TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

SUBASH BASNET

INVESTIGATIONS OF OUTDOOR MOBILITY PATTERNS OF TAXICABS IN URBAN SCENARIOS

Master of Science Thesis

Examiner:

Assoc. Prof. Elena Simona Lohan

Examiners and topic approved by

The Faculty Council of the Faculty of

Computing and Electrical Engineering

Approved on 04th November 2015.

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Master's Degree Program in Information Technology
BASNET, SUBASH: **Investigations of Outdoor Mobility Patterns of Taxicabs in Urban Scenarios**
Master of Science Thesis, 67 pages
March 2017
Major: Communication Systems and Networks
Examiner:
Associate Professor Elena Simona Lohan, Laboratory of Electronics and Communication Engineering, Tampere University of Technology (Finland).
Keywords: Urban computing, Smart City, Mobility Pattern, Taxicab, Global Positioning System (GPS), GPS traces.

This thesis investigates various outdoor mobility patterns of taxicabs in urban environments based on open-data real traces and it proposes a suitable outdoor mobility model to fit the provided measurement data. This thesis is processing user traces of taxicabs of two major cities: Rome and San Francisco downloaded from CRAWDAD open-source repository, which is responsible for sharing data from real networks and real mobile users across the various research communities around the world.

There are numerous sources of collecting traces of users in a city, such as mobile devices, vehicles, smart cards, floating sensors etc. This thesis presents a comparative analysis of the mobility patterns of various taxicabs from Rome and San Francisco cities based on data collected via GPS-enabled mobile devices. Finding suitable mobility models of taxicabs to represent the travelling patterns of users moving from one location to another with respect to their varying time, location and speed can be quite helpful for the advanced researches in the diverse fields of wireless communications, such as better network planning, more efficient smart city design, improved traffic flows in cities. Also other applications such as weather forecasting, cellular coverage planning, e-health services, prediction of tourist areas, intelligent transport systems can benefit from the information hidden in user traces and from being able to find out statistically valid mobility models.

The work here focused on extracting various mobility parameters from the crowdsourced open-source data and trying to model them according to various mobility models existing in the literature.

The measurement analysis of this thesis work was completed in Matlab.

PREFACE

This Master of Science thesis has been written for the completion of Master of Science (Technology) in the Laboratory of Electronics and Communication Engineering at Tampere University of Technology (TUT), Tampere, Finland.

Firstly, I would like to thank my examiner and supervisor, Associate Professor Elena Simona Lohan, for her long-term support throughout my thesis work and providing me with a thesis topic. It could be difficult to complete my thesis work without her great supervision, guidance and attention.

Secondly, I would like to thank my wife Sini Basnet for her continuous love and support. I would like to pay my deepest gratitude towards my parents whose everlasting care and motivation helped me to complete my studies. Also, I would like to thank my brother Sushil Basnet and sister Sarita Karki for always been encouraging throughout my studies.

Finally, I would like to thank all my friends especially Kumud Koirala for his help during my thesis.

I would like to dedicate my thesis work to my parents.

Tampere, March, 2017
Subash Basnet

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

2D	Two Dimensional (model)
3D	Three Dimensional (model)
3G	Third Generation of mobile communication
4G	Fourth Generation of mobile communication
ARIMA	Auto Regressive Integrated Moving Average
BM	Brownian Motion
CORSIM	Corridor Simulation
CRAWDAD	Community Resource for Archiving Wireless Data At Dartmouth
DSRC	Dedicated Short Range Communication
DTN	Delay Tolerant Network
GPS	Global Positioning System
GPRS	General Packet Radio Service
GSM	Global System for mobile communication
ICT	Information and Communications Technology
IoT	Internet of Things
ITS	Intelligent Transportation System
KL	Kullback Leibler
M2M	Machine to machine
MANET	Mobile Ad hoc NETwork
MSD	Mean Squared Displacement
PARAMICS	Parallel Microscopic Simulation
PDF	Probability Density Function
RFID	Radio Frequency Identification
RSU	Roadside Unit

SLAW	Self-similar Least Action Walk
SVM	Support Vector Machine
TRANSIMS	Transportation Analysis and Simulation System
UCC	Unified Communications and Collaboration
UWB	Ultra-wideband
V2I	Vehicle-to-Infrastructure
V2V	Vehicle-to-Vehicle
VANET	Vehicular Ad hoc NETwork
VMS	Variable Messaging Systems
VRC	Vehicle-to-Roadside
α	Flight length factor
β	Pause time factor
d_i	Euclidean distance
e_i	Elapsed time
k	Cut-off value
l	Length of flight
mc_{trip}	Spatial mean centre
p_i	Pause time
P	Probability density function
q_i	Duration of travel
Δr_0	Reference parameter
s_i'	Average speed of user
Δt_f	Duration of flight
Δt_p	Duration of pause time
θ	Direction of the flight
γ	Diffusion coefficient

1 INTRODUCTION

This chapter briefly describes the topic of the thesis. The background, motivation and its scope are also concisely explained. Additionally, the objectives and the research approach of this thesis study are portrayed as well as the Author's contributions.

1.1 Background and Motivation

With a swift development of wireless networking technologies and embedded systems, an uprising number of computing devices and sensors are surrounded in our daily lives. Consequently, a lot of information related to user mobility such as location, motion, and behaviours of vehicles, can be easily extracted. Using the information collected by moving taxis is one method to study more about user mobility patterns in an urban scenario. Nowadays, GPS devices are fitted into taxis and their trace data, therefore, can be easily retrieved. For instance, many taxi companies in China have to install GPS tools in each of their taxis to facilitate their administrative tasks. As a result, it helps to retrieve the current and historical taxi traces data for the prediction of urban user mobility [1].

Ubiquitous computing has mostly been experimented in both the rural areas such as forests, vineyards, or glaciers and also in the urban environments such as smart houses or rooms where every single sensor, person, transportation unit, house or any street can be taken as a computing component. But urban areas are more complicated to study because of a dense population and dynamic behaviour of people who rapidly take part and leave out of the system with different output of data usage compared to day or night time [2]. Along with the increasing population of human civilization, there is an emerging need for urban planning that should incorporate land use planning as well as transportation planning thereby developing the physical, economical and the social environments of society [3]. Urbanization is increasing in a rapid manner in many developing countries where bigger cities have already started urban reconstruction, their renewal and suburbanization. That's why, we need competent technologies so that they can soundly ascertain urban dynamics and produce truthful information to urban planners. Indeed, developed countries usually possess bigger challenges for the urban planning as they have countless taxicabs in urban areas. For instance, the number of taxicabs in Mexico City, Beijing, Tokyo and Seoul are all over 60,000 respectively. Similarly, there are more than 10,000 licensed taxicabs in around 30 cities, including New York City, Shanghai, Hong Kong, London, and Paris correspondingly.

For the efficient communication and monitoring, taxis are usually armed with GPS sensors that help them to deliver their location data to a central server in a uniform time interval, e.g. 1~2 minutes. Clearly, most of the taxicabs in the major developed cities in the world holding GPS devices produce massive data of trajectories everyday [4] [5] [6].Moreover, taxicabs with GPS devices are considered to be ubiquitous mobile sensors which continuously sense the city's traffic information on roads and also the user mobility patterns [3].

