

**Regularity of public transport usage  
A case study of bus rides in Lisbon, Portugal**

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

**DOI**

[10.5038/2375-0901.19.4.10](https://doi.org/10.5038/2375-0901.19.4.10)

**Publication date**

2016

**Document Version**

Final published version

**Published in**

Journal of Public Transportation

**Citation (APA)**

Foell, S., Phithakkitnukoon, S., Veloso, M., Kortuem, G., & Bento, C. (2016). Regularity of public transport usage: A case study of bus rides in Lisbon, Portugal. *Journal of Public Transportation*, 19(4), 161-177. <https://doi.org/10.5038/2375-0901.19.4.10>

**Important note**

To cite this publication, please use the final published version (if applicable).  
Please check the document version above.

**Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights.  
We will remove access to the work immediately and investigate your claim.

---

# Regularity of Public Transport Usage: A Case Study of Bus Rides in Lisbon, Portugal

*Stefan Foell, The Open University, UK*

*Santi Phithakkitnukoon, Chiang Mai University, Thailand*

*Marco Veloso, Instituto Politécnico de Coimbra, Portugal*

*Gerd Kortuem, Delft University of Technology, The Netherlands*

*Carlos Bento, University of Coimbra, Portugal*

---

## Abstract

*This paper presents an analysis of regularity in public transport usage based on a large-scale bus transportation data of Lisbon, Portugal. By exploring the combined information from the bus boarding history of riders and bus arrivals at each bus stop, an analysis of individual bus usage was performed. Daily and weekly patterns were extracted, from which it was observed that a rider takes, on average, 2 trips, visits 1.93 distinct stops, and uses 1.55 distinct bus lines daily. Inter-trip time analysis revealed a daily cycle, and a study of the interaction between riders and bus infrastructure explored how usage was concentrated on particular bus lines and stops.*

**Keywords:** *Public transit; bus data mining; smart card data; urban computing; transport usage patterns.*

## Introduction

With fast-growing urbanization, collective transportation systems (such as buses, trains, and subway systems) become significantly important, as they enable continuous movement of a large quantity of inhabitants while also saving energy and reducing carbon emissions. In addition, public transportation information can provide useful data that reflect citizen needs and their daily patterns. Therefore, urban planners should pay close attention to these transportation modalities to learn from the information about a city's pulses of activities and improve the existing systems to meet passenger demands. If the public transportation infrastructure fails to evolve and adapt to user

behavior, the use of the public transportation may drop, and the increase of individual vehicles may occur, causing more traffic congestion, energy consumption, and pollution.

The development and adoption of new technologies such as smart card systems provide an exceptional opportunity to collect relevant information regarding the use of transportation systems. Several studies have taken advantage of the available information, most commonly to provide online information about bus scheduling or an estimation of waiting time according to timetable charts. Previous research has tackled data-centric problems in the public transport domain, but has focused predominantly on the performance metrics of the transport system itself, not on how individual users rely on public transport systems as part of their daily routines.

By combining large-scale data collected by Automated Fare Collection (AFC) and Automated Vehicle Location (AVL), we analyzed the regularity of bus usage, focusing on rider patterns and choices when using the public bus transport system. First, we uncovered typical daily and weekly transport behaviors according to the frequency of usage of bus lines and stops. Then, by quantifying the similarity of travel across different days of a week, we discovered characteristic temporal structures. Finally, by analyzing sequential travel decisions, we inferred typical periodicities of the bus ride behavior and identified temporal dependencies between bus boardings.

## **Related Work**

The development of novel public transport information systems has been the focus of active research over recent years. Mobile transport applications such as OneBusWay (Ferris et al. 2010b), Tiramisu (Zimmerman et al. 2011), PATH2GO (Zhang et al. 2011) or MOVE-ME (Cunha and Galvão 2014) have been proposed to give smartphone users access to travel information from virtually anywhere. Other applications provide the best travel information according to user location (Weigang et al. 2005) or using social networks to provide feedback and improve the user experience (Nunes et al. 2011). More recently, researchers have proposed personalized transport information that proactively recommends transport updates to individual travelers ahead of time and without requiring active user intervention (Ferris et al. 2010a).

However, personalization concepts that are based on an understanding of transport usage routines are not incorporated into these applications. With the unprecedented availability of large amounts of digital data produced by sensors integrated into public transport systems, novel opportunities have emerged to mine transport behavior patterns that could make these applications behave more intelligently.

Traditionally, data mining in the area of public transport systems focuses primarily on ridership demand estimation and optimization of public transport management and operation. For instance, Ceapa et al. (2012) used AFC data to estimate crowd levels at London Underground stations and predict events of overcrowding. To estimate intra-city travel flows, Smith et al. (2012) employed a gravity model to approximate the variance in travel demand between two underground stations in London. To study the accessibility of the London Underground system for persons with disabilities, Ferrari et

al. (2013) combined information from journey planning with a demand model gained from transport usage data. To establish a low-cost solution for congestion detection and traffic flow analysis, Bejan et al. (2010) leveraged on-bus probe data to analyze journey times experienced by road users. To improve reliability of public transportation Matias et al. (2010) studied the optimum number of schedules. The authors applied the Dynamic Time Warping distance with a k-means clustering and were able to identify different profiles between weekends and weekdays in non-scholar periods.

Recently, the focus of data mining studies has expanded to the analysis of individual transport usage. Instead of characterizing aggregate demand or travel flows, this direction of research seeks to improve the understanding of transport usage patterns linked with individual users. For instance, Lathia et al. (2012) demonstrated that travel histories can be exploited to improve journey planning information. By incorporating variances in transport behavior among individual users, travel times can be estimated that are more accurate than those provided by official schedules. Further, Lathia et al. (2011) proposed a ticket recommendation system to help travelers make the best ticket purchase decision. Analysis reveals that significant monetary savings potentially could be achieved when recommending tickets that match the user's travel needs. Foell et al. (2013) proposed a machine learning approach to predict travel intentions of riders. Based on features that characterize temporal usage pattern, a prediction was made if a user will be an active rider on a future day or not.

