THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Spatiotemporal Analysis of Urban Mobility Dynamics: A Case Study of Bicycle Sharing System

Ida Bagus Irawan Purnama

MSc in Information Technology,

The University of Nottingham, UK

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2017

School of Information and Electrical Engineering

Abstract

Understanding human mobility dynamics is of fundamental significance for many applications, and a wide range of data-driven mobility studies have been conducted using different datasets. Mobility traces which provide digital records of individual mobility allow analysis of individual mobility patterns, trends, and anomalies. Bicycle Sharing Systems (BSS) with origin-destination (OD) sensing systems that record departure and arrival times of each trip are among the most promising urban transport systems which do have such digital data available. BSS allow users to choose their own origin, route, and destination as well as travel time based on their needs. This flexibility leads to uncertainty on the operator side in terms of system use, and this thesis explores both uncertainty and regularity in demand to gain new insights for improving BSS deployments, services, and operations. Using BSS data from two cities (London and Washington DC), this thesis focuses on three main topics: station neighbourhood analysis, individual next place prediction, and prediction of system demand from system-level to individual station-level.

Stations neighbourhood analysis aims to reveal the quality of connections among nearby stations by examining users' behaviour in choosing other stations when their commonly visited station is disturbed because it is of out of service (shutdown) or in an imbalanced state (full or empty). Two methods are proposed to conduct this analysis which are spatial-mobility-motifs and station temporary shutdown. Two metrics are also proposed to measure the quality of connections which are impact distance and usage transformation. Results show that 300 metres of travel distance is the impact distance of a station shutdown as measured by at least 20% usage change for nearby stations. 300 metres is also the most common distance that appears during motif analysis when users choose nearby stations within a neighbourhood. Results from these both analyses could be used to help BSS operators identify potentially ineffective stations and isolated stations. 300 metres can also be used as a standard distance between stations when deploying a new system or redesigning the existing network topology.

User clustering aims to group users with similar mobility behaviour. Information theory is then used to measure the next-location predictability of each cluster. The goal is to identify highly predictable users so that useful services might be offered based on their predicted next place. Two temporal clustering metrics are proposed which are total trips (1 feature) and hourly trips across the day (24 features). These metrics adequately reflect the frequency and the regularity of user mobility. Three clusters are identified with distinct spatiotemporal characteristics which are named *casual users*, *regular users*, and *commuters*. Entropy analysis demonstrates that

commuters follow the basic entropy ordering rule that more history provides more predictability. Since real entropy is close to the conditional entropy for commuters, this suggests that the next location is strongly determined by the previous sequence of stations. Predictability, which is the theoretical upper bound of prediction accuracy, is approximately 80% for commuters. The accuracy of predicting destination given trip origin and user information is analysed at different times and for different clusters, using first and second order Markov models. Using previous trip history enhanced with aggregate data for trips without individual history, the highest prediction accuracy of 80% is achieved for commuters during the morning peak hours. Similar approaches are employed for return-to-next-pickup prediction, but their accuracy is less than the pickup-to-return accuracy. Trip prediction information could be used for a *user-based notification system* that can proactively notify highly predictable users in advance about information relevant to their likely destination.

Aggregate BSS usage at system level follows a regular daily and weekly pattern, combining commuting behaviour with recreation use. Being able to predict system-wide usage can enable better planning of redistribution and maintenance activities by operators. Rather than predict system-wide use for each hour directly, it is conjectured that greater prediction accuracy can be gained by predicting the deviation from the regular weekly pattern. Results show that the deviation-based prediction using machine learning predictors can significantly improve the prediction performance for both London and Washington DC in comparison with naïve approaches based on recent historical averages. Accuracy is also significantly improved compared to previously published BSS machine learning predictors. The RRMSE results from the best predictor are 16.9% in London using Bayesian Ridge Regression and 16.7% in Washington DC using Random Forest Regression for a week of validation data. Using these same predictors over two weekly test sets achieves 13.8% and 14.1% in week 1, and 27.5% and 22.7% in week 2, which is an anomalous week before Christmas. In all cases these results are much better than historical average prediction. The most important input features are the one-previous-hour deviation, followed by the two-previous-hour deviation. The effect of weather is already present in the previous hour inputs, and so separate weather inputs do not add much additional prediction information. Station-level prediction has significant error across a whole day, but predicting peak hour use in busy stations is much better than using historical averages, and this could help BSS operators to better predict unexpectedly heavy use of certain stations at certain times.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

I. B. I. Purnama, N. Bergmann, R. Jurdak, and K. Zhao, "Characterising and Predicting Urban Mobility Dynamics by Mining Bike Sharing System Data," in *the 11th IEEE International Conference on Ubiquitous Intelligence and Computing (UIC)*, pp. 159-167, 2015, DOI: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.46

Publications included in this thesis

No publications included.

Contributions by others to the thesis

Thiago de Sousa Silveira, an intern at CSIRO in 2014, authored the Python coding of the waypoint distance calculations between origin and destination BSS stations, which is used in section 4.3.1 of the thesis.

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

Acknowledgements

Firstly, I would like to express my sincere gratitude and immeasurable appreciation to my principle supervisor, Prof. Neil Bergmann (UQ), and all my co-supervisors from CSIRO, Prof. Raja Jurdak, Dr. Branislav Kusy, and Dr. Kun Zhao for their continuous support, immense knowledge, technical insights, word of encouragement, valuable feedback, and writing advice. Their guidance helped me in all the times of my study to achieve all the research milestones.

Besides my supervisors, I would like to thank Thiago de Sousa Silveira who helped me to solve the Python code for the waypoint distance calculation.

My special thanks also goes to DIKTI that provided me full scholarships, my institution (PNB) that let me go to pursue my PhD degree, and UQ as well as CSIRO that provided me excellent research facilities and study environments.

I would also like to thank my research group colleagues, house mates, and Diktiers for all discussions, companionship, and fun that we have had in the last three and half years. You made most of my times joyful.

Last but not the least, I would like to thank my lovely family in Bali for supporting me throughout my years of study. This accomplishment would not have been possible without all of you. Thank you!

Keywords

urban mobility, bicycle sharing system, mobility characterization, clustering, entropy, predictability, next-location prediction, Markov chain model, cyclostationary-based prediction, machine learning

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 080504, Ubiquitous Computing 20%

ANZSRC code: 080607, Information Engineering and Theory, 80%

Fields of Research (FoR) Classification

FoR code: 0805, Distributed Computing, 20%

FoR code: 0806, Information Systems, 80%

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ABR	Adaboost Regressor
AI	Artifial Intelligent
AR	Auto Regressive
ARIMA	Auto-Regressive Integrated Moving Average
ARMA	Auto-Regressive Moving Average
ARMSE	Average Root Mean Squared Error
AVG	Average
BN	Bayesian Network
BPNN	Back Propagation Neural Network
BRR	Bayesian Ridge Regressor
BSS	Bicycle Sharing System
CDF	Cumulative Distribution Function
CDR	Call Details Record
CTRW	Continous Time Random Walk
DTW	Dynamic Time Warping
EM	Expectation Maximization
ER	Error Rate
GA	Genetic Algorithm
GBDT	Gradient Boosting Decision Tree
GBM	Gradient Boosting Method
GBRT	Gradient Boosting Regression Tree
GIS	Geographic Information System
GMT	Greenwich Mean Time
GPS	Global Positioning System
GTBR	Gradient Tree Boosting Regression
HA	Historical Average
HMM	Hidden Markov Model
HP-KNN	Hierarchical Prediction K-Nearest Neighbour
HP-MSI	Hierarchical Prediction Multi Similarity-based Inference
HT	Historical Mean
iBPNN	improved Backpropagation Neural Network
KM	K-Means
LCHS	London's Cycle Hire Scheme
LDA	Latent Dirichlect Allocation
LMO	Louvain Modularity Optimization
LV	Last Value
MA	Moving Average
MAE	Mean Absolute Error

MAPE	Mean Absolute Percentage Error
MCD	Maximal Clique Detection
ME	Mean Error
ML	Machine Learning
MM	Markov Model
MMC	Mobility Markov Chain
MMM	Mixed Markov Model
MSE	Mean Squared Error
NAB	Normalized Available Bikes
NAS	Normalized Activity Score
NN	Neural Network
OD	Origin-Destination
PCC	Pearson's Correlation Coefficient
PDF	Probability Density Function
PR	Preferential Return
QoS	Quality of Service
RF	Random Forest
RFID	Radio Frequency Identification
RFR	Random Forest Regression
RMSD	Root Mean Squared Difference
RMSE	Root Mean Squared Error
RMSLE	Root Mean Squared Logarithmic Error
ROG	Radius of Gyration
RQ	Reseach Question
RR	Ridge Regression
SA	Simulated Annealing
sIB	sequential Information Bottleneck
SLD	Straight Line Distance
sMAPE	symmetric Mean Percentage Error
STD	Standard Deviation
STDBSCAN	Spatiotemporal Density-Based Spatial Clustering of Applications with Noise
SVM	Support Vector Machine
SVR	Support Vector Regression
SWT	Sliding Windows Technique
TGAMs	Two-stage Generalised Additive Models
WPD	Waypoint Distance
WRMSE	Weighted Root Mean Squared Error
WT	Waiting Time

CHAPTER 1
INTRODUCTION

Understanding human mobility dynamics is indispensable for a range of applications from urban planning [1], traffic forecasting [2] and transit systems [3] to public health [4, 5], epidemic prevention [6, 7], emergency response [8], and location-based services [9]. Human mobility modelling is possible because humans naturally move with a certain degree of spatial and temporal regularity in their daily routines. On the other hand, human mobility also contains a degree of irregular or random movement. Individuals might explore unfamiliar places, follow a new route, use different travel methods, or they may move in familiar places but at unusual times. High regularity of movement equates to high predictability, while high randomness brings high uncertainty or entropy. There are multiple complex, interrelated factors that affect both the regularity and the randomness of human movement. Therefore, understanding the regularity of human mobility is a challenging problem [10]. Improving the prediction of individual and population mobility has wide potential applications in the areas mentioned above.

Urban populations need to be served by effective and efficient transportation systems, such as roads, cycle paths, public transport and parking facilities, to support activities such as commuting to work, shopping, leisure and tourism. The pulse of urban activities that reflects their underlying spatial and temporal characteristics can be inferred from human mobility dynamics associated with those activities [11]. The study of urban mobility dynamics involves understanding where, when, and how citizens move at city-wide scale and at subregion levels. If individual or group mobility patterns can be captured appropriately, these can be analysed to provide insights about urban mobility patterns at different spatiotemporal scales. This task entails capturing records of individual mobility in order to properly analyse their patterns, trends, and anomalies. Unfortunately, not all transportation systems enable regular capture of such data, and some of them only provide aggregate data on fixed routes and schedules without capturing the fine-grained individual mobility behaviour.

Bicycle Sharing Systems (BSS) with origin-destination (OD) sensing systems that record the departure and arrival times of a one-way individual trip are among the most promising urban transportation systems which have such fine-grained mobility data available. BSS are a subtype of on-demand transport networks that include taxis, hailing services (e.g. Uber), and

ridesharing (e.g. carsharing). Unlike conventional public transport systems (e.g. subways and buses) with fixed routes, schedules, and transit stops, BSS allow users to choose their own route and schedule [12]. Compared to taxis which also have that flexibility, BSS are more individualised because taxis can carry a group of people and need drivers, but taxis are more flexible in terms of origin and destination. BSS also are either faster or competitive with taxis in terms of travel time in dense urban areas [13]. Accordingly, BSS trips are well aligned with inner city travellers' mobility.

This flexibility obviously brings advantages and challenges such as high efficiency on the user's side as well as uncertainty on the operator's side [14]. This uncertainty arises because users may pick up and return their rented bikes whenever and wherever they want. However, as humans tend to move in certain regular patterns on a daily basis, the likely system use could potentially be predicted from the movement behaviour history which is embedded in users' previous trip data. This could be further understood by considering some external factors which spatially and temporally align with that trip data, such as local weather. In addition, uncertainty is involved not only in when and where pickups and returns occur, but uncertainty also comes from the individual routes which are followed by users. It is very hard to trace user trajectories between stations because BSS are not usually equipped with GPS (Global Positioning System) trackers. Furthermore, the uncertainty also arises when a station faces a perturbation (e.g. temporary shutdown), or when it is in an imbalanced state (either full or empty). How users respond to such circumstances, what the impacts for other stations are, how to properly measure this impact, and how to use this impact knowledge to improve BSS operations, are all questions where there are not clear answers.

Recently, most BSS research studies have conducted their analysis and prediction at an aggregate level [15-17] to observe the global and local trends, for example at city and cluster scale, since almost all BSS public datasets contain information about trips, but these are not linked to individual users. This study uses what we believe is the only public BSS dataset that provides information on (anonymized) individual users. It covers approximately 6 months of BSS system use in London in 2012. In addition, this study will also investigate prediction of aggregate BSS system use using some new techniques that will be shown to significantly improve prediction accuracy. Three major studies will be undertaken.

First, relationships between neighbouring stations will be used to understand spatial characteristics of BSS, such as how temporary closure of a station affects its neighbours.

Second, different techniques will be used to explore the predictability of individual user mobility. Third, different machine-learning techniques will be explored to predict the system-wide BSS usage, the usage in neighbourhood clusters, and the usage at individual stations. The motivations of these investigations are first to improve the design of BSS systems by using the spatial insights, second to be able to identify predictable users in order to provide useful advice and assistance, and third to assist BSS operators to better predict unusual usage patterns and to plan responses to these. As well as these practical BSS motivations, it is also expected that the BSS data analysis will enhance our existing understanding of human mobility patterns in general.

The remainder of the thesis is organized as follows. Chapter 2 presents the literature review that critically reviews existing work related to human mobility and BSS analysis. Chapter 3 explains the research gaps, research questions, and research tasks in detail. Chapter 4 presents the dataset pre-processing and spatiotemporal preliminary data analysis. Chapter 5 discusses the results of station neighbourhood ties analysis. Chapter 6 discusses the user clustering and next-place prediction results. Chapter 7 discusses the results of deviation-based prediction over the daily and weekly patterns of BSS data. Chapter 8 presents the conclusions, original contributions and describes possible future work.

CHAPTER 2
LITERATURE REVIEW

There has been a long history of interest in human mobility, but the difficulty has always been in how to monitor the movements of humans. Trip data from BSS databases is one relatively new way to track human movement, and there has been significant recent interest in using BSS data to study mobility. Additionally, the recent explosion of interest in Data Analytics caused by the availability of big data sets means that researchers are also interested in how Data Analytics can improve BSS operations. So this chapter reviews the literature corresponding to two major topics: ***human mobility studies*** and ***BSS studies***. The human mobility review section will present the generic characteristics, models, metrics, limitations, and predictability of human mobility. These studies have been done using various sources of human mobility data, as well as proxies in which human location and movement is approximated by the movement of devices (such as phones) or artefacts (such as banknotes). After first providing an overview and some history about BBS, the BSS review section will mainly discuss research that relates to the spatiotemporal analysis and prediction of BSS data.

Even though human mobility studies have been conducted using a wide variety of data sources which have different characteristics, there are some generic mobility metrics and methods that will be applicable for studying BSS mobility. By reviewing BSS studies alongside other human mobility studies, the gaps can be identified where human mobility metrics and methods that have not implemented yet in BSS studies. How these methods can potentially improve BSS services, deployment, and operation will be able to identified, so that the research questions and methodology can be formulated in the next chapter.

2.1. Human Mobility Studies

The majority of studies of human mobility exploit the high degree of regularity and predictability of future locations of individuals where movement ranges are mostly dictated by daily routine [18, 19]. To understand the nature of human mobility dynamics, a broad range of data-driven studies have been conducted. As synthetic data has limited scope to capture the fine detail of real human mobility [20], most recent mobility studies have been driven from various sources of real world data. These data mobility traces use data such as banknote tracking [21, 22], call detail records (CDR) of mobile phones [23-27], taxi data [28, 29],

railway system data [30, 31], transit system and smart card data [3, 32], GPS-based traces [29, 33], and social media with geo-tagging [9, 10, 34]. Such digital information reflects the daily mobility activity in certain ways that correspond to the visited locations of proxies or individuals at specific times [35]. For instance, the mobility of banknotes corresponds to the geographic circulation of notes from person to person, while the mobility from CDR analyses reflects the mobile-phone position in terms of the nearest cellular base station. Kang et al. [36] summarized three desirable traits of mobile positioning: *large sample size*, *high temporal and spatial resolution*, and *high spatiotemporal dynamics*. Although the available datasets significantly differ in their features, granularity, and resolution, the results agree on a number of quantitative characteristics and metrics of human mobility [26]. For example, mobile-phone and banknote studies both result in a power-law distribution of distance travelled.

This section will review some human mobility topics related to this research, namely displacement distribution, waiting time, radius of gyration, preferential return, mobility motifs, entropy and predictability, and mobility prediction. Later in the thesis, in Chapter 4, these same analyses will be applied to BSS data to investigate whether such analyses are able to provide new insights and understanding of BSS usage.

2.1.1. Displacement Distribution of Human Mobility

Individual human trajectories are generally characterised by heavy-tailed distributions, a distribution with a "tail" that is heavier than an exponential, that show the complexity of human mobility [23, 26]. The heavier the trajectory tail is, the larger the probability of getting one or more very large values in its distribution is. Using dollar-bill tracking as a proxy from the *WheresGeorge* online game, Brockmann [21] studied the scaling law of human mobility. This work revealed that the probability of a bill traveling a certain distance within a certain time falls as an inverse power law $P(r) \sim 1/(r^{1+\mu})$, where $\mu = 0.6$. The consecutive geographical mobility of a bank note is similar to a class of random walks known as *Lévy flight* in which the probability distribution of step-length is heavy-tailed.

In another study, Song et al. [26] used mobile-phone traces in CDRs and defined *jump size* Δr as the displacement between consecutive locations showing the distance travelled by an individual. The probability $P(\Delta r)$ has a heavy-tailed characteristic, $P(\Delta r) \sim |\Delta r|^{-1-\alpha}$ where $0 \leq \alpha \leq 2$. They also suggested the relevance of *Lévy flight* or *continuous-time random-walk* (CTRW) models for human mobility [26]. This CTRW model is widely used in the random-walk

community [22]. However, real human trajectories are not random and do not follow highly reproducible scaling laws. This is because humans have a significant propensity to return to the locations they visited frequently before, such as their home, workplace, recreation area, or shopping centres. Good mobility models should describe the recurrence and temporal periodicity inherent to human mobility, in contradiction with random walk models such as CTRW. Random methods alone cannot be the basis of a modelling framework which captures the basic features of human mobility [26].

Barbosa et al. [37] investigated the characteristics of human trajectories by exploring mobile phone position data and the Brightkite[1] data in Brazil using a rank-based approach of visited locations. They proposed two rank variables which are the *frequency* rank and the *recency* rank. Both of those ranks were measured from the accumulated sub-trajectories. Since the authors had an interest in individual trajectories, they only considered the data that corresponded to the user's displacement by filtering the recurrent observations in one location. For each individual, this produces a time series of trajectories through the observation period. Based on their observations, they concluded that human trajectories are biased towards recently visited locations [37].

The limitations of using mobile positioning are analysed by Kang et al. [36]. They suggested that a good positioning device should collect individual's geo-position continuously through time. However, mobile-phone mobility data only contains position information when a communication using that device happens. Therefore, the extent to which actual mobility can be represented and revealed from mobile phone data needs to be tested appropriately. Their results show that, although the mobile trajectories as a sampling of real trajectories have a lot of missing detail, they can be used to estimate the actual profiles of individual movement over a long time period.

Meanwhile, Wu et al. [9] used 15 million social media check-in data to construct the displacement distribution of individuals in Shanghai, China with area of observation of 50 km x 35 km. They assumed the observed area to be divided into square lattices (500 x 500 metres). They combined the movement-based approach with the activity-based approach to reproduce intra-urban mobility. This is possible because check-in data is able to indicate the travel purpose of users by demand-tags that are mostly associated with the venue where they are

[1] Brightkite was a location-based social discovery networking launched in 2007 and closed in 2011.

checked in. Then, the authors implemented an agent-based modelling mechanism that produces simulated patterns which fit well with the real distribution of observed movements.

Using GPS-based traces, Zignani and Gaito [33] were able to extract common points of interest, called geo-locations. From those geo-locations, they offered a definition of geo-community which describes the spatial and social context relations of human mobility. Then, they conducted a statistical analysis to show the fundamental qualities of human movements. Because the GPS points are not spread homogeneously, they defined different types of locations by observing the inclination of GPS points that tend to meet in few regions. They applied two clustering methods, namely *the density-based clustering* method and *the hierarchical agglomerative* method. This analysis identified the distance distribution covered by individuals both within and between geo-locations including the pause or waiting time. They found that the hierarchical clustering method performed better.

All these works from different source of data suggest generic spatial displacement characteristics in human mobility which are heavy-tailed distributions and not random. They follow a certain quality of regularity and are biased towards recently visited locations because people have a tendency to return to the locations they visited frequently before. Users also have common points of interest, and certain waiting times as described in the next section.

2.1.2. Waiting Times Between Mobility

Another quantity for describing the heavy-tailed distributions of human mobility is *waiting times* $P(\Delta t)$ defined as the time a user spends at one location that shows the time between a displacement and the next displacement [26], or a time between consecutive trips that are expected to vary across individuals [30]. Again, using CDR data, Song et al. [26] depicted the heavy-tailed distributions of waiting times as $P(\Delta t) \sim |\Delta t|^{-1-\beta}$ where $0 \leq \beta \leq 1$. However, Hasan et al. [35] found that this distribution of waiting time is not generally true for all types of locations.

Similarly for other mobility data, waiting times can be an idle time between calls [23] if a mobile phone is used as a mobility proxy, or it can be the time when a bank note is saved by an individual before it transferred to others [22], or it can be the time for a taxi driver to wait for the next passenger [28], or vice versa the time for a passenger to wait for a taxi to arrive [29]. This waiting time distribution may reveal useful insights about mobility patterns. For example,

a larger waiting time of taxi passengers suggests a lower availability of taxis [29]. In contrast, a shorter waiting time of taxi drivers means a higher availability of passengers.

Another waiting time analysis was conducted by Barbosa et al. [37], but in this study they observed the time interval (in hours) between visiting the same location. They found two important features about human mobility characteristics. First, peaks are experienced at intervals of 24 hours. This captures that temporal regularity where humans tend to revisit to the previously visited places as part of their daily routines. Second, that return probability shows very rapid decays as the time increases. They presented a different outlook for human mobility examination in which this temporal aspect plays a much more important role than the inter-event times [37]. Papalardo and Simini [19] stated that the waiting time is the temporal mechanism showing the distribution of time between two successive journeys. However, it does not model the tendency of human to be in certain locations at specific times.

2.1.3. Radius of Gyration

Another feature for describing the complexity of human mobility that also follows a heavy-tailed distribution is *radius of gyration* (RoG). It is understood as the characteristic distance covered by an individual when observed up to certain time [23]. In other words, it describes the characteristic travel distance of an individual in a certain time period, usually on a daily basis [24]. It is formulated as $RoG = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\overline{r_i} - \overline{r_{cm}})^2}$ where r_i is the i^{th} position recorded for an individual user, $i = 1,\ldots,N$, and $\overline{r_{cm}} = \frac{1}{N}\sum_{i=1}^{N} \bar{r}_i$ is the geometric centre of the trajectory [23]. Using 6 months observation (T) of cellular-phone data, Gonzales et al. [23] classified the RoG of individuals into three categories which are mostly small ($RoG_{small}(T) \leq 3$ km), medium ($20 \leq RoG_{medium}(T) \leq 30$ km), or large ($RoG_{large}(T) \geq 100$ km). The RoG adheres to a power-law distribution with an exponential cut-off. Since the RoG follows a heavy-tailed distribution [24], this indicates that, even though most of the individual travels are confined to less than 3 km, there are a few users who regularly travel hundreds of kilometres. Similarly, Song et al. [26] found the growth (as the time interval increases) of radius of gyration was very slow and that it also follows a heavy-tailed distribution.

Later, by assuming that individual travel speed is constant and that individuals have fixed commonly-visited locations such as home and workplace where they spend most of their time, Xiao-Yong et al. [38] calculated RoG and also showed that the typical area of individual daily

movement is an ellipse. It is skewed if the travel distance is increased [25]. They conducted their analysis using *Mobidrive* data which is a travel diary that records travel behaviour of 360 people in Germany day by day over a 6 week survey period. They simplified the daily travel of individuals into three subsequent activities which are commuting from home to workplace, going to leisure activities, and returning back home. They found that most people have similar orderings of activities even though the times and leisure venues vary.

For such human mobility quantities above, the probability distributions of *jump size* P(Δr), *waiting times* P(Δt), and *radius of gyration* P(RoG) show a heavy-tailed distribution characteristic where human mobility patterns are mostly concentrated in a region of a few kilometres for certain time durations. There are a few individuals who travel much further, and also a few individuals who wait much longer than the normal waiting times for their next movement. These few outliers result in distributions having heavier tails than a simple power law distribution. Furthermore, in terms of spatial context, some regions have unique spatial ties to other regions which could vary over time. This suggests the complexity of spatiotemporal human mobility patterns cannot be fully predicted by straightforward rules or models.

2.1.4. Preferential Return

As individuals tend to visit similar places as part of their daily routine, the concept of *preferential return* (PR), proposed by Song et al. [24], offers a well-designed model for the visitation frequency distribution for returning to previously visited locations. On the other hand, they also identified *exploration* for visiting a new location. In preferential return, they defined the probability Π_i for returning to a location i as $\Pi_i \propto f_i$, where f_i is the frequency of visitation to that location [24]. This PR and exploration reproduce a scaling property of human mobility in which more visits will occur if a location is discovered earlier [39].

Incorporating a recency-based mechanism by including a bias towards recently visited locations, Barbosa et al. [37] proposed an extension for the preferential return mechanism with a temporal perspective. They tested the respective relationship of the probability of return using a different rank analysis. They claimed their approach is based on an experiential proof that if the time of last visit to a location is longer, then the probability of finding that user in that location is lower. In other words, a user has tendency to return to recently visited locations. Furthermore, they suggested that the probability of visitation to specific place is proportional to the number of previous visitations to that place.

Using both vehicle tracked GPS and mobile phone data, Papparaldo et al. [18] identified two distinct classes of individuals: *explorers* and *returners*. They claimed that existing models cannot describe the existence of these two classes. Then, they proposed what they claimed is a more realistic model that would be able to capture the empirical findings of those two classes. They used RoG to understand how the *k*-th most frequent places of an individual govern the characteristic distance covered by that person. The role of the *k*-th most frequent places was investigated by comparing the probability distribution of RoG_{total} and RoG_k where k = 2,….,10. The correlation between *k*-th RoG and total RoG lets them measure the level of similarity between recurrent and overall mobility patterns. They found that populations are split into two typical classes. Returners limit their mobility to a few locations, and their recurrent patterns are comparable to the overall ones. Instead, explorers cannot be restricted to limited locations.

2.1.5. Human Mobility Motifs

Human mobility can also be characterized by the trips among a sequence of visited places [40]. As humans mostly move in daily routines, a daily mobility motif can be defined as the equivalent spatial class of directed network [41] that represents the traces of those visited locations on the daily basis. A directed graph is an ordered pair *G = (V, E)* where *V* is the set of nodes (or vertices) representing BSS stations, and *E* is a set of ordered pairs of nodes (i.e., directed edges) representing trips. This exhibits a unique daily trace of individuals from one location to other locations during a day.

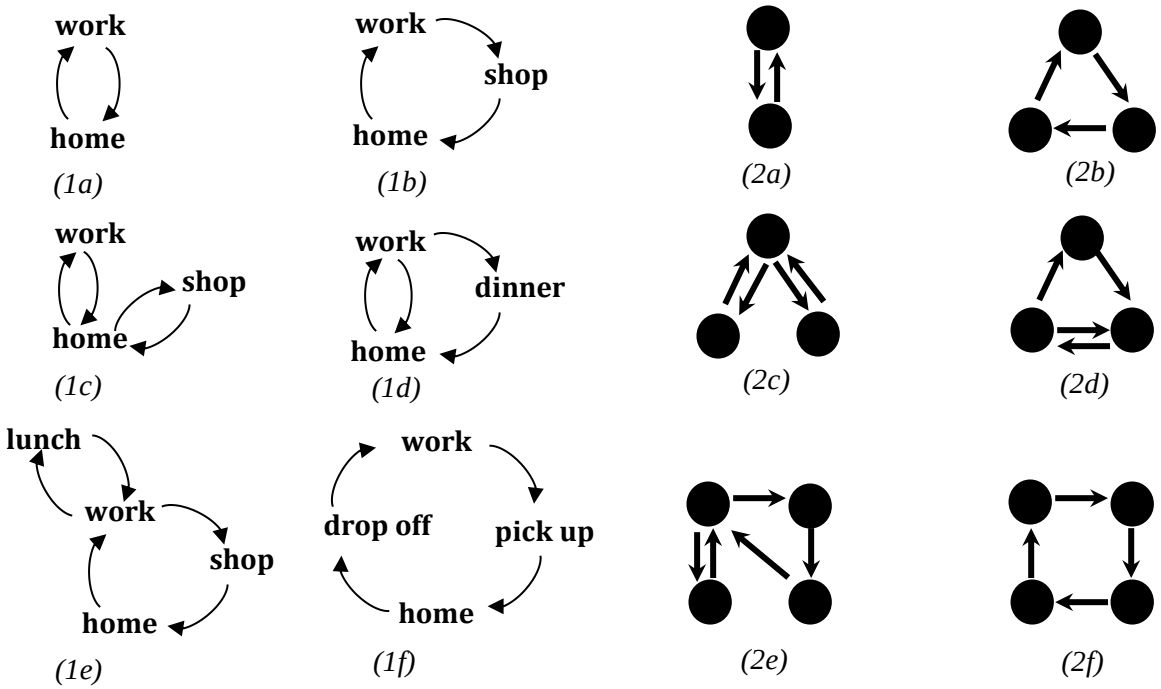


Figure 2.1. Example of daily mobility motifs in real world ***redrawn*** *from Jiang et al. [41].*

Among a population, there should be similar daily motifs as individuals have similar common places to visit (work, home, shop, etc.). Figure 2.1 shows the real world activity pattern structures (1a-1f) with the corresponding highly abstract daily motifs format (2a-2f) [41]. Schnieder et al. [40] used mobile phone and survey data to find the mobility daily motifs of individuals in Paris and Chicago. Using 0.5% occurrence as a minimum threshold that should appear in the dataset, they found 17 unique networks that represent motifs, as shown in Figure 2.2. This is already sufficient to capture up to 90% of mobile phone and survey population in both cities.

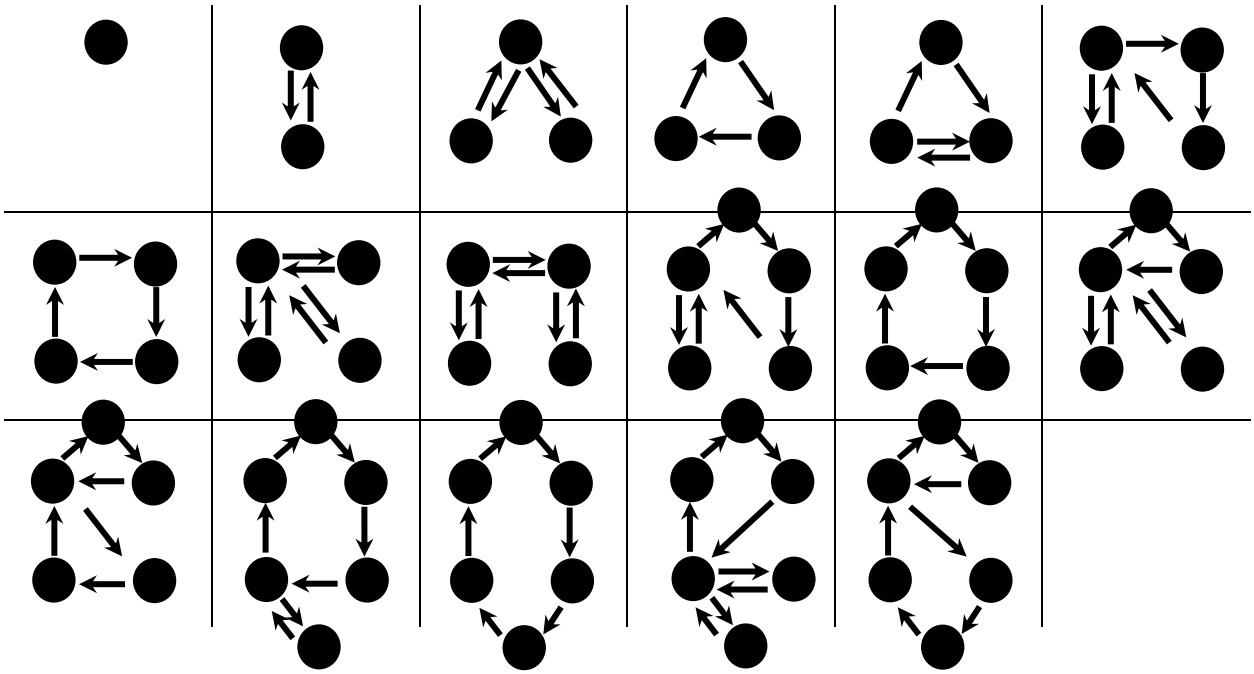


Figure 2.2. Daily mobility motifs in Paris and Chicago ***summarised*** *and* ***redrawn*** *from Schnieder et al [40].*

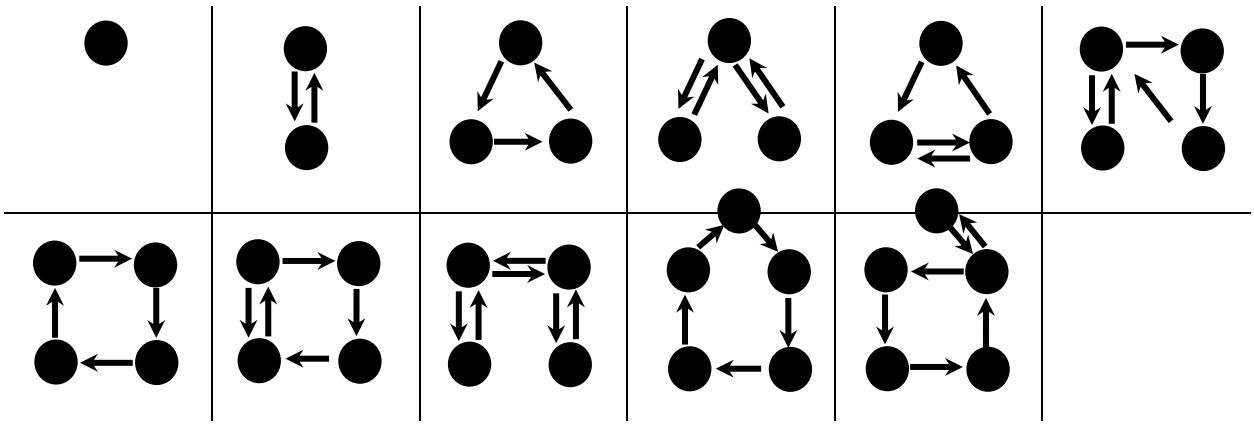


Figure 2.3. Daily mobility motifs in Singapore ***summarised*** *and* ***redrawn*** *from Jiang et al. [41].*

Later, Jiang et al. [41] adopted a similar approach using Singapore CDR cell-phone as well as survey data to uncover the Singaporean residential mobility motifs. The found from the phone data that on an average weekday, the most frequent motif is that 33% of Singapore residents visited 2 places, followed by 30% 3 places, 14% 4 places, 13.5% stayed at home (1 place), 5.5% 5 places, 2.1% 6 places, and less than 2% visited more than 6 places, as shown in Figure 2.3. These motifs cover around 90% of the population. While for survey data in that same study, 2-nodes is the most dominant motifs with 55% of the population.

To the best of our knowledge, this motif analysis has not been investigated yet with BSS data. Popular sequences of visited BSS stations for individual users on a daily basis are not known. Use of this spatial motif analysis in BSS design and deployment may have potential to assist in BSS system operations, and the potential of this analysis is worthy of further investigation.

2.1.6. Entropy and Predictability

Being able to predict a traveller's next location from their current location would allow useful information to be relayed to the traveller about their travel. The usefulness of this information will depend on the accuracy of the next location prediction. This section explores fundamental concepts about trip predictability.

In information theory, *Entropy* (S) is a fundamental quantity to measure the uncertainty or randomness of movement, and it can be used to capture the degree of predictability. Entropy summarises the information that is present in the sequence of locations, characterising a time series [42]. Theoretically, there are four different measurements of entropy, *Random entropy* (S^{Rand}), *Shannon* entropy (S^{Shan}), *Conditional entropy* (S^{Cond}), and *Real entropy* (S^{Real}). All of those will be bound by the relationship: $S^{Real} \leq S^{Cond} \leq S^{Shan} \leq S^{Rand}$. Random entropy captures the randomness of mobility by considering only the number of distinct locations visited by a user. This means each location is considered as having an equal probability. Shannon entropy, also known as temporal-uncorrelated entropy, counts the probability of visiting each distinct station. This demonstrates the heterogeneity of visitation patterns. Conditional entropy captures the correlation between one location and the subsequent location in the time series. This considers frequency and the order in which the locations were visited. Real entropy fully captures the spatiotemporal order that presents in user mobility, not only the frequency and

order but also the time spent at each location. The detailed formulas for each of these entropies are given later in Chapter 6 when these are used for BSS analysis.

Predictability (Π) is the measurement of users' future whereabouts [24]. Fano's inequality is used to introduce Π_{max} as the *fundamental limit of predictability*. This is useful in the scenario where a random variable Y is known to estimate the value of a correlated random variable X. It relates the probability of error in estimating X to its *conditional entropy* S(X|Y). Here, it would predict correctly the user's next location based on the history of locations with a maximum probability of Π_{max}. The accuracy that can be attained by a predictability algorithm will be influenced by the inherent characteristics of the users' movement patterns [43].

A theoretical limit of predictability has been demonstrated in recent studies. Song et al. [24] posed 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 records. They found a 93% potential predictability in user mobility, Figure 2.4. This high percentage indicates that there is a huge potential to explore the regularities of human mobility using mobile phones. 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%, Figure 2.5. Similarly, using a smartphone-based study of 500 users in Finland, Qin et al [44] demonstrated that patterns and entropy relate to the degree of activities and locations with 78% predictability.

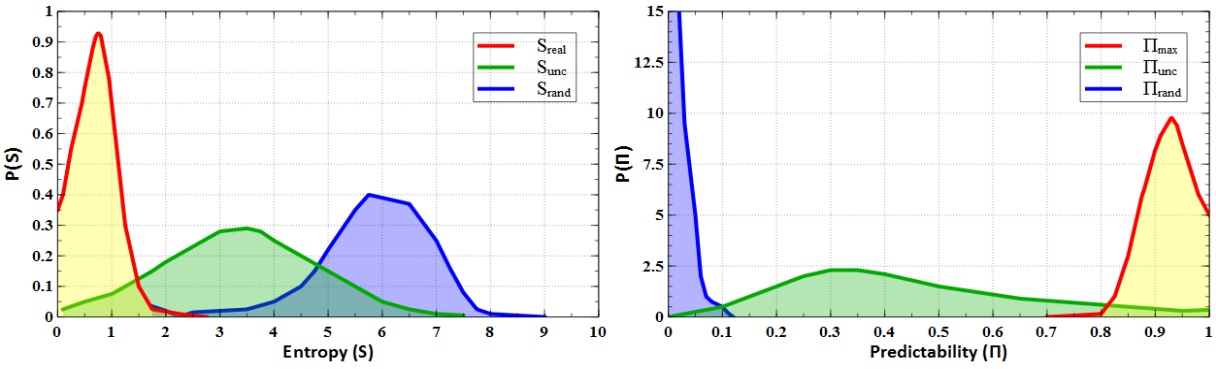


Figure 2.4. Entropy and predictability ***redrawn*** *from Song et al. [24].*

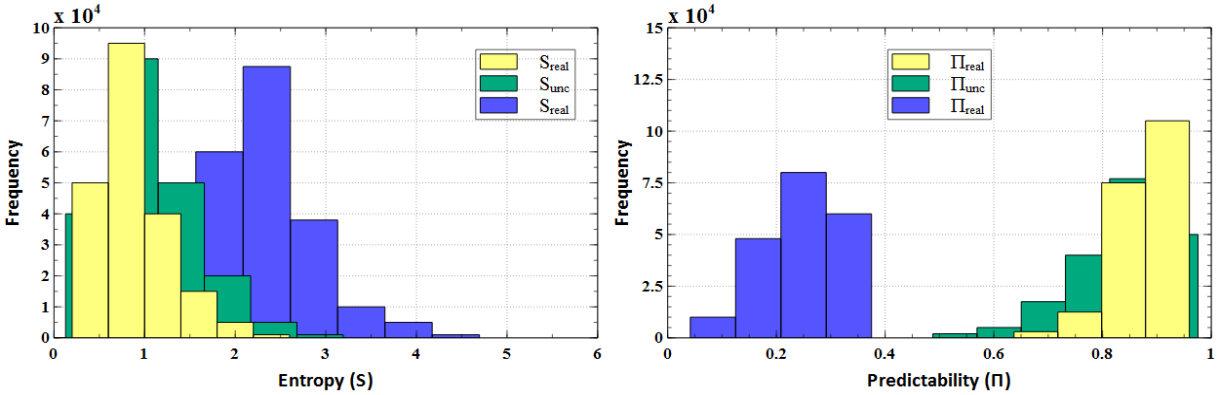


*Figure 2.5. Entropy and predictability **redrawn** from Lu et al. [25].*

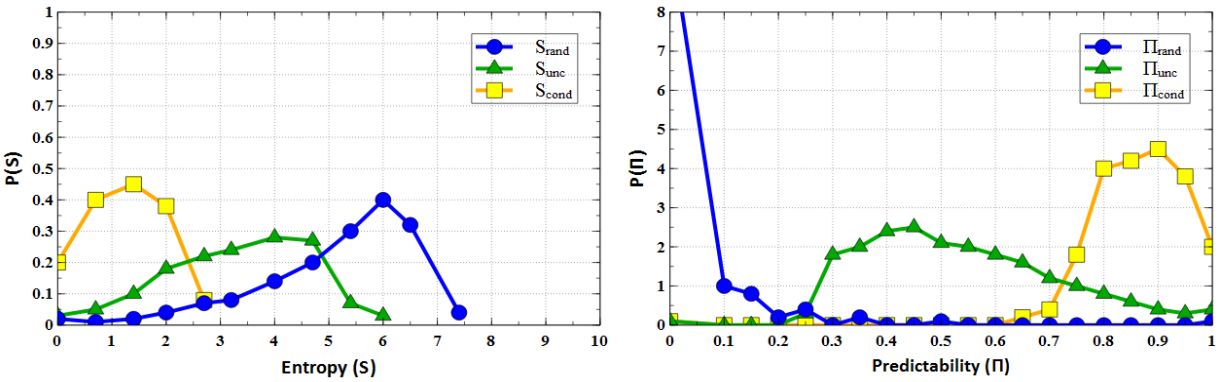


*Figure 2.6. Entropy and predictability **redrawn** from Sinatra and Szell [45].*

Recently, Sinatra and Szell [45] 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, the online world Pardus[2]. Here, individuals are not performing physical movements, but rather, navigate a virtual avatar. They found that movements in virtual human lives follow the same high levels of predictability as real world mobility, Figure 2.6. To some extent, the future movement of players can be predicted well if the temporal correlation of visited places is accounted for.

However, to the best of our knowledge, no studies have yet investigated the concept of entropy and predictability in BSS data, except our own work conducted as part of this thesis [46]. As a specific mode of transport in urban areas, it is expected that the predictability bounds of BSS data have unique features that are different from mobile phone data [46].

[2] www.pardus.at

2.1.7. Human Mobility Prediction

In human mobility prediction, numerous spatiotemporal prediction approaches for different mobility metrics have been proposed. A simple prediction technique for flows between locations uses the historical average at similar times in the past. Regression techniques use a parametric model to predict future locations based on the current and past system state, and the model parameters are often determined by machine-learning techniques, i.e., optimising the parameters to best fit past data. Many different regression models are described later in this section. Techniques from time series analysis, such as *Auto-Regressive Moving Average* (ARMA) and *Auto-Regressive Integrated Moving Average* (ARIMA) are often used. Mobility-specific parametric models, such as Gravity and Radiation models have been proposed. For individual next place prediction, Markov-chain models use a transition matrix to estimate the probabilities of the next system state based (i.e. next location) based on the current state. A prediction algorithm could be considered to be good when it shows a better prediction than the baseline historical average can achieve [47]. Prediction algorithms are used to predict various human mobility metrics such as traffic flows, commuting patterns, next location, travel times, passenger numbers, future mobility trends, and mobility classification.

This section reviews a number of studies in the prediction of different mobility metrics in different scenarios in order to understand the range of techniques used, the range of metrics that are predicted, the performance metrics used to evaluate the accuracy of the prediction, and the accuracies that are achieved for different situations.

In time series regression-based models, Li et al. [28] proposed an improved ARIMA model to predict the spatiotemporal variations of taxi passenger numbers in an extracted hotspot using 4000 taxis' GPS traces in Hangzhou, China, over a year. Their prediction performance achieves 5.8% error. When used to predict the location of the likely passengers, they can decrease the distance travelled and time taken by 6.4% and 37.1 % respectively. The key success of their scenario is the clustering of the pick-up and set-down events of passengers. ARIMA models assume the future value of a variable as a linear function of several historical observations with random errors. Later, Moreira-Matias et al. [29] proposed three distinct short-term prediction models, Time-Varying Poison, Weighted Time-Varying Poison, and an ARIMA model, as well as an adaptive (sliding windows) ensemble of time series models to predict the spatial distribution of taxi-passengers in order to improve taxi-driver mobility intelligence in Porto, Portugal. Their major contribution to the area is due to the adaptive

characteristics of their approach in streaming data, while other works mainly conduct their experiments using an offline testbed. Using streaming data from 441 taxis, first, a histogram series is made to aggregate the information, and then three time-series forecasting techniques were conducted to make a prediction. As a result, they can achieve a very good performance where the maximum value of error was 28.23%. Meanwhile, the sliding-window ensemble is always the best model where the prediction error that can be achieved was always lower than 26%.

Massuci et al. [48] tested the Gravity and Radiation models for commuting patterns as well as for public transportation flows in England and Wales at national level and city scale. The Gravity model observes the flows between origin and destination based on their distance and population where the flow is proportional to the product of the OD populations and inversely proportional to the power law of their distance. The Radiation model originates from a particle diffusion model with emission and absorption rates. The flow can be estimated by considering the population in the circular radius of OD where the circle centre itself is the middle point of OD. Overall, the gravity model shows better performance. They found that for large cities the original Radiation model underestimates the flows of commuting. After introducing a normalization factor to generalize the radiation model, a competitive result can be achieved. Meanwhile, Ren et al. [49] used a Radiation model to predict the commuting flows in spatial networks based on cost-based generalization using US census and highway traffic data. Compared with real traffic, they found that their model captures the normal distribution of the traffic flows. It achieves a high Pearson Correlation Coefficient (PCC), 0.75, based on travel time costs.

Asahara et al. [50] proposed a variant of a Markov model called the *Mixed Markov Model* (MMM) which is an extension of a standard *Markov Model* (MM) and a *Hidden Markov Model* (HMM) to predict pedestrian movement in Osaka, Japan. Based on their observations, the two previous models were not generic enough to encompass all types of mobility. The MMM was proposed due to the existence of similar mobility behaviour among certain pedestrians. They achieved 74.4% as the highest prediction accuracy. Later, Gambs et al. [51] adopted the concept of a *Mobility Markov Chain (MMC*) and extended it to n-MMC in order to incorporate the *n* previously visited locations to predict the next location. They used data from three dataset which are Phonetic (October 2009 to January 2011), Geolife (April 2007 to October 2011), and Synthetic. They found that the prediction accuracy of the next location is in the range of 70% to 95% as soon as n = 2.

Machine learning is a type of functional approximation [52, 53] based on a parametric model, where the model parameters are determined by examining training data consisting of sets of input features and their corresponding outputs. Typically, the parameters are chosen so that the error is minimised between the real outputs in the training set and the machine-learning estimation of those outputs based on the corresponding input features. In the area of machine learning techniques in mobility prediction, Zhang and Haghani [52] employed a *Gradient Boosting* regression tree method (GBM) to improve the freeway travel time prediction in Maryland, US, using 2012 RITIS (Regional Integrated Transportation Information System) data. Then, they compared their GBM performance with *Random Forest* (RF) and ARIMA methods. Here, GBM uses a boosting method to generate a decision tree sequentially in order to minimize certain loss functions and improve the prediction accuracy at the same time. Specifically, this improvement goal is done by introducing a new weak learner sequentially and putting emphasis on it to compensate the shortcomings of current weak learners. By this technique, they analysed the prediction performance from 105,408 freeway travel time records. Using *mean absolute percentage error* (MAPE) as a metric, their GBM solution performed better than RF and ARIMA. The results are 2.01%, 2.04%, and 2.03% MAPE for prediction 5 minutes ahead using GBM, RF and ARIMA respectively, 2.77%, 2.78%, and 2.90% for prediction 15 minutes ahead, and 2.82%, 2.85%, and 3.01% for prediction 30 minutes ahead.

Later, Lopez et al. [54] proposed *Support Vector Machine* (SVM) based prediction to predict individual mobility behaviour for different modes of transport such as bike, car, bus, foot and train that are most likely to be used for travelling. SVM is a discriminative classifier algorithm based on the concept of a decision plane to classify a linearly separable dataset with decision boundaries. Their data come from crowdsourced data using a dedicated smartphone app in the city of Leuven, Belgium, collected from January to April 2015. It consists of 17,040 validated trips from 292 users and is divided into two datasets, 75% for training and 25% for testing. They used 11 input variables which are *User ID, Trip ID, Start time, Stop time, Start Location, Stop Location, Distance, Transportation mode, Trip purpose, Working day identification,* and *Holiday identification*. Using all these features, the prediction accuracy is 82%, and they used a confusion matrix to explain the existence of misclassifications between transport classes.

Baumann et al. [55] analysed the performance of 18 prediction algorithms focussing on their capability to predict the location transitions where individuals move between two places.

They observed 37 individuals' mobility traces over 1.5 years. They found high average accuracy for next-place prediction but not for predicting transition between two places. They proposed an algorithm called MAJOR by combining 18 methods considered in their analysis into a single algorithm. Then, they made the final prediction using the majority vote from all those algorithms. The spatiotemporal metrics which are *current location, previous location, time of the day, day of the week,* and *weekday/weekend* are defined, so that the ability of predictor to capture those transitions can be characterised. With MAJOR, they could achieve high accuracy of up to 87% for both next-place prediction and transitions prediction.

For prediction of the next value in a time series, the most useful history is often immediately before that value, and so the features that are input to a predictor will often consist of the current value and the previous N values, which is called a sliding window of size N+1. As the time of the predicted value advances, the set of input features is a window of previous values that slides through time to keep pace with the predicted value. In other cases, one might be categorizing values in a time series, e.g. to see if they appear anomalous. In that case the sliding window may consist of data values before and after the sample in the time series.

Moreira-Matias et al. [29] used a sliding window to measure the error of their streaming taxi data prediction before a new prediction is done for the next period. Each new prediction was used to update the average of the overall prediction. In their scenario, they considered 4 hours as the sliding window size.

Meanwhile, Li et al. [28] employed the sliding window mechanism to scan and filter the incorrect records from a trajectory using a set of criteria. A record will be rejected if it does not meet a defined criterion. Similarly, Chen et al. [56] implemented a sliding window for detecting the anomalous events when frequency during a certain hour is much higher than the adjacent hours on the same day. They chose 3 hours as the window size and slide it along the observed data to scan the centre of window and flag any values that are much higher than their neighbours.

Table 2.1 summarizes the state-of-the-art in human mobility prediction including data source, timespan, prediction metrics, methods, and performance assessment. It can be seen that most of the studies in Table 2.1 involve individual based predictions, so that the individual identities of moving entities are essential, such as taxi ID, mobile phone ID, pedestrians ID, and census ID. This is different to recently published studies in BSS prediction that are mostly using aggregated system-wide predictions as will be shown in subsection 2.3.7.

Table 2.1. Summary of existing work in human mobility predictions.

Author	Data source	Timespan	Prediction metrics	Method	Performance
Li et al. [28]	4000 taxis' GPS traces in Hangzhou, China	A year	The spatiotemporal variations of taxi passenger numbers in an extracted hotspot	An improved ARIMA model	5.8% error
Moreira-Matias et al. [29]	441 taxis, in Porto, Portugal	Streaming data	The spatial distribution of taxi-passengers	Three time-series models and an adaptive ensemble of those three time series models.	Error is lower than 26% using an adaptive ensemble model.
Massuci et al. [48]	England and Wales population census	2001	The flow of commuting pattern	Radiation and Gravity model	Gravity model shows a better performance
Ren et al. [49]	US census and highway traffic data		The commuting flows in spatial networks based on cost-based generalization	Radiation model	A high Pearson Correlation Coefficient (PCC), 0.75, based on travel time costs.
Asahara et al. [50]	1337 pedestrians in Osaka Japan	February 2010	Pedestrian movement	Mixed Markov Model	Prediction accuracy 74.4%
Gambs et al. [51]	Three dataset from Phonetic, Geolife, and Syntetic	Phonetic from Oct 2009 to Jan 2011 Geolife from Apr 2007 to Oct 2011	Next place prediction	n^{th}-Mobility Markov Chain	Prediction accuracy of the next location is in the range of 70% to 95% as soon as n = 2
Zhang and Haghani	Regional Integrated Transportation Information System (RITIS) in Maryland, US	2012	Travel time prediction	Gradient Boosting, Random Forest, ARIMA	2.01%, 2.04%, & 2.03% MAPE (5 mins ahead), 2.77%, 2.78%, & 2.90% (15 mins ahead), and 2.82%, 2.85%, and 3.01% (30 mins ahead)
Lopez et al. [54]	A dedicated smartphone app in the city of Leuven (292 users)	January to April 2015	Individual mobility behaviour for different modes of transport	*Support Vector Machine* (SVM)	Best prediction accuracy is 82%
Baumann et al. [55]	37 individuals' mobility traces from their mobile phone	1.5 years	Next-place prediction and transition prediction	*MAJOR* which is a combination of 18 algorithm	Best prediction accuracy is 82%

2.2. BSS Overview

The first generation of BSS was launched in Amsterdam in 1965 [57]. Recently, the fourth generation of BSS with fully automated operation has been widely implemented as a sustainable transportation system in many cities. There has been significant growth from 375 systems comprising 236,000 bikes in May 2011 to 535 systems with an estimated fleet of 517,000 bikes in April 2013 [58]. These numbers have further increased to 712 systems with 806,200 bikes in June 2014 [59]. This massive growth of BSS is related to the promotion of healthier mobility choices in crowded cities as well as to reduce traffic congestion and air pollution. It has also been introduced to be a simple solution to address the under-served destinations and the "first or last mile" connection problem in getting citizens from major transportation hubs such as bus or train stations to their final destination such as workplaces or home, or vice versa. BSS will also prevent people from being troubled with private bike ownership issues such as routine maintenance, parking, storage, and theft. Some BSS share their trip data repositories for public access. In London and New York, for instance, publically available trip data describes up to one million trips a month in summer.

The growth in global uptake of BSS illustrates the usefulness and popularity of such systems, however such systems are not without problems. For example, the only two BSS schemes in Australia, Brisbane and Melbourne, have not attracted as much use as anticipated [60]. In Brisbane, there were only 200,000 trips over 20 months [61]. In Mumbai India, the BSS was closed due to lack of use and failure to implement the model on a sufficiently large scale [62].

2.2.1. BSS as a Complex System

BSS stations are typically spread non-homogeneously over an urban area with a density of one station every few hundred metres. The short inter-station distances are because a BSS rental is intended for a short one-way individual trip within a city. Users can be registered regular users or casual users. A trip occurs when a user picks up a bike in one station, rides it on his or her preferred route and returns that bike to a vacant docking slot of another station in the system. The system-wide mobility pattern can then be described as a dynamic network. This network is formed by the stations and a large traffic flow between stations over time. In each station, usage can be measured. Thus, from a network science viewpoint, BSS can be analysed as a complex system composed of interconnected stations that exchange bikes [63].

If the availability of bikes and empty docking slots cannot meet the instantaneous demand level, users may not find available bikes to rent, or may not be able to return the bikes to their preferred stations. Significantly more pickups than returns, or vice versa, brings a station to an imbalanced state [63], where the station is full (and unavailable for returns) or empty (and unavailable for pickups). This is an intrinsic problem of BSS because of its natural one-way renting mechanism. Imbalance will obviously decrease the efficiency and service level of the system. Redistributing bikes manually from highly loaded stations to the empty ones using service trucks is a critical task to keep the system as balanced as possible. Manual redistribution significantly increases system cost. For example, in Taipei city, its BSS reached a deficit of at least $NT 1 million after running for one year, and redistribution was one of its most expensive costs [64]. 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 [65]. These high costs and time-consuming operations will be further compounded if the redistribution scenario is reactive, so that bikes are only redistributed after imbalances occur. Accordingly, proactive redistribution is needed [66]. There are many studies that have been conducted for bikes redistribution scenarios and optimization of vehicle routing using both static and dynamic based approaches [67-69]. Static repositioning is conducted during the night when the usage is very low and the system is nearly idle, while the dynamic rebalancing is performed during the day to deal with forthcoming shortages. However, this study will not directly deal with the optimization of the redistribution problem, but it is more about the understanding of users' mobility behaviours that affect the stations usage and other aspects of services and operations. Sections 2.2.2, 2.2.3 and 2.2.4 very briefly introduce the areas that the research is this thesis will address, so that the subsequent literature review has some context.

2.2.2. Station Neighbourhood Ties

As a dynamic network, stations in BSS are not independent and should be relationships among nearby stations so that if something happens in one station, it will affect other stations. In a BSS operational context, when a station is out of service because of shutdown, or when it is in an imbalance state because it is full or empty, it will impact to other stations. This impact could depend on the behaviour of the users and also the topology of the system. There has not been any published research on these BSS spatial ties, and so this thesis proposes using some of the techniques from other (non-BSS) mobility studies such as mobility motifs, to fill in these

knowledge gaps. In addition, the application for BSS design and deployment is also investigated.

2.2.3. BSS Individual Mobility Behaviour

In many current studies, aggregated BSS analysis is more popular rather than individual-based ones. This is often because of the lack of any individual identification in most BSS publicly available trips data due to privacy issues [63]. Some studies have used the scrapped stations usages from BSS websites [11, 70], which lacks individual information. However, the dynamics of BSS are directly inherited from users' individual behaviours. Each user may have unique movement styles and preferences that lead to diverse trip frequency, duration, speed, waiting times, motifs, distance, and direction. Meanwhile, the same regularities and patterns are likely to be associated with the same user type: commuters, casual users, tourists, and night workers [12, 71-73], for example. If the homogenous users can be grouped into certain clusters, it would be possible to measure cluster predictability level and use their collective trends to make a cluster-based prediction than using non-homogenous of whole users. It is also expected that the same user types have similar responses to external factors such as hour of the day, day of the week, nearby points of interest, station spatial layout, and weather [17]. How users move both spatially and temporally over the BSS, therefore, has been a subject of several previous studies [30, 71, 72, 74]. However, there is still a room to further investigate entropy and predictability for BSS. As described earlier, predictability is the theoretical inverse of entropy (or randomness). From an information theory perspective, the performance of a prediction algorithm is limited by the predictability metric that is inherent to the data [24, 43].

To differentiate users, recent studies [72, 74] employed naïve approaches using demographic and subscription status, such as registered users with an annual subscription and unregistered users with a limited period subscription. There is a risk in creating user types with non-uniform movement patterns that potentially contains outliers. Some unregistered users may have regular trips similar to registered users, and vice versa. This thesis hypothesises that regularities should be associated with how frequently and regularly users travel, rather than on their registration and demographic status, and this will be explored later in Chapter 6. On the other hand, some other studies used spatial [71] and temporal mobility pattern [75], but none of them conducted further analysis to measure their homogeneity and predictability. Therefore, this study proposes users clustering based on their actual temporal behaviour and conducts

further analysis about their homogeneity and predictability as well as the application for BSS operation.

2.2.4. BSS Aggregated Mobility Behaviour

At the system-wide level, the BSS aggregated mobility pattern is the sum of many different individual trips with certain dynamics over time. Using the assumption of the hourly usage which consists of a constant (i.e. stationary) underlying weekly pattern (i.e. cycle) plus a disturbance to that pattern caused by certain factors, this study proposes a new predictor that estimates the current disturbance from the underlying seasonal weekly pattern. This can be extended to an underlying weekly pattern that itself changes slowly over the seasons. Meanwhile, most BSS studies prefer to predict the absolute values of hourly usage [15-17].

Predicting system-wide behaviour involves all stations in the system, and there are many internal and external features that could be possible features for enhancing prediction. Data sizes are large - for instance there were 573 BSS stations and 566,000 users in the London BSS Data in 2012 [46] that this project will analyse. Calculating all possible features over hundreds of BSS stations and hundreds of thousands of users is very computationally expensive.

Over the last decade, various data-driven analyses on BSS have been done from different perspectives. These have used either publically available shared-data that mainly contains trip information, or the scraped data from BSS websites that take snapshots of station states at regular intervals, or survey data that contains the BSS users' opinions, experiences and demographic data. Station usage analyses are intended to identify the fluctuation of demand and availability of bikes or docking slots, while trip-based analyses are commonly intended to reveal the mobility dynamics and individuals' behaviours, and survey-based analyses are typically related to investigating quality of the BSS services as will be presented more detail in the following section.

2.3. Previous BSS Studies

Shared BSS data mostly come from cities in Europe and the USA, with only limited data from Asia and Australia. The majority of BSS analyses use data from big cities such as London [19, 31, 46, 72, 75-79], Washington DC [12, 17, 27, 47, 56, 76, 80-85], Paris [86-94], and New York [17, 81, 89, 95-97]. Other studied data sets are from Chicago [73, 80, 98, 99], Lyon [63, 65, 74, 100], Boston [12, 76, 80, 101], Barcelona [11, 70, 102], Hangzhou [15, 16, 103],

Brisbane [61, 83], Minneapolis [76, 104], Vienna [105, 106], Denver [76, 84], Pisa [64, 107], Dublin [14, 108], Minnesota [84], Seville [102], Montreal [109] Helsinki [110], Vancouver [111], Nanjing [112], and Castellon [113]. In addition to BSS data, some of those studies also used weather data as a feature of their analyses. For spatial visualisation purposes, most of these studies used data superimposed on city maps.

Many different topics are covered in these studies, but this review will focus on BSS generations and problems, system design and implementation impact, spatiotemporal analysis, user and station clustering, mobility models, weather effects, prediction, and journey advisors.

2.3.1. BSS Generations and Problems

In terms of system operation, there have been four generations of BSS, with significant evolution and improvement across generations. The first generation was introduced in July 1965 in Amsterdam, and was called the white bike or free bike system [39, 57]. Initially, fifty white painted bikes were placed throughout the city for free use. However, they were often stolen and damaged because they were left unlocked, and this system was soon abandoned. Later, there were two other cities that also implemented a similar free bike-sharing scheme, La Rochelle in France and Cambridge in the UK in 1974 and 1993 respectively [57].

The second generation of BSS was launched in January 1995 in Copenhagen, Denmark. This scheme had used a coin-deposit system which enabled users to pick up a bike by depositing a coin into a dock, then returning the bike to a dock where they received back the deposit coin [39]. Soon after this, many cities in Europe and US introduced similar bicycle sharing schemes using this coin-deposit system. The weakness of this system was the customer anonymity so that, similar to the first generation, the bikes were subject to theft.

The third generation was started in 1998 in Rennes, France when IT-based systems were used. This system had the capability of reading RFID (Radio Frequency Identification) tags on bikes, and accepting credit or debit cards for hire as well as for membership. In this generation, the user accountability was improved considerably. Recently, the fourth generation has added features such as on-line station availability, special pricing for self-rebalancing, and integrated billing systems with other transportation means [57].

In the fourth generation, the imbalance state between the availability of bikes and vacant docking slots still exists as an intrinsic problem and major concern. The one-way trips in BSS can result in asymmetric flows [95] which in turn give non-uniform distribution of bikes

amongst stations. Vogel et al. [105] adopted a data mining approach to gain insight into the complex bike activity patterns in Vienna. They revealed imbalance states in bike distribution that can be understood in terms of system structure and activity dynamics. Several studies have analysed and proposed some methods to address the imbalance issue from different points of view, such as the optimization of fleet routing and the number of fleets [67], a proposal for giving incentives to users to rebalance [68], implementing imbalance prediction [17, 27, 69], and proposing journey advice for users [108, 114, 115]. Some of those studies will be reviewed in the next subsections.

Generally, all of the proposed solutions have the same goal which is to guarantee a certain *quality of service* (QoS) level of BSS. This service level expresses the users' satisfaction with the BSS. Knowing the BSS QoS level is crucial because successful implementation of BSS requires the ability to cope with a fluctuating demand [116]. Pfrommer et al. [69] proposed a simple measurement of BSS QoS where the service level is equal to potential customers minus no-service events divided by potential customers. Here, the no-service events correspond to users who could not pick up or return bikes because of unavailable resources. Raviv and Kolka [116] introduced a user dissatisfaction function which measures the performance of a BSS station based on the quality of repositioning. They stated that there are two repositioning modes. Static repositioning is conducted during the night when the usage is very low and the system is nearly idle, while the dynamic mode is performed during the day to deal with forthcoming shortages. In another study, Singla et al. [68] proposed self-balancing by giving incentives to users to assist with rebalancing. They employed differential pricing policies and provided alternative routes to users for picking up or returning their bikes. Using the Boston bike data from July 2011 to October 2012, they compared their incentive policy simulation with the already running truck policy, and showed a potentially favorable result. They claimed that their work is the first work studying the dynamic incentives for BSS users.