The research of user mobility modelling has been started since a decade ago. Similar to mobile phone IDs, taxicabs GPS trajectories possess significant information. Ziebart et al. [7] studied various navigation services related with the driver's behaviour. Yuan et al. [6] produced navigation services reasoning on taxi drivers historical GPS footprints. Liu et al. [8] invented the schemes of taxi drivers comparing the performance of top drivers and normal drivers. Similarly, Zheng et al. [3] revealed flawed urban planning using GPS trajectories of taxicabs travelling in the urban areas. These GPS trajectories help to appraise efficiency of the urban planning such as for newly constructed roads and subway lines in a city, and they can help the urban planners tackling new challenges which were not identified before. Furthermore, Zhang et al. [9] proposed a new model of irregular driving patterns from taxi GPS traces, focusing on automatic detection of taxi frauds or road network changes in urban areas.

Also, there have been studies for guiding the passengers or taxi drivers to make their lives more convenient. Phithakkitnukoon et al. [10] worked on prediction of distribution of empty taxis around the city with a naïve Bayesian classifier, which counts several factors such as weather, day of the week, and time of day to improve performance. Moreover, Chang et al. [11] anticipated the demands of taxicabs in urban areas. They filtered the historical data with recent circumstances, e.g., location, time, and weather. After that, those filtered data are clustered and mapped to road names semantically. But the authors of [11] did not take account of the distribution of vacant taxis around the clusters they provide, which is away from the real demand.

1.2 Objectives and Scope of the Thesis

The objective of this thesis has been to investigate the outdoor mobility patterns of taxicabs in urban areas. This thesis analysed urban taxi mobility traces obtained from two major developed cities: San Francisco and Rome. This thesis work is processing datasets from CRAWDAD, a community resource for archiving wireless data at Dartmouth university [12] [13] . The study of mobility of taxicabs, governed by the drivers' behaviour and passengers' destination places, has drawn lot of attention in recent years. Among various mobility models, trace-driven taxicab mobility models contain detailed information but with too much overhead also. Generally, most of the basic mobility models fail to comprise the required social and geographical features such as the transition among various regions, hot spots etc. Besides this, it is observed that most of the mobility oriented research works have paid attention on estimation of geographical layouts but not necessarily modelling of them. This thesis work mainly focuses on investigation of mobility patterns of taxicabs with geographical information in the urban areas fitting with appropriate mobility model.

1.3 Author’s contributions

The author’s major contributions to this thesis work are outlined as follows:

1. Literature studies about smart city, mobility trace data, vehicular communication and various existing user mobility models.
2. Pre-processing and conversion of measurement data, downloaded from CRAWDAD repository, into suitable format for further analysis.
3. Extracting various user mobility parameters from measurement data in order to investigate how the users’ movement patterns are and what mobility model can approximately fit into it from MATLAB software.
4. Analysing and comparing measured mobility parameters of each dataset used in this thesis work.
5. Comparing output mobility parameters with few existing user mobility models.
6. Analysis of the output mobility model parameters in the context of random walk model.

Some challenges observed during this thesis study are outlined as below:

- The conversion of the data formats of both datasets of cities: Rome and San Francisco into desired computable form in MATLAB platform was one of the challenges of this thesis.
- Since both the datasets are large in volume, processing of datasets in appropriate formats is time consuming.
- The inaccuracy of data is observed in some extents during simulation and analysis stages.

1.4 Research Approach

This thesis work is based on simulations which are implemented on MATLAB software. Initially, the datasets downloaded from [12] [13] were converted into required format and then various mobility parameters such as speed, pause time, probability of return at the same point, travel time, step, etc. are computed and analysed by simulation. The overall steps of processing datasets used in this thesis work are described in Figure 1-1 below.

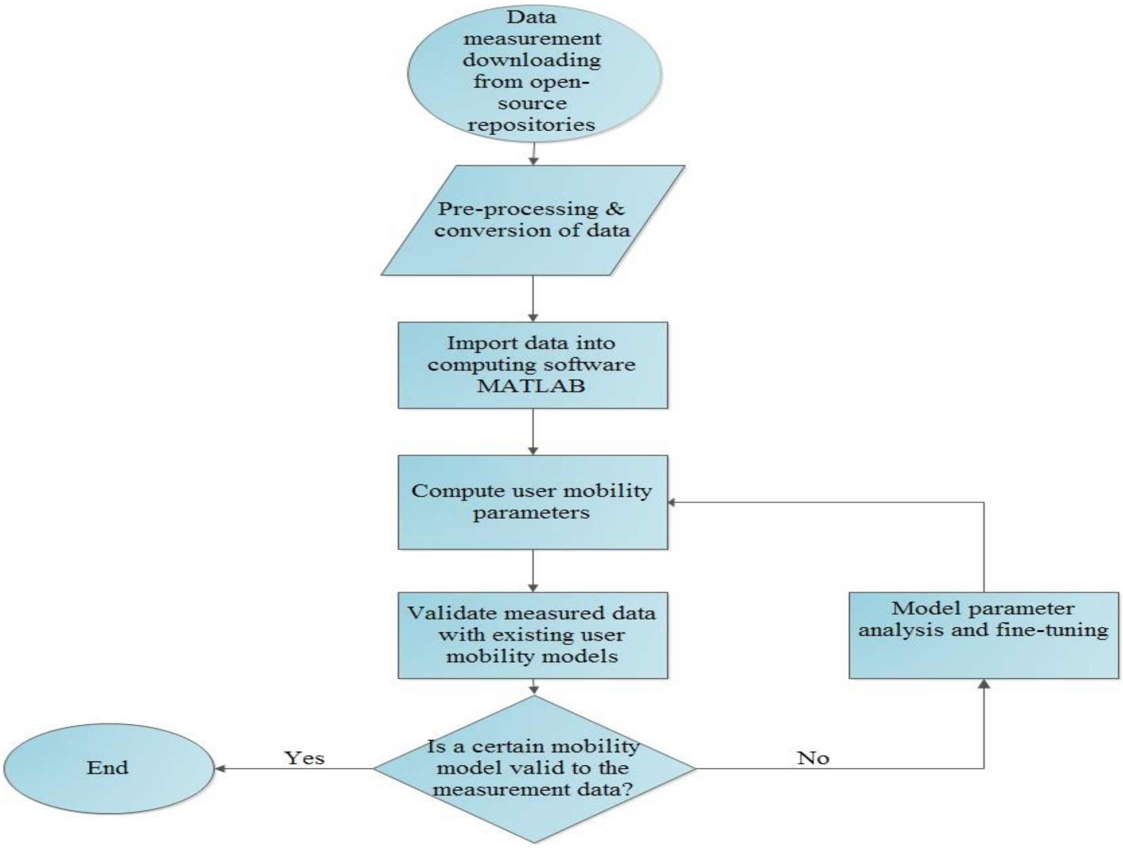

Figure 1-1 Main steps of processing overall dataset in this thesis work

The target for this thesis study has been to compare statistically the mobility parameters collected by taxi drivers in two major cities: Rome and San Francisco, and then to come to a conclusion for optimum mobility model of the taxicabs in urban areas among the studied models. The overall structure of this thesis is summarised as below:

- Chapter 2 describes the concept of smart city, its challenges and applications. Besides this, it also explains about trace data and data mining techniques.
- Chapter 3 explains various mobility models used for outdoor car traffic. Further, it describes the types of mobility parameters related to this thesis work.
- Chapter 4 outlines the concept of vehicular communication and its applications, their routing protocols, and the research issues resolved by various authors which are similar to this thesis study.
- Chapter 5 covers the description of measured datasets and the procedure of its analysis, conversion and some map plotting.
- Chapter 6 describes the measurement analysis and the final results of the measured datasets.
- Chapter 7 explains the summary of main findings of this research work and ideas for future directions along with the open challenges.