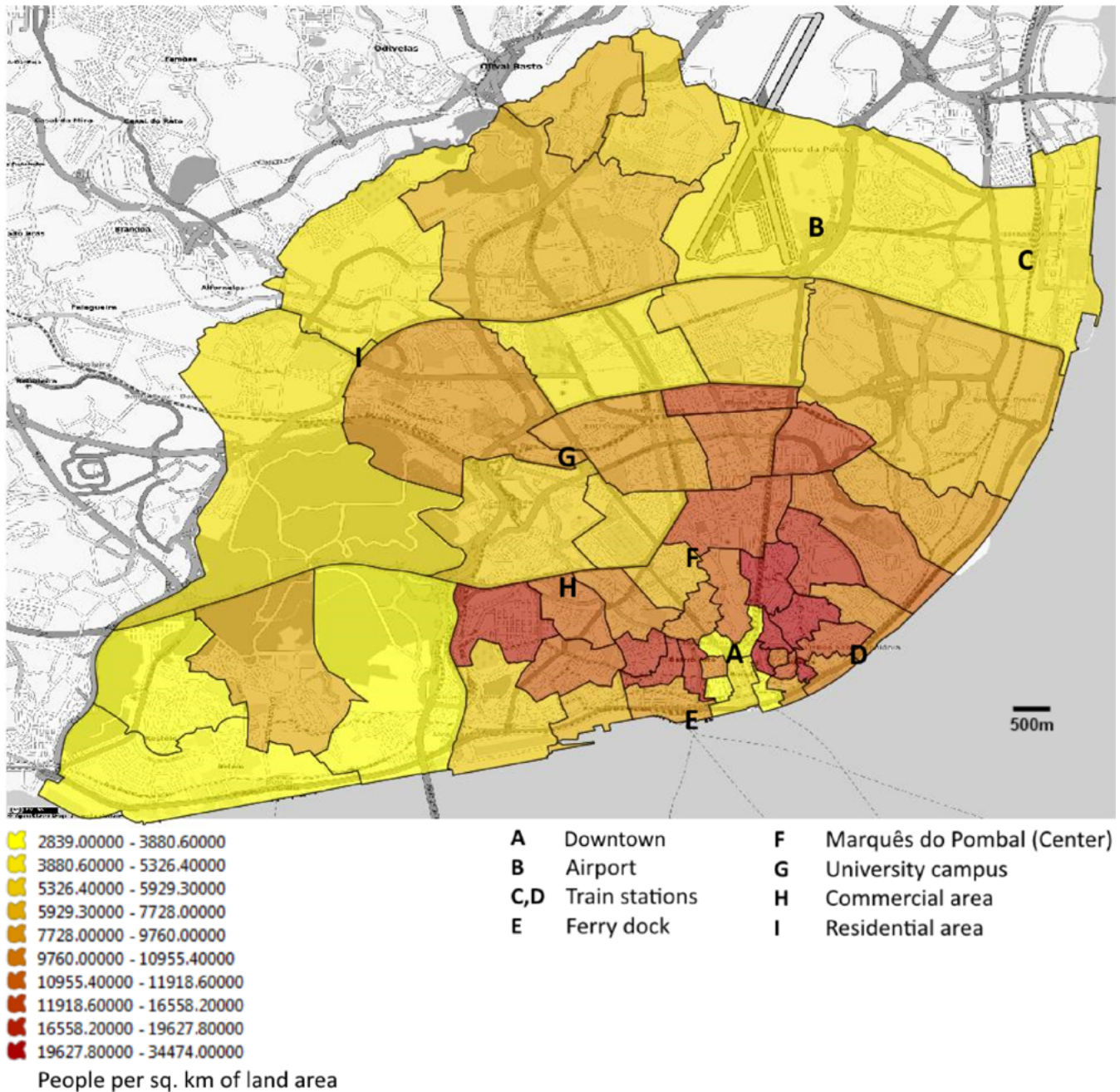
In this work, we expand previous works (Foell et al. 2015) by analyzing specific aspects of individual bus usage. In particular, we investigated behaviors of daily bus usage of individual riders, looking at both bus stop and bus line access patterns, uncovering pattern similarities on daily and weekly use and inter-trip time behaviors.

## **Dataset Description**

### ***Study Area***

Lisbon, Portugal, as of 2010, consisted of 53 parishes, an area of around 110 km<sup>2</sup>, and a population of 800,000 habitants, as represented in Figure 1.

The city's downtown is the central area, which includes the oldest and smallest parishes with the greatest population density (red); touristic, historic, and commercial areas; and the interface for several public transportation services (bus, subway, train, and ferry). Encircling the city center are residential areas surrounding business areas with lower population density (yellow). Major infrastructures (e.g., airport, industrial facilities) are located on the city's outskirts. The public transportation system consists on bus, subway, train, and ferry.



**FIGURE 1.** Lisbon municipality and population density

All transportation systems (train, subway, bus and taxi) provide station hubs in the city center, enabling a multimodal transportation system. However, the train system has routes only near the riverside, connecting the city to other districts. The main public transportation is bus and subway, with 235 million and 180 million passengers, respectively, in 2010,<sup>1</sup> using a radial route. These two transportation systems were

<sup>1</sup> INE, *Instituto Nacional de Estatística*, Statistics Portugal, <https://www.ine.pt/>.

developed independently and overlap in some routes. However, the subway system is limited to the city of Lisbon, and the bus system provides connections with the outskirts. Passengers usually use either the bus or subway system, not both, in the same commute.