2.3.2. BSS System Design and Implementation Impact

To properly set up a bike-sharing system, the system should be configured in a way that meets the users' needs. Methodologies proposed by Dell'Olio et al. [117] consider the users, system and policy aspects in designing BSS. They estimated the potential demand for BSS, as well as the willingness of users to pay for travelling within a city, and designed suitable locations for stations and the pricing policies of a sharing system. In another study, allowing for the interests of both system planners and users, Lin et al. [118] proposed a mathematical

model to determine the number and locations of the stations, the network structure of bicycle paths that connect between stations, as well as the travel paths for users between each pair of origin and destination stations. This work addressed the system design problem in an integrated view incorporating setup cost, reallocation cost and travel cost.

Meanwhile, Eluru and Imani [99] focussed on examining the influence of bicycle infrastructure (number of stations and station capacity), land use, and built environment on bicycle usage. First, they considered bicycle infrastructure as exogenous or produced by external factors in modelling demand. In cases where the bicycle infrastructure is closely related to the land-use and urban form, it is important to recognise that developing models treating the bicycle infrastructure as exogenous to the dependent variable (bicycle demand) might lead to incorrect and biased model estimations. Then, they addressed that challenge by proposing an econometric framework to jointly model the decision processes under consideration.

The implementation of BSS has a significant impact on human mobility in an urban area. Quantitatively, Jäppinen et al. [110] modelled the potential impact of the BSS implementation on public transport travel times in Greater Helsinki, Finland, based on the population and 16 important destinations in the city. As BSS is intended to solve last mile mobility problem, they compared total travel times between using public transport + BSS and using public transport only. They found that the mixed scenario, public transport extended with BSS, could reduce travel times by more than 10% on average, or around 6 minutes per individual trip. They stated that although the time savings per individual may not appear remarkable, the total summed across all potential users will be considerable. For example, if the daily public transport trips are around 500,000 to or from the city centre (Helsinki Region Transport, 2010), and if there are only 5000 trips that use BSS, this equates to 500 hours saving in travel time. While in New York City, Faghih-Imani et al. [13] found that BSS also are either faster or competitive with taxis in terms of travel time in a dense urban area.

However, these existing works have not used information from existing BSS data to identify the best distance between nearby stations that can be used as a standard of BSS design and deployment. As will be shown later, some relatively complex data analytics can give some insight into suitable station spacing.

2.3.3. BSS Spatiotemporal Analysis

From a temporal perspective, investigating the behaviour pattern of a bike-sharing system is very helpful to understand the mobility characteristics of a city which can reflect urban activity dynamics over time. Borgnat et al. [63] have explored BSS from signal processing and data analysis perspectives. They modelled the time evolution of the Velo'v dynamics movement in Lyon, France. They varied the aggregated time scale form 15 minutes to 2 hours to find a good trade-off between resolution of detail and fluctuations in distributions, and then selected 1 hour as the appropriate aggregation time scale. Using that one hour scale, they showed that the BSS temporal pattern is mostly ***cyclostationary*** over the week. A cyclostationary temporal pattern is a periodic pattern that is repeated at regular intervals such as daily, weekly or yearly. If there are N time bins in one cycle, then the time series consisting of all the points in the same bin (e.g. 9 am Mondays) form a time series, and if each of those N time series are statistically stationary, the periodic time series is called cyclostationary.

This BSS pattern can be divided into two group patterns: *weekdays* and *weekend* days. Their results on weekdays show three usage peaks, in the morning, at lunchtime and late afternoon, whereas on weekend, usage is concentrated in the afternoon.

Unlike Borgnat et al. [63] who used origin-destination trip data, O'brien et al. [75], in a later study, examined the footprint of docking stations activities to conduct a global view of BSS data for generating insights into sustainable transportation systems from 38 systems all over the world: 16 in Europe and the Middle East, 11 in Asia, 2 in Australasia and 9 in the Americas. The bicycle sharing systems that they studied have at least 40 docking stations and a clean feed of data. Their concern was to look at the temporal changes in bicycle distribution within those stations and to analyse the variation of occupancy rates over time. Looking at the diurnal and weekly variations in usage, they compared and contrasted temporal patterns and used it as one basis of classification between cities. Elsewhere in Europe, Froehlich et al. [11] provided a temporal analysis of Barcelona's bike station usage patterns to identify shared behaviours across stations and show how these behaviours relate to location, neighbourhood and time of the day. They demonstrated the potential of using BSS as a data source to gain insights into city dynamics and aggregated human behaviour. Their temporal results revealed a repeating three-pronged spike in station activity during the weekday, which corresponds to the morning, lunch, and evening commutes. Still in Europe, Ciancia et al. [79] presented a descriptive PDF of cycling times in London. They found one salient feature of cycling times

which is that 7% of all trips are longer than 30 minutes (the free-use trip time). Some trips are up to two hours, which is more than enough to travel between any two stations in the service area of BSS in London. This range of long trip times fits a so-called fat-tailed distribution, where for trips longer than 30 minutes, the PDF of cycling times is proportional to t^{-a} where $a > 0$ (for London, a =3.1).

From a spatial perspective, BSS can be seen as a directed network graph $G = (V, E)$ where nodes (V) represent stations, edges (E) represent the flow between stations, and edge weights correspond to the inter-station trip numbers [80]. Accordingly, many graph theoretic methods can be applied to BSS networks to examine their connectivity. Bargar et al [80] applied three graph algorithms in their spatial analyses which are *Maximal Clique Detection* (MCD), *Louvain Modularity Optimization* (LMO), and *Spatiotemporal Density-Based Spatial Clustering of Applications with Noise* (ST-DBSCAN). MCD was used to determine the largest interconnection link in the network from a subset of trip information. LMO was used to find groups of stations that essentially have no perfect cliques but are still highly interconnected. ST-DBSCAN was used to cluster similar trips, since it has the capability to integrate temporal features and other non-spatial features of data into Density-Based Spatial Clustering by defining density using neighbours' states. Meanwhile, Zhou [98] constructed a similarity graph from bike flows then used the fast-greedy algorithm to discover the spatial community of those flows. The algorithm goal is to optimize the modularity function which is one index which defines the network structure. The higher the modularity index is, the denser the node connection is within a community and the rarer it will be with outside nodes.

Froehlich et al. [11] presented a visualization of stations that show spatial dependencies, e.g. uphill stations tend to be empty and less active stations are located at the edge or outer ring of the bike-sharing network. Meanwhile, O'brien et al. [75] used one kilometre around each docking station as a buffer area approximation that could possibly influence that station. This distance is a compromise between the maximum straight-line distance to a bike station that someone would likely walk and the minimum distance that a user would be likely to cycle outside the boundary of the buffer [75].

Generally, GPS tracking is not installed on bikes in a BSS, so that the actual trajectory of trips cannot be traced. The only positions that can be sensed are the pickup geolocation (origin) and the return geolocation (destination). Subsequently, almost all BSS studies simply use the Euclidean distance which is the shortest straight line distance in the plane rather than compute

the actual travelled distance for their trajectory analysis. O'brien et al. [75] assumed that the cycling trajectory within a city usually is not significantly longer than the straight line distance, so using Euclidean distance is still reasonable. Austwick et al. [76] used this measure because at least the Euclidean distance is free from a set of hypotheses about route choice and routing mechanisms. Faghih-Imani and Eluru [99] also used the shortest distance even though they stated that the actual journeys may involve a different path. A different approach was used by Jensen et al. [100] to get the BSS trip distance by looking at the distance measured by counters installed on the bicycles in Lyon, so that they got the precise trip distance as well as time travelled. Using that approach, the average trips distance and time travelled was 2.49 km and 14.7 minutes respectively, giving an average speed 10.16 km/hour. Using the actual distances, they also observed the average speed at certain hours of the day. For example, average speed was 14.5 km/hour during early weekday mornings.

Padgham [31] investigated the convergence and divergence of the flux flows using 351 stations from the London BSS data. He found that the human mobility in collective patterns arise from a mixture of both diffusive and directed movement. Meanwhile, Borgnat et al. [63] analysed BSS spatial patterns to understand how the flows are distributed spatially along a network in which the bike stations are deployed uniformly within a city. They identified areas where stations receive more or less incoming and outgoing bikes as well as their flow directions. A matrix of flows between stations was constructed and modelled as a directed graph. Here, they added the time dimension to the flow matrix, $T[n,m](t)$, so that the weights of the flow from station n to station m, at time t, will be the number of trips $T[n,m]$. They also revealed that spatial and temporal dependencies exist between stations.

In order to understand London bikes flows in a way that allows relevant details to be perceived, Wood et al. [77] proposed three visualisation approaches: *Flow Maps* (using curved flow symbols to show the flow structures), *Gridded View of Stations* (maintaining the geographical relationship to depict docking station status spatially and temporally), and *Origin-Destination Maps* (visualise the OD matrix directly while keeping geographic context). Then, they compared those to four existing approaches which are *Semi-opaque Euclidean Flows Vector* (using straight line between OD drawn opaquely), *Flows Density Mapping* (transforming linear flows into a continuous surface), *Edge Bundled Flows Vector* (using graphical aggregation of occupied adjacent pathways), and *OD Matrix Visualisation* (coping with the congestion problem by ascribing equal graphic weight for short and long journeys). They concluded that *Origin-Destination Maps* complement the general idea for visualizing

origin-destination flows that is able to display both longer and shorter trip flows simultaneously. Also, it geographically shows flows in a manner which is unbiased and scalable.

2.3.4. BSS Users and Station Clustering

Understanding the behaviour and characteristics of BSS users has been investigated by several researchers. Beecham et al. [71] have used spatial analysis, namely density-estimation, in classifying the commuting behaviour in London BSS data to identify the potential commuting cyclists and their plausible workplaces. They compared the terminating and originating journeys within the same vicinity of derived workplaces between peak-times in the morning and in the afternoon [71]. For each user, first, an empirically-defined workplace, or set of workplaces, is created. Then, all trips that arrive at this workplace in the morning and depart from this workplace in the evening are labelled as commutes [71].

Beyond commuters, in another study, Lathia et al. [72] have analysed another type of user which is the casual user. Unlike commuters from the previous study, a casual user here is simply defined as an unregistered user as opposed to a registered user that may travel more frequently. Here, they studied the impact of bike hire policy change in December 2010 for casual users from using a registration key to access the system to simply using a debit or credit card to do so. Specifically, they investigated how the policy change affected the system's usage throughout the city. They found that quicker access to the system has a significant correlation with greater weekend usage for casual users. On the other hand, it also reinforces the weekday commuting trend [72].

Similar to Lathia et al. [72] who used registration data to classify users, Vogel et al. [74] have used the period of users' membership data (annual, weekly or daily) of Velo'v BSS, Lyon, where annual membership contains users' demographic information such as age, gender and postcode. However, they stated that using only subscription data to classify users may result in improper or biased classes because they may not be generic enough to capture similar behaviour. Then, using a k-means clustering method, they clustered the annual membership users into nine classes based on cycling patterns distributed according to the intensity and the regularity of use. 21 attributes are defined, the first eight corresponding to weekly activity while the others correspond to annual activity. Some of those features are averaged number of trips made per week, average number of trips made on weekdays, total number of trips made

over the year, number of trips made for all months, percentage of movements on the busiest weekday, and percentage of movements on the busiest month. Although their clustering method exhibits nine classes of users, they then propose that it is fairly easy to interpret them into only four clearly separated categories which they call *user of heart* (intensive, regular users), *assiduous users*, *multimodal users*, and *sporadic users*.

In another study, O'brien et al. [75] collated temporal characteristics of 38 BSS from all over the world which are the number of the peaks per day for weekdays and weekends, the relative difference between weekend and weekday usage, and average load factor. Using these temporal characteristics, they then proposed four user demographic categories, *commuters* who use bikes from home to workplaces during weekdays, *utility users* who use bikes on weekdays for shopping and errands, *leisure users* who use bikes generally on weekends for fun and exercise, and *tourist users* who use bikes for exploring the city.

Table 2.2. Summary of existing works in BSS users clusters.

Author	BSS Data	Clustering method	Users cluster
Beecham et al. [71]	London	Spatial analysis (density-estimation)	Commuting Pattern
Lathia et al. [72]	London	Registration status	Casual users
Vogel et al. [74]	Lyon	Membership data	Annual, Weekly, Daily
		K-mean based on cycling pattern of annual users	User of heart, Assiduous users, Multimodal users, and Sporadic users
O'brien et al. [75]	18 BSS from all over the world	Temporal characteristics	Commuters, Utility users, Leisure users, and Tourist users

Table 2.2 summarises the limited existing works in BSS user clustering. There are at least three metrics that are used for clustering which are registration or membership status, spatial features, and temporal characteristics. However, none of these studies conduct further analysis about their homogeneity, how predictable the clusters are, and what the practical benefit for BSS operation is by identifying those clusters.

For stations, segmenting the bike stations into several sections or clusters is useful for various operational purposes such as monitoring, prediction and rebalancing. This clustering problem has been addressed in a number of studies. Froehlich et al. [11] investigated a hierarchical clustering technique called *dendrogram clustering* over each station's DayViews.

Here, a DayView is calculated by averaging station data that matches certain criteria into a 24 hour window, discretised into five-minute bins (288 bins/day). They then built two sets of clusters: one based on weekday *Activity Score DayViews* ("Activity Cluster") and the other on weekday *Available Bicycle DayViews* ("Bicycle Cluster"). In both cases, a normalized weekday DayView representation was created for each station and a similarity matrix constructed to store the DTW (*Dynamic Time Warping*) distance between each cluster. Finally, their clustering algorithm returned five activity clusters and six bicycle clusters based on flows.

Borgnat et al. [63] clustered stations in communities and clustered flows of activity between stations at finer time-scales. First, to understand the impact of the inhomogeneity of the city on the long-term activity of individual stations, they looked for groups of stations exchanging many bicycles. This amounts to detecting communities of stations in a network. Second, in order to uncover the main properties of flows on the Velo'v station network, a k-means algorithm is run on *T[n,m](t)* for *t* equal to the 19 selected time-features and *(n,m)* being 1046 pairs of stations. They then produce four well-separated clusters.

Vogel et al. [105] did cluster analysis in order to group stations according to their normalized bike pickup and return activity. The goal is that data objects within a group are similar to each other and different from objects in other groups. They applied three clustering algorithms which are k-means (KM), *Expectation Maximization* (EM) and *sequential Information Bottleneck* (sIB). The EM algorithm extends the KM paradigm, while sIB which is an agglomerative clustering method originally designed for cluster analysis of documents is used because it is capable of dealing with high dimensionality data. Here, data objects are assigned to k clusters whereas the number of clusters has to be chosen beforehand. Based on initial partitioning, objects are relocated by minimizing the distance of objects within clusters and maximizing the distance of objects in different clusters. Cluster validation indices measure if a structure found with cluster analysis is adequate. Then, they used three indexes for cluster validation, *Davies-Bouldin-Index*, *Dunn-Index* and *Silhouette-Index*. Their result shows that according to the elbow criterion[3] their three algorithms yield the best cluster for $k = 5$.

Etienne and Oukhellou [88] proposed a generative model based on Poisson mixtures to analyse the patterns in different areas of Paris using different functions, considering the latent factors of each station. To handle the event discrepancy between stations, they introduced station scaling factors [88]. They found k = 8 as a good trade-off between the cluster

[3] A naïve procedure to determine the optimal number of clusters that identifies where adding more clusters has limited impact on node-centroid distances.

complexity and interoperability. Then, they named those eight clusters as *spare-time 1, spare-time2, parks, railway stations, housing, employment 1, employment 2,* and *mixed*. Meanwhile, using similar Paris bike data and also a similar Poisson mixtures based approach, Randriamanamihaga et al. [87] proposed a generative count-series model adapted from Poisson mixtures model to discover temporal-based clusters over OD flow data. This approach reveals how areas with different usage interact over time by considering latent factors. For each edge, these latent factors determine the cluster memberships. In other words, this method can be applied to cluster the edges of temporal weighted-graph based on the temporal characteristics. Their results presented four cluster labels, *weekend joyriding, night life, morning works,* and *early bird works*, based on socio-economics information across OD flows which are density of populations, employment, and commercial zones.

Another clustering method is proposed by Xu et al. [119] which is an improved k-means algorithm to segment stations in Hangzhou based on optimised *Simulated Annealing* (SA). Here, the optimised SA algorithm was used to assign the preliminary cluster centre to the k-means algorithm. In k-means, the value of k as an input to the algorithm is typically based on some criteria such as the prior knowledge, the desired purpose of clusters, and type of clusters. The closeness or similarity is a common profile to be used to group the stations where in k-means the closeness is computed by Euclidean distance. On the other hand, SA is a probabilistic method to solve different combinatorial optimization problems both unconstrained and bound-constrained. While the traditional k-means clustering has sensitivity to the preliminary cluster centre, using the initial centre obtained by the improved algorithm will avoid the blind search in the initial stage of k-means and reduce the number of iterations. Their results exhibit that the proposed method is more efficient and robust than traditional k-means clustering.

Table 2.3 summarises the stations clustering using various methods. Typically, the number of clusters is less than 10, so each cluster has many stations. For example, in Paris with 1208 fixed stations in July 2007 [87, 88], if clusters are set to 8 as proposed by [88], then each cluster has an average of 151 stations. Furthermore, because they are clustered by activity profiles, one region could consist of different cluster members. On the other hand, if operators employ region-based monitoring and distribution, smaller, geographical based clusters may be more useful, because they could be easier to manage.

Table 2.3. Summary of existing works in BSS station clustering.

Author	BSS Data	Clustering method	Stations cluster
Froehlich et al. [11]	Barcelona	A hierarchical clustering technique called *dendrogram clustering* over each station's DayViews	6 activity clusters, 5 bicycle clusters
Borgnat et al. [63]	Lyon	K-mean of activity flow between stations	4 well-separated clusters
Vogel et al. [105]	Vienna	Normalized bike pickup and return activity using k-means (KM), *Expectation Maximization* (EM) and *sequential Information Bottleneck* (sIB)	Three algorithms yield the best cluster for k = 5
Etienne and Oukhellou [88]	Paris	Poisson mixtures, considering the latent factors of each station	k = 8 as a good trade-off between the cluster complexity and interoperability
Randriamanamihaga et al. [87]	Paris	A generative count-series model adapted from Poisson mixtures model	4 cluster labels
Xu et al. [119]	Hangzhou	Improved k-means based on optimised *Simulated Annealing* (SA)	More efficient than traditional k-mean

2.3.5. BSS Mobility Models

BSS with stations and bike exchange between stations can be modelled as a Markov-chain model. Here, each station is a state, and bike exchange is a transition between states where its probability can be computed. With discrete time and finite states, BSS with n stations can be completely described by n x n transition probability matrix (P) where P_{ij} denotes the one step probability of trips from station S_i to station S_j. Crisostomi et al. [101] described P as a row-stochastic and non-negative matrix. Then, since the entity of each row is a probability, each row sums to 1. They assumed that a station is related to two states. BS_i state associates with a parked bike at that station, and TB_i state refers to a bike that is moving from that station to any other station. Accordingly, at every time window (they use one second for simulation), a bike in a BS_i state can either move to the travelling state or remain in the same parking state. For a bike in a TB_i state, it can either keep moving (remains in TB_i state because the destination has not been reached yet) or may change to a parking state BS_j at any destination station. This model can be seen in Figure 2.7.a.

Gast et al. [94] proposed a Markovian model for modelling a single station to show its behaviour. Initially, they observed Kendal's notation for queuing networks where a station is modelled as time-inhomogeneous M/M/1/k queue. There were two assumptions in this model.

First is memory-less transitions and Poisson processes of user and bike arrival. Second is independence between stations. Practically, this is not true because when a station is full no bike can dock there; and these queued arrivals will often divert to nearby stations. Conversely, if a station is empty, no bike can depart from there which will reduce the arrival rate of other stations. Alternatively, more realistic assumptions could consider each trip from origin (i) to destination (j) with departure and arrival intensity for each process. Unfortunately, this makes the model and parameter fitting more complex but with little gain in modelling accuracy, so that they restricted their model to one where each station behaves as an independent M/M/1/k queue, Figure 2.7.c.

Another proposed BSS mobility model is using *Latent Dirichlect Allocation* (LDA) which is a three-level Hierarchical Bayesian Model for discrete data. This statistical model was originally developed to analyse document collections that consist of bags of words. Montoliu [113] used a topic model based on LDA to uncover the mobility pattern of a BSS in Castellon, Spain, in an unsupervised manner. Topic models are statistical generative models which represent the mixture of topics in documents. They are learned in latent space because they involve latent variables and are useful for modelling task. Topic models have an ability to characterise bags, the representation of discrete data. Then the author decoded the time period into three symbols, *increment* (↗), *decrement* (↘), and *no change* (↔) to characterise the station behaviour based on the availability of bikes during the day, Figure 2.7.b. Using this scenario, the latent topics that described mobility model can be effectively revealed.

Using a similar LDA approach, Côme et al. [93] investigated the sizeable OD matrices of Paris BSS data using LDA to discover the spatiotemporal behaviour of the system. First, a few OD templates were extracted. Then, they were interpreted as typical and temporally localized as a demand profile. They defined k (5) OD templates, *home ↔ work commute*, *lunch time*, *work ↔ home commute*, e*vening behaviour*, and *spare time*. Using these templates, they observed the stations that receive or lose more bikes than the average in the OD templates. Their results show that just a few OD templates can be used to summarise demand profiles for the system.

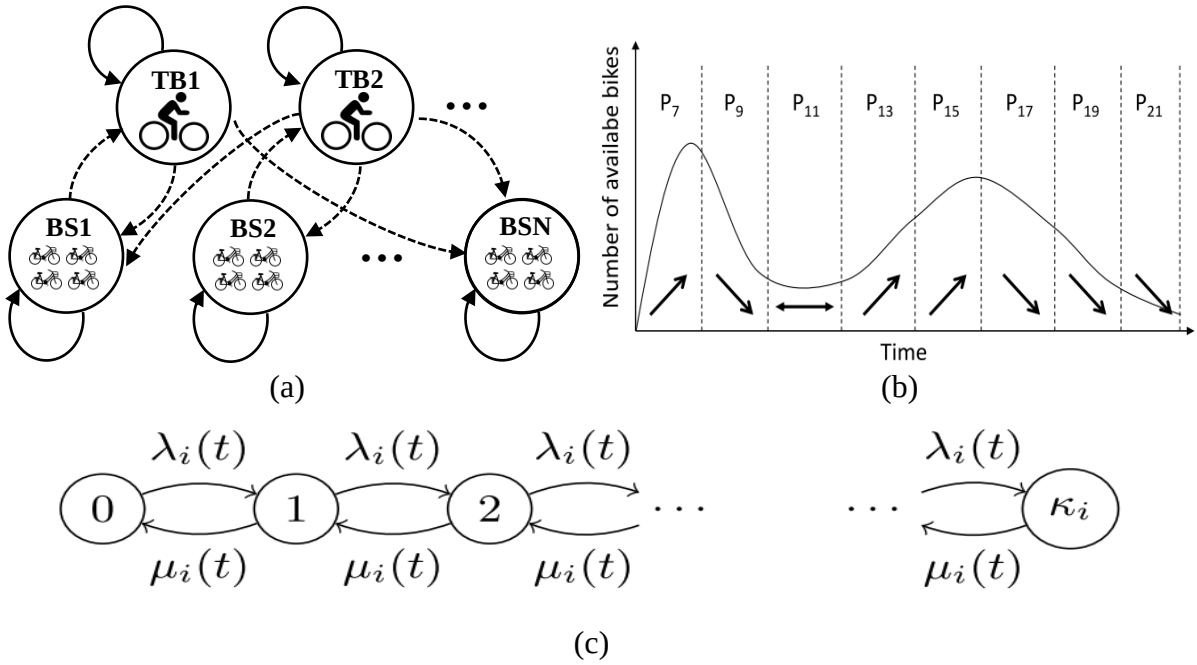


Figure 2.7. The redraw of (a) BSS Markov model [101], (b) LDA model [113], and (c) M/M/1/κ_i model [94].

2.3.6. Weather Effects on BSS

The impact of weather to the spatiotemporal dynamics of BSS has been investigated in some studies. Using 20 months of Brisbane bike data that comprised 285,714 trips, Corcoran et al. [61] examined the trip flows in three levels which are *system-wide, suburb to suburb*, and *station to station* with respect to weather. They employed the flow-comap to observe to what degree the weather conditions will modify the dynamics of BSS usage both spatially and temporally. They highlighted their results that both *rain* and *winds* are significantly correlated to the trip numbers at the system-wide level. The total numbers of trips significantly decrease for stronger winds and rainfall, while temperature was found not to have a significant impact in influencing the number of trips. This is not so surprising because Brisbane lies in the sub-tropical climate zone with relatively small temperature variations.

Meanwhile, Gebhart and Nolan [82] conducted the same weather impact investigation for BSS trips in Washington DC. They used more weather variables: *temperature, rainfall, thunderstorm, wind, snow, fog,* and *humidity* levels. They linked those variables to hourly number of users and duration of use. Their results show the effect of cold temperatures, rainfall, and high level of humidity to diminish both the possibility of using BSS and the duration of trips are quite significant. In contrast, the effects of thunderstorms, fog, wind, and snow are not statistically significant.

In Canada, Gallop et al. [111] modelled bicycle traffic with respect to the weather in Vancouver. They observed some weather variables such as *temperature, relative humidity, wind speed, clearness, fog,* and *precipitation (drizzle, rain, snow)*. As they claimed that the travelling decision of users to go with bikes is based on current weather rather than forecasted weather, their analysis is therefore focussed only on actual weather. Then, they used an ARIMA model to conduct the analysis in three stages, *identification*, *estimation*, and *diagnosis*. Their results confirm that weather has a substantial impact on bike usage where *temperature*, *humidity*, *rain,* and *rain* in the previous 3 hours are all found to be significant. However, they suggested that the impact is overemphasised in the model shown by the failure to interpret the complex patterns of serial correlation.

These three weather impact studies for BSS come with similar conclusion that there are certain weather variables which influence individuals in using BSS. However, the variables that have a substantial impact are different among different cities.

2.3.7. Various Predictions in BSS

Prediction is one of the most important topics in BSS research because it can help operators to plan their bike redistribution, or users to plan their journeys, or researchers to find the best prediction method for a particular scenario. Similar to the work on human mobility prediction in subsection 2.1.7 above, numerous prediction methods have been applied to BSS data. They have been investigated for various prediction purposes such as prediction of traffic flows [15, 17], the available of bikes and/or vacant docking slots [47, 70], number of trips [16], trips duration [73], level of demand [62, 85], over-demand prediction [81], pairwise demand prediction [96], number of usage patterns [27], next place prediction [46], cycle lane usage [109], station occupancy [120], the potential destination station and arrival time [73], waiting time for the next available bike [14], and pair of pickup and return stations that close to the origin and destination places of users [73, 96].

Froehlich et al. [11] compared four simple predictive models to predict the availability of bikes at each station in Barcelona. They used 13 weeks of scraped website data of bike and vacant slots availability that was sampled every 2 minutes. Those models are *Last Value* (LV), *Historical Mean* (HM), *Historical Trend* (HT) and *Bayesian Network* (BN). They applied clustering techniques to identify shared behaviours across stations and showed how these behaviours relate to location, neighbourhood, and time of day. They then compared the

experimental results from those four predictive models of near station usage. They have shown that fairly simple predictive models are able to predict station usage with an average error of only two bicycles and can classify station states (*full*, *empty*, or *in-between*) with 80% accuracy up to two hours into the future.

Using the same Barcelona data, Kaltenbrunner et al. [70] adopted a statistical model to predict the number of the available bikes and vacant docks for each station. They conducted an analysis of the activity cycle that can be obtained from the number of bicycles available at the nearby stations. First, they focussed on the local cycles, one for every station. Later, they aggregated these cycles to infer global activity cycles followed by examining the usefulness of these cycles to predict future numbers of bicycles in the stations. They also implemented time series analysis methods in the form of an *Auto-Regressive Moving Average* (ARMA) model. As its name implies, an ARMA model incorporates two fundamental models: an *Auto-Regressive* (AR) component which is able to exploit relevant information related to the autocorrelation nature of the time series, and *Moving Average* (MA) model which is able to incorporate information from additional sources of information generally denominated "input". The ARMA model is trained by means of an optimization procedure aiming at minimizing the fitting error within a selected training dataset. Their results reveal that the dynamics of neighbouring stations definitely have an important influence on the ability of predicting bicycle availability at a given station. Their experimentation also shows that considering a number of surrounding stations between 5 and 20 will provide good predictive power. Then, they evaluated how the prediction error increases as the time interval for predictions is increased. At 30 minutes prediction interval, the average prediction error is below than 1 bicycle, and then it reaches the maximum number of 3 bicycles after 1 hour prediction interval.

Differently from Froehlich et al. [11] and Kaltenbrunner et al. [70] who predicted the availability of bikes and vacant docks, Borgnat et al. [63] used the data from Lyon's BSS to predict the hourly number of rented bicycles by taking into account factors that are external to the cyclic pattern. Here, they predicted the entire traffic in each hour of the day by a combination model: the non-stationary amplitude $A_d(d)$ for a given day added to the fluctuation $F(t)$ at a specific hour. For prediction of $A_d(d)$, they looked for explanatory factors among weather, seasons, number of subscribed users, number of bicycles available, and specific conditions such as holidays. While for prediction of $F(t)$, they used a standard empirical spectrum analysis. They showed that $F(t)$ is well modelled by an auto-regressive process of

order 1 with exogenous input. Using this scheme can decrease the standard deviation of the error of the global prediction from 210 bicycles to 120 bicycles per hour.

Meanwhile, Zeng et al. [97] proposed a global feature-based model to improve BSS demand prediction in New York City. They used a *Gradient Boosting Decision Tree* (GBDT) and a *Neural Network* (NN) as feature extractors and employed four predictors: *Linear Regression, Decision Tree, Random Forest* and *Support Vector Regressor*. They examined three approaches: a *city centric model* which uses a single predictor using global data from all stations' data, a *station centric model* which uses an individual predictor for each station, and a *hybrid model* which is the station centric model with global features. They used three evaluation metrics: MAE, RMSE, and RMSLE. Using 12 months of training data and 3 months of testing data, they demonstrated that using global features from GBDT and NN in a hybrid model improves the prediction performance, and they also showed that the best predictor is RF.

Similar to Zeng et al. [97], Singvhi et al. [96] also used New York BSS data to predict the pairwise bike demand in morning rush hours (7 am to 11 am) during weekdays, as the system is highly driven by commuters at that time. They also considered other external data as covariates such as weather, taxi data, aggregated neighbours, precipitation, and day of the week. Using a regression model and RMSE, they demonstrated that examining the pairwise trips at neighbourhood stations level can significantly improve the prediction performance, compared to considering only individual stations.

Again, Li et al. [17] used New York BSS data in comparison with Washington DC BSS data. They proposed a hierarchical model, which contains a bipartite clustering algorithm, a multi-similarity-based inference model, and a check-in inference algorithm, to predict the check-out/in of each station cluster in a bike-sharing system, based on historical bike data and meteorology data. They evaluated their model using an RMLSE metric. They obtained the performances which are significantly beyond other methods such as HA (*Historical Average*), ARMA, GBRT (*Gradient Boosting Regression Tree*), HP-KNN (*Hierarchical Prediction K-Nearest Neighbour*), HP-MSI (*Hierarchical Prediction Multi Similarity-based Inference*), especially under anomalous conditions.

In Europe, Chen et al. [14] presented a class of algorithm, *Two-stage Generalised Additive Models,* (TGAMs) intended for demand and availability based prediction on various time scales. Specifically, it estimates the distribution of waiting times for the next available bike or car parking space if the present availability is zero. To test their algorithm, they provided two

case studies in Dublin, Ireland, using BSS data and city parking spaces. Then, they compared their algorithms to LV, HA, and ARMA. Taking the exogenous variables, weather, and time of day, their TGAMs lead to significantly improved performance. They claimed that their predictive algorithm can be used for uncertainty-aware journey planning especially for the needs to wait for the availability of resources such as bikes/docking slot or city parking lot spaces.

In China, Hangzhou is the first city which implemented a BSS, and currently they have one of the largest BSS in the world with around 3000 stations and 60000 bikes. Using Hangzhou BSS data from July to December 2011, Xu et al. [15] proposed a hybrid prediction model that combined the normalization process, improved k-means clustering, and sixth order polynomial smooth *Support Vector Machine* (SVM) to predict the traffic flows. Experimentally, they compared their hybrid model to a *Back Propagation Neural Network* (BP-NN) and pure SVM. They used *Error Rate* (ER) as the performance metric, with results of 8.23%, 5.17%, and 3.57% for BP-NN, pure SVM, and hybrid SVM respectively.

Focussing on using only *Random Forest* (RF), Patil et al. [62] examined demand prediction of BSS data from Washington DC. RF is an ensemble technique in machine learning, also called bagging or bootstrap aggregation [52], which combines many weak learners so as to create a strong learner [85]. It trains learners on a resampled version of training data. Then, using a tuning process to determine the optimal parameters, the authors achieved a better result using RMLSE as a performance metric than the result for RF without the tuning process. In another study, using a similar RF algorithm, Yang et al. [16] conducted traffic prediction of bike check-out and check-in using Hangzhou data. Using CDF and RMSLE as performance metrics, they compared their RF algorithm with three baseline predictors, HA, ARMA, and HP-MSI. They used the first 20 days of each month as a training set and the remaining days for their test set. They proposed case studies for check-out and check-in prediction for rainy summer weekday and sunny winter weekday. Their results show that the RF predictor outperforms the three baseline predictors in most scenarios.

Using more predictors, Giot and Cherrier [47] employed five regression systems to predict BSS usages up to a day ahead in Washington DC using two years of trip data, 2011 and 2012. Those regressors are:

- *Gradient Tree Boosting Regressor* (GTBR) uses an ensemble of weak learners. GBR incrementally builds the regression function to optimize the loss function or minimize the

error metric. Stage by stage, GBR introduces new weak learners to compensate for the shortcomings of current weak learners. Each new learner is a regression tree that is fitted to the negative gradient to the loss function. Each GBR has several hyperparameters that include the number of trees, the depth (or number of leaves), and the shrinkage (or learning rate).

- *Bayesian Ridge Regressor* (BRR) is a variation on Linear Regression with non-linear terms. BRR includes regularization to handle the trade-off between bias (under-fitting due to insufficient model order) and variance (overfitting due to excessive model order) in Linear Regression with an L2 term. This will prevent overfitting to training data by favouring a simpler model and lead to better generalization with lower regression coefficients.

- *Support Vector Regressor* (SVR) which relies on kernel functions to minimize the loss function in order to get most deviations less than a margin of tolerance or threshold. For a non-linear problem, SVR transforms the data into a higher dimensional feature space to make it possible to perform the linear separation. Selecting a particular kernel type and kernel function parameters is usually based on the distribution of input values of the training data and application-domain knowledge.

- *AdaBoost Regressor* (ABR) which is short for Adaptive Boosting uses several decision tree regressors that are fitted iteratively with increasing weights for successive regressors. The latest regressors can fit more detail as the number of boosts is increased with the most difficult samples. This boosting technique allows the regressor to fit the data with less error than a single decision tree.

- *Random Forest Regressor* (RFR) which uses an ensemble approach to building a strong learner from a set of weak learners, which are random regression trees on various subsets of the training set. It employs the averaging of the output from all those weak regression trees as the final regression value. Bootstrapping is used to tune the subset size from the original choice. Therefore, RFR is a type of additive model that makes predictions by combining decisions from a sequence of base models.

They compared the performance of those regressors using RMSE to three baseline classifiers which are *Mean Value, Mean Hour*, and *Last Hour*. They also chose five ranges for feature importance which are *very relevant* (weight > 100), *relevant* (100 ≥ weight > 10), *average* (10 ≥ weight > 0), and *not relevant* (weight ~ 0). Their results show that the

regressors' performances are better than the intuitive baseline system. The best two performing regressors are RR and ABR with the most relevant feature being the bike usage one hour ago.

Table 2.4. Summary of existing works in BSS predictions.

Author	Data source	Timespan	Prediction metrics	Method	Performance
Froehlich et al. [11]	Barcelona	13 weeks	The availability of bikes at each station	HM, LV, HT, BN	Avg Err: 17%, 9%, 9%, and 8%
Kaltenbrunner et al. [70]	Barcelona	7 weeks	The available bikes and vacant docks	ARMA	Mean absolute error: 1.39
Borgnat et al. [63]	Lyon	2 years + 8 months	Hourly number of rented bikes	Linear regression	Mean relative error: 12%
Zeng et al. [97]	New York	1 year	Bike demand prediction	LR, DT, RF, SVR	RMSE: 24.1, 40.1, 25.6, 58.3
Singvhi et al. [96]	New York	3 months	A pairwise bike demand	A regression model	RMSE: 0.42
Li et al. [17]	New York & Washington DC	6 months	The check-out/in of each station cluster	A hierarchical model	ER is reduced by 0.03 beyond baseline method
Xu et al. [15]	Hangzhou	6 months	Traffic flows	BP-NN, pure SVM, and hybrid SVM	*Error Rate* (ER): 8.23%, 5.17%, and 3.57%
Patil et al. [62]	Washington DC	2 years	Demand prediction	RF	RMSLE: 0.5
Yang et al. [16]	Hangzhou	1 years	Bike check-out and check-in	RF, ARMA, HA, and HP-MSI	RMSLE: 0.42, 0.48, 0.46, and 0.46 (check out)
Giot and Cherrier [47]	Washington DC	2 years	BSS usage	ABR, RR, SVR, RFR, and GTBR	RMSE: 102, 79, 336, 336, 312

Table 2.4 summarises the prediction scenarios with different methods, targets, and performance evaluations. One similarity among them is that they all focussed on prediction at aggregated demand or usages based either at station, cluster, or system-wide level. This could be because these metrics are directly related to the BSS operation. On the other hand, trip prediction for other mobility modalities have concentrated more on individual prediction as shown in subsection 2.1.7. Currently, individual user based prediction in BSS has not been widely explored either in terms of its accuracy or in terms how to use this prediction information to improve system operations.

2.3.8. BSS Journey Advisor

One potential application of prediction in BSS is for helping users to plan and navigate their trips. Yoon et al. [108] proposed a personal journey advisor application for BSS in Dublin. For a given origin and destination, their application suggests the best pair of stations to be used to pickup and return the bikes. This is in order to usefully minimize the overall walking and biking travel time as well as maximizing the probability to find available bikes at the first station and vacant return slots at the second one. An example, in Dublin some bike stations can experience no bikes or no empty slots for 3-4 hours a day. Reducing this imbalance is an optimization problem. To solve it, they modelled the real mobile renters' behaviour in terms of travel time and used the predicted availability at every bike station to choose the pair of stations which maximizes their measure of optimality. To develop the application, they built a spatiotemporal prediction system that is able to estimate the number of available bikes for each station in short and long term intervals, outperforming already developed solutions. The prediction system is based on an underlying spatial interaction network among the bike stations and takes into account the temporal patterns included in the data. They applied a modified ARIMA model by considering spatial interaction and temporal factors to predict the available bikes/docks at each station. One of their contributions is to deal with spatiotemporal prediction by using signals from neighbouring stations and seasonal trends.

Recently, Yang and Zhang [115] have proposed a novel travel adviser to predict the number of available bikes after a given period of time so as to optimize users' travel choice in Barcelona. They used an *improved Backpropagation Neural Network* (iBP-NN) and *Genetic Algorithm* (GA) as their prediction algorithm and included a novel prediction model considering the impact of surrounding stations. They used two parameters, *Normalized Available Bikes* (NAB) and *Normalized Activity Score* (NAS) to indicate the time pattern of bike stations. Here, NAB can effectively reflect the percentage of available bikes, while NAS can effectively indicate how active a station is at a given time *t*. By considering bicycle numbers at surrounding stations which can be explicitly factored into the prediction model, the algorithm results in significant gains in terms of prediction accuracy. Their experimental outcomes demonstrate that their novel approach can appropriately handle the BSS non-linear prediction problem.

Meanwhile, Zhao et al. [114] developed *GreenBicycling*, a smart-phone application to provide mainly context-aware BSS information as well as to promote healthier lifestyle choices in Hangzhou, China. It provides simple interfaces for users to know the number of bikes currently at any stations or near any location, the likely number of docking slots upon arrival, and the shortest path and the distance between two specific stations. They used a BP-NN as a predictor. While for the healthy lifestyle context, they provide a quantitative calorie estimate for journeys.

These three applications depend on the OD input from users so that the applications can give information and projections related to the specified origin and destination. However, if destinations can be accurately predicted when a user picks up a bike, the application could give information proactively. This should be possible for highly predictable users with sufficient trip history to give accurate predictions, and this area is worth further investigation.

2.4. Review Summary

The review sections above provide an overview of recent work on human mobility as well as on BSS studies. In human mobility studies, many methods, models, and metrics have been proposed from various sources of real world data to reveal the spatiotemporal characteristics as well as the limitations of human mobility. A certain degree of regularity in human mobility is the basis for the majority of the studies. Similarly in BSS studies, a wide variety of spatiotemporal analyses have been conducted using BSS data from many cities. Those are intended to help the operators provide the best service. However, there are still some approaches in human mobility studies that have the potential to be implemented in BSS data as a complementary investigation to existing studies such as mobility motifs, entropy, and predictability.

Generally, there are at least three BSS entities that have been used as the topics of analysis which are stations, users, and external factors. For stations, there are studies about system design and deployment that are mostly from surveys. However, there is still a room for improvement, especially in using the user movement behaviours that can be revealed from their trip data to determine the practical distance between stations that could have significant impact for users' station preferences. For users, most studies undertook analysis at an aggregate level instead of being individually based. This could be due to non-availability of user identification in most BSS shared data because of privacy concerns. Further investigations are also needed to

explore about to what extent this individual based analysis can be used to improve the BSS services. While for demand or usage prediction, there are three level of analyses that can be conducted which are system-wide, clusters, and stations level. Further studies are needed to explore to what extent these three levels of analysis can assist with the BSS operation. Similar daily patterns of usage are observed on weekdays and different patterns on weekends. This periodic nature of BSS dynamics on a daily basis and on a weekly basis could be a promising technique to be further explored as a basis for prediction.

The next chapter will investigate the research gaps that can be identified from this literature review in more detail, and then the research questions for this thesis will be formulated. Subsequent chapters will present the research methodology, detailed results and critical analysis needed to answer each of the research questions in turn.

CHAPTER 3
GAPS AND RESEARCH QUESTIONS

The problems of spatiotemporal data-driven analysis using BSS data can be viewed from three aspects: stations, users, and external factors. Stations, which are the core part of the system, are spread non-homogenously across a city, and they have different capacities and distances between each other. Users have different behaviours and they are the source of the dynamicity of the system. On the other hand, external factors such as seasons and weather will influence when users hire bikes adding to the uncertainty of system use. Therefore, an understanding of how to incorporate the stations, users, and external factors together to characterise current behaviour, predict future behaviour, and use the results to improve the BSS deployment, services, and operation remain challenging. This chapter will first analyse the gaps in the current state-of-the-art based on the previous Literature Review chapter. Next, these gaps will drive the formulation of the key research questions (RQ) for the thesis. Then the individual research tasks associated with each RQ will be described.

3.1. Gap Analysis

From the literature review of human mobility and BSS studies in the previous chapter, there are gaps that can be identified which suggest directions for further research. In some cases, existing analytical methods for human mobility studies in other fields have not been applied to the analysis of BSS data. Furthermore, the impact of human mobility analysis needs to be more widely applied to practical understanding of BSS system operation.

3.1.1. Gaps in spatial analysis

While temporal aspects of BSS data have been widely analysed, there is scope for more investigations in the spatial analysis aspects, especially for how nearby stations affect each other, what is referred to in this thesis as stations' **neighborhood ties**. As a complex dynamic network, if a station is out of service (e.g. temporary shutdown) or in an imbalance state (e.g. full or empty), the nearby stations and stations which have high number of connections to that station are likely to be affected. The spatial distances over which stations influence their neighbours, and the metrics that can be used to analyze these influences, are not clearly known from existing work.

Another example of neighborhood ties is when users choose to visit a station close to another station they often visit. For example, a user might visit any one of a number of stations close to their place of work to pick up a bike. In this case, spatial motifs analysis, which has not been previously investigated for BSS studies, may give useful insights. For example, a worker may have a mobility motif over one day of **home → work → home**. However, the BSS stations that are used near home and near work might not be the same on both trips. Analysis of the distance between different stations used for such a motif, may give useful insights into the distances that users are likely to travel to go to an alternate station. As well as simply understanding stations' spatial links, the question of how to use this spatial neighborhood ties knowledge in BSS design and deployment has not been investigated. Knowing how far users are likely to travel to find alternative stations because of shutdowns or imbalances can assist in the location of stations, and could be used to provide notifications to users about alternate nearby stations.

Nearby stations are determined by the distance or time that a user has to travel between those stations, either on foot or by bicycle. Most BSS studies in the literature review have used the straight-line, Euclidean distance to measure the separation of stations. One has used Manhattan distance [73], which maps well to grids of roads that are aligned with the Manhattan distance axes, but gives poor estimates for roads that are not so aligned. Both of these distance measures give a poor estimate of inter-station travel time when there are obstacles between the stations, such a railway line or river that require a roundabout route between stations. Waypoint distance, based on the shortest feasible route using available paths, may provide a better alternative for BSS spatial analysis.

3.1.2. Gaps in users analysis and prediction

Understanding groups of users who share common behavior is important because their spatiotemporal collective trends can be potentially used to improve the BSS services, such as providing customized notifications. Some studies have used the explicit subscription and demographic information to cluster users but these groups of users may have greatly different usage patterns. One study has used spatial density estimation to cluster users, others have used temporal information, such as the average number of trips in certain period of times. These may give more homogenous behaviour of users in a cluster. However, after clustering, no studies have undertaken detailed further investigations regarding the characteristics of each

cluster to understand their homogeneity and regularity in terms of spatiotemporal mobility as well as entropy and predictability.

If some users are highly predictable then customized, personalized notifications can be provided based on their expected usage. For example, users can be notified if their expected destination is likely to be imbalanced. The ability of existing clustering approaches to capture highly predictable users has not been investigated. Current BSS studies have not used collective trends of clusters to assist in predictions of future use. The predictability of individual next locations, and whether these next locations are highly determined by users' current locations or previous trip history, are not known.

Using mobility predictions from statistics for a whole cluster maybe useful for providing information to users who have no history for particular locations. These trends may vary over the course of a day. So, the ensemble of next location prediction using individual history and population trends combined with temporal features for every station appears to be a useful area for further investigation. Additionally, how to use this knowledge about users in the high predictability clusters to improve the quality of BSS services is also worthy of further investigation.

3.1.3. Gaps in system level analysis and prediction

As BSS system usage exhibits a regular cyclostationary pattern [63] over the week, this pattern may provide a basis for prediction at a system-wide level. The current system usage can be thought of as a combination of this regular weekly pattern, plus some current perturbations, e.g. due to bad weather. So prediction can be framed as the problem of finding the underlying weekly pattern and predicting this perturbation. In terms of conventional time-series analysis, if the week is divided into 7 x 24 hours, then the BSS usage consists of 168 stationary processes, and the usage at, say, 9 am - 10 am Mondays forms one stationary process. The whole time series is called cyclostationary. One particular hour's usage is a random number drawn from the statistics of that hour's stationary process. So it seems useful to investigate techniques from time-series analysis such as estimating the underlying the historical reference, and then seeing effects such as bad weather as a predictable perturbation from the underlying historical reference.

Although this technique of seeing usage as an underlying historical trend plus a perturbation is common in many time-series forecasting problems, it has not been implemented

in BSS studies previously, and its usefulness is unknown. Previous BSS studies have directly predicted the total usage value [15-17]. There appears to be scope to make better prediction of system usage, and to investigate what factors and features significantly influence the prediction performance. It is also useful to understand whether this technique can be implemented at different levels of BSS system: system-wide, subsystem level (clusters) and individual station level. There is also a need to investigate which performance metrics are most appropriate for giving insight into prediction accuracy at these different levels of operation. Most importantly, how improving these three levels of prediction can help to improve BSS operations also requires investigation.

3.2. Research Questions and Tasks

Based on the gap analysis above, this study formulates four research questions (RQ) that form the focus of this thesis. Each RQ below is followed by the general methodology that will be applied to answer that question. The more specific methodologies for individual tasks will be described at the beginning of each subsequent result chapter. In addition, preliminary data analysis is first conducted to understand the basic spatiotemporal characteristics of the data.

RQ1: ***What insights can be gained from the BSS stations' neighbourhood ties?***

This investigation will focus on two spatial metrics which are spatial distance of mobility motifs and spatial impact of temporary stations shutdown. First, the most common BSS mobility motifs will be determined followed by the calculation of the distance between stations to identify what inter-station distance corresponds to a neighbourhood. Second, some stations shutdown cases will be examined, and then the impact distance for nearby stations will be calculated. To measure the impact distance, looking at changes in usage before, during, and after shutdown will be investigated. This will give an insight about how the influence of a station decays with distance, and identify typical impact distances. If these two approaches give similar distances, then this distance will give useful insights about how users respond if they have to choose other stations instead of their commonly visited stations. Because this investigation involves distance, a comparison of the usefulness of the widely used Euclidean distance and the proposed waypoint distance will be also undertaken.

RQ2: ***To what extent can clustering identify highly predictable users, what are the maximum limits of predictability, and how can these be achieved?***

If users can be grouped into clusters so that frequent, highly predictable users are all in one cluster, then additional individualised notifications can be sent to those based on their expected behaviour. This might, for example, request users to return their bicycle to a particular station in a neighbourhood which is currently almost empty.

User clustering in this study aims to cluster users mainly based on the temporal similarity of their mobility behaviour. The total trips for a user will show how frequently an individual uses BSS since frequent users have more historical data on which to base future prediction. The pattern of trips across each day will reflect the regularity of daily routines. Users with a regular routine are likely to be more predictable. In order to decide on exactly what temporal characteristics might identify frequent users with daily routines, some preliminary data analysis will first be carried out to understand the daily usage patterns of the BSS. Then, users can be clustered based on their temporal characteristics, and appropriate labels will be given for each cluster.

To test if this proposed clustering technique can capture the highly predictable users, entropy and predictability analysis will be applied. From an information theory perspective, the entropy results will identify clusters where the predictability of a user's next location is improved by consideration of their past trip history. If predictability is improved by using past history, then those clusters are said to have Markovian traits. The predictability results will identify the maximum prediction accuracy that could be achieved. An ensemble predictor will be investigated which uses individual trip history where available and collective trends of clusters where history is absent for that user, in order to predict individual trips. To understand the dynamic of prediction accuracy, the results will be presented on an hourly basis, and also on a daily basis. This will indicate if there are particular times of day when trip predictability is high, and so individual personalised notifications are likely to give useful information to users at those times. To understand the performance of the prediction algorithms, prediction accuracy will be compared to the theoretical limits identified by the predictability level of each cluster.

RQ3: ***To what extent can the cyclostationary pattern of bicycle sharing systems be used to conduct and improve the prediction of BSS usage and which factors are most effective for good prediction?***

The detailed scenario of the proposed deviation-based prediction in the cyclostationary pattern of the BSS will be investigated to identify the regularity of the underlying patterns, the best historical reference values, and the factors which affect perturbations to those patterns. Since the perturbations are unlikely to be a linear combination of factors, and since there are many factors that could be taken into account, the prediction will use machine learning approaches. This needs dataset splitting for training, validation, and testing. Feature selection and feature importance will also be investigated. Prediction will be investigated at system-wide, cluster level, and individual stations. In addition to the London dataset, the same techniques will be applied to the Washington DC BSS data, to investigate how generalizable these techniques are. The absolute and relative level of prediction errors will be used to analyse the prediction performance.

RQ4: ***How can the stations' neighbourhood ties and highly predictable clusters knowledge, as well as the system-wide predictions at different levels, benefit the BSS deployment, services, and operations?***

This RQ investigates the practical application of answers from the previous three RQs. First, the stations' neighbourhood ties results could be useful for BSS deployment and design because the expected outcome from RQ1 is the appropriate distance between stations within a neighbourhood. Second, identifying a cluster of highly predictable users from RQ2 enables appropriate individualized notifications that would improve user experience and station operations. Third, the three level predictions from RQ3 could be beneficial for BSS operations to optimize proactive rebalancing. The potential practical contributions of each RQ will be illustrated with examples and the total practical contribution of this work will be summarised in this RQ.

CHAPTER 4
PRELIMINARY DATA ANALYSIS

Intra-city scale mobility typically has trips with relatively short duration and short distance characteristics. Understanding mobility first requires measuring mobility, and one method for doing this is to use the origin and destination information of trips. One useful source of digital urban mobility information is BSS trip data which has exact OD information for each bicycle trip. The first step of this investigation, described in this chapter, will be exploratory data analysis and visualization of BSS trip data, which will give insights into subsequent data analysis and prediction. This preliminary data analysis will investigate the underlying properties, metrics, patterns, and trends of BSS mobility dynamics over time and space at both individual and aggregate level.

This data exploration primarily consists of two parts: temporal and spatial analysis. Each will be supported by relevant visualization to highlight significant aspects of the data. Firstly, temporal analyses will explore mobility patterns at hourly, daily, weekly and monthly time scales to look for any consistent and regular patterns at different time resolutions. Knowledge of times when the BSS experiences high demand over the course of the day and the size of that demand will point to times that are likely to cause usage imbalance. The complexity of the distribution of trip durations will be investigated. Not only individual trip durations, but also the intervals between trips may have useful information such as identifying daily commuting behaviour. Secondly, spatial analyses aim to observe geographical differences in the behaviour of BSS users. Trips will be analysed as to whether they follow any preferred flows and directions, what the typical ranges of travel are, and to what extent they cause imbalance states at stations. To understand whether particular stations are subject to frequent visitations, revisitation analysis will be conducted, so that the *exploration* and *preferential return* ratio [24] can be understood. All these spatial analyses could be useful to identify the areas where high demand frequently occurs.

Analyses such as idle times, distance expansion growth and revisitation need an individual sequence of trips for individual users, which is often not publically available. Hence, this exploratory data analysis is based on London BSS data from August-November 2012, which includes individual, anonymized user IDs associated with each trip as described in the next

section. This exploratory data analysis will provide some insights for further analyses in the research questions as depicted in the research workflow in Chapter 3.

4.1. Datasets

4.1.1. Main dataset

This section presents the main dataset which is the individualized user trip history of *London's Cycle Hire Scheme* (LCHS)[4] from August – December 2012.

There are six major data fields as shown in Table 4.1: 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 files, Table 4.2 and Table 4.3 respectively, which are linked to each station and user in Table 4.1.

Table 4.1. London bike data structure.

User ID	Pickup Data		Return Data		Duration (minutes)
	Station	***Pickup time***	***Station***	***Return time***	
1465	251	2012-08-01 06:34	506	2012-08-01 06:40	5.75
1507	239	2012-08-01 07:05	44	2012-08-01 07:15	9.95
1465	506	2012-08-01 16:45	251	2012-08-01 16:51	6.00

Table 4.2. Station geolocation.

Station ID	Street Location	Geolocation	
		Latitude	Longitude
44	Bruton Street	51.510737	-0.144165
239	Warren Street	51.524438	-0.138019
251	Brushfield Street	51.518908	-0.079249

Table 4.3. User type.

User ID	Type	Description
1465	1	Registered
1507	1	Registered
7086221	0	Unregistered

For characterization and prediction studies, the dataset will be divided into training and testing datasets. Training dataset, D1, contains the 2012 summer to autumn trips from 1st August 2012 to 30th November 2012 (122 days~17 weeks), while testing dataset, D2, spans from 1st to 23rd December 2012 (23 days). Originally, the dataset covers 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 trips with an unrealistic duration (< 1 min or > 24 hrs). This eliminates data that is not valid for the analysis. After removing the 5.25% of affected data, the dataset has 2,805,718 trips with 566,456 users.

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

While more recent datasets are available for London, and for other cities, this 2012 dataset is the only publically available data that includes unique user IDs associated with each trip. Data from other BSS datasets does not identify individual users, and so cannot be used for individual data-driven clustering or prediction. Therefore this five-month London dataset is used for the majority of the analysis in this study.

The London dataset itself categorizes users into two classes: *unregistered* and *registered* users. This relates to how they subscribe and use the system. Most of the users, 89%, are *unregistered* users who correspond to 43% of the trips, while 11% of *registered* users have 57% of trips. This uneven division comes from the 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. *Registered users* are much more frequent riders compared to *unregistered* ones.

For system prediction in RQ3, another BSS data set will be used for comparison which is from Washington DC BSS[5] in the same period, which does not identify individual users. This is to investigate the more general applicability of the system-wide prediction approaches.

4.1.2. Complementary Dataset

Another dataset used in this study mainly for RQ3 are daily historic records of weather[6] in Central London as well as Washington DC. There are four features: *Temperature* in °C, *Humidity* in %, *Wind speed* in km/hr, *Rainfall level* in mm/hr. These weather logs are used as an independent data stream for validation of clustered users' behaviour when dealing with weather conditions as well as for input features for system-wide prediction.

4.2. Temporal Analysis

This temporal analysis section will investigate BSS dynamics at various time resolutions beginning with usage density distributions at hourly, daily, weekly, and monthly resolutions. This is followed by waiting times and trip duration analysis. The analysis will identify periods when usage is low, moderate, high, or reaches a peak with their corresponding level, and will also investigate the periodicity of these patterns. Furthermore, the characteristic of trip duration will also be investigated to see if it follows the heavy-tailed distribution that has been observed in other mobility studies.

[5] Downloaded from the capital bike share website (https://www.capitalbikeshare.com/system-data)
[6] Downloaded from the wunderground website (www.wunderground.com)

4.2.1. Daily Patterns

The daily BSS usage pattern for the chosen period is shown in Figure 4.1, which shows the density distribution of three BSS entities which are *bikes*, *users* and *trips* on a daily basis along 122 days ~ 17 weeks ~ 4 months of the learning period (01 Aug – 30 Nov 2012). Data has been cleaned and preprocessed as described in subsection 4.1.1. While numbers of bikes in the system tend to be constant, users and trips are highly variable with a generally decreasing trend starting from the last week of September. Usage on weekdays is more than on weekends, and trip numbers are higher than user numbers.

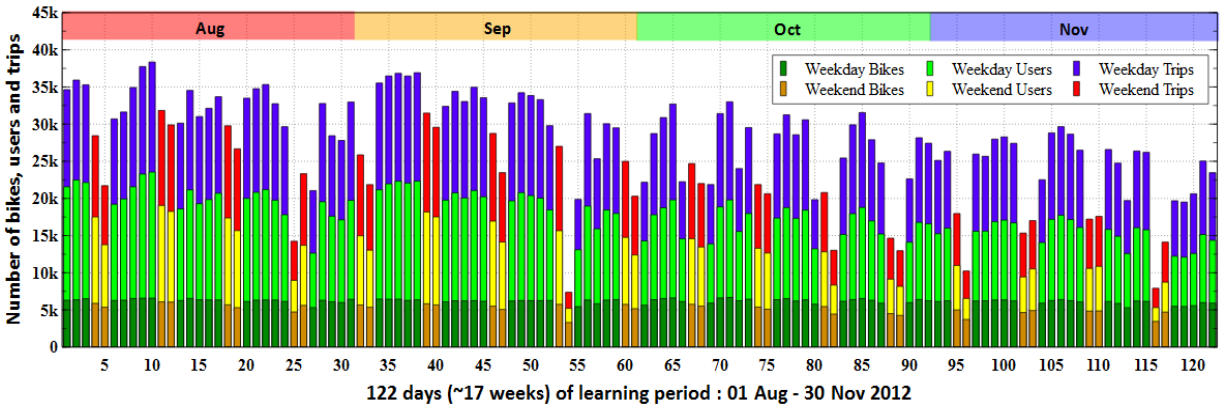


Figure 4.1. Daily numbers of bikes, users and trips.

These facts indicate at least three preliminary propositions that relate to the daily contexts. Firstly, there is significantly more usage on weekdays, suggesting that a significant proportion of use is associated with urban commuting. Secondly, many users hired bikes more than once a day, and so the idle times between trips may be useful to investigate. Thirdly, usage decreases towards the end of the year, i.e. towards winter. This could be related to the weather.

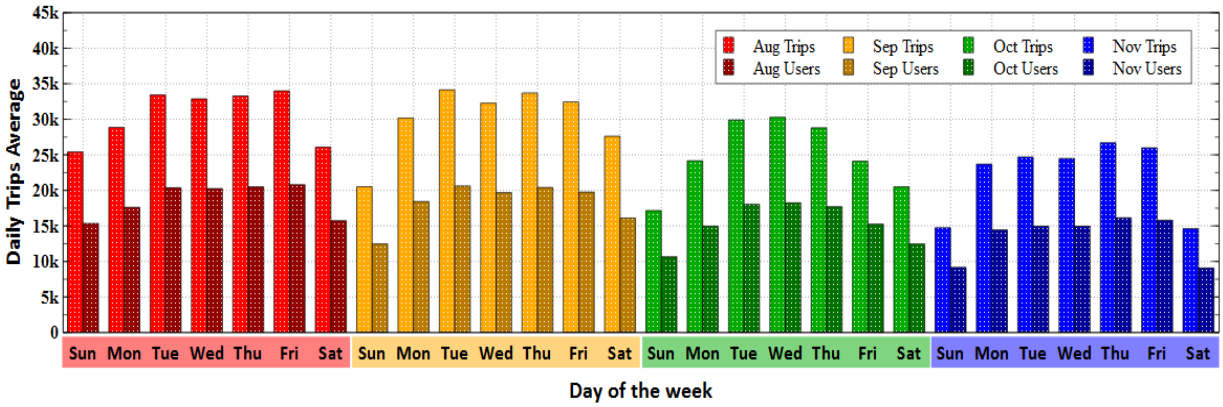


Figure 4.2. Weekly averages of trips and users per month.

Figure 4.2 summarizes the distribution of trips and users by day of the week for the four months under consideration. Generally, the monthly trips average (N) is ordered $\mathbf{N^{Aug} > N^{Sep} > N^{Oct} > N^{Nov}}$. The average number of trips is about 38% more than the average number of users, nearly stable for all days in all months. This means that at least 62% of daily users make only one trip.

4.2.2. Hourly Patterns

Hourly patterns show how usage varies over the course of one day. Figure 4.3 displays the first three weeks of data, where red colors are weekends and green are weekdays. Here, weekends have only one peak in the middle of the day, while weekdays have two peaks, in the morning and afternoon. This pattern is similar in the rest of the data. Therefore, hourly pattern is ***cyclostationary*** (Borgnat et al. [63]), i.e. it contains similar repeating patterns on a daily basis both for weekdays (two peaks) and a different daily pattern on weekend days (one peak). Having hourly sharp usage peaks may produce asymmetric flows in the system that potentially create imbalance states in bike distribution if no effective redistribution is undertaken. The daily averages of these hourly patterns grouped by month are presented in Figure 4.4.

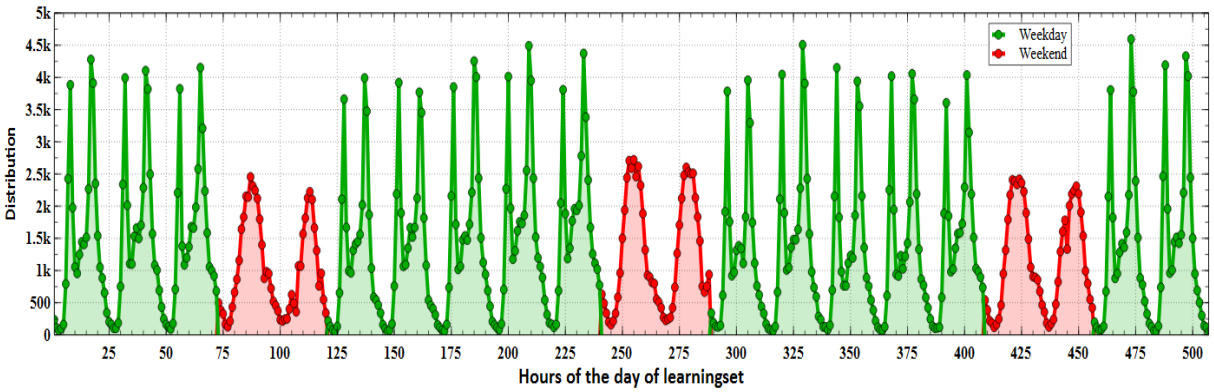
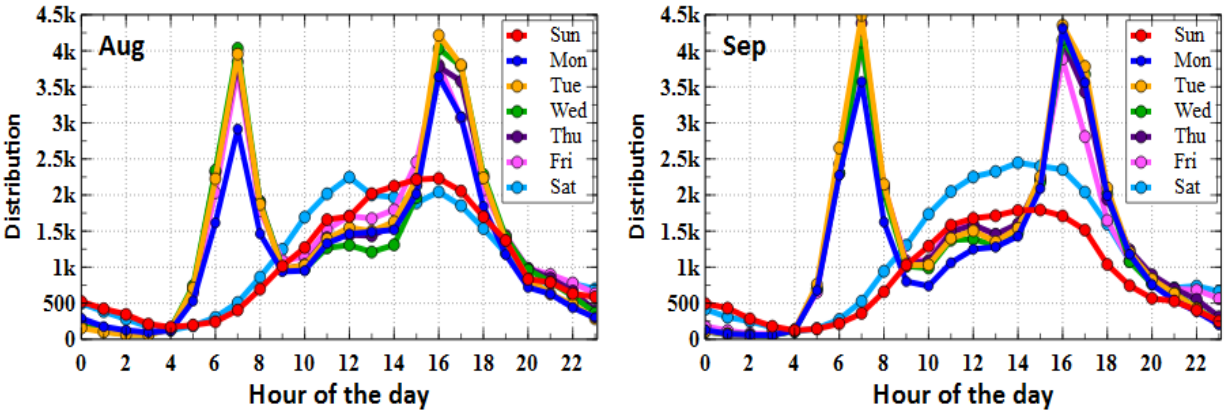


Figure 4.3. Hourly trip patterns.