2 SMART CITY

This chapter focuses on the concept of smart city, its challenges and applications. Additionally, it also describes about trace data and data mining techniques in the context of smart cities.

2.1 Concept of smart city

A smart city can be characterized as an innovative city that uses vital Information and Communication Technologies (ICTs) to ensure social, economic and environmental developments. To meet the requirements of smart city, first one should sense the real-time data of citizens and all the city infrastructure, process it and finally generate statistically meaningful data out of the processed measurements [14]. In a smart city concept, an important task is the efficient planning of urban developments and traffic. Such careful planning could reliably simplify many obstacles, such as traffic bottlenecks, environment pollution, social health vulnerabilities, security and economy threats. In a smart city concept, we do not only utilize the information technology platform, but also human manpower, public investments and environmental resources [15]. Figure 2-1 describes the various innovative technologies used to develop the infrastructure of a smart city. Further, Figure 2-1 explains that by integrating different technologies to the trace data of various objects such as human, vehicles, animals etc. can develop productive applications and hence lead to progress towards a smart city. For example, technologies like M2M (Machine to Machine) and UCC (Unified Communications & Collaboration) play a significant role to support towards smart city development. The M2M belongs to technologies which let both the wireless as well as wired systems to communicate with other machines of the same ability without any human intervention. The vehicle to vehicle communication uses M2M technology for various purposes such as information sharing of accidents on the road, road maps exchanges, route information transfers and so on. Similarly, UCC refers to a multipurpose conferencing software that combines numerous methods such as text, audio, video and virtual white boards and makes them accessible via a single interface. In this thesis, we are processing on trace data of various taxicabs from two different cities: Rome and San Francisco to investigate the mobility patterns of users in urban environment where various technologies could be used between those taxicabs for sharing information to each other.

A smart city can be broadly defined as an instrumented, interconnected and intelligent city. Firstly, instrumentation means here to sense and manipulate real-world data by means of various devices, for example, sensors, kiosks, meters, cameras, mobile phones, tablets, implanted medical devices and many more along with networks of human sensors. Secondly, interconnection here means processing those trace data into an enterprise computing platform and transmitting them through various city services. Finally, intelligence represents here preparing those data for analysing, modelling and optimising for the efficient operational decisions in the business areas [16].

The word ‘‘smartness’’ can be technically defined as one capable of understanding, learning, computing, and self-responsiveness. Hence, first significant step to make a city smart, can be sensing, analysing and processing the pre-detected data among its dynamic environment. The sensed data which is also referred as traces yield very important information of dynamic objects such as human, vehicles, animals and others with their mobility patterns. There are Global Positioning System (GPS) and many localization technologies applied in the infrastructure of most of the cities that make sensing and mining of data very easy. Those traces extracted from the moving objects basically represent a temporal sequence of spatial points and their timestamps [15].

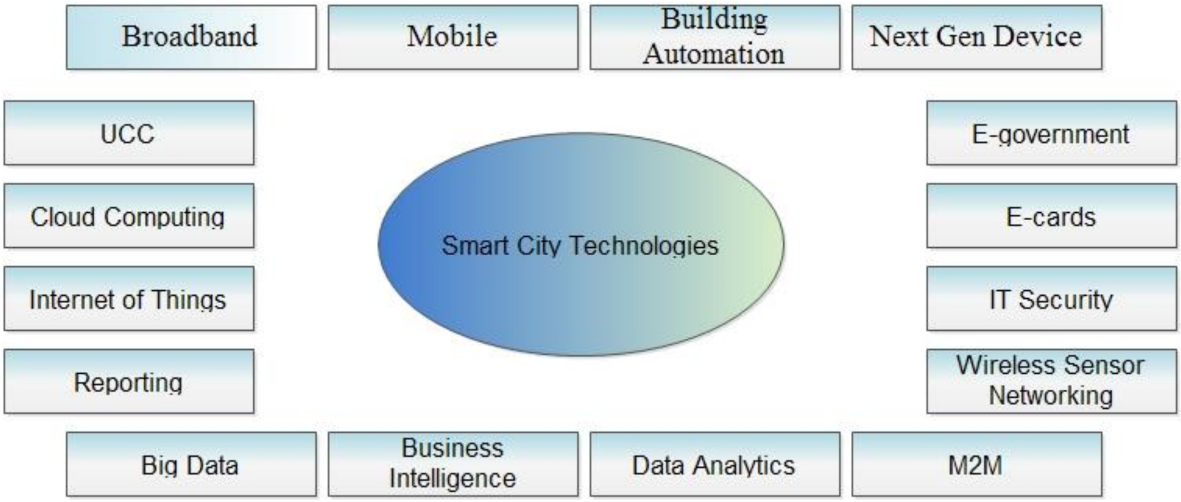

Figure 2-1 Various technologies used in smart city implementation

2.2 Trace Data

It’s a complicated procedure to collect and extract huge trace data of dynamic objects in a city. Depending on the types of sensors and devices, the trace data sources mainly can be classified into four groups described as below:

- **Traces based on mobile devices**: They are portable and capable of providing traces of their owners’ location easily by means of various technologies, for example, mobile devices having GPS, Wi-Fi, GSM, and Bluetooth chipsets on them can be used to collect such user traces.

- **Traces based on vehicles:** Since many of the vehicles are furnished with GPS devices already, it’s easy to navigate both the vehicles own location and the passengers as well as drivers’ trajectories.

- **Traces based on smart cards**: Smart cards used for example for banking purposes or for transportation plans can also help and be used to extract the location of users.

- **Traces based on floating sensors**: Floating sensors refers to sensors which contain radio frequency identification (RFID) tags which are capable of mining their own traces, for example, trash-tracking. A trash-tracking is a tracking system for observing the waste-removal chain based on the attachment of ‘‘trash tags’’ to waste objects [17]. Moreover, these sensors are based on various localization technologies which is illustrated more in details in Table 2.1

Based on the technology behind the data collection, the trace data types in Table 2.1 are compared with four important characteristics which are briefly discussed as below [15]:

- **Reference**: It describes the two types of location of trace data: a) absolute location and b) relative location. The absolute location represents the exact location details of trace data, for example, latitude 30°12’45’’N, longitude 120°10’19’’E. On the other hand, the relative location means location with respect to a reference point, for example, 1 km south-east from a fixed benchmark. This thesis study has trace data which represents absolute location details of targeted cities.

- **Expression**: It describes whether the trace data is a physical location (e.g. 30°12’45’’N, 120°10’19’’E) or a symbolic location label based on the descriptive of the location itself, such as a city or a named room (e.g. Zhejiang University, Hangzhou).

- **Precision**: It is the accuracy of location of the trace data.

- **Coverage**: It describes the valid range of communication between transmitter and receiver.