### **Bus Data**

Large-scale data of bus transportation was provided by Carris,<sup>2</sup> the largest bus operator in Lisbon. The bus system comprises 2,328 bus stops and 105 bus lines, operated by 773 vehicles. Each bus line has vehicles moving in opposing directions, and 56.7% of bus stops are shared by different bus lines. From the top 5 bus stops shared by more than 10 bus lines, 4 are located downtown. Figure 2 shows the location of bus stops and the corresponding number of rides started at each stop, represented by the radius.

**FIGURE 2.**  
Lisbon map, showing  
popularity of bus stops



*Radius scaled according to number of rides that started at the stop.*

The data were collected from April 1–May 30, 2010, resulting in almost 9 weeks of bus usage tracking (61 days). For the purpose of the study, we investigated two different datasets, A and B, where A refers to a record of AFC data and B entails AVL data. Dataset A provides the bus boarding history of passengers identified by the IDs of their smart cards, without any personal information from the user. Dataset B supplies bus probe data that contain entries of recorded bus arrivals for each stop along the bus route, but no user-related travel data.

<sup>2</sup> Carris, <http://www.transporteslisboa.pt/>.

To analyze the transport behavior of bus riders, we combined both datasets into ride histories, which include spatio-temporal ridership information. Whereas dataset A gives an insight into the buses taken by riders, no information is given about the departure bus stops. Using the bus probe data from dataset B, we linked the start of each bus ride to the closest matching bus arrival. More precisely, to identify the correct departure stop, we looked for the smallest time difference between a rider's bus boarding event and the arrival of the same bus at any stop. As part of this procedure, we removed rides that could not be linked due to inconsistencies—i.e., for some boardings, no matching record of the same bus was present, the time gap was too large, or duplicate entries were observed in which smart cards were logged twice or more times upon bus boarding.

This process produced a cleaned dataset of complete bus ride information. Even though the data were associated with individual riders, only anonymous user data were provided. Formally, the dataset consisted 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,  $t \in T$  indicates the bus boarding time,  $s \in S$  is the ride's departure location (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 of two months. The rides started at  $|S| = 2110$  distinct bus stations and were taken with  $|L| = 96$  distinct bus lines. For each individual bus user  $u$ ,  $H_u$  denotes the user's bus ride history,  $S_u$  is the set of bus stops, and  $L_u$  is the set of lines used by  $u$ .

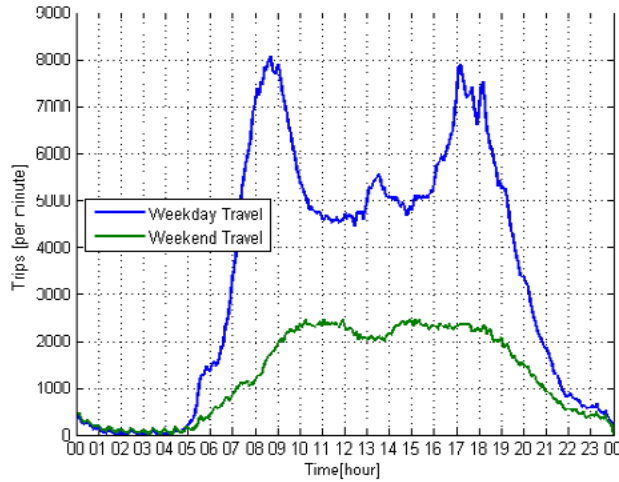
## Regularity of Bus Ridership

The main goal of this work was to explore patterns of bus ridership. This section analyzes different viewpoints in travel behaviors, from the rider perspective as well as the bus infrastructure. In the following sections, we explore the temporal distribution of travels, the usage of distinct bus lines and stops, how the bus infrastructure is used, the similarity of travel patterns on consecutive days, and rider behavior between trips.

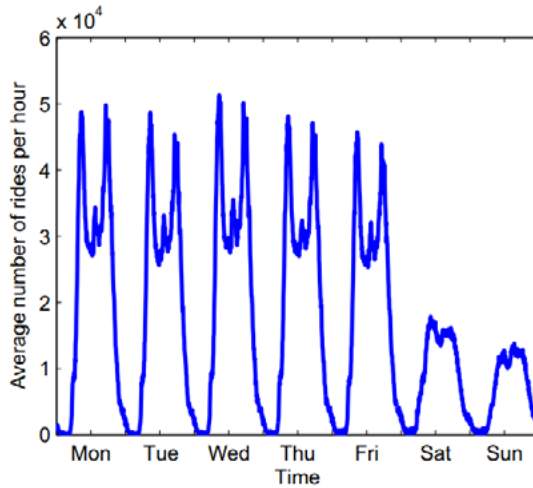
### *Daily and Weekly Distribution of Ridership*

The study began with the exploration of the average weekly distribution of bus rides, plotted in Figure 3 (daily) and Figure 4 (weekly). As expected, bus is used predominantly on weekdays. Daily activity is characterized by two peaks of travel movements (morning and evening), which correspond to inbound and outbound commuting on weekdays and a small increase of activity during lunch time. In the morning, 21% of travel activity takes place between 7:30 and 10:00 AM, and in the evening, 21% of all weekday travels are taken between 4:30 and 7:00 PM.

**FIGURE 3.**  
Daily distribution of average  
ridership demand



**FIGURE 4.**  
Weekly distribution of  
average ridership demand  
(Mon–Sun). Significant peaks  
of high volume ride activity  
on weekdays and lower and  
more uniformly-distributed  
ridership on weekends.



On weekends, the number of rides is more equally distributed across the days. After the morning increase, around 10:00 AM, the decrease of rides takes place only in the evening. However, a distinctive characteristic of the rides allows the differentiation of Saturdays from Sundays: for the former, the bulk of the rides takes place before 12:00 noon; for the latter, the highest demand falls into the afternoon hours. Moreover, when compared to weekdays, the increase in trips tends to start later in the day on weekends, from 10:00 AM onwards. Interestingly, the daily temporal pattern of bus ridership shares similar characteristics with other urban transportation system, i.e., taxi service (Phithakkitnukoon et al. 2010).