There are two sharp peaks on weekdays and only a moderate peak on weekend days signaling busy times. On weekday mornings, the peak occurs between 5 am to 9 am and in the afternoon it occurs from 3 pm to 7 pm, while it is distributed thorughout the middle of the day from 10 am to 6 pm on weekends. The weekday peaks are at times when people usually travel to their workplaces in the morning and leave their workplaces in the evening. In other words, a commuting characteristic is clearly shown by BSS hourly patterns on weekdays. The fact that there are only two peaks also shows a socio-cultural aspect where not many people use bikes

during weekday lunch times in London, which is different to a study conducted by Froehlic et al. [11] in Barcelona, Borgnat et al. [63] in Lyon, and Cóme et al. [86] in Paris that found three spikes during weekdays, in the morning, at lunchtime and late afternoon. On weekends, a moderate peak appears in the middle of the day indicating leisure use. Again, this is different from weekend patterns of Barcelona which has two peaks around midday and in the afternoon [11], but it is somewhat similar to Lyon which has one peak concentrated in the afternoon.

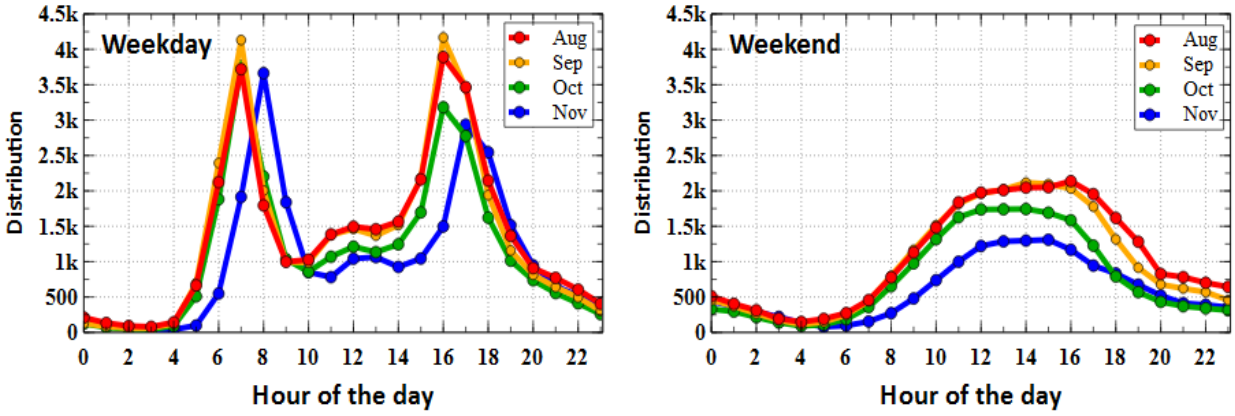


(a) August *(b) September*

(c) October *(d) November*

Figure 4.4. Average of hourly trip patterns per day of the week for each month.

In the weekday afternoons of all months, the afternoon commuting peak is lower but broader, showing that there is a greater spread of commuting times in the afternoon peak compared to the morning peak. BSS data uses GMT (*Greenwich Mean Time*) or Universal Time for recording trip data. During August – October, the UK uses British summer time (1 hour ahead of GMT), so the November peak (blue line) is shifted by 1 hour when UK time returns to GMT as shown more clearly in Figure 4.5.

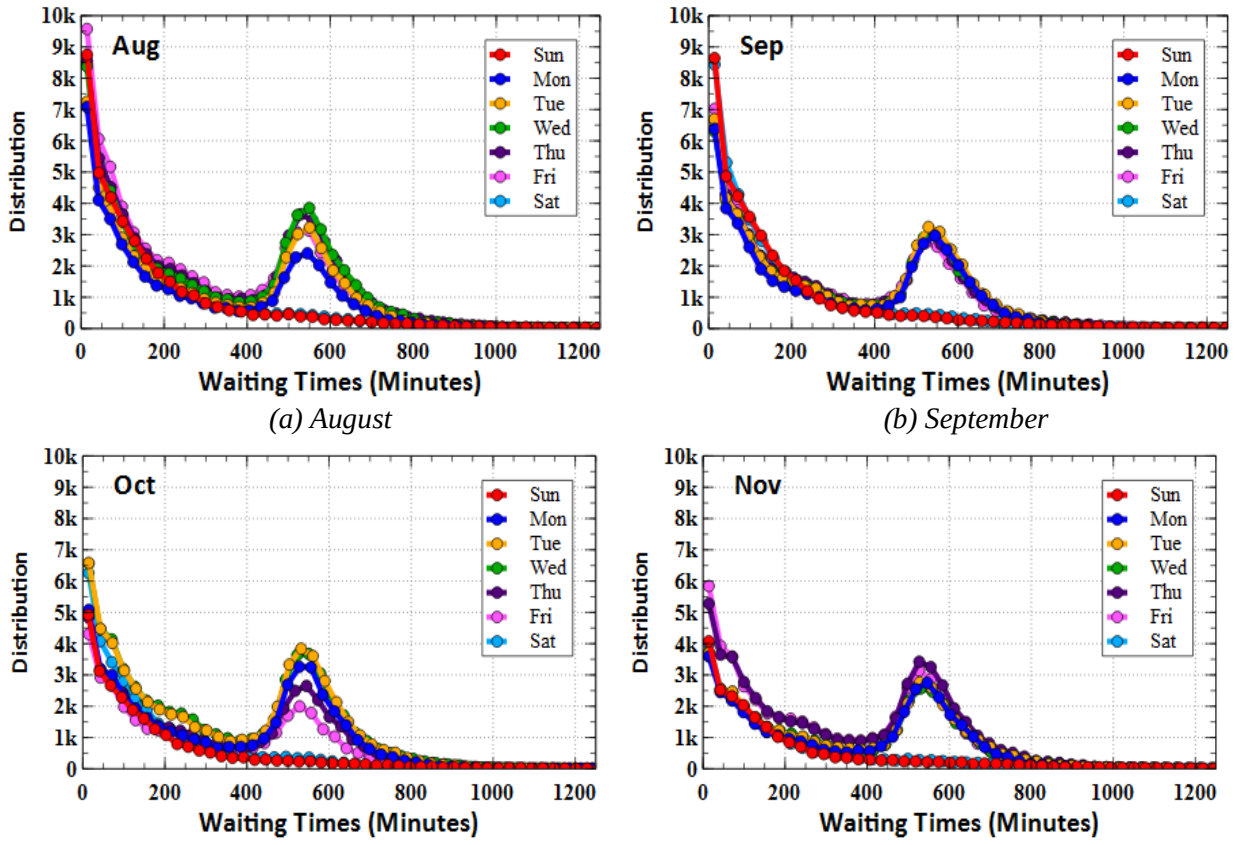


(a) Weekday *(b) Weekend*

Figure 4.5. Weekday and weekend average of hourly trips patterns with one hour shifted on November.

4.2.3. Waiting Times

As shown in subsection 4.2.1, there are a proportion of users who have more than one trip a day. Here, the time between trips is defined as the *waiting time* (WT) which specifically is the period between one return and the next pickup of an individual within one day.



(a) August *(b) September*

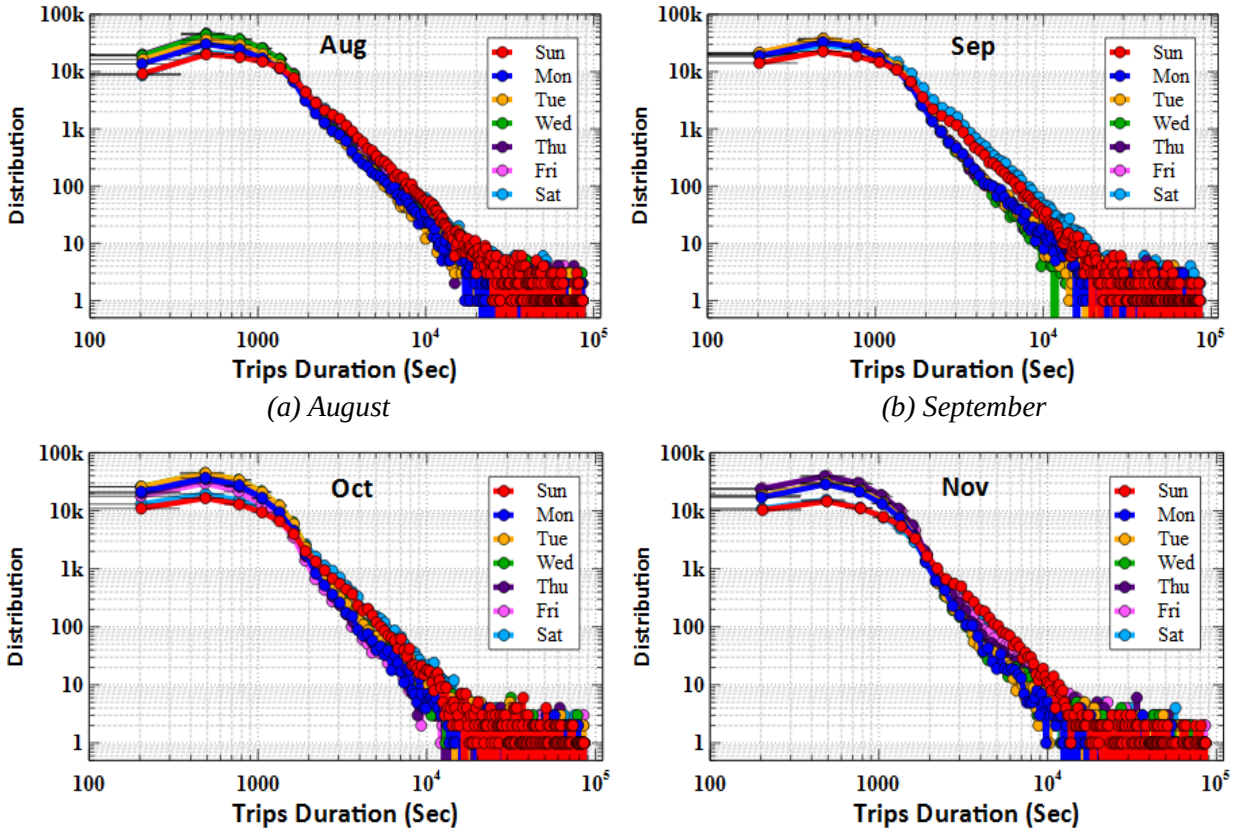
(c) October *(d) November*

Figure 4.6. Daily waiting times patterns.

For all days in all months as shown in Figure 4.6, there are many waiting times lower than 100 minutes (~1.5 hours). However, there is also a peak of WT between 400 – 750 minutes (6.7 – 10.8 hours) with the peak around (500 minutes ~ 8.33 hours) on weekdays, while this peak does not appear at all on weekend days. The length of this waiting time conforms to the common working time of around 8 hours on weekdays. Again, this suggests commuting behaviour where there are many people use a bike to travel to work in the morning and from work in the afternoon. People who have this characteristic ("commuters") will be further analysed in Chapter 6.

4.2.4. Trip Duration

Trip duration is the time from picking up a bike to returning it back to the system. Here, it is calculated in seconds. Figure 4.7 shows the trip durations (in log-log scale) are mostly short and have a *fat tail* on the right side. These so-called *heavy-tailed characteristics* mostly occur after 10000 seconds (~2.78 hours). This means that there are many short trip durations, and few, but non-negligible, long trip durations (perhaps tourist using the bike for sightseeing over several hours). This shows the complexity of human mobility.



(a) August *(b) September*

(c) October *(d) November*

Figure 4.7. Daily trip duration patterns in log-log scale.

Statistically, Table 4.4 lists trip duration averaged by the days of the week. In all cases, the *standard deviation* (STD) is much larger than the *average* (AVG), around twice the average, which is typical of fat tail distributions as illustrated above. Furthermore, there are three other trends that could be highlighted from the table. First, trip duration (T) on weekend days is longer than on weekdays, $\mathbf{T^{WD} < T^{WE}}$. On weekends either people tend to travel further in distance or they travel more slowly. This will be examined in the distance and speed analysis in the next sub-section. Second, duration (T) in August is higher than September, September is higher than October, and October is higher than November, $\mathbf{T^{Aug} > T^{Sep} > T^{Oct} > T^{Nov}}$. This shows that people tend to travel for less time as winter approaches. Third, all the average figures are less than 30 minutes (1800 seconds), the limit of charge-free usage. Based on the data shown in Figure 4.7, more than 92% of trips are less than 1800 seconds. So the free rental period has a significant effect on usage characteristics.

Table 4.4. Average and standard deviation of daily trip duration.

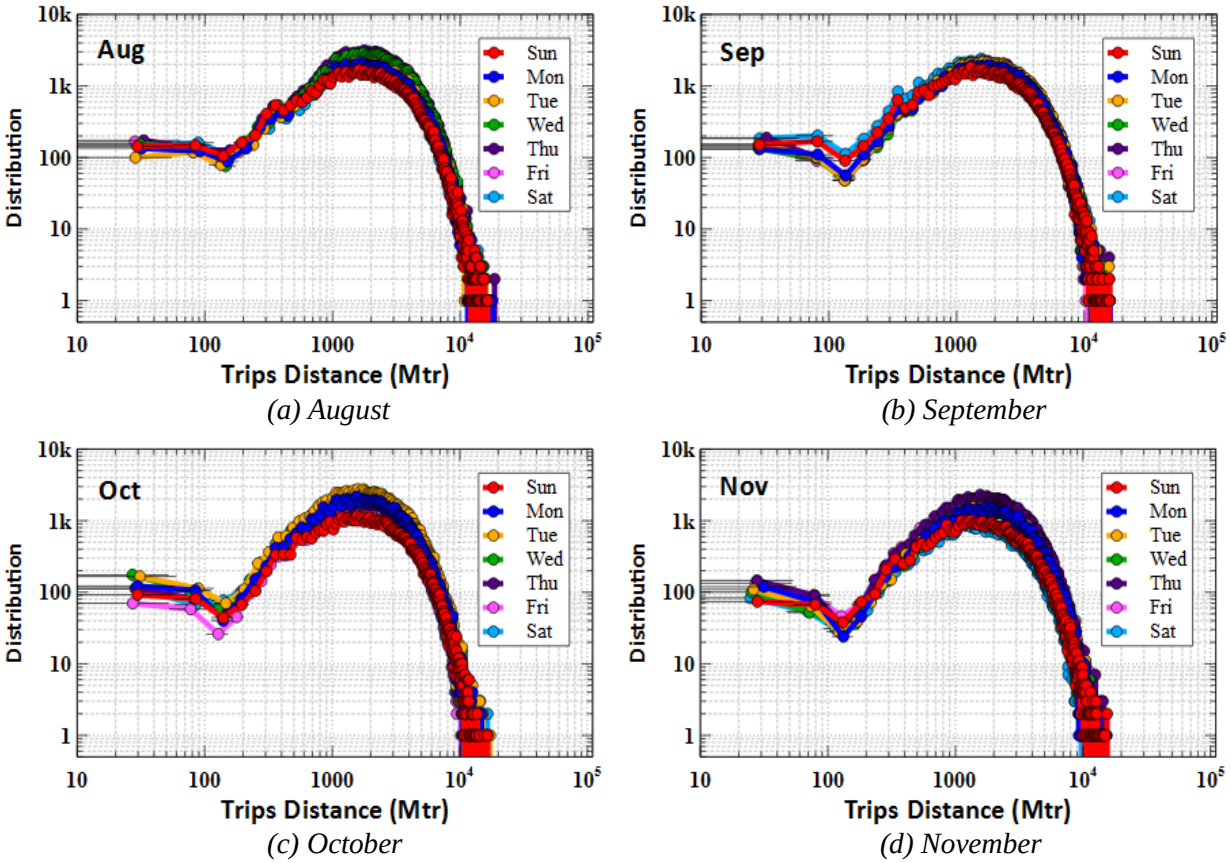
No	Day	Average (seconds)				Standard Deviation (seconds)			
		Aug	Sep	Oct	Nov	Aug	Sep	Oct	Nov
1	Mon	1119.8	952.6	850.7	814.1	2286.8	1901.6	1932.7	1619.7
2	Tue	1066.3	931.7	858.3	807.9	2176.6	1805.2	1739.7	1595.0
3	Wed	1050.1	912.5	860.8	819.0	2147.5	1742.7	1690.1	1783.0
4	Thu	1060.0	925.5	854.7	831.4	2116.0	1841.0	1744.1	1753.8
5	Fri	1112.1	943.3	859.1	842.0	2435.6	2054.4	2122.7	1737.1
6	Sat	1502.0	1315.5	1172.3	1045.4	2966.5	2541.9	2598.9	2558.4
7	Sun	1530.0	1290.3	1187.3	1118.9	2923.8	2618.3	2695.3	2460.0
Avg Weekday		**1081.6**	**933.1**	**856.7**	**822.9**	**2232.5**	**1869.0**	**1845.9**	**1697.7**
Avg Weekend		**1516.0**	**1302.9**	**1179.8**	**1082.1**	**2945.1**	**2580.1**	**2647.1**	**2509.2**

4.3. Spatial Analysis

Spatial analysis focusses on the dynamics of mobility metrics that relate to geographical distribution of movement. This gives knowledge about how far people travel, flows and directions, where the most pickups, returns and imbalances occur, and which OD routes are most popular. This section begins with trip distance followed by station usage activity, trip link analysis, and revisited stations.

4.3.1. Trip Distance

From the origin and destination geo-location of stations, the Euclidian distance which is origin destination *straight line distance* (SLD) can be computed. However, in practice, the trip distance is determined by the path which rider chooses. Unlike the fat-tail trip duration distribution, the distance distributions as shown in Figure 4.8 are less fat-tailed but more skewed. This suggests the variability of distance is not as much as variability in the duration distribution.



(a) August *(b) September*

(c) October *(d) November*

Figure 4.8. Daily trip distance per month in log-log scale.

The average and standard deviation of daily trip distance are listed in Table 4.5. The standard deviation for all days and months is less than the average, around 40%. Comparing each month, people tend to travel further in August, $\mathbf{D}^{\mathbf{Aug}} > \mathbf{D}^{\mathbf{Sep}} > \mathbf{D}^{\mathbf{Oct}} > \mathbf{D}^{\mathbf{Nov}}$. This monthly decreasing trend is similar with trip durations in each month. Conversely, weekday and weekend trends are different for distance and duration. Here, people have a tendency to travel further on weekdays than weekends, $\mathbf{D}^{\mathbf{WD}} > \mathbf{D}^{\mathbf{WE}}$. This means that people ride faster (further distance in shorter time) on weekdays. This phenomenon is expected because weekday commuters are more hurried while weekend leisure users are more relaxed.

Table 4.5. Average and standard deviation of daily trip distance.

No	Day	Average (metres)				Standard Deviation (metres)			
		Aug	Sep	Oct	Nov	Aug	Sep	Oct	Nov
1	Mon	2836.2	2752.7	2705.5	2706.2	1692.2	1646.0	1592.1	1580.8
2	Tue	2863.4	2773.9	2710.5	2692.2	1676.8	1647.1	1602.5	1571.0
3	Wed	2898.7	2762.2	2699.0	2679.0	1679.3	1646.9	1588.6	1564.1
4	Thu	2875.8	2746.4	2692.1	2673.2	1671.9	1628.6	1583.0	1564.2
5	Fri	2839.2	2705.5	2621.5	2638.7	1662.4	1609.1	1547.2	1550.9
6	Sat	2813.5	2662.3	2582.7	2509.1	1720.2	1681.2	1608.4	1571.9
7	Sun	2788.3	2645.3	2622.7	2592.1	1744.8	1691.1	1638.3	1634.1
Avg Weekday		**2862.7**	**2748.1**	**2685.7**	**2677.8**	**1676.5**	**1635.5**	**1582.7**	**1566.2**
Avg Weekend		**2800.9**	**2653.8**	**2602.7**	**2550.6**	**1732.5**	**1686.2**	**1623.3**	**1603.0**

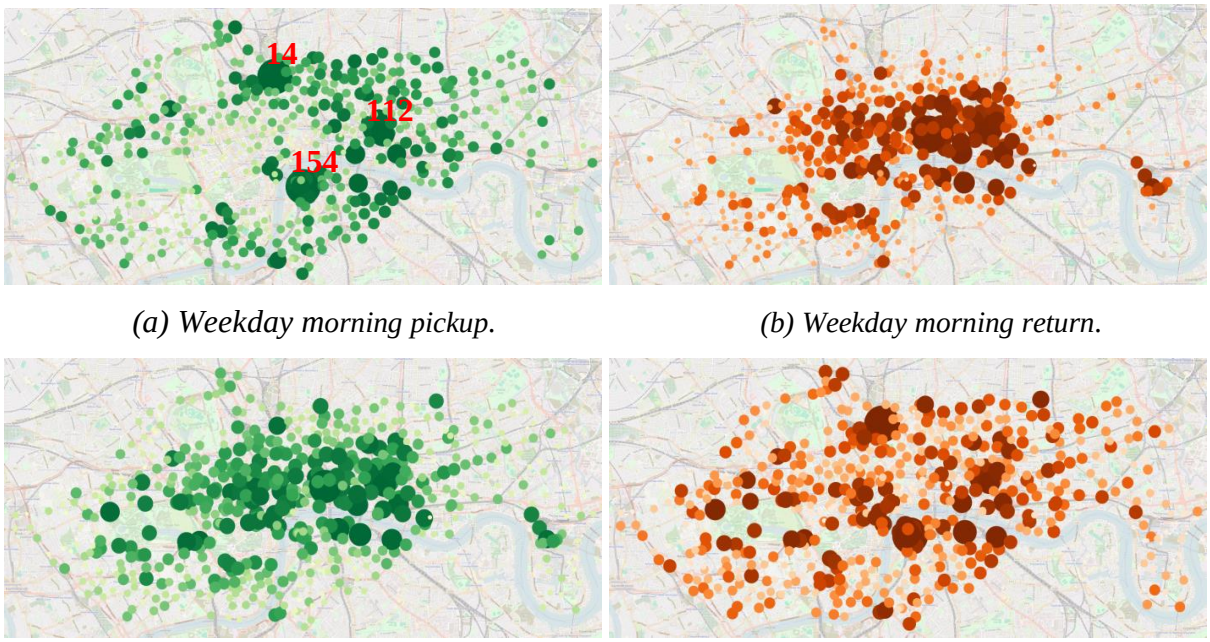
4.3.2. Station Activity

The availability of bikes and vacant docking slots at stations can be seen as a dynamic process determined by *one state* and *three activities*. The state is *available bikes*, and activities are *pickup*, *return* and *redistribution* processes. As the London BSS data only contains the pickup and return information, this section will analyse the weight and balance of those two activities spatially during peak times, when stations have highest demand. Figure 4.9.a and b show the pickup and return average of the weekday morning peak (5 am to 9 am), while Figure 4.9.c and d are for the afternoon peak (3 pm to 7 pm) activities. All of these examples are in August.

On weekday mornings, pickup is higher at the outer areas of central London. At the same time, for return activity, it is centered on or convergent to the inner areas. This outer to inner flow can be stated as an ***inward flow***. This flow is not surprisingly as commuters mostly come from suburbs and may use public transport to stations then take bikes for their last mile to their destination. Conversely, such activities in the afternoon are the opposite of the morning pattern which is from inner to outer. This inner to outer flow can be defined as an ***outward flow***.

In the morning pickup or afternoon return figures, there are three major bike stations that have very high use, denoted by large circles and station IDs. They are **14) Kings Cross Station** which is a major London railway terminus on the north edge of central London, **112) Liverpool Street Station** which is a central London railway terminus and connected London Underground station in the north-eastern corner of the city of London, and **154) Waterloo Station** which is a central London railway terminus and London underground station complex.

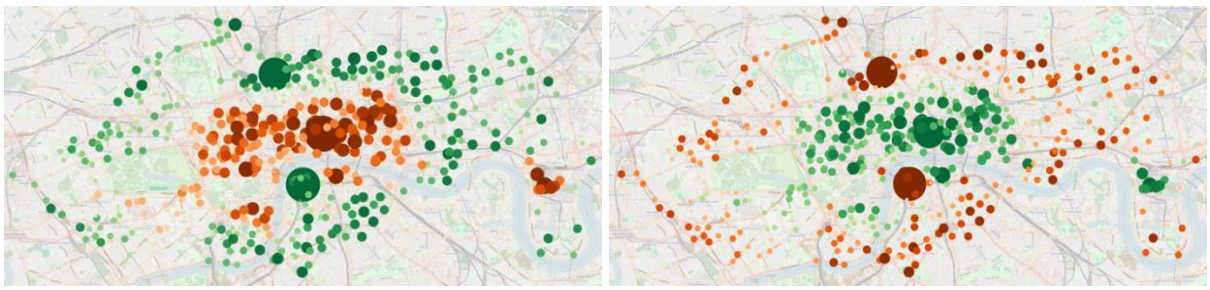
Those bike stations are susceptible to imbalance because of high demand. This state is shown in Figures 4.10 which localizes the pickup vs return balance, weighted by the circle size. Those are calculated by the absolute value of the difference between pickup and return average in each station.



(a) Weekday morning pickup. *(b) Weekday morning return.*

(c) Weekday afternoon pickup. *(d) Weekday afternoon return.*

Figure 4.9. Stations activities in the weekday peak times.



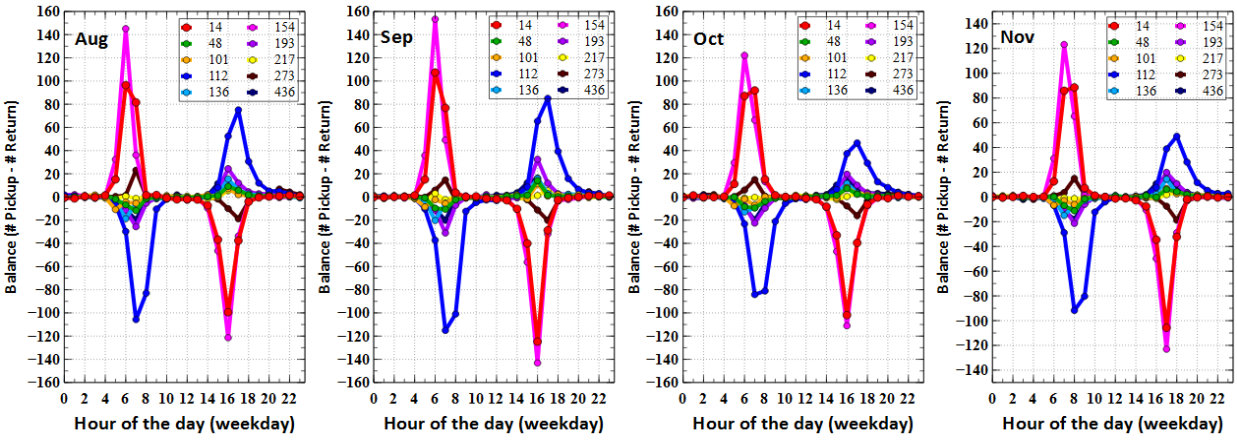
(a) Weekday morning balance. *(b) Weekday afternoon balance.*

Figure 4.10. Stations balance in the weekday peak times.

The green circles in Figure 4.10.a denote that there is more pickup than return activity in the morning peak hours in those stations, mostly in outer stations. This produces a lack of bikes and more empty docking slots. This is defined as ***positive imbalance***. At the same time, the opposite situation where return is more than pickup occur in most inner stations. This is shown by the brown circles which result in a lack of empty slots and more bikes. This is called as

negative imbalance. In some stations the afternoon imbalances are smaller than the morning ones, Figure 4.10.b.

To examine the range of imbalance level on weekdays, the hourly imbalances of ten stations averaged over each of four months are presented in Figures 4.11. The three aformentioned major stations, 14, 112, and 154, have large imbalance levels. In the mornings in August, stations 14 and 154 suffer from **pickup > return** (positive imbalance) at 6 am, reaching an imbalance level +100 for station 14 and +150 for station 154. One hour later, at 7 am, station 112 receives many more returns than pickups, **return > pickup** (negative imbalance), reaching the imbalance level -110. This circumstance is reversed in the afternoon where station 14 and 154 have negative imbalances at 4 pm, while station 112 has a positive imbalance at 5 pm. Generally, their afternoon imbalance levels are less than in the morning. There are also three other stations which have imbalance levels around ±20. They are station 273 which has positive imbalance and station 193 and 136 which have negative imbalances in the morning and vice versa in the afternoon. Other remaining stations have imbalances less than ±5. Due to monthly usage variation, there is increasing imbalance in September and decreasing in October and November.

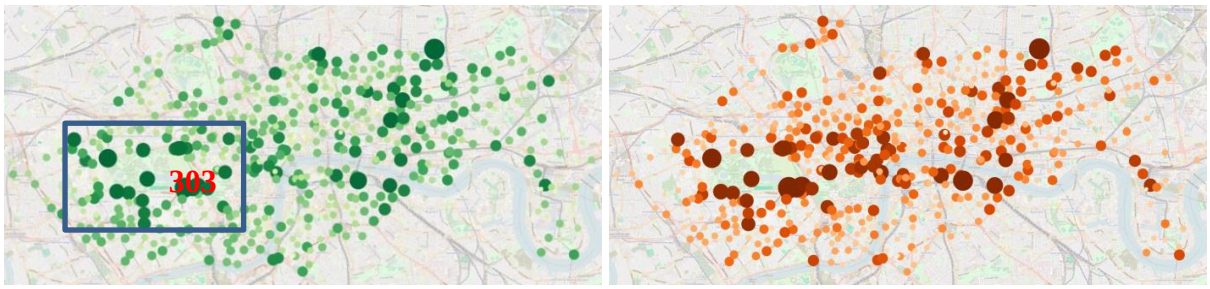


(a) August *(b) September* *(c) October* *(d) November*

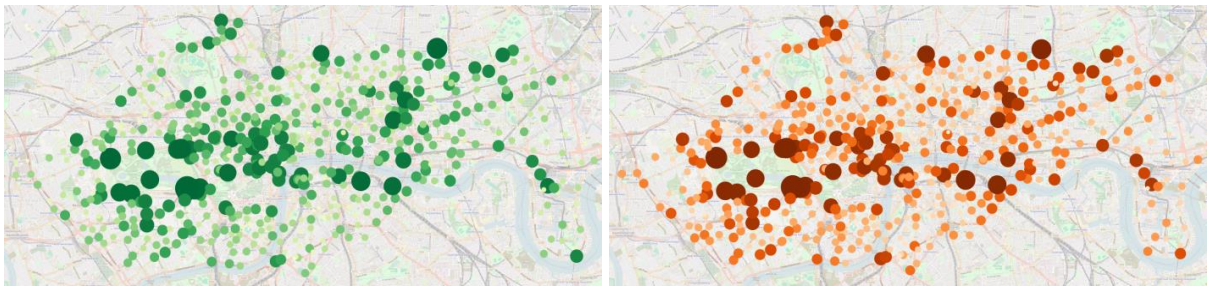
Figure 4.11. Hour of the day balance (#pickup - #return) of 10 stations on weekdays.

The same station activities analysis is conducted for weekend peak hours as shown in Figure 4.12. Here, pickup activity in the morning is more spread than the weekday mornings, and there are not highly dominant stations like on weekdays. One high activity area is shown around **Hyde Park**, shown as a rectangular area. This is a popular recreation area. The activity patterns between pickup and return both in the morning and afternoon are not very different. In the other words, no ***inward*** nor ***outward flow*** exists on weekends. As a result, their balance is

quite uniform as shown in Figure 4.13. Therefore, the weekend flow can be defined as a ***uniform flow***. The redistribution task on weekends is not as essential, because by this *uniform flow* the user ***self-balancing*** occurs.

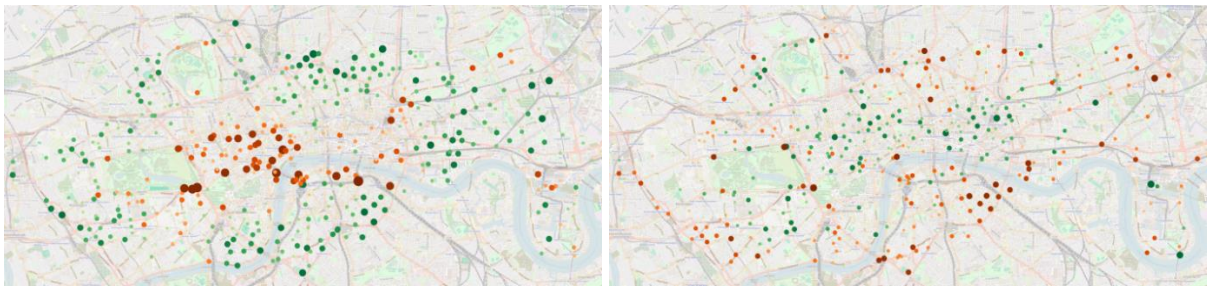


(a) Weekend morning pickup *(b) Weekend morning return*



(c) Weekend afternoon pickup *(d) Weekend afternoon return*

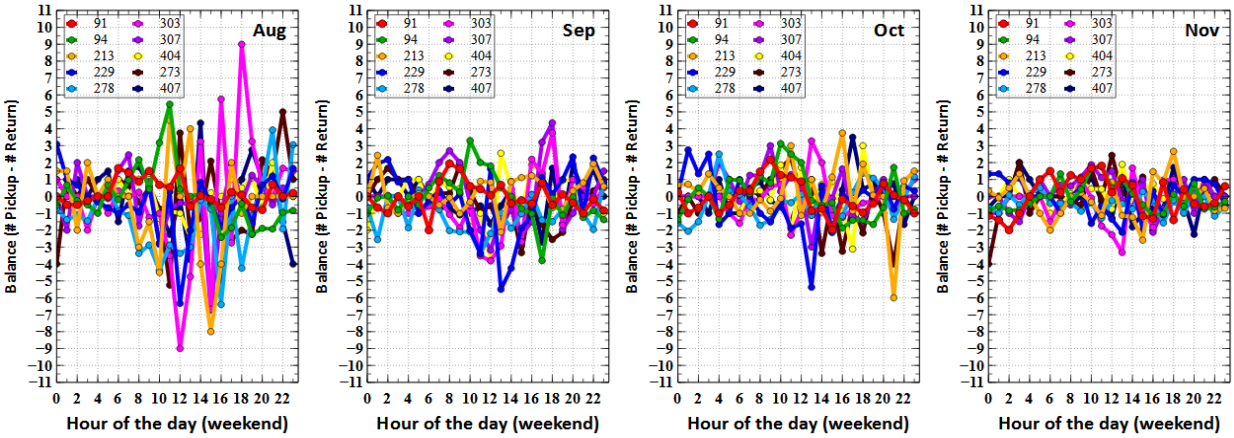
Figure 4.12. Stations activities in the weekend peak times.



(a) Weekend morning balance *(b) Weekend afternoon balance*

Figure 4.13. Stations balance in the weekend peak times.

Unlike the imbalance level on weekdays that have dominant stations showing huge imbalances, the weekend imbalances are relatively small and random as shown in Figures 4.14. Only in August there are stations with imbalance more than ±6, station 303 and 213, with maximum value ±9, while in November they are the least where all are less than ±4. Note that the imbalance term in this section just considers the pickup and return number per hour. The real imbalance should be calculated in consideration of station capacity and the available bikes or docking slots at associated hours.



(a) August *(b) September* *(c) October* *(d) November*

Figure 4.14. Hour of the day balance (#pickup - #return) of 10 stations on weekends.

4.3.4. OD Link Analysis

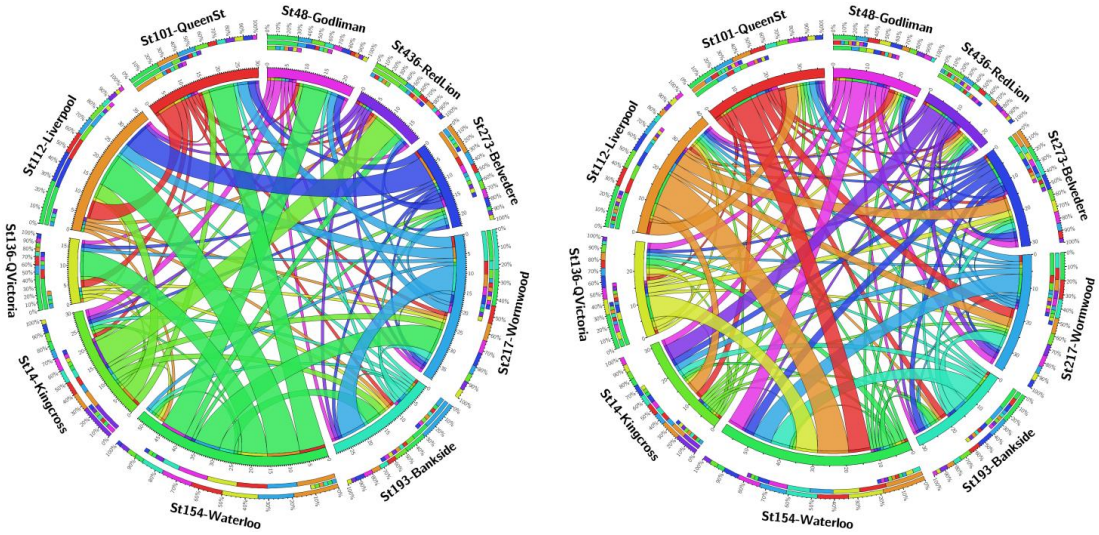
The daily averages of specific OD trips are shown in Figure 4.15. These figures display only links (OD pairs) with more than 2 trips. Here, the darker the line the more trips between that OD pair are represented. The *inward flow* in the weekday morning, Figure 4.15.a, produces dominant links in the centre, mostly between the three aforementioned large stations. The *outward flow* in the afternoon, Figure 4.15.b, is more spread covering a larger area. However, the three large stations links still remain with additional large links around Hyde Park. On weekends, Figure 4.15.c and d, both the morning and afternoon usage is highest around Hyde Park with no significant link between the three peak weekday stations.

(a) Weekday morning *(b) Weekday afternoon*

(c) Weekend morning *(d) Weekend afternoon*

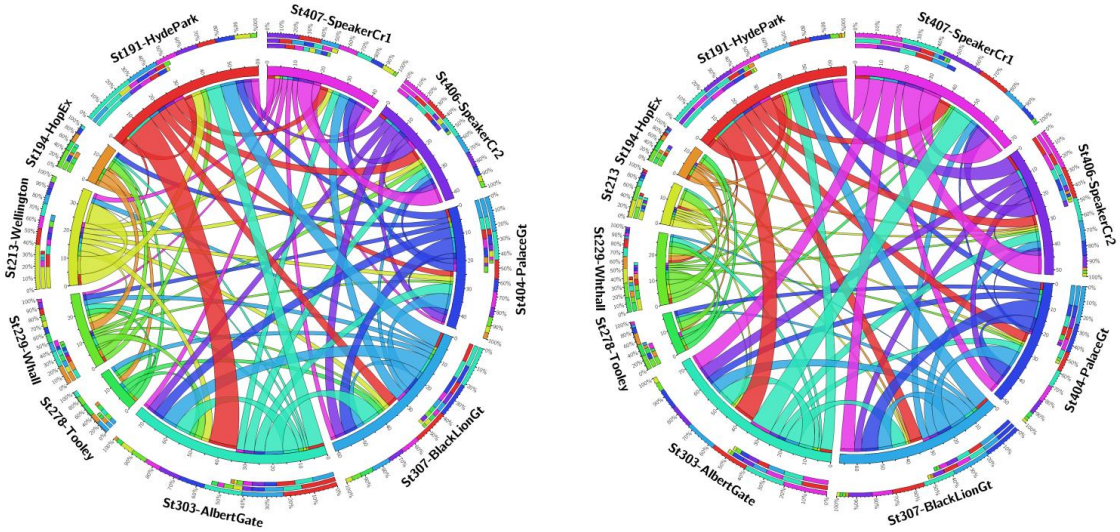
Figure 4.15. Daily average of OD link.

Another approach to observe the OD links is using a circular plot. The circular visualization in Figure 4.16 shows OD links between the 10 busiest stations. Figure 4.16.a shows that major flows in weekday mornings are from station 154 (Waterloo Station) shown by the green link. It has five fat links to station 48, 101, 112, 136 and 217, while station 14 (Kings Cross Station) has only one fat link to station 436. Other fat links are from station 217 to 193, 273 to 112, 101 to 112, and 14 to 112. Here, station 112 (Liverpool St Station) becomes a main destination from other stations. Conversely, in the afternoon it turns into a main origin to many stations and station 154 then becomes a main destination as shown by many colours coming into it.



(a) Weekday morning *(b) Weekday afternoon*

Figure 4.16. OD link of 10 stations during weekday peak times.



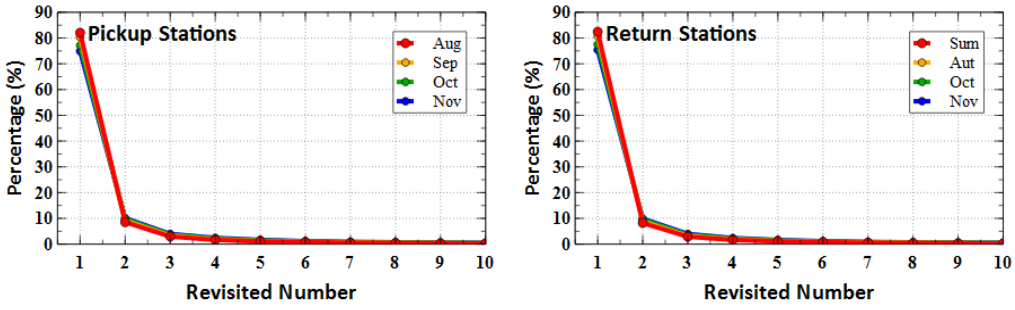
(a) Weekend morning *(b) Weekend afternoon*

Figure 4.17. Link of 10 stations during weekend peak times.

On weekends, as shown in Figure 4.17, most stations serve as half origin and half destination for both mornings and afternoons. For example, consider station 191 (Hyde Park) and station 303 (Albert Gate). Half of 191's links are red (outgoing from 191) and half of 303's links are light blue (outgoing from 303). At the same time, the other half of links to 191 and 303 are multicolour indicating these stations are destinations for these trips.

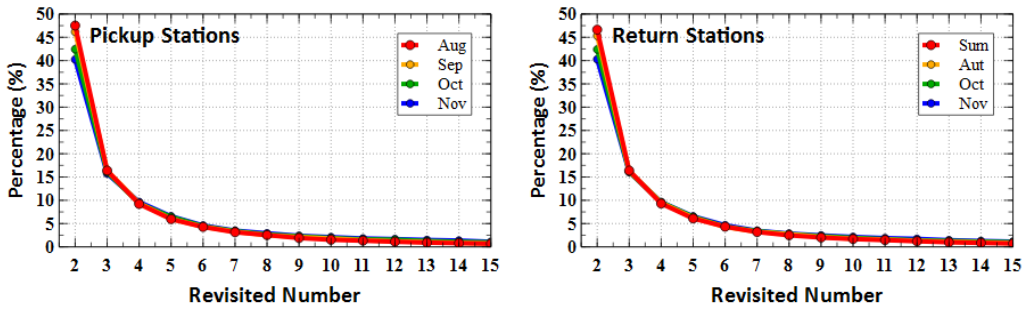
4.3.5. Revisited Stations

Figure 4.18 shows the number of visits observed within the month where a pickup visit is considered as different to a return visit. For each user, numbers of visits for all visited stations are first aggregated per month. This actually shows the relations between each user and each visited station weighted by numbers of visits. Then, for all users and stations, those numbers are averaged. Referring to the *exploration* and *preferential returns* terms by Song et al [24], around 80% are only one visit per station per month, and called as the *exploration* because they are a first time visit, while the rest 20% containing revisited stations are referred to as *preferential returns*. For revisited analysis purposes, Figure 4.19 shows revisited stations where pickup or return occurs at the same station during a month. The trends are almost similar for all months in which two pickups or returns at the same station a month are between 40% and 45%. The others are around 15%, 10%, and 5% of three, four and five times revisited, while six visitations are also close to 5%. The remainders which are more than six revisited are less than 5%. This implies there are frequently visited stations for particular users and indicates a certain degree of regularity. How that regularity corresponds to certain users, and whether that regularity can be measured as well as predicted will be a major topic in Chapter 6.



(a) Pickup station *(b) Return station*

Figure 4.18. Percentage of revisited number of stations.



(a) Pickup station *(b) Return station*

Figure 4.19. Percentage of revisited number of stations (at least 2 visit a month).

4.4. Preliminary Data Analysis Significance Summary

This chapter has investigated the characteristics, dynamicity, and regularity of human mobility in an urban area through intensive spatiotemporal analyses of BSS data. There are several significant observations that have been revealed in a spatiotemporal mobility context.

The number of trips each day is about 38% more than the number of users. This means that there are a significant proportion of users who use BSS more than once a day. There is a decreasing trend of daily usage as winter approaches where Aug > Sep > Oct > Nov. The trip average on weekdays is around 20% more than the average on weekend days. In hourly patterns, there is a ***cyclostationary*** [63] pattern on a daily basis both for weekdays with two sharp peaks and weekends with one moderate peak signaling busy times. The presence of only two peaks also shows a socio-cultural aspect where lunchtime mobility in London is less than in some other cities. Having hourly sharp usage peaks produce asymmetric flows in the system. Peak hours in the weekday morning is from 5 am to 9 am, while in the afternoon is from 3 pm to 7 pm. For the weekend, its midday peak is from 10 am to 6 pm. There is a one hour time shift in November because *British Summer Time* ends.

Waiting times patterns show two types of usage. There are a significant number of short waiting times, less than 1.5 hours, showing a characteristic similar to a negative exponential distribution. On the other hand, there is also a significant set of waiting times between 6.7 and 10.8 hours on weekdays that reflect a commuting pattern. In this range, waiting times show a shape similar to a normal distribution. This daily waiting time can only be captured if users have more than one trip a day. Trip durations are mostly less than 30 minutes, the limit of charge-free usage. It is found that more than 92% of trips are less than 30 minutes. So the free rental period has a significant effect on usage characteristics. The average trip duration on weekdays is less than on the weekends. Trip duration is shorter as winter approaches, Aug >

Sep > Oct > Nov. Trip duration shows highly heavy-tailed characteristics that mostly occur after 10000 seconds (~2.78 hours). This means that there are many short trip durations, and few, but non-negligible, long trip durations. This temporal metric shows the complexity of human mobility.

The distribution of distance is less fat-tailed but more skewed than the duration distribution. The average trip distance on weekdays is more than on weekends and there is a change as winter approaches where Aug > Sep > Oct > Nov. For station activity, there is an ***inward flow*** from the outer to inner stations in weekday morning. While in the weekday afternoon, there is an ***outward flow*** from the inner to outer stations with a wider destination area. In weekend mornings and afternoons, there is a more even flow between inner and outer stations, called a ***uniform flow***. There are three dominant stations on weekdays which are King Cross, Waterloo, and Liverpool St. stations. For the imbalance state, there is the potential for a lack of bikes (***positive imbalance***) in outer stations and lack of empty slots (***negative imbalance***) in inner stations for weekday morning peak times, vice versa in the afternoon. On weekends, the system is largely ***self-balancing*** because of the *uniform flow*. Link weight follows the station activity observations in which fat links are connected to the three busy stations on weekdays, while on the weekends busy stations are around Hyde Park. Referring to the *exploration* and *preferential returns* terms by Song et al [24], there are around 80% of visits (pickups or returns) within a month which are the only visit of that user to that station. In mobility terms**,** this is called ***exploration*** because these are first time visits, while the remaining 20% of visits are revisited stations, and this is referred to as ***preferential returns***.

CHAPTER 5
STATION NEIGHBOURHOOD ANALYSIS

Actual human mobility consists of trips to places of social significance to individuals. For example, a person may travel from home to work in the morning and work to home in the afternoon. The BSS trip is just a portion of the total trip, and the choice of origin and destination BSS stations is not unique, even for repetitions of the same total trip. In other words, even though the overall source and destination (home, work) are the same, there are a number of BSS stations, close to home, that could be used, and a variety of BSS destination stations close to work. A choice of stations is useful if stations are unavailable due to closure for maintenance, or temporarily unavailable because of imbalance. In that case, a nearby station can be used. Users could be advised about alternate nearby stations before they begin their journey. BSS operators could estimate the effect of station unavailability on nearby stations. However, it is unclear within which station use is spatially correlated, and how close is "nearby"? That is the question that this chapter addresses.

As a dynamic network, certain neighbourhood ties or spatial correlation should be exist among BSS stations so that disturbances at one station will affect other stations. The level of the impact will be influenced by the willingness of users to choose alternate nearby stations, and by the regularity of trip destinations. A reasonable preliminary assumption is that if users have to choose other nearby stations rather than their usual station, the station substitution preferences will depend on a certain "nearby-ness" or proximity distance. At an aggregate level, this distance may provide insight about how many nearby stations get affected and to what degree, when a station is disturbed. Currently the spatial ties between BSS neighbourhood stations have not been investigated in the literature. Hence, this chapter will investigate these neighbourhood ties in terms of distance and disturbance level from two perspectives. The first is from the individuals' perspective using mobility motifs analysis, which is the analysis of users' daily trip patterns. Second is from the stations' perspective using the temporary station shutdown analysis. Here, the *station-usage-based* method is proposed to compute the usage change *before*, *during*, and *after* shutdown. The distance from mobility motifs analysis will be compared with the impact distance of shutdown stations. Possible impacts on BSS operation will also be investigated in this chapter.

Survey data can provide social contexts to individual daily trips, for example, ***home → work → shop-near-home***. On the other hand, the trace of individuals in BSS data can only be

represented by the sequence of their daily visited stations. For the given example above, the station trace for a user could be ***44 → 50 → 45***, or ***21 → 75 → 23*** for another user, where the numbers represent the stations ID. For a particular user, on a particular day, the visited stations can be labelled with letters, where A replaces the first visited station, B the next, etc. If a station is revisited in the daily mobility trace, then the previous letter is re-used. If the station IDs in the above examples are labelled, then both of the traces become ***A → B → C***. In this case, the distance between ***A*** and ***C*** could reflect whether A and C represent the same overall destination (e.g. ***home***) or different overall destinations. In other words, if the station choices of consecutive trips are different from the previous ones, then the distances between stations may give knowledge about station neighbourhood ties, and whether the choice of stations is confined to a certain neighbourhood distance. Since the daily *spatial-mobility-motifs* of BSS have not been previously analysed, this work can also contribute to the literature on motif models to complement the existing mobility motif results for cell phones and mobility survey data [40, 41]. Previously, in motif analysis, each destination became the origin of the next trip, and so a connected directed graph uniquely patterns. With BSS, each trip has a unique origin and destination. So, this study also proposes a new technique of consecutive labels (A, B, C) on motif nodes to make BSS mobility motifs clearer to understand.

In BSS operation, the temporary shutdown of a station is sometimes required due to reasons such as maintenance, redesign, or special events. This shutdown obviously will change the topology of the existing network and may impact on nearby stations. This may affect the quality of service, especially for individuals who make their trips regularly via a particular station. Individuals' responses could be different. They may try to find alternative nearby stations or they may use other modes of transportation. This might lead to a loss of users, especially if the shutdown station has a significant role in the network, and no nearby stations are an immediate substitute. In this case, the significance of a station can be expressed in terms of location, usage (pickup and return), and number of links (trips) with other stations in the network. This leads to some further questions: how the shutdown impact is for nearby stations, how to properly measure the impact, to what extent other stations can be an automatic substitute for the shutdown station, and how this shutdown knowledge can best be used for the BSS operation, design, and deployment.

When a station shutdown is analysed, which nearby stations will be included in that analysis is an essential first step. Considering all the stations in the network seems too large as an impact scope since the shutdown will most likely only strongly affect nearby stations. To

decide on the relevant impact scope, the proposed *station-usage-based* method uses an approximate radius derived from the trip distance and walking distance suggested by [75] as a preliminary radius of observation. Then, the usage transformations *before*, *during*, and *after* shutdown are conducted for all stations in this set to see the impact distance. Once the impact distance is found, it is also used to identify what we define as the *ineffective* as well as the *isolated* stations including the related recommendations for BSS operator actions.

In most BSS datasets, the only positions that can be provided are the origin and destination station geo-locations, so that the two-dimensional Euclidean distance between these is widely used for the spatial analysis. However, as a simple straight line, Euclidean distance cannot capture variations in the actual travel distances which are affected by road layout. In this chapter, the usefulness of waypoint distance (i.e., distance between points along a feasible path via roads) is compared with Euclidean distance and with Manhattan distance. Finally, the work in this chapter is used to answer RQ1 and a part of RQ4.

5.1. Methodology

This section begins with the waypoint distance description and its difference from Euclidean and Manhattan distances. This is followed by the method for transforming trip data to the *daily motifs*. Then, the selection of nearby stations is presented. Finally, the concept of *station-usage-based* analysis by means of usage transformation is proposed. Here, a shutdown station is identified from the usage dataset because it does not have any usage (pickup or return) for a period of several days where a number of shutdown cases will be investigated in section 5.3.

5.1.1. Waypoint Distance

BSS usefully capture the individual mobility with clear geo-location of origin-destination, albeit without the real route of each trip. Determining the real route of BSS users, rather than using the straight-line two-dimensional *Euclidean distance* [11, 75], is not possible unless each bike is equipped with a GPS tracker which it is not the case here. However, a better estimate of the distance travelled should be possible. Users will most likely follow the road network and many will choose the shortest route. For this reason, this study will infer a trip distance by selecting the most likely road segments with the shortest distance between OD from a series of ***route points*** given by Google Maps API and MapQuest API. This is called the ***waypoint distance***, which is described in Figure 5.1. A waypoint refers to an intermediate point on a path

at which the direction of travel is changed. Then, a route is defined as a sequence of straight-line segments from origin, via the waypoint, to the destination. To the best of our knowledge, this is the first BSS study which adopts this waypoint distance approach.

From Figure 5.1, the waypoint distance can be formulated from a series of waypoints $(P_1, P_2, \dots, P_m)$ between OD as the sum of Euclidean distances between consecutive points as follows:

$$WaypointDist_{OD} = \sum_{i=1}^{M+1} e_i \tag{5.1}$$

Here, M is number of waypoints between OD and e_i is the Euclidean distance between each pair of points starting from O and ending at D.

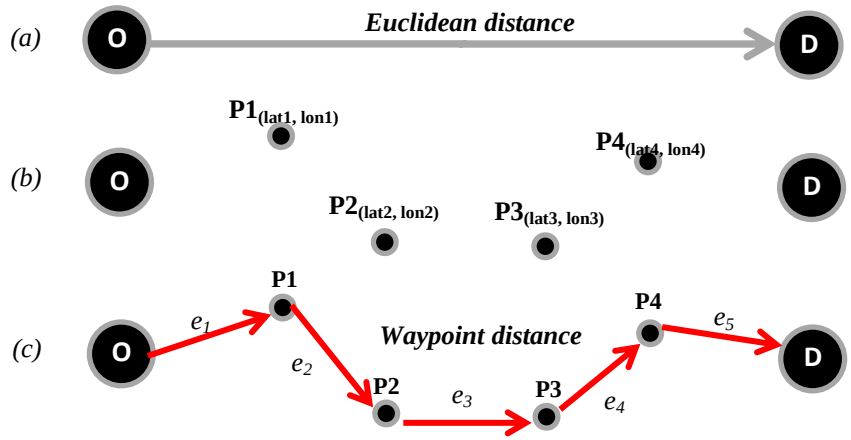


Figure 5.1. (a) Euclidean distance between OD, (b) Four waypoints (P1, P2, P3 and P4) between OD, (c) Waypoint distance (e1 + e2 + e3+ e4 + e5) between OD.

This waypoint distance is different from the *Manhattan distance* which calculates distance as the x-distance plus y-distance based on strictly vertical and horizontal paths which parallel along the axes (x,y) [121, 122]. For a set of axes (x,y), the Manhattan distance between OD could be:

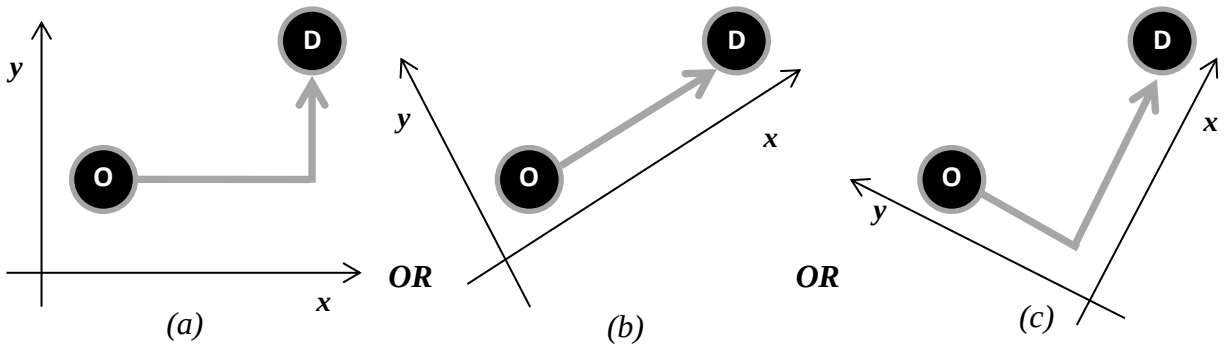


Figure 5.2. The illustration of axes dependence of Manhattan distance.

Experience from other researchers [122] finds that the Euclidean distance typically underestimates the road distance, and the Manhattan distance overestimates the distance. Both distance measures significantly underestimate the road distance if there is a significant obstacle between origin and destination, such as a railway line or river. The waypoint distance is used here to provide a more accurate measurement of which BSS stations are nearby each other.

It is not expected that all BSS users will follow the shortest waypoint route to reach a certain destination from a certain origin because they have the freedom to choose their own routes. They may be sightseeing, for example. However, the waypoint approach provides a practical estimate of the shortest distance compared to Euclidean and Manhattan distances. The waypoint distance is useful when comparing effects on stations of a nearby shutdown station, since it is the shortest travel distance that is important. Waypoint distance can give a different ordered set of nearby stations to a shutdown station. For example, if the order of nearby stations from station **A** using Euclidean distance is **B-C-D-E**, it could be **C-B-E-D** using waypoint distance. The real example of this case from London BSS is given in subsection 5.3.1.

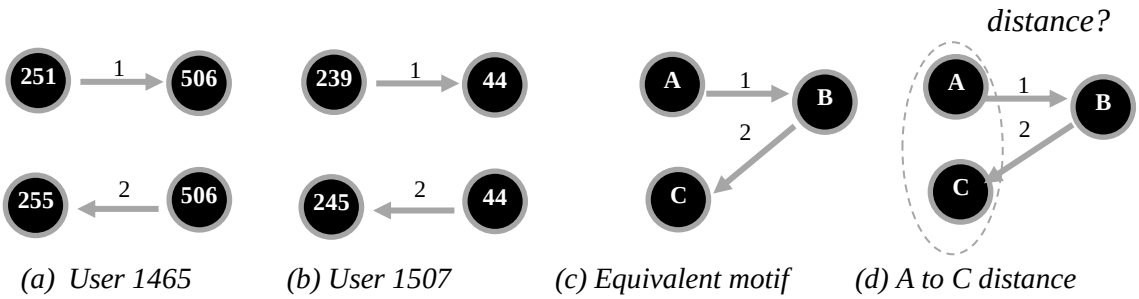
5.1.2. Spatial Mobility Motifs

Spatial mobility motifs represent OD trajectories or trace patterns of users over one day in a graphical form. More formally, a motif is represented as a directed graph and defined as $G = (V, E)$, which consists of a set of V nodes or vertices representing BSS stations and a set E of directed edges which represent trips between stations by one user during one day. Two mobility motif graphs are said to be equivalent if there is a one-to-one mapping between the nodes and edges in the two graphs. Equivalent graphs are said to represent the same mobility motif.

Therefore, even though different users visit different stations, common spatial patterns could be inferred if those OD stations are labelled consecutively by the stations visited over a day. Here, the first daily pickup station for each user will be labeled with *A*. The subsequently visited stations in that day will be labeled either with a new label (*B, C, D, ...*) if that station has not yet been visited or with the previously used label corresponding to a station that has been visited. Figure 5.3 illustrates how a similar daily motif is drawn from two users with different OD trips as listed in Table 5.1.

Table 5.1. Daily trips example of two users.

User ID	Pickup Data		Return Data	
	Station	***Pickup time***	***Station***	***Return time***
1465	251	2012-08-01 06:34	506	2012-08-01 06:40
1465	506	2012-08-01 16:45	255	2012-08-01 16:51
1507	239	2012-08-01 07:05	44	2012-08-01 07:15
1507	44	2012-08-01 17:00	345	2012-08-01 17:20



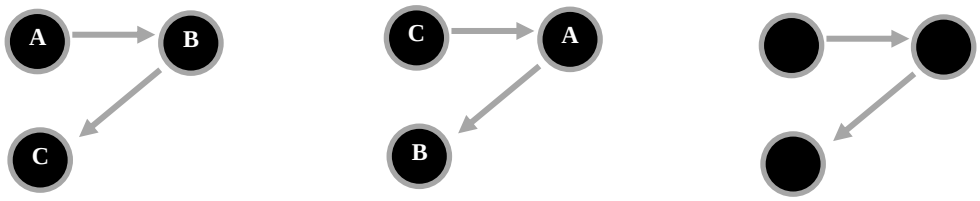
(a) User 1465 *(b) User 1507* *(c) Equivalent motif* *(d) A to C distance*

Figure 5.3. From stations traces to equivalent motif ***A→B B→C.***

It can be seen from Table 5.1 that the pickup station for the second trip of each user is the return station of their previous trip. Therefore, if visited stations are labelled in alphabetical order *(A,B,C,D,E,F,…)*, their motifs become **A→B B→C**, Figure 5.3.c. In this case, directed edges stand for a trip from pickup to return station, and the numbers on edges are the trip sequence numbers.

The motifs in Figure 5.3 above may represent a simpler total trip motif, such as ***home → work → home,*** with the user choosing different stations to leave and return home. Looking at the distances between A and C in motifs like Figure 5.3.d across all of the BSS trips may give some understanding of what distances typically corresponding to nearby stations.

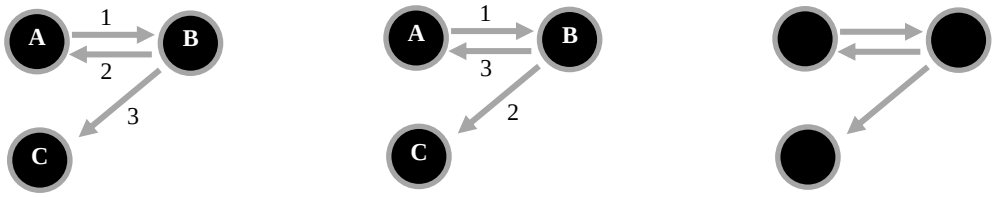
The labels assist in distinguishing different motifs which cannot be distinguished just from unlabeled edges and nodes as used in previous motifs analysis in [41] and [40]. For example, two labelled graphs A→B B→C and A→B C→A, Figure 5.4.a&b, represent different motifs with labelled graphs, but would be indistinguishable with unlabelled graphs, as in Figure 5.4.c.



(a) (b) (c)

Figure 5.4. The labelled and unlabelled motifs.

Not only node labels, but edge labels also have an important role to distinguish different motifs such as **A→B B→A B→C** and **A→B B→C B→A** as shown in Figure 5.5.a&b. Even with node labels, these motifs could not be distinguished.



(a) *(b)* *(c)*

Figure 5.5. The labelled (with edge numbers) and unlabelled motifs

5.1.3. Impact Distance

A preliminary assumption for the maximum distance of the significant effects of a station shutdown is needed to determine the set of nearby stations that will be analysed. In this case, the median trip distance on section 4.3.1 will be employ with an assumption that most users travel are confined by that distance. Another measure of impact distance from a station is taken from another human mobility study which is a typical walking distance suggested by O'brien et al. [75].

During station shutdown, it is proposed that the set of stations which are affected by a shutdown, called the nearby stations, are those that are within a specific distance, called the impact radius, of the shutdown station.

5.1.4. Station Usage Changes

There are two steps to understanding the impact of a station shutdown on the BSS operations. The first step is to estimate the maximum reasonable impact radius, and therefore the set of possible impacted stations. The second step is to understand the usage changes, or transformations, in those nearby stations to determine which stations are most affected, and therefore what the actual impact radius is. Two measures of change of usage are proposed: ***before-to-during (BtoD)*** and ***during-to-after (DtoA)*** shutdown. For a shutdown length of D days, usage is analysed over 5 periods of D days as shown in Figure 5.6. This will give at least two comparisons to see the uniformity of the changes. When the system is in normal operation without a shutdown, both backward and forward windows comparisons should show negligible usage changes. Meanwhile, the similar length of periods will allow direct usage comparisons

over the period of evaluation. For example, if the length of a shutdown is 7 days, then the average of usage in 1 to 7 days before *(B_1toD)*, as well as 8 to 14 days before *(B_2toD)*, are compared with the average of usage during shutdown.

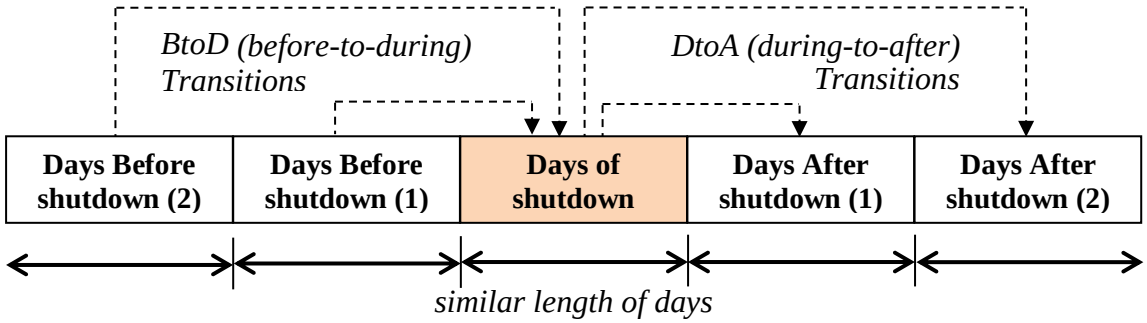


Figure 5.6. The usage changes before-to-during and during-to-after.