It is significant to find the location accuracy of sensed trace data that can be attained using localization methods as described above. Moreover, the traces can be categorised in various forms such as centralised or distributed, GPS based or GPS free, fine grained or coarse grained, stationary or mobile traces and range-based or range-free and so on. Our thesis work is processing GPS based trace data. The trace data have different meaning based on the technology they are using for the sensing and processing steps. There are various localization methods but some are frequently used for trace data, for example, GPS, WiFi, GSM, Bluetooth and RFID, UWB(Ultra-wideband) [18].

Table 2.1 various popular localization technologies

Technology	**Data**	**Reference**	**Expression**	**Accuracy**	**Coverage**
GPS	Geographic coordinate	Absolute	Physical	1–5 meters (95-99%)	Outdoors
WiFi	Access point ID + signal strength or local coordinate	Relative	Symbolic/ Physical	1–20 meters	<100 meters from an access point
Cell Tower	Cell tower ID + signal strength or geographic coordinate	Relative/ Absolute	Symbolic/ Physical	50–200 meters in cities	Cell coverage. 5–30 km from a cell tower.
Bluetooth	Device ID	Relative	Symbolic	1-20 m	5–10 meters for Class 1; 20–30 meters for Class 2
RFID	Reader’s ID/position	Relative	Symbolic/ Physical	1-5 m	1 meter for passive RFID; 100 meters for active RFID
UWB	Individual’s ID/position	Absolute	Physical	1-10 meters	Indoor/Outdoor

2.3 Methods of Trace Analysis and Mining

The process of trace analysis and mining of trace data can be accomplished in many ways depending on the required research scenarios. Primarily, the main methods for trace mining are described below.

- **Clustering**: This method is used for extracting hotspots, moving objects traces, and processing various objects with same kind of traces. It contains two types of objects: points and traces. Clearly, points are the group of objects with similar features and dimensions whereas traces contain objects with differing features and facets. One of the critical problems in this technique is measurement of same or different characteristics of points and traces [15].

- **Ranking:** It is one of the popular methods used for analysing the mobility patterns, locating specific persons from the social networks. In this method, there are some samples of data and with their features as input and then a model is mapped with a range for rendering in an orderly manner. The required characteristics of trace data can be attained through various aspects such as location string, geometric curve, geographic route, co-presence, time information, and sequence of semantic labels [15].

- **Classification:** In this method, there is a class of samples, and a model developed by training set of those samples along with their respective class labels. It consists of three major phases: feature extraction, training process, and the testing stage. Clearly, during analysis of unique activity from trace data, those traces are referred as samples whereas individual activity is a class itself. This technique includes both the daily living activities and social events also. There are many instances of this method such as finding daily living activities, abnormality detection, and transportation behaviour [15].

- **Regression:** In this technique, a continuous function is developed by means of dependent variables and single or multiple independent variables. It is used to observe the trace-driven social networks along with a definite mobility model designation. It is used for correlating the social networks as well as for predicting the urban area transportation. For example, it is used for the supply of public bicycles, traffic conditions, finding amount of passengers and so on [15].

- **Physical Statistical Modelling:** It is very important technique for finding statistical models of user mobility unfolding the physical laws and many detailed parameters such as number of visiting places, frequency of visits, step length and so on. For example, power law for step length and continuous-time random walk for mobility model uses this method [15].

The basic layout of trace mining methods for a smart city is illustrated in Figure 2-2 given below. This figure explains about the general process of sensing the trace data from various sources of moving objects in the city and then mining those data with various techniques as mentioned above in an appropriate form. This entire process can be described briefly as combination of three important phases: trace processing, knowledge gathering and applications delivery.

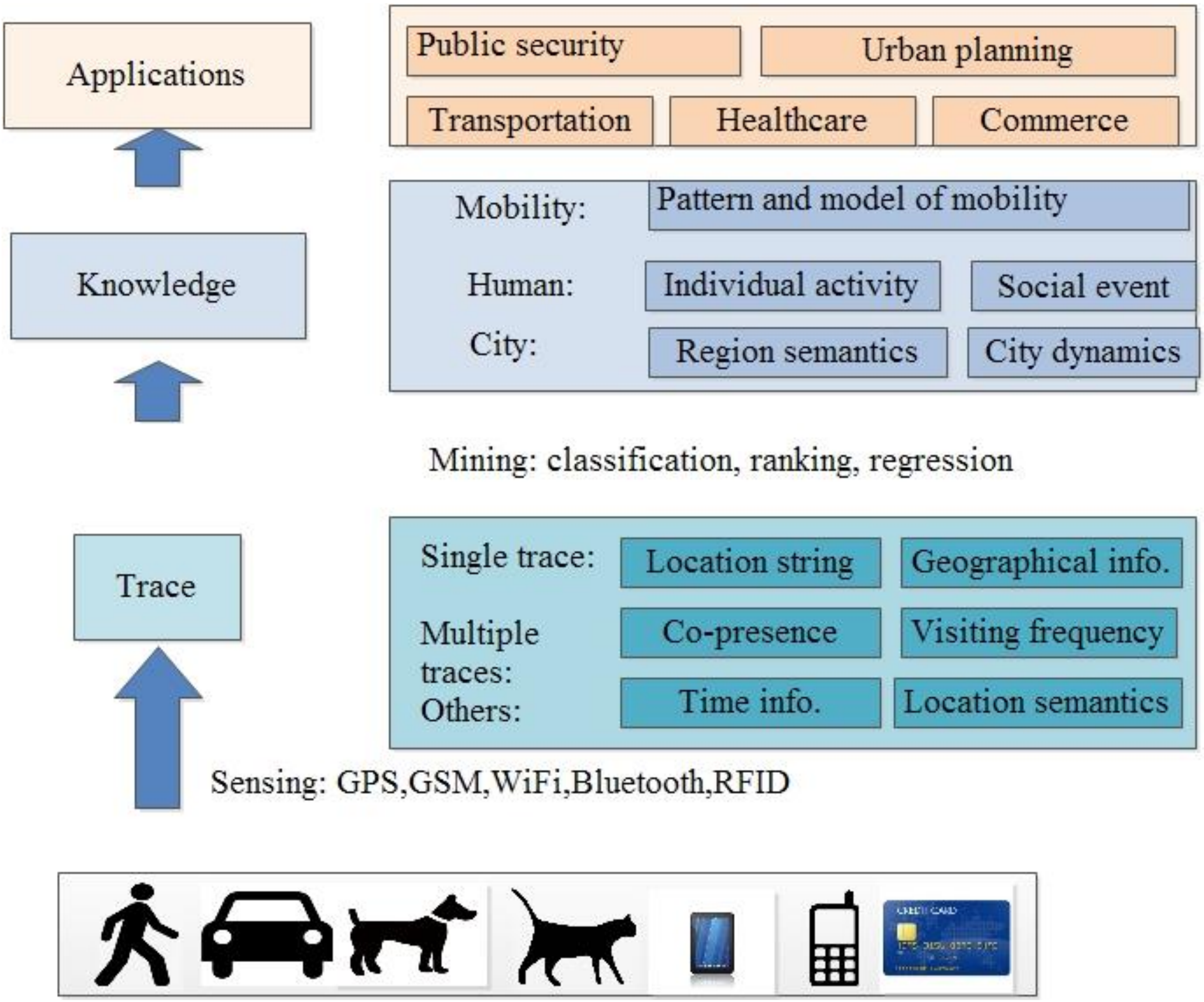

Figure 2-2 Basic layout of trace data mining for a smart city

Smart city is an expanded form of a digital as well as an intelligent city which needs a proper combination of the urban trace data and management systems along with the latest ICT technologies, for example, Internet of Things (IoT), cloud computing and many more. Additionally, the so-called fourth generation (4G) mobile communication technology has already helped a lot to accelerate the broadband applications with the high-speed data transmission which is a strong base of smart city infrastructure design. Similarly, IoT is an emergent technology which helps to identify, locate and manage network with the mutual transmission of trace data between various protocol oriented devices like sensors, RFIDs, GPS and so on. Moreover, cloud computing also plays vital role here. It works in a distributed computing platform with very efficient data processing resulting reliable data monitoring of human daily activity like electricity, gas, water and many more [19].

