Characterising and Predicting Urban Mobility Dynamics By Mining Bike Sharing System Data

Ida Bagus Irawan Purnama, Neil Bergmann
School of ITEE, University of Queensland
Brisbane, Australia
i.purnama@uq.edu.au, n.bergmann@uq.edu.au

Raja Jurdak, Kun Zhao
Commonwealth Scientific Industrial and Research Organization (CSIRO), Pullenvale, Australia
Raja.Jurdak@csiro.au, Kun.Zhao@csiro.au

***Abstract* — Understanding the spatio-temporal characteristics of human mobility in urban areas is invaluable especially for traffic management and urban planning. An opportunity to characterize and predict urban mobility is provided by mining *Bike Sharing System* (BSS) trip data spatially and temporally. This study focuses on identifying highly predictable BSS users, revealing their mobility characteristics and predicting their next-place movements. In undertaking user classification, we compare between naïve subscription based classification, namely *registered* and *unregistered users*, and trip number based classification, namely *commuter, regular* and *casual users*. Our analysis of 13 weeks of London BSS data shows that users with high predictability are those with 50 or more historical trips and with commuting patterns that can be fairly represented by *registered users*. However, because of the diversity of usage patterns, the spatio-temporal mobility characteristics of *registered users* are not as uniform as *commuters* who are the fastest, farthest travelled, most consistent, and most predictable riders with Markovian qualities. Using only a first order Markov predictor, the prediction of trip destination based on trip origin in weekday morning is up to 78% predictable for *commuters*.**

Keywords — urban mobility dynamics; characterization; predictability; entropy; data mining; bike sharing system; commuter; next-place prediction; Markov predictor.

I. INTRODUCTION

Understanding human mobility is of fundamental significance for a range of applications from intelligent transportation systems and urban planning to epidemic spread, location-based services and other socioeconomic dynamics. In transportation systems, it is important to observe and predict where, when and how people move at an individual level in order to better manage the system.

Most transportation systems only provide aggregate data on fixed routes without capturing the fine-grained individual trip data. Bike Sharing Systems (BSS) with origin-destination (OD) sensing systems that record only departure and arrival time of a one-way individual trip are among the most promising transportation systems which have such data available. Unlike other transport systems with fixed schedules and routes, BSS give more freedom to users to choose their own route and time individually [8]. Even without GPS tracking, the BSS OD sensing system can usefully capture the spatio-temporal features of individual mobility and allow detailed quantitative analyses at an individual level. However, only limited work has attempted to model mobility as well as prediction for a single mode of transport that captures freedom of movement choices for individual users.

A. BSS Background

Recently, BSS have been implemented widely as a sustainable transportation system in many cities globally with a significant growth from around 375 systems comprising 236,000 bikes in May 2011 to around 535 systems with estimated fleet of 517,000 bikes in April 2013 [1]. Some BSS share their trip data repositories for public access. In London, for instance, publically available trips data can exceed one million a month in summer. Practically, a trip occurs when a user picks up a bike in one station and returns it to an empty docking slot of another station. Then, a system-wide mobility pattern formed by a large number of individual trips between stations over time can be described as a dynamic network. Here, stations are represented by nodes and individual trips or traffic flows between stations are embodied by time-varying edges. Accordingly, as user numbers grow, the complexity of the network will dynamically increase.

If the availability of bikes and empty docking slots cannot serve the demand level, users may not find available bikes to rent, or may not be able to return the bikes to their preferred stations. This is called an imbalanced state [14] which can potentially decrease the efficiency as well as the service level of the system. This imbalanced state is usually resolved by redistributing the bikes manually from full stations to empty ones using service trucks which increase system cost. For example, in Taipei city, its BSS reached a deficit of $NT millions after running for one year, and redistribution is one of the most expensive costs [2]. In Paris's Vélib, the operational cost for redistributing a bike is about $3, and in Barcelona's Bicing, 50% of 230 service staffs are assigned only to the bike redistribution task [2]. These high costs and time-consuming operations will be further compounded if the redistribution scenario is reactive since it redistributes bikes after imbalances occur. Thus, if there is a fundamental spatio-temporal analysis that could identify and categorise users' mobility behaviours, foresee their next locations, and use their recurrent pattern to predict future traffic, this may improve the BSS operation proactively. Since daily human mobility pattern follows clear statistical regularities [11], imbalance can be addressed before it occurs by taking advantage from understanding the spatio-temporal mobility characteristics and predictability of BSS users.

978-1-4673-7211-4/15 $31.00 © 2015 IEEE
DOI 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.46

B. BSS User Mobility

BSS users may have different movement habits and preferences: commuters, tourists, and night workers for example. Meanwhile, the same regularities and patterns are likely to be associated with the same user type. How different users move both spatially and temporally over the BSS therefore has been a subject of some previous studies [3,4,5,6]. However, none of these studies consider distinctions among the BSS user types for predicting their mobility based on their uniform predictability. Recent studies [5,6] usually employ naïve approaches using subscription and demographic status to differentiate users which risks creating users group with outlier and non-uniform movement patterns. For example, some unregistered users that generally have random patterns may have regular trips similar to registered users, and vice versa. In this case, regularities are associated with how frequently a user travels, not on their registration status. Moreover, the performance of the mobility prediction algorithm should be benchmarked by the upper bound of predictability metric that is inherent to the data, while recent studies do not rely on this information theory analysis.

In a temporal context, trip history shows the frequency of user mobility and intervals between two consecutive trips. A certain waiting time may implicitly indicate the user type. If waiting times are between 6 to 8 hours and it occurs frequently only in weekdays, it can be expected that the user is a commuter, even without considering the spatial context or departure-arrival stations. Similarly, users who have few trips with short and uncertain amount of waiting time and random route might be tourists. Thus, if we can group these kinds of users into certain classes, it would be possible to measure predictability and analyse prediction in more detail. Here, prediction means the accuracy of predicting the next user's place which can be the next consecutive station, next return station only, or next pickup station only.

C. Paper Contribution and Structure

This data driven study mainly aims (1) to analyze the degree of randomness and predictability of BSS users; (2) to define new users classes based on predictability results and to validate this classes with spatio-temporal analysis; (3) to propose the next-place prediction algorithm based on the result of user classification; and (4) to explore whether usage based classification can improve the predictability.

The rest of the paper is organized as follows. Section 2 reviews current related work, followed by technical background in section 3. Section 4 describes the datasets, and section 5 reports users classification followed by spatio-temporal analysis and results. The prediction method and accuracy are presented in section 6, and the paper finishes with conclusions and future work in section 7.

