2009, pp. 1–35.

[8] B. Ferris, K. Watkins, and A. Borning, "OneBusAway: A transit traveler information system," in *Mobile Computing, Applications, and Services*. Berlin, Germany: Springer-Verlag, 2010, pp. 92–106.

[9] N. Lathia, C. Smith, J. Froehlich, and L. Capra, "Individuals among commuters: Building personalised transport information services from fare collection systems," *Pervasive Mobile Comput.*, vol. 9, no. 5, pp. 643–664, 2013.

[10] H. Nishiuchi, J. King, and T. Todoroki, "Spatial–temporal daily frequent trip pattern of public transport passengers using smart card data," *Int. J. Intell. Transp. Syst. Res.*, vol. 11, no. 1, pp. 1–10, Jan. 2013.

[11] M. Munizagaa and C. Palmab, "Estimation of a disaggregate multimodal public transport Origin Destination matrix from passive smartcard data from Santiago, Chile," *Transp. Res. C, Emerg. Technol.*, vol. 24, pp. 9–18, Oct. 2012.

[12] K. Chu and R. Chapleau, "Enriching archived smart card transaction data for transit demand modeling," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2063, pp. 63–72, 2008.

[13] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transp. Res.*, vol. 19, no. 4, pp. 557–568, Aug. 2011.

[14] L. Ferrari, M. Berlingerio, F. Calabrese, and B. Curtis-Davidson, "Measuring public-transport accessibility using pervasive mobility data," *IEEE Pervasive Comput.*, vol. 12, no. 1, pp. 26–33, Jan.–Mar. 2013.

[15] A. I. Bejan *et al.*, "Statistical modelling and analysis of sparse bus probe data in urban areas," in *Proc. 13th IEEE ITSC*, 2010, pp. 1256–1263.

[16] C. Smith, D. Quercia, and L. Capra, "Anti-gravity underground?" in *Proc. 2nd PURBA*, 2012, pp. 1–8.

[17] S. Tao, J. Corcoran, I. Mateo-Babiano, and D. Rohde, "Exploring bus rapid transit passenger travel behaviour using big data," *Appl. Geogr.*, vol. 53, pp. 90–104, Sep. 2014.

[18] I. Ceapa, C. Smith, and L. Capra, "Avoiding the crowds: Understanding tube station congestion patterns from trip data," in *Proc. ACM SIGKDD Int. Workshop Urban Comput.*, 2012, pp. 1–18.

[19] N. Lathia and L. Capra, "Mining mobility data to minimise travellers' spending on public transport," in *Proc. ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2011, pp. 1–9.

[20] S. Foell *et al.*, "Mining temporal patterns of transport behaviour for predicting future transport usage," in *Proc. 3rd PURBA*, 2013, pp. 1239–1248.

[21] X. Ma, Y.-J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders' travel patterns," *Transp. Res. C, Emerg. Technol.*, vol. 36, pp. 1–12, Nov. 2013.

[22] L. Zhang, S. D. Gupta, J.-Q. Li, K. Zhou, and W.-B. Zhang, "Path2Go: Context-aware services for mobile real-time multimodal traveler information," in *Proc. 14th IEEE ITSC*, 2011, pp. 174–179.

[23] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.