2.4 Challenges in a Smart City

Trace data represent many kinds of information such as user mobility, their daily activities, social events, regions semantics and many dynamics of a city. But throughout processing of those data, it can evolve many issues and challenges that are briefly described below [15].

- **Mobility- Patterns, Models and Prediction:** Identifying the mobility patterns of dynamics of a city is one of the major issues. The mobility patterns of a single user and multiple users follow different laws and models. Moreover, interpreting those patterns into suitable models is another challenging task. Usually, there is always contradiction between a hypothetical model and an empirical model of data leading the required outcome into erroneous ones. That's why, it is important to consider including both models for the accurate design of user data models. Similarly, it is still hard to predict different mobility models, the good way to proceed ahead is considering both the periodic modulations and spatial correlations while performing predictions. Wang et al. [20] studied day-to-day mobility traces of 6 Million mobile phone subscribers and detected the similarity between individuals' movements, their social connectivity and the intensity of communications between them were intensely associated with each other accomplished by both the empirical analysis and predictive models. Moreover, merging both mobility and network measures, authors observed that the prediction accuracy could be noticeably enhanced in supervised learning.

- **Human Behaviour- From Location to Activity and Event:** The low-level user activities such as walking, sitting and standing, can easily be discovered with trace data. On the other side, extracting trace data from social activities such as movement of crowd including epidemics or planned crimes, can be challenging task. It is complex to mine comprehensive information from the low-level user activities. Additionally, the various machine learning techniques and other complex algorithms can detect user behaviour in limited areas but not in all. Phithakkitnukoon et al. [21] produced an activity-aware map that covers most likely activities connected with a definite region in the map founded on POIs (Points of Interests) information. By means of activity-aware map, authors extracted users' distinct daily movement patterns from investigating a huge mobile phone data of closely one million records. As a result, the correlation analysis displayed a strong correlation in regular activity patterns inside the group of people who allocated common work area's profile.

- **Region Characteristics- Semantics and Significance:** The region characteristics often reflects the user mobility with the traces. For example, during holidays there are many people in popular sites of city compared to workdays. Clearly, the best example for this scenario are the hotspots. The hotspots reflect the regular and repetitive user activities in a visual way and helps to extract high-level data traces. In [1] , authors suggested an improved ARIMA (Auto-regressive Integrated Moving Average) based prediction method to foretell the spatial-temporal deviations of passengers in a hotspot which counts holidays, weekdays and divides day with series of time-scales leading to reliable tracking of users' mobility.

- **City Dynamics- Expression, Visualization, and Evaluation:** City dynamics consists of many types of data extracted from various fields in a city such as traffic flow, epidemic spread, urban growth and so on. The main challenge here is to process diverse trace data with different sources and formats. For example, sources of traces can be mobile phones, smart cards, vehicles and various sensors. Additionally, there are different localization technologies such as GPS, WiFi, Bluetooth and so on representing either symbolic or relative location details. Hence, it is quite beneficial to consider from all perspectives as mentioned above. The authors in [22] recommended a robust system for smart cities and urban planning using an IoT-produced Big Data analysis. The projected architecture comprises of four tiers which perform various operations such as collection, aggregation, communication, processing, and interpretation. The whole structure is settled using Hadoop technologies with spark to attain real-time processing.

- **Social Relation- Trace-Based Social Analysis:** Trace data can reflect social relationships of people. For example, people who visit together in a same location can be considered to have social relation. On the other hand, there is an issue in this case as people can often visit same places without knowing each other. Hence, it is required to discover appropriate trace-driven social relation correlating their locations. Yang et al. [23] recommended the Heterogeneous Human Walk (HHW) mobility model founded on the investigation of how overlapping community structure and individuals' local degrees promote to heterogeneous human mobility status. The HHW model creates the k-clique overlapping community structure via simple algorithms which captures the spatio- temporal uniformity of users' movement behaviors and provides statistical attributes of both social networks as well as individual user traces.

- **Privacy- Disclosure and Protection:** Trace data contains tremendous information of users such as their physical, social and personal activities with the location details. Hence, if those trace data are leaked or accessed by unauthorized persons then it may cause many harmful issues. Moreover, the existing localization techniques such as GPS, GSM, WiFi etc. and can leak users' ID or other types of personal information. Consequently, traces can disclose privacy of people connecting to uninvited services and applications. As a precaution, there has to be some proxy systems between users and applications so that they can protect privacy and make it secure communication. The authors in [24] applied sanitization technique, that enhances uncertainty to the mobility trace data and eliminates some sensible information. Moreover, authors compared various clustering algorithms and heuristics which can be applied as inference attacks, and assess their effectiveness for the identification of point of interests, as well as their flexibility to sanitization schemes such as sampling and perturbation.

2.5 Applications of smart cities

The trace data attained from various fields can help a smart city in numerous ways. Clearly, mobility data can optimise traffic systems of a city, social and personal data can improve public health, security and commerce, and region semantics can boost the urban planning and computing. The five major applications for a smart city are explained slightly below.

2.5.1 Smart Transportation

Trace data for vehicles often extract traffic information either explicitly or implicitly. Explicitly, it explains information about traffic conditions, road maps, transportation supply and so on. On the other hand, it implicitly displays information about traffic accidents, driving strategy, route navigation and many more. The vehicle trace data can also be helpful for guiding the popular tourist places, hotels, shopping malls and also for best travel route among various locations in a city. Moreover, their trace data can help to fulfil the future demand of public transportation resources with proper prediction of allocation or dispatching [15]. The various benefits of smart transportation for users in a smart city have been described below.

- **Smart Parking (e.g. SFpark):** Better parking availability suitable with price and proximity.

- **Smart ticketing (e.g. Oyster card, Suica):** Faster and easier payment for transport services.

- **Real time journey planner (e.g. Citymapper, moovit):** Easier travel planning in real time basis from one location to another.

- **Command & control centre (e.g. Minnesota Urban Partnership Agreement):** Transmits in-journey information by means of Variable Messaging Systems (VMS) for clear travel expectations and route plans.

- **Bicycle sharing service (e.g. Barclays Cycle Hire, Vélib', Bicing):** Provides alternative solution of travel offering 'grab and go' bikes all over the city.

- **Car sharing service (e.g. Lyft, Zipcar):** Provides 'grab and go' cars all over the city.