II. Related Work

Human mobility studies using data driven approaches have been conducted from various data sources such as banknotes [23], mobile phone [10,25], social media [16], taxi data [9], GPS-based traces [22], and BSS [14,19,20]. Those studies are mostly centred on spatio-temporal statistical, visualization and network analysis. Focusing in urban areas, BSS data can show the population density of a city and usefully capture the individual mobility with clear geo-location of origin-destination, albeit without the real route of each trip. Estimating the real route of BSS users, instead of only using Euclidean distances[1] [13,19], remains an open challenge. As a mode of transportation, BSS users will most likely follow the road network. For this reason, our work infers the trip distance by selecting the most likely road segment between OD.

BSS as complex networks have been studied to reveal the dynamics of a city. Borgnat et al. [14] have explored BSS from signal processing and data analysis perspectives. They modelled the time evolution of the Velo'v dynamics movement in Lyon, Paris, showing that it is mostly cyclostationary[2] over the week, and then disentangling the spatial patterns to understand the flows. This study gives insights on the users' social behaviours when using BSS. In a later study, O'Brien et al. [19] have conducted a global view of BSS data for generating insights into sustainable transportation systems from 38 systems located in Europe, Asia, Australasia and the Americas. They mainly analysed the variation of occupancy rates over time and comparison across the system's extent to infer the likely demographics and intentions of user groups.

Several researchers have attempted to characterise BSS users and their behaviours. Beecham et al. [3] have used spatial analysis, namely density-estimation, in classifying the commuting behaviour of London bicycle data to identify the potential commuting cyclists and their plausible workplaces. The authors compared the terminating and originating journeys within the same vicinity of derived workplaces between peak-times in the morning and in the afternoon [3]. Previously, in an effort to infer commuting behaviour, Lathia et al.[4] have made sensible assumptions from London underground data that people who do commute should have at least two trips per working day with a frequent repetition in the same origin-destination pair. However, more than half of people did not fit this spatial closed-loop model criterion. Therefore, it may be better to use this in combination with temporal attributes to better classify users.

Beyond commuters, in another study, Lathia et al.[5] have also analysed casual user behaviour before and after the London bike hire policy was changed in December 2010, from using a registration key to access the system to simply using a debit or credit card to do so. Unlike commuters from the previous study, casual user here is simply defined as an unregistered user as opposed to a registered user that may travel more frequently. Similar to Lathia et al.[5] who used clear casual data, Vogel et al.[6] have used the period of users' membership data (annual, weekly or daily) of Velo'v, Lyon, where annual membership contains clear users' demographic information such as age, gender and postcode. However, using only subscription data to classify users can possibly get improper classes because they may not be generic enough to capture similar behaviour.

[1] A straight-line distance from the origin to destination station.

[2] A process having statistical properties that varies cyclically with time.

In terms of predictability, Song et al.[10] came with a fundamental question: What is the role of randomness in human behaviour and to what degree are human actions predictable? Then, they explored the limit of predictability by measuring the entropy of each individual's trajectory among anonymised mobile phone users. They found a 93% potential predictability in user mobility. Later, Lu et al.[25] measured the movement uncertainties of 500,000 individual travel patterns among mobile phone users in Cote d'Ivoire by considering the frequencies and temporal correlations of individual trajectories. They found that the theoretical maximum predictability is as high as 88%. Recently, Sinatra and Szell [11] have applied entropy and predictability to measure the behavioural actions and mobility of a large number of players in the virtual universe of a massive multiplayer online game. They found that movements in virtual human lives follow the same high levels of predictability as offline mobility. However, to our knowledge, no studies have implemented the concept of entropy and predictability in BSS data yet. As a specific mode of transport in urban areas, we expect the predictability bound of BSS data has unique features that are different from mobile phone data.

In term of mobility prediction, several spatio-temporal prediction algorithms have been proposed. Asahara et al.[27] proposed a variant of a Markov model called the *Mixed Markov-chain Model* (MMM) which is an extension of standard *Markov Models* (MM) and *Hidden Markov Models* (HMM). Based on the authors' observation, the two previous models were not generic enough to encompass all types of mobility [17]. The MMM was proposed due to the existence of similar mobility behaviour among certain pedestrians. Later, Gambs and Killijian [17] adopted the concept of *Mobility Markov Chain (MMC*) and extended it to n-MMC in order to incorporate the *n* previously visited locations from GPS-enabled devices. They found that the prediction accuracy of the next location is in the range of 70% to 95% as soon as n = 2. Here, based on our predictability analysis, we show that a first-order Markov predictor performs well for a highly predictable group of BSS users, namely *commuters*.

III. TECHNICAL BACKGROUND

In this section, we provide some preliminary technical definitions that we use to characterize user mobility, namely displacement flow, radius of gyration, entropy, predictability, prediction accuracy and Markov Model.

A. Displacement flow

Displacement flows can be described by a vector of the origin-destination station, time of departure-arrival, and the corresponding type of user-type [21]. Let $X_{d,h}^{(u,v)}$ be the number of bikes going from origin $u \in \{1,\ldots, S\}$ to destination station $v \in \{1,\ldots, S\}$, day $d \in \{1,\ldots, D\}$ and hour $t \in \{1,\ldots, 24\}$ [21]. With a one-hour aggregation window, the flows will be:

$$X_d^{(u,v)} = (X_{d,1}^{(u,v)}, X_{d,2}^{(u,v)}, \ldots, X_{d,24}^{(u,v)}) \quad (1)$$

B. Radius of Gyration (RoG)

RoG is a spatial feature that quantifies the spatial stretch of an individual trajectory [7]. This shows how far a user moves from his/her geometric centre.

$$RoG = \sqrt{\frac{1}{N}\sum_{i=1}^{n} (\overline{r_i} - \overline{r_{cm}})^2} \quad (2)$$

Where N is number of trips, r_i is distance of trip i, and $\overline{r_{cm}} = \frac{1}{N}\sum_{i=1}^{n} \bar{r}_i$ is the geometric centre of the trajectory [7].

C. Entropy

Entropy is commonly used to measure the degree of randomness in time-series data. Here, entropy is applied to measure the randomness of human mobility denoted by a sequence of visited locations. Following [10,11,12], there are four different representations of entropy: *random* (S^{Rand}), *Shannon* (S^{Shan}), *conditional* (S^{Cond}), and *real* (S^{Real}). All will be bound by the relationship $S^{Real} \leq S^{Cond} \leq S^{Shan} \leq S^{Rand}$.