The percentage of changes for each nearby station can be calculated as:

$$B_n toD\ (\%) = 100 * \left(\frac{Usage\ D - Usage\ B}{Usage\ B}\right), \qquad DtoA_n\ (\%) = 100 * \left(\frac{Usage\ A - Usage\ D}{Usage\ A}\right) \quad (5.5)$$

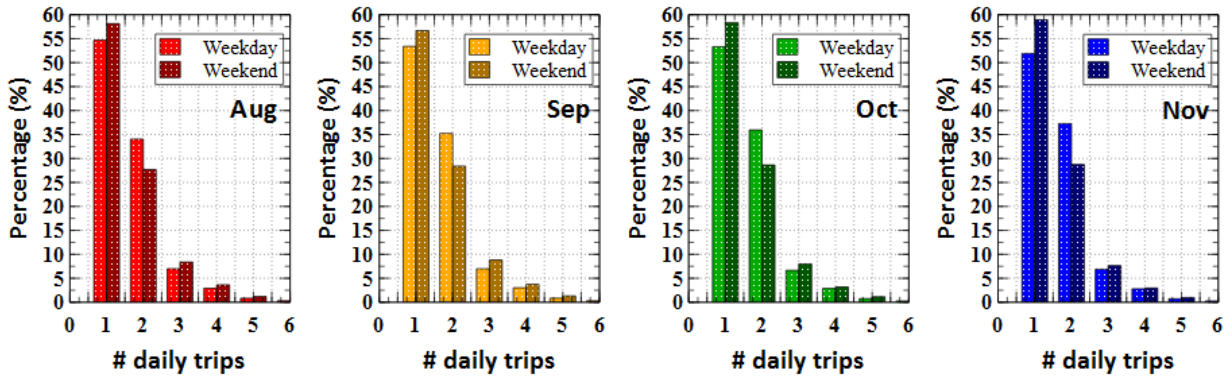
It is expected that $B_n toD$ will be positive with a decreasing impact as distance increases because the nearby stations will probably receive more use during the shutdown. Conversely, $DtoA_n$ will be negative because the nearby stations will return back to their normal state when the shutdown is finished. If a stations has little or no changes, then the shutdown has no impact on that station.

5.2. Spatial Mobility Motifs Analysis

Using the methodology in section 5.1.2, this section investigates the characteristics of BSS mobility motifs to understand the daily movement patterns of BSS. By looking at common motifs, common BSS usage patterns can be identified. This may aid BSS system operations, but also will be useful in understanding human mobility more generally. As reviewed in Chapter 2, previous human mobility motif studies were conducted by Schineider et al. [40] and Jiang et al. [41] using surveys and mobile phone datasets. Adopting the concept of motifs from network theory, they consider a daily network pattern as a motif if that network is found in more than 0.5% of the dataset [40]. Using this threshold, they found 17 and 11 unique daily mobility networks respectively in analogy to motifs in complex networks, where this threshold is also used in this section to find BSS motifs.

5.2.1. Daily Trips Count

The extraction of BSS mobility daily motifs in this section starts from the calculation of how many trips each individual makes per day. Figure 5.7 summarizes the percentage of users who make a certain number of trips on a daily basis averaged by weekday and weekend, and observed per month. In all months, the majority of users make only one trip a day reaching around 55% of users on weekday and 60% on weekend. Then, users with two trips a day are around 35% on weekday and 30% on weekend, and users with three trips a day are around 5% on weekday and 8% on weekend. The remainders are more than three trips. This study will only consider up to three trips a day which covers more than 90% of users. As the number of trips increases, the number of possible motifs increases exponentially, so numbers in particular motifs beyond three trips are negligible. The details of all the different motif types are listed in section 5.2.2.



(a) August *(b) September* *(c) October* *(d) November*

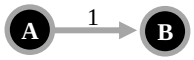
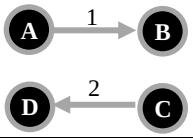
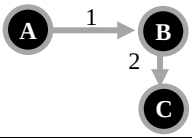
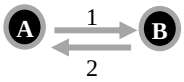
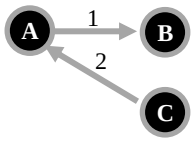
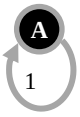
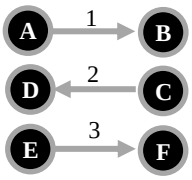
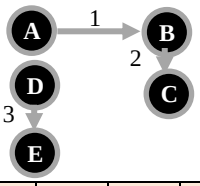
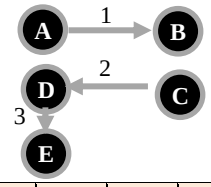
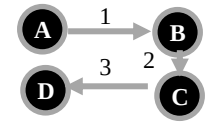
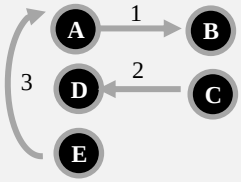
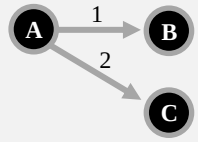
Figure 5.7. Percentage of number of daily trips per user.

5.2.2. Daily Motifs Type

Observing from one to three trips a day, there are 216 network patterns found as candidate motifs. They are 2 networks for one trip a day, 15 networks for two trips a day and 199 networks for three trips a day. However, only a few of them are popular networks. Table 5.2 shows the 12 top networks based on their appearance on weekdays and weekends for the four months period.

Considering 0.5% as a minimum threshold [40], only 10 daily network patterns can be considered as common motifs. Two motifs are from one trip a day: **A→B** and **A→A**, four motifs are from two trips a day: **A→B C→D, A→B B→C, A→B B→A** and **A→B C→A**, and four other motifs from three trips a day: **A→B C→D E→F**, **A→B B→C D→E**, **A→B C→D D→E**, and **A→B B→C C→D**.

Table 5.2. Twelve top networks (10 as motifs of more than 0.5%)

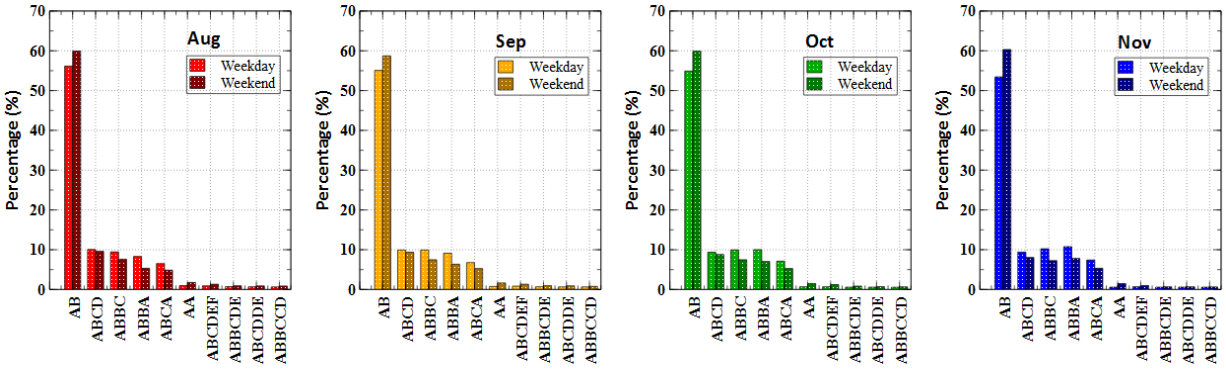
No	1				2				3				4			
Motif	A→B 				A→B C→D 				A→B B→C 				A→B B→A 			
%	Aug	Sep	Oct	Nov	Aug	Sep	Oct	Nov	Aug	Sep	Oct	Nov	Aug	Sep	Oct	Nov
WD	56.15	55.08	54.90	53.44	10.06	9.93	9.37	9.39	9.44	9.93	9.96	10.24	8.31	9.13	10.01	10.77
WE	59.93	58.72	59.91	60.32	9.60	9.37	8.81	8.05	7.62	7.48	7.48	7.26	5.33	6.35	7.04	7.81
No	**5**				**6**				**7**				**8**			
Motif	A→B C→A 				A→A 				A→B C→D E→F 				A→B B→C D→E 			
%	Aug	Sep	Oct	Nov	Aug	Sep	Oct	Nov	Aug	Sep	Oct	Nov	Aug	Sep	Oct	Nov
WD	6.52	6.76	7.12	7.41	0.98	0.76	0.68	0.59	0.92	0.84	0.68	0.69	0.72	0.68	0.59	0.59
WE	4.87	5.26	5.27	5.34	1.77	1.64	1.50	1.50	1.32	1.32	1.23	0.99	0.93	0.95	0.85	0.67
No	**9**				**10**				**11**				**12**			
Motif	A→B C→D D→E 				A→B B→C C→D 				A→B C→D E→A 				A→B A→C 			
%	Aug	Sep	Oct	Nov	Aug	Sep	Oct	Nov	Aug	Sep	Oct	Nov	Aug	Sep	Oct	Nov
WD	0.68	0.65	0.58	0.60	0.65	0.65	0.57	0.59	0.37	0.38	0.38	0.39	0.35	0.35	0.33	0.37
WE	0.90	0.91	0.71	0.67	0.87	0.78	0.66	0.64	0.43	0.50	0.48	0.43	0.48	0.47	0.45	0.45

In all months, the motifs can be categorized by three groups based on their percentage range as shown in Figures 5.8. The first is the most dominant one, **A→B,** which is 54% of all weekday trips and 59% on weekends.

The second group consists of four motifs which span from 5% to 10%. All of them are from 2 trips a day (motif no 2, 3, 4, and 5). They are **A→B C→D** where there is no similar or recurrent visited stations, **A→B B→C** where the second pickup is same as the previous return, **A→B B→A** where the second trip is exactly the reverse of the previous trip, and **A→B C→A** where the last return comes back to the first pickup.

In the third group, there are five motifs with a small percentage range between 0.5% and 1.5%. There is one *roundtrip* or *self-loop* from one trip a day which is motif **A→A**, while others come from the three trips a day. Motif **A→B C→D E→F** is a motif with no recurrent stations which means users who have this motif must visit six different stations a day. Then,

motif **A→B B→C D→E** with one revisited station where the second pickup is the previous return, followed by motif **A→B C→D D→E** also with one revisited station. The last motif is **A→B B→C C→D** with two revisited stations.



(a) August *(b) September* *(c) October* *(d) November*

Figure 5.8. Percentage of daily motifs.

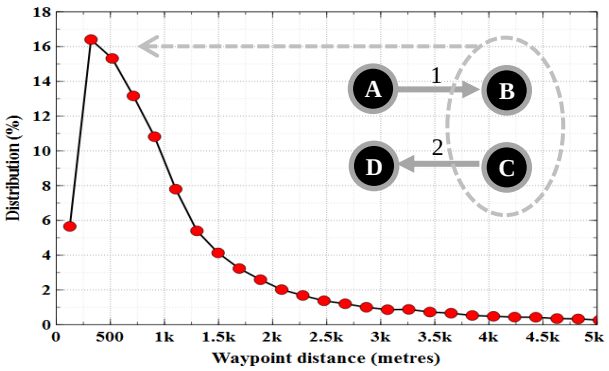
The following insights can be noted from this data. Only 10 of 216 motifs are common (>0.5%). This is fewer than common motifs from cell phone and mobility survey data. There are only 3 motifs without daily revisited stations, motif 1, 2 and 7, but their total percentage is high, around 66.4%. The number of visited stations in the 10 motifs varies from 1 to 6 different stations. In the next section, analysis of these motifs will be used to estimate the typical distance between nearby alternate stations for trips.

5.2.3. Distance Analysis of Daily Mobility Motifs

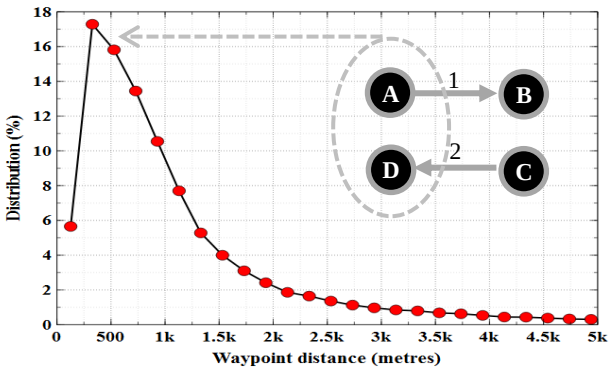
It is our conjecture that a significant proportion of users who make BSS trips between social destinations such as ***home***, or ***work***, will sometimes use different origin or destination BSS stations. So in some significant proportion of trips with, say, three different stations, such as **(A→B B→C)**, it will be the case that **A** and **C** are different BSS stations used for the same social location. Looking at the distribution of distances between **A** and **C** will give insight into the typical distances between nearby stations.

In particular, these distance observations are made for motif no 2 **(A→B C→D)**, no 3 **(A→B B→C)** and no 5 **(A→B C →A)** as shown in Figure 5.8. Results show that for motifs 2 and 5, Figure 5.9.a and d, they have tendency to pick up bikes for the second trip close to the previous station where they returned the bike for the first trip, with most common inter-station distances of 300 m to 500 m. An inter-station distance of 100 m is much less common, perhaps due to the fact that not many pairs of stations are this close to each other.

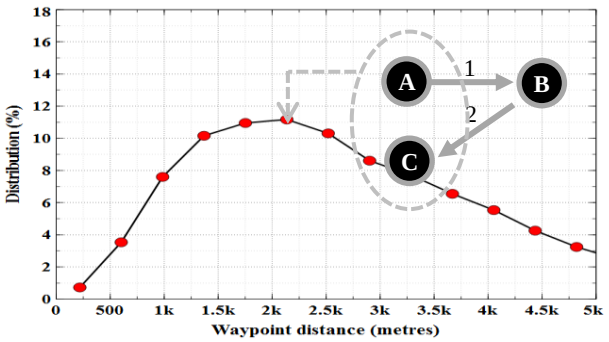
Similar characteristics occur for returning bikes on the second trip, Figure 5.9.b. This means users tend to pick up and return the bikes in a similar area. However, a different result occurs for motif no 3, Figure 5.9.c, where their distance between first pickup and the second return stations are about 2 kilometres which means these are usually two quite separate trips. By looking at the peaks in Figures 5.9.a, b, and d, these results show that if users choose a different station in a previously visited area, the inter-station distance in this neighbourhood is most commonly around 300 m.



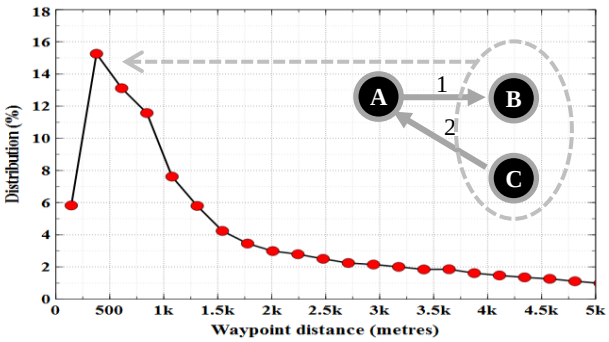
(a) WP distance distribution of B to C motif no 2



(b) WP distance distribution of D to A motif no 2



(c) WP distance distribution of C to A motif no 3



(d) WP distance distribution of B to C motif no 5

Figure 5.9. Distance distribution of nearby OD stations based on daily motifs.

5.3. Shutdown Stations Analysis

Usage changes in nearby stations when a station is shut down. An example of this usage transformation is shown in Figure 5.10 with one shutdown station and corresponding daily usage patterns of nearby stations. In this example, the shutdown station is station 360 (11 days of shutdown, days 40 to 50) which is denoted with the red circle. It can be seen that there are usages transformation in nearby stations which vary as a function of distance. A significantly increased usage occurs in station 177 and 316, while a slight increase also happens in station 359 and 320, as shown in Figure 5.11. By looking at a number of shutdown cases, the impact radius is analysed using the proposed methods in the subsection 5.1.4.

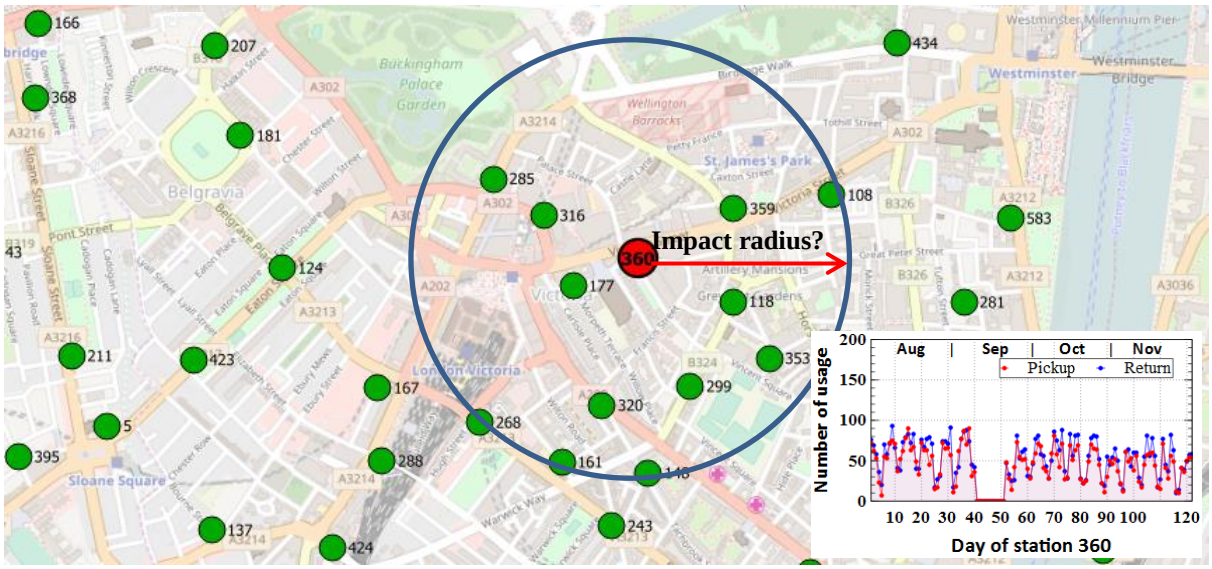


Figure 5.10. The usage pattern of shutdown station and its nearby stations geolocation.

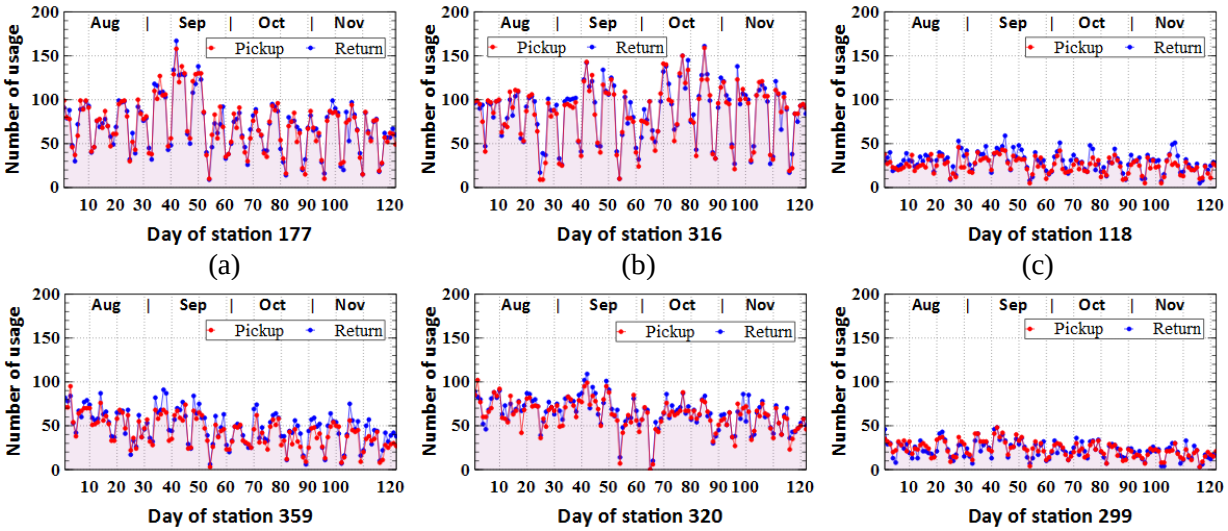
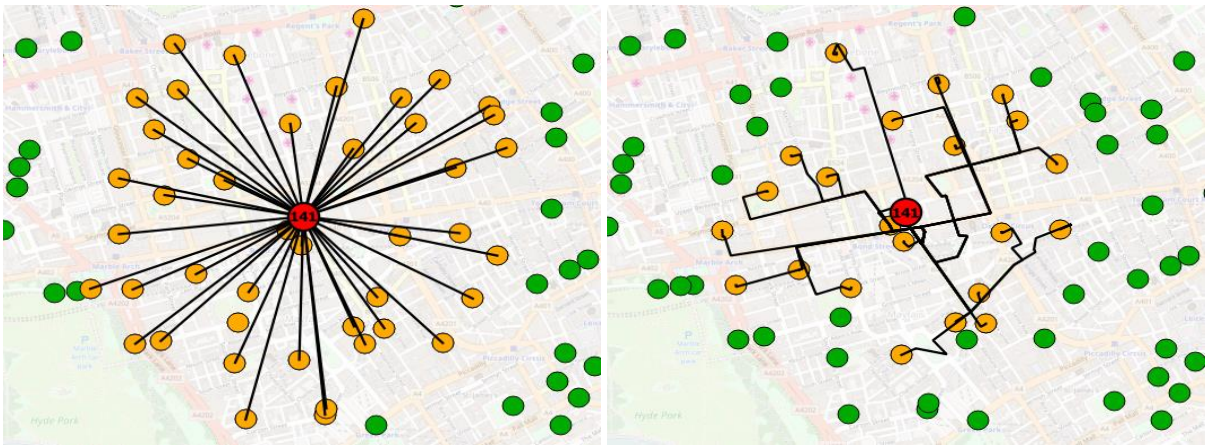


Figure 5.11. Daily usage patterns of nearby stations from the shutdown station.

5.3.1. Nearby Stations Set

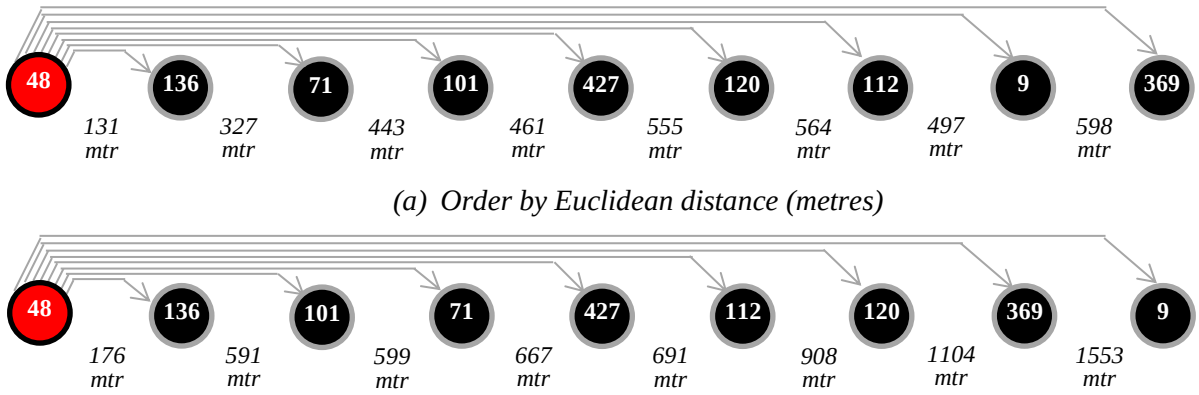
Using the trip distance curves in Figure 4.8 of previous chapter which have a peak distance (median distance) of around 1 km, and together with the walking distance suggested by O'brien et al. [75] which is also 1 km, this 1 km is used as an initial radius from the shutdown station. Then, eight shutdown cases are observed measuring distance both with Euclidean and waypoint distance from the shutdown station. The set of nearby stations example based on formula 5.1 can be seen in Figure 5.12 for station 141 (red circle with station ID). It can be seen using the same 1 km distance, the stations reached by a waypoint distance of 1 km contain fewer stations (22 stations) than by Euclidean distance (44 stations).



(a) Euclidean distance *(b) Waypoint distance*

Figure 5.12. Stations set considering 1 km distance to the shutdown station.

The list of nearby stations ordered by inter-station distance, gives also a different list ordering when using Euclidean and waypoint distance. For example, the nearby stations order from station 48 based on Euclidean distance, Figure 5.13.a is different with waypoint distance, Figure 5.13.b. This different order may give different results when observing the usages transformation.



(a) Order by Euclidean distance (metres)

(b) Order by waypoint distance (metres)

Figure 5.13. Nearby stations order from the central based on Euclidean and waypoint distance

5.3.2. Daily Usages Transformation

Implementing the *station-usage* analysis using equation 5.5 for all the shutdown cases, the results of daily pickup transitions of nearby stations *before-to-during* and *during-to-after* shutdown can be seen in Figure 5.14 and Figure 5.15 for Euclidean and waypoint distance order respectively. Observing for pickup on weekdays, figures on the left side (*a, c, e, g, i, k,* and *m*) present the transitions from *before-to-during* ($B_n toD$) shutdown, while on the right side (*b, d, f, h, j, l,* and *n*) present the transitions from *during-to-after* ($DtoA_n$) shutdown. Red circles

represent the transitions relative to one window before (B_1toD) and after shutdown ($DtoA_1$), and blue circles represent the transitions relative to two windows before (B_2toD) and after shutdown ($DtoA_2$).

The figures can be interpreted as: when a station is shut down for a certain period of days and if B_ntoD figures (left side) give significant positive values, it means the stations are impacted with increased usage, because a number of users choose nearby stations as substitutes for the shutdown station. Within the impact radius stations are significantly affected. For more distant stations, the effect should be less. Similar behaviour occurs when the shutdown station is re-activated. If $DtoA_n$ figures (right side) give significant negative values, it indicates the nearby station is impacted, because users who previously choose the nearby stations come back to use the re-activated shutdown station.

The approximate impact radius can be observed to be a few hundred metres from the shutdown stations. Generally, using Euclidean distance order, Figure 5.14 shows that the affected stations seem in the radius of 200 m. While using waypoint distance order, Figures 5.15 give an impact radius of around 300 m.

However, looking in Table 5.3 shows that the nearest stations from the shutdown, within the 200 m Euclidean distance are not always impacted. This can be seen for station 514 (Euclidean: 194 m) and 112 (Euclidean: 205 m) where no transformations occur. Their waypoint distances are 325 m and 483 m respectively. By contrast, all stations which are less than 300 m of waypoint distance are impacted. For B_ntoD, the affected stations get increased usages from 20% to 80% (e,g,i,k,m). Similarly for $DtoA_n$, they get similar decreased usages. This fact gives an insight that a waypoint distance of 300 m is a good estimator of the limit of the distance users will walk to an alternate nearby station.

It can also be seen from Table 5.3 that not all stations impact their neighbours during shutdown. For example, station 112 with very high daily usage (282.5 and 264.4 for one and two windows before shutdown) does not increase the usage of station 393, the nearest station, even though their Euclidean distance is 205 m. Their waypoint distance is 485 m which is further than users normally will walk to an alternate station. On the other hand, station 197 which has daily usage around 75 to 85, much smaller than station 112, has a significant impact on its two nearby stations because their waypoint distances are less than 300 m. This means that the waypoint distance is a more reliable estimate of impact radius than the Euclidean distance and also better at predicting the relative impact to nearby stations.

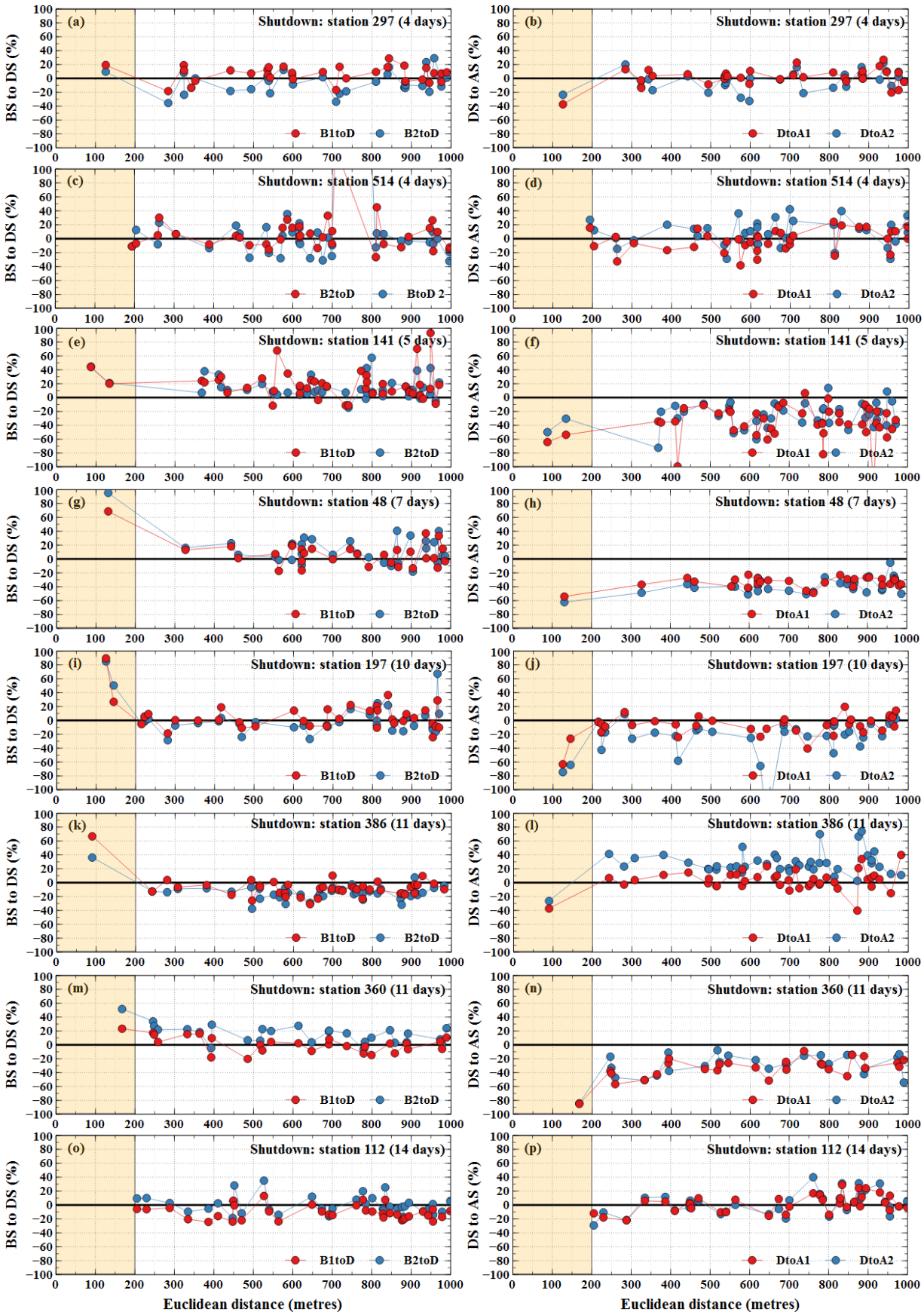


Figure 5.14. Average daily pickup transition (%) of nearby stations ***before-to-during*** *and* ***during-to-after*** *shutdown ordered by* ***Euclidean distance.***

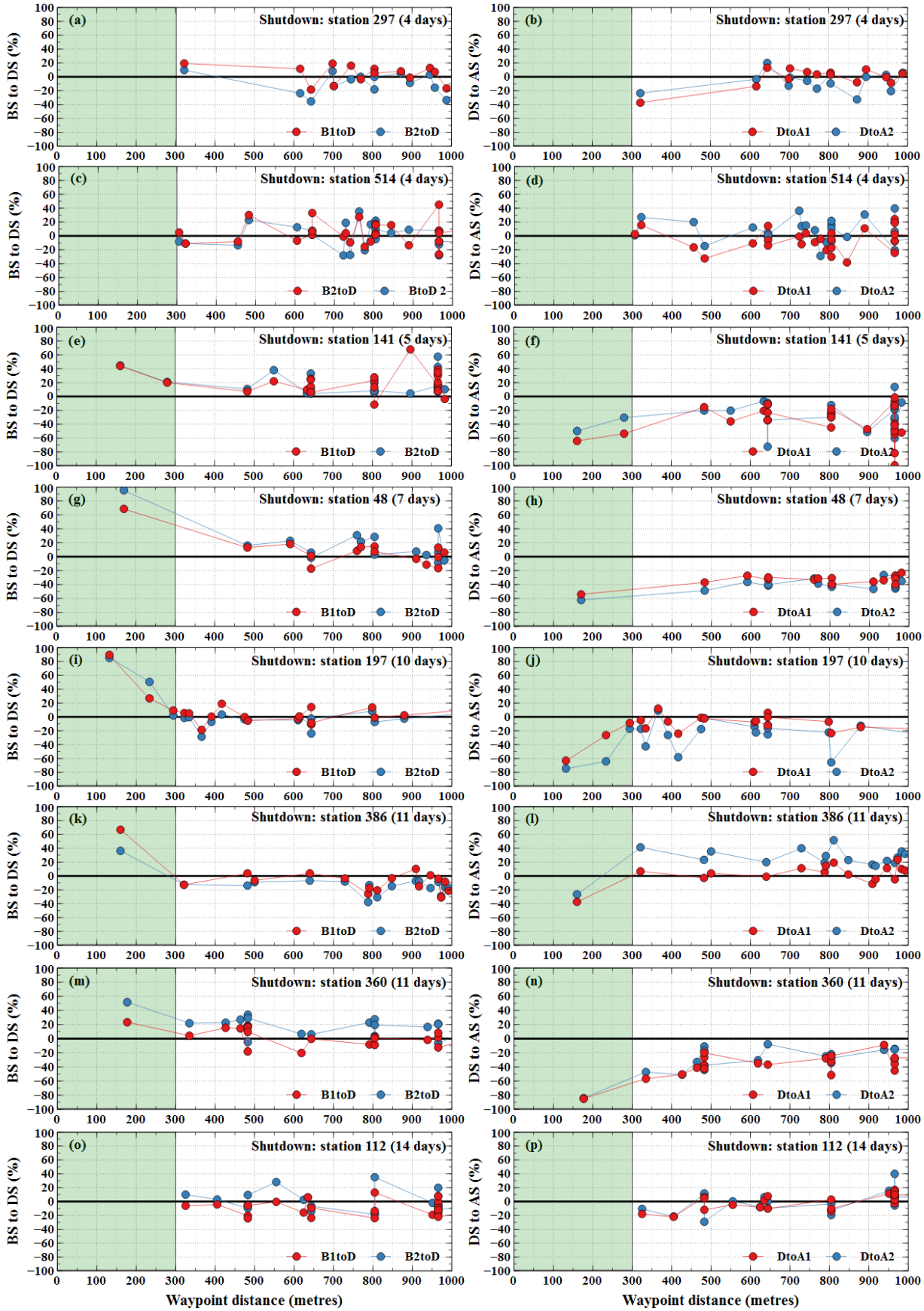


Figure 5.15. Average daily pickup transition (%) of nearby stations ***before-to-during*** *and* ***during-to-after*** *shutdown ordered by* ***waypoint distance.***

*Table 5.3. The transitions of **before-to-during** and **during-to-after** for average daily pickup (five closest stations to the shutdown station).*

Shutdown Station ID	Nearby Station ID	Euclidean Distance (metres)	Waypoint Distance (metres)	Average daily pickup					% of transition (Formula 5.5)			
				Before 2	Before 1	During Shutdown	After 1	After 2	B_2toD (%)	B_1toD (%)	DtoA$_1$ (%)	DtoA$_2$ (%)
297 (4 days)				**17**	**13.7**	**0**	**15.2**	**10.5**				
	549	126	322	31	28.5	34	24.75	27.5	9.7	19.3	-37.4	-23.6
	548	284	644	15.5	12.25	10	11.5	12.5	-35.5	-18.4	13.0	20.0
	324	324	698	49	44.5	53	51.5	47	8.2	19.1	-2.9	-12.8
	371	324	616	19	13	14.5	12.75	14	-23.7	11.5	-13.7	-3.6
	235	343	702	48.5	48.5	42	47.75	41.5	-13.4	-13.4	12.0	-1.2
514 (4 days)				**33.6**	**35.6**	**0**	**27.0**	**31.0**				
	400	194	325	41.0	41.0	36.5	43.3	50.0	-11.0	-11.0	15.6	27.0
	403	204	605	32.0	38.7	36.0	32.5	41.0	12.5	-6.9	-10.8	12.2
	210	259	306	62.0	54.3	57.0	58.5	57.5	-8.1	4.9	2.6	0.9
	99	262	483	48.0	45.3	59.0	44.5	51.5	22.9	30.1	-32.6	-14.6
	121	305	644	42.7	42.7	45.5	42.8	44.5	6.6	6.6	-6.4	-2.2
141 (5 days)				**37.3**	**33.6**	**0**	**24.3**	**27.6**				
	301	87	161	67.3	67.7	97.4	59.3	65.0	44.7	43.9	-64.2	-49.8
	106	134	280	45.3	45.7	54.8	35.7	42.0	20.9	20.0	-53.6	-30.5
	210	368	644	72.0	62.0	77.0	57.3	44.7	6.9	24.2	-34.3	-72.4
	6	375	550	45.7	51.7	63.0	46.3	52.3	38.0	21.9	-36.0	-20.4
	116	411	644	93.7	99.3	124.6	92.7	111.3	33.0	25.4	-34.5	-11.9
48 (7 days)				**76.6**	**82.4**	**0**	**62.5**	**69.9**				
	136	131	170	70.2	73.2	123.4	56.6	46.6	75.8	68.6	-54.1	-62.2
	71	327	483	128.8	132.2	149.4	94.2	76.6	16.0	13.0	-36.9	-48.7
	101	443	591	200.4	208.2	245.8	178.8	156.4	22.7	18.1	-27.3	-36.4
	427	461	644	146.8	153.6	155.4	104.8	90.8	5.9	1.2	-32.6	-41.6
	120	555	805	84.0	80.6	86.4	52.2	52.2	2.9	7.2	-39.6	-39.6
197 (10 days)				**74.8**	**86.7**	**0**	**78.5**	**81.1**				
	173	125	132	52.4	51.1	96.9	59.4	55.5	84.9	89.4	-63.1	-74.5
	377	144	234	48.8	48.7	64.3	50.9	39.2	50.4	26.7	-26.4	-64.1
	154	215	483	302.5	306.1	289.3	282.4	283.8	-4.4	-5.5	-2.4	-1.9
	361	223	322	66.8	62.3	65.7	62.6	56.0	-1.6	5.5	-4.9	-17.3
	273	223	334	91.8	86.9	91.3	78.3	64.0	-0.5	5.1	-16.7	-42.6
386 (11 days)				**72.3**	**66.8**	**0**	**49.4**	**73.1**				
	383	90	160	47.8	39.0	65.0	40.7	47.8	36.1	66.7	-37.4	-26.5
	192	242	321	65.0	65.1	56.8	60.6	80.2	-12.6	-12.8	6.7	41.3
	109	280	482	86.0	71.4	74.1	72.0	91.3	-13.8	3.8	-2.8	23.2
	244	307	500	42.9	41.7	39.0	40.4	52.8	-9.0	-6.5	3.7	35.3
	260	380	729	41.9	39.9	38.4	42.7	53.8	-8.2	-3.5	11.1	39.9
360 (11 days)				**58.6**	**72.7**	**0**	**52.7**	**61.1**				
	177	167	177	86.9	106.9	131.7	71.1	71.4	51.6	23.2	-85.1	-84.3
	316	246	483	86.4	98.6	115.6	83.9	98.7	33.8	17.2	-37.8	-17.1
	118	248	464	29.3	32.4	37.1	26.3	27.9	26.9	14.4	-41.2	-33.1
	359	258	335	52.0	60.9	63.3	40.4	43.0	21.8	4.1	-56.7	-47.3
	299	332	427	30.4	32.3	37.2	24.7	24.7	22.5	15.3	-50.6	-50.9
112 (14 days)				**264.4**	**285.5**	**0**	**181.8**	**190.8**				
	393	205	483	65.0	75.4	71.1	63.5	55.0	9.4	-5.7	-12.0	-29.3
	546	230	325	50.2	58.8	55.2	46.8	49.9	10.0	-6.1	-17.9	-10.6
	27	288	405	59.2	63.6	60.9	49.9	50.1	2.9	-4.2	-22.0	-21.6
	66	335	483	111.7	126.9	101.0	107.6	112.7	-9.6	-20.4	6.1	10.4
	67	386	483	68.6	85.9	65.0	68.4	73.6	-5.2	-24.3	5.0	11.7

A combined graph for all the shutdown stations is presented in Figure 5.16 where the waypoint distance order gives a better representation of impact decay over distance. It can be seen that the nearest station using waypoint distance gets the highest impact, while this is not the case in the Euclidean distance order.

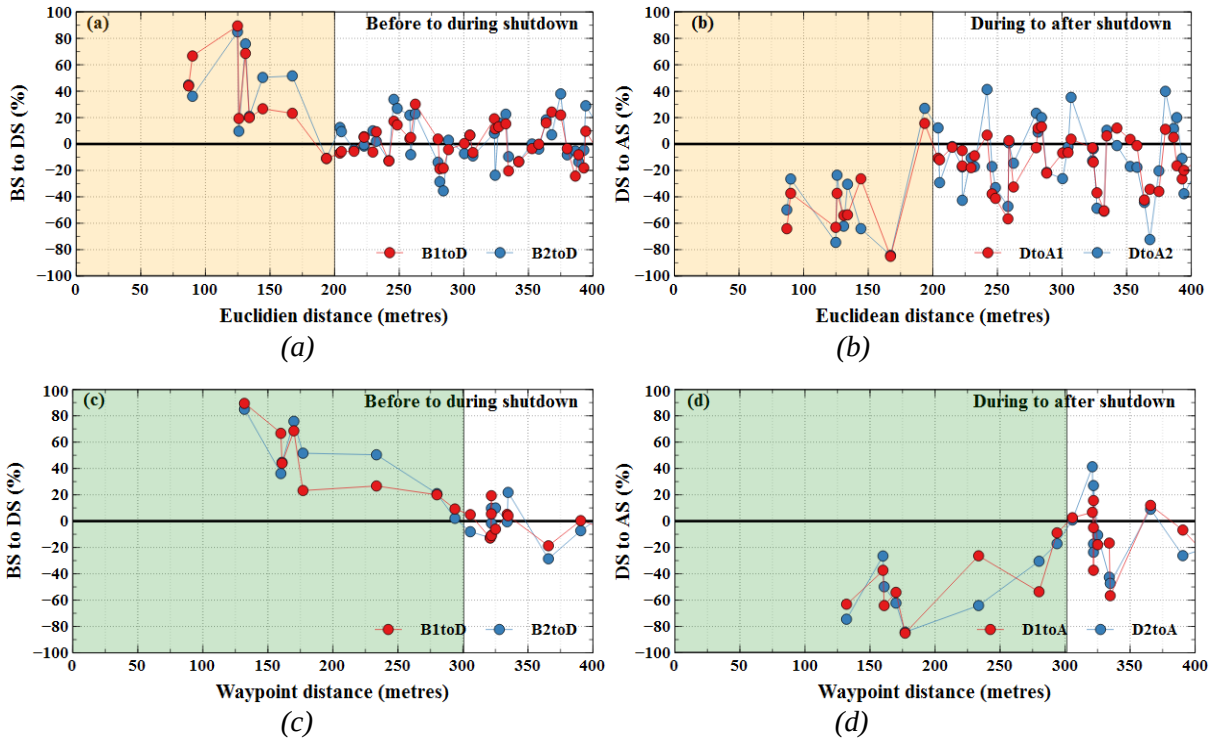


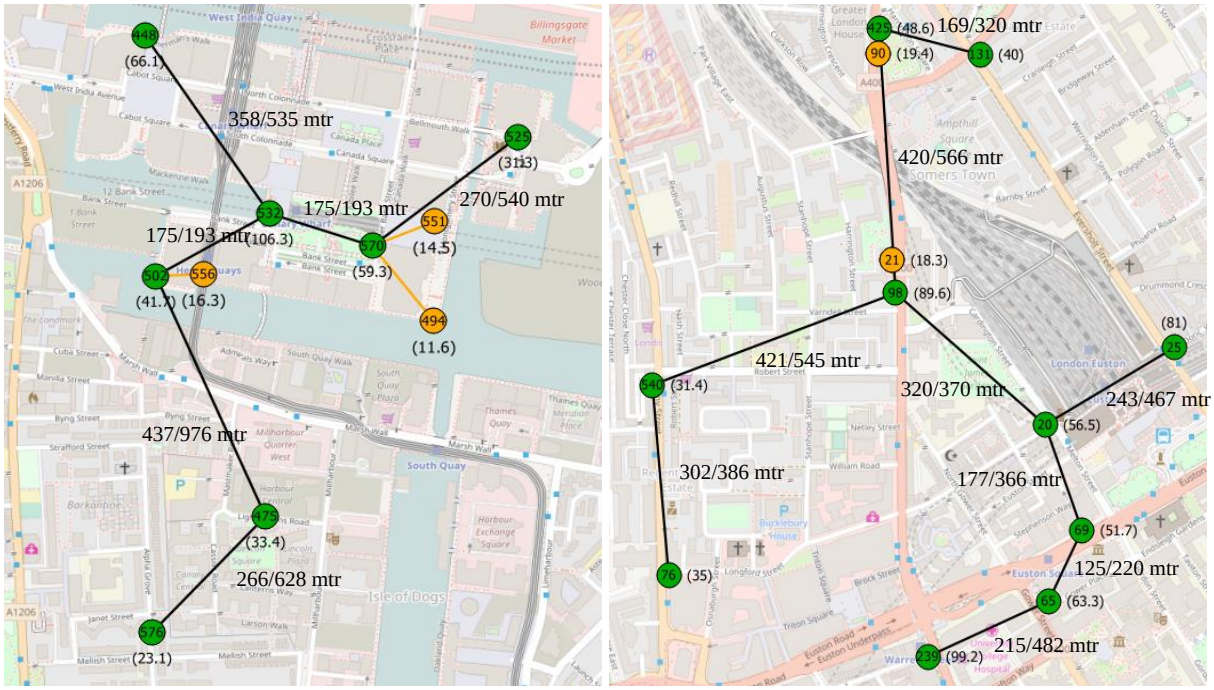
Figure 5.16. The recaps of average daily pickup transitions (%) of nearby stations ≤ 400 metres.

5.4. The Impact Distance Application

5.4.1. Ineffective stations

Knowledge from the shutdown impact distance as well as the users' spatial-mobility-motifs can be applied for detecting ineffective stations. The previous results of both analyses show that 300 m is the limit of the distance for choosing an alternate station. Hence, an ineffective station can be identified based on this distance combined its usage relative to its nearby stations. If a station has low usage, less than a threshold **(α)**, and if its removal from the network still gives inter-station distances in the range of **300 m** between the remaining nearby stations, then this station can be labelled ***ineffective***. This is because its removal is unlikely to have a big impact on the network. Its users can still be handled by nearby stations within 300 m as substitute stations. Removing ineffective stations and reallocating their resources can give

better overall system utilisation. Two cases are presented in Figure 5.17 with related Euclidean/waypoint distance and daily average usages.



(a) *(b)*

Figure 5.17. Ineffective stations example based on distance.

If α is set less than an average of 20 uses per day (subjective value) in the area shown in Figure 5.17.a, station 494, 551, and 556 are candidate ineffective stations. Now, by observing their distance to the nearest stations, they could be eliminated or amalgamated with the nearest stations. In this case, station 494 and 551 can be amalgamated with station 570, while station 556 can be amalgamated with station 502. Removing these three stations still give inter-station distances below than 300 m for the remaining stations.

In the area shown Figure 5.17.b, station 21 and 90 are candidate ineffective stations. In this case, station 90 can be eliminated or amalgamated with station 452 because there is station 131 that can be a backup for station 452. Similarly, station 21 can be eliminated amalgamated with station 98 because there is station 20 that can be a backup for station 98. In addition, these examples also indicate that if two stations are very close, one of them will be more dominant or receive more usage than the other. One of the next examples will show where a station could not be removed because the 300 m distance would be violated.

5.4.2. Isolated stations

An ***isolated*** station can be defined as a station that has high usage, and has no other stations within 300 m waypoint distance. An example of an isolated station is shown in Figure 5.18. Station 419 has high usage compared to its nearby stations. Its nearest station is station 245 at 585 m waypoint distance. If this isolated station is shut down or is full or empty, there is no nearby station within 300 m that can be a substitute. So, adding a new nearby station within 300 m is recommended.

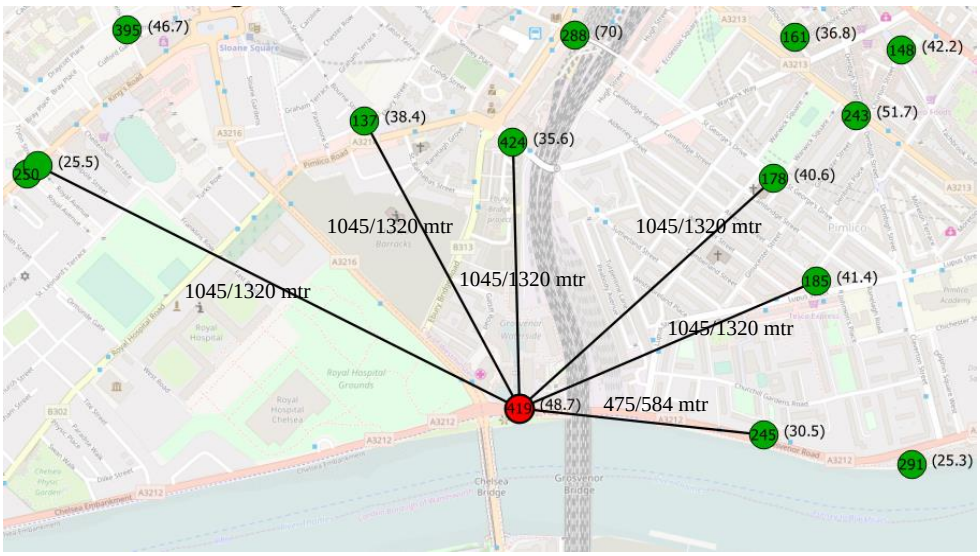


Figure 5.18. The isolated station example based on distance.

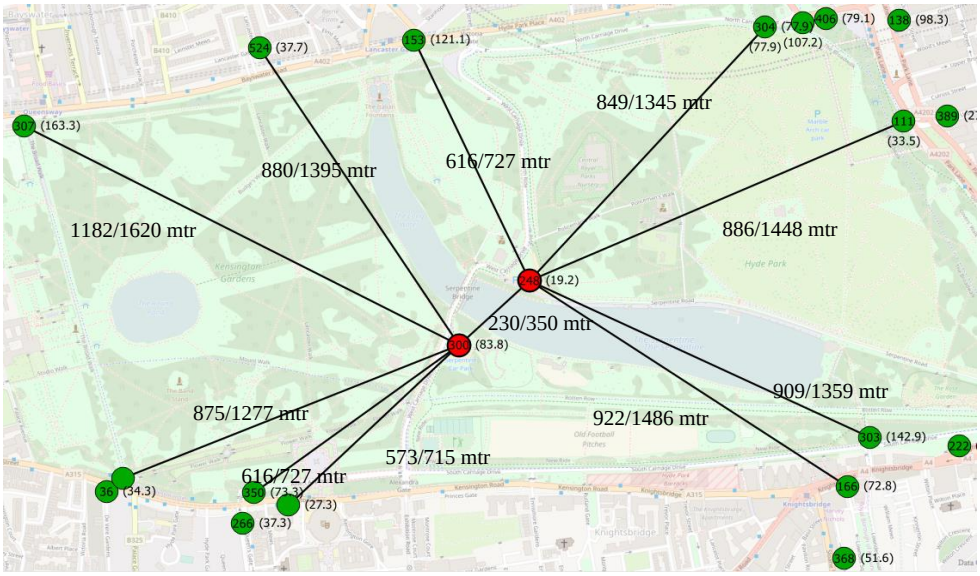


Figure 5.19. Two isolated stations example in Hyde Park.

An example of relatively isolated stations which provide mutual backup is shown in Figure 5.18. Stations 300 and 248 are in the centre of Hyde Park. In this case, even though station 300

and 248 are isolated with long distances to others, but they are still quite close to each other (230 m and 350 m for Euclidean and waypoint distance respectively). So, if one of them is shut down or full or empty, the other one still acts as a substitute. Reducing their separation would be better, because now it is around 350 m waypoint distance.

5.5. Station Neighbourhood Significance Summary

The results in this chapter show that nearby stations has neighbourhood ties, and that 300 m is the waypoint distance where these ties are significant. Two approaches, the spatial mobility motifs and the temporary shutdown station analysis, support this finding. Ten motifs are found where 300 m waypoint distance is the most common inter-station distance for users who choose different stations in the same area. Similarly for the shutdown stations analysis, 300 m waypoint distance is the distance with significant impact for usage changes in nearby stations. These changes decay from 80% to 20% as inter-station distance increases to 300 m.

This work has potential practical application in BSS system design, operation and maintenance. This impact distance knowledge can be used by the BSS operator to plan for station shutdown by ensuring other stations within 300 m can effectively cope with increased usage. Combined with usage information, this impact distance knowledge could be to identify ineffective stations in the network that can be eliminated. Another application is to identify isolated stations with high usage where a new nearby station in 300 m is recommended to provide system reliability, so that other alternate stations are sufficiently nearby to cope with station unavailability.

Finally, when a new BSS is planned for a city, this work provides additional planning insights. An inter-station distance of a maximum of 300 m will provide reasonable alternative stations during station unavailability. Furthermore, station distance should be calculated using waypoint distance, not Euclidean distance. Stations very close to each other are not advised – one is likely to be ineffective. The identified ineffective stations in the analysis in 5.4.1 were all close to a much more popular station. Also, from the spatial motifs distance distribution, Figure 5.19, a very close station (100 m or so) is not a dominant alternate choice for users.

CHAPTER 6
USER CLUSTERING AND NEXT PLACE PREDICTION

One of the common goals of human mobility studies is to be able to predict future trips, either at the level of individuals, or as aggregate movements across the area under study. Being able to predict the next location for individual users can potentially improve services to the users, for example suggesting nearby stations if their predicted target station is likely to be full when they arrive. Results presented in the preliminary data analysis chapter show that BSS has spatial and temporal regularities as well as significant randomness. The system wide regularity is likely to translate into individual trends in patterns of usage for some frequent users. Trips by some other classes of users are likely to be much less predictable, particularly when those users do not have a long trip history. Identifying users who demonstrate high regularity in the form of consistent temporal patterns seems significant for prediction and operation. Intuitively, users who regularly use the BSS for home-work commuting are likely to make similar trips at similar times on work days. To identify such users, this study will use a clustering approach based on appropriate temporal features of their trip data. Such features might correspond to patterns such as daily commuting. This chapter investigates prediction of individual users' next locations. Meanwhile, the next chapter deals with the different problem of predicting system-wide usage.

To the best of our knowledge, no previous research has investigated individual user trip predictability and prediction using BSS data, and so the results presented here are new for BSS mobility data. Previous work in BSS prediction has concentrated on the system-wide based predictions [47, 63, 70]. The London 2012 BSS data is the only publically available data tagged with individual user information.

The user-based analysis in this chapter consists of five sub-topics: user clustering, cluster characterization, cluster entropy and predictability, user next-location prediction, and practical applications. First, user clustering aims to classify users based on similar movement behaviour that is reflected in the regularity of their trip patterns. Since temporal regularity is more meaningful for frequent BSS users, *the total number of trips* as well as *the number of hourly trips* will be proposed as clustering features. Total trips will show how frequently an individual uses the BSS, while the hourly trip patterns will reflect the travel regularity within a user's daily routine. It is expected that clustering on these temporal patterns will provide more

homogeneous classes of users, rather than basing clusters solely on subscription categories. In addition, it also provides the ability to have more than two clusters. The proposed clustering will be compared to the user clusters from existing studies [74] and [75].

Second, to identify differences between clusters, the spatiotemporal metrics in Chapter 4 and 5 will be analysed by cluster. This cluster characterization analysis will highlight specific mobility behaviour of different groups of users at an aggregate level. Metrics will examine how users in different clusters use the BSS hourly and daily, their waiting times before the next trip, how their use is affected by season, how quickly they ride, their spatial extent as measured by their RoG, and their spatial daily motifs. These characteristics will allow meaningful labels for these clusters.

Third, the randomness and regularity of each cluster is examined using techniques from information theory, and *entropy* and *predictability* will be calculated [24]. This provide an upper bound to the potential prediction accuracy that can be achieved by a prediction algorithm [25]. Different entropy measures can be used to determine whether users' future trips have a strong spatiotemporal correlation with past trips, and depend only upon the current location, not on the sequence of trips that preceded it. If this is the case, then a Markov model should be a useful predictor.

Fourth, different prediction scenarios will be used to predict the next user location either for *pickup-to-return* or *return-to-pickup*. When a user visits a new station, their past history cannot be used for prediction. Subsection 4.2.5 showed that only around 20% of trips contain revisited stations. Accordingly, *population-based prediction* per cluster that represents the collective trends will be used to make a prediction if an *individual-based prediction* is not possible. In addition, to capture finer temporal resolution, the trip history will be further subdivided based on *day of the week* and *time of the day* rather than using the whole history as one OD transition matrix. Further analysis will determine which method gives better prediction accuracy. The dynamics of prediction accuracy over time and the correlation strength of *pickup-ride-return* and *return-wait-pickup* will also be examined to get insights into the quality of prediction. This will also be compared to individual prediction results from previous mobility studies from other modalities.

Fifth, some possible applications of this work in BSS operations will be presented, such as identifying the most common stations, shortest routes, and visiting times of highly predictable

users, and using this information to provide individualized notifications for those users. Finally, the results of this chapter will be used to answer RQ2 and a part of RQ4.

6.1. Technical Background

Before addressing the five topics mentioned above, this section presents the technical background mostly from information theory. It begins with entropy, followed by predictability, prediction accuracy, Markov models, the next place prediction scenario, and k-means clustering.

6.1.1. Entropy

Entropy is commonly used to capture the degree of randomness in a list of visited locations in which there are temporal scales of variability between locations. Here, entropy is applied to measure the randomness of BSS mobility denoted by the sequence of visited locations where pickup and return stations are both counted as visited locations without considering the routes in between. Following [24, 42, 45], there are four different representations of entropy.

a. The random entropy (S_i^{Rand}) for an individual user i only considers the number of distinct BSS stations, N, visited by that user.

$$S_i^{Rand} = log_2 N \tag{6.1}$$

Because $log_2 0$ is undefined, we need $N>0$. Since all users visit at least one station, $N > 0$ for all users.

b. The Shannon entropy (S_i^{Shan}) for an individual user i counts the probability of visiting each distinct visited station, j, in his/her visitation history, summed across all stations that are visited at least once.

$$S_i^{Shan} = -\sum_{j=1}^{N} p_{ij} \, log_2 \, p_{ij} \tag{6.2}$$

where p_{ij} = *number of visits to station j by user i / total visits for all stations visited by user i.* This will ensure that p_{ij} is always > 0.

c. The conditional entropy (S_i^{Cond}) for an individual user i captures the correlation between visiting one BSS station x_{t-1} with the subsequent station x_t in the time series of

locations. Here, t denotes the integer order in the sequence of visited stations, so that x_{t-1} is the previously visited station before station x_t.

$$S_i^{Cond} = -\sum_{x_t \in X_i} \sum_{x_{t-1} \in X_i} p_i(x_{t-1}, x_t) log_2\, p_i(x_t | x_{t-1}) \quad (6.3)$$

where X_i is the set of all stations visited by an individual user i, $p_i(x_{t-1},x_t)$ is the probability of visiting the ordered pair of visited stations, x_{t-1} and x_t by user i, while $p_i(x_t|x_{t-1})=p_i(x_{t-1},x_t)/p_i(x_{t-1})$ is the probability of visiting the visited station x_t at time-ordered t given a preceding visited station, x_{t-1} by user i. Only pairs that appear in a user's history are used to ensure $p_i(x_t|x_{t-1}) > 0$.

d. The real entropy evaluates the randomness based on the full spatiotemporal 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 of locations. More precisely, for a sequence of length n, the estimated value of entropy is

$$S_i^{Real} = \left(\frac{\sum_{m=2}^{n} l_m}{n\, log_2 n}\right)^{-1} \quad (6.4)$$

Where l_m is the length of the shortest sequences of locations starting at position m that does not appear in the part of sequences up to position $m - 1$.

6.1.2. Predictability

An important measure is predictability Π which is the upper bound of the accuracy for a prediction algorithm to correctly predict the user's next location [24]. For instance, Π = 0.4 means that the user's next location is 40% predictable at most, while at least 60% of his/her next locations are random and unpredictable. The predictability Π_i of user i is subject to Fano's inequality [123] and can be related to the user's entropy S_i by:

$$S_i^{\bullet} = H(\prod_i^{\bullet}) + (1 - \prod_i^{\bullet}) log_2(N_i - 1) \quad (6.5)$$

with $H(\prod_i^{\bullet})$ being the binary entropy function which is defined as the entropy of a Bernoulli process with the probability of success $\prod_i^{\bullet}$ that can take only two values: 1 (success) and 0 (failure).

$$H(\prod_i^{\bullet}) = -\prod_i^{\bullet} log_2 \prod_i^{\bullet} - (1 - \prod_i^{\bullet}) log_2 (1 - \prod_i^{\bullet}) \quad (6.6)$$

where $\bullet$ is a placeholder for any type of entropy, and N_i is the total possible locations visited by user i based on his/her history. In other words, given the entropy S, we can find the predictability Π by solving Equation (6.5) numerically. In this thesis, the solution is obtained

by the *fsolve function* of the optimization package in Python where it returns the roots of the (non-linear) equations defined by func(x) = 0 given a starting estimate.

The different predictability values give upper bounds for prediction accuracy, based on using different information. Random predictability represents the accuracy possible by randomly selecting one of out the set of possible locations as the prediction, Shannon predictability gives the prediction accuracy of selecting the most popularly visited location for that user, conditional predictability gives the accuracy possible by basing the prediction on the one previously visited station, and real predictability gives the accuracy possible by using the complete history.

6.1.3. Prediction Accuracy

Prediction accuracy of a predictor can be defined as a ratio between the number of correct predictions over the total number of predictions [51]. Here, the correct prediction is set by discrete validation prediction in which the prediction outcome is binary (true or false).

$$P_{Acc} = \frac{P_{true}}{P_{all(true+false)}} \tag{6.7}$$

6.1.4. Markov Model

A Markov Model uses the current state (locations) to determine the likelihoods of the subsequent state (the possible next locations). This predictor provides a simple approach to capture sequential dependence, and is defined by a set of states $S = \{S_1, \dots, S_n\}$, a transition matrix $T = \{T_{1,1}, \dots, T_{n,n}\}$, and a vector of initial probabilities $P = \{p_{1,1}, \dots, p_{n,n}\}$ [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 [51]. Here, a state can be a location (if only one previous location is considered) or can be a sequence of previously visited locations. This will define the order of Markov Model. A first order model states equal to the one location, and a second order model has states which are ordered pairs of visited stations.

Once a Markov model is built, then the next state (location) is predicted from the current state based on the highest transition probability.

A Markov Model can be represented either as a directed graph, Figure. 6.1, or a probability transition matrix, Table 6.1. It is shown that the total probability from one state to all other states is 1.

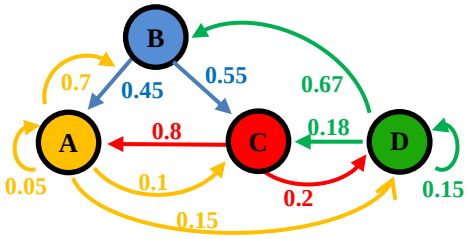


Figure 6.1. Graph representation example of transition states by nodes and edges.

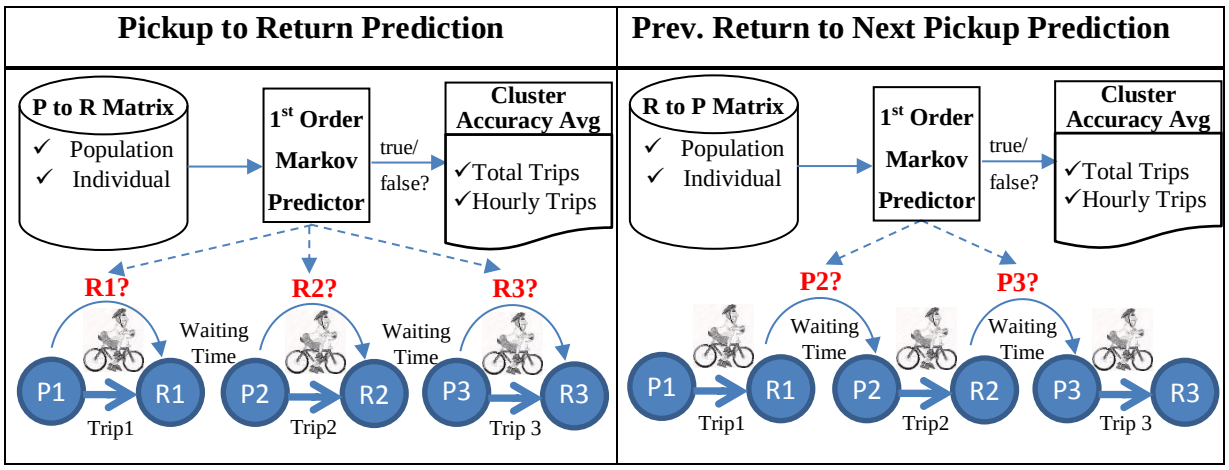
Table 6.1. The probability transition matrix example.