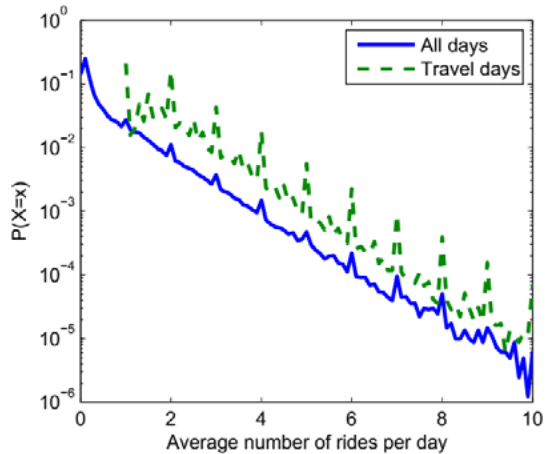
### ***Distribution of Individual Ridership***

Each individual passenger has a specific travel pattern. To determine those patterns, we computed the probability distribution of individual ridership demand (Figure 5). We used two measures for analyzing individual ridership demand:  $f_u$  is the average number of rides taken per day, including non-travel days (starting with the user's first ride), and  $f'_u$  is the average number of bus rides per actual travel day.



**FIGURE 5.**

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



We observed that 78% of passengers traveled by bus less than one time per day. On average,  $f_u$  is 0.61 rides per day, corresponding to 4.4 bus rides per week. This result was significantly affected by two groups of passengers: for 12.9% of riders, only one ride was recorded, and 12.8% of the users took two rides). Nonetheless, 50% of the most active riders had an average of 1.12 rides per day, whereas for the top 10%, 2.65 rides per day was observed. However, as noted, these results include non-travel days.

Exploring another scenario, we computed the number of bus rides per actual travel day ( $f'_u$ ). On average, each passenger took two daily trips, which is intuitive (i.e., commutes, traveling from home to work and vice-versa). As Figure 5 shows, the distribution of  $f'_u$  is characterized by visible peaks around integer (whole-number) frequencies. These patterns still hold for larger ride frequencies, even though the probability of occurrence decreases exponentially. As a consequence, we can conclude that bus ride demand is heterogeneously distributed across the population.

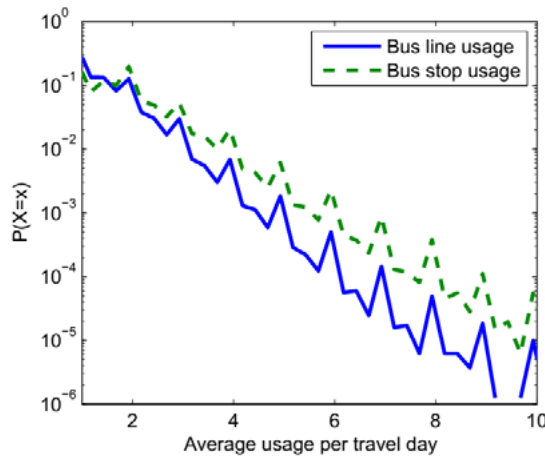
### **Bus Stop and Line Usage**

To investigate the adequacy of the bus service to passenger needs and general mobility patterns, we examined rider interactions with the bus infrastructure—more specifically, where passengers boarded the bus and what lines were taken. On average, passengers visited 1.93 distinct stops each day and used 1.55 distinct bus lines.

To determine individual mobility patterns, we computed the probability distribution of a passenger’s average daily usage (Figure 6). We observed that a larger fraction of passengers visited a higher number of bus stops than bus lines. Moreover, although passengers rarely visited the same bus stop on the same day, the same bus line was seen more often in a user’s daily ride history repeatedly, e.g., for taking return trips.

**FIGURE 6.**

Probability distribution of rider average daily usage, differentiating between stops (green) and lines (blue) used. Distribution of bus stop usage more skewed than bus line usage.



To further investigate the relationship of travel frequency and the usage of bus infrastructure (distinct bus stops and bus lines), we used a linear regression. For the case of bus stop visits, the data can be fitted with a linear equation  $s_u = 0.9449 * f'_u + 0.0115$ . Based on the fitted slope, we can conclude that there is almost a one-to-one correspondence between the number of bus rides and the number of distinct bus stops observed on the same day—i.e., every time a passenger takes a bus ride on the same day, a distinct bus stop is used. This relationship holds for all users, as most of the variation in observed bus stop visits can be explained by the travel activity ( $R^2 = 0.9944$ ).

In case of bus lines, the relationship can be fitted with a linear equation  $l_u = 0.6096 * f'_u + 0.3507$ , with more variation in the data ( $R^2 = 0.7656$ ) when compared with the bus stops visited. This can be explained by the fact that the decision whether a certain bus line is suitable depends on the origin and destination of a ride. The fitted slope implies that there is a 40% chance that a ride is taken with the bus line used before on the same day.

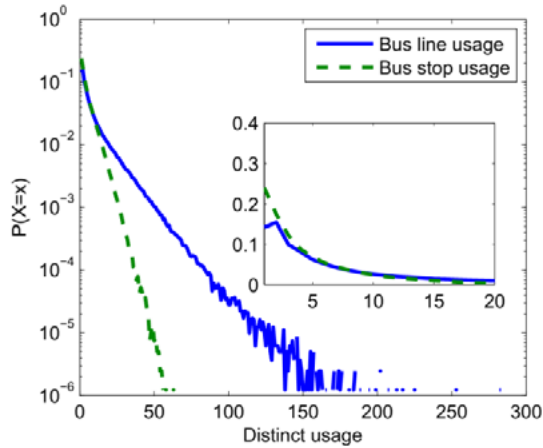
### Travel Scope

In the previous section, we analyzed the daily interaction between passengers and the bus infrastructure (bus stops and bus lines) to understand the daily patterns of the riders. In this section, we analyze the subsets of the transport infrastructure that are relevant for the rider’s mobility requirements, termed “travel scope.”