1) The random entropy only considers the number of distinct visited bike stations, N, by an individual user, i.

$$S_i^{Rand} = log_2 N_i \quad (3)$$

2) The Shannon entropy counts the probability of visiting each distinct station, p_i, by an individual user, i.

$$S_i^{Shan} = -\sum_{i=1}^{N} p_i \, log_2 \, p_i \quad (4)$$

3) The conditional entropy captures the correlation between one bike station x_{t-1} with the subsequent station x_t in the time series.

$$S_i^{Cond} = -\sum_{x_t \in X_t} \sum_{x_{t-1} \in X_t} p_i(x_{t-1}, x_t) log_2 \, p_i(x_t | x_{t-1}) \quad (5)$$

$p_i(x_{t-1},x_t)$ is the probability of visiting the ordered pair of stations, x_{t-1} and x_t, $p_i(x_t|x_{t-1})=p_i(x_{t-1},x_t)/p_i(x_{t-1})$ is the probability of visiting station x_t at time t given a preceding station, x_{t-1}.

4) The real entropy evaluates the randomness based on the full spatio-temporal information of the sequence: frequency, visitation order and time spent. It is estimated using a *Lempel-Ziv* (LZ) algorithm estimator that searches for repeated sequences.

$$S_i^{Real} = \left(\frac{\sum_{i=2}^{n} l_i}{n \, log_2 n}\right)^{-1} \quad (6)$$

Where l_i is the length of the shortest location at position i that does not appear in the part of sequence up to position i - *1*.

D. Predictability

Predictability can be stated as a measure of the users' future whereabouts that could be potentially achieved. Naturally, the predictability is inversely proportional to the entropy where low entropy is linked to high predictability and vice-versa. By formula inversion, it can be computed as a function of any kind of entropy S as in (7) below:

$$S_i^{\bullet} = -\Pi_i^{\bullet} log_2 \Pi_i^{\bullet} - (1 - \Pi_i^{\bullet}) log_2 (1 - \Pi_i^{\bullet}) + (1 - \Pi_i^{\bullet}) log_2 (N_i - 1)$$

Where $\bullet$ is a placeholder for any type of entropy.

E. Prediction Accuracy

Prediction accuracy of a predictor can be defined as a ratio between the numbers of correct predictions over the total number of predictions [17].

$$P_{Acc} = \frac{P_{correct}}{P_{all(true+false)}} \quad (8)$$

Here, the correct prediction is set by discrete validation prediction in which the prediction outcome is binary (true or false).

F. Markov Model (MM)

A Markov model uses the previously visited locations to predict the next location. This predictor provides a simple approach to capture sequential dependence, and is defined by a set of states $S = \{S1, \ldots, S_n\}$, a transition matrix $T = \{t_{1,1}, \ldots, t_{n,n}\}$, and a vector of initial probabilities $P = \{p_{1,1}, \ldots, p_{n,n}\}$, [26]. Each transition $t_{i,j}$ has a probability $p_{i,j}$ assigned to it that corresponds to the probability of moving from state S_i to state S_j [17]. MM can be represented either as a directed graph, Fig. 1, or a transition matrix, Table I.

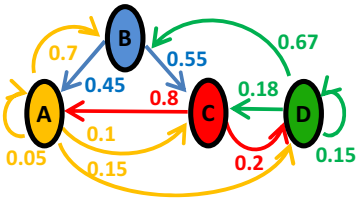


Fig. 1. Graph representation of transition states by nodes and edges.

TABLE I. THE PROBABILITY TRANSITION MATRIX

State	A	B	C	D
A	0.05	0.7	0.1	0.15
B	0.45	0	0.55	0
C	0.8	0	0	0.2
D	0	0.67	0.18	0.15

IV. DATASETS

With the technical background in hand, this section discusses our main dataset which is user trips history of *London's Cycle Hire Scheme* (LCHS)[3]. There are six major data fields as shown in Table II: user identifier, pickup/return bike stations, start/end timestamp, and trip duration. Stations' geo-location (latitude, longitude) and user registration type are given in separate data, which are linked to each station and user in Table II.

TABLE II. LCHS TRIPS DATA STRUCTURE EXAMPLE

User ID	Pickup Data		Return Data		Duration (minutes)
	Station	*Pickup time*	*Station*	*Return time*	
xxxx	251	2012-08-01 06:34	506	2012-08-01 06:40	5.75
xxxx	239	2012-08-01 07:05	44	2012-08-01 07:15	9.95

We divide our dataset into training and testing datasets. Training dataset, D1, contains the 2012 summer to autumn trips from 1 Aug 2012 to 31 Oct 2012 (92 days~13 weeks), while testing dataset, D2, spans from 1 to 11 Nov 2012 (11 days). Originally, all data cover 2,961,183 trips linked to 566,888 users that were collected from 569 bicycle stations in Central London. The dataset then has been cleaned to exclude looping trips (trips with identical origin and destination in which the distance could not be computed) and trips with an unrealistic duration (< 1 min or > 24 hrs). This eliminates data that is not valid for our analysis. After removing the affected data, its record declines by 5.25% to 2,805,718 trips with 566,456 users.

The dataset itself categorizes users into two naïve classes: *unregistered* and *registered* users. This relates to how they subscribe and use the system. Most of the users, 89%, are *unregistered* users who have only 43% trips, while 11% of *registered* users have 57% trips. This inverse percentage comes from their average trips per user which is only 2.41 trips per *unregistered* user with a standard deviation of 2.45 and 26.71 trips per *registered user* with a standard deviation of 35.26. This clearly shows how *registered users* mostly are frequent riders in contrast with *unregistered* ones.

Another dataset used in this study is a daily historic record of rainfall level[4] in Central London. This rainfall log is used as an independent data stream for cross-validation of classified users' behaviour when dealing with weather conditions. The aim is to compare the response of group of users with respect to the rainfall in order to highlight any differences.

V. ANALYSIS AND RESULTS

In this section, we first analyse the distribution of users by their predictability level and use their average distribution to define the related number of trips. This is initially done to guide the definition of homogeneous user groups based on similar regularity patterns. Temporal and spatial analyses are then used to characterize the differences of the mobility patterns among the groups. To show the diversity of usage role in identifying the highly predictable users, entropy and predictability analysis is used with the classified users. Finally, the Markovian trait of the highly predictable users is tested by analysing the similarity of real and conditional predictability.

A. Preliminary Analysis

The full spatio-temporal information of origin-destination sequences[5] can be applied to get an initial insight of prediction level using real predictability, Π^{real}. Accordingly, we first analyse the distribution of users by their predictability, Fig. 2.a, and its correspondence to the number of trips, Fig. 2.b.