State	A_{i+1}	B_{i+1}	C_{i+1}	D_{i+1}
A_i	0.05	0.7	0.1	0.15
B_i	0.45	0	0.55	0
C_i	0.8	0	0	0.2
D_i	0	0.67	0.18	0.15

6.1.5. Next Place Prediction Scenarios

The first prediction scenario is the first order Markov Model which considers only one location, either pickup or return station, depending on the desired prediction of the next station activity (*pickup-to-return* or *return-to-pickup* prediction). To predict the next return station, the predictor will only observe one previous pickup station, Figure 6.2.a. To predict the next pickup station, the predictor will only observe one previous return station, Figure 6.2.b. In this case, the predictor will search for the highest probability value in the OD transition matrix to find where that user is most likely returning his/her bike or most likely picking up a new bike.

The OD transition matrix will be constructed based on individual history. However, prediction can also be attempted when a user visits a new station with no trip history. In this case, the population-based history will be used. Using this approach, predictions can be attempted for all trips in a test set. This will also show if using the population-based history can assist the accuracy. The prediction accuracy itself will be presented per user cluster in two temporal forms, on a daily basis within the test set and on an hourly basis on weekdays to see the accuracy dynamics over time.



(a) *(b)*

Figure 6.2. Prediction scenario of the first order Markov Model.

The second scenario is the second order Markov Model as shown in Figure 6.4. Here, instead of considering only one previous station, this model examines the two consecutive previous stations as states in the transition matrix. This pair will make the OD transition matrix much larger. The example of the second order Markov Model for three stations (A, B, C) is shown in probability transition diagrams, Figure 6.3, and probability transition matrix, Table 6.2. This approach may improve the accuracy for users with sufficient trips to build such a transition matrix.

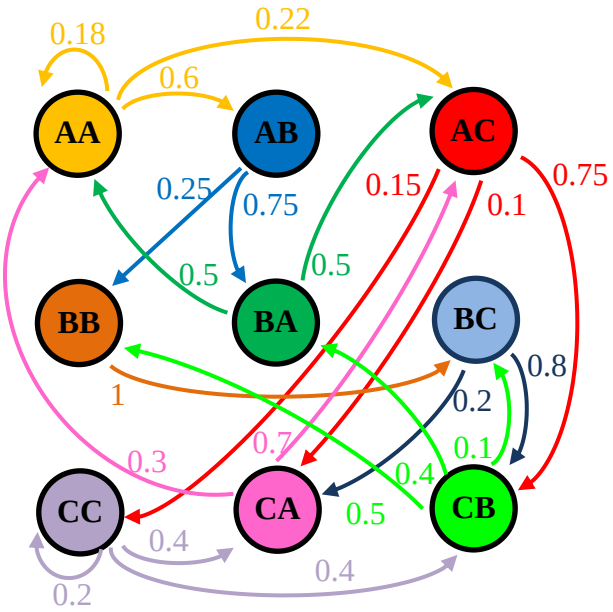
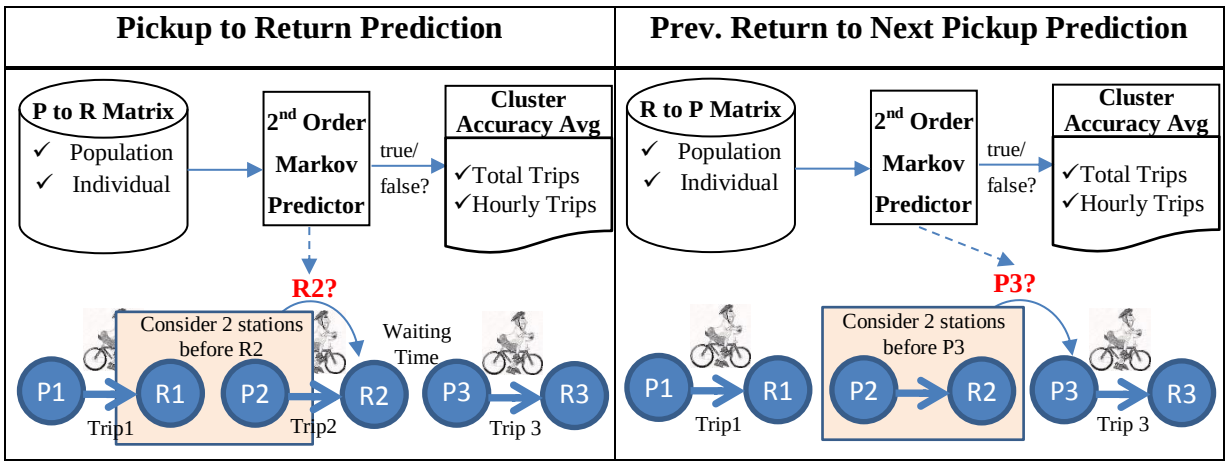


Figure 6.3. The second order probability transition states.

Table 6.2. The second order probability transition matrix example.

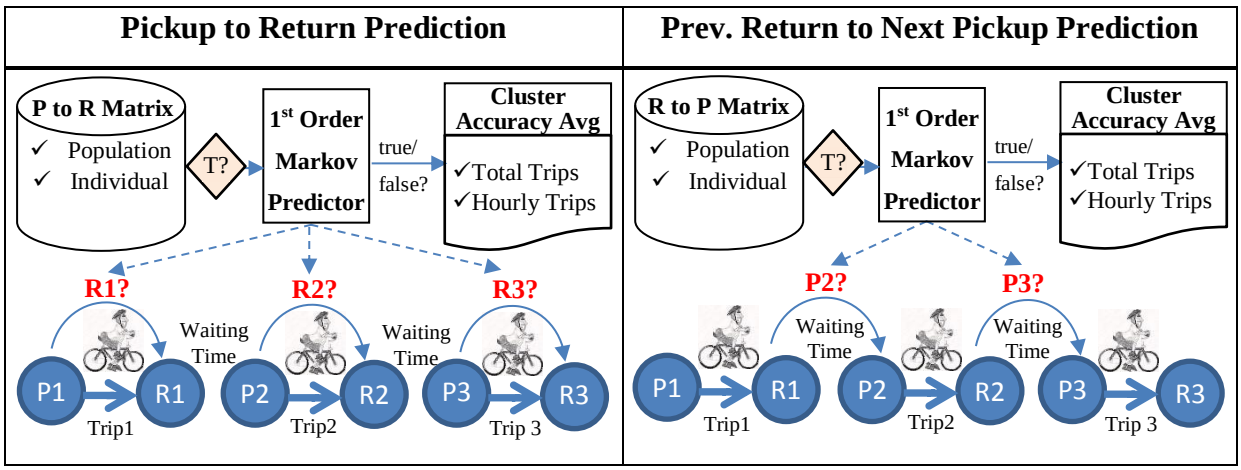
State	A_{i+1}	B_{i+1}	C_{i+1}
AA_i	0.18	0.6	0.22
AB_i	0.75	0.25	0
AC_i	0.1	0.75	0.15
BB_i	0	0	1
BA_i	0.5	0	0.5
BC_i	0.2	0.8	0
CC_i	0.4	0.2	0.4
CA_i	0.25	0.25	0.5
CB_i	0.3	0.5	0.2



(a) *(b)*

Figure 6.4. Prediction scenario of the second order Markov Model.

The third and fourth scenarios are made by taking into account the temporal aspect of trip data, daily or hourly, as shown in Figure 6.5. Here, separate OD transition matrices will be constructed based on either *the day of the week* (third scenario) or *peak times of the day* (fourth scenario). For day of the week, the OD matrix will be divided into seven, while for peak times of the day it will be divided into three: OD matrix in the morning peak (5 – 9 am), afternoon peak (3 – 7 pm) and the times out of those peaks. Then the algorithm will examine the day or time when the trip activity occurs and look to the associated OD Matrix. For example, if a user takes a bike on Monday, then the daily based predictor uses the Monday OD matrix, or if a user takes a bike at 7 am, then the hourly based predictor only uses the morning OD matrix. This subdivided matrix approach may be able to increase the accuracy by capturing users' temporal routines.



(a) *(b)*

Figure 6.5. Prediction scenario of the first order Markov Model using peak time and daily filter.

6.1.6. K-Means

K-means is one of the simplest unsupervised learning algorithms which aims to partition a group of data points into a small number of clusters. By defining k centres first, one for each cluster, each point from the given data will be associated with the initial nearest centroid or centre of the cluster. Then, iterative refinement will be employed to re-calculate the new k centroids and a new binding has to be done between each point and the nearest new centroids. This looping process will be done continuously until centroids do not move anymore. If $X = \{x_1, x_2, \ldots, x_n\}$ is a set of feature vectors, then the k-means algorithm attempts to minimize the squared distance function: $O = \sum_{i=1}^{k} \sum_{x \in G_i} (\|x - \mu_i\|)^2$ in order to cluster those n feature

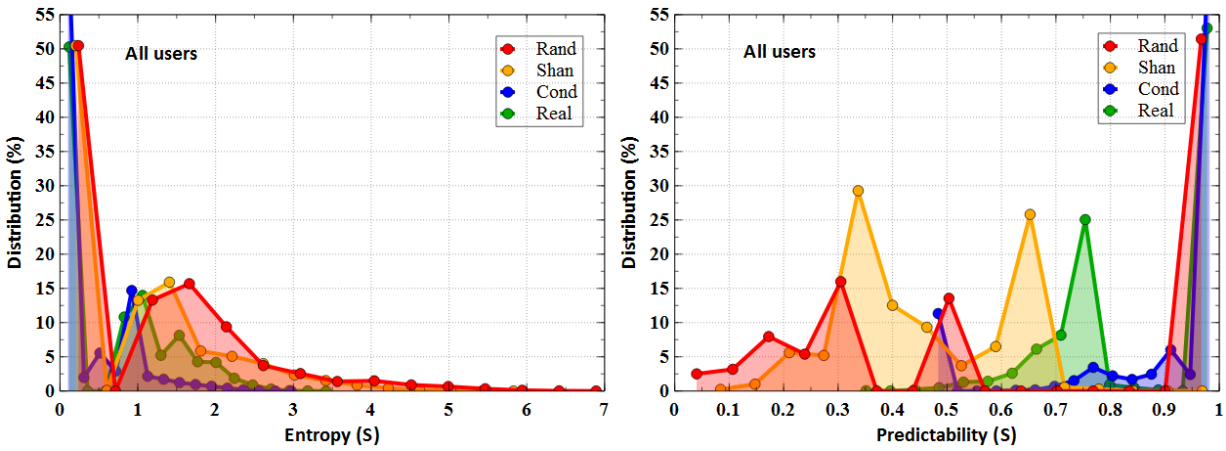
vectors into k clusters, namely $G_1 \ldots G_k$ where μ_i is the centroid of cluster G_i. Commonly, the components of the feature vectors are normalized to give them equal weight in clustering.

6.2. Preliminary Entropy and Predictability

To measure the randomness of visitation patterns or the uncertainties of movements among users, the four types of *entropy* (S) will be compared. The inverse of entropy yields *predictability* (Π) that expresses how predictable a user's movements are.

6.2.1. Randomness and Regularity of All Users

Using formulas 6.1 to 6.6, all types of entropy and predictability are computed for all users to get the preliminary insights of their distributions as shown in Figure 6.6. As the maximum value of entropy is nearly 7 shown by random entropy, the bins sizes are set to 15 so that users are placed into bins of entropy around (0 to 0.5), and then the percentages are shown for each bin. Similar bins sizes are also implemented for predictability distribution.



(a) *(b)*

Figure 6.6. Entropy and predictability of all users.

It can be seen that the entropy distributions do not follow the rule of entropy order, $S^{Real} \leq S^{Cond} \leq S^{Shan} \leq S^{Rand}$, and the predictability distributions are jagged and hard to analyse. This suggests that BSS users have wide ranges of entropy (from 0.1 to 6.9) and predictability (from 0.05 to 0.95) that reflect the variety of randomness as well as regularity. Users with a low number of trips, e.g., one trip will give very low entropy as well as very high predictability which relates to the peaks left (entropy) and right (predictability). If users with similar mobility behaviour, either those with high randomness or those with high regularity, are separated into different clusters, then highly predictable users may be able to be more easily identified.

6.2.2. Randomness and Regularity of Subscription-based Users

Most BSS trip data separates users based on their subscription status, such as *registered* and *unregistered* users. Registered users might have an annual subscription, while unregistered users just provide payment details each time they hire. Table 6.3 shows that around 90% of users are unregistered users. However, their trips only cover 36.5% of total trips. This means that the remaining 10% who are registered users have 63.5% of total trips. There is a large difference in the average trip numbers per user, 1.9 and 32.4 respectively for unregistered and registered users. Furthermore, the standard deviations of trip numbers are slightly higher than their averages. This will produce a fat-tail in the right side of their distribution as shown in Figure 6.7. This figure also demonstrates the overlaps in distribution in which some registered users have a small number of trips, and some unregistered users have a quite high number of trips. This may lead to the inhomogeneous characteristics within clusters. In other words, some unregistered users show registered user characteristics and vice versa. So, they become outliers in their own subscription group.

Table 6.3. The statistics of users by subscription.

Users Types	Number of trips per user				% of Users	% of Trips
	Min	Avg	Stdev	Max		
Unregistered	1	1.9	2.1	162	90.33%	36.50%
Registered	1	32.4	44.6	1054	9.67%	63.50%

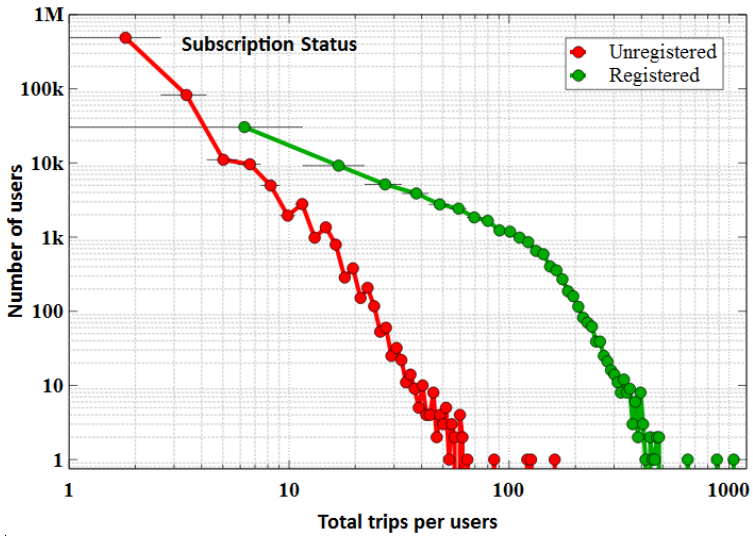


Figure 6.7. Total trips distribution per user by subscription in log-log scale (bins 100).

Using this subscription status, the entropy and predictability distributions can then be separated as shown in Figure 6.8. Unregistered users show almost similar jagged distributions compared to the previous distributions for all users. Meanwhile, registered users show a better normal distribution, but they still do not follow the entropy rule order and still contain peaks in the lower side of entropy (left side) as well as in the higher side of predictability (right side) corresponding to users with a small number of trips. Subscription status does separate predictable users and so different user clustering is proposed in the next section.

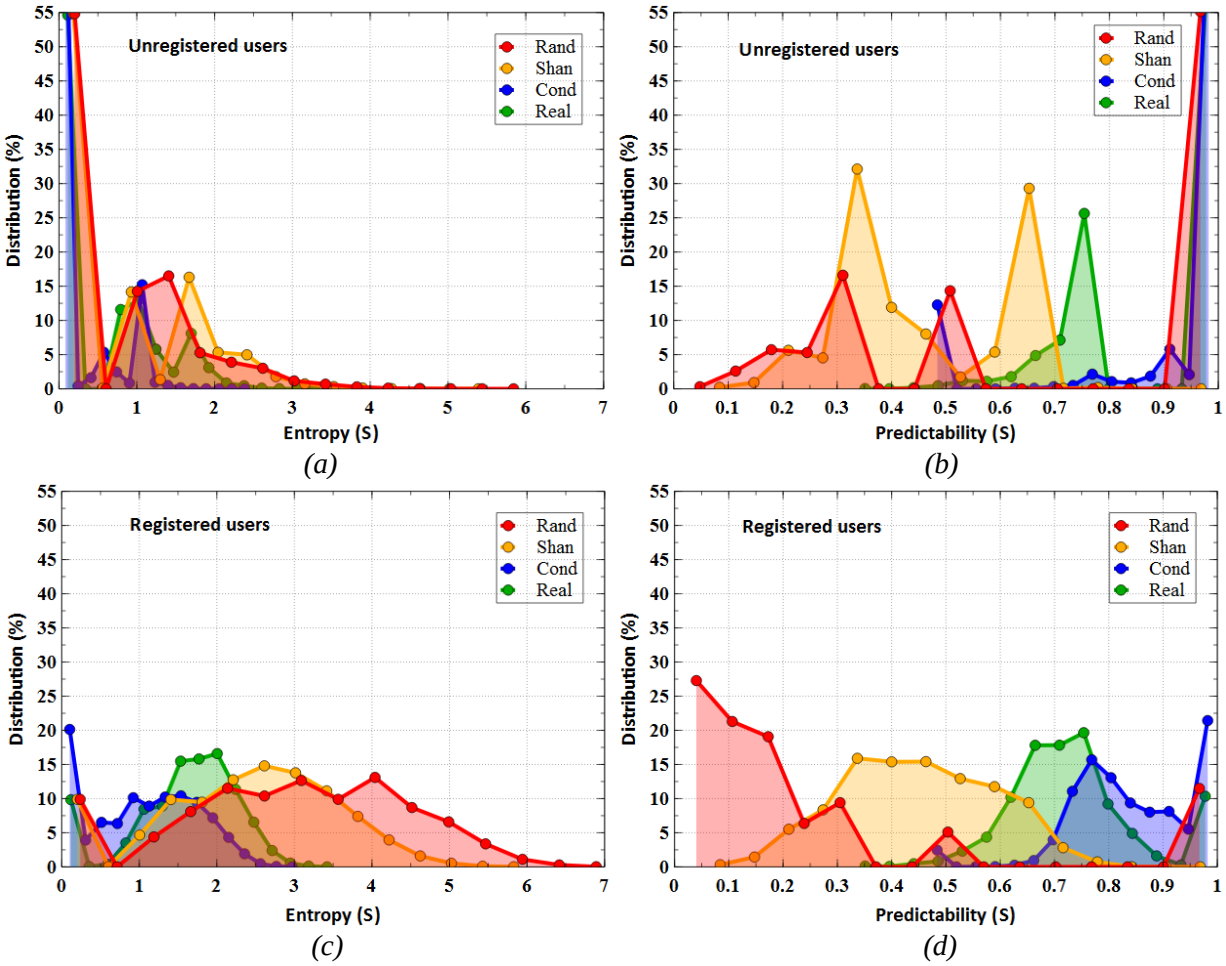


Figure 6.8. Entropy and predictability of unregistered users (a,b) and registered users (c,d).

6.3. User Clustering

The previous preliminary entropy and predictability results show that there are still outliers in the groups of users by subscription. On the other hand, group-based analysis and prediction benefits from homogenous user groups. Hence, this section proposes user clusters using two temporal approaches which are ***total trip clustering*** using upper and lower bound thresholds and ***hourly trip clustering*** using k-means, as described in detail below. Table 6.4 summarizes

the percentages, average and upper and lower bound threshold of each user cluster. Firstly, the preliminary labels are simply given using alphabetic order which are **cluster A, B,** and **C** or **D, E, and F**. Later, each cluster will be labelled based on their spatiotemporal characteristics. Three clusters are chosen since we anticipate that users are either frequent users, rare users, or somewhere in between. It is expected that frequent users will have obvious different spatiotemporal characteristics to rare users, while the outliers of both will be grouped into one middle cluster.

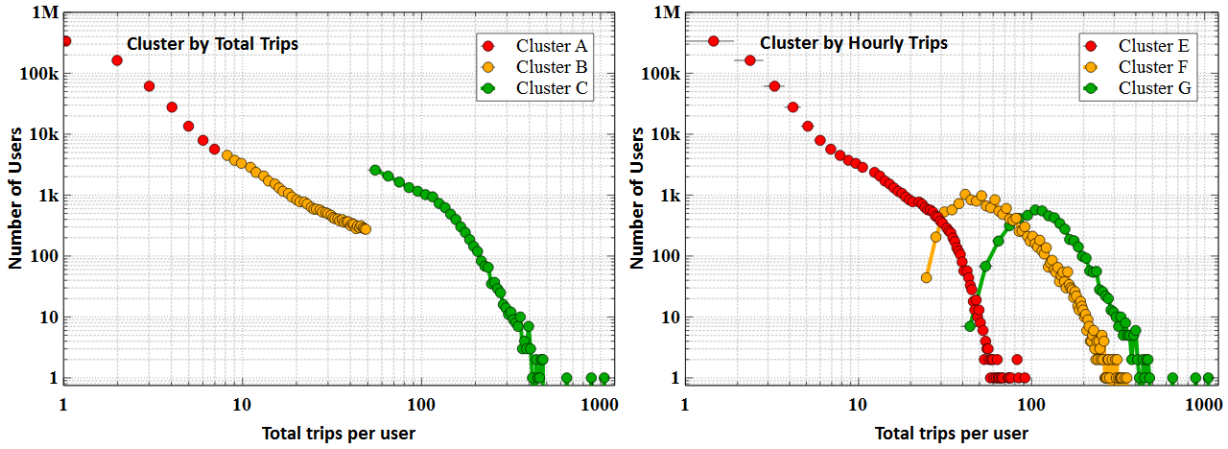
Table 6.4. The statistics of users clustering.

Clusters Users	% of Subscription		Number of trips per user				% of Users	% of Trips
	Unreg	Regist	Min	Avg	Stdev	Max		
Cluster by Total Trips:								
Cluster A	96.02 %	3.98 %	1	1.8	1.1	7	91.86 %	33.65%
Cluster B	35.47 %	64.53 %	8	18.5	10.9	49	2.15 %	22.51%
Cluster C	0.17 %	99.83 %	50	100.6	49.2	1054	5.99 %	43.84%
Cluster by Hourly Trips:								
Cluster D	92.97 %	7.03 %	1	2.5	3.8	92	97.15 %	50.50 %
Cluster E	0.39 %	99.61 %	23	68.7	34.9	357	0.77 %	29.04 %
Cluster F	0.02 %	99.98 %	39	131.7	55.9	1054	2.08 %	20.45 %

For the three proposed user clusters based on total trips, the thresholds for the clusters are listed in Table 6.4, and has been published in [46]. The threshold less than 8 captures lowest third of trips, more than 50 captures 99% of registered users, and leaves outliers to the middle cluster. Here, **cluster A** is intended for users who have few trips, and they could be very hard to predict because of a lack of history data for learning. On the other hand, **cluster C** is intended for users who have a lot of trips and is expected to be the most predictable user group. **Cluster B** is intended to accommodate users who have mixed characteristics between **cluster A** and **cluster C**.

Using 50 trips as the lower threshold for cluster C [23], they are only around 6% of users but have around 44% of trips. Almost all members of this cluster come from registered users, only 0.17% come from unregistered users. Cluster A, on the other hand, dominate the users' population where the upper threshold of this categorization is at 7 [46]. Using this threshold, they are almost 92% of users, where 96% of them are from unregistered users and only 4% are from registered users. However, their trips which are 33.65% of the total are still less than cluster C trips. Cluster B, with total trips between 8 and 49, is the cluster with the least members and has only 2.15% of the users and 22.51% of the trips. Using this threshold

technique, there is (by definition) no overlap in trip number distributions as shown in Figure 6.9.a.



(a) *(b)*

Figure 6.9. Total trips distribution per user cluster in log-log scale.

The second proposed user clustering approach in this study is based on the pattern of the hourly trips using k-means. This aims to observe whether the number of the hourly trip in daily basis can give a significant differentiation of user clusters. Here, total trips *(T)* per user are averaged *(A)* per hour of the day for the whole learning dataset as:

$$T_{DailyTrips} = A_{h0} + A_{h1} + A_{h2} + \cdots + A_{h23} \qquad (6.8)$$

As a result, each user has 24 hours of trips that are used as 24 input features for k-means clustering without any scaling or normalization. As listed in Table 6.4 and shown in Figure 6.9.b, this approach gives a different breakup. Most unregistrered users are still in one Cluster (D), but now registered users are split between Clusters E and F. The next section will present the characterization of each cluster to investigate their usefulness for prediction. Afterwards, an appropriate label can be given to each of these clusters.

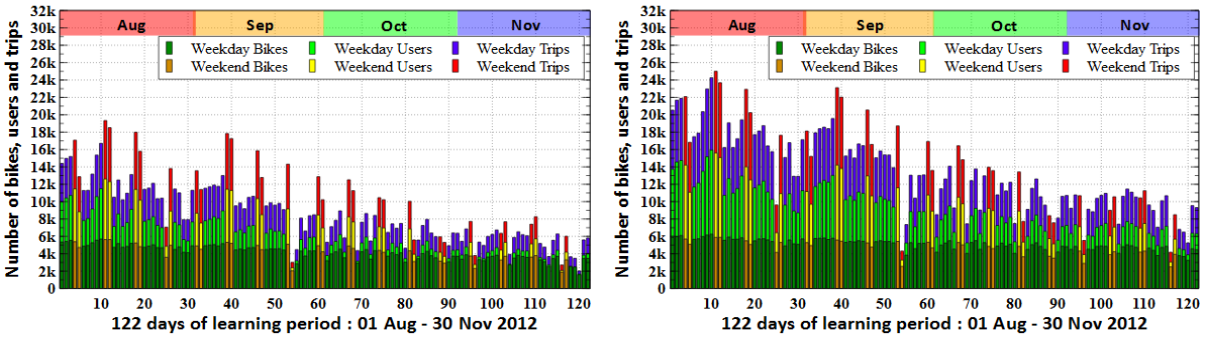
6.4. Cluster Characterization

This section will characterize each cluster using the spatiotemporal analysis methods that have been presented in the preliminary data analysis and also stations' neighbourhood ties chapters to find any significant differences among them so that they can be labelled. Analysis starts with cluster daily pattern analysis mainly to reveal if clusters have strong indications of typical commuting patterns, and how usage patterns vary along the period of study. This is followed by analysis of cluster hourly patterns and waiting times. These analyses aim to further

explore the commuting patterns as well as the variability of trips at a smaller time scale. Features of trip speed, users' distance growth via RoG analysis, and the spatial motifs of each cluster will be explored. Finally, after all spatiotemporal characteristics are highlighted then the associated label is given for each cluster.

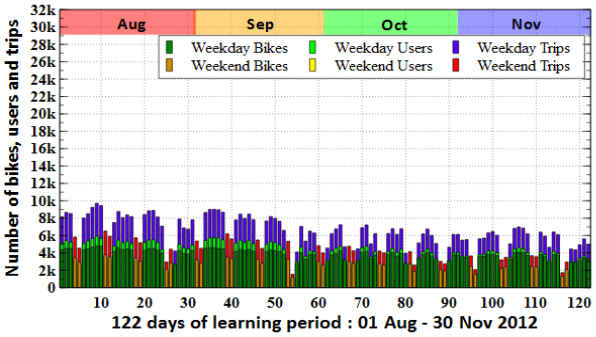
6.4.1. Cluster Daily Pattern

Daily pattern analysis in Chapter 4 explained some trends of temporal metrics. There is more usage on weekdays than on weekends, a proportion of users travelled more than once a day, and usage decreases towards the end of the year. It is expected that each proposed user cluster has distinct behaviour for these contexts. Their daily usage patterns are presented in Figure 6.10.

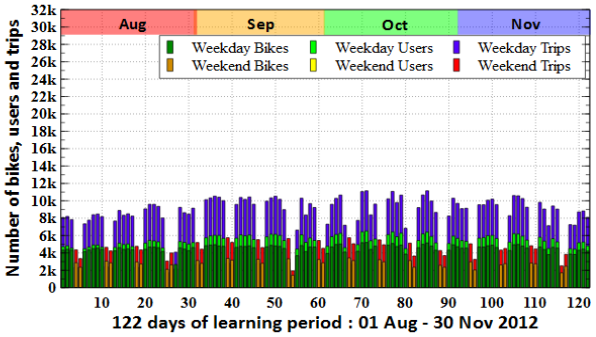


(a) Cluster A-by-total

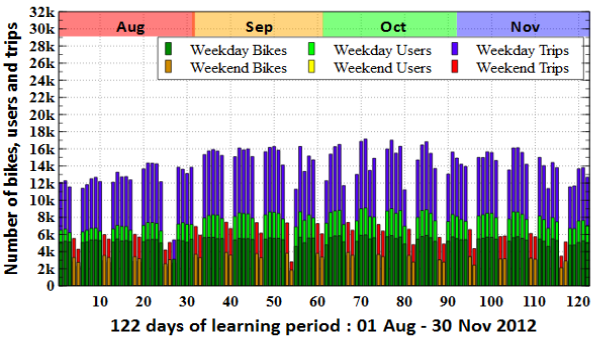
(b) Cluster D-by-hour



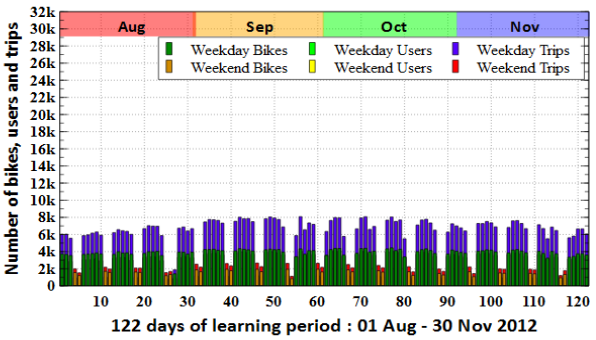
(c) Cluster B-by-total



(d) Cluster E-by-hour



(e) Cluster C-by-total



(f) Cluster F-by-hour

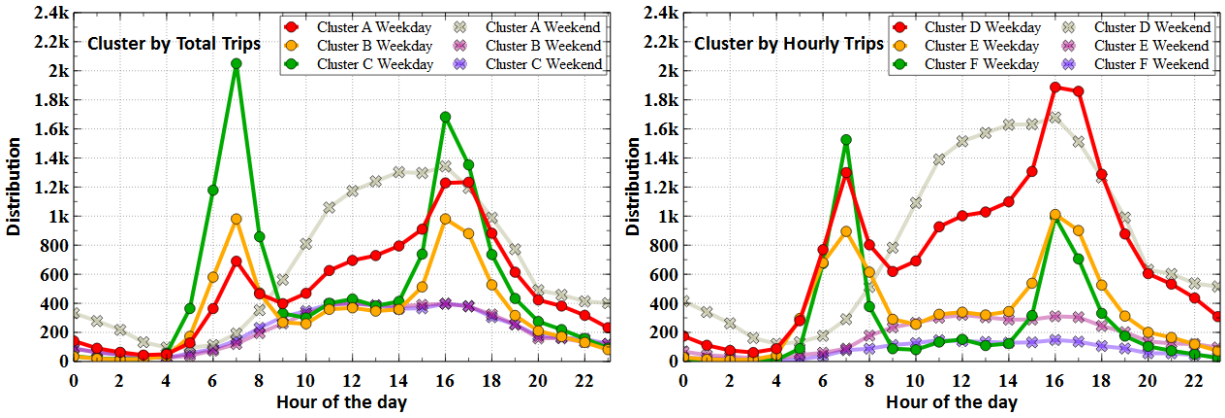
Figure 6.10. Daily trips and user number of each cluster.

It can clearly be seen that clusters A/D and C/F have different daily trip patterns. Both of clusters A/D, Figure 6.10.a&b, are strongly affected by season, decreasing towards winter, while both of clusters C/F, Figure 6.10.e&f, are relatively more stable. In terms of weekday and weekend usage, cluster C show clear commuting patterns where weekday usage is much more than weekend usage and clusters B and E also show this pattern. Meanwhile, clusters A and D show the opposite trend where weekend usage is more than weekdays.

The size of the cluster A and D differ by only 6.29%, but there is significant difference in average number of trips, as shown in Figures 6.10.a&b. The commuting pattern in clusters B,C,E and F appears because their average trips per user are large enough to establish that pattern.

6.4.2. Cluster Hourly Pattern

The clearest temporal pattern given by the overall hourly trip patterns in the preliminary data analysis chapter were identifying peak times and commuting usage on weekdays. In this section, clusters C and F strongly show these traits in the morning and afternoon peak times with low usages in the middle of the day, shown by the green lines in Figure 6.11. Cluster B shows a similar tendency to cluster C but with lower peaks. On the other hand, clusters A and D show a small peak in the morning, then after 9 am they gradually increase until reaching a peak at 4 pm and 5 pm in the afternoon. On weekends, trips are dominated by cluster A/D with a broad peak from 9 am to 8 pm, while cluster F has the least number of trips, as shown in Figure 6.11.b. This means that clusters C/F are very active on weekdays and inactive on weekends.

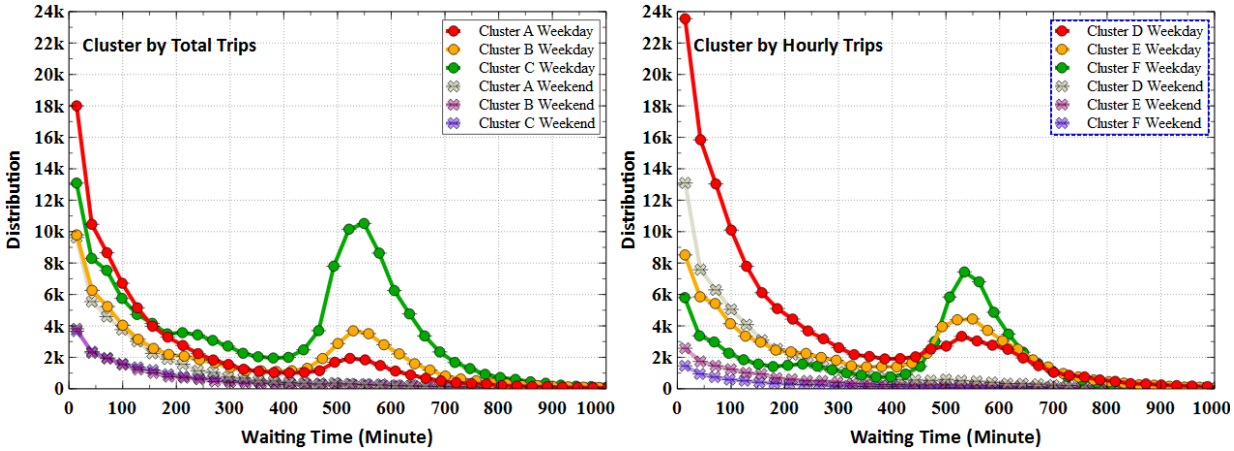


(a) *(b)*

Figure 6.11. Weekday and weekend hourly trip patterns per cluster.

6.4.3. *Cluster Waiting Time*

Waiting time is one of the temporal metrics that can also show commuting patterns on weekdays. Shown by the green lines in Figures 6.12, both of cluster C and F have a high number of waiting times between 500 and 700 minutes (7-10 hours) on weekdays followed by cluster B and E suggesting the normal office hours length. Meanwhile, cluster A and D have the lowest number of waiting times between 500 and 700 minutes, but they have the highest short waiting times, WT < 100 minutes, while cluster C and F is the least. Again, this tells that cluster C and F show a strong commuting pattern.

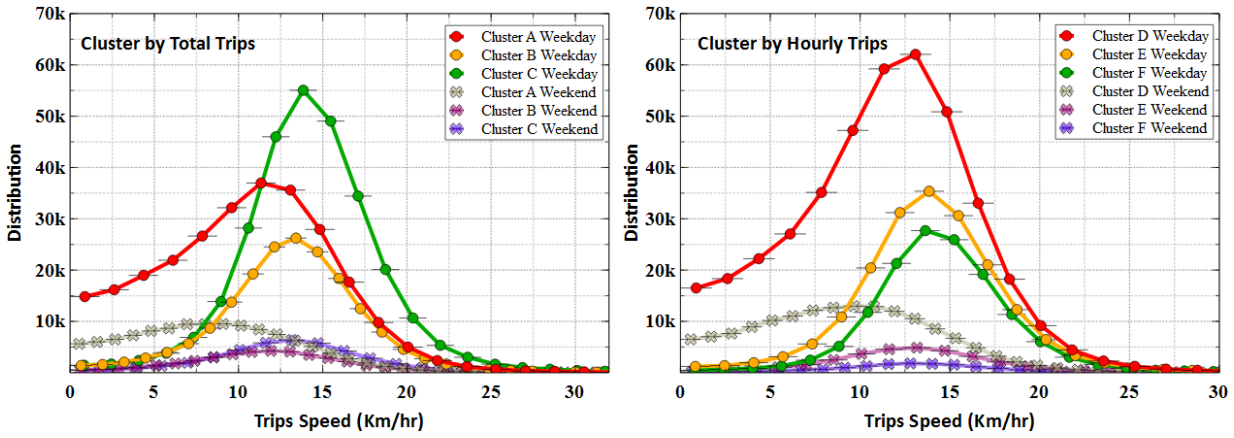


(a) *(b)*

Figure 6.12. Weekday and weekend waiting time patterns per cluster.

6.4.4. *Cluster Trip Speed*

Using the waypoint distance as explained in Chapter 5, the variability of trip speed based on user clusters in the morning peak time of August are presented in Figure 6.13. Each cluster has different trip speed, and Table 6.5 lists their average for the day of the week.



(a) *(b)*

Figure 6.13. Weekday and weekend trip speed patterns per cluster.

As expected, trip speed on weekdays is higher than on weekends. The fastest day for all clusters is on Friday. Cluster C are faster than cluster B, and cluster B are faster than cluster A. Cluster C-by-totals are the fastest riders, 14.86 km/hr, while cluster A-by-totals riders are the slowest, 10.80 km/hr. Furthermore, the very slow speed on the left side of distribution mostly belongs to cluster A shown by the red lines, where is it expected that riders have done lots of sightseeing between origin and destination rather than travelling the shortest route (which is used for speed estimation).

Table 6.5. The average of daily trips speed per user cluster.

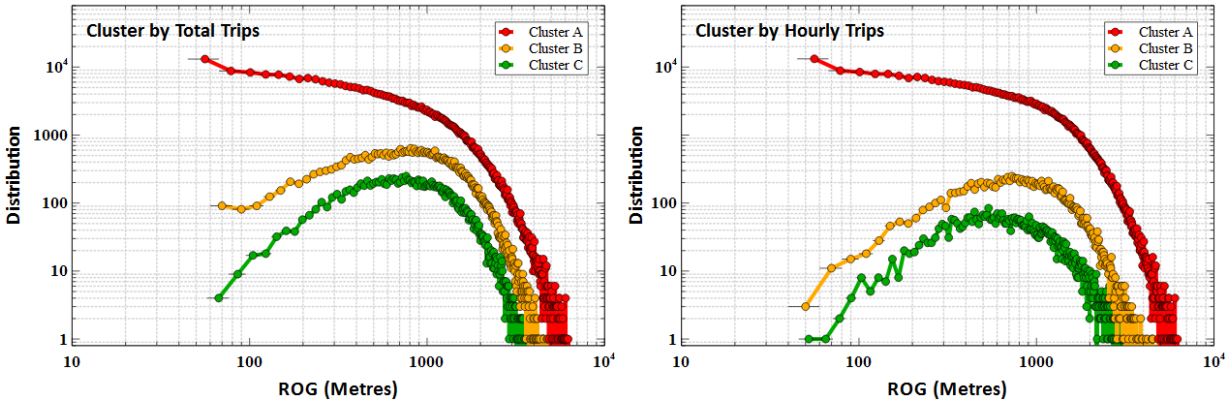
No	Day	Average Speed (km/hr)					
		Cluster by Total Trips			Cluster by Hourly Trips		
		A	B	C	A	B	C
1	Mon	12.83	14.03	14.68	13.40	14.60	14.62
2	Tue	13.07	14.17	14.85	13.58	14.73	14.82
3	Wed	13.23	14.08	14.83	13.60	14.69	14.80
4	Thu	13.28	14.23	14.89	13.72	14.74	14.88
5	**Fri**	**13.41**	**14.31**	**15.04**	**13.80**	**14.88**	**15.03**
6	Sat	10.92	13.50	14.21	11.78	14.21	14.12
7	Sun	10.68	13.74	14.51	11.60	14.64	14.12
Avg Weekday		**13.17**	**14.16**	**14.86**	**13.62**	**14.73**	**14.83**
Avg Weekend		**10.80**	**13.62**	**14.36**	**11.69**	**14.42**	**14.12**

One study conducted by Jensen et al. [100] in Lyon, France, got a precise distance using a counter installed on the bicycle. Then, by using those real distances and duration, they got the average speed on early weekday mornings of 14.5 km/hr. Using waypoint distance in this study, cluster B and C give a similar speed result on weekdays, where trips are mostly commuting.

6.4.5. Cluster RoG

Radius of gyration (RoG) calculations of each cluster show the distinctive skewness as depicted in Figures 6.14. If mobility data captures a reasonable number of trips for users, then one would expect the RoG curve to show a peak in spatial extent at a characteristic distance related to common trip length [46]. The RoG curves for clusters A and D do not show this characteristic, rather the RoG shows a similar shape to the plot of trip distance. This suggests

that for many users in these clusters, there is insufficient data to clearly identify that user's mobility patterns, and poor prediction accuracy might be expected.



(a) *(b)*

Figure 6.14. ROG patterns of the user clusters in log-log scale.

6.4.6. Cluster Motifs

The common spatial trace pattern of each cluster can be seen in the percentage of motifs as shown in Figure 6.15 for six top motifs. Here, the percentages are computed per cluster for motifs which appear in that cluster.

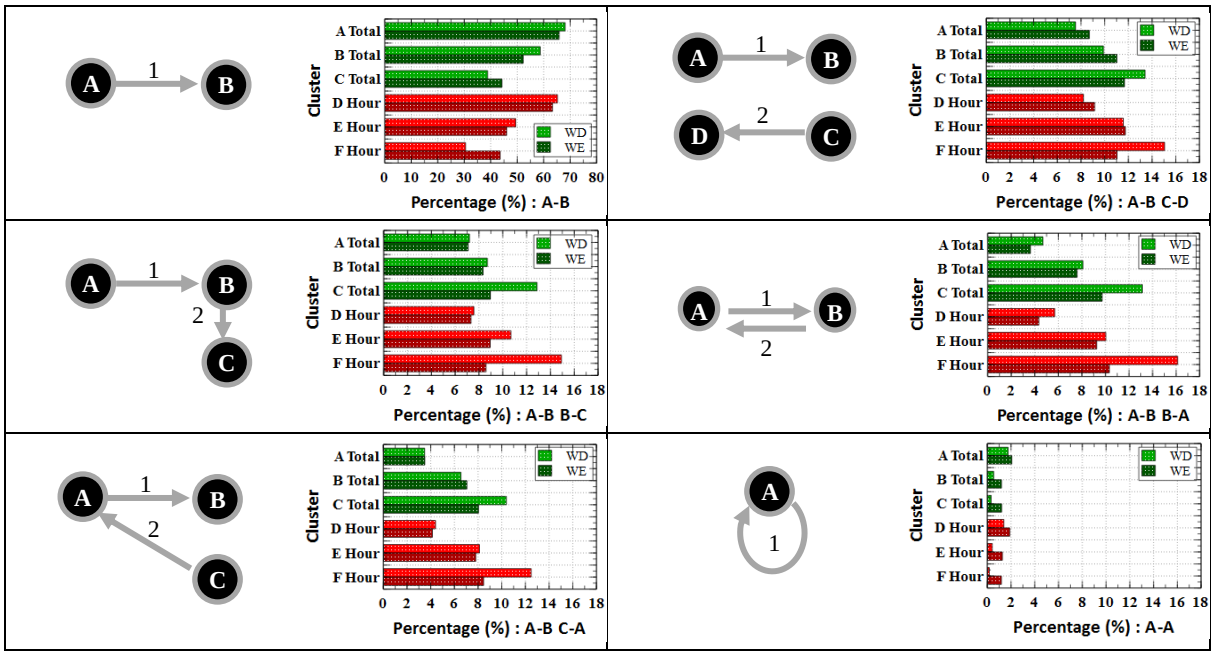


Figure 6.15. Cluster daily spatial-mobility-motifs.

The fact that the four two-trip motifs are all above 10% during weekdays for clusters C and F suggests that daily patterns are complex, and perhaps less predictable. One might expect

a simple **home → work → home** social pattern to usually correspond to a **A→B→A** motif for BSS usage. This appears not to be the case, and this may affect prediction accuracy. Also, the motif diagrams show that clusters A and D have similar patterns, and clusters C and F have similar patterns, as expected, since their membership is similar (A/D low use, C/F high use). More surprisingly, B and E are similar to each other and different from the others. Hence only three labels will be used – one for A/D, one for B/E and one for C/F.

6.4.7. Cluster Label

The previous spatiotemporal analyses of each cluster show that there are distinctive characteristics mainly between clusters A/D and C/F. Clusters C/F show strong commuting patterns where they are more active on weekday with similar high peak time both in the morning and the evening, and less active on weekends. They also reflect the most frequent users and relatively stable toward season. In addition, they also have waiting times on weekdays which are close to the office hours, ride faster than others, and show more commuting motifs. Therefore, clusters C/F will be labelled as ***commuters***. Clusters B/E show quite similar behaviour to clusters C/F, but they are less frequent than cluster C and will be labelled as ***regular users***. Conversely, clusters A/D show seasonal and sightseeing traits which are active on weekend and weekday afternoons, the slowest riders, and highly affected by season. Therefore, they are labelled as ***casual users***. These labels, ***commuters, regular,*** and ***casual users,*** will be used to the rest of the analyses in this chapter. To differentiate clusters either from total trips or hourly trips, their name will be written as cluster-by-total or cluster-by-hour, for example commuters-by-total or commuters-by-hour.

On the other hand, in term of cluster labels, Vogel et al. [74] proposed four clusters focussing only on annual users which are ***user of heart, assiduous users, multimodal users,*** and ***sporadic users***. Similarly, O'brien et al. [75] also proposed four clusters which are ***commuters, utility users, leisure users,*** and ***tourist users***. However, none of them conducted further analysis to understand how predictable each cluster is.

6.5. Entropy and Predictability of Users by Cluster

Examining entropy and predictability by user cluster will show whether the user clustering approach can give a substantial difference to how prediction might be done. Entropy can also be used to infer the significance of the Markov Chain transition probabilities, i.e. whether the next station is highly predicted from the current station. As stated earlier, predictability can be

used as a theoretical upper bound of the prediction that could be possibly achieved using a suitable prediction algorithm [46].

6.5.1. Entropy

The entropy computation needs a sequence of visited places. Here, the sequence of visited stations per user in the learning set is used without distinguishing pickup and return activities *Random, Shannon, Conditional* and *Real Entropy* are computed based on equations in section 6.1.1. Hence, each user will have four metrics of entropy as displayed in Figure 6.16.

Casual-by-total and casual-by-hour users, Figure 6.16.a&b, show jagged histograms for all types of entropy which make them hard to analyse. The large proportion of very low entropy values corresponds to a small number of trips and prediction accuracy for these users will be unlikely to be high. [46]. This low entropy spike was previously shown to also be present for registered users in the preliminary entropy analysis subsection 6.1.2. Some registered users with low trips behave more like unregistered users. In the new clusters, these anomalous users are correctly clustered in the casual user clusters.

On the other hand, the entropy distribution of regular users and commuters are smoother, showing normal distribution form, Figure 6.16.c-f. Entropy of commuter-by-total, commuters-by-hour and regular-by-hour clusters satisfy the basic entropy ordering rule: $\mathbf{S^{Rand}} \geq \mathbf{S^{Shan}} \geq \mathbf{S^{Cond}} \geq \mathbf{S^{Real}}$. This also suggests that hourly-based clustering is better than total-based clustering from the entropy perspective, in terms of identifying different groups of potentially predictable users. Note that estimation of different types of entropy and the above inequality becomes exact only for infinitely long sequences where that all location and transition probabilities can be accurately calculated [46].

To interpret what useful insights are provided by entropy, one cluster is chosen: the entropy of commuters-by-hour, Figure 6.16.f. Here, the means of $\mathbf{S^{Rand}}$, $\mathbf{S^{Shan}}$, $\mathbf{S^{Cond}}$ and $\mathbf{S^{Real}}$ are 4.5, 3, 2 and 1.5 consecutively. Since the $\mathbf{S^{Rand}}$ mean is 4.5, this indicates that the next bike station for a user could randomly be found in any of $2^{S^{rand}} \approx 2^{4.5} \approx 23$ stations. This high random possibility is a result of considering only the distinct visited stations. On the other hand, 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. So $\mathbf{S^{Shan}}$ gives fewer high likelihood next station options than S^{Rand}. Similarly, if the sequence order of station visitation is taken into account, then the conditional entropy greatly reduces to $\mathbf{S^{Cond}} \approx 2^2 \approx 4$ stations. Finally, by considering the whole history, real entropy can give the smallest next place possibility which is

$S^{Real} \approx 2^{1.5} \approx 2.8 \approx 3$ stations. Since real entropy is close to conditional entropy, this suggests that entropy is strongly determined by location history, with most information in just the one last visited station. So Markov transition probabilities can be used for prediction.

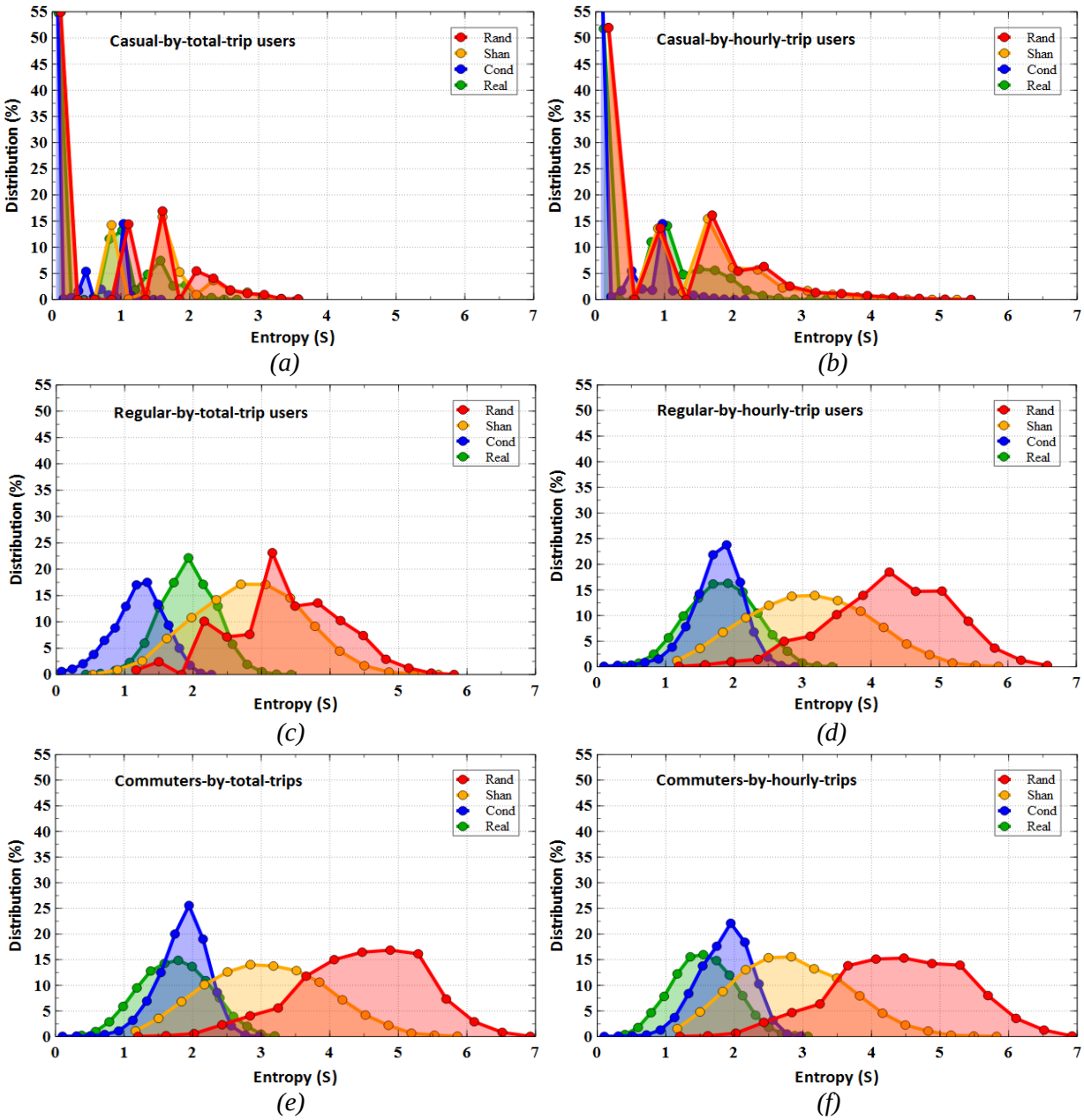


Figure 6.16. Random, Shannon, Conditional and Real Entropy of each group of users.

6.5.2. Predictability

Predictability is the inversion of entropy (which can be thought of as unpredictability). Figures 6.17 show the predictability distribution as the inverse of the entropy and Table 6.6 presents their peak value. Focusing on real predictability, Π^{Real}, commuters have the highest

values which are around 0.78 for commuters-by-total, and 0.80 for commuters-by-hour. Regular users only have 0.67 for regular-by-total and 0.75 for regular-by-hour, while for casual users, it is jagged and hard to analyse.

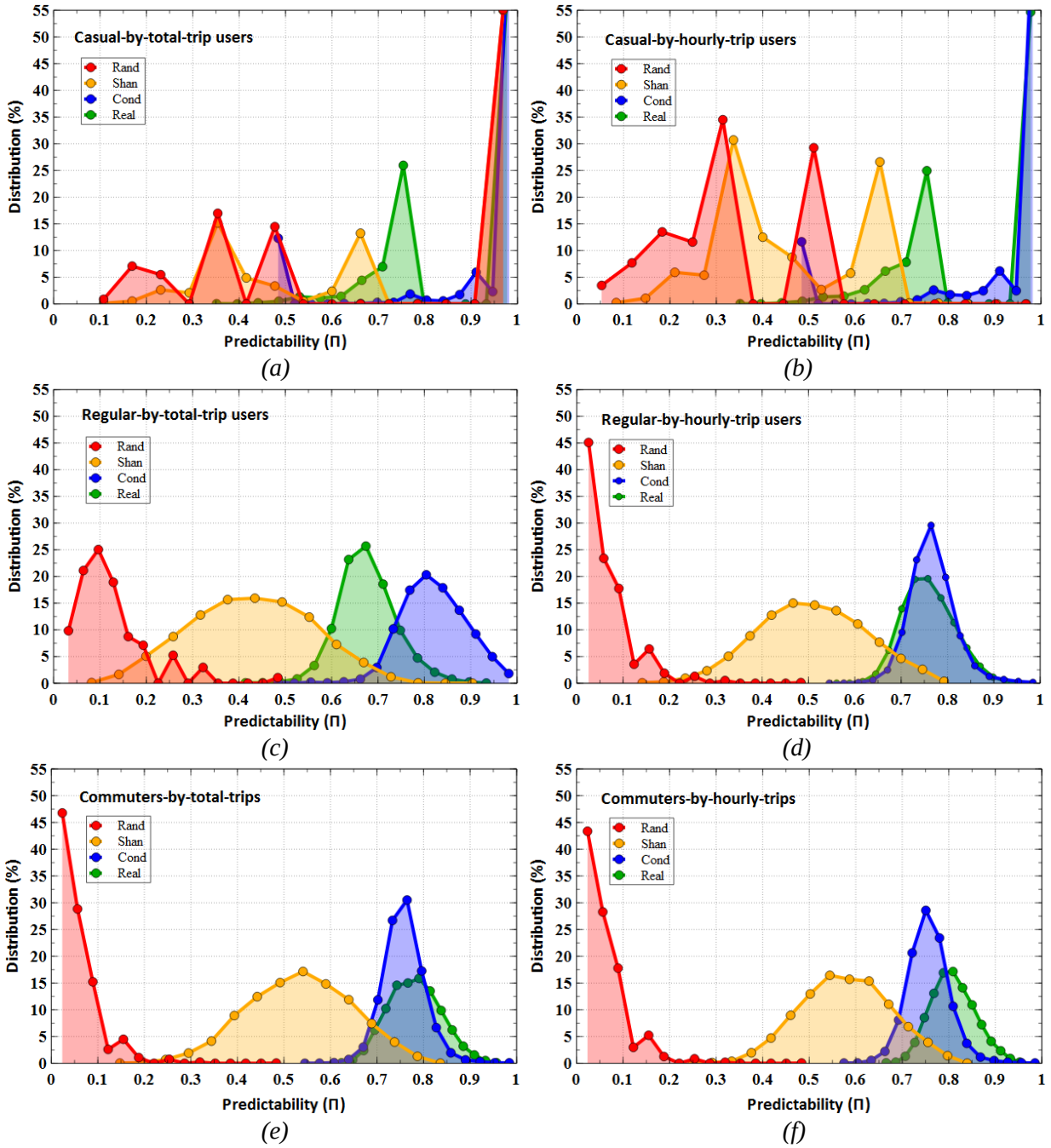


Figure 6.17. Random, Shannon, Conditional and Real predictability of each group of users

All these predictability values indicate there is a possibility that, respectively, around 78% and 80% of commuters-by-total' and commuters-by-hour' next station whereabouts could be predicted using a good prediction algorithm, while the remaining 20% and 22% of cases are

hard to predict. In this case, predictability provides a theoretical upper bound of prediction algorithm performance [46]. More specifically, for actual prediction accuracy, this is a target that could possibly be achieved by a good algorithm [25].

Table 6.6. Peak predictability of commuters and regular users.

Cluster	Peak Predictability		
	Shannon	**Conditional**	**Real**
Regular-by-total	0.43	0.80	0.67
Regular-by-hour	0.51	0.76	0.75
Commuters-by-total	0.54	0.76	0.78
Commuters-by-hour	0.54	0.75	0.80

Other studies have investigated the fundamental regularity of human mobility using different mobility modalities, but this is the first study to investigate the predictability of individual BSS users. Therefore, this work adds to previous studies based on different mobility modalities.

6.5.3. Markovian traits

The real predictability is close to the conditional predictability, $\mathbf{\Pi^{Real} \sim \Pi^{Cond}}$, and this strongly suggests most of the information about the likely next location is contained in the current location, with a weak dependence on previous history. The prediction problem can be posed where the actual predictability can be represented by the conditional predictability [46]. Considering only the last station yields almost the same predictability as considering the entire trip history. In this case, a Markov model predictor where states correspond to locations could achieve close to 78% to 80% prediction accuracy, especially for commuter-by-total, commuter-by-hour and regular-by-hour users. On the other hand, casual users will be hard to predict.

The predictability of BSS users can be compared to other predictability studies using these information theory methods but using other mobility modalities. For mobile phone data, Song et al. [24], Lu et al. [25], and Qin et al [44] found 93%, 88%, and 78% of predictability respectively. The high predictability of Song et al. [24] and Lu et al. [25] could be due to mobility tracking using mobile phone considering the nearest cellular base station as a position. Hence, even though an individual moves around near the same base stations, he/she will be considered to be in the same place. Predictability of BSS users in this study is close to the result of Qin et al [44] which is 78%. However, Song et al. [24] and Qin et al [44] did not

continue their work to the prediction to show whether their high predictability results can be achieved in practice. Meanwhile, Lu et al. [25] implemented a Markov Chain model to conduct prediction, and they could achieve an accuracy at their predictability level using the first order Markov model.

6.6. Users Next Place Prediction

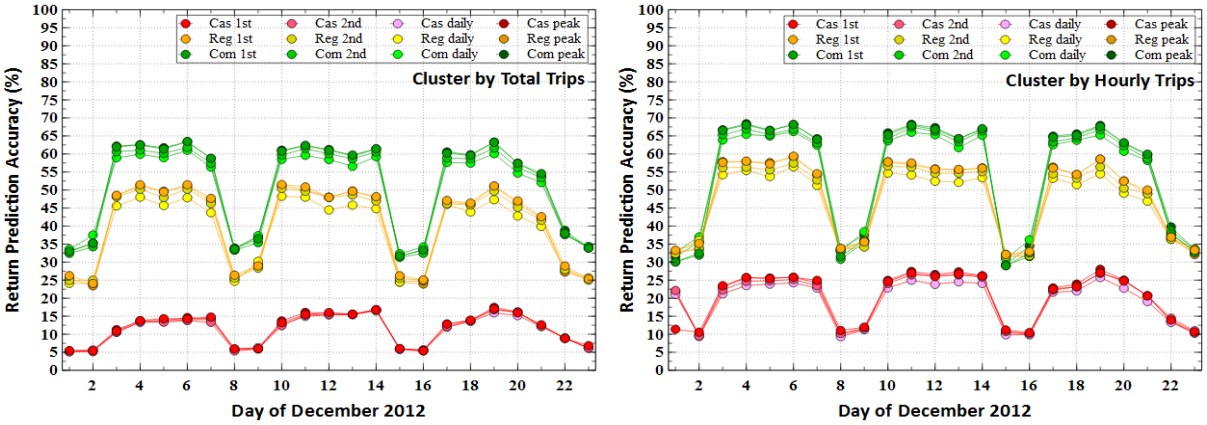
In this section, the Markov Chain based predictor will be constructed to predict the user's next location based on their trip history ensemble with the collective trends of the cluster for trips with unavailable history. Four types of Markov predictor based on their OD matrix selection as proposed in subsection 6.1.5 will be applied to ***pickup-to-return*** as well as ***return-to-pickup*** prediction. First, using the whole trip history as one OD probability matrix, the first order Markov Model will be used. Second, it will be extended to the second order model to see whether the higher order can help to increase the accuracy. Third and fourth, the splitting OD matrix approaches based on *day-of-the-week* and *peak-times-of-the-day* will be investigated as a possibility to improve the accuracy.

Separate transition matrix probabilities are calculated for each user based on all their trips in the training period. Consider the first order predictor based on all trips, for pickup-to-return prediction, the transition probabilities for A → B are calculated on the number of trips that start at A and end at B for that user throughout the training set and the highest probability will be used. Similarly, for return-to-pickup prediction, the transition probabilities for A → B are calculated on the number of trips where the previous trip ended at A and the next trip starts at B for that user throughout the training set and the highest probability will be used. Then for prediction, each pickup is predicted using the most likely transition from the previous return location in the return-to-pickup matrix, and each return is predicted by the most likely transition from the pickup-to-return matrix.

6.6.1. Pickup to Return Prediction Accuracy

The return prediction is first conducted for all users without cluster on a daily basis within 23 days of the testing period, and using formula 6.6 the accuracy is calculated which is only 43% on weekday and 18% on weekend. Then, the return prediction is conducted for each user cluster using the proposed scenarios as in subsection 6.1.5. The results show that this cluster-based prediction accuracy is higher than prediction without clustering. The daily accuracy of each cluster for individual-based method is shown in Figure 6.18, while the cluster average

accuracy for whole period and splitting over weekday and weekend is given in Table 6.7. There are four general trends of the results which are prediction accuracy on weekdays is higher than on weekends, the commuters are more accurately predicted than regular and casual users, cluster by-hourly-trips give better accuracy than cluster by-total-trips, and the ensemble of first order Markov Model with *peak-times-of-the-day* matrix give the highest accuracy.



(a) *(b)*

Figure 6.18. Daily pickup-to-return prediction accuracy.

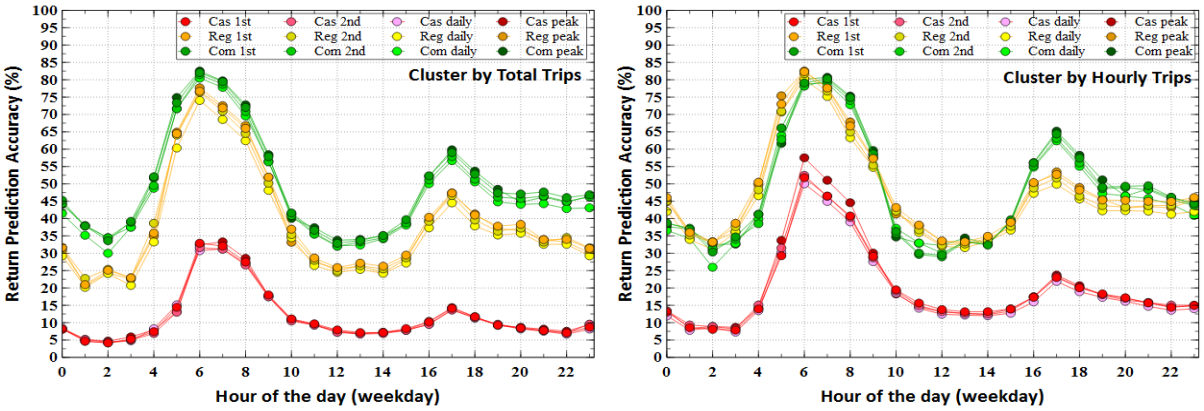
Table 6.7. The average of pickup-to-return prediction accuracy for each method.

Method	Total-based clusters						Hourly-based clusters					
	Weekday			Weekend			Weekday			Weekend		
	Cas.	Reg.	Com.	Cas.	Reg.	Com.	Cas.	Reg.	Com.	Cas.	Reg.	Com.
Individual-based prediction												
1st Order Markov full matrix	51.3	56.1	64.1	34.7	35.6	39.9	53.2	60.2	68.9	35.4	40.1	38.1
2nd Order Markov full matrix	49.5	54.5	62.7	34.7	34.4	38.9	51.4	58.5	67.9	34.5	38.8	37.7
1st Order Markov daily matrix	49.6	52.9	61.4	34.0	34.9	40.8	49.8	56.8	66.9	34.5	40.1	40.9
1st Order Markov peak matrix	51.5	55.9	64.2	35.1	34.6	40.2	52.9	60.2	69.1	34.6	39.4	40.0
Individual + collective trends prediction												
1st Order Markov full matrix	14.5	48.7	60.6	6.2	26.4	34.4	24.9	56.1	65.7	11.4	34.1	32.8
2nd Order Markov full matrix	14.0	47.7	59.3	6.0	25.9	33.8	24.5	54.6	64.7	12.5	33.3	32.5
1st Order Markov daily matrix	14.0	45.5	58.0	6.2	25.7	35.1	23.1	52.9	63.7	11.8	34.0	35.1
1st Order Markov peak matrix	14.5	48.4	60.6	6.2	25.5	34.6	25.1	56.0	65.8	12.2	33.5	34.4

However, the results suggest that the prediction accuracy on a daily basis is still lower than the highest predictability level that was calculated in section 6.4.2 which is around 80% for commuters-by-hour. In daily basis, the maximum accuracy that can be achieved here by this cluster is around 70% using the ensemble of first order Markov Model with *peak-times-of-the-*

day matrix. This can be seen in days 4, 6, 11, 14, and 19 as shown in Figure 6.18.b. This suggests that *peak-times-of-the-day* matrix can slightly improve prediction, while it is not the case for the second order Markov Model and the *day-of-the-week* OD matrix. Furthermore, implementing the collective trends to predict trip without history cannot improve the accuracy significantly.