- **Taxi booking service (e.g. Hailo, Uber):** Offers taxi services any time and any place [14].

2.5.2 Smart Urban Planning

Urban planning is another important use of smart city. The trace data can help for urban planning in many ways. Firstly, it can support to plan for infrastructure designing and distributing in a required areas of city. This is possible because of information attained about visiting frequencies of people. Additionally, trace data predict the usage of lands too. That is why, the urban planning can be evaluated whether it is excessive or less according to the public demands. For example, number of vehicles on the road can predict to optimise the traffic systems [15].

2.5.3 Smart Public Health

Trace data can also be useful for mining records of patients, their social and private activities, thereby handling their diseases in a systematic way. Similarly, many contagious diseases can be controlled with the prediction of co-presence and density of people via tremendous amount of trace data. By monitoring daily user behaviour, it is easier to make people alert from obesity and many types of chronic conditions. Hence people can be healthier and improvise their work hours and exercise schedule tracked by adequate use of their trace data by service providers [15].

2.5.4 Smart Public Security

User trace data can detect social activities, abnormal crowd of people, criminals' tracks, and reaction of people in disasters and so on. Moreover, the daily activities of people normally have repetitive and similar routines and hence their unusual activities can be detected with the traces. The criminals and other gangs, suspicious as well as lost individuals can also be located by means of their trace data. As a result, the public is secured either individually or in society [15].

2.5.5 Smart Commerce

The field of commerce can also be improvised with the user trace data. The trace of frequently visited places of people can be very helpful in many ways such as check-in information, which contains both the cash flows and business affairs. Hence, many commercial advertisements can be fixed according to peoples' movements. The trace data, therefore, can help to fix the advertisements timing and place enhancing the business policies. Additionally, the user trace in shopping malls can predict duration and interest of people in specific orders as well [15].

3 USER MOBILITY MODELS

This chapter describes different mobility models of users with various features and types. Besides this, it covers some important mobility parameters that were investigated during this thesis work.

3.1 Types of mobility model

The study of patterns of user mobility has always been a challenging task in the last decade. Since mobility is very difficult to design and implement, one should consider various factors such as velocity, distance, longitude, latitude and so on of users' relative position and time at an individual level. With the growing popularity of mobile devices, many researchers got engaged in developing mobile applications, systems and various services in recent years. As a result, there has been given less priority on designing realistic user mobility leading towards poor performance on mobile systems. Nowadays, urban infrastructures are mostly equipped with real-time location-aware devices such as mobile phones, GPS, smart cards, etc., which provide huge amount of real-time location traces thereby supporting to develop various user mobility models.

In Mobile Ad hoc NETworks (MANETs), there is limited amount of real trace data and researchers often have to depend on synthetic data. On the other hand, Vehicular Ad hoc NETworks (VANETs), which is also a subset of MANETs, provides huge traffic traces representing users' strict patterns of movements in different time intervals and geographical spaces. One of the main differences between VANET and MANET is that mobile nodes move randomly whereas vehicles do not because nodes (vehicles) are already powered. Moreover, their communication time is bounded and speed is circumscribed by the traffic control system. Due to nodes high mobility nature and recurrent topology fluctuations, there is big effect in overall network performance in VANET. Hence, for the reliable network performance, proper management of mobility in vehicular networks on various models are very essential. VANET consists of two kinds of communication: 1) vehicle-to-vehicle (V2V) and 2) vehicle-to-infrastructure (V2I) communication. The V2V is a direct or multihop communication among vehicles whereas V2I characterises communication between vehicles and foundation of Roadside unit (RSU). The V2V is related with Dedicated Short Range Communication (DSRC) with short bandwidth and high performance. Similarly, V2I possesses GPRS/3G, WiFi or WiMax [25].

The trace-driven taxicab mobility models contain detailed information, which is derived by driver's behaviours and also passenger's destinations, resulting in too much overhead. Firstly, the taxicab driver's path of movement can be limited to mostly popular places demanded by passengers at a fixed time interval. Secondly, the patterns of mobility of taxicab can differ on the type of nearby regions such as roads, close vehicles, climate, traffic lights and many more factors. Finally, the mobility patterns of a taxicab can be also different in different traffics of a daytime or night-time of a day. Hence, the mobility patterns of VANETs are very complex in nature. The trace-based mobility models are experimented on real life systems which possess accurate information. Besides this, they contain enormous participants and sufficiently large observation period. But synthetic-based mobility models try to extract realistic modelling of user movements without using real traces of data [26]. Mainly, the mobility models for VANET can be classified as given in Figure 3-1 below and described in the next sub-sections.

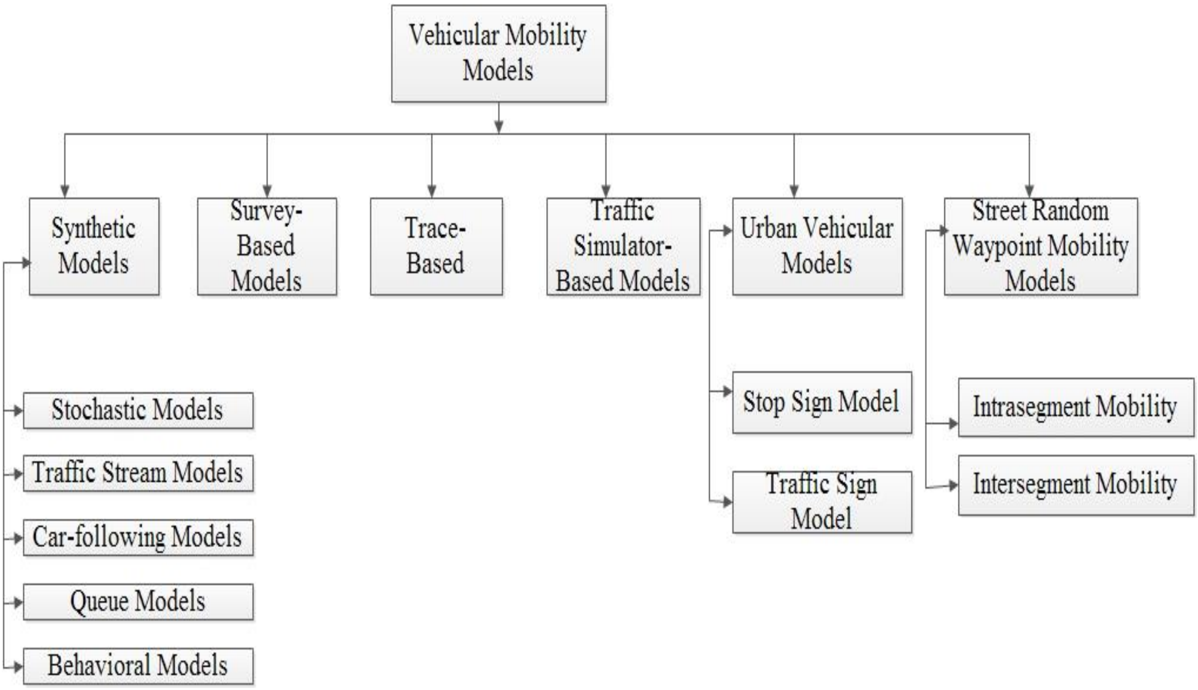

Figure 3-1 Types of vehicular mobility models

3.2 Synthetic mobility models

A mobility model should simulate the movements of users in a realistic way. That means, there should be varying speed and direction of users within a specific time interval. The movements of users may not be in just a straight line with a uniform speed during the simulation process. Here are some important types of synthetic mobility models described below.