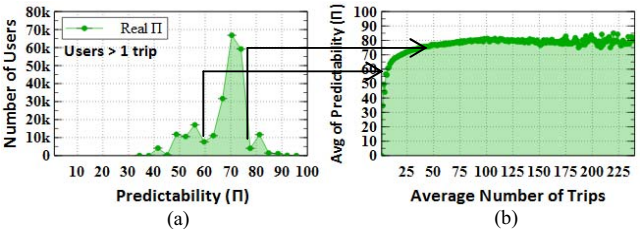


Fig. 2.(a) Real predictability of users, (b) Average of predictability per # trips

[3] Downloaded from TFL website (www.tfl.gov.uk) which provides a public open access with email sign-up permission.

[4] Downloaded from the nw3 weather station web (nw3weather.co.uk)

[5] Sequences of all origin-destination station in a timeline.

Fig. 2.a shows three peaks in the predictability that can represent low, medium and high predictability where the predictability bounds are between 60 and 78. If we then map these predictability bounds to the average predictability by number of trips, we will get trip number threshold around 7 as lower and 50 as upper thresholds, Fig. 2.b. In a previous predictability study, Lu et al.[25] also found that more than 50 historical observations give a clear difference in prediction accuracy between stationary and non-stationary trajectories. This means that a fairly good predictability could be achieved with a sufficient number of trips. Because the length of a trip sequence is formed by the total number of trips, the *individual total trip* (ITT) is then defined as an aggregated metric for each individual for user classification purposes.

B. User Classification

Using ITT values of 7 and 50 from Fig. 2.b above, we get the distribution of users and trips as in Table III, and we define them as *casual users* (ITT ≤ 7), *regular users* (7 < ITT < 50), and *commuters* (ITT ≥ 50) for low, medium and high predictability users respectively. Table III clearly shows that most *casual users* and *trips* come from *unregistered users* and most *commuters* come from *registered users*, while *regulars users* are evenly split between *registered* and *unregistered users* and accomadate the outliers from these two classes.

TABLE III. BOUNDS AND PERCENTAGES OF CLASSIFIED USERS

Class	Bound	Users		Trips	
		Unregistered	*Registered*	*Unregistered*	*Registered*
Casuals	ITT ≤ 7	95.93 %	4.07 %	94.58 %	5.42 %
Regulars	7 < ITT < 50	52.02 %	47.98 %	34.28 %	72.39 %
Commuters	ITT ≥ 50	0.08 %	99.92 %	0.05 %	99.94 %

C. Temporal Analysis

In order to validate the users' classification, Fig.3 describes three temporal attributes of all users which are day of training period (daily pattern), time of the day (hourly pattern) and waiting time (minutely pattern). Daily patterns display the cumulative number of users and their trips for every single day with additional daily rainfall level along the 92 days of training period. Hourly patterns describe the hourly cumulative number of trips in each weekday and weekend along the training period. Waiting time shows the idle times in minutes between two consecutive trips both in weekday and weekend. The daily rainfall level is used as an external independent attribute to observe the response of classified users to the weather conditions.

In daily patterns, the number of *casual users* and trips as well as *unregistered* ones highly depends on the weather as shown between days 50 to 90, Fig. 3.a,c. When the rainfall was high, the trips decrease significantly. For both users, weekend trips are higher than weekday trips. Meanwhile, *commuters* as well as *registered users*' behaviour show a relatively stable pattern even on rainy days, Fig. 3.b,e. Their weekend trips are almost half of the weekday trips. This shows they do not use the BSS during weekends and travel regularly during weekdays. On the other hand, the differences are not so significant for *regular users*.

In hourly temporal patterns, again, contrasting patterns are shown between *commuters* and *casual users* as well as *registered* and *unregistered* users, while *regular users* have patterns between the two (Fig. 3.f-j). In particular, *commuters* have two sharp *peak times* (PT) for all days during weekdays which are **PT1: 5:00-9.00 am** and **PT2: 15:00-19.00 pm**, and only a small number of trips in middle of weekdays. These two peaks could be related to the rush trips hours in the

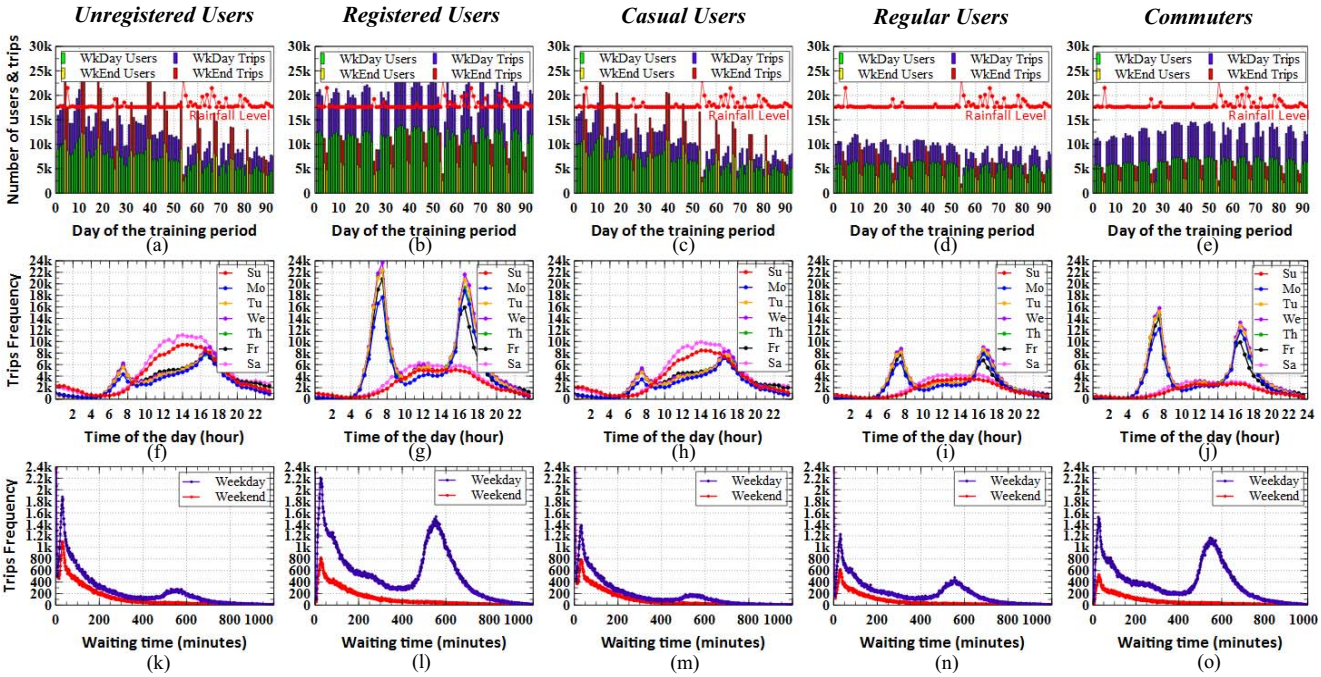


Fig. 3.(a-e) Day of training period distribution, (f-j) Time of the day distribution, (k-o) Waiting time distribution.

morning when most *commuters* go to their workplaces and in the afternoon when they go back home. Conversely, *unregistered* and *casual users* have fewer trips during peak hours in weekdays because their leisure use of BSS peaks from midday to afternoon of weekends.