It cannot be expected that accuracy will be stable across the hours of the day, since commuters, as the most predictable users, do not spread their trips homogenously across every hour during the day. To see which hours of the day significantly contribute to shape the daily accuracy dynamics, Figures 6.19 show the average prediction accuracy per user cluster in an hourly basis on weekdays.



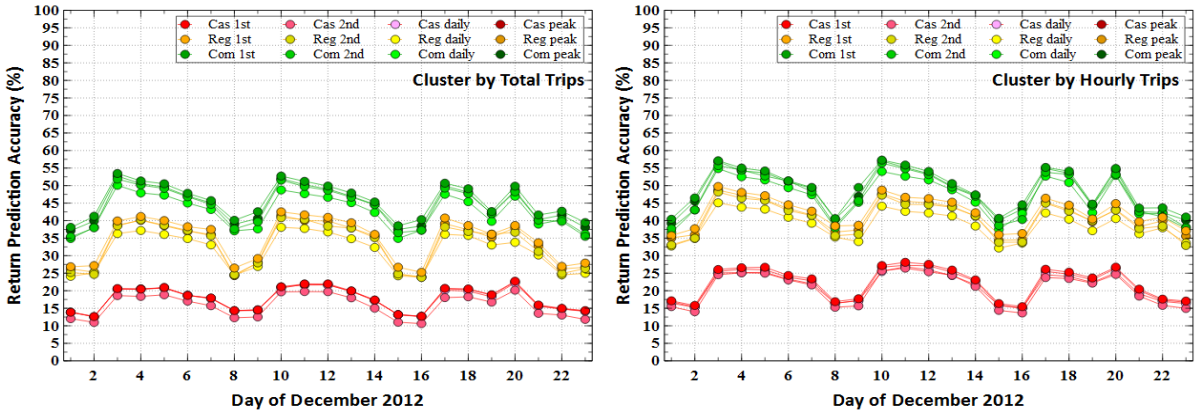
(a) *(b)*

Figure 6.19. Hourly pickup-to-return prediction accuracy.

It can be seen that the peak periods from 5 am to 9 am when commuters are dominant in the system have the highest accuracy, reaching 78%-80%, similar to the theoretical predictability. This morning peak period could contribute most to keep the daily accuracy high because other hours are less predictable. The least predictable time is at midday and early morning.

6.6.2. Return to Pickup Prediction Accuracy

Similar approaches of pickup-to-return prediction in the previous subsection are implemented for return-to-pickup prediction in this section to understand whether it also has similar trends. It can be seen from Figure 6.20, the highest prediction accuracy every week are always on Mondays, day 3, 10 and 17, while other weekdays are lower than Monday with gradually decreasing patterns over the week where **Monday > Tuesday > Wednesday > Thursday > Friday**. However, all of those are lower than return prediction accuracy.



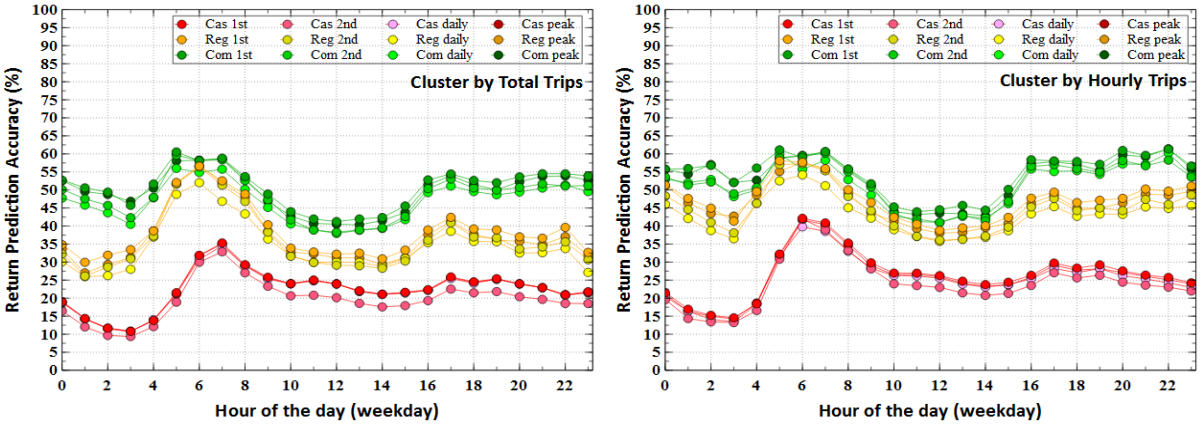
(a) *(b)*

Figure 6.20. Daily return-to-pickup prediction accuracy.

Table 6.8. The average of return-to-pickup prediction accuracy for each method.

Method	Total-based clusters						Hourly-based clusters					
	Weekday			Weekend			Weekday			Weekend		
	Cas.	Reg.	Com.	Cas.	Reg.	Com.	Cas.	Reg.	Com.	Cas.	Reg.	Com.
Individual-based prediction												
1st Order Markov full matrix	37.0	45.2	51.6	30.6	35.0	46.0	40.9	48.8	54.7	33.4	43.4	49.3
2nd Order Markov full matrix	36.4	43.6	50.4	30.2	33.1	43.3	39.5	47.3	53.8	31.9	40.9	46.5
1st Order Markov daily matrix	35.9	40.3	48.2	30.0	31.5	41.8	37.5	44.3	52.1	30.8	39.1	45.5
1st Order Markov peak matrix	36.3	43.7	50.6	29.9	32.8	44.4	39.5	47.5	54.1	31.6	41.6	48.0
Individual + collective trends prediction												
1st Order Markov full matrix	20.0	39.0	48.6	13.8	27.1	40.4	25.4	45.1	52.3	16.7	37.6	43.3
2nd Order Markov full matrix	17.8	37.5	47.4	11.8	25.3	37.5	23.7	43.6	51.3	14.9	34.9	40.4
1st Order Markov daily matrix	19.7	35.1	45.5	13.8	24.8	37.0	24.0	41.2	49.9	16.2	34.3	40.3
1st Order Markov peak matrix	19.8	37.8	47.8	13.7	25.7	39.0	24.9	44.0	51.8	16.3	36.2	42.2

Among different prediction scenarios, the results show that the first order Markov Model gives the highest accuracy followed by the second order Markov Model, then the peak-based and daily-based OD matrix. This suggests that temporal aspects of the OD matrix which are day and peak time cannot improve the accuracy of return-to-pickup prediction. The accuracy of hourly return-to-pickup prediction follows the daily tendencies which are lower than pickup-to-return prediction as shown in Figure 6.21. The highest accuracy is in the morning peak time, 5 am to 9 am. Again, clusters by-hourly-trip give better accuracy than the others.



(a) *(b)*

Figure 6.21. Hourly return-to-pickup prediction accuracy.

The results of both pickup and return prediction above, where the pickup prediction is mostly below than 60% while the return prediction can reach 80% in weekday morning, suggest that the correlation between ***return-waiting-pickup*** is less than ***pickup-ride-return***. This fact suggests that once people pickup bikes they are likely more predictable with their intended destination, compared to the next trip that they will make. Overall, the results show that user clustering by-hourly-trips can give better prediction accuracy than clustering by-total-trips.

Recall, that if there is no entry in the transition matrix, i.e. a user visits a new station, then the predictor uses a collective population-based matrix. This matrix can be one matrix for all users over all times. It can be specific to each cluster for all times, or it can be specific to a cluster and the time of day (morning-peak, afternoon-peak, other). How the collective trends of clusters can actually help the prediction can be seen from Figure 6.22.

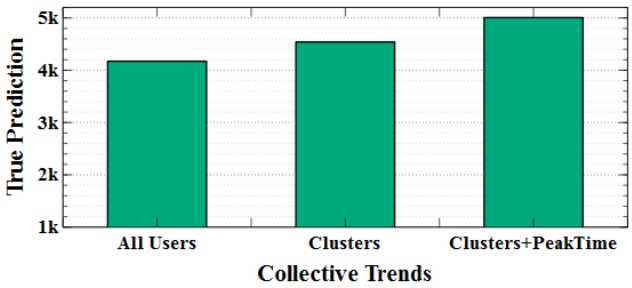


Figure 6.22. The True prediction by population (collective trends).

Figure 6.22 shows that there is increase around 8.8% of *True* prediction number for trips without history if using the clusters collective trends instead of using all population trends (without clustering). Then, if the clusters collective trends are divided by peak time, correct

predictions increase around 20%. This suggests that collective trends of clusters in peak time can be used to improve the *True* prediction number, even though this is only 6% of all trips without history.

6.7. Practical Application

If a user's next location can be reasonably accurately predicted, then personalized notifications can be sent to that user relevant to their expected trip. If trips are unpredictable, then sending notifications is more likely to be useless and annoying.

So the first step is to identify which trips are most predictable. These trips can be observed from the high predictable users which are commuters. It is recommended that only these users are targeted for personalized notifications. Furthermore, times of trips affects predictability, and it is also recommended to send notifications only during peak times when the accuracy is higher than other time slots.

Next is the nature of notifications. If stations will be shut down or likely to be full or empty at particular times ahead, or if the shortest routes to the likely destination are congested, an advance notification can be sent automatically or proactively to these highly predictable users as they start their trip. The notification can include the possible alternative nearby stations, routes, or time of travel. This is possible because the common visited stations, shortest routes and visiting times of those users are mostly known from their regular history. This *user-based notification system* will make the system more intelligent, and it can complement the journey advisor systems proposed by Yoon et al. [108] and Yang and Zhang [115] which use *station-usage analysis* as the basis of their advisory system.

If a user has several higher-probability next destinations which are close to each other (within 300 m, as indicated in the station neighbourhood discussion in Chapter 5), then the notification could suggest an alternate station that the user is known to also use, and may be almost as convenient. Even if the predicted destination is not full, the system might suggest a preferred nearby alternate destination to assist with user-based station rebalancing, perhaps offering an incentive to use the alternate destination.

6.8. Next Place Prediction Significance Summary

This chapter has first investigated how users are properly clustered using their temporal features and labelled using their spatiotemporal characteristics. Then, their randomness shown

by *entropy* and the limit of their regularity shown by *predictability* are measured to get the upper bound of predictability that is achievable in prediction.

Results suggest that the proposed temporal clustering technique using hourly trip numbers that reflect the frequency and regularity of mobility per hour on a daily basis can properly capture the homogeneous users in terms of spatiotemporal characteristics and predictability. Two group of users show obviously different behaviours, while a third group shows behavior which combines aspects of those two. Comparing to other predictability studies in information theory fields, using mobile phone data, Song et al. [24], Lu et al. [25], and Qin et al [44] found predictability of 93%, 88%, and 78%, respectively. The upper bound of predictability for commuters in this study which is 80% is close to the result of Qin et al [44]. However, Song et al. [24] and Qin et al [44] did not continue their work to actual prediction to show whether their high predictability results can be achieved by a predictor. Meanwhile, Lu et al. [25] implemented a Markov Chain (MC) model to conduct prediction, and they achieved an accuracy similar to their predictability level using a first order MC model. In this study, prediction using the first order Markov Model at different times of day can achieve prediction accuracy similar to the predictability level, especially for commuters during the peak times on weekdays. This proposed technique uses an ensemble which combines the collective trends of the user's cluster to predict trips without history for that user, and this improves accuracy compared to just using individual history.

Highly predictable users can be provided with personal notifications that can complement the journey advisor systems proposed by Yoon et al. [108] and Yang and Zhang [115]. This proposed personal notification may assist with *user-based station rebalancing*. For example, if a highly predictable user has several higher-probability next destinations which are close to each other (within 300 m, as indicated in the station neighbourhood discussion in Chapter 5), then an alternate station that the user is known to also use, and may be almost as convenient, could be suggested even if the predicted destination is not full. Incentives might be provided to encourage this user-based rebalancing.

CHAPTER 7
SYSTEM-WIDE PREDICTION

The previous chapter dealt with issues about predicting the behaviour of BSS users, and their predictability, and how this might be used to enhance their experience. This chapter deals with the predictability of aggregate system use, and is about issues that affect the BSS operator. Estimating system-wide usage at particular times on particular days is useful for BSS operators in order to ensure, as far as possible, that there are sufficient bicycles available to service that demand. Good demand prediction will enable operators to better plan rebalancing and maintenance activities.

This chapter investigates a prediction method for system level usage based on the cyclostationary traits that are strongly evident in hourly BSS patterns over the week [63]. The assumption here is that the hourly usage consists of a consistent (i.e. statistically stationary) underlying weekly pattern (i.e. cycle) plus a disturbance to that pattern caused by certain factors. This can be extended to an underlying weekly pattern that itself changes slowly over the seasons, so that the normal or average weekly pattern in winter is different to that in summer. Rather than predict the absolute values of hourly usage, this new predictor estimates the current disturbance from the underlying seasonal weekly pattern. If the estimation is positive, it means that the current state of BSS is busier than the historical reference, and if negative, usage is lower than average. Although relatively common in other time series forecasting studies, no previously published studies of BSS usage have used this type of approach. This technique is commonly used to model time series such as the daily temperature within a yearly cycle. For example, the daily maximum temperature in London on the 1st of June in previous years is a reasonable estimation of the maximum temperature for the same date in this year. In this case, there are 365 interleaved stationary processes where each of them takes a new value once per year that is usually similar to the previous year. Similarly, if BSS prediction uses hourly bins within a weekly cycle (24/7), there will be 168 interleaved stationary processes that must be taken into account.

This chapter will analyse the prediction of BSS usage at three levels (system-wide, cluster-based, individual). By measuring the prediction performance at different levels, it will be possible to analyse if there is a spatial correlation in prediction performance so that areas that are better predicted can be identified. The broadest level for prediction is the aggregate of all

bike stations in the BSS and can be treated as one entity, called the *system-wide* level. At this level, the only prediction variable is trips that are counted from either the number of pickups or number of returns, because at this level each pickup results in one return (although not necessarily in the same hour). In other words, trips in this study are defined as the number of bikes that are rented in the system within each one hour period. At the middle level, the system can be divided into sub-systems called *clusters* that consist of a group stations with similar features in a region. At this cluster level, the prediction can be in terms of three different parameters that represent the usage of the cluster. These metrics are *pickup,* which is bikes out from the cluster, *return,* which is bikes into the cluster, and *balance* between return and pickup (pickup minus return). The finest level of prediction is individual bike *stations* which have the same prediction variables as *clusters*. At this station level, the spatial correlation between stations can be investigated. Most of the existing BSS studies undertake prediction at system-level [10, 11, 21, 49, 70, 81, 90], only a few predict at a cluster level [65] and at the station level [2, 87], and none at all investigate all three levels in one study. It would be expected that the system usage on an hourly basis will become more chaotic or unpredictable as prediction moves from system-wide to cluster to station level.

While some existing BSS prediction studies use signal processing and data mining approaches [11, 63, 75, 105], this study will analyse machine learning techniques as an alternative to those approaches. This chapter will investigate how the proposed prediction scenario can be implemented using machine learning predictors at each level, how the underlying stationary patterns will be estimated, which external factors should be taken into account for prediction, how to properly measure prediction performance, and how their performance compares to existing similar studies [47]. The practical implications of this prediction approach that bring to the BSS operation will also be analysed. Finally, the work in this chapter will be used to answer RQ3 and a part of RQ4.

This chapter is organized as follows. It starts with a methodology section, followed by analyses of prediction results of system-wide, cluster and station levels. It concludes with an analysis of the practical significance of the results.

7.1. Methodology

This methodology section begins with an explanation of the seasonal-based prediction scenario. This is followed consecutively by dataset splitting and pre-processing, machine

learning predictor and feature selection, feature importance, a sliding window technique for underlying pattern estimation, and performance analysis metrics.

7.1.1. Deviation-based Prediction Scenario

The deviation-based prediction used in this study is defined as the prediction of current state based on a deviation from the recent historical reference at the same time bin. Here, time dependant features (individual pickup and return times) will be organised chronologically into a series of time bins. To give a good trade-off between the resolution of details and fluctuations, following Borgnat et al. [63], the prediction bin is every 1 hour. In the rest of this chapter, each hour will be referred to by its starting time, so Tuesday 9 am means the hour from Tuesday 09:00:00 to 09:59:59. Hence, there will be 24 stationary processes on a daily basis and 168 on a weekly basis. This needs a clean dataset with one number of trips each hour. At the system-wide level, the only prediction target is number of trips in this hour, while in cluster and station levels, prediction targets are number of pickups, returns, or balance between them. This prediction scenario is illustrated in Figure 7.1, for the case where the historical pattern is the previous week's trips.

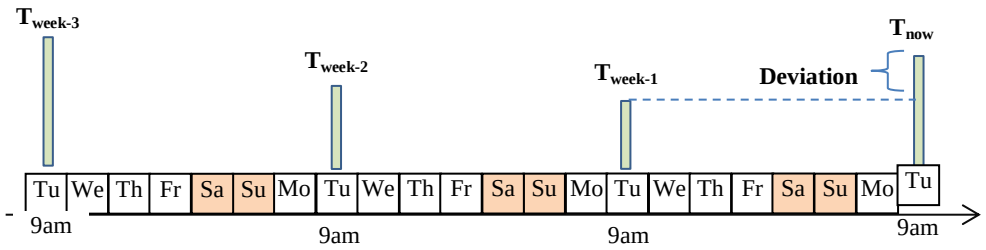


Figure 7.1. The deviation-based prediction scenario.

An important first step in the prediction is to investigate how the historical average of usage is derived, and this depends on how the periodic pattern is considered. One could consider the 24*7=168 cyclostationary bin approach described earlier, where historical averages are based on previous usage at that same time in previous weeks. Another choice would be to assume that there is one daily pattern for weekdays, and one for weekends, and so the reference would be the same time at the previous day. The last choice is that the pattern depends on very local temporal variations, and so the historical reference is the previous hour.

Specifically, consider estimation of the next hour's usage. Figure 7.1 shows that the number of current trips on Tuesday at 9 am can be predicted as the number of trips at the same day and time a week ago plus or minus a deviation. In addition to the previous week, there are

two other temporal features that could be possible as reference points, viz., the previous hour and the previous day. If the previous hour is a reference point, the number of trips on Tuesday at 9 am can be calculated as a number of trips at 8 am with some deviations. If the previous day is used, the reference point will be Monday at 9 am. Using this technique, the predictor predicts the deviation, and the current state will be estimated by adding this predicted deviation to the reference points that are already known.

Possible estimates of the historical patterns can use not just one reference point but the average of several historical reference points. Such averages may give a better estimate of the underlying seasonal pattern by cancelling out disturbances. For example, using three previous weeks, the reference point will be the average of those three previous same-day-&-time values. However, using many historical values, such as a whole year, may hide seasonal changes, and also hide underlying trends such as increasingly popularity of the BSS. The best choice of cyclostationary pattern estimator will be explored in this chapter.

Based on the aforementioned explanations, this deviation-based prediction scenario over the cyclostationary pattern of BSS data based on *N* previous weeks can then be formulated as follows:

$$T_{PredNow} = D_{Pred} + \frac{1}{N}\sum_{i=1}^{N} T_{week.i} \tag{7.1}$$

If the predictor considers hourly and daily history, the week term in the formula can be substituted by hour or day. Specifically for daily references, since weekdays and weekend days have different patterns, the daily basis will consider these separately. For example, Monday's prediction will be based on Friday, not on Sunday, Sunday will be based on Saturday, and Saturday will be based on the previous Sunday. For a whole day, prediction consists of a set of 24 estimates of references and deviations for each hour as shown below:

$$T_{PredDay} = \{(T_{Ref\_h0} + D_{Pred\_h0}), (T_{Ref\_h1} + D_{Pred\_h1}), \dots, (T_{Ref_{h23}} + D_{Pred_{h23}})\} \tag{7.2}$$

7.1.2. Dataset Selection and Splitting

Unlike the two previous chapters that required user IDs for individual analysis, the prediction task in this chapter looks at aggregate station usage, so that other BSS datasets without user ID can be used. Here, two BSS datasets from London and Washington DC with trips from August to December 2012 will be used. This investigates whether the proposed

prediction method is generic enough to be applicable to a different dataset without knowing its patterns in advance (Washington DC dataset). In machine learning prediction tasks, it is common practice to split data into at least two sets which are for training and testing. However, in this study data will be divided into three sets. The first set is four months for training (1st August – 30th November), the second set is one week for validation (1st – 7th December), and the third set is two weeks for testing (8th – 21st December). Predictors will be tested first with the validation set to find the best features and hyperparameter settings. Then the best predictor will be used with the test set to judge the performance of the predictor. The dataset has been cleaned to exclude trips with unrealistic durations (< 1 minute or > 24 hours). For London, the remaining data are 2,805,718 records with 566,456 users and 573 stations. For Washington DC, the remaining data are 891,297 records and 194 stations with no individual user information. The London BSS is significantly larger than Washington DC BSS.

Another complementary data source used in this study is an hourly historic record of weather[7] in Central London and Washington DC. This weather log consists of temperature, humidity, wind speed and rainfall. It was already seen earlier in Chapter 6 that casual users in London are significantly reduced on rainy days and on colder days. This dataset will be used to investigate whether these weather features can improve the prediction performance.

7.1.3. Machine Learning Predictors

As the target output is a numerical value, the problem can be stated as a regression problem, i.e. applying a best-fit model to a series of numerical values. BSS patterns are not linearly related to prediction features, its dataset is large, and many factors can be taken into account in prediction, so this prediction scenario will be implemented using machine learning techniques. There are many different machine learning algorithms that could be used, each with different hyperparameter spaces to explore. The key research question here is not to decide on which machine learning algorithm is best, but rather to decide which historical average assists prediction most, and which sets of features are most helpful. Also, by using the same algorithms as other researchers, we can more easily compare our prediction results to this other work.

Following the approach from Giot and Cherrier [47] who employed five regression systems to predict the BSS usages in Washington data, this cyclostationary-based prediction

[7] Downloaded from the wunderground website (www.wunderground.com)

scenario also employs all those regressors that have been reviewed in Chapter 2 which are *Gradient Boosting Regressor* (GBR), *Bayesian Ridge Regressor* (BRR), *Support Vector Regressor* (SVR), *AdaBoost Regressor* (ABR), *Random Forest Regressor* (RFR), plus *Decision Tree Regressor* (DTR), to see which regressor gives the best performance for the proposed cyclostationary-based prediction scenario in London as well as Washington data. Later, prediction results from this study will be compared to the previous work [47] for Washington data to see whether the cyclostationary-based prediction is better than existing work without this cyclostationary approach.

Other regressors, such as Artificial Neural Networks or Linear Regression with non-linear features could also be explored, but the above regressors have been chosen because these regressors have a relatively small hyper-parameter space that needs to be searched compared to the very broad range of parameters for techniques like Neural Networks.

7.1.4. Naïve Predictors

The prediction results from all these machine learning regressors will be compared to two naïve approaches as a baseline benchmark. This aims to see whether the complex machine learning techniques make better predictions than the simple naïve historical approaches. Those two naïve approaches are first based on *historical average* (HA) and second based on *deviation average* (DA) from reference points. The historical average approach assumes that deviations from cyclostationary patterns are unpredictable and uses an estimated deviation of zero. Meanwhile, the deviation average approach assumes that changes in usage patterns are seasonal (e.g. reducing usage towards winter), and so the historical trend is slowly increasing or decreasing based on the deviations in the recent past.

Because many different approaches are being explored, it is necessary to devise some clear terminology, hence the following definitions. The length of the historical average (HA) will be first made from the whole learning set which is four months (HA4Month), then for the one last month (HA1Month), and the one last week (HA1Week). This is to see whether longer or shorter historical averages make any difference. Then, for the deviations average (DA), three reference values will be used: one (DA1Ref), the average of two (DA2Ref), and the average of three (DA3Ref) previous deviations. As there are 3 types of references, hour (Hr), day (Dy) and week (Wk), the deviation average prediction will be conducted for each of these references. Accordingly, there are 9 combinations of DA predictors which are DA1RefHr,

DA1RefDy, DA1RefWk, DA2RefHr, DA2RefDy, DA2RefWk, DA3RefHr, DA3RefDy, and DA3RefWk.

7.1.5. Feature Selection and Feature Importance

Extracting, transforming and selecting features are some of the crucial preliminary tasks in machine learning prediction. In this cyclostationary scenario, the day/time of the estimated output and the previous reference points are key features. The current prediction time features are *hour of the day, day of the week,* and *month of the year* for the current prediction. Here, year and public holidays are not considered because the data in the validation and test sets are in the same year and no holidays appear in that data.

Several other potential features are investigated such as the previous state, weather, and percentage of unregistered users. The first are the deviations at one and two hours ago, motivated by the fact that very recent usage figures might indicate whether this particular day is a busy day or not for the BSS, or for this cluster or station. The second is the weather conditions (temperature, humidity, wind speed and rain) to see the role of external factors. The third is the percentage of unregistered users who used the BSS one and two hours ago at each level to see whether their ratio has an impact. The unregistered users' ratio is considered because their usages are more varied than registered users. For the station level prediction, the state of the nearby neighbourhood will be added to see the spatial correlation between stations.

Given three data sets (training, validation, testing), first, the predictor will be trained using the training set. Second, the hyper-parameters will be tuned on the validation set, using a grid search. The detailed results from this grid-search of the hyper-parameter space and the chosen hyper-parameters for each regressor are presented in the appendix. Third, a feature importance test will be applied to rank the features automatically in terms of their impact in shaping the prediction output. For example, for the random forest regressor, the method is to keep track of the reduction in impurity or mean-square error that is attributed to each feature as the data falls through the trees in the forest. The feature importance technique that is used is the *gini importance* or *mean decrease impurity* and is defined as the total decrease in node impurity (weighted by the probability of reaching that node which is approximated by the proportion of samples reaching that node) averaged over all trees of the ensemble [124]. Fourth, the performance will be tested on the unseen test set. By adding the validation set to the training set as a new training set, prediction will be conducted on the test set with the selected best

regressor. In this case, the regressor is selected from the validation results that give the best performance with tuned hyper-parameters as well as with the most significant features.

7.1.6. Sliding Windows Technique

The cyclostationary pattern of BSS actually has a seasonal component which changes from season to season, so that the pattern relevant to the predicted time is the history closest to the current time. Furthermore, there is a tension between a long training set which reduces effect of disturbances, but tends to reduce ability to adapt to the seasonal or popularity trends, and a smaller training set which can react to recent changes but is more sensitive to disturbances. To overcome this issue, a *sliding window technique* (SWT) is proposed for this cyclostationary-based prediction where a fix-length training set time window will move forward behind the test set on a daily basis. In other words, as each day passes, the last predicted day will become a new member of the training set to predict the next day, while the first day of training set will be dropped. This can be seen in Figure 7.2.

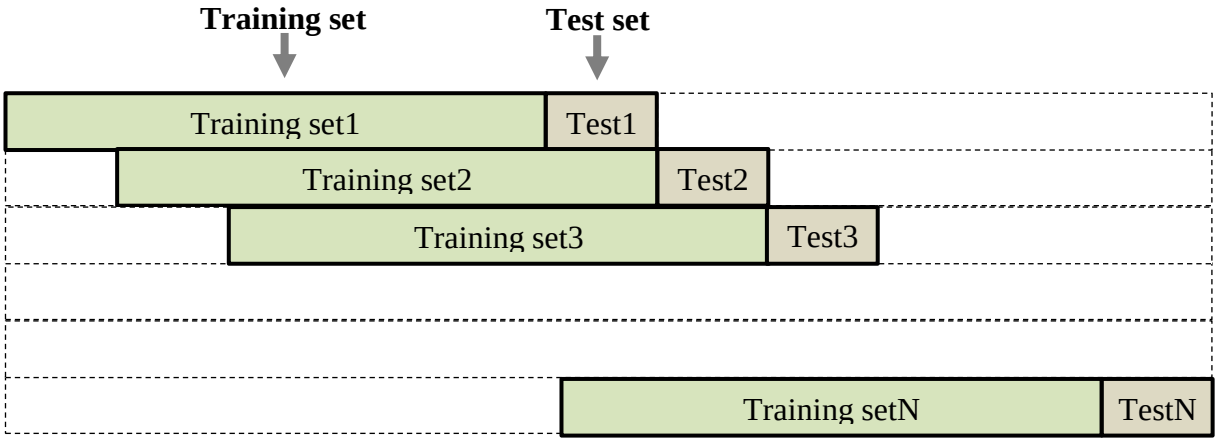


Figure 7.2. The sliding window technique.

This technique means that each prediction uses the most recent usage data available, allows the predictor to track seasonal trends, and long term increased usage trends, and each training set will be a similar length. Consequently, the predictors must be retrained every time they move forward. The best size for the sliding window will be experimentally determined.

7.1.7. Performance Analysis

Once the prediction has been done, performance analysis is needed to interpret how well the prediction fits with reality by comparing the predicted values with the actual ones. The well-known *root mean square error* (RMSE) will be employed as a basic performance metric. The advantage of using RMSE is that it provides an error metric that has the same unit as the prediction output. The RMSE formula is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (7.4), \qquad RRMSE_{(\%)} = 100 * \left(\frac{RMSE}{Trips_{groundTruth}}\right) \qquad (7.5)$$

where n is the number of samples to predict, yi is the ground truth and $\hat{y}_i$ is the prediction of sample i, and $Trips_{groundTruth}$ is the average of y_i.

While most BSS prediction studies only use the RMSE prediction accuracy metric, this study will also use a transformation of RMSE to a relative metric (RRMSE(%)), which is calculated by dividing RMSE by the ground truth of average predicted trips. The error ratio of prediction relative to the ground truth can be easier to interpret. For example an RMSE of 10 indicates a good estimate if the average correct value is 1000 and a poor estimate if the average correct value is 2. Here, RRMSE figures of 1% and 500% are more informative. Another alternative "relative" error measure is MAPE, which averages the absolute value of each error relative to the individual correct value. However, at the level of individual prediction, many actual values are zero, and MAPE is undefined (or infinite). The reason for using a relative measure is to be able to compare data with different scales (in this case London and Washington DC). RRMSE also preserves the same ordering of accuracy as RMSE for different predictors for one city, which MAPE does not necessarily do.

In the cluster and station levels, in addition to RMSE and RRMSE, the RRMSE *range* is used which is the categorization of relative error based on certain ranges of scores. This gives another informative measure for practical applications which indicates how often the predictions are "good" or "bad". Following the daily sliding windows method, each sliding window will yield a daily RMSE from 24 hourly bins. Therefore, the performance over the whole day will be computed as one average of daily RMSE as follows:

$$RMSE_{avg} = \frac{1}{24}\sum_{i=1}^{24} RMSE_i \qquad (7.6)$$

7.2. System-Wide Prediction Implementation

For the system-wide scale, three approaches to prediction are compared: naïve prediction based on historical average of the same day-of-the-week and same hour-of-the-day with three different lengths of average (HA4Month, HA1Month, and HA1Week), naïve prediction based on past deviations of one, average of two, and average of three for hour, day, and week references (DA1RefHr, DA1RefDy, DA1RefWk, DA2RefHr, DA2RefDy, DA2RefWk, DA3RefHr, DA3RefDy, and DA3RefWk), and machine learning prediction based on past deviations (ABR, BRR, DTR, GBR, SVR, and RFR) with input variables being hour of the day, day of the week, month of the year, the previous one and two hour states, weather (temperature, humidity, wind speed and rain), and percentage of users.

All these predictors will be tested in London and Washington data. Comparison will also be made with the work from Giot and Cherrier [47] for Washington data. Following the methodology defined in the previous section, the predictions are done first on the validation set. Then, after getting the best features and predictor with tuned hyper-parameters, the predictions are conducted using the test set. Finally, the performance metrics are analysed and comparisons are conducted among the three approaches.

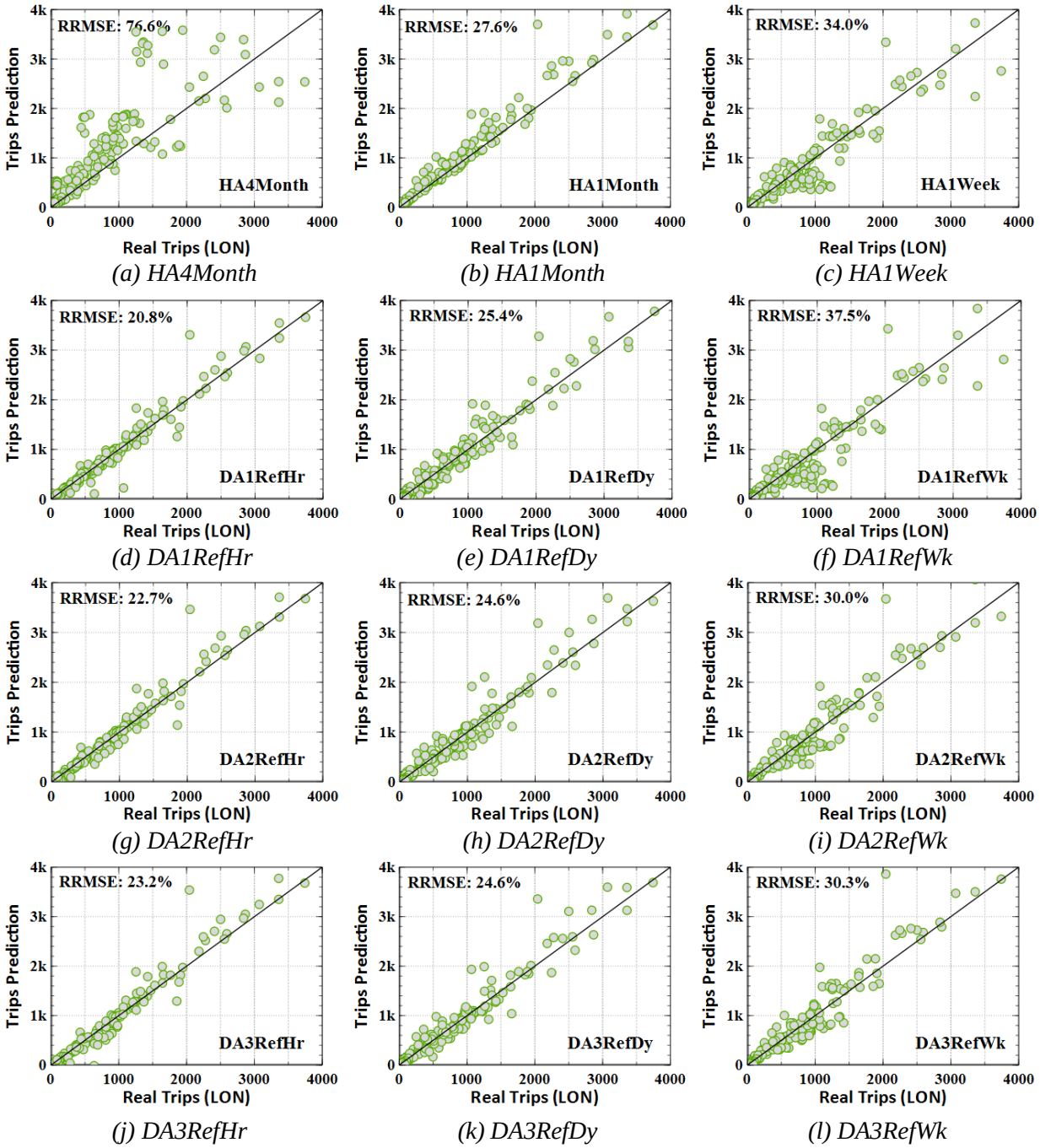
7.2.1. Naïve Prediction Results

The naïve prediction is conducted using the *weekly-daily-hourly* basis that is applied to predict the current hour trip number in the validation set. Here, the weekly-daily-hourly basis means there are 24 hours bins for each day of the week which is equal to 168 hourly average bins. First, RMSE is computed, and then the RRMSE is calculated from the hourly trip average in the validation set which is 816.25 in London and 231.25 in Washington DC. All the prediction performance in RMSE and RRME are presented in Table 7.1.

Table 7.1. Naïve prediction RMSE and RRMSE results.

No	Predictor	London Data		Washington Data	
		RMSE	**RRMSE (%)**	**RMSE**	**RRMSE (%)**
Historical Average					
1	HA4Month	625.2	76.6	69.6	30.1
2	HA1Month	225.3	**27.6**	54.6	**23.6**
3	HA1Week	277.5	34.0	100.6	43.5
Deviation Average					
4	DA1RefHr	169.8	**20.8**	43.9	**19.0**
5	DA1RefDy	207.3	25.4	70.1	30.3
6	DA1RefWk	306.1	37.5	108.2	46.8
7	DA2RefHr	185.3	**22.7**	44.6	**19.3**
8	DA2RefDy	200.8	24.6	86.3	37.3
9	DA2RefWk	244.9	30.0	81.6	35.3
10	DA3RefHr	189.4	**23.2**	45.8	**19.8**
11	DA3RefDy	200.8	24.6	86.7	37.5
12	DA3RefWk	247.3	30.3	57.6	24.9

It can be seen from Table 7.1 that for historical average (HA) based prediction (lines 1-3 in the table), the best predictor for both cities is one month historical average (line 2) with RRMSE 27.6% and 23.6% for London and Washington data respectively. This suggests that shorter and longer averages perform worse. In London itself, the end of daylight saving (refer to Chapter 4 subsection 4.2.2) obviously has a significant effect on the four months average to reduce the accuracy. Using just the previous week does not even out any weekly disturbances, and does not represent the underlying trend. Individual weekly disturbances will give high error. Therefore, in deviation average (DA) based predictions, only one month deviation averages are used. For one reference (lines 4-6), average of two references (lines 7-9), and average of three references (lines 10-12), the best predictions all come from the hour reference. Among all, the one hour reference (DA1RefHr) is the best with RRMSE 20.8% and 19.0% for London and Washington data respectively. This indicates that DA predictors are better than HA predictors.



(a) HA4Month *(b) HA1Month* *(c) HA1Week*

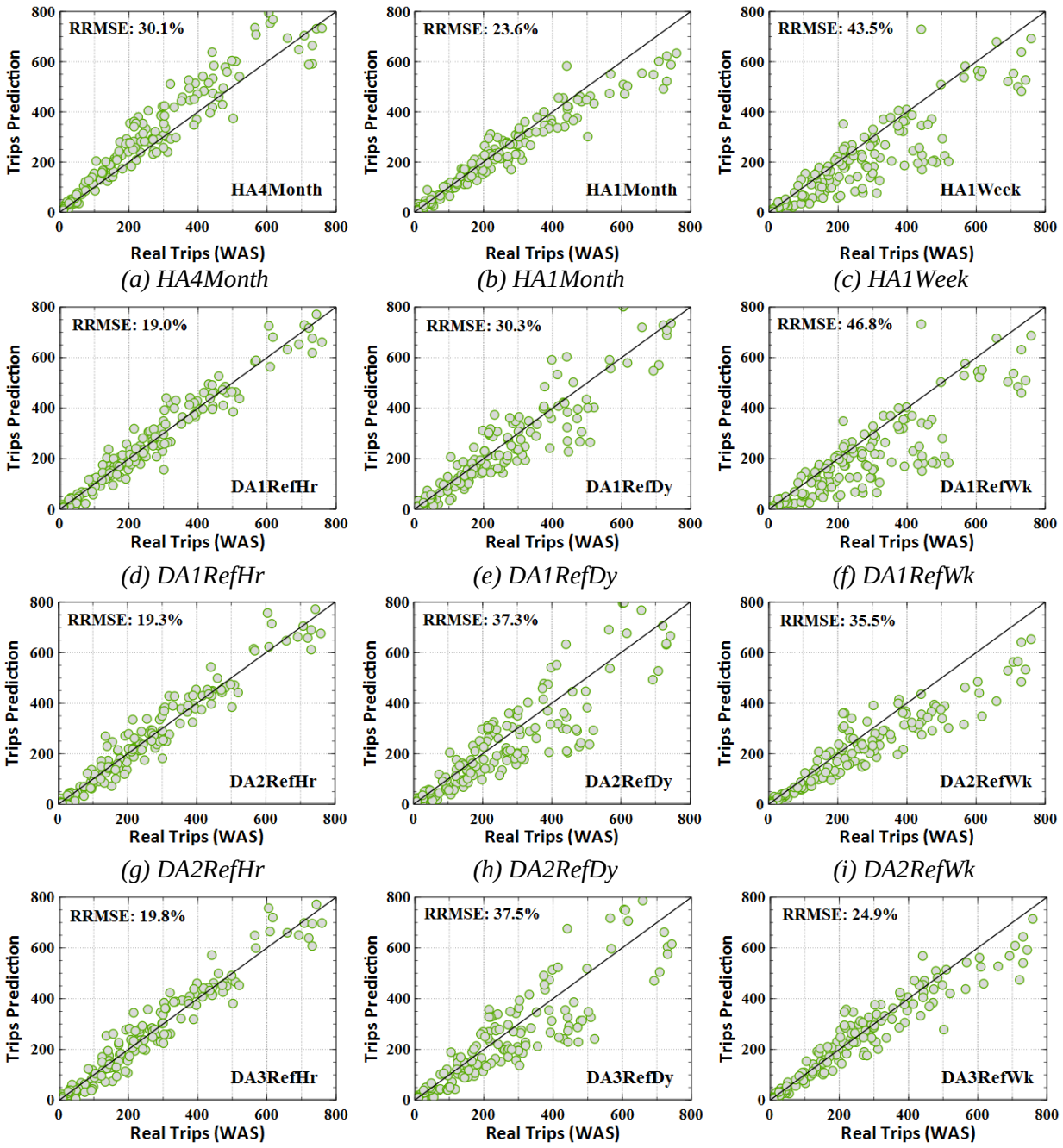
(d) DA1RefHr *(e) DA1RefDy* *(f) DA1RefWk*

(g) DA2RefHr *(h) DA2RefDy* *(i) DA2RefWk*

(j) DA3RefHr *(k) DA3RefDy* *(l) DA3RefWk*

Figure 7.3. Naïve prediction Vs real trips (London) for 168 hours prediction (light green circles) with binning (blue circles). (a-c) HA, (d-c) DA1Ref, (g-i) DA2Ref references, and (j-l) DA3Ref.

To visually observe how close the prediction results compare to the ground truth, the visual comparisons of the predicted trips vs real trips are presented in Figure 7.3 for London and Figure 7.4 for Washington data. The light green circles represent all predicted points (168 hours ~ green circles) along 7 days in validation set. A lower RRMSE will be where the circles are closest to the diagonal line corresponding to perfect prediction, and the visually best predictor is DA1RefHr, Figure 7.3.d and 7.4.d, which agrees with the results from Table 7.1.



(a) HA4Month *(b) HA1Month* *(c) HA1Week*

(d) DA1RefHr *(e) DA1RefDy* *(f) DA1RefWk*

(g) DA2RefHr *(h) DA2RefDy* *(i) DA2RefWk*

(j) DA3RefHr *(k) DA3RefDy* *(l) DA3RefWk*

Figure 7.4. Naïve prediction Vs real trips (Washington) for 168 hours prediction (light green circles) with binning (blue circles). (a-c) HA, (d-c) DA1Ref, (g-i) DA2Ref references, and (j-l) DA3Ref.

As the deviation based prediction using previous hour reference are better than day and week references in both cities, this suggests that the state one hour previously is very significant as a prediction feature for the current state in this prediction scenario. In other words, using this naïve approach, a closer reference point to the intended state is better. Based on this result, the machine learning approach in the next section will only focus on using the

deviation-based approach. The main goal is to investigate whether more complex prediction methods yield better results.

7.2.2. Coefficient Correlation and Feature Importance

How each feature correlates independently with the predicted metric is evaluated with a Pearson's Coefficient Correlation test, while the importance of each feature in machine learning prediction is evaluated with a feature importance test. Both tests are done with the training dataset. Note that BRR does not provide feature importance, it has a coefficient weight for each parameter so that it is represented by a numerical value instead of a percentage. The test results for each reference are shown in Table 7.2 as percentages except for BRR.

Following the Evans range of correlation [125], for the Current Time, all features have *very weak* (0 – 19%) and *weak* correlation (20 – 39%). This is because only a couple of these many time features are active for any one measurement. Any one time feature (eg. hr1) is only active for a small percentage of examples. Only the one previous hour state based on week and day reference have a *strong* (60 – 79%) and *very strong* (80 – 100%) correlation respectively. Then, their correlations decrease for the two previous hour state.

For weather features, most of them are *very weak*. However, this could be that the effect of weather is already present in the previous hour inputs, and so separate weather inputs do not add much additional information. Similarly, the percentages of unregistered users have a *very weak* correlation. The feature importance for ML also exhibits a similar trend to the Pearson correlation.

Table 7.2. Pearson's Coefficient Correlation and ML Feature Importance or Coefficient.

Features	Pearson's Coeff Correlation						ML Feature Importance/Coef					
	LON			WAS			LON			WAS		
	Hrs	Day	Week	Hrs	Day	Week	Hrs	Day	Week	Hrs	Day	Week
							RF	BR	GB	RF	GB	RF
	%	%	%	%	%	%	%	Coef	%	%	%	%
State of times (Month of the year)												
• Aug	0	16	13	0	0	35	0.2	-0.08	0.0	0.1	0.3	0.2
• Sep	0	-4	-2	0	2	-1	0.1	-0.14	0.6	0.1	0.3	0.1
• Oct	0	-7	-6	0	-2	-18	0.0	0.14	0.2	0.2	0.4	0.2
• Nov	0	-5	-5	0	-1	-16	0.1	0.10	0.3	0.3	0.1	0.1
• Dec	nan	nan	nan	nan	nan	nan	0.1	0.01	0.1	0.0	0.0	0.0
State of times (Day of the week)												
• Mon	0	-16	-4	0	-12	-2	0.2	-0.04	0.2	0.1	0.5	0.2
• Tue	0	6	-4	0	-1	-3	0.3	0.02	0.1	0.1	0.3	0.2
• Wed	0	9	2	0	16	3	0.1	-0.03	0.1	0.1	0.3	0.2
• Thu	0	5	3	0	-3	2	0.1	-0.02	0.3	0.1	0.2	0.1
• Fri	0	0	4	0	1	2	0.1	0.02	0.1	0.3	0.4	0.2
• Sat	0	5	0	0	13	1	0.3	0.02	0.2	0.2	0.3	0.2
• Sun	-1	-9	-2	-1	-15	-4	0.0	0.00	0.0	0.4	0.6	0.1
State of times (Hour of the day)												
• h0	-4	-1	0	-7	0	-2	0.0	0.01	0.0	0.1	0.3	0.0
• h1	-2	-1	0	-5	0	-2	0.0	0.01	0.0	0.2	0.0	0.0
• h2	-1	-1	0	-3	0	-2	0.0	0.01	0.0	0.0	0.0	0.0
• h2	-1	-1	0	-2	0	-2	0.0	0.01	0.0	0.0	0.0	0.0
• h4	0	-1	0	-1	0	-2	0.0	0.01	0.0	0.0	0.0	0.0
• h5	8	-1	-1	3	0	-2	0.1	0.00	0.2	0.1	0.0	0.0
• h6	26	0	-2	12	0	-1	0.0	-0.01	2.2	0.1	0.1	0.1
• h7	35	2	-2	31	0	0	0.2	0.00	3.6	0.5	2.1	1.5
• h8	-18	2	5	31	0	3	0.6	0.05	1.3	0.4	0.8	0.5
• h9	-22	0	2	-31	0	1	0.0	-0.01	0.0	4.7	0.7	0.7
• h10	-3	0	0	-9	0	0	0.0	-0.01	0.1	0.1	0.2	0.1
• h11	7	0	-1	8	0	0	0.1	-0.01	0.0	1.1	0.1	0.2
• h12	4	0	0	10	0	1	0.4	0.00	0.1	0.4	0.3	0.1
• h13	-1	0	0	-1	0	1	0.5	-0.01	0.0	1.1	0.2	0.1
• h14	2	0	-1	-3	0	1	0.3	-0.02	0.1	0.6	0.1	0.1
• h15	9	1	-2	3	0	2	0.2	-0.02	0.4	0.5	0.2	0.1
• h16	29	2	-4	14	0	2	0.8	-0.03	3.5	1.0	0.3	0.4
• h17	-3	2	1	32	-1	3	5.7	0.01	0.5	5.9	1.1	0.8
• h18	-25	1	2	-10	-1	2	5.7	0.00	0.3	2.1	0.5	0.4
• h19	-19	0	1	-26	-1	0	0.2	-0.01	0.3	0.2	0.4	0.3
• h20	-10	0	0	-19	0	0	0.2	0.00	0.0	1.7	0.1	0.1
• h21	-5	-1	0	-12	0	-1	0.0	0.00	0.0	0.0	0.2	0.1
• h22	-3	-1	0	-8	0	-1	0.0	0.01	0.0	0.0	0.1	0.0
• h23	-4	-1	0	-8	0	-1	0.0	0.01	0.0	0.0	0.1	0.0
One previous hour states												
• Deviation to ref points	44	83	77	31	75	87	35.3	0.91	42.3	29.3	63.0	74.5
• Percentage of unreg users	-9	8	9	-7	14	25	4.9	-0.09	7.4	4.5	3.7	2.3
• Temperature	-3	14	17	-8	1	34	0.5	-0.97	3.0	0.8	2.2	1.4
• Humidity	5	-20	-19	17	-7	-3	0.5	-0.32	2.5	1.0	2.0	1.3
• Wind speed	0	-1	-8	-9	-19	-16	0.3	-0.86	3.6	0.6	2.4	1.6
• Rain status	2	-25	-23	1	-15	-16	0.1	-0.15	0.9	0.2	0.2	0.1
Two previous hours states												
• Deviation to ref points	-25	64	55	-19	52	71	39.7	-0.19	13.0	30.9	4.9	3.9
• Percentage of unreg users	16	8	7	-14	11	21	0.9	0.03	4.7	7.1	3.5	2.6
• Temperature	-3	13	16	-9	1	33	0.5	-0.14	2.9	1.0	2.4	1.4
• Humidity	6	-17	-17	19	-5	-1	0.4	-0.41	2.2	1.0	2.2	1.5
• Wind speed	0	-1	-8	-8	-17	-16	0.3	-0.36	2.0	0.6	1.8	1.5
• Rain status	7	-23	-21	1	-16	-17	0.1	-0.19	0.6	0.1	0.3	0.3

7.2.2. Machine Learning Prediction in Validation Dataset

The machine learning prediction in this section uses the historical references that give the best performance in naïve deviation-based prediction for each reference type (hour, day or week). In London, they are one previous hour (DA1RefHr), the average of two days (DA2RefDy), and the average of two weeks (DA2RefWk). While in Washington DC, they are one hour (DA1RefHr), one day (DA1RefDy), and the average of three weeks (DA3RefWk).

Following the proposed feature selection described earlier, the primary time features consist of the current month, day and hour. The format of these features is constructed as a binary value (1 or 0) for each category. Hence, the feature current time will be as follow:

$$Current\ Time = Aug, Sep, Oct, Nov, Dec, Sun, Mon, Tue, Wed, Thu, Fri, Sat,$$
$$h1, h2, h2, h3, h4, h5, h6, h7, h8, h9, h10, h11, h12, h13,$$
$$h14, h15, h16, h17, h18, h19, h20, h21, h22, h23, h24$$

Here, there are **36 fields** in the Current Time: 5 fields for months, 7 fields for days of the week, 24 fields for hour of the day. For example, if the reference point is on Tuesday, 4th December at 9 am, then the Current Time features with value 1 are only *Dec*, *Tue*, and *h9*, while others will be 0. These features can be extended if the data covers a complete 12 months, or if the holidays along the learning period are included as a feature, or if four seasons are taken into account.

To see the effect of combining features, the ML prediction is conducted in three rounds. In the first round, the one previous hour metrics such as one previous hour deviation, one previous hour of casual user percentages, and one previous hour of weather (temperature, humidity, wind speed and rain) are added so that **42 features** are used. In the second round, the two previous hour's metrics are added so that **48 features** are used. In the third round, the features which have strong and very strong correlation only are considered to see whether using fewer, better features can give a better result. Using six regressors as explained in the Methodology Section, Table 7.3 presents RMSE and RRMSE for each reference type of each round for both cities including the features are used.

The first round prediction using Current Time (36 features) and one previous hour metrics (6 features) can achieve better performance than the naïve predictors. The best prediction error is 17.1% RRMSE for London using SVR with two day reference and 17.4% for Washington DC using RFR with three week reference.

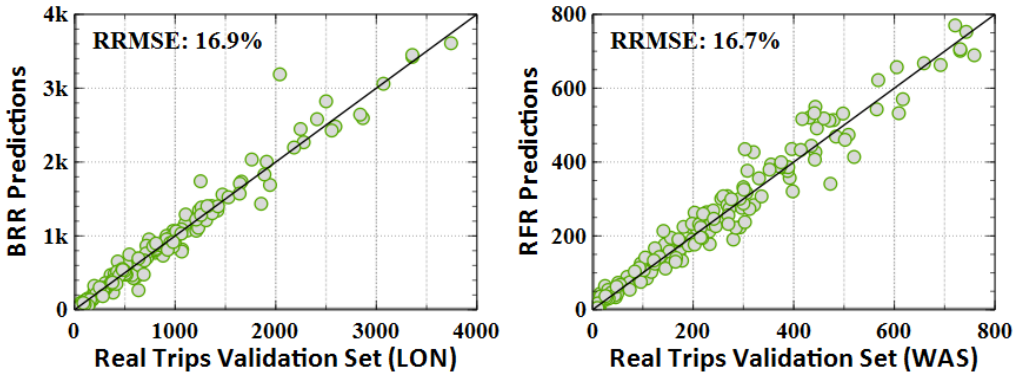
Table 7.3. RMSE and the percentage of RMSE of the system-wide prediction.

Ref.	Prediction of London												Prediction of Washington DC											
	RMSE						RRMSE (%)						RMSE						RRMSE (%)					
	AB	BR	DT	GB	RF	SV	AB	BR	DT	GB	RF	SV	AB	BR	DT	GB	RF	SV	AB	BR	DT	GB	RF	SV
1st	**Features:** ***Current Time*** *(Aug, Sep, Oct, Nov, Dec, Sun, Mon, Tue, Wed, Thu, Fri, Sat, h1, h2, h3, h4, h5, h6, h7, h8, h9, h10, h11, h12, h13, h14, h15, h16, h17, h18, h19, h20, h21, h22, h23, h24),* ***one previous hour metrics*** *(deviation1, casual percentage1, temperature1, humidity1, wind speed1 and rain1)*																							
HR	497	660	396	337	259	619	60.9	80.8	48.5	41.3	31.7	75.8	74.4	78.3	61.4	53.1	41.4	119.3	32.1	33.8	26.5	22.9	17.9	51.5
DY	285	141	369	183	171	140	34.9	17.2	45.2	22.4	20.9	17.1	58.8	46.6	85.4	48.7	48.7	46.6	25.4	20.1	36.9	21.0	21.0	20.1
WK	370	164	284	154	195	159	45.3	20.1	34.8	18.9	23.9	19.5	60.2	44.4	71.3	40.7	40.4	43.5	26.0	19.2	30.8	17.6	17.4	18.8
2nd	**Features:** ***Current Time*** *(Aug, Sep, Oct, Nov, Dec, Sun, Mon, Tue, Wed, Thu, Fri, Sat, h1, h2, h3, h4, h5, h6, h7, h8, h9, h10, h11, h12, h13, h14, h15, h16, h17, h18, h19, h20, h21, h22, h23, h24),* ***one previous hour metrics*** *(deviation1, casual percentage1, temperature1, humidity1, wind speed1 and rain1),* ***two previous hour metrics*** *(deviation2, casual percentage2, temperature2, humidity2, wind speed2 and rain2)*																							
HR	339	528	339	217	157	521	41.5	64.7	41.5	26.5	19.3	63.9	66.3	75.6	49.1	50.4	38.5	113.3	28.7	32.7	21.3	21.8	**16.7**	49.0
DY	166	138	334	171	151	139	20.3	**16.9**	40.9	20.9	18.5	17.0	56.6	47.4	82.0	46.3	47.2	46.4	24.5	20.5	35.5	20.1	20.4	20.1
WK	234	164	192	158	169	159	28.6	20.1	23.5	19.4	20.7	19.5	67.1	44.7	62.4	39.4	39.3	43.3	29.1	19.4	27.0	17.1	17.0	18.7
3rd	**Features:** ***Current Time*** *(Aug, Sep, Oct, Nov, Dec, Sun, Mon, Tue, Wed, Thu, Fri, Sat, h1, h2, h3, h4, h5, h6, h7, h8, h9, h10, h11, h12, h13, h14, h15, h16, h17, h18, h19, h20, h21, h22, h23, h24),* ***the strong & very strong features*** *(deviation1, deviation2)*																							
HR	274	508	226	482	143	530	33.5	62.2	27.7	59.1	17.5	64.9	72.2	74.9	54.7	68.5	39.2	113.7	31.2	32.4	23.6	29.6	**16.9**	49.1
DY	225	138	231	151	147	139	27.5	**16.9**	28.3	18.5	18.0	17.0	69.6	47.7	74.4	58.1	44.1	46.6	30.1	20.6	32.1	25.1	19.0	20.1
WK	227	157	244	175	171	158	27.8	19.2	29.9	21.4	21.0	19.3	76.4	44.6	61.3	51.5	39.4	43.6	33.0	19.3	26.5	22.2	17.0	18.8
Noted: The best prediction from Giot and Cherrier [47] for Washington data using RMSE metric													**102**	**79**	**-**	**312**	**336**	**336**						

The second round prediction by adding the two previous hour metrics to the first round prediction gives an improvement in both cities for almost all predictors and references. This means that two previous hour metrics can improve the prediction. The best performance that can be achieved in this round is 16.9% for London using BRR with two day reference and 16.7% for Washington DC using RFR with one hour reference.

The third round prediction by choosing only the features that have strong and very strong correlation, which are the one and two previous hours deviation (Table 7.2), gives almost similar results to the second round prediction. The best performance that can be achieved in this round is also 16.9% for London using BRR with two day reference and 16.9% for Washington DC using RFR with one hour reference. All results suggest that the one and two previous hours deviation have a role to improve the prediction performance of the current state.

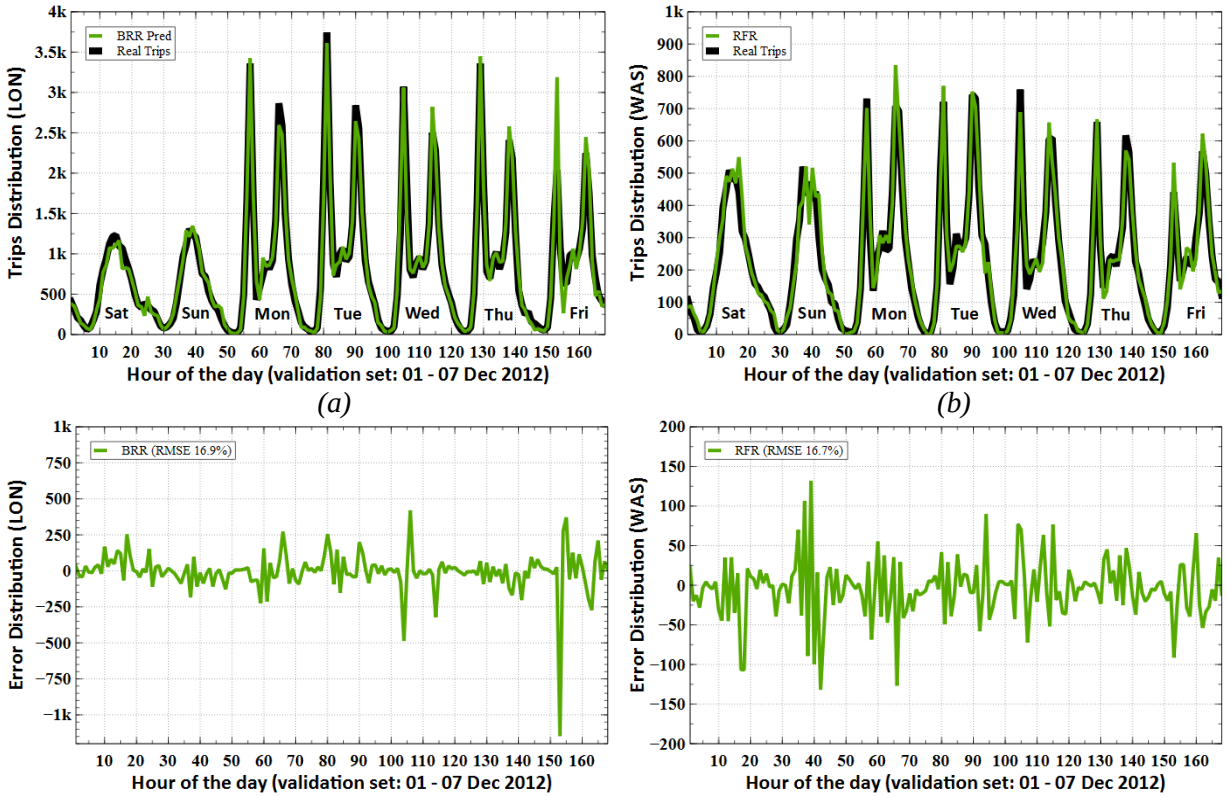
Compared to the existing works from Giot and Cherrier [47] using Washington DC data as can be seen in the bottom of Table 7.3, their smallest RMSE is 79 by Ridge Regression. Meanwhile, using similar Ridge Regression, this study achieves RMSE 75.5. The smallest RMSE is 38.5 using Random Forest, while their Random Forest gives an RMSE of 336. This means that the proposed deviation-based prediction in this study is much better.



(a) *(b)*

Figure 7.5. The best performance of ML prediction round three.

For the best predictor, Figure 7.5 suggests that the distribution between real trips as a ground truth and the prediction is quite linear if trips for London are less than 1800 and Washington DC are less 300 per hour. This is because most hourly trips occurred below than those points so that there are enough data to learn. However, that visualization cannot give the time information of prediction. Hence, in order to see at what times the prediction gives over-estimations and under-estimations, Figure 7.6 visualizes the time series patterns of errors.



(a) *(b)*

(c) *(d)*

Figure 7.6. (a-b) Real trips Vs Best ML prediction in validation set (BRR for London and RFR for Washington DC), (c-d) Error distribution (real trips minus ML prediction).

It is clearly seen in both cities that over and under estimations mostly happen in peak times. In London, for example, there is a slight underestimate on Monday and Tuesday afternoon, but on Wednesday and Thursday afternoon the predictor overestimates. Its overestimate increases on Friday morning because the actual trips drop and do not follow a common level with the rest of the week at that time. This shows that there is a significant decrease in using BSS at that time that cannot be easily predicted.

7.2.4. Machine Learning Prediction in Testing Dataset

After obtaining the best predictor, the best references, and the strong and very strong features, now prediction is done for the two weeks of the test set to see the generalization of the model. Here, the only strong and very strong features are used because they give almost similar RRMSE with using all features. The ML results are shown in Figure 7.7 for the test set week 1: 8th -14th December (Figure 7.7.a&b) and week 2: 15th – 21st December (Figure 7.7.c&d), and the comparisons to the naïve approach are given in the next paragraph.

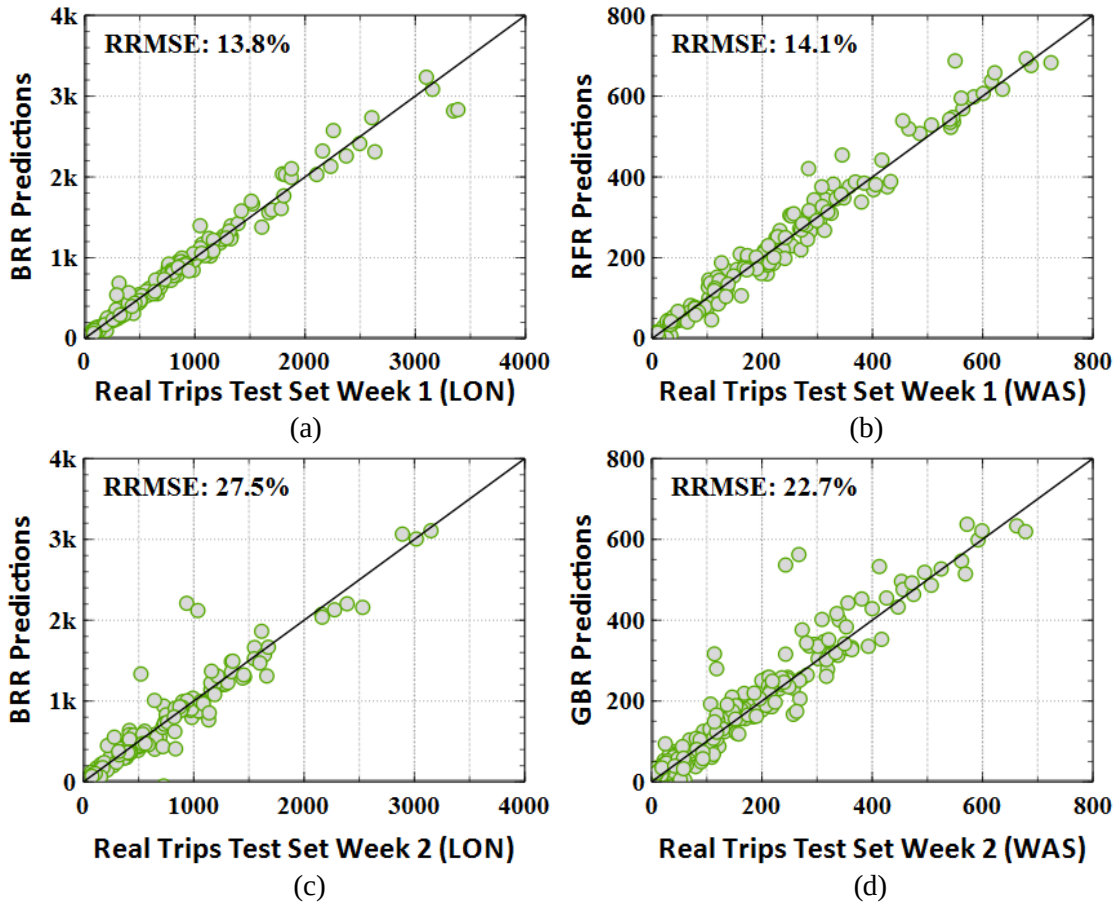


Figure 7.7. The ML prediction of testing set: (a,b) week 1 and (c,d) week 2.