We are interested in exploring the quantity of bus lines and bus stops used by distinct passengers to identify limits in transportation activities of different users. Figure 7 shows the proportion of distinct bus lines ( $|L_u|$ ) and bus stops ( $|S_u|$ ) used by the rider. Most passengers use a small part of the bus infrastructure. In the case of bus lines, 70% of riders use only 1–5 lines, and 20% of riders use 5–8 lines (from a total of 2,110 possible bus lines). In the case of bus stops, the scope is larger and varied: 70% of all riders visit 1–9 stops, and 20% are seen at 10–23 different stops (from a total of 96 bus stops).

**FIGURE 7.**

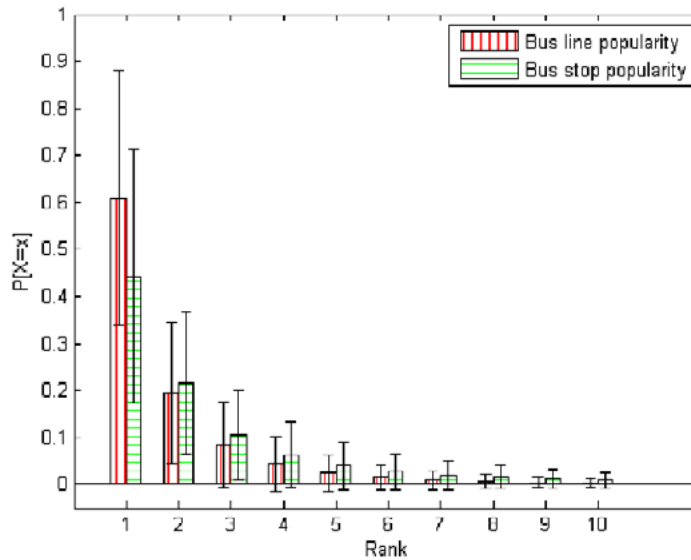
Characterization of scope of rider bus usage—probability distribution that certain number of distinct lines and stops used by any rider.



It is also important to understand if the full range of bus lines and stops are used equally or if some elements of that network are used only occasionally. To analyze if there is any skew in the usage, we computed a ranked distribution of the average popularity of the top 10 most frequently-used stops and lines in a user’s ride history, as depicted in Figure 8.

**FIGURE 8.**

Ranked average popularity of top 10 most frequently-used bus lines (red) and stops (green) in user ride history. For each rank, mean and standard deviation of popularity are shown.



To achieve this outcome, we first created, for each user  $u$ , an ordered vector  $r_u = [p_1, p_2, \dots, p_n]$  of usage probabilities  $p_i$ , that measure the fraction of rides associated with the user’s  $i$ -th most frequently-used bus stop/line, and then averaged the result across all users. As shown on 0, the usage popularity quickly drops after higher ranks. We can conclude that, on average, the use of the bus infrastructure is concentrated on a limited number of bus lines and stops. Whereas the most popular bus line is accessed with a probability of 0.61 on average, the second most popular line has a probability of only 0.19 (third rank 0.8, fourth rank 0.04). In contrast, for bus stops, the popularities decrease at a slower rate, which means that a larger number of different stops is relevant: the most popular bus stop is visited with a probability of 0.44 on average, and

the second ranked bus stop 0.2 (third rank 0.1, fourth rank 0.06). This is in line with our previous analysis, where we emphasized that more distinct bus stops are involved in rides than distinct lines, on average.

### Similarity of Travel across Week

Previous sections have shown an apparent temporal pattern. Therefore, we examined the similarity of passenger behavior on different days for which we used the cosine similarity to compare vectorized representations of a rider's bus usage on different days.

For the similarity coefficient, it should be relevant to consider which same bus stops and lines have been taken on different days, and their usage count should be in a comparable order on those days. We transformed the bus line usage patterns on a specific day in a vector space model (VSM). A similar procedure was applied to bus stops visits.

For given two days  $d_i, d_j \in D, d_i \neq d_j$  and the set of all used bus lines  $L_u(d_i, d_j) = L_u(d_i) \cup L_u(d_j)$  on these days, we derive a travel vector  $tr_{u,i,j}(d_i) = \langle n_1, n_2, \dots, n_{|L_u(d_i, d_j)|} \rangle$ , which encodes the number of rides of  $u$  taken with various lines. More precisely, the  $i$ -th entry,  $n_i \in \mathbb{N}, 1 \leq i \leq |L_u(d_i, d_j)|$ , of the travel vector denotes the number of rides taken with bus line  $l_i \in L_u$  on the selected day,  $d_i$ . For each pair of days  $d_i, d_j \in D, d_i \neq d_j$ , we can then compute the similarity, defined as:

$$C_u(d_i, d_j) = 1 - \frac{tr_{u,i,j}(d_i) \cdot tr_{u,i,j}(d_j)}{\|tr_{u,i,j}(d_i)\| \cdot \|tr_{u,i,j}(d_j)\|} \quad (1)$$

which corresponds to the cosine distance.