We define *waiting time* (WT) is the time between the end of a trip and the next trip, but WT in this study is counted only on a daily basis. Therefore, the WT between the last trips on particular day to the first trip on the next day will not be counted. This will ensure that a WT captures the daily idle time patterns. In weekdays, *commuters* as well as *registered users* show a spike between 450 – 690 minutes (≈ 7.5 – 11.5 hours) that does not appear in weekends. This WT aligns well with the common working hours during the week and completely disappears on the weekend. Meanwhile, *casual* and *unregistered users* are dominated by short waiting times.

D. Spatial Analysis

Spatial analysis aims to show the characteristics of displacement shifts mainly *distance* (D) and *radius of gyration* (RoG). Combining *distance* (D) with *duration* (T), *speed* (S) can then be calculated. All these displacement metrics are summarized in Table IV and averaged over peak times (PT) and out of peak (OFP) with the exception of RoG because of its continuity traits. To show the complexity of displacement attributes that usually have a fat tail distribution, these are viewed using a logarithmic scale, Fig. 4.a-c, except for speed with numeric scale as shown in Fig. 4.d.

TABLE IV. DISPLACEMENT METRIC PER PEAK TIME AND OUT OF PEAK

Groups	**Naïve Classification**						**Trip Number Based Classification**								
Users	**Unregistered**			**Registered**			**Casual**			**Regular**			**Commuter**		
Metric	**D**	**T**	**S**	**D**	**T**	**S**	**D**	**T**	**S**	**D**	**T**	**S**	**D**	**T**	**S**
Unit	***km***	***min***	***k/h***	***km***	***min***	***k/h***	***km***	***min***	***k/h***	***km***	***min***	***k/h***	***km***	***min***	***k/h***
PT1[a]	3.2	18.6	13.5	2.9	13.3	14.6	3.1	19.6	13.2	2.9	14.3	14.2	3.0	12.8	14.9
PT2[b]	2.8	28.1	9.8	2.8	14.1	13.9	2.8	31.2	9.8	2.7	16.5	12.6	2.8	13.2	14.3
OFP[c]	2.7	30.1	9.4	2.4	13.9	13.2	2.7	32.9	9.4	2.5	17.3	11.8	2.5	12.6	13.8
Avg	2.7	28.5	9.9	2.7	13.8	13.8	2.7	31.2	9.8	2.6	16.5	12.5	2.7	12.8	14.3
St.dev	1.7	140	5.3	1.6	70	4.6	1.7	184	5.2	1.6	97	4.9	1.6	42	4.5
RoG	570.1 (m)			1006.8 (m)			485.4 (m)			1009.9 (m)			1029.4 (m)		

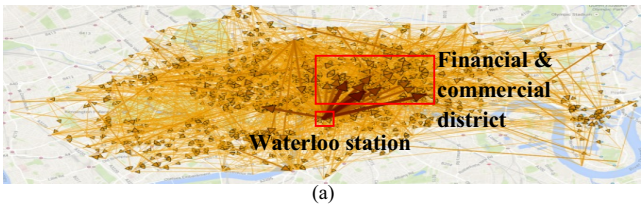
[a]PT1: 05:00 – 09:00 am, [b]PT2: 15:00-19:00 pm, [c]OFP: Out of PT1 and PT2

Results show that, during the morning peak (PT1) which is usually a rush hour on weekdays, all users ride faster and farther than at any other times. On average, *commuters* are the fastest riders, while *casual users* are the slowest, $S^{cas} < S^{unr} < S^{reg} < S^{rgt} < S^{com}$. Moreover, *casual users* spend more time riding than others, and *commuters* spend the least, $T^{com} < T^{rgt} < T^{reg} < T^{unr} < T^{cas}$. This could be because *commuters* tend to be in hurry especially in the morning, while *casual users* such as tourists tend to ride more slowly for sightseeing.

In orbits of movement, the RoG of *commuters*, *registered* and *regular users* show a skewed decay over increasing trips distance. The majority of trips were confined to a distance of 1000 metres. *Casual* and *unregistered users* have shorter radius, $RoG^{cas} < RoG^{unr} < RoG^{reg} < RoG^{rgt} < RoG^{com}$, because they have very few trips, and their distributions do not have a characteristic length scale like others.

E. Route Estimation

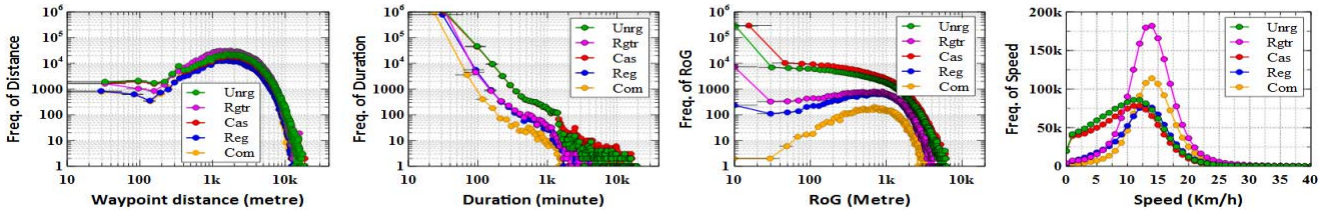
We also calculate and visualise the waypoint or network distance, Fig. 5.b, using an application that queries some options for paths to two online map servers: Google Maps and MapQuest. This can estimate the bike trip routes instead of only using straight line distances, Fig. 5.a. O'Brien et al.[19] did not consider this waypoint distance in calculating speed. Using this new approach, we found that the average trip length difference increasing by 22.64%. As a result, this waypoint distance method yields an average speed which is in close agreement with the experiment that has been done by Jensen et.al.[15] for BSS in Lyon, France. They got a precise distance as well as speed, 14.5 km/h, using a counter installed on the bicycle, while in this study a realistic speed average for commuters can be inferred which is approximately 14.9 km/h in the morning, Table IV. However, not all waypoints are available in map servers. This can be seen from a few long straight lines that are still exist, Fig. 5.b.