The ML results show that for the testing set week 1 there is an improvement in performance from the validation set. Now, RRMSE for London is 13.8% and Washington DC is 14.1% while in the validation set it is 16.9% and 16.7% for London and Washington DC respectively. On the other hand, for testing set week 2 there is a downgraded performance to 27.5% for London and 22.7% for Washington DC. The reason of these phenomena can be explained using Figure 7.7 and 7.8. However, all the ML results are still better than the naïve approaches (DA1RefHr) which are 20.5% (week 1) and 44.5% (week 2) for London, and 16.9% (week 1) and 24.1% (week 2) for Washington DC.

Figure 7.8 shows that week 1 of testing set (8^{th} – 14^{th} December) for both cities look normal, with no significant uncommon patterns happening in actual trips (black line). The only underestimated predictions occur similarly on Monday and Tuesday for both cities and slight overestimates for the days after that in London. On the other hand, there are the unusual patterns in week 2 in London, Figure 7.9.a, starting from Wednesday afternoon where trips in peak times decrease significantly different from the previous Monday and Tuesday. This trend is followed by the days after Wednesday. While in Washington DC, it happens on Friday, Figure 7.9.b. Those decreases could be because a Christmas holiday effect had already begun.

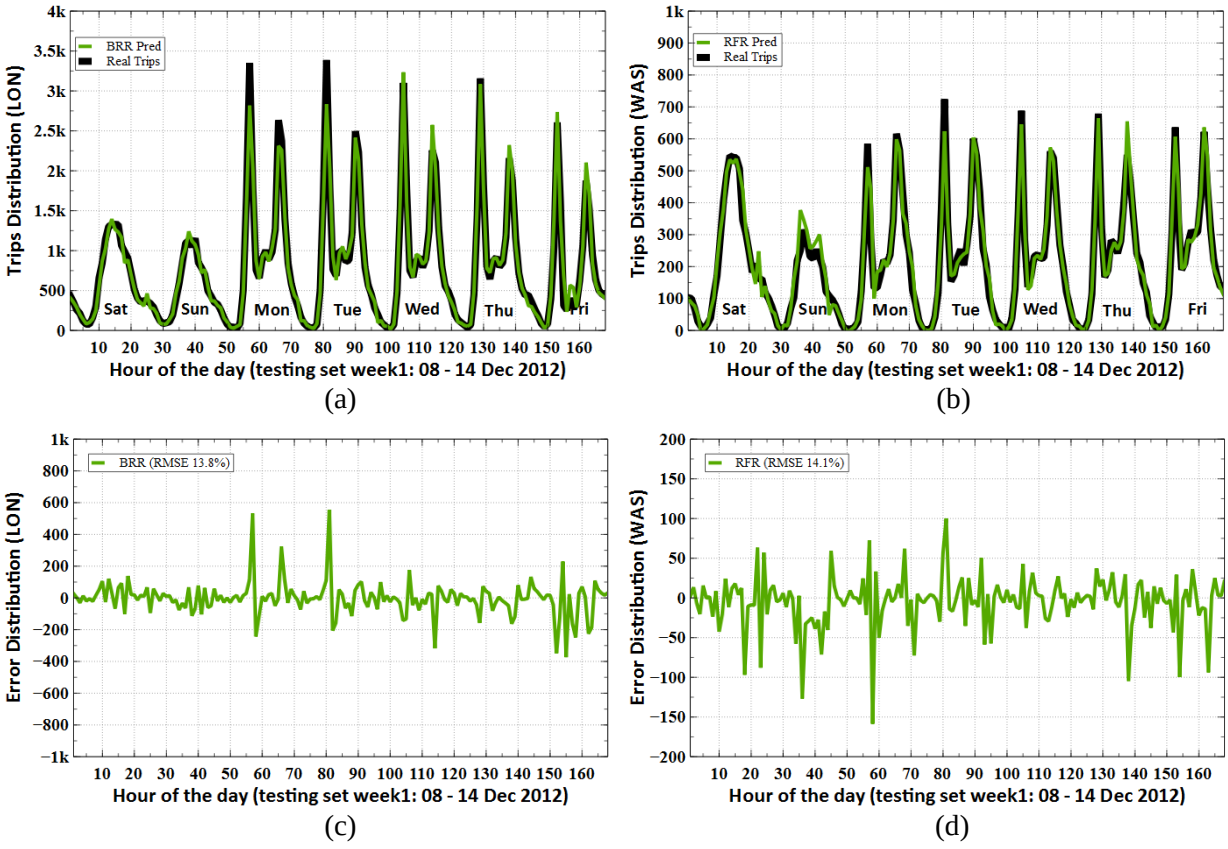


Figure 7.8. The ML prediction of test set week 1 (a,b) and error distribution (c,d).

Comparing week 2 of the testing set (15^{th} – 21^{st} December) patterns in both cities, there are interesting phenomena occurring where the predictor shows different responses to the sudden outlier patterns, Figure 7.9. In London, when real trips (black line) on Wednesday afternoon suddenly dropped and the predictor cannot predict it well, the predictor overestimates, but the day after (on Thursday afternoon) the predictor can predict it correctly. Similarly, when trips on Thursday morning dropped, then on Friday morning a reasonable prediction can be made so that there is no significant error on Friday. This daily basis adjustment works because of the implementation of one-day reference in London. While in Washington DC, its adjustment is even more responsive since its prediction scenario uses the previous hour reference. This can be seen on Friday in Figure 7.9.b. When trips on Friday morning and afternoon drop to less than the previous days, the predictor can still give a good prediction following the actual trips so that no significant errors happen, unlike London that needs a day to wait to respond to the sudden outlier patterns and return to the right level.

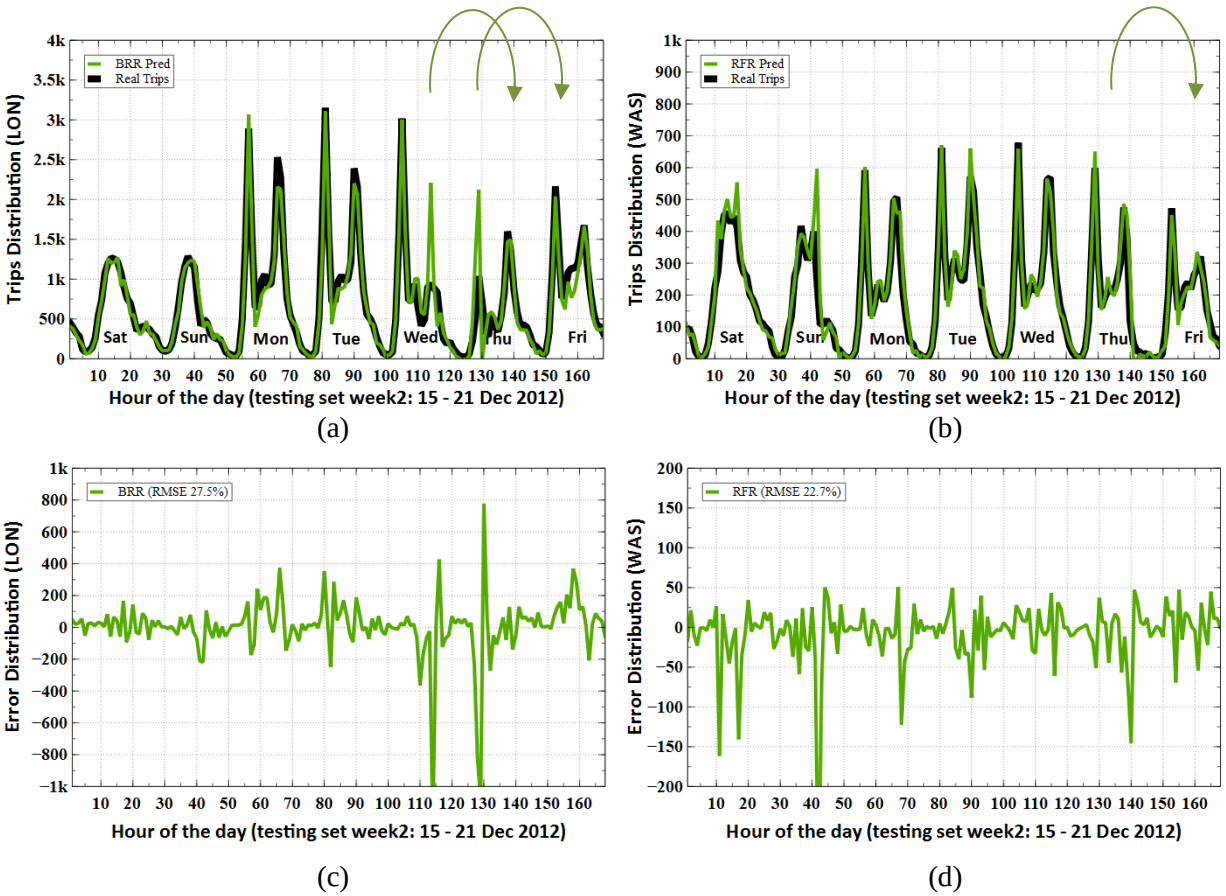


Figure 7.9. The ML prediction on testing set week 2 (a,b) and error distribution (c,d).

7.3. Cluster Prediction

Clustering is a technique to identify groups whose members exhibit similar behaviour. Clustering in this study aims not to get the optimal clusters, but more to implement the proposed cyclostationary prediction scenario into smaller BSS sub-systems. This is to test whether localized hourly prediction can achieve a similar performance level to the system-wide prediction. From a practical operational viewpoint, the question is whether one can predict unbalanced bicycle use in a certain geographic area that might need proactive rebalancing.

Here, clustering based on station geolocation is needed so that the clusters members will be in similar region or close each other. The prediction will also give an insight which parts of the system are better predicted, whether the ML approach is still better than the naïve approach, and whether cluster-based prediction performance is similar to the system-wide performance. As stated in the Methodology section, the prediction metrics at the cluster level can be divided into three outputs, pickups (bike out), returns (bike in) and balance (out minus in), which represent different aspects of the cluster activities.

7.3.1. K-Means Clustering

In k-means clustering, the number of clusters, k, can vary from 1 to the number of stations in the system using geographical location as the features. Then, the average distances to the centre of each cluster are calculated. Figure 7.10 shows the relationship between number of clusters and their average radius. If 1 km is selected as the average radius to the centre based on a reasonable walking distance, the approximate number of clusters is 75 (seventy five). This approximation is implemented in both cities as they show almost similar cluster numbers.

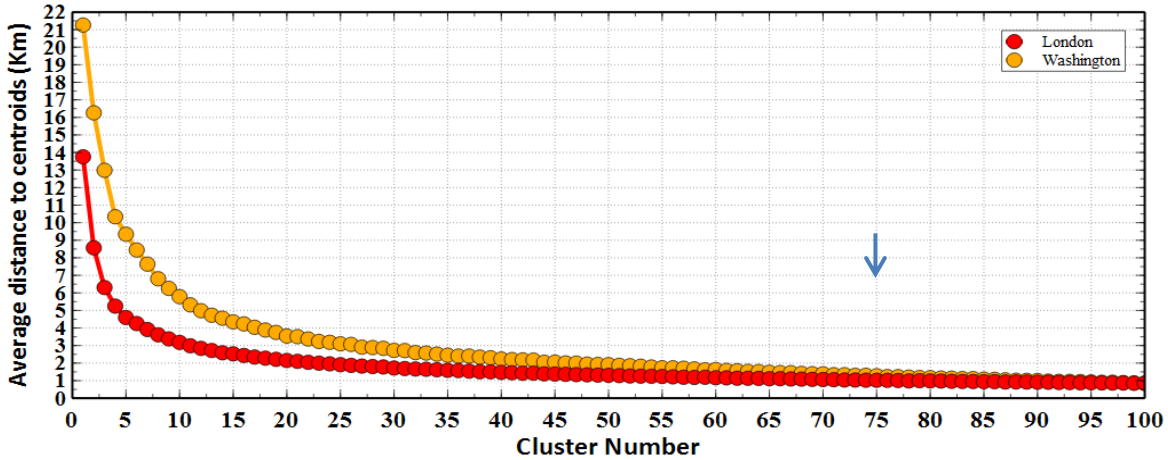
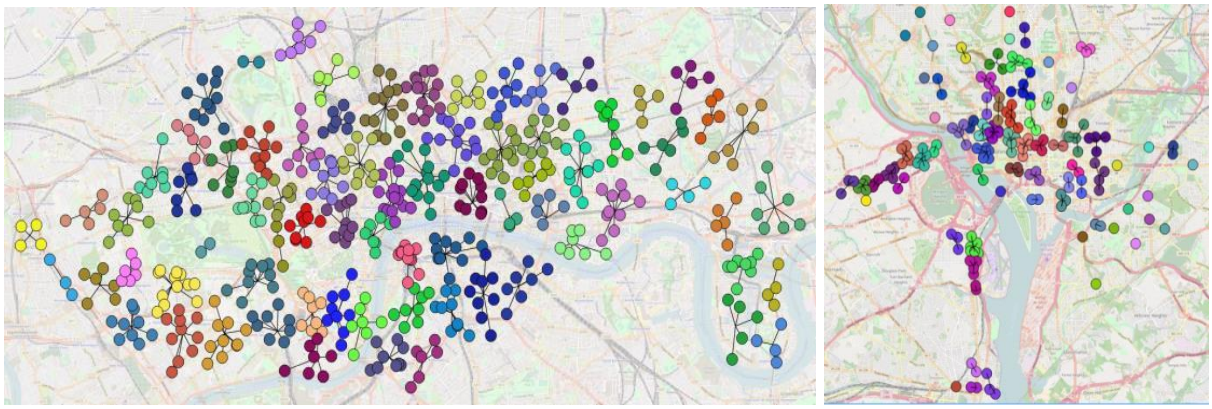


Figure 7.10. Cluster number vs distance to the centre.

Statistically, using 75 as number of clusters gives the minimum, average and maximum numbers of stations in a cluster in London as 2, 7.65 and 15, while in Washington DC they are 1, 2.58 and 7 respectively. Their distributions on the map can be seen in Figure 7.11. This region-based clustering gives cluster members close to each other. Other studies about station clusters have used station activities profile to give a small number of clusters of similar usage patterns.



(a) (b)

Figure 7.11. Map of BSS cluster station in (a) London and (b) Washington DC .

7.3.2. Cluster Prediction RRMSE-Range

At this cluster level, predictions are conducted using the best predictor algorithm from the system-wide prediction which is BRR for London and RFR for Washington DC. Then, they are compared to naïve prediction (DA1RefHrs). Three weeks of predictions are made where each week contains three different activities, pickup, return, and balance, where balance is number of pickup minus number of return.

Here, the RRMSE of 75 clusters is categorized in the form of error ranges to visualize performance across all the clusters, and the each range bin describes a 20% band of RRMSE because the prediction performance of system-wide prediction is approximately 20%. For example, if clusters have RRMSE 25% or 30%, then they will be categorized in the range bin 20% to 40%. This will give an insight into how the errors differ across clusters. The results of the RRMSE-range are summarized in Table 7.4 for the ML approach and Table 7.5 for the naïve approach with the heat map showing the number of stations that fall in the associated range, and where darker red means more clusters in that band.

The results show that prediction at cluster level cannot reach the good performance of system-wide prediction since none of the results has RRMSE below 20%. Generally, for the

ML approach in Table 7.4, the RRMSE of pickup and return prediction are relatively similar, mostly in the range of **40-60%** for London and **60-80%** for Washington DC. On the other hand, the prediction of balance is less than the individual pickup and return prediction. They are mostly in the range of **80-120%** for London and **140-200%** for Washington DC. This suggests that the prediction in London with larger station numbers for each cluster is better than in Washington DC. The prediction using ML approach is still better than using naïve prediction, Table 7.5, where numbers of clusters with RRMSE in the range 20% to 40% and 40% to 60% using the ML approach are higher than using the naïve approach.

Table 7.4. RRMSE-Range of 75 clusters using BRR (London) and RFR (Washington DC).

Error Range		Number of Clusters in LONDON (BRR)									Number of Clusters in WASHINGTON DC (RFR)								
		Validation set			Test set week 1			Test set week 2			Validation set			Test set week 1			Test set week 2		
>	≤	Out	In	Bal	Out	In	Bal	Out	In	Bal	Out	In	Bal	Out	In	Bal	Out	In	Bal
0%	20%																		
20%	40%	20	18		20	22			1		2	2		1	1			1	
40%	60%	38	38	5	37	34	6	38	45		8	11		12	10		10	7	
60%	80%	10	12	11	12	11	11	24	16	4	15	13	2	12	13		11	15	1
80%	100%	3	3	20	1	3	20	7	8	11	8	7	5	4	7	6	8	6	4
100%	120%	2	2	18	3	3	19	3	2	24	8	5	5	8	5	6	5	5	6
120%	140%	1		9		1	6	1	2	19	3	5	10	6	2	5	6	4	8
140%	160%		2	9	1	1	10	1	1	12	4	2	10	5	9	8	5	4	6
160%	180%	1		2	1		2	1		4	5	5	10	6	3	15	2	2	13
180%	200%			1			1			1	5	5	11	3	3	11	5	3	10
> 200%											17	20	22	18	22	24	23	28	27
#Clusters		**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**

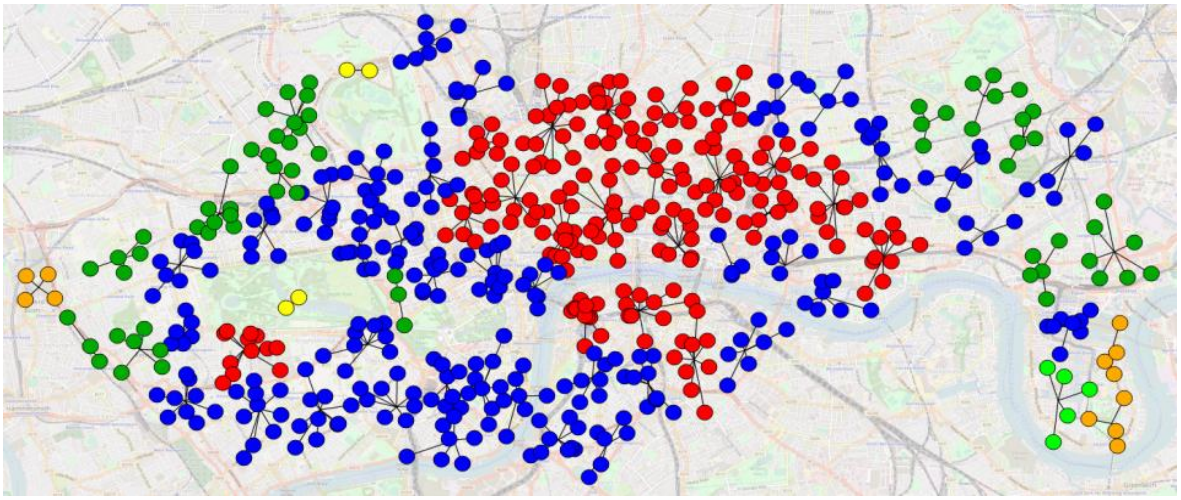
Table 7.5. RRMSE-Range of 75 clusters using naïve prediction (DA1RefHrs).

Error Range		Number of Clusters in LONDON (Naïve)									Number of Clusters in WASHINGTON (Naïve)								
		Validation set			Test set week 1			Test set week 2			Validation set			Test set week 1			Test set week 2		
>	≤	Out	In	Out	Out	In	Bal	Out	In	Bal	Out	In	Bal	Out	In	Bal	Out	In	Bal
0%	20%																		
20%	40%	1	3		3	1						2							
40%	60%	33	31	2	33	36	2	1	3		7	6		10	7		4	5	
60%	80%	22	20	6	19	20	7	24	27		9	12		5	9		11	11	
80%	100%	11	13	7	13	11	10	31	29	5	8	8	1	11	10		6	9	1
100%	120%	4	4	17	2	2	13	12	9	9	10	3	5	5	3	7	6	2	5
120%	140%	2	1	18	1		19	3	2	14	5	5	4	7	4	4	6	5	3
140%	160%		1	5	2	4	8	1	3	17	2	4	7	3	6	3	3	3	5
160%	180%	1		11	1		4	2		15	4	1	6	3	1	8	3	1	6
180%	200%		1	5			8	1	2	6	3	3	8	2	4	4	5	5	2
> 200%		1	1	4	1	1	4			9	27	31	44	29	31	49	31	34	53
#Clusters		**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**

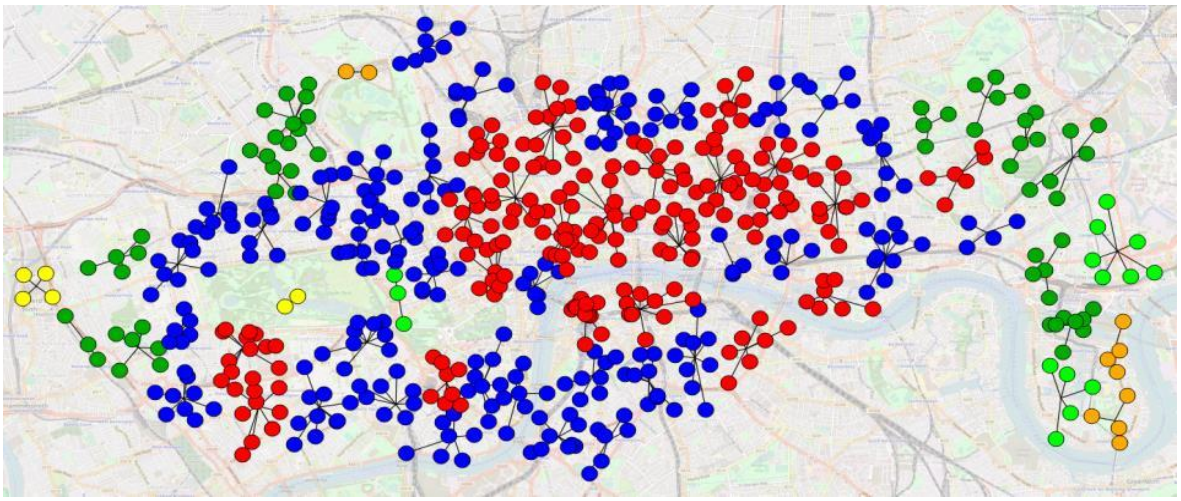
The high RRMSE could be due to the low numbers of pickups and returns at certain hours at cluster level, such as in the early morning, midday, and at night. The percentage of relative

error will be high. This suggests that hourly based prediction is not practical at this smaller subsystem level. This will be even clearer with station level prediction.

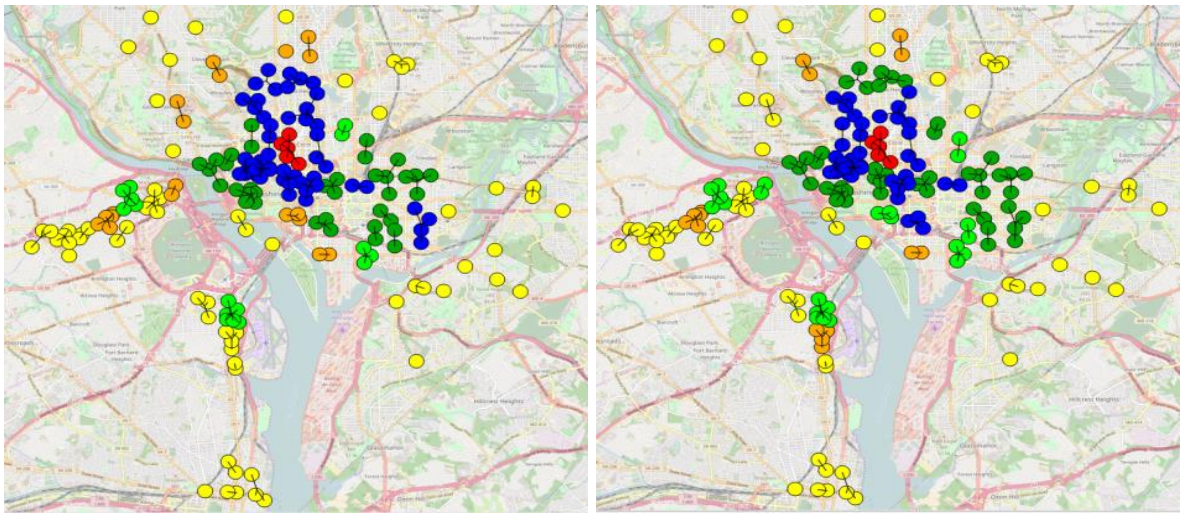
To see the relationship of prediction range with the spatial distribution of clusters, Figures 7.12 show the RRMSE-Range of pickup and return prediction on the map of London (a, b) and Washington DC (c, d). This visualization aims to see whether clusters with certain prediction range sit in certain places. Here, red denotes less than 40%, blue is 40.1-60%, dark green is 60.1-80%, light green is 80.1-100%, orange 100.1-120% and yellow is more than 120.1%.



(a) Pickup (London)



(b) Return (London)



(c) Pickup (Washington DC) *(d) Return (Washington DC)*

Figure 7.12. The RRMSE-Range of pickup and return on the map where red is less than 40%, blue is between 40.1% and 60%, green is between 60.1% and 80%, light green is between 80.1% and 100%, gold is between 100.1% and 120%, and yellow is more than 120.1% .

It can be seen for both cities that the smaller RRMSE clusters are in the centre of the map. Stations near the city centre are better predicted than outer stations. Stations with higher use exhibit more consistent usage patterns compared to those clusters which lead to better prediction.

Overall, cluster-based prediction of hourly usage does not appear to be sufficiently accurate to be useful for BSS operations.

7.4. Station Prediction

Station level prediction can be seen as a particular example of cluster prediction where one cluster has one member station. Therefore, all analyses will be similar to the previous cluster analyses but at a much larger scale. Before the prediction results are presented, the station usage pattern and station usage range on an hourly basis will be analysed to get a high-level view of how their hourly usages are distributed, and how the low usages at certain hours, early morning, midday, and at night, will potentially produce high relative errors. On the other hand, the peak time usages may be larger and more predictable. The station hourly usage prediction will be presented first to show the potential disadvantage of hourly based prediction at station level. Then, the peak hours based prediction will be proposed that could be more useful than hourly based prediction across the whole day. Imbalance is more likely during high usage periods, and effective prediction during these peak times will be of most practical benefit.

7.4.1. Station Hourly Usage Pattern

The averages of station hourly usage for each station are shown in Figure 7.13 by one point per station, ordered from highest to lowest. As explained in subsection 7.1.4, those actually include many zeros in the data because many stations, especially the small stations in the outer part of the city, do not receive pickup and return every hour, especially out of peak hours. Usage in these very quiet times is not predictable, but it is also not very useful, since it has little effect on BSS operations. This will be further shown in the following hourly usage distribution analysis.

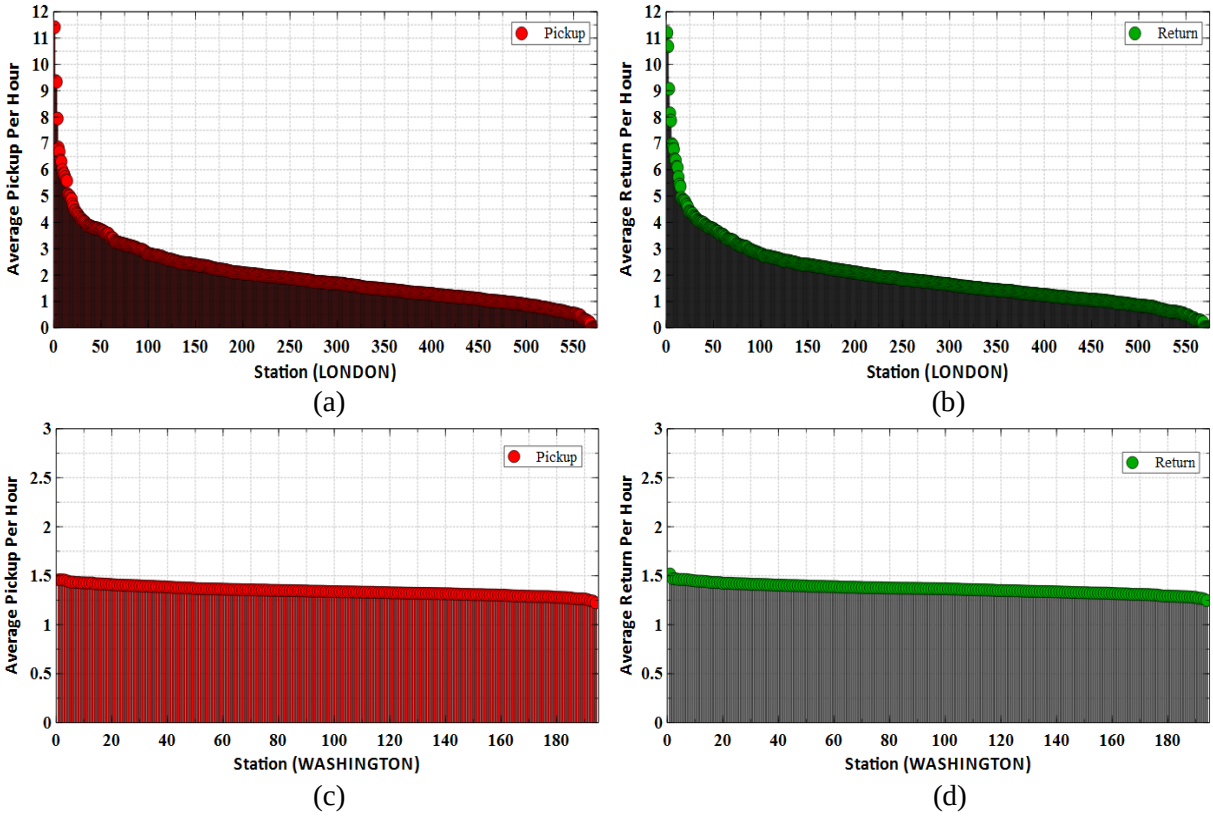


Figure 7.13. The average of hourly pickup and return of all stations.

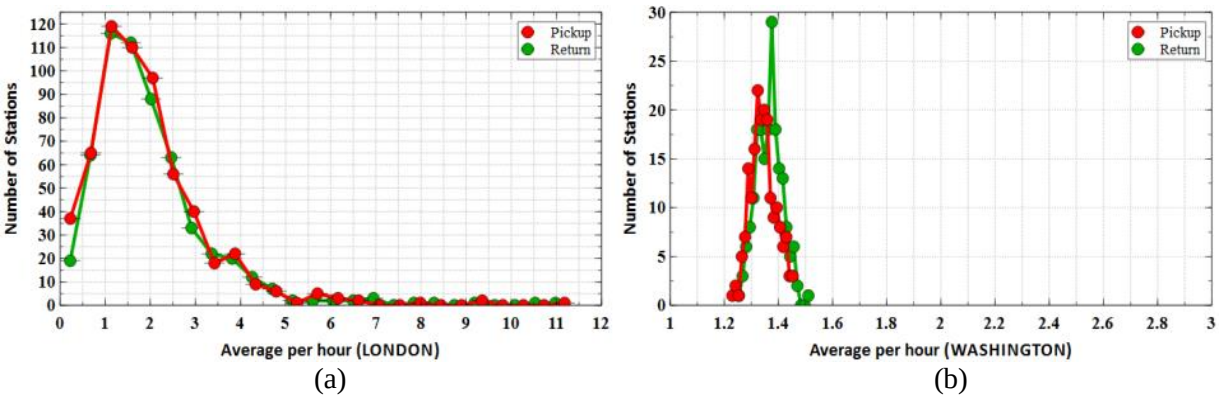


Figure 7.14. The average histogram of hourly pickup and return of all stations.

As can be seen, the hourly usage averages of stations in London are more varied than Washington DC which looks more homogenous at average of pickup and return around 1.3 per hour. The histograms of both averages have means which are similar at around 1 to 1.5 per hour as shown in Figure 7.14. Then, to define the range bins of a table of station use, two times the mean of the histogram is used, i.e. 2. The resulting hourly heat map of the number of stations based on their hourly average range can be seen in Table 7.6 and 7.7.

It is shown that in London before 5 am almost all stations have hours usage of less than 2, while even at peak times in the morning and afternoon, there are more than one hundred stations with usage less than 2 bikes. In Washington DC, on the other hand, all stations receive usage less than 2 bikes before 7 am, while during the afternoon peak the majority of station numbers increase only from range 0-2 to range 2-4, and only 14 stations have an hourly average more than 4 as shown in Table 7.7. Based on this distribution, it seems it is not sensible to predict usage before 5am in London and before 7am in Washington DC. To check this, the next section will first predict whole hours followed by the prediction of peak hours.

Table 7.6. The heat map table of number of stations based on their hourly average (London).

Range		**Hour of the day (LONDON)**																							
>	≤	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
0	**2**	572	573	573	573	573	531	293	114	141	317	429	339	302	314	305	240	151	130	192	333	450	501	521	554
2	**4**	1					34	172	125	183	208	129	192	201	197	194	177	139	151	199	170	99	60	47	17
4	**6**						5	56	125	138	40	13	34	51	48	57	95	96	104	103	49	20	9	3	2
6	**8**						1	30	83	61	4	2	7	15	11	13	36	61	70	39	14	3	3	2	
8	**10**							7	54	23	3		1	3	2	3	15	37	46	20	5	1			
10	**12**							4	30	13	1			1	1	1	6	37	23	10	1				
12	**14**						1	3	15	5							2	11	21	3					
14	**16**							1	11	2							2	13	7	4					
16	**18**							2	2									10	8	2	1				
18	**20**								4	2								6	1						
> 20							1	5	10	5								12	12	1					

Table 7.7. The heat map table of number of stations based on their hourly average (Washington DC).

Range		**Hour of the day (WASHINGTON DC)**																							
>	≤	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
0	**2**	194	194	194	194	194	194	194	194	15	190	194	189	94	99	124	93	1			20	187	194	194	194
2	**4**									179	4		5	100	95	70	101	193	180	194	174	7			
4	**6**																		14						

7.4.2. Station Hourly Usage Prediction

Similar to cluster prediction, the ranges of station prediction use a 20% range and are compared between the ML and naïve approaches. The results are shown in Table 7.8 and 7.9.

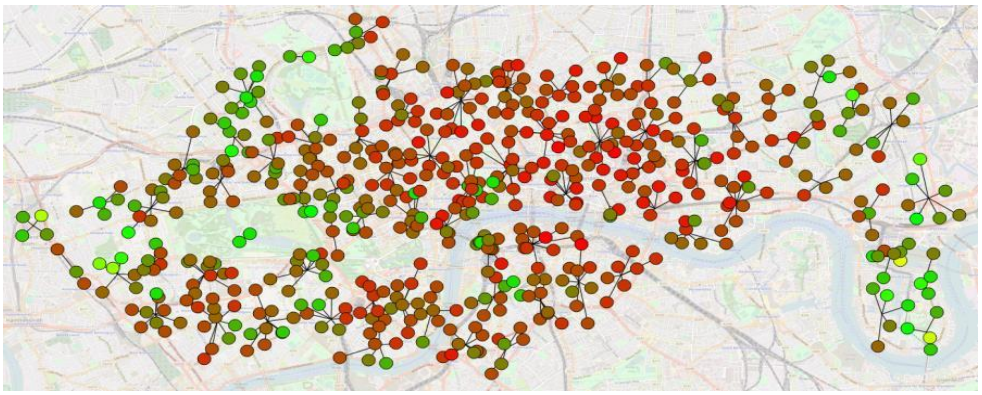
Table 7.8. RRMSE-Range using BRR (LON) and RFR (WAS. DC).

Error Level		LONDON (BRR)									WASHINGTON DC (RFR)								
		Validation set			Test set week 1			Test set week 2			Validation set			Test set week 1			Test set week 2		
>	≤	Out	In	Bal	Out	In	Bal	Out	In	Bal	Out	In	Bal	Out	In	Bal	Out	In	Bal
0%	20%																		
20%	40%			1															
40%	60%	6	5	1	5	5	2		1										
60%	80%	38	46	1	39	45	1	11	14	2									
80%	100%	119	109	1	116	91	6	60	64	1									
100%	120%	153	135	36	142	139	34	137	120	5									
120%	140%	79	84	98	96	107	92	113	131	53	4	4		8				2	
140%	160%	69	68	160	55	61	155	86	77	141	11	2		34	8		2	5	
160%	180%	37	48	133	37	38	133	53	49	174	15	17		42	21	4	14	13	
180%	200%	27	24	71	30	25	77	39	51	104	27	32		29	29	26	31	27	13
200%	220%	15	13	30	15	21	31	21	14	49	36	25	1	33	32	53	31	22	29
220%	240%	9	13	19	15	12	13	14	15	15	20	30	3	28	19	42	39	25	51
240%	260%	3	6	7	3	7	9	13	9	8	17	16	6	8	15	32	22	24	49
260%	280%	3	3	3	2	1	4	3	4	7	10	13	16	7	10	21	17	29	23
280%	300%	3	3		4	4	2	6	3	2	9	9	32	4	13	10	14	14	13
> 300%		2	3	1	5	4	2	4	3	2	28	32	43	0	8	0	2	6	0

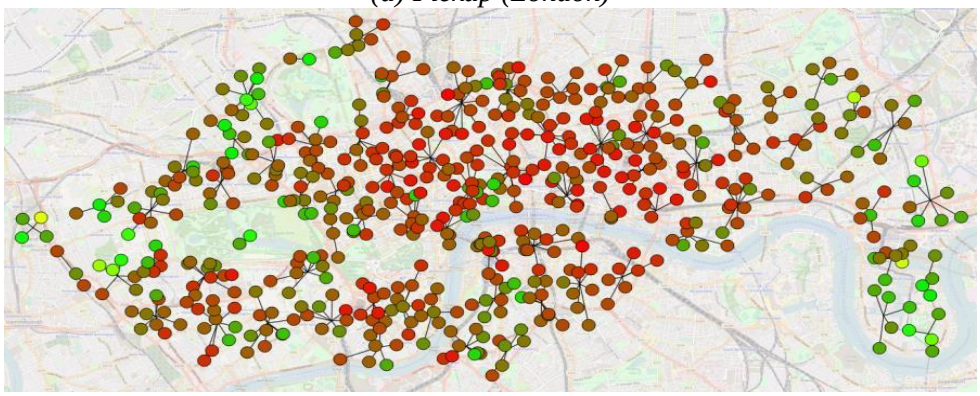
Table 7.9. RRMSE-Range using naïve prediction (DA1RefHr).

Error Level		LONDON (Naive)									WASHINGTON DC (Naive)								
		Validation set			Test set week 1			Test set week 2			Validation set			Test set week 1			Test set week 2		
>	≤	Out	In	Bal	Out	In	Bal	Out	In	Bal	Out	In	Bal	Out	In	Bal	Out	In	Bal
0%	20%																		
20%	40%																		
40%	60%	6	6	2	2	3													
60%	80%	69	68	1	50	44	2	4	1										
80%	100%	142	141	10	115	109	12	34	40										
100%	120%	128	121	66	158	149	52	115	121	9									
120%	140%	88	81	156	73	85	131	112	104	54									
140%	160%	52	63	139	55	59	172	100	94	134									
160%	180%	30	34	106	37	33	98	54	74	171				4					
180%	200%	16	17	48	25	36	44	57	41	93				24					
200%	220%	14	12	21	17	14	20	32	28	49	3			26		7			
220%	240%	6	6	11	12	8	14	18	21	24	6	2	1	23		17			
240%	260%	5	2	1	5	5	7	12	12	16	5	2	4	14	1	29			
260%	280%	5	3	1	3	2	4	8	10	2	15	7	7	12	7	26	1		6
280%	300%		4	1	3	6	3	5	7	8	8	12	17	9	9	34	2	2	10
> 300%		5	8	3	11	13	7	15	13	6	157	171	165	82	177	81	191	192	178

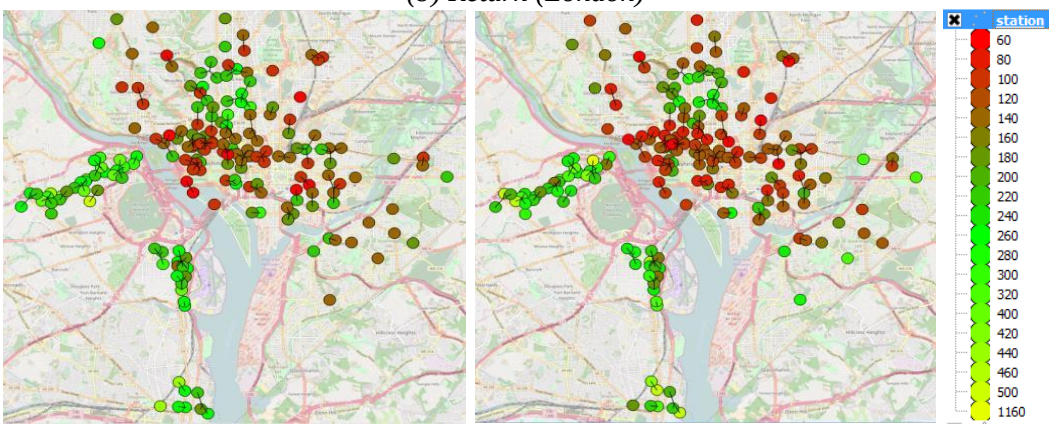
At this station level, the performance decreases significantly compared to the cluster level. It can be seen that RRMSE for pickup and return for most stations in London are in the range of 100-120% followed by 80-100%, 120-140%, and 140-160%. While in Washington DC, results are even poorer in the range of 200-240%. The spatial distribution of RRMSE per station is presented on the maps of London and Washington DC, Figure 7.15. Stations in the inner cities (Red Circles) have better range than the outer ones (Green Circles) for both cities. The red circle of return is spread broader than the green ones. The colour legend that corresponds to the RRMSE levels is shown in Figure 7.15.e.



(a) Pickup (London)



(b) Return (London)



(c) Pickup (Washington DC) *(d) Return (Washington DC)* *(e)*

Figure 7.15. Station RRMSE of pickup and return on the map where the upper bound of RRMSE is shown in (e).

7.4.3. Station Peak-Hour Usage Prediction

The low hourly prediction accuracy at station level, shown by high RRMSE, can be explained with an example: if a station receives no pickup or return at certain hours but the predictor predicts 1 or 2, this gives 100% or 200% error respectively which is a very high RRMSE. This suggests that RRMSE is not an appropriate metric for small BSS subsystems. As stated earlier in this section, prediction on an hourly basis is not practical for BSS operations, especially for redistribution purposes. There would be a very high cost if stations are visited by the operator to redistribute the bikes every hour or many times in a day. Ideally, stations should be balanced just before peak hours. Therefore, the peak-hour based prediction to predict the total usages during the coming peak hour seems more useful and will be tested in this section.

As shown in the Preliminary Data Analysis (subsection 4.2.2), the peak hours occur between 5 am to 9 am in the morning and 3 pm and 7 pm in the afternoon. Accordingly, the prediction is made two times a day which are at one hour before those peak hours using the cyclosationary deviation-based method as previously proposed for system-wide. The ten busiest stations of each city are predicted for the two weeks of the testing dataset, a one day previous reference is used, and the best ML predictors in system-wide prediction, BRR for London and RFR for Washington DC, are employed. Furthermore, the nearest station usage is added as an input feature to capture the spatial dependency features. Then, the prediction error is compared to the naïve approaches, DA1RefD, as shown in Table 7.10 for **RRMSE (%)** and Table 7.11 for RMSE **(rounded)** where red colour means naïve is better than ML, blue means both are similar, and blck means that ML is better.

Results in both cities show that, in the majority of cases, ML predictors give better prediction (smaller RRMSE) than the naïve approaches. Predictions for pickup in the morning peak mostly give less error than the return prediction. Conversely, predictions for return in the afternoon peak mostly give less error than the pickup prediction. When there is no anomalous usage in week one, the ML predictor gives RRMSE mostly less than 20% for both cities. The increasing errors occur in week two when there is anomalous usage at the end of the week as shown in subsection 7.2.4.

For balance predictions, this has higher RRMSE than pickup and return predictions. This is because balance itself comes from pickup minus return that makes its value smaller, and RRMSE relatively higher. However, in terms of RMSE, balance errors are quite similar to pickup or return errors, and some of them are even smaller as shown in Table 7.11.

Table 7.10. ***RRMSE (%)*** *of peak hours prediction using Machine Learning (BRR for London & RFR for Washington DC) and Naïve Prediction (DA1RefD).*

Station ID	Machine Learning Prediction						Naïve Prediction					
	Pickup		Return		Balance		Pickup		Return		Balance	
	Week1	Week2	Week1	Week2	Week1	Week2	Week1	Week2	Week1	Week2	Week1	Week2
LONDON Morning Peak (5am-9am)												
14	13.4	48.6	18.8	46.7	13.3	61.4	13.4	72.9	32.9	57.0	11.7	77.8
45	16.7	32.0	20.2	39.8	60.9	62.5	20.1	56.0	32.4	74.0	76.1	100.0
95	21.2	38.4	19.3	30.8	58.4	61.7	24.7	34.1	22.5	34.6	77.8	82.3
101	18.5	28.3	18.7	33.9	29.6	85.3	23.1	51.9	16.8	61.5	49.3	68.3
104	20.9	33.5	20.5	41.3	35.4	72.4	16.7	23.2	27.3	41.3	88.6	88.5
154	5.8	14.8	20.3	59.3	5.0	13.0	19.8	46.8	30.4	35.6	14.9	16.9
194	21.6	38.0	16.8	35.9	50.6	90.8	31.2	57.0	22.4	49.4	75.9	77.8
270	19.2	34.7	23.2	28.7	30.0	64.8	21.7	44.2	32.5	45.9	35.0	71.3
341	18.7	35.7	24.4	49.2	40.0	53.9	30.0	44.6	44.8	71.1	80.0	53.9
374	20.1	34.8	21.9	51.2	27.4	28.6	48.3	60.2	49.2	42.7	51.7	57.1
LONDON Afternoon Peak (3pm-7pm)												
14	25.4	36.8	14.7	21.5	13.1	24.5	36.3	55.3	18.1	26.0	29.5	31.2
45	32.5	33.0	19.3	29.2	70.1	76.1	37.1	52.8	27.1	34.0	87.6	91.3
95	22.5	32.6	17.9	49.4	77.8	71.8	16.9	32.6	26.8	41.2	97.3	71.8
101	13.6	27.1	18.7	22.8	67.7	73.7	18.2	51.8	28.1	42.7	90.3	92.1
104	14.8	50.8	24.8	39.9	84.0	82.3	44.5	22.6	27.3	21.5	98.0	82.3
154	22.2	40.2	9.6	27.9	10.6	27.2	50.0	80.3	16.7	34.3	11.9	34.4
194	16.0	38.9	23.2	29.1	82.9	88.8	20.1	49.3	23.2	43.6	36.8	98.6
270	46.7	29.4	19.4	25.7	61.5	33.3	57.0	35.3	24.9	32.1	66.2	40.0
341	18.1	47.3	26.8	41.7	55.2	61.7	32.6	52.0	36.8	50.0	92.1	102.9
374	25.9	54.4	20.1	32.2	21.6	35.3	77.8	85.5	42.7	85.9	23.2	58.8
WASHINGTON DC Morning Peak (5am-9am)												
31101	21.9	53.9	18.9	35.9	39.8	76.9	26.2	70.0	37.8	71.8	34.1	92.3
31103	17.3	20.4	58.3	45.1	15.2	14.9	25.9	24.4	87.5	90.3	15.2	19.9
31110	25.0	23.7	35.6	18.2	46.6	62.2	30.0	47.4	47.4	36.4	69.9	77.8
31214	26.1	10.4	51.9	37.5	52.5	18.0	31.3	26.1	64.9	50.0	61.2	35.9
31229	48.6	20.2	52.5	47.7	36.8	32.6	44.5	48.6	70.1	79.5	52.6	70.5
31239	35.9	14.3	36.1	16.4	63.6	38.4	41.9	28.6	39.7	32.7	54.5	67.1
31241	40.2	31.0	34.6	24.5	36.8	48.9	44.2	22.2	25.9	29.4	92.1	97.7
31600	26.9	15.8	47.7	26.4	24.4	35.0	26.9	26.3	63.6	52.8	32.5	35.0
31612	27.8	22.9	46.7	63.7	25.2	14.0	33.3	28.7	93.5	63.7	31.5	35.0
31619	28.3	22.7	56.0	50.0	28.3	32.0	21.2	29.1	84.0	75.0	24.3	36.0
WASHINGTON DC Afternoon Peak (3pm-7pm)												
31101	24.5	47.0	19.7	27.5	24.8	22.8	48.9	67.9	37.0	58.1	44.7	51.2
31103	18.9	42.7	25.2	49.6	79.3	75.6	33.0	42.7	33.6	55.2	92.5	94.5
31110	42.0	14.8	15.4	20.7	95.4	100.0	51.3	34.5	30.8	33.1	47.7	60.0
31214	22.0	17.4	17.6	27.8	88.5	92.3	40.3	30.4	30.2	47.2	88.5	76.9
31229	30.9	22.1	18.7	14.7	36.8	47.6	54.9	29.5	29.1	29.5	57.9	61.2
31239	27.1	34.8	24.0	33.6	35.0	41.2	23.7	29.8	28.0	33.6	52.5	61.7
31241	30.4	26.9	22.0	28.2	73.8	96.0	19.0	34.6	25.7	34.5	110.7	82.3
31600	19.3	33.6	25.3	16.0	64.5	70.0	38.6	56.0	25.3	24.0	82.9	84.0
31612	37.3	38.9	24.8	29.2	42.4	87.5	46.7	46.7	34.8	38.9	74.2	98.5
31619	38.2	49.0	15.9	17.8	25.5	28.9	44.6	70.0	28.6	39.1	31.8	64.9

Table 7.11. ***RMSE (rounded)*** *of peak hours prediction using Machine Learning (BRR for London & RFR for Washington DC) and Naïve Prediction (DA1RefD).*

Station ID	Machine Learning Prediction						Naïve Prediction					
	Pickup		Return		Balance		Pickup		Return		Balance	
	Week1	Week2	Week1	Week2	Week1	Week2	Week1	Week2	Week1	Week2	Week1	Week2
LONDON Morning Peak (5am-9am)												
14	20	56	4	9	17	60	20	84	7	11	15	76
45	5	8	5	7	4	5	6	14	8	13	5	8
95	6	9	6	8	3	3	7	8	7	9	4	4
101	8	12	10	16	3	5	10	22	9	29	5	4
104	10	13	12	20	4	9	8	9	16	20	10	11
154	10	24	2	5	8	20	34	76	3	3	24	26
194	9	14	9	16	6	7	13	21	12	22	9	6
270	8	11	5	5	6	10	9	14	7	8	7	11
341	5	8	6	9	2	3	8	10	11	13	4	3
374	10	11	4	6	9	6	24	19	9	5	17	12
LONDON Afternoon Peak (3pm-7pm)												
14	7	8	22	24	16	22	10	12	27	29	36	28
45	7	5	5	6	4	5	8	8	7	7	5	6
95	8	9	6	12	4	4	6	9	9	10	5	4
101	6	11	8	8	3	4	8	21	12	15	4	5
104	7	18	10	13	6	4	21	8	11	7	7	4
154	4	7	16	39	16	34	9	14	28	48	18	43
194	8	15	14	14	9	9	10	19	14	21	4	10
270	9	5	7	8	13	5	11	6	9	10	14	6
341	5	10	8	10	3	3	9	11	11	12	5	5
374	4	7	16	15	14	12	12	11	34	40	15	20
WASHINGTON DC Morning Peak (5am-9am)												
31101	5	10	1	2	7	10	6	13	2	4	6	12
31103	4	5	2	2	3	3	6	6	3	4	3	4
31110	5	4	6	2	2	4	6	8	8	4	3	5
31214	5	2	4	3	6	2	6	5	5	4	7	4
31229	12	5	3	3	7	6	11	12	4	5	10	13
31239	6	2	10	4	7	4	7	4	11	8	6	7
31241	10	7	8	5	2	3	11	5	6	6	5	6
31600	5	3	3	2	3	4	5	5	4	4	4	4
31612	5	4	1	2	4	2	6	5	2	2	5	5
31619	8	7	2	2	7	8	6	9	3	3	6	9
WASHINGTON DC Afternoon Peak (3pm-7pm)												
31101	5	9	8	9	5	4	10	13	15	19	9	9
31103	4	8	3	9	6	4	7	8	4	10	7	5
31110	9	3	4	5	6	5	11	7	8	8	3	3
31214	6	4	7	10	11	12	11	7	12	17	11	10
31229	9	6	9	6	7	7	16	8	14	12	11	9
31239	8	7	6	6	2	2	7	6	7	6	3	3
31241	8	7	6	9	2	7	5	9	7	11	3	6
31600	4	6	8	4	7	5	8	10	8	6	9	6
31612	4	5	5	6	4	8	5	6	7	8	7	9
31619	6	7	5	5	4	4	7	10	9	11	5	9

7.5. Practical Applications

More accurate prediction of BSS activity at system-wide, cluster or individual station level allows BSS operator to more accurate plan their systems operations. For example, at a particularly busy time for the system as a whole, more human resources can be deployed for bike redistribution to ensure that the system can provide sufficient resources to meet demand (bikes for pickups, empty docking slots for returns). At less busy times, less redistribution can be scheduled, saving human resources and cost.

Predicting higher or lower use in particular clusters, or in individual stations can assist by targeting the best sources and destinations of bikes for redistribution. Additionally, if particular stations are likely to be imbalanced, notifications at those stations on electronic billboards can point users to the closest stations (within a 300 m neighbourhood) which are most likely to have either bikes or docking slots available.

The above analysis has shown that usage in clusters and at individual stations on an hourly basis across the whole day cannot be sufficiently accurately predicted to provide useful operational intelligence. However, taking the example of individual stations, it has been shown that usefully accurate information can be obtained if the total usage across the morning peak and the afternoon peak at busy stations is predicted. Accurate usage for busy stations at busy times will enable BSS operators to better plan bike redistribution twice a day, before each of the peaks.

7.6. Summary and Significance of Results

This chapter has investigated deviation-based prediction referenced from the cyclostationary pattern of BSS usage data. Prediction has been investigated at system-wide, cluster, and station levels using machine learning approaches. There are several significant outcomes in terms of the implementation and performance of these proposed prediction scenarios at each level in comparison with naïve predictions and previously published prediction results.

The RRMSE of different machine learning approaches and different methods for calculating the historical baseline are investigated with a validation dataset. For London, using the historical average for prediction, the best RRMSE is 27.6% using the historical baseline of the last month (HA1Month). This is reduced to 20.3% by using the best historical deviation

prediction which is the one previous hour deviation (DA1RefHr) prediction. Using machine learning to predict deviations gives an RRMSE of 16.9% by using a BRR predictor with the average of the two days as the deviation reference. Similar trends occur in Washington DC. There is an improvement from 23.6% when using the historical average of last month (HA1Month) to 19.0% when using the historical deviation prediction with one previous hour deviation (DA1RefHr) and to 16.7% using an RFR predictor with one previous hour baseline for calculating deviation.

These results show that the machine learning predictors can improve the prediction performance by similar levels for both cities. Here, the single very strong feature identified for machine learning is the one-previous-hour deviation, followed by the two-previous-hour deviation as a strong feature. The effect of weather is already present in the previous hour inputs, and so separate weather inputs do not add much additional prediction information.

These results are compared to the existing works from Giot and Cherrier [47] using similar data from Washington DC, but predicting usage directly (rather than deviation-based prediction). Their smallest RMSE for the next hour prediction is 79 using Ridge Regression. Meanwhile, using similar Ridge Regression, this study achieves an RMSE of 75.5. More importantly, using a Random Forest Regressor gives an RMSE of 38.5, while their Random Forest result is 336. This means that the proposed deviation-based prediction in this study gives considerably better results than attempting to estimate usage directly.

The best scenarios (predictor, reference, and feature) are then applied to a 2 week test set. For week 1 the machine learning predictors achieve an RRMSE of 13.8% in London and 14.1% in Washington DC, and for week 2 approaching the Christmas holidays, 27.5% in London and 22.7% in Washington DC. These test set results show that the machine learning approaches give useful improvements in performance compared to naïve historical-average approaches.

The error in the “normal” week 1 of about 14% is of the same order as the 12% error given by Borgnat et al [63] for Barcelona for hourly number of rented bikes, and much better than the results of Yang et al [16] for Hangzhou for bike check-in and check-out, where their RMSLE of 0.42 to 0.48 corresponds to a relative error that is likely to be greater than 60%. While direct comparisons are not possible in different cities, our results still appear promising.

As a general principle, it is recommended to use the previous hour reference for the cyclostationary-based prediction scenario in BSS as implemented for Washington DC. Its

adjustment to sudden outlier patterns is faster than previous-day and previous-week references. When seasonal fluctuations are evident, the sliding window technique is also recommended. This will enable the training set to be as close as possible in time to the testing set so that the system is more responsive to seasonal changes in demand, however it does need more frequent retraining.

At the level of cluster and station level prediction, the accuracy of hourly usage prediction with machine learning, measured by RRMSE across the whole day is not very good. There are high prediction errors outside of peak hours, such as early morning, midday and at night when the usages are very low. For example, if a station receives 1 pickup but the predictor predicts 2, this gives 100% RRMSE.

Therefore, morning and afternoon peak times prediction is then proposed for busy stations. This gives much better results. From an operational viewpoint, accurate prediction of busy stations at busy times is most useful, since this can give an estimation of potential upcoming imbalances before peak hours occur so that proper redistribution can be done if high usage is predicted, or redistribution costs can be saved when predicted use is low.

CHAPTER 8
CONCLUSIONS, CONTRIBUTIONS AND FUTURE WORK

8.1. Conclusions

Intensive data analyses have been undertaken to investigate aspects of mobility dynamics that are buried in BSS data in order to answer the four research questions as proposed in Chapter 3. The answers to each research question are given below.

8.1.1. RQ1 (What insights can be gained from the BSS stations neighbourhood ties?)

Through intensive spatial analysis in Chapter 5, several useful insights have been gained.

Spatial motifs can give insights into common daily mobility patterns. Since conventional unlabeled motifs as used in previous studies have some ambiguous interpretations, a labelled motif notation has been developed that ensures that patterns are uniquely interpreted. Based on 0.5% threshold from all possible motifs that may appear, the 10 common spatial network patterns can be considered as motifs in BSS mobility. In addition, distance analysis for certain motifs reveals that users typically select a station from within a 300 m neighbourhood of their origin or destination.

In terms of distances between stations, this study introduces the notion of the waypoint distance for use in BSS studies, which gives a better shortest route representation between OD stations than the widely-used Euclidean distance. This waypoint distance also gives a more accurate distance than using the Manhattan distance, which is highly dependent on the choice of XY axes. When studying the impact distance of a station shutdown, waypoint distance also gives a more reliable measure for determining affected nearby stations. Results show that there is typically increased usage of 20% to 80 % *before-to-during* shutdown for nearby stations less than 300 m from the shutdown station. Conversely, a similar percentage of decreased usage occurs *during-to-after* shutdown for those 300 m nearby stations.

The results from these two different approaches suggest that there is a strong relationship between disturbances at one station and other nearby stations within 300 m waypoint. These disturbances could be a temporary shutdown or station imbalance. If a station is unavailable, users tend look for alternate stations within 300 m. So this distance should be considered when BSS station locations are chosen during system design.

8.1.2. RQ2 (To what extent can clustering identify highly predictable users, what are the maximum limits of predictability, and how can these be achieved?)

Chapter 6 has investigated how users can be clustered using their temporal characteristics and then labelled after observing their mobility behaviour using various spatiotemporal metrics. These clusters identify users with similar usage characteristics, especially those who are highly predictable. While one study [74] used various temporal features, this study uses two different feature sets for clustering: total trips (1 feature) and the number of hourly trips across the day (24 features). This can adequately reflect the trip frequency and the trip regularity of the users. Results show the distinct spatiotemporal characteristics of the proposed clusters which are labelled as *casual users, regular users,* and *commuters*.

Casual users show seasonal and recreational traits and *they* have relatively few trips and short waiting times. They are more active on weekends and weekday afternoons, strongly affected by season (decreasing as winter approaches) and they are the slowest riders, suggesting sightseeing rather than simple transport. ***Commuters*** show resilient commuting patterns where weekday usages are much more than weekend usage, and usage is less affected by the season. They have two weekday usage peaks in mornings and afternoons, and also a high proportion of waiting times that correspond with daily working hours. They are the fastest riders and demonstrate a RoG skewed towards small distances showing a characteristic distribution of length scale. Meanwhile, ***regular users*** have traits that are a mixture of commuters and casual users.

Casual users have irregular histograms for all types of entropy. This suggests that a large number of users with few trips results in a small number of discrete entropy values [46]. Instead, the entropy distribution of regular users and commuters are smoother, showing a typical form for these histograms. The entropy of commuters follows the basic entropy ordering rule: $\mathbf{S^{Rand} \geq S^{Shan} \geq S^{Cond} \geq S^{Real}}$, while for regular users only the hourly-based cluster closely follows this rule. This suggests that hourly-based clustering is better than subscription and total trips based clustering for identifying homogeneous user groups.

Since real entropy is close to the conditional entropy for commuters, this suggests that entropy is strongly determined by the sequence of recently visited stations. This indicates that the trip data has *Markovian traits* where the actual predictability can be represented by the conditional one [46]. In other words, the next location of a BSS commuter is dependent on the one previously visited location, and it does not depend on locations further in the past.

The predictability results show that a Markov model predictor for commuters' next locations has an upper bound of 80% for prediction accuracy, and that about 20% of next locations are effectively random and unpredictable. Predictability of BSS users in this study is close to the predictability using mobile phone data conducted by Qin et al [44]. However, they did not continue their work to the prediction to show whether their predictability results can be achieved. Meanwhile, Lu et al. [25] implemented a Markov Chain (MC) model to conduct prediction, and they could achieved an accuracy similar to their predictability level using a first order Markov Chain model.

Using a first order Markov Chain predictor, prediction accuracy is better for commuters than regular and casual users, and cluster by-hourly-trips gives better prediction accuracy than by-total-trips. In pickup-to-return prediction, using the ensemble of a first order Markov Model with peak time OD matrix and with commuter collective trends for trips that are not in a user's history, a prediction accuracy is achieved which corresponds to the predictability bound of 80% during the morning peak when commuters are dominant in the system. Similar approaches are implemented for return-to-pickup prediction, but their accuracy is less than the pickup-to-return. Other techniques like the second order MM and the daily matrix do not improve the accuracy. The results of both return and pickup predictions above show that the correlation of the ***pickup-ride-return*** is stronger than the ***return-waiting-pickup***. This fact suggests that once people pickup bikes especially in the morning peak of weekdays they are likely more predictable with their intended destination. Potential uses of this prediction are discussed under RQ4 below.

8.1.3. RQ3 (To what extent can the cyclostationary pattern of bicycle sharing systems be used to conduct and improve the prediction of BSS usage and which factors are most effective for good prediction?)

Chapter 7 describes a deviation-based prediction method using cyclostationary patterns of BSS data. Predictors are implemented at system-wide, cluster, and station levels using machine learning approaches. Results suggest that the deviation-based prediction using machine learning predictors can improve the prediction performance in comparison to naïve approaches based on recent historical averages. Results are significant better than results in previously published studies [47]. The best RRMSE that can be achieved by different ML techniques are 16.9% in London using BRR and 16.7% in Washington DC using RFR in a validation set. Using these ML techniques, good results are achieved in two weeks of testing data - 13.8 %

and 14.1% in week 1, and 27.5 % and 22.7% in an anomalous week 2. Using similar data from Washington DC, Giot and Cherrier [47] achieved a smallest RMSE for the next hour prediction of 79 by Ridge Regression. Meanwhile, using similar Ridge Regression, this study achieves RMSE 75.5. The best ML predictor in this study has an RMSE of 38.5 using Random Forest, compared to Giot & Cherrier's Random Forest results of 336. This shows that the proposed deviation-based prediction in this study is a significant improvement over previous BSS prediction methods.

It is also found that the very strong feature for ML prediction is the one-previous-hour deviation, followed by the two-previous-hour deviation as a strong feature. The effect of weather is already present in the previous hour inputs, and so separate weather inputs do not add much additional prediction information. Therefore, it is recommended to use the previous hour reference for the deviation-based prediction in BSS. Its adjustment to sudden outlier patterns is faster than previous-day and previous-week references. Furthermore, when the seasonal fluctuations are evident, the sliding window technique should be implemented. This will enable the new training sets to adapt the predictor quickly to seasonal changes, but it does require frequent retraining.

Comparing system-wide, cluster, and station level prediction, the results show good prediction accuracy at the system-wide level. The inner-city clusters and stations are better predicted than the outer ones as they tend to receive more pickups and returns which make their cyclostationary patterns more constant. Hourly-based cluster and station predictions do not give useful prediction accuracy, however, predicting usage for the busiest stations at the busiest times does give useful information for optimizing BSS operations.