- **Stochastic Models**: In this model, vehicles are moving arbitrarily selected speed and their mobility patterns resemble randomly moving nodes on a graph. Their mobility patterns are random in nature as they follow unplanned paths over the graph either independently or in a group [25].

- **Traffic Stream Models**: In this model, the vehicles’ movement are reflected as hydrodynamic phenomenon [25].

- **Car-Following Models:** The model analyses activities of each driver depending on the vehicles state. Here the state represents position, speed and acceleration of the nearby vehicles [25].

- **Queue Models:** This model takes account of two major things: the queued vehicles and the roads which are assumed as queue buffers [25].

- **Behavioral Models:** In this model, behavioural rules regulate every movement of vehicle. Those rules are executed by social impacts, reasonable conclusions or actions that obey a stimulus-reaction activity [25].

Some popular synthetic mobility models are briefly described below.

3.2.1 Random walk mobility model

This mobility model is one of the simplest mobility model producing just the random movement patterns of nodes or entities which are unpredictable in nature [26]. It is also referred to as Brownian walk [27]. In this model, an entity moves from one location to another by selecting a direction and speed in a random way. The ranges of such speed and direction are selected from pre-defined sets, [*SpeedMin, SpeedMax*] and [0, 2π] respectively. In this model, every entity moves in either a constant time interval t or a constant distance travelled d, where new direction and speed are calculated eventually. The movement continues for a definite amount of time intervals or distance, and the entire process is repeated a predefined number of times. As soon as entity reaches the simulation boundary, it gets reflected with an angle determined by the incoming direction heading towards new path.

There are various forms of the Random Walk Mobility Model which includes 1-D, 2-D, 3-D, and d-D walks. Among them, the 2-D Random Walk Mobility Model is considered of more importance since Earth's surface is also represented as 2-D form. In 1921, Polya concluded that a random walk on a 1-D or 2-D surface returns to the origin with a complete certainty, i.e. a probability of 1.0 [28]. That means, the random walk mobility model simulates the movements of entities around their origin points without caring the entities travelling away never to return.

The Random walk mobility model is simple to implement but it has disadvantages too. It is a memoryless model since it doesn't store previous data of speed and position to decide for future implementations. As a result, it generates unrealistic mobility patterns with random pauses and sharp turns. The Random Gauss-Markov Mobility Model can resolve this issues [26].

3.2.2 Random Waypoint Model

The random waypoint mobility model introduces pause times between changes in direction and/or speed, and was first projected by Broch et al. [29]. It is a very popular mobility model for many researchers and is considered as a strong base for developing other mobility models. In this model, an entity starts moving from an initial point within the simulation area by choosing destination point such as (x, y) and speed v randomly from a predefined range of speeds [*SpeedMin, SpeedMax*]. Then the entity travels into newly selected destination points with the given range of speeds. As soon as the moving entity finds its destination, it takes a certain pause for a certain time interval, and again the same process continues. It is worth to note that the mobility patterns of entity in random waypoint mobility model is similar to the random walk mobility model when the pause time is zero and [*MinSpeed, MaxSpeed*] = [*SpeedMin, SpeedMax*].

The relationship between speed and pause time of moving entities in the random waypoint mobility model is complex. Since fast moving entities having longer pause times results steadier network compared to slower entities with shorter pause times. Hence, for a better performance evaluation, proper mobility parameters should be taken into account in this model [26].

Although random waypoint model is popular among many mobility models, it has some disadvantages as well. There is clustering of entities in the centre of simulation area resulting decrease in average number of neighbours possessed by each entity. It is called ‘‘density wave’’. Additionally, there is vast decrease in average speed of entity as the duration of simulation is increasing. This effect is called ‘‘speed decay’’ [30].

3.2.3 Levy Walk Mobility Model

Levy walk mobility model also resembles a random walk model where step lengths, possessing certain probability distribution, are introduced instead of just the random steps. The nature of those steps is random and unsystematic. The distribution of displacements over all users can be estimated by a truncated power-law given below.

$$P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} \exp\left(\frac{-\Delta r}{\mathrm{k}}\right) \tag{3.1}$$

Where the described parameters are

- P is probability density function
- β is a user-dependent exponent and its value is 1.75± 0.15 (i.e. mean ± standard deviation)
- Δr is travelled distance or jump size
- Δr_0 is a reference parameter and its value is 1.5 Km
- k is a cut-off value which can vary in different experiments [31].

According to Brownian Motion (BM), diffusion of tiny particles can be defined as a mean free path (or flight) and a mean pause time between flights. Here, flight means the longest travel distance in a straight-line from one point to another without any directional change or pauses. The amount of aggregate displacement of any particle from its starting point is referred to as mean squared displacement (MSD) and such mobility possesses normal diffusion. Levy walks are one of those random mobility models which can describe abnormal mobility representing super-diffusion where MSD is directly proportional to t^{γ} if $\gamma > 1$. Similarly, when $\gamma = 1$, it is called normal diffusion, and if $\gamma < 1$, it is called subdiffusion [32].

Every step of a moving entity can be expressed as combination of four important parameters such as step S = (l; θ; Δt_f; Δt_p).

Where parameters described above are

- $l>0$ is length of flight
- θ is direction of the flight
- $\Delta t_f > 0$ is duration of flight
- $\Delta t_p >= 0$ is duration of pause time

The entity in the Levy walk model randomly selects the direction within a uniform distribution of angle from [0, 360], and its flight length as well as pause duration from probability distributions p(l) and ψ (Δt_p) correspondingly. Further, it selects α (flight length factor) and β (pause time factor) within a specified range as an important levy distribution coefficient [32].

3.2.4 Mobility models for terminal mobility in cellular systems

The study of mobility models for mobile terminals plays a vital role in the field of cellular systems. For instance, we can compute performance analysis of location updates of mobile terminals efficiently by means of its mobility models research work. Earlier researchers have studied and recommended various mobility models for cellular systems. Thomas et al. mentioned in [33] that mobile terminals demonstrate random speeds and autonomous random directions as they move on. Hong and Rappaport stated in [34] that mobile terminals display the random change in their speed and directions while crossing boundaries. Similarly, Guerin in [35] proposes that mobile terminals only change their speeds and direction when it reaches a certain random time or travels a random distance. The terminal mobility is denoted by two main parameters: the probability density functions (pdf's) of speed and the direction of terminal's motion. The three popular mobility models which are often used in analysis of performance parameters in cellular communication are described briefly below [36].

- **Model A:** The mobile terminals possess independent random variables as speed and direction which are constant. Besides this, the terminals' direction is distributed uniformly between (0, 2π) [33].

- **Model B:** In this model, speed and direction of mobile terminals are random and independent as in Model A. But whenever terminals cross boundary, they reproduce values of speed and direction in a random fashion which are autonomous to each other. The newly produced directions fall in the range of $(-\pi/2, \pi/2)$.

- **Model C:** In this model, terminals' speed and direction are random and independent too. But they are reproduced randomly after crossing a certain time interval or travelling a random distance. The newly produced direction fall into range of $(0, 2\pi)$. Similarly, the random time interval displays a mean T and a mean D for random distance [35].