(a)

(b)

Fig. 5. (a) OD Euclidian distance, (b) Estimated waypoint route.



(a) (b) (c) (d)

Fig. 4. Spatial analysis distribution: (a) Waypoint distance, (b) Trips duration, (c) Radius of Gyration, (d) Trips speed.

F. Entropy and Predictability

In this section, to investigate the randomness of the visitation patterns, entropy and predictability for all users based on formulas (3) to (7) are used to calculate, Fig. 6.a-j. Firstly, without considering users' activity (pickup or return), entropy and predictability of *casual* and *unregistered users* suggest that the large number of users with few trips give rise to a huge volume of discrete predictability values, since the number of visitation pattern configuration for these users is limited. This makes their histograms jagged and hard to analyse, Fig. 6.a,c,f,h. The figures for *commuters* and *regular users* are smooth and satisfy the basic entropy ordering rule: $S^{Rand} \geq S^{Shan} \geq S^{Real}$, because they have sufficiently long trip histories compared to *casual* and *unregistered users*. Note that estimation of different types of entropy and the above inequality becomes exact only for infinite sequence.

Since the S^{rand} mean value in *commuters* is around 4.5, this indicates that a typical user who will go to the next bike station could randomly be found in any of $2^{Srand} = 2^{4.5} \approx 22$ stations. This random possibility is a result of considering the distinct visited stations only. If visitation frequency is counted, then the uncertainty will be shown in Shannon entropy with the mean value of 3, $S^{Shan} \approx 2^3 \approx 8$ stations. Similarly, if the sequence orders of visitation, repetition or frequency and time spent are taken into account, then the real entropy greatly reduces to $S^{Real} \approx 2^{1.8} \approx 3.48 \approx 4$ stations. This ordering result is consistent with the findings of a previous study [10].

The predictability graphs in Fig. 6.f-j show the inversely proportional relationship with entropy. *Commuters* have the highest real predictability in which the mean value of Π^{real} is around 78, Fig. 6.j, followed by *regular* and *registered users* which are 66 and 68 respectively. This indicates there is a possibility that 78% of commuters' whereabouts could be predicted using a suitable prediction algorithm, while 22% of cases appear random and hard to predict. Here, predictability provides an upper bound of prediction algorithm performance. Using same method, we also calculate predictability for the sequence of visited stations based on their activity, pickup or return only. This approach can increase the commuters' predictability up to 2 % from the first approach which does not differentiate the activity of visited stations.

G. Markovian Traits

To identify if the recurrent patterns of commuters' mobility has a close relation with their previous trip patterns, we then compare real entropy to conditional one, Fig. 7.a. The result shows that $S^{Real} \sim S^{Cond}$.

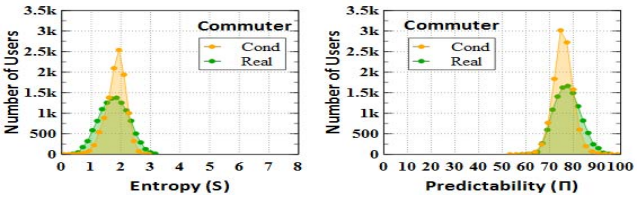


Fig. 7.(a) Cond and real entropy, (b) Cond and real predictability

This suggests the Markovian traits where actual entropy can be represented by conditional entropy. This means that predictability depends on the previous location, and is less affected by earlier trip history. The conditional predictability, Π^{cond}, also gives an upper bound on prediction accuracy of 78%. Again, predictability shows strong Markovian characteristics, $\Pi^{real} \sim \Pi^{cond}$. This indicates considering the last station yields the same predictability as considering the entire of trips. A Markov model predictor should achieve close to this 78% prediction accuracy.

VI. Prediction

In this section, the first order Markov Model (MM1) predictor is applied to predict the next location (station) in the predefined testing dataset, D2: 1 - 11 November 2012, using probability matrix from learning dataset, D1: 1 August – 31 October 2012. We propose three scenarios, Fig. 8.a-c: (a) prediction of the next station that uses the order of visited stations without differentiating between pickup and return which has predictability around 78% for *commuters*; (b) pickup to return prediction that predicts the return station from a given pickup station; and (c) the last return to the next pickup prediction as a converse of scenario b. Here, scenario b and c implement the difference activity of visited stations which has predictability around 80% for *commuters*.