Note that we subtracted the cosine similarity (right) from 1 to produce values in the interval [0,1] to obtain a ranking in ascending order, where 1 represents the highest and 0 is the lowest similarity score. Based on individual users' similarity scores, we constructed a population-wide similarity matrix  $S_c$  of size  $|D| \times |D|$ . Each entry  $(i, j), i \neq j$ , of matrix  $S_m$ , stores the average similarity value:

$$S_c(i, j) = \frac{\sum_{u \in U_{i,j}} C_u(d_i, d_j)}{|U_{i,j}|} \quad (2)$$

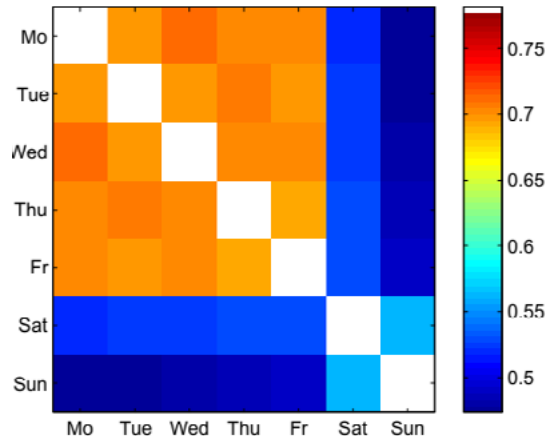
among the days  $d_i, d_j \in D, d_i \neq d_j$ , across all users  $u \in U_{i,j} \subset U$ , who have taken at least one ride on both days (note that  $S_c$  is symmetric across all day pairs, i.e.,  $S_c(i, j) = S_c(j, i)$ ).

The heatmap in Figure 9 displays the similarity matrix between different days of the week; the color denotes the intensity of the similarity. Clearly, two distinct groups arise, with stronger (weekdays) and weaker (weekends) travel similarities among their elements. On weekdays, travel behavior follows a regular pattern, explaining the similarity between days. The similarity of the Monday/Wednesday pair and the Tuesday/Thursday pair appear to be relatively high, suggesting a stronger link between these pairs, which are two days apart. In contrast, weekend bus rides are highly distinctive

from weekdays rides, especially on Sunday, which significantly deviates from the bus usage pattern observed on other days.

**FIGURE 9.**

Similarity matrix for weekly bus line usage. Similarity of ride behavior among different weekdays measured using cosine distance over vector of observed bus line boardings.

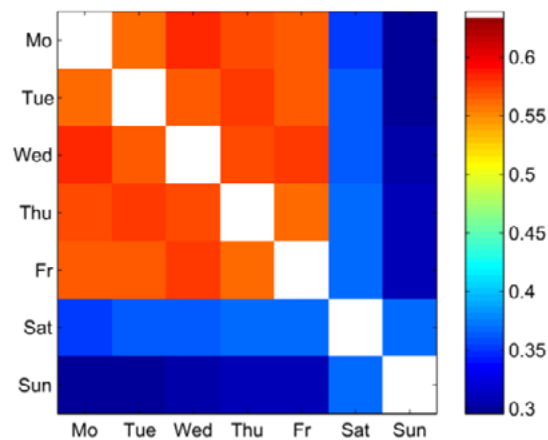


These observations are in line with empirical observations: usually, on weekdays, passengers tend to have a more rigorous schedule (e.g., work, school), which generates a more regular pattern. On the other hand, on weekends, passengers have a more diverse and less restrict schedule of activities (e.g., shopping, attending cultural or sporting events), creating a more random pattern.

A similar approach was applied to bus stop visits, and a similar outcome was obtained. The heatmap in Figure 10 displays the similarity matrix between different days of the week for bus stop visits. The similarity patterns follow the same trend observed for bus line usage. However, the absolute similarities values prove to be lower since bus stop visits exhibit a less certain usage signature, as shown previously.

**FIGURE 10.**

Similarity matrix for weekly bus stop visits. Analogous to 9, similarity scores represent values of cosine distance, but in this case applied to bus stop usage vectors.



To summarize the findings, we aggregated the computed cosine distances in Table 1. Days are grouped in different categories with the correspondent intra-group (all days, weekdays, weekends) and inter-group (between weekdays and weekends) average similarities.

**TABLE 1.**  
Similarities of Bus Usage over  
Different Parts of a Week

Day Subset	Bus Line Usage Similarity	Bus Stop Visit Similarity
All days	0.60	0.48
Weekdays only	0.70	0.57
Weekends only	0.56	0.37
Weekdays/weekends	0.50	0.33

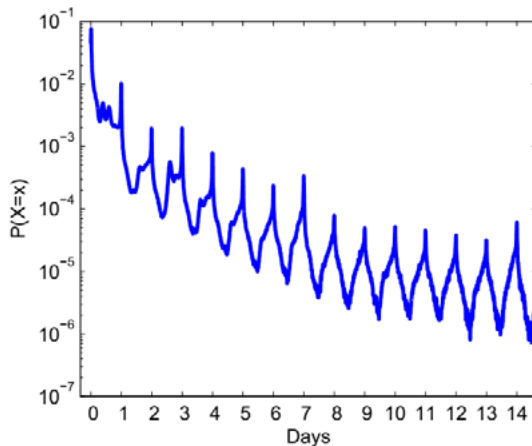
Values indicate average cosine distance

**Periodicity of Travel Behavior**

In addition to passenger patterns of using bus infrastructure, we explored their travel behavior between bus trips, i.e., the riders’ inter-trip times (elapsed time between two consecutive bus boardings). Note that this time period does not correspond to bus waiting times, but to a temporal measure, characterizing public transit access periods, and, hence, a measure for regularity of travel.

The distribution of the inter-trip times is shown on Figure 11 (from 23,447,595 consecutive buses, considering all individual rides) and indicates that inter-trip times are frequently short (12% of all observations fall within the interval  $t \leq 20$  minutes, and inter-trip times are within  $t \leq 30$  which account for 18%). This can be explained by the change between bus lines in a single journey. Previously, we observed that, on average, each passenger used 1.55 distinct bus lines daily. Since each passenger visits, on average, 1.93 distinct stops each day, this can also be an indication that the observed inter-trip time could include additional waiting time and short walk to a different bus stop.