8.1.4. RQ4 (What do the station neighbourhood ties and high predictable clusters knowledge, as well as the system-wide predictions at different levels, bring to the BSS deployment, services, and operations?)

Results from Chapters 5, 6, and 7 all have potential impact on BSS systems operations. This section collects these potential uses in order to answer RQ4.

A waypoint distance of 300 m is found to be the distance that users will travel to an alternate station in the same neighbourhood. This knowledge can be used by BSS operators when a station is temporarily shutdown by focusing the availability of resources for nearby stations within 300 m waypoint distance. Combined with the average usage data, another

possible application is to identify the ineffective stations in the network that can be eliminated because its deletion will not significantly impact system availability. Yet another possible application is to identify isolated stations with high usage where a new nearby station within 300 m is recommended. This new station is intended as a backup if the main station is shut down or imbalanced. For a new BSS in new cities, this 300 m waypoint distance between stations can be used as a planning guideline to avoid isolated stations and ineffective station locations.

The knowledge from user clustering and next place prediction can be used to identify the likely destination of predictable users. The most predictable users are identified as commuters during the morning peak. One possible application is a user-specific notification system that can proactively notify highly predictable users. For example, if their predicted destination station will be shut down or full or empty at particular times ahead, or if there are delays on the route, a notification can be sent automatically when the user starts a trip. The notification can include the possible nearby stations, routes, or time of travel. This might, for example, request users to return their bicycle to a particular station in a neighbourhood (within 300 m, as indicated in the station neighbourhood analysis result) which is currently almost empty to assist with user-based station rebalancing. This *user-based notification* will make the system more efficient, and it can complement the existing journey advisor systems which are not user specific.

More accurate prediction of BSS activity at system-wide, cluster or individual station level allows BSS operators to more accurate plan their systems operations. For example, at a particularly busy time for the system as a whole, more human resources can be deployed for bike redistribution to ensure that the system can provide sufficient resources to meet demand (bikes for pickups, empty docking slots for returns). At less busy times, less redistribution can be scheduled, saving human resources and cost.

Predicting higher or lower use in particular clusters, or in individual stations can assist by targeting the best sources and destinations of bikes for redistribution. Additionally, if particular stations are likely to be imbalanced, notifications at those stations on electronic billboards can point users to the closest stations (within a 300 m neighbourhood) which are most likely to have either bikes or docking slots available.

The analysis in this study has shown that usage in clusters and at individual stations on an hourly basis across the whole day cannot be sufficiently accurately predicted to provide useful

operational intelligence. However, taking the example of individual stations, it has been shown that usefully accurate information can be obtained if the total usage across the morning peak and the afternoon peak at busy stations is predicted when the system is highly driven by commuters at that time. Accurate usage for busy stations at busy times will enable BSS operators to better plan bike redistribution twice a day, before each of the peaks, and will also enable better scheduling of appropriate human resources for these redistribution activities.

8.2. Original Contributions

This thesis makes the following contributions.

The first contribution has been in identifying the neighbourhood in which BSS stations affect each other. The analysis of trip behaviour has identified typical BSS mobility motifs, and in turn analysis of these motifs has identified 300 m as the typical neighbourhood of a station. Furthermore, it has been shown that 300 m waypoint distance is a better measure of neighbourhood than measures using Euclidean distance or Manhattan distance. Analysis of station shutdowns has also confirmed the neighbourhood of a station as 300 m waypoint distance. This distance has implications for BSS design and operations.

The second contribution is the identification of highly-predictable BSS users. Using novel temporal clustering features, a highly predictable class of users is identified, referred to as commuters. Using information theory of entropy and predictability, a first order Markov Chain predictor is proposed, which combines individual and system-wide information. It is shown that during peak times, this predictor has a prediction accuracy close to the theoretical predictability of 80%. This information could enable user-specific notifications to improve system efficiency.

The third contribution is in the area of system usage. At system-level, a predictor based on deviations from historical average usage patterns, and using machine learning prediction techniques is used to predict total system usage on an hourly basis. It is compared to predictors based on historical averages of use, historical deviations in use, and machine learning approaches by other researchers based on direct prediction of system use. The new predictor is shown to have significantly better prediction accuracy. The usage estimation at station level also identify that useful predictions can be made with the machine learning approaches for the busiest stations at the busiest times. System prediction can aid with better planning of bike redistribution.

8.3. Future Work

This mobility analysis has been undertaken with a limited time span of BSS data (London 2012), which is the only publically available data that we found with individual user identification. A longer data set over several years would allow a more detailed analysis of usage variations over the whole year, and also investigate the year-to-year trends in system usage.

There is also no publically available information on the methods that BSS operators use to rebalance bike stations, so the direct application of this mobility analysis to system rebalancing has not been possible. The application of the insights from this work to real BSS operations would also be a useful future direction. The first possibility is to investigate whether the types of user-specific notifications suggested do indeed improved system efficiency and user satisfaction. Second, if the prediction of system usage can be done for several hours ahead, then the station level prediction can then be extended to assist operators to plan their redistributions as efficiently as possible.

REFERENCES

[1] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in *Proceedings of the 13th international conference on Ubiquitous computing*, pp. 89-98, 2011, DOI: 10.1145/1921591.1921596.

[2] R. Kitamura, C. Chen, R. M. Pendyala, and R. Narayanan, "Micro-simulation of daily activity-travel patterns for travel demand forecasting," *Transportation*, vol. 27, pp. 25-51, 2000, DOI: 10.1023/A:1005259324588.

[3] L. Moreira-Matias, O. Cats, J. Gama, J. Mendes-Moreira, and J. F. de Sousa, "An online learning approach to eliminate Bus Bunching in real-time," *Applied Soft Computing*, vol. 47, pp. 460-482, 2016, DOI: 10.1016/j.asoc.2016.06.031.

[4] D. W. MacPherson and B. D. Gushulak, "Human mobility and population health: new approaches in a globalizing world," *Perspectives in Biology and Medicine*, vol. 44, pp. 390-401, 2001, DOI: 10.1353/pbm.2001.0053.

[5] A. Wesolowski, "Quantifying Human Movement Patterns for Public Health," Carnegie Mellon University, 2014. Available: http://repository.cmu.edu/dissertations/329/

[6] V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, and A. Vespignani, "Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions," *PLoS medicine*, vol. 4, p. e13, 2007, DOI: 10.1371/journal.pmed.0040013.

[7] S. Riley, "Large-scale spatial-transmission models of infectious disease," *Science*, vol. 316, pp. 1298-1301, 2007, DOI: 10.1126/science.1134695.

[8] Q. Wang and J. E. Taylor, "Quantifying human mobility perturbation and resilience in Hurricane Sandy," *PLoS one*, vol. 9, p. e112608, 2014, DOI: 10.1371/journal.pone.0112608.

[9] 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*, vol. 9, p. e97010, 2014, DOI: 10.1371/journal.pone.0097010.

[10] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, "A tale of many cities: universal patterns in human urban mobility," *PloS one*, vol. 7, p. e37027, 2012, DOI: 10.1371/journal.pone.0037027.

[11] J. Froehlich, J. Neumann, and N. Oliver, "Sensing and Predicting the Pulse of the City through Shared Bicycling," in *International Joint Conference on Artificial Intelligence IJCAI*, pp. 1420-1426, 2009, http://www.nuriaoliver.com/bicing/IJCAI09_Bicing.pdf.

[12] R. Jurdak, "The impact of cost and network topology on urban mobility: A study of public bicycle usage in 2 US cities," *PloS one*, vol. 8, p. e79396, 2013, DOI: 10.1371/journal.pone.0079396.

[13] A. Faghih-Imani, S. Anowar, E. J. Miller, and N. Eluru, "Hail a cab or ride a bike? A travel time comparison of taxi and bicycle-sharing systems in New York City," *Transportation Research Part A: Policy and Practice*, vol. 101, pp. 11-21, 2017, DOI: 10.1016/j.tra.2017.05.006.

[14] B. Chen, F. Pinelli, M. Sinn, A. Botea, and F. Calabrese, "Uncertainty in urban mobility: Predicting waiting times for shared bicycles and parking lots," in *the 16th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pp. 53-58, 2013, DOI: 10.1109/ITSC.2013.6728210

[15] H. Xu, J. Ying, H. Wu, and F. Lin, "Public bicycle traffic flow prediction based on a hybrid model," *Applied Mathematics and Information Sciences*, vol. 7, pp. 667-674, 2013, DOI: 10.12785/amis/070234.

[16] Z. Yang, J. Hu, Y. Shu, P. Cheng, J. Chen, and T. Moscibroda, "Mobility Modeling and Prediction in Bike-Sharing Systems," in *the 14th ACM Annual International Conference on*

Mobile Systems, Applications, and Services, pp. 165-178, 2016, DOI: 10.1145/2906388.2906408.

[17] Y. Li, Y. Zheng, H. Zhang, and L. Chen, "Traffic prediction in a bike-sharing system," in *the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, p. 33, 2015, DOI: 10.1145/2820783.2820837.

[18] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabási, "Returners and explorers dichotomy in human mobility," *Nature communications*, vol. 6, 2015, DOI: 10.1038/ncomms9166.

[19] L. Pappalardo and F. Simini, "Modelling individual routines and spatio-temporal trajectories in human mobility," *arXiv preprint arXiv:1607.05952*, 2016, http://arxiv.org/pdf/1607.05952.pdf.

[20] A. Hess, K. A. Hummel, W. N. Gansterer, and G. Haring, "Data-driven human mobility modeling: a survey and engineering guidance for mobile networking," *ACM Computing Surveys (CSUR)*, vol. 48, p. 38, 2016, DOI: 10.1145/2840722.

[21] D. Brockmann, "Following the money," *Physics World*, vol. 23, p. 31, 2010, http://iopscience.iop.org/article/10.1088/2058-7058/23/02/37/pdf.

[22] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, pp. 462-465, 2006, DOI: 10.1038/nature04292.

[23] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779-782, 2008, DOI: 10.1038/nature07850.

[24] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, pp. 1018-1021, 2010, DOI: 10.1126/science.1177170.

[25] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific reports*, vol. 3, 2013, DOI: 10.1038/srep02923.

[26] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, pp. 818-823, 2010, DOI: 10.1038/nphys1760.

[27] A. K. Datta, "Predicting bike-share usage patterns with machine learning," Master Thesis, Department of Informatics, University of Oslo, 2014. Available: http://www.duo.uio.no/handle/10852/42177

[28] 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," *Frontiers of Computer Science*, vol. 6, pp. 111-121, 2012, DOI: 10.1007/s11704-011-1192-6.

[29] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting taxi–passenger demand using streaming data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, pp. 1393-1402, 2013, DOI: 10.1109/TITS.2013.2262376.

[30] N. Lathia, C. Smith, J. Froehlich, and L. Capra, "Individuals among commuters: Building personalised transport information services from fare collection systems," *Pervasive and Mobile Computing*, vol. 9, pp. 643-664, 2013, DOI: 10.1016/j.pmcj.2012.10.007.

[31] M. Padgham, "Human movement is both diffusive and directed," *PloS one*, vol. 7, p. e37754, 2012, DOI: 10.1371/journal.pone.0037754.

[32] J. Khiari, L. Moreira-Matias, V. Cerqueira, and O. Cats, "Automated setting of Bus schedule coverage using unsupervised machine learning," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 552-564, 2016, DOI: 10.1007/978-3-319-31753-3_44.

[33] M. Zignani and S. Gaito, "Extracting human mobility patterns from gps-based traces," in *IEEE Wireless Days (WD), International Federation for Information Processing (IFIP)*, pp. 1-5, 2010, DOI: 10.1109/WD.2010.5657695.

[34] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth, "Understanding human mobility from Twitter," *PloS one*, vol. 10, p. e0131469, 2015, DOI: 10.1371/journal.pone.0131469.

[35] S. Hasan, C. M. Schneider, S. V. Ukkusuri, and M. C. González, "Spatiotemporal patterns of urban human mobility," *Journal of Statistical Physics*, vol. 151, pp. 304-318, 2013, DOI: 10.1007/s10955-012-0645-0.

[36] C. Kang, Y. Liu, Y. Mei, and L. Xu, "Evaluating the representativeness of mobile positioning data for human mobility patterns," *GIScience, Columbus*, 2012, http://www.giscience.org/past/2012/proceedings/abstracts/giscience2012_paper_111.pdf.

[37] H. Barbosa, F. B. de Lima-Neto, A. Evsukoff, and R. Menezes, "The effect of recency to human mobility," *EPJ Data Science*, vol. 4, p. 21, 2015, DOI: 10.1140/epjds/s13688-015-0059-8.

[38] Y. Xiao-Yong, H. Xiao-Pu, Z. Tao, and W. Bing-Hong, "Exact solution of the gyration radius of an individual's trajectory for a simplified human regular mobility model," *Chinese Physics Letters*, vol. 28, p. 120506, 2011, DOI: 10.1088/0256-307X/28/12/120506.

[39] S. D. Parkes, G. Marsden, S. A. Shaheen, and A. P. Cohen, "Understanding the diffusion of public bikesharing systems: evidence from Europe and North America," *Journal of Transport Geography*, vol. 31, pp. 94-103, 2013, http://innovativemobility.org/wp-content/uploads/2015/07/Diffusion-of-Public-Bikesharing-Systems.pdf.

[40] C. M. Schneider, V. Belik, T. Couronné , Z. Smoreda, and M. Gonzalez, "Unraveling daily human mobility motifs," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* pp. xxxviii-xxxviii, 2013, DOI: 10.1098/rsif.2013.0246.

[41] S. Jiang, J. Ferreira, and M. C. Gonzales, "Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore," *IEEE Transactions on Big Data*, 2016, DOI: 10.1109/TBDATA.2016.2631141.

[42] R. Gallotti, A. Bazzani, M. Degli Esposti, and S. Rambaldi, "Entropic measures of individual mobility patterns," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2013, p. P10022, 2013, DOI: 10.1088/1742-5468/2013/10/P10022.

[43] P. Baumann and S. Santini, "On the use of instantaneous entropy to measure the momentary predictability of human mobility," in *the 14th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 535-539, 2013, DOI: 10.1109/SPAWC.2013.6612107.

[44] S.-M. Qin, H. Verkasalo, M. Mohtaschemi, T. Hartonen, and M. Alava, "Patterns, entropy, and predictability of human mobility and life," *PloS one*, vol. 7, p. e51353, 2012, DOI: 10.1371/journal.pone.0051353.

[45] R. Sinatra and M. Szell, "Entropy and the predictability of online life," *Entropy*, vol. 16, pp. 543-556, 2014, DOI: 10.3390/e16010543.

[46] I. B. I. Purnama, N. Bergmann, R. Jurdak, and K. Zhao, "Characterising and Predicting Urban Mobility Dynamics by Mining Bike Sharing System Data," in *the 11th IEEE International Conference on Ubiquitous Intelligence and Computing (UIC)*, pp. 159-167, 2015, DOI: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.46.

[47] R. Giot and R. Cherrier, "Predicting bikeshare system usage up to one day ahead," in *IEEE symposium on Computational intelligence in vehicles and transportation systems (CIVTS)*, pp. 22-29, 2014, DOI: 10.1109/CIVTS.2014.7009473.

[48] A. P. Masucci, J. Serras, A. Johansson, and M. Batty, "Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows," *Physical Review E*, vol. 88, p. 022812, 2013, DOI: 10.1103/PhysRevE.88.022812.

[49] Y. Ren, M. Ercsey-Ravasz, P. Wang, M. C. González, and Z. Toroczkai, "Predicting commuter flows in spatial networks using a radiation model based on temporal ranges," *Nature Communications 5*, 2014, DOI: 10.1038/ncomms6347.

[50] A. Asahara, K. Maruyama, A. Sato, and K. Seto, "Pedestrian-movement prediction based on mixed Markov-chain model," in *the 19th ACM SIGSPATIAL international conference on advances in geographic information systems*, pp. 25-33, 2011, DOI: 10.1145/2093973.2093979.

[51] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez, "Next place prediction using mobility markov chains," in *the First Workshop on Measurement, Privacy, and Mobility*, p. 3, 2012, DOI: 10.1145/2181196.2181199.

[52] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 308-324, 2015, DOI: 10.1016/j.trc.2015.02.019.

[53] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144, 2016, DOI: 10.1145/2939672.2939778.

[54] A. J. Lopez Aguirre, I. Semanjski, and S. Gautama, "Forecasting travel behaviour from crowdsourced data with machine learning based model," in *the Fifth International Conference on Data Analytics*, pp. 93-99, 2016, http://biblio.ugent.be/publication/8116649.

[55] P. Baumann, W. Kleiminger, and S. Santini, "The influence of temporal and spatial features on the performance of next-place prediction algorithms," in *ACM international joint conference on Pervasive and ubiquitous computing*, pp. 449-458, 2013, DOI: 10.1145/2493432.2493467.

[56] L. Chen, D. Yang, J. Jakubowicz, G. Pan, D. Zhang, and S. Li, "Sensing the pulse of urban activity centers leveraging bike sharing open data," in *the 11th IEEE International Conference on Ubiquitous Intelligence and Computing (UIC)*, pp. 135-142, 2015, DOI: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.43.

[57] S. Shaheen, S. Guzman, and H. Zhang, "Bikesharing in Europe, the Americas, and Asia: past, present, and future," *Transportation Research Record: Journal of the Transportation Research Board*, pp. 159-167, 2010, DOI: 10.3141/2143-20.

[58] J. Larsen, "*Bike-sharing programs hit the streets in over 500 cities worldwide,*" 2013, http://www.earth-policy.org/plan_b_updates/2013/update112, Access Date: 15/2/2014.

[59] S. A. Shaheen, E. W. Martin, A. P. Cohen, N. D. Chan, and M. Pogodzinsk, "Public Bikesharing in North America During a Period of Rapid Expansion: Understanding Business Models, Industry Trends & User Impacts," *Mineta Transportation Institute (MTI) Report 12-29*, 2014, http://transweb.sjsu.edu/PDFs/research/1131-public-bikesharing-business-models-trends-impacts.pdf.

[60] C. J. O'brien O, Batty M, "Mining bicycle sharing data for generating insights into sustainable transport systems," *Journal of Transport Geography*, vol. 34, pp. 262-273, 2014, DOI: 10.1016/j.jtrangeo.2013.06.007.

[61] J. Corcoran, T. Li, D. Rohde, E. Charles-Edwards, and D. Mateo-Babiano, "Spatio-temporal patterns of a Public Bicycle Sharing Program: the effect of weather and calendar events," *Journal of Transport Geography*, vol. 41, pp. 292-305, 2014, DOI: 10.1016/j.jtrangeo.2014.09.003.

[62] A. Patil, K. Musale, and B. P. Rao, "Bike share demand prediction using Random Forest," *International Journal of Innovative Science*, vol. 2, 2015, http://ijiset.com/vol2/v2s4/IJISET_V2_I4_195.pdf.

[63] P. Borgnat, P. Abry, P. Flandrin, C. Robardet, J.-B. Rouquier, and E. Fleury, "Shared bicycles in a city: A signal processing and data analysis perspective," *Advances in Complex Systems*, vol. 14, pp. 415-438, 2011, DOI: 10.1142/S0219525911002950.

[64] J.-H. Lin and T.-C. Chou, "A geo-aware and VRP-based public bicycle redistribution system," *International Journal of Vehicular Technology*, 2012, DOI: 10.1155/2012/963427.

[65] P. Borgnat, P. Abry, P. Flandrin, and J.-B. Rouquier, "Studying Lyon's Vélo'v: a statistical cyclic model," in *European Conference on Complex Systems*, 2009, http://hal.univ-grenoble-alpes.fr/ensl-00408147/document.

[66] R. Regue and W. Recker, "Proactive vehicle routing with inferred demand to solve the bikesharing rebalancing problem," *Transportation Research Part E: Logistics and Transportation Review*, vol. 72, pp. 192-209, 2014, DOI: 10.1016/j.tre.2014.10.005.

[67] J. Schuijbroek, R. Hampshire, and W.-J. van Hoeve, "Inventory rebalancing and vehicle routing in bike sharing systems," *European Journal of Operation Research*, 2013, DOI: 10.1016/j.ejor.2016.08.

[68] A. Singla, M. Santoni, G. Bartók, P. Mukerji, M. Meenen, and A. Krause, "Incentivizing Users for Balancing Bike Sharing Systems," in *Association for the Advancement of Artificial Intelligence*, pp. 723-729, 2015, http://las.inf.ethz.ch/files/singla15incentivizing.pdf.

[69] J. Pfrommer, J. Warrington, G. Schildbach, and M. Morari, "Dynamic vehicle redistribution and online price incentives in shared mobility systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, pp. 1567-1578, 2014, DOI: 10.1109/TITS.2014.2303986.

[70] 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*, vol. 6, pp. 455-466, 2010, DOI: 10.1016/j.pmcj.2010.07.002.

[71] R. Beecham, J. Wood, and A. Bowerman, "Studying commuting behaviours using collaborative visual analytics," *Computers, Environment and Urban Systems*, vol. 47, pp. 5-15, 2014, DOI: 10.1016/j.compenvurbsys.2013.10.007.

[72] N. Lathia, S. Ahmed, and L. Capra, "Measuring the impact of opening the London shared bicycle scheme to casual users," *Transportation research part C: emerging technologies*, vol. 22, pp. 88-102, 2012, DOI: 10.1016/j.trc.2011.12.004.

[73] J. Zhang, X. Pan, M. Li, and S. Y. Philip, "Bicycle-sharing system analysis and trip prediction," in *the 17th IEEE International Conference on Mobile Data Management (MDM)*, pp. 174-179, 2016, DOI: 10.1109/MDM.2016.35.

[74] M. Vogel, R. Hamon, G. Lozenguez, L. Merchez, P. Abry, J. Barnier, P. Borgnat, P. Flandrin, I. Mallon, *et al.*, "From bicycle sharing system movements to users: a typology of Vélo'v cyclists in Lyon based on large-scale behavioural dataset," *Journal of Transport Geography*, vol. 41, pp. 280-291, 2014, DOI: 10.1016/j.jtrangeo.2014.07.005.

[75] O. O'brien, J. Cheshire, and M. Batty, "Mining bicycle sharing data for generating insights into sustainable transport systems," *Journal of Transport Geography*, vol. 34, pp. 262-273, 2014, DOI: 10.1016/j.jtrangeo.2013.06.007.

[76] M. Z. Austwick, O. O'brien, E. Strano, and M. Viana, "The structure of spatial networks and communities in bicycle sharing systems," *PloS one*, vol. 8, p. e74685, 2013, DOI: 10.1371/journal.pone.0074685.

[77] J. Wood, A. Slingsby, and J. Dykes, "Visualizing the dynamics of London's bicycle-hire scheme," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 46, pp. 239-251, 2011, DOI: 10.3138/carto.46.4.239.

[78] P. Aeschbach, X. Zhang, A. Georghiou, and J. Lygeros, "Balancing bike sharing systems through customer cooperation-a case study on London's Barclays Cycle Hire," in *the 54th IEEE Annual Conference on Decision and Control (CDC)*, pp. 4722-4727, 2015, DOI: 10.1109/CDC.2015.7402955.

[79] V. Ciancia, D. Latella, M. Massink, and R. Pakauskas, "Exploring spatio-temporal properties of bike-sharing systems," in *IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshops (SASOW)* pp. 74-79, 2015, DOI: 10.1109/SASOW.2015.17.

[80] A. Bargar, A. Gupta, S. Gupta, and D. Ma, "Interactive visual analytics for multi-city bikeshare data analysis," in *the 3rd International Workshop on Urban Computing (UrbComp)*, 2014, http://www2.cs.uic.edu/~urbcomp2013/urbcomp2014/papers/Bargar_Bikesharing.pdf.

[81] L. Chen, D. Zhang, L. Wang, D. Yang, X. Ma, S. Li, Z. Wu, G. Pan, T.-M.-T. Nguyen, *et al.*, "Dynamic cluster-based over-demand prediction in bike sharing systems," in *the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 841-852, 2016, DOI: 10.1145/2971648.2971652.

[82] K. Gebhart and R. B. Noland, "The impact of weather conditions on bikeshare trips in Washington, DC," *Transportation*, vol. 41, pp. 1205-1225, 2014, DOI: 10.1007/s11116-014-9540-7.

[83] M. Ahillen, D. Mateo-Babiano, and J. Corcoran, "Dynamics of bike sharing in Washington, DC and Brisbane, Australia: Implications for policy and planning," *International journal of sustainable transportation*, vol. 10, pp. 441-454, 2016, DOI: 10.1080/15568318.2014.966933.

[84] R. Rixey, "Station-level forecasting of bikesharing ridership: Station Network Effects in Three US Systems," *Transportation Research Record: Journal of the Transportation Research Board*, pp. 46-55, 2013, DOI: 10.3141/2387-06.

[85] M. Lin, "Application of machine learning techniques to forecast bike rental demand in the capital bikeshare program in Washington, D.C.," *Final Project, UC Berkeley*, 2015, DOI: 10.13140/RG.2.1.1433.7766.

[86] E. Côme, A. Randriamanamihaga, and L. Oukhellou, "Spatio-temporal usage pattern analysis of the Paris Shared Bicycle Scheme: a data mining approach," in *Transport Research Arena (TRA) 5th Conference: Transport Solutions from Research to Deployment*, 2014, http://trid.trb.org/view.aspx?id=1320228.

[87] A. N. Randriamanamihaga, E. Côme, L. Oukhellou, and G. Govaert, "Clustering the Vélib'origin-destinations flows by means of Poisson mixture models," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2013, https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2013-95.pdf.

[88] C. Etienne and O. Latifa, "Model-based count series clustering for bike sharing system usage mining: a case study with the Vélib'system of Paris," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, p. 39, 2014, DOI: 10.1145/2560188.

[89] N. Bonnotte, R. Cherrier, R. Delassus, and Y. Alouini, "Real-time data analytics and optimization of shared-vehicles networks," in *the 22nd Intelligent Transport Systems (ITS) World Congress*, 2015, http://www.qucit.com/wp-content/uploads/sites/3/2016/05/201510_Final-Paper_Qucit_ITSBIkes.pdf.

[90] Y. Chabchoub and C. Fricker, "Classification of the vélib stations using Kmeans, Dynamic Time Wraping and DBA averaging method," in *the International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)*, pp. 1-5, 2014, DOI: 10.1109/IWCIM.2014.7008802.

[91] Y. Han, E. Côme, and L. Oukhellou, "Towards bicycle demand prediction of large-scale bicycle sharing system," in *the 93rd Transportation Research Board Annual Meeting*, 2014, http://www.comeetie.fr/pdfrepos/TRBYUFEI.pdf.

[92] R. Nair, E. Miller-Hooks, R. C. Hampshire, and A. Bušić, "Large-scale vehicle sharing systems: analysis of Vélib'," *International Journal of Sustainable Transportation*, vol. 7, pp. 85-106, 2013, DOI: 10.1080/15568318.2012.

[93] E. Côme, N. A. Randriamanamihaga, L. Oukhellou, and P. Aknin, "Spatio-temporal Analysis of Dynamic Origin-Destination Data Using Latent Dirichlet Allocation: Application to Vélib'Bike Sharing System of Paris," in *the 93rd Transportation Research Board Annual meeting*, p. 19p, 2014, http://hal.archives-ouvertes.fr/hal-01052951.

[94] N. Gast, G. Massonnet, D. Reijsbergen, and M. Tribastone, "Probabilistic forecasts of bike-sharing systems for journey planning," in *the 24th ACM International on Conference on Information and Knowledge Management*, pp. 703-712, 2015, DOI: 10.1145/2806416.2806569.

[95] E. O'Mahony and D. B. Shmoys, "Data Analysis and Optimization for (Citi) Bike Sharing," in *the 29th AAAI Conference on Artificial Intelligence*, pp. 687-694, 2015, http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9698/9314.

[96] D. Singhvi, S. Singhvi, P. I. Frazier, S. G. Henderson, E. O'Mahony, D. B. Shmoys, and D. B. Woodard, "Predicting bike usage for new york city's bike sharing system," in *AAAI Workshop on Computational Sustainability*, 2015, http://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/viewFile/10115/10185.

[97] M. Zeng, T. Yu, X. Wang, V. Su, L. T. Nguyen, and O. J. Mengshoel, "Improving Demand Prediction in Bike Sharing System by Learning Global Features," in *Knowledge Discovery (KDD) Conference on Machine Learning for Large Scale Transportation Systems (LSTS)*, 2016, DOI: 10.1145/1235.

[98] X. Zhou, "Understanding spatiotemporal patterns of biking behavior by analyzing massive bike sharing data in Chicago," *PloS one*, vol. 10, p. e0137922, 2015, DOI: 10.1371/journal.pone.0137922.

[99] A. Faghih-Imani and N. Eluru, "Analyzing Destination Choice Preferences in Bicycle-Sharing Systems: Investigation of Chicago's Divvy System," in *Transportation Research Board (TRB) Annual Meeting, Washington DC*, 2015, http://docs.trb.org/prp/15-2959.pdf.

[100] P. Jensen, J.-B. Rouquier, N. Ovtracht, and C. Robardet, "Characterizing the speed and paths of shared bicycle use in Lyon," *Transportation research part D: transport and environment*, vol. 15, pp. 522-524, 2010, DOI: 10.1016/j.trd.2010.07.002.

[101] E. Crisostomi, M. Faizrahnemoon, A. Schlote, and R. Shorten, "A Markov-chain based model for a bike-sharing system," in *IEEE International Conference on Connected Vehicles and Expo (ICCVE)*, pp. 367-372, 2015, DOI: 10.1109/ICCVE.2015.12.

[102] R. C. Hampshire and L. Marla, "An analysis of bike sharing usage: Explaining trip generation and attraction from observed demand," in *the 91st Transportation Research Board Annual Meeting*, pp. 12-2099, 2012, http://nacto.org/wp-content/uploads/2012/02/An-Analysis-of-Bike-Sharing-Usage-Explaining-Trip-Generation-and-Attraction-from-Observed-Demand-Hampshire-et-al-12-2099.pdf.

[103] K. Zhao, S. Tarkoma, S. Liu, and H. Vo, "Urban Human Mobility Data Mining: An Overview," *IEEE International Conference on Big Data*, 2016, DOI: 10.1109/BigData.2016.7840811.

[104] X. Wang, G. Lindsey, J. E. Schoner, and A. Harrison, "Modeling bike share station activity: effects of nearby businesses and jobs on trips to and from stations," *Journal of Urban Planning and Development*, vol. 142, p. 04015001, 2015, DOI: 10.1061/(ASCE)UP.1943-5444.0000273.

[105] P. Vogel, T. Greiser, and D. C. Mattfeld, "Understanding bike-sharing systems using data mining: Exploring activity patterns," *Procedia-Social and Behavioral Sciences*, vol. 20, pp. 514-523, 2011, DOI: 10.1016/j.sbspro.2011.08.058.

[106] C. Rudloff and B. Lackner, "Modeling demand for bikesharing systems: neighboring stations as source for demand and reason for structural breaks," *Transportation Research Record: Journal of the Transportation Research Board*, pp. 1-11, 2014, DOI: 10.3141/2430-01.

[107] W. Zeng, C. W. Fu, S. M. Arisona, and H. Qu, "Visualizing interchange patterns in massive movement data," in *Computer Graphics Forum*, pp. 271-280, 2013, DOI: 10.1111/cgf.12114.

[108] J. W. Yoon, F. Pinelli, and F. Calabrese, "Cityride: a predictive bike sharing journey advisor," in *the 13th IEEE International Conference on Mobile Data Management (MDM)*, pp. 306-311, 2012, DOI: 10.1109/MDM.2012.16.

[109] R. Wenger, H. Zheng, and S. Dimitrov, "Biking Lane Usage Prediction," *McGill University Internal Report*, 2014, http://rl.cs.mcgill.ca/comp598/fall2014/comp598_submission_102.pdf.

[110] S. Jäppinen, T. Toivonen, and M. Salonen, "Modelling the potential effect of shared bicycles on public transport travel times in Greater Helsinki: An open data approach," *Applied Geography*, vol. 43, pp. 13-24, 2013, DOI: 10.1016/j.apgeog.2013.05.010.

[111] C. Gallop, C. Tse, and J. Zhao, "A seasonal autoregressive model of Vancouver bicycle traffic using weather variables," *i-Manager's Journal on Civil Engineering*, vol. 1, p. 9, 2011, http://docs.trb.org/prp/12-2119.pdf.

[112] J. Zhao, J. Wang, and W. Deng, "Exploring bikesharing travel time and trip chain by gender and day of the week," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 251-264, 2015, DOI: 10.1016/j.trc.2015.01.030.

[113] R. Montoliu, "Discovering mobility patterns on bicycle-based public transportation system by using probabilistic topic models," in *Ambient Intelligence-Software and Applications*, ed: Springer, pp. 145-153, 2012, DOI: 10.1007/978-3-642-28783-1_18.

[114] Y. Zhao, L. Chen, C. Teng, S. Li, and G. Pan, "Greenbicycling: A smartphone-based public bicycle sharing system for healthy life," in *IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, pp. 1335-1340, 2013, DOI: 10.1109/GreenCom-iThings-CPSCom.2013.232.

[115] M. Yang and X. Zhang, "A Novel Travel Adviser Based on Improved Back-propagation Neural Network," *the 7th IEEE International Conference on Intelligent Systems, Modelling and Simulation*, 2016, DOI: 10.1109/ISMS.2016.15.

[116] T. Raviv and O. Kolka, "Optimal inventory management of a bike-sharing station," *IIE Transactions*, vol. 45, pp. 1077-1093, 2013, DOI: 10.1080/0740817X.2013.770186.

[117] L. Dell'Olio, A. Ibeas, and J. L. Moura, "Implementing bike-sharing systems," *Municipal Engineer*, vol. 164, p. 89, 2011, DOI: 10.1680/muen.2011.164.2.89.

[118] J.-R. Lin and T.-H. Yang, "Strategic design of public bicycle sharing systems with service level constraints," *Transportation research part E: logistics and transportation review*, vol. 47, pp. 284-294, 2011, DOI: 10.1016/j.tre.2010.09.004.

[119] H. Xu, J. Ying, F. Lin, and Y. Yuan, "Station Segmentation with an Improved K-Means Algorithm for Hangzhou Public Bicycle System," *Journal of Software*, vol. 8, pp. 2289-2296, 2013, DOI: 10.4304/jsw.8.9.2289-2296.

[120] A. Sarkar, N. Lathia, and C. Mascolo, "Comparing cities' cycling patterns using online shared bicycle maps," *Transportation*, vol. 42, pp. 541-559, 2015, DOI: 10.1007/s11116-015-9599-9.

[121] T. Sherwood, E. Perelman, G. Hamerly, and B. Calder, "Automatically characterizing large scale program behavior," in *ACM SIGARCH Computer Architecture News*, pp. 45-57, 2002, DOI: 10.1145/635506.605403.

[122] R. Shahid, S. Bertazzon, M. L. Knudtson, and W. A. Ghali, "Comparison of distance measures in spatial analytical modeling for health service planning," *BMC health services research*, vol. 9, p. 200, 2009, DOI: 10.1186/1472-6963-9-200.

[123] T. M. Cover and J. A. Thomas, "Elements of information theory," *John Wiley & Sons*, 2012, DOI: 10.1002/047174882X.ch11.

[124] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, "Classification and regression trees," *CRC press*, 1984.

[125] J. D. Evan, "Straightforward statistics for the behavioral sciences," *Pacific Grove, CA: Brooks/Cole Publishing*, 1996, DOI: 10.12691/jas-4-1-1.

APPENDIX A

A1. The search space of hyper-parameters in validation set

Regressor	Hyper-parameters space
ABR	param_grid = {“n_estimators”: [100,200,300,400,500,600,700,800,900,1000], “learning_rate”: [1,2,3,4]}
BRR	param_grid = {“ n_iter”: [100,200,300,400,500,600,700,800,900,1000], “tol”: [0.001, 0.01, 0.1, 1]}
GBR	param_grid = {“n_estimators”: [100,200,300,400,500,600,700,800,900,1000], “learning_rate” : [0.001, 0.01, 0.1, 1]}
SVR	param_grid = {“cache_size”: [100,200,300,400,500,600,700,800,900,1000], “C”: [0.001, 0.01, 0.1, 1], “kernel”:[“linear”]}
RFR	param_grid = {“n_estimators”: [100,200,300,400,500,600,700,800,900,1000]}

A2. RMSE of each regressor in validation set for all search spaces

Based on Table 7.1, prediction is done in 3 rounds for each reference where the features for **round 1** are current times and one previous hour state, **round 2** are current times, one and two previous hours state, **round 3** are current times, the strong and very strong features.

A2.1. ABR Hyper-parameters for London

Reference	Prediction round	learning_rate	n_estimators									
			100	200	300	400	500	600	700	800	900	1000
Hour	1	1	437	386	390	385	446	392	371	489	403	359
		2	476	505	477	532	569	534	382	446	488	463
		3	540	515	1459	665	1042	725	715	490	571	578
		4	1075	1130	535	1692	1536	1430	1846	1828	1556	1740
	2	1	317	404	401	399	387	340	354	405	383	366
		2	356	357	376	383	410	377	371	340	398	392
		3	249	306	301	301	295	274	309	313	370	275
		4	912	784	998	1475	740	1366	1547	1608	718	358
	3	1	254	256	260	248	258	268	250	258	263	260
		2	243	252	263	264	250	257	261	265	260	262
		3	240	244	258	250	261	267	428	256	254	234
		4	419	1608	1488	1679	300	682	1732	636	529	1002
Day	1	1	191	216	206	291	263	167	213	224	197	192
		2	286	269	274	306	325	306	224	293	319	321
		3	274	282	288	316	309	293	284	326	301	281
		4	196	1140	461	759	1362	441	794	922	713	1266
	2	1	232	273	270	287	299	252	236	180	300	203
		2	254	251	312	296	315	264	286	297	288	304
		3	270	265	283	252	289	287	265	277	287	272
		4	898	1278	486	1504	302	327	631	401	634	626
	3	1	383	284	361	410	396	382	279	311	405	387
		2	396	395	438	427	422	379	429	421	426	278
		3	251	313	359	371	321	318	332	332	336	338
		4	202	458	427	339	841	766	433	671	1045	1423
Week	1	1	232	213	238	271	292	255	242	227	218	199
		2	329	471	389	395	455	415	424	447	360	333
		3	428	525	414	455	469	470	461	475	491	483
		4	795	1004	734	628	1092	1019	636	1038	497	1009
	2	1	252	314	353	359	329	318	355	347	379	271
		2	258	312	336	329	306	315	355	335	341	349
		3	310	328	357	298	338	310	299	301	316	284
		4	1234	1046	660	874	359	946	802	1034	755	555
	3	1	218	212	216	240	208	220	200	199	211	212
		2	237	248	249	275	257	256	255	254	275	266
		3	241	240	278	258	264	282	271	253	369	274
		4	556	1836	1340	315	537	326	1673	1233	667	499

A2.2. BRR Hyper-parameters for London

Reference	Prediction round	toll	n_iter									
			100	200	300	400	500	600	700	800	900	1000
Hour	**1**	**0.001**	595	595	595	595	595	595	595	595	595	595
		0.01	595	595	595	595	595	595	595	595	595	595
		0.1	595	595	595	595	595	595	595	595	595	595
		1	547	547	547	547	547	547	547	547	547	547
	2	**0.001**	543	543	543	543	543	543	543	543	543	543
		0.01	543	543	543	543	543	543	543	543	543	543
		0.1	543	543	543	543	543	543	543	543	543	543
		1	543	543	543	543	543	543	543	543	543	543
	3	**0.001**	498	498	498	498	498	498	498	498	498	498
		0.01	498	498	498	498	498	498	498	498	498	498
		0.1	498	498	498	498	498	498	498	498	498	498
		1	498	498	498	498	498	498	498	498	498	498
Day	**1**	**0.001**	140	140	140	140	140	140	140	140	140	140
		0.01	140	140	140	140	140	140	140	140	140	140
		0.1	140	140	140	140	140	140	140	140	140	140
		1	140	140	140	140	140	140	140	140	140	140
	2	**0.001**	138	138	138	138	138	138	138	138	138	138
		0.01	138	138	138	138	138	138	138	138	138	138
		0.1	138	138	138	138	138	138	138	138	138	138
		1	138	138	138	138	138	138	138	138	138	138
	3	**0.001**	138	138	138	138	138	138	138	138	138	138
		0.01	138	138	138	138	138	138	138	138	138	138
		0.1	138	138	138	138	138	138	138	138	138	138
		1	138	138	138	138	138	138	138	138	138	138
Week	**1**	**0.001**	163	163	163	163	163	163	163	163	163	163
		0.01	163	163	163	163	163	163	163	163	163	163
		0.1	163	163	163	163	163	163	163	163	163	163
		1	163	163	163	163	163	163	163	163	163	163
	2	**0.001**	164	164	164	164	164	164	164	164	164	164
		0.01	164	164	164	164	164	164	164	164	164	164
		0.1	164	164	164	164	164	164	164	164	164	164
		1	164	164	164	164	164	164	164	164	164	164
	3	**0.001**	157	157	157	157	157	157	157	157	157	157
		0.01	157	157	157	157	157	157	157	157	157	157
		0.1	157	157	157	157	157	157	157	157	157	157
		1	157	157	157	157	157	157	157	157	157	157

A2.3. GBR Hyper-parameters for London

Reference	Prediction round	learning_rate	n_estimators									
			100	200	300	400	500	600	700	800	900	1000
Hour	**1**	**0.001**	539	513	491	472	457	444	435	427	420	415
		0.01	416	363	341	332	334	339	345	351	350	349
		0.1	334	351	360	367	373	388	386	403	414	405
		1	753	708	725	709	721	724	675	715	727	719
	2	**0.001**	524	491	460	431	409	388	371	357	345	334
		0.01	333	261	243	242	240	235	229	227	223	222
		0.1	210	235	256	278	269	267	271	272	274	278
		1	504	491	508	498	517	485	490	559	491	496
	3	**0.001**	525	496	468	442	419	397	380	365	352	341
		0.01	340	265	241	230	221	215	211	207	205	201
		0.1	211	173	167	166	163	165	164	163	161	161
		1	206	204	203	203	202	204	204	203	201	204
Day	**1**	**0.001**	186	186	184	181	176	173	170	167	164	162
		0.01	162	163	167	170	172	175	179	183	184	190
		0.1	185	210	257	270	267	267	272	268	271	270
		1	334	336	344	346	357	350	362	364	351	349
	2	**0.001**	186	186	184	181	177	173	170	167	165	163
		0.01	163	164	163	161	161	165	167	168	169	167
		0.1	167	189	194	202	213	219	218	223	235	230
		1	364	361	361	386	361	392	392	361	377	387
	3	**0.001**	178	170	164	159	154	151	148	146	144	143
		0.01	143	142	144	144	143	143	143	143	143	145
		0.1	147	152	154	158	159	161	161	161	163	163
		1	207	216	217	214	215	211	210	209	218	217
Week	**1**	**0.001**	206	197	189	182	177	172	169	166	164	162
		0.01	162	157	156	159	164	169	175	179	180	181
		0.1	192	227	256	264	276	283	276	277	268	280
		1	387	338	335	398	338	344	346	354	406	343
	2	**0.001**	207	198	191	185	181	177	174	172	169	168
		0.01	168	161	159	164	168	172	178	183	188	194
		0.1	186	201	205	199	211	200	210	210	213	219
		1	436	480	438	444	475	450	440	439	473	473
	3	**0.001**	206	197	188	182	176	171	167	165	162	161
		0.01	160	155	155	156	157	158	159	161	162	163
		0.1	165	165	164	164	164	159	164	162	164	170
		1	202	210	209	222	216	223	217	223	220	219

A2.4. SVR Hyper-parameters (Kernel Linear) for London

Reference	Prediction round	C	cache_size									
			100	200	300	400	500	600	700	800	900	1000
Hour	**1**	**0.001**	563	563	563	563	563	563	563	563	563	563
		0.01	565	565	565	565	565	565	565	565	565	565
		0.1	565	565	565	565	565	565	565	565	565	565
		1	555	555	555	555	555	555	555	555	555	555
	2	**0.001**	516	516	516	516	516	516	516	516	516	516
		0.01	509	509	509	509	509	509	509	509	509	509
		0.1	504	504	504	504	504	504	504	504	504	504
		1	502	502	502	502	502	502	502	502	502	502
	3	**0.001**	521	521	521	521	521	521	521	521	521	521
		0.01	520	520	520	520	520	520	520	520	520	520
		0.1	519	519	519	519	519	519	519	519	519	519
		1	515	515	515	515	515	515	515	515	515	515
Day	**1**	**0.001**	140	140	140	140	140	140	140	140	140	140
		0.01	140	140	140	140	140	140	140	140	140	140
		0.1	140	140	140	140	140	140	140	140	140	140
		1	140	140	140	140	140	140	140	140	140	140
	2	**0.001**	139	139	139	139	139	139	139	139	139	139
		0.01	139	139	139	139	139	139	139	139	139	139
		0.1	139	139	139	139	139	139	139	139	139	139
		1	140	140	140	140	140	140	140	140	140	140
	3	**0.001**	139	139	139	139	139	139	139	139	139	139
		0.01	139	139	139	139	139	139	139	139	139	139
		0.1	139	139	139	139	139	139	139	139	139	139
		1	139	139	139	139	139	139	139	139	139	139
Week	**1**	**0.001**	158	158	158	158	158	158	158	158	158	158
		0.01	158	158	158	158	158	158	158	158	158	158
		0.1	159	159	159	159	159	159	159	159	159	159
		1	161	161	161	161	161	161	161	161	161	161
	2	**0.001**	158	158	158	158	158	158	158	158	158	158
		0.01	159	159	159	159	159	159	159	159	159	159
		0.1	159	159	159	159	159	159	159	159	159	159
		1	162	162	162	162	162	162	162	162	162	162
	3	**0.001**	158	158	158	158	158	158	158	158	158	158
		0.01	158	158	158	158	158	158	158	158	158	158
		0.1	158	158	158	158	158	158	158	158	158	158
		1	158	158	158	158	158	158	158	158	158	158

A2.5. RFR Hyper-parameters for London

Reference	Prediction round		n_estimators									
			100	200	300	400	500	600	700	800	900	1000
Hour	1		273	292	295	293	290	299	291	287	293	291
	2		212	220	210	206	209	206	210	205	199	206
	3		144	144	142	142	144	144	142	146	146	146
Day	1		177	166	176	176	178	173	172	171	175	173
	2		163	159	158	160	154	162	156	157	157	157
	3		147	148	149	146	149	146	146	147	147	149
Week	1		204	178	186	185	187	187	188	188	187	187
	2		171	169	175	171	174	176	173	168	176	170
	3		164	162	159	163	162	162	162	163	163	161

A2.6. ABR Hyper-parameters for Washington

Reference	Prediction round	learning_rate	n_estimators									
			100	200	300	400	500	600	700	800	900	1000
Hour	1	1	72	73	67	70	70	69	72	75	68	78
		2	67	74	81	72	83	77	68	76	79	79
		3	75	74	79	74	74	73	78	73	72	73
		4	189	145	105	134	161	160	170	316	111	218
	2	1	66	74	78	78	78	78	80	79	77	79
		2	73	79	73	74	73	74	73	71	73	76
		3	63	68	64	64	68	65	69	67	66	65
		4	135	98	156	137	140	174	200	148	114	192
	3	1	69	74	75	72	74	75	72	70	78	75
		2	73	72	72	76	73	75	75	71	72	74
		3	67	71	67	73	68	71	71	67	73	70
		4	99	75	134	140	151	94	113	114	109	105
Day	1	1	53	58	56	64	61	63	62	55	64	64
		2	55	56	57	59	60	60	57	61	59	59
		3	71	58	65	62	60	59	60	61	62	64
		4	134	167	100	145	111	180	260	150	111	69
	2	1	59	59	64	63	64	62	62	63	63	63
		2	58	59	61	64	62	60	62	63	63	61
		3	65	72	70	69	66	71	70	73	67	72
		4	107	145	139	96	94	89	209	106	200	267
	3	1	64	74	75	64	69	71	70	67	56	68
		2	83	65	75	67	66	62	59	67	62	61
		3	70	74	77	75	76	77	77	77	75	76
		4	293	242	106	273	391	154	210	167	112	214
Week	1	1	60	65	58	62	59	77	65	62	61	62
		2	60	65	67	67	58	73	63	69	64	63
		3	62	66	53	53	56	59	56	55	54	54
		4	62	138	71	184	89	169	127	281	148	62
	2	1	71	69	83	71	71	74	82	76	72	69
		2	80	63	69	64	65	67	72	69	67	68
		3	58	54	49	53	52	55	54	55	53	53
		4	321	91	196	217	243	64	258	97	171	254
	3	1	86	75	78	82	74	77	81	68	71	72
		2	93	93	89	89	78	76	82	84	94	90
		3	85	93	88	79	82	83	89	89	84	82
		4	337	158	100	231	121	126	153	341	112	92

A2.7. BRR Hyper-parameters for Washington

Reference	Prediction round	toll	n_iter									
			100	200	300	400	500	600	700	800	900	1000
Hour	**1**	**0.001**	79	79	79	79	79	79	79	79	79	79
		0.01	79	79	79	79	79	79	79	79	79	79
		0.1	79	79	79	79	79	79	79	79	79	79
		1	79	79	79	79	79	79	79	79	79	79
	2	**0.001**	76	76	76	76	76	76	76	76	76	76
		0.01	76	76	76	76	76	76	76	76	76	76
		0.1	76	76	76	76	76	76	76	76	76	76
		1	76	76	76	76	76	76	76	76	76	76
	3	**0.001**	75	75	75	75	75	75	75	75	75	75
		0.01	75	75	75	75	75	75	75	75	75	75
		0.1	75	75	75	75	75	75	75	75	75	75
		1	75	75	75	75	75	75	75	75	75	75
Day	**1**	**0.001**	47	47	47	47	47	47	47	47	47	47
		0.01	47	47	47	47	47	47	47	47	47	47
		0.1	47	47	47	47	47	47	47	47	47	47
		1	47	47	47	47	47	47	47	47	47	47
	2	**0.001**	47	47	47	47	47	47	47	47	47	47
		0.01	47	47	47	47	47	47	47	47	47	47
		0.1	47	47	47	47	47	47	47	47	47	47
		1	47	47	47	47	47	47	47	47	47	47
	3	**0.001**	47	47	47	47	47	47	47	47	47	47
		0.01	47	47	47	47	47	47	47	47	47	47
		0.1	47	47	47	47	47	47	47	47	47	47
		1	48	48	48	48	48	48	48	48	48	48
Week	**1**	**0.001**	45	45	45	45	45	45	45	45	45	45
		0.01	45	45	45	45	45	45	45	45	45	45
		0.1	45	45	45	45	45	45	45	45	45	45
		1	45	45	45	45	45	45	45	45	45	45
	2	**0.001**	45	45	45	45	45	45	45	45	45	45
		0.01	45	45	45	45	45	45	45	45	45	45
		0.1	45	45	45	45	45	45	45	45	45	45
		1	45	45	45	45	45	45	45	45	45	45
	3	**0.001**	45	45	45	45	45	45	45	45	45	45
		0.01	45	45	45	45	45	45	45	45	45	45
		0.1	45	45	45	45	45	45	45	45	45	45
		1	45	45	45	45	45	45	45	45	45	45

A2.8. GBR Hyper-parameters for Washington

Reference	Prediction round	learning_rate	n_estimators									
			100	200	300	400	500	600	700	800	900	1000
Hour	**1**	**0.001**	119	113	107	102	98	94	90	87	84	82
		0.01	82	66	60	56	54	53	52	51	50	50
		0.1	51	47	47	47	47	46	45	45	45	45
		1	55	58	59	59	60	60	59	59	60	60
	2	**0.001**	118	112	106	100	96	92	88	85	83	80
		0.01	80	63	58	56	54	54	53	53	53	52
		0.1	51	48	46	45	45	45	45	44	44	44
		1	56	57	62	59	62	58	62	61	58	59
	3	**0.001**	118	112	106	101	96	92	89	86	83	81
		0.01	81	65	58	54	51	49	48	47	46	45
		0.1	45	43	42	41	41	42	42	43	43	44
		1	51	53	49	50	52	50	52	52	50	50
Day	**1**	**0.001**	68	65	62	60	58	56	55	53	52	51
		0.01	51	48	47	47	48	48	48	48	48	47
		0.1	48	50	51	51	52	52	52	52	52	53
		1	51	55	60	61	63	61	62	61	61	62
	2	**0.001**	68	65	62	60	58	56	55	53	52	51
		0.01	51	48	47	46	47	47	48	48	48	48
		0.1	48	48	47	47	48	47	47	48	47	48
		1	59	67	68	64	69	63	70	63	65	64
	3	**0.001**	68	65	62	60	58	56	54	53	52	51
		0.01	51	47	46	45	45	44	44	43	43	43
		0.1	43	41	42	43	43	43	44	45	45	46
		1	48	56	55	56	62	61	61	61	58	62
Week	**1**	**0.001**	58	56	54	52	51	50	48	47	47	46
		0.01	46	43	42	41	41	41	41	41	41	41
		0.1	41	42	42	41	42	42	42	42	42	42
		1	51	54	56	56	57	56	58	58	58	58
	2	**0.001**	58	56	54	52	51	50	48	47	47	46
		0.01	46	43	42	41	41	41	41	41	40	40
		0.1	40	39	39	39	40	40	40	41	41	41
		1	53	56	58	60	60	59	61	61	61	61
	3	**0.001**	58	56	54	53	51	50	49	48	47	46
		0.01	46	43	42	41	41	40	40	39	39	39
		0.1	40	39	40	40	41	41	42	42	42	42
		1	52	54	58	58	59	59	60	59	60	61

A2.9. SVR Hyper-parameters (Kernel Linear) for Washington

Reference	Prediction round	C	cache_size									
			100	200	300	400	500	600	700	800	900	1000
Hour	**1**	**0.001**	122	122	122	122	122	122	122	122	122	122
		0.01	123	123	123	123	123	123	123	123	123	123
		0.1	120	120	120	120	120	120	120	120	120	120
		1	104	104	104	104	104	104	104	104	104	104
	2	**0.001**	116	116	116	116	116	116	116	116	116	116
		0.01	116	116	116	116	116	116	116	116	116	116
		0.1	114	114	114	114	114	114	114	114	114	114
		1	100	100	100	100	100	100	100	100	100	100
	3	**0.001**	118	118	118	118	118	118	118	118	118	118
		0.01	117	117	117	117	117	117	117	117	117	117
		0.1	114	114	114	114	114	114	114	114	114	114
		1	97	97	97	97	97	97	97	97	97	97
Day	**1**	**0.001**	47	47	47	47	47	47	47	47	47	47
		0.01	46	46	46	46	46	46	46	46	46	46
		0.1	46	46	46	46	46	46	46	46	46	46
		1	46	46	46	46	46	46	46	46	46	46
	2	**0.001**	47	47	47	47	47	47	47	47	47	47
		0.01	47	47	47	47	47	47	47	47	47	47
		0.1	46	46	46	46	46	46	46	46	46	46
		1	46	46	46	46	46	46	46	46	46	46
	3	**0.001**	47	47	47	47	47	47	47	47	47	47
		0.01	47	47	47	47	47	47	47	47	47	47
		0.1	46	46	46	46	46	46	46	46	46	46
		1	46	46	46	46	46	46	46	46	46	46
Week	**1**	**0.001**	44	44	44	44	44	44	44	44	44	44
		0.01	44	44	44	44	44	44	44	44	44	44
		0.1	44	44	44	44	44	44	44	44	44	44
		1	43	43	43	43	43	43	43	43	43	43
	2	**0.001**	44	44	44	44	44	44	44	44	44	44
		0.01	44	44	44	44	44	44	44	44	44	44
		0.1	43	43	43	43	43	43	43	43	43	43
		1	42	42	42	42	42	42	42	42	42	42
	3	**0.001**	44	44	44	44	44	44	44	44	44	44
		0.01	44	44	44	44	44	44	44	44	44	44
		0.1	44	44	44	44	44	44	44	44	44	44
		1	43	43	43	43	43	43	43	43	43	43

A2.10. RFR Hyper-parameters for Washington

Reference	Prediction round		n_estimators									
			100	200	300	400	500	600	700	800	900	1000
Hour	1		42	42	42	42	42	43	42	42	42	42
	2		39	38	38	39	39	38	39	38	38	39
	3		41	41	41	41	41	41	41	41	41	41
Day	1		48	48	48	48	48	48	47	48	48	48
	2		48	48	48	48	47	48	47	47	47	47
	3		44	45	43	44	44	44	43	44	44	44
Week	1		43	42	42	43	42	42	43	42	42	42
	2		40	40	40	40	40	40	40	40	40	40
	3		41	40	40	39	39	40	39	40	40	39

A.3. The best hyper-parameters of each ML regressor for system-wide prediction

Regressor	Ref.	Features	Hyper-parameters
LONDON			
ABR	**Hour**	Round 1	{'base_estimator':None, 'learning_rate':1.0, 'loss':'linear', 'n_estimators':1000, 'random_state':None}
		Round 2	{'base_estimator':None, 'learning_rate':3.0, 'loss':'linear', 'n_estimators':600, 'random_state':None}
		Round 3	{'base_estimator':None, 'learning_rate':3.0, 'loss':'linear', 'n_estimators':100, 'random_state':None}
	Day	Round 1	{'base_estimator':None, 'learning_rate':1.0, 'loss':'linear', 'n_estimators':600, 'random_state':None}
		Round 2	{'base_estimator':None, 'learning_rate':1.0, 'loss':'linear', 'n_estimators':800, 'random_state':None}
		Round 3	{'base_estimator':None, 'learning_rate':4.0, 'loss':'linear', 'n_estimators':100, 'random_state':None}
	Week	Round 1	{'base_estimator':None, 'learning_rate':1.0, 'loss':'linear', 'n_estimators':1000, 'random_state':None}
		Round 2	{'base_estimator':None, 'learning_rate':1.0, 'loss':'linear', 'n_estimators':100, 'random_state':None}
		Round 3	{'base_estimator':None, 'learning_rate':1.0, 'loss':'linear', 'n_estimators':800, 'random_state':None}
BRR	**All Refs.**	All Rounds	{ 'alpha_1':1e-06, 'alpha_2':1e-06, 'compute_score':False, 'copy_X':True, 'fit_intercept':True, 'lambda_1':1e-06, 'lambda_2':1e-06, 'n_iter':100, 'normalize':False, 'tol':1, 'verbose':False}
DTR	**All refs.**	All Rounds	{'criterion':'mse', 'max_depth':None, 'max_features':None, 'max_leaf_nodes':None, 'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'presort':False, 'random_state':None, 'splitter':'best'}
GBR	**Hour**	Round 1	{'alpha':0.9, 'init':None, 'learning_rate':0.01, 'loss':'ls', 'max_depth':2, 'max_features':None, 'max_leaf_nodes':None,'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':400, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
		Round 2	{'alpha':0.9, 'init':None, 'learning_rate':0.1, 'loss':'ls', 'max_depth':4, 'max_features':None, 'max_leaf_nodes':None,'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':100, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0,

			'warm_start':False}
		Round 3	{'alpha':0.9, 'init':None, 'learning_rate':0.1, 'loss':'ls', 'max_depth':4, 'max_features':None, 'max_leaf_nodes':None, 'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':900, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
	Day	Round 1	{'alpha':0.9, 'init':None, 'learning_rate':0.01, 'loss':'ls', 'max_depth':2, 'max_features':None, 'max_leaf_nodes':None,'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':100, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
		Round 2	{'alpha':0.9, 'init':None, 'learning_rate':0.01, 'loss':'ls', 'max_depth':4, 'max_features':None, 'max_leaf_nodes':None,'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':400, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
		Round 3	{'alpha':0.9, 'init':None, 'learning_rate':0.01, 'loss':'ls', 'max_depth':4, 'max_features':None, 'max_leaf_nodes':None, 'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':200, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
	Week	Round 1	{'alpha':0.9, 'init':None, 'learning_rate':0.01, 'loss':'ls', 'max_depth':2, 'max_features':None, 'max_leaf_nodes':None,'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':300, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
		Round 2	{'alpha':0.9, 'init':None, 'learning_rate':0.01, 'loss':'ls', 'max_depth':4, 'max_features':None, 'max_leaf_nodes':None,'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':300, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
		Round 3	{'alpha':0.9, 'init':None, 'learning_rate':0.01, 'loss':'ls', 'max_depth':4, 'max_features':None, 'max_leaf_nodes':None, 'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':300, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
SVR	**Hour**	All Rounds	{'C':1, 'cache_size':100, 'coef0':0.0, 'degree':3, 'epsilon':0.1, 'gamma':'auto', 'kernel':'linear', 'max_iter':-1, 'shrinking':True, 'tol':0.001, 'verbose':False}
	Other Ref.	All Rounds	'C':0.001, 'cache_size':100, 'coef0':0.0, 'degree':3, 'epsilon':0.1, 'gamma':'auto', 'kernel':'linear', 'max_iter':-1, 'shrinking':True, 'tol':0.001, 'verbose':False}
RFR	**Hour**	Round 1	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 100, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
		Round 2	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 900, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
		Round 3	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 300, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
	Day	Round 1	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 200, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}

		Round 2	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 500, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
		Round 3	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 400, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
	Week	Round 1	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 200, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
		Round 2	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 800, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
		Round 3	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 300, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
WASHINGTON DC			
ABR	**Hour**	Round 1	{'base_estimator':None, 'learning_rate':2.0, 'loss':'linear', 'n_estimators':100, 'random_state':None}
		Round 2	{'base_estimator':None, 'learning_rate':3.0, 'loss':'linear', 'n_estimators':100, 'random_state':None}
		Round 3	{'base_estimator':None, 'learning_rate':3.0, 'loss':'linear', 'n_estimators':100, 'random_state':None}
	Day	Round 1	{'base_estimator':None, 'learning_rate':1.0, 'loss':'linear', 'n_estimators':100, 'random_state':None}
		Round 2	{'base_estimator':None, 'learning_rate':2.0, 'loss':'linear', 'n_estimators':100, 'random_state':None}
		Round 3	{'base_estimator':None, 'learning_rate':1.0, 'loss':'linear', 'n_estimators':900, 'random_state':None}
	Week	Round 1	{'base_estimator':None, 'learning_rate':3.0, 'loss':'linear', 'n_estimators':300, 'random_state':None}
		Round 2	{'base_estimator':None, 'learning_rate':3.0, 'loss':'linear', 'n_estimators':300, 'random_state':None}
		Round 3	{'base_estimator':None, 'learning_rate':1.0, 'loss':'linear', 'n_estimators':800, 'random_state':None}
BRR	**All Refs.**	All Rounds	{ 'alpha_1':1e-06, 'alpha_2':1e-06, 'compute_score':False, 'copy_X':True, 'fit_intercept':True, 'lambda_1':1e-06, 'lambda_2':1e-06, 'n_iter':100, 'normalize':False, 'tol':1, 'verbose':False}
DTR	**All refs.**	All Rounds	{'criterion':'mse', 'max_depth':None, 'max_features':None, 'max_leaf_nodes':None, 'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'presort':False, 'random_state':None, 'splitter':'best'}
GBR	**Hour**	Round 1	{'alpha':0.9, 'init':None, 'learning_rate':0.1, 'loss':'ls', 'max_depth':2, 'max_features':None, 'max_leaf_nodes':None,'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':700, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
		Round 2	{'alpha':0.9, 'init':None, 'learning_rate':0.1, 'loss':'ls', 'max_depth':4, 'max_features':None, 'max_leaf_nodes':None,'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':800, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0,

			'warm_start':False}
		Round 3	{'alpha':0.9, 'init':None, 'learning_rate':0.1, 'loss':'ls', 'max_depth':4, 'max_features':None, 'max_leaf_nodes':None, 'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':400, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
	Day	Round 1	{'alpha':0.9, 'init':None, 'learning_rate':0.01, 'loss':'ls', 'max_depth':2, 'max_features':None, 'max_leaf_nodes':None,'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':300, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
		Round 2	{'alpha':0.9, 'init':None, 'learning_rate':0.01, 'loss':'ls', 'max_depth':4, 'max_features':None, 'max_leaf_nodes':None,'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':400, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
		Round 3	{'alpha':0.9, 'init':None, 'learning_rate':0.1, 'loss':'ls', 'max_depth':4, 'max_features':None, 'max_leaf_nodes':None, 'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':200, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
	Week	Round 1	{'alpha':0.9, 'init':None, 'learning_rate':0.1, 'loss':'ls', 'max_depth':2, 'max_features':None, 'max_leaf_nodes':None,'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':100, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
		Round 2	{'alpha':0.9, 'init':None, 'learning_rate':0.1, 'loss':'ls', 'max_depth':4, 'max_features':None, 'max_leaf_nodes':None,'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':200, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
		Round 3	{'alpha':0.9, 'init':None, 'learning_rate':0.1, 'loss':'ls', 'max_depth':4, 'max_features':None, 'max_leaf_nodes':None, 'min_samples_leaf':1, 'min_samples_split':2, 'min_weight_fraction_leaf':0.0, 'n_estimators':200, 'presort':'auto', 'random_state':None, 'subsample':1.0, 'verbose':0, 'warm_start':False}
SVR	**All Ref.**	All Rounds	{'C':1, 'cache_size':100, 'coef0':0.0, 'degree':3, 'epsilon':0.1, 'gamma':'auto', 'kernel':'linear', 'max_iter':-1, 'shrinking':True, 'tol':0.001, 'verbose':False}
RFR	**Hour**	Round 1	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 100, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
		Round 2	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 200, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
		Round 3	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 100, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
	Day	Round 1	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 100, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
		Round 2	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2,

			'min_weight_fraction_leaf' : 0.0, 'n_estimators': 500, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
		Round 3	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 300, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
	Week	Round 1	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 200, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
		Round 2	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 100, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}
		Round 3	{'bootstrap': True, 'criterion': 'mse', 'max_depth': None,'max_features':'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split' : 2, 'min_weight_fraction_leaf' : 0.0, 'n_estimators': 400, 'n_jobs' : 1, 'oob_score' : False, 'random_state' : None, 'verbose': 0, 'warm_start' : False}