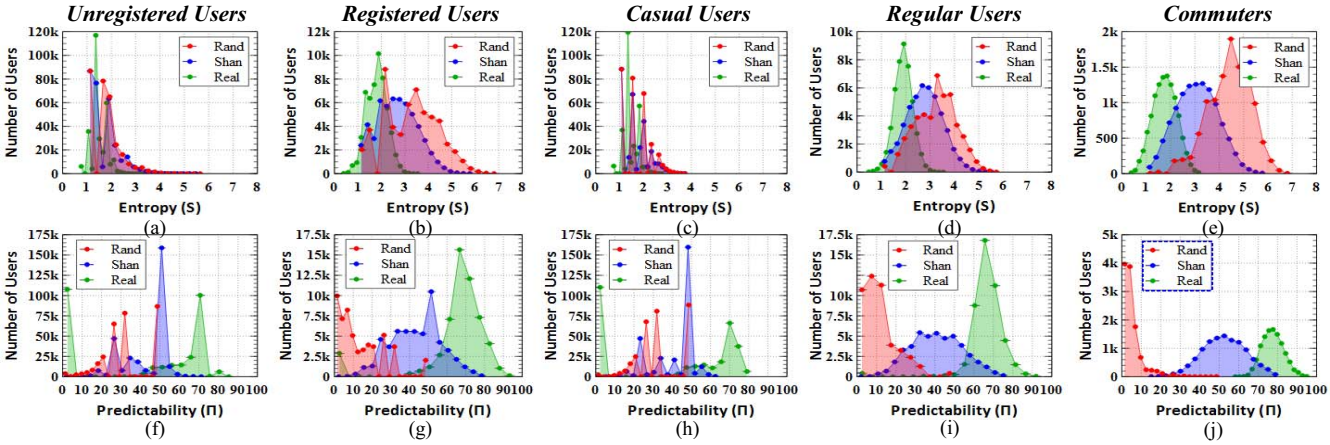
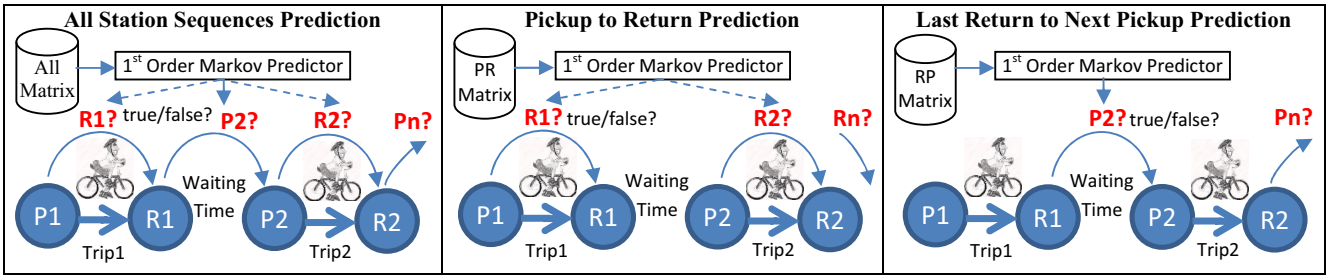
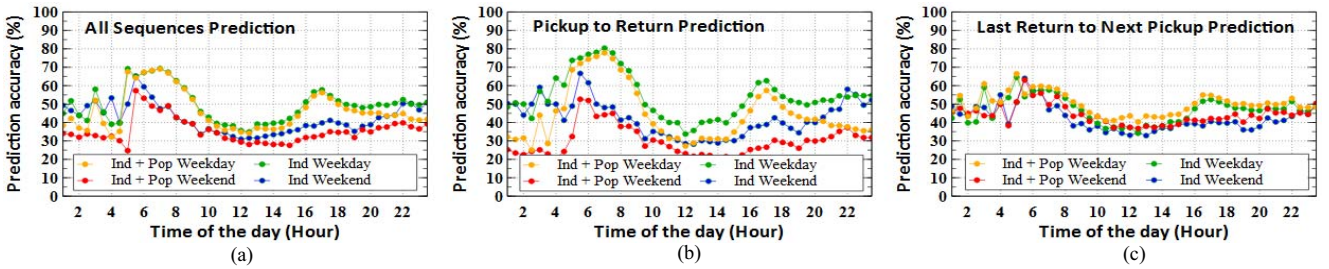


Fig. 6.(a-e) Entropy distribution, (f-j) Predictability distribution.



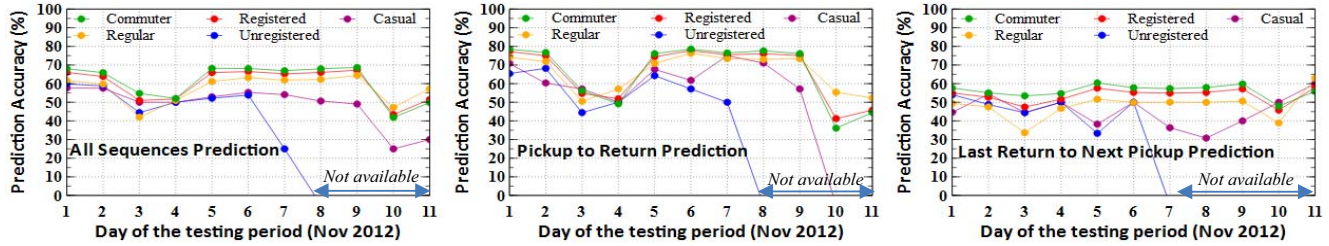
(a) (b) (c)

Fig. 8. Three prediction scenarios using first order Markov Model predictor.



(a) (b) (c)

Fig. 9. Hourly accuracy of individual and population prediction on weekday and weekend



(a) (b) (c)

Fig. 10. Daily accuracy of individual based prediction per user group in the morning peak time.

Fig. 8 shows that the first order Markov Model only uses the probability information of *the one previous state* from *the intended state*. This means that if the intended state is the next station then the previous state is the current station. For each input of a station, the predictor will use the probability matrix of trip history to or from that station. Then, it will give the highest probability to whichever next station a particular user is most likely to visit. For the first scenario, it will look at all visited stations, while the other two will look the pickup to return matrix or the last return to the next pickup matrix only. However, not all users have a trip history with a particular station. To address this issue, we employ two approaches: (1) using population based matrix which will be combined with individual based matrix for users who have trip history; (2) using individual based matrix only which means we ignore the unavailable trip history. These two approaches are intended to see whether population based data can improve or decrease the prediction accuracy.

The results for prediction accuracy (the proportion of correct predictions) are presented in Fig. 9 and 10, showing the hourly prediction dynamics (averaged over the same hour for the 11 testing days) and the daily dynamics (averaged per users across the morning for each of the 11 testing days).

A. Hourly Prediction

Fig. 9.a shows the hourly prediction accuracy for the first prediction scenario using individual and population approaches, separated into weekdays and weekends. Generally, prediction accuracy in weekdays is higher than in weekends, and in the morning is higher than at other times. In particular, the highest accuracy is around 70% at 5:00 and 7:00 am. This accuracy is still below than predictability upper bound of 78%. On the other hand, the second scenario, Fig. 9.b, gives a higher result. It reaches 80% at 7:00 am. Finally, Fig. 9.c shows that accuracy for the last return to the next pickup is less than the reverse one, around 60% at 7:00 am, but the population prediction can slightly increase the accuracy (orange dots are higher than green dots). Weekend and weekday are almost similar in this scenario. All these results indicate that trips from pickup to return stations are more predictable than idle transition state from return to pickup stations.

B. Daily Prediction

Focusing only in the morning when the prediction accuracy is higher, Fig. 10 gives daily morning prediction accuracy per user class. In weekdays, *commuters* show almost constant accuracy nearly 70%, 78%, and 60% in the three

scenarios respectively. This is closely followed by *registered users*. From this result, in terms of user classification, the *registered user class* is sufficient to represent highly predictable users, with only a small additional benefit from using the *commuters class*.