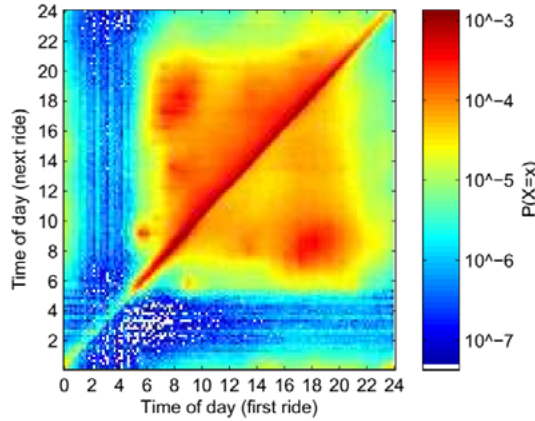
**FIGURE 11.**  
PDF of inter-trip times across  
all trips. Distribution reveals  
high number of interchanges  
with short inter-trip times and  
characteristic daily cycles of  
consecutive bus boardings



Considering the inter-trip times taking place on the same day, we can observe two (daily) peaks, at 9.5 hours and 14.5 hours after the last trip. The former is consistent with a typical daily commute, taking place when a passenger returns home after a working day (e.g., first trip to work at 9:00 AM, second trip returning home at 6:00 PM). The latter is complementary to the first peak and covers the following overnight period (e.g., 6:30PM– 9:00AM).

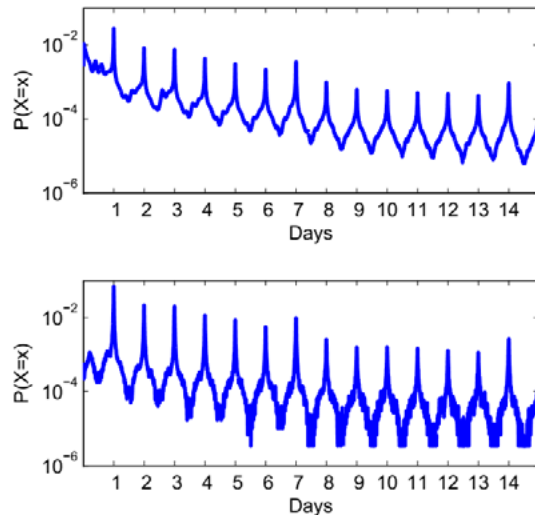
By observing the distribution over a wider time window, we can identify that one-day cycles have the largest probability of occurrence. Surprisingly, this suggests that the usual commuting pattern (home–work–home) was not observed, as only one bus ride was observed daily. To further explore this finding, we examined the joint distribution of the trip starting times of consecutive bus rides (Figure 12). The probability density is at the highest around the diagonal, which demonstrates that subsequent bus rides often are conducted at the same time of the day. This is especially true for the morning and evening periods of a day, when most rides take place.

**FIGURE 12.** PDF of timings of consecutive bus rides. High probability density on diagonal demonstrates that subsequent bus rides often taken at same time of day.



Since passengers, on average, board at the same hour on consecutive days, one can hypothesize that bus riders take the same route daily with a certain purpose. To test this, we computed the probability distributions of inter-trip times of rides taken with the same lines and rides boarded at the same stops, shown on Figure 13. A clear cyclic pattern of travel is observed, supporting the idea that bus rides that involve the same lines and stops often are connected to a specific and regular trip purpose (e.g., commuting to work or school).

**FIGURE 13.** PDF of inter-trip times with strong cyclic patterns for bus rides taken with same bus lines (top) and departing from same stop (bottom).



## Conclusions

In this work, we studied the regularity of bus usage in Lisbon, Portugal. By mining two months' bus ride data, we were able to reveal some usage behavioral patterns, as follows.

There are more rides on weekdays than on weekends, most possibly due to the usual commutes, and, on average, a user takes 2 daily trips, visits 1.93 distinct stops, and uses 1.55 distinct bus lines. Weekdays travel patterns hold a strong similarity, whereas weekends (especially Sundays) reveal a very distinctive pattern. These behaviors are a strong evidence of the usual commutes.

Inter-trip times are mostly short (18% of observations fell within the interval  $t \leq 30$  minutes). The second trip of the day usually takes place at 9.5 hours and 14.5 hours after the last trip. Subsequent bus rides often are taken at the same time of day (especially in the morning and evening periods).

On average, the use of the bus infrastructure is concentrated on a particular number of bus lines and stops. In the case of bus lines, 70% of riders use between 1–5 lines, and 20% use 5–8 lines (a maximum of 3.8% of the lines available), whereas in the case of bus stops, 70% of all riders visit 1–9 stops, and 20% were seen at 10–23 different stops (up to 23% of bus stops available).

This work has leveraged on the availability of bus ride histories for better understanding the regularity in bus usage behavior. In contrast to existing data mining studies of transport usage that are mostly concerned with aggregate travel characteristics, e.g., travel demand estimation, we analyzed travel behavior patterns of individual bus riders. Understanding individual travel behavior patterns is important for the development of novel personalized transport information systems that can provide proactive assistance to transport users. Our results provide a basis to develop a robust predictive algorithm, which is part of our future work.

## References

- Bejan, A., R. Gibbens, E. Evans, A. Beresford, J. Bacon, and A. Friday. 2010. "Statistical Modelling and Analysis of Sparse Bus Probe Data in Urban Areas." *Proceedings of the 13th Intl. IEEE Conf. on Intelligent Transportation Systems (ITSC '10)*.
- Ceapa, I., C. Smith, and L. Capra. 2012. "Avoiding the Crowds: Understanding Tube Station Congestion Patterns from Trip Data." *Proceedings of the ACM SIGKDD Intl. Workshop on Urban Computing*.
- Cunha, J., and T. Galvão. 2014. "State of the Art and Future Perspectives for Smart Support Services for Public Transport. Service Orientation in Holonic and Multi-Agent Manufacturing and Robotics." *Studies in Computational Intelligence*, 544: 225-234.



- Ferrari, L., M. Berlingerio, F. Calabrese, and B. Curtis-Davidson. 2013. "Measuring Public-Transport Accessibility using Pervasive Mobility Data." *IEEE Pervasive Computing*, 12: 26-33.
- Ferris, B., K. Watkins, and A. Borning. 2010a. "OneBusAway: A Transit Traveler Information System." *Mobile Computing, Applications, and Services*. Berlin Heidelberg: Springer.
- Ferris, B., K. Watkins, and A. Borning. 2010b. "OneBusAway: Results from Providing Real-Time Arrival Information for Public Transit." *Proceedings of the 28th Intl. Conf. on Human Factors in Computing Systems (CHI '10)*.
- Foell, S., G. Kortuem, R. Rawassizadeh, S. Phithakkitnukoon, M. Veloso, and C. Bento. 2013. "Mining Temporal Patterns of Transport Behaviour for Predicting Future Transport Usage." *Proceedings of the 3rd Workshop on Pervasive Urban Applications (PURBA)*.
- Foell, S., G. Kortuem, R. Rawassizadeh, S. Phithakkitnukoon, M. Veloso, and C. Bento. 2015. "Predictability of Public Transport Usage: A Study of Bus Rides in Lisbon, Portugal." *IEEE Transactions on Intelligent Transportation Systems*, 16:5: 2955-2960.
- Lathia, N., and L. Capra. 2011. "Mining Mobility Data to Minimise Travellers Spending on Public Transport." *Proceedings of the ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*.
- Lathia, N., C. Smith, J. Froehlich, and L. Capra. 2012. "Individuals among Commuters: Building Personalised Transport Information Services from Fare Collection Systems." *Pervasive and Mobile Computing*.
- Matias, L., J. Gama, J. Mendes-Moreira, and J. Sousa. 2010. "Validation of Both Number and Coverage of Bus Schedules using AVL Data." *13th International IEEE Conference on Intelligent Transportation Systems (ITSC)*: 131-136.
- Nunes, A., T. Galvão, and J. Pitt. 2011. "Using Social Networks for Exchanging Valuable Real Time Public Transport Information among Travelers." *IEEE 13th Conference on Commerce and Enterprise Computing (CEC)*: 365-370.
- Phithakkitnukoon, S., M. Veloso, C. Bento, A. Biderman, and C. Ratti. 2010. "Taxi-Aware Map: Identifying and Predicting Vacant Taxis in the City." *Proc. Aml 2010, First International Joint Conference on Ambient Intelligence*: 86-95.
- Smith, C., D. Quercia, Lici. 2012. "Anti-Gravity Underground?" *Proceedings of the Second Workshop on Pervasive Urban Applications (PURBA)*.
- Weigang, L., W. Koendjibharie, R. Juca, Y. Yamashita, and A. Maciver. 2005. "Algorithms for Estimating Bus Arrival Times using GPS Data." *IEEE 5th International Conference on Intelligent Transportation Systems*.
- Zhang, L., S. Gupta, J. Li, K. Zhou, and W. Zhang. 2011. "Path2Go: Context-Aware Services for Mobile Real Time Multimodal Traveler Information." *Proceedings of the 14th IEEE Intl. Conf. on Intelligent Transportation Systems (ITSC '11)*.

Zimmerman, J. et al. 2011. "Field Trial of Tiramisu: Crowd-Sourcing Bus Arrival Times to Spur Co-Design." Proceedings of the Annual Conf. on Human factors in Computing Systems (CHI '11).

## About the Authors

**STEFAN FOELL** (*stefan.foell@open.ac.uk*) is a Research Associate at the Ubiquitous Computing and Sustainability Lab at The Open University, UK. He received a Diploma in Computer Science from the Technical University Berlin, Germany. Before joining The Open University, he was a Research Associate at the University of Stuttgart, Germany. His research interests are in urban computing, big data, and smart cities. Currently, he is working in the European Research project GAMBAS on personalized mobile transport guides.

**SANTI PHITHAKKITNUKON** (*santi@eng.cmu.ac.th*) is with the Department of Computer Engineering and Excellence Center in Infrastructure Technology and Transportation Engineering (ExCITE) at Chiang Mai University, Thailand. His research is in the area of urban informatics. He received B.S. and M.S. degrees in Electrical Engineering from Southern Methodist University and a Ph.D. in Computer Science and Engineering from the University of North Texas. Before joining Chiang Mai University, he was a Lecturer in Computing at The Open University, a Research Associate at Newcastle University, and a Postdoctoral Fellow at the MIT SENSEable City Lab.

**MARCO VELOSO** (*mveloso@dei.uc.pt*) is an Adjunct Professor at the Polytechnic Institute of Coimbra and a Researcher at the Center for Informatics and Systems of the University of Coimbra, where he is a member of the Ambient Intelligence laboratory. His research explores the use of data mining techniques on big data for smart cities and intelligent transportation systems. He received a Ph.D. in Science and Information Technology and M.S. and B.S. degrees in Informatics Engineering from the University of Coimbra.

**GERD KORTUEM** (*gerd.kortuem@open.ac.uk*) is Professor of Internet of Things at the Design Engineering department, faculty of Industrial Design Engineering at Delft University of Technology, The Netherlands. His research focuses on the design of digital technologies to tackle key societal issues such as health, energy, and transportation and covers digital cities, urban data mining, the Internet of Things, human computer interaction, and wearable computing. He currently runs research projects on smart energy, intelligent transport and citizen innovation.

**CARLOS BENTO** (*bento@dei.uc.pt*) is an Aggregated Associate Professor at the University of Coimbra. His 100 publications comprise papers in international journals and conferences and book chapters. He is the director of the Ambient Intelligence Lab of CISUC (Amllab), and director of the Laboratory on Informatic Systems at Instituto Pedro Nunes (IPN). Over the past years, his research addressed artificial intelligence and ubiquitous computing for smart sustainable cities and intelligent transport systems.