VII. CONCLUSION

This study has characterised BSS mobility patterns and has identified a group of highly predictable users. Using predictability as a benchmarking metric, we show that a simple first-order Markov predictor performs well for individual within this user group. We have investigated the difference between using the naïve subscription status and more detailed and temporal trips data to classify BSS users. Spatio-temporal analysis was used to reveal users mobility dynamics and predictability, and then predict their next place movement. Using trip number based classification to separate outliers, the results show strong distinction of spatio-temporal mobility characteristics among users. In the morning peak time, all users ride faster and farther than at other times. In particular, *commuter*s are the fastest, farthest, lest affected by rainfall, and show the most consistent trip patterns. On the other hand, *casual user*s are the slowest, influenced most by rainfall, have the shortest radius of gyration and most random use pattern. In a prediction context, *registered users* can represent the highly predictable users, but *commuters* are still the slightly superior. In addition, the waypoint distance can be used as a new method to represent the near real distance instead of the straight line distance.

For users with a sufficiently long history, we find that the conditional entropy contains as much information as the real entropy, which indicates that the visiting sequence lacks long-range correlation, and the user's next location is mostly dependant on his/her current location. Therefore, we used a simple first-order Markov predictor that has achieved favourable prediction accuracy, which is just slightly lower than the theoretical limit. Future work will explore the use of higher order Markov Model predictors [17] or more advanced methods such as neural networks to achieve higher accuracy [28], and the effect of special events for BSS usage patterns.

REFERENCES

[1] J. Larsen, "Bike-Sharing programs hit the streets in over 500 cities worldwide," Earth Policy Institute, 2013.

[2] J. H. Lin and T. C. Chou, "A Geo-Aware and VRP-Based Public Bicycle Redistribution System," International Journal of Vehicular Technology, 963427, 2012.

[3] R. Beecham, J. Wood, and A. Bowerman, "Studying commuting behaviours using collaborative visual analytics," Computer, Environment and Urban System, 47:5-15, 2014.

[4] N. Lathia, C. Smith, J. Froelich, and L. Capra, "Individuals among commuters: Building personalised transport information services from fare collection systems," Pervasive and Mobile Computing, 9(5): 643-664, 2013.

[5] N. Lathia, S. Ahmed, and L. Capra, "Measuring the impact of opening the London shared bicycle scheme to casual users," Transportation Research Part C, 22: 88-102, 2012.

[6] M. Vogel, R. Hamon, G. Lozenguez, L. Merchez, P. Abry, J. Barnier, P. Borgnat, P. Flandrin, I. Mallon, and Robardet, "From bicycle sharing system movements to users: a typology of Velo'v cyclists in Lyon based on large-scale behavioural dataset," Journal of Transport Geography, 2014.

[7] M. C. Gonzales, C. A. Hidalgo, and A-L. Barabasi, "Understanding individual human mobility patterns," Nature, 453(7196):779-782, 2008.

[8] R. Jurdak, "The impact of cost and network topology on urban mobility: a study of public bicycle usage in 2 U.S. cities," PLoS ONE 8(11): e79396, 2013.

[9] X. Li, G. Pan, Z. Wu, G. Qi, S. Li, D. Zhang, W. Zhang, and Z. Wang, "Prediction of urban human mobility using large-scale taxi traces and its applications," Front. Comput. Sci 6(1):111-121, 2011.

[10] C. Song, Z. Qu, N. Blumm, and A-L. Barabasi, "Limits of Predictability in Human Mobility," Science, Vol 327, 1018, 2010.

[11] R. Sinatra and M. Szell, "Entropy and predictability of online life." Entropy, 16:543-556, 2014.

[12] R. Gallotti, A. Bazzani, M. D. Esposti, and S. Ramdaldi, "Entropic measures of individual mobility patterns," Journal of Statistical Mechanism: Theory and Experiment, P10022, 2013.

[13] J. Froehlich, J. Neumann, and N. Oliver, "Sensing and predicting the pulse of the city through shared bicycling," IJCAI'09. AAAI Press, 14020-1426, 2009.

[14] P. Borgnat, C. Robardet, J-B. Rouquier, P. Abry, E. Fleury, and P. Flandrin, "Shared bicycles in a city: A signal processing and data analysis persepective," Advances in Complex Systems, 14 (3): 415-438, 2010.

[15] C. Jensen, J-B. Rouquier, N. Ovtracht, and C. Robardet, "Characterizing the speed and paths of shared bicycle use in Lyon," Transportation Research Part D, 15: 522-524, 2010.

[16] L. Wu, Y. Zhi, Z. Sui, and Y. Liu, "Intra-Urban human mobility and activity transition: evidence from social media check-in data," PLoS ONE, 5: e97010, 2014.

[17] S. Gambs, M-O. Killijian, and M.N.d.P. Cortez, "Next place prediction using mobility Markov Chains," MPM'12 Proceedings of the 1st Workshop on Measurement, Privacy, and Mobility, 3, 2012.

[18] E. Cóme, A. Randriamanamihaga, and L. Oukhellou, "Spatio-temporal usage pattern analysis of the Paris shared bicycle scheme: a data mining approach," Transport Research Arena, Paris, 2014.

[19] O. O'Brien, J. Cheshire, and M. Batty, "Mining bicycle sharing data for generating insights into sustainable transport systems," Journal of Transport Geography, 34: 262-273, 2014.

[20] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs, "Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system," Pervasive and Mobile Computing, 6(4):455-466, 2010.

[21] A. Randriamanamihaga, E. Come, L. Oukellou, and G. Govaert, "Clustering the V´elib' origin-destinations flows by means of poisson mixture models," ESANN proceedings, Computational Intelligence and Machine Learning, 2013.

[22] M. Zignani and S. Gaito, "Extracting Human Mobility Patterns From GPS-based Traces," IEEE, 978-1-4244-9229-9/10, 2010.

[23] D. Brockmann, "Following the Money," Pysics World, 2010.

[24] S-M. Qin, H. Verkasalo, M. Mohtaschemi, T. Hartonen, and M. Alava, "Patterns, Entropy, and Predictability of Human Mobility and Life," PLoS ONE, 7(12): e51353, 2012.

[25] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the Limit of Predictability in Human Mobility", Scientific Reports 3: 2923, DOI: 10.1038/srep02923.

[26] V. V. Mayil, "Web navigation path pattern prediction using first order Markov Model and Depth first Evaluation," International Journal of Computer Applications (0975 – 8887), 45:16, 2012.

[27] A. Asahara, A. Sato, K. Maruyama, and K. Seto, "Pedestrian-movement Prediction based on Mixed Markov-chain Model," In Proceedings of the 19th International Conference on Advances in Geographic Information Systems, pages 25-33, USA, 2011.

[28] J. Lin, K. Zhao, B. Kusy, J. Wen, and R. Jurdak, "Temporal embedding in convolutional neural networks for robust learning of abstract snippets", arXiv.1502.05113, 2015.