3.2.5 SLAW mobility model

Recently, different measurement analyses of user mobility traces have invented many noteworthy statistical parameters of user mobility. To be specific these incorporate truncated power-law distributions of flights, pause-times, inter-contact times, fractal way-points, and heterogeneously characterized regions of individual mobility. Inappropriately, there is no any mobility model that efficiently apprehends all of the features mentioned above. But there is a new mobility model called SLAW (Self-similar Least Action Walk) that generates synthetic walk traces supporting all those features which are outlined below [37].

- **(F1) Truncated power-law flights and pause-times:** The distances of straight line trips of user with no any directional changes or pauses possess a truncated power-law distribution [31] [32] [38].

- **(F2) Heterogeneously bounded mobility areas:** There can be broadly numerous mobility zones for various people. But, usually human travel only in their limited regions of mobility [31].

- **(F3) Truncated power-law inter-contact times (ICTs):** The times passed between two consecutive contacts of a single person can again be demonstrated by a truncated power law distribution [39].

- **(F4) Fractal waypoints:** The destination points or also known as waypoints of user can be designed by the fractal points. That means, users often like to travel towards more popular places [40].

All the characteristics mentioned above from F1-F4 are essentially connected to each other. For example, the fractal waypoints (F4) persuade power-law flights (F1) and the heavy-tail flights inside a restrained area (F2) come with an outcome of truncated power-law ICTs (F3). All these features are significant for analysing network performances in cellular communications such as Delay-Tolerant Networks (DTN). In SLAW model, users tend to follow mobility patterns confined in a communal environment where they tend to interact as a common meeting point. Besides this, by designing power-law flights as well as fractal waypoints, this model can also articulate on regular and also spontaneous travel patterns existing in the regular users' mobility. SLAW model isn't limited for just investigating mobile network parameters but also widely used in urban planning, traffic estimating, biological and mobile virus breakdowns where user mobility is a vital key. This model is easy to execute like in the case of earlier random walk models because it takes limited parameters as input like size of walk-about area, number of users, and the Hurst value which is treated to produce the fractal waypoints [37].

3.3 Survey-Based Mobility Models

From surveys, we can extract important information of macroscopic mobility patterns. In this model, surveys are accomplished to obtain different mobility patterns by characterizing regular user activities in urban areas. The collected information from such survey can be school, work, restaurant, break time, duration of meeting time, frequency and similarly other everyday user behaviours. As an advantage, this model can project mobility models which are unable to solve mathematical derivations. On the other side, it can only represent a coarse grain mobility patterns instead of exact user movements [25].

3.4 Trace-Based Mobility Models

As described earlier, trace-based mobility models represent real mobility traces collected by different measurements campaigns. In this thesis, we are dealing with this model collecting data traces of taxicabs from two different major cities Rome and San Francisco from CRAWDAD (Community Resource for Archiving Wireless Data At Dartmouth) [12] [13]. It is a complicated procedure to design a mobility model and then authenticate it applying those traces into it. On the other hand, we can make it easier and save time by directly applying trace-based model which possesses real-time traces of mobility patterns. This model is recently getting popular because of availability of data traces of mobility patterns which are abundant and easy to access through various sources. As a disadvantage, it is complex to simplify mobility patterns which aren’t straight noticed by the traces but it can be resolved somehow with the help of few mathematical models [25].

By using traces, different researchers have succeeded to unearth mobility models which represent genuine movement patterns of walkers. For example, Teduce et al. in [41] extracted a mobility model with real data traces of campus LAN at ETH Zurich. Firstly, they prepared a simulation area allocated by small squares and then the possibility of shifts between adjacent squares was investigated by means of access point’s information. Similarly, Yoon et al. in [42] produced a probabilistic mobility model by merging WiFi and access points’ information of users along with the map where traces were perceived. They extracted a discrete time Markov Chain that displays information of both previous and current location as well as start and end point of path. Inappropriately, their research didn’t include the concept of correlations of nodes.

Kim et al. in [43] extracted locations of people from coarse grained wireless traces to spot common areas. They concluded that users’ speed and pause times obey a log-normal distribution. Additionally, they also derived that nodes' movement almost follows the path of roads which cannot be designed with a uniform distribution. Again they also missed out correlation of nodes here. Hou et al. in [44] designed a mobility model with a semi-Markov process that included connection time instants of access points and defined various time scales of users' mobility. Interestingly, this model survived to explain correlation between users' mobility which associated timed location prediction algorithm and succeeded to exactly foretell users' forthcoming locations. Similarly, Chaintreau et al. in [39] introduced the concept of inter-contact times of users' wireless devices by means of coarse-grained wireless traces along with the trail testbeds with iMotes. They concluded that the inter-contact times displays a heavy tail nearly a power law distribution unlike the usual so-called exponential distribution. But Srinivasan et al. in [45] explained that the inter-contact times of students in a university campus exhibited in the order of magnitude of few hours.

By analysing different researchers' study on trace-based mobility models, it is observed that the speed and pause time of users typically obeys a log-normal distribution instead of a uniform distribution, and the inter-contact times have to be designed by a power law distribution instead of an exponential distribution [46].

3.5 Traffic Simulator-Based Mobility Models

In this model, a detailed traffic simulator plays a role to take out the mobility traces of vehicles. The mobility models are constructed with the help of enhanced synthetic models and dominant authentication procedure by consuming real data traces or surveys of users' activities. There are few popular traffic simulator systems such as Parallel Microscopic Simulation of Road Traffic (PARAMICS), Corridor Simulation (CORSIM), Transportation Analysis and Simulation System (TRANSIMS) which have capacity to design microscopic traffic monitoring structure of urban environments. The traffic simulators have a huge collection of parameters resulting complexity for the alignment of this mobility model [25].

3.6 Urban Vehicular Mobility Models

In this model, the streets are considered as significant units which compel nodes to move in definite paths regardless of their end points. There are two types of urban mobility models which are described below [47].

3.6.1 Stop Sign Model (SSM)

This model describes the mobility of vehicles in the street containing stop signs in every crossroads. When vehicles appear at a junction point, they wait for a specific interval of time, and then proceed towards destination upholding a certain distance from nearby vehicles.

3.6.2 Traffic Sign Model (TSM)

In this model, vehicles follow a queue system with allocation of certain wait time to move from one intersection to other in the street. When a vehicle is waiting to cross any intersection point, rest following vehicles move further until the first one leaves [25].

3.7 Street Random Waypoint (STRAW) Mobility Models

In this model, a simple car following model is taken into consideration circumscribed in a genuine traffic scenario. In each direction there is at minimum one lane available where vehicles may travel. This model operates on a random street placement design in order to conclude preliminary locations of vehicles. In this model, a vehicle rests in a lane just in front of the intersection of street. Further, if there is already a vehicle in the same lane, rest consecutive vehicles are positioned just behind the remaining one. This model can be classified into following two types explained below [48].

3.7.1 Intra-segment mobility

This model obeys some regulations, by which vehicles alter its speed, are described below:

- Vehicles slow down its speed and stop before the intersection if there is no space on the part of street where it will continue.

- If there are already vehicles placed ahead of present vehicle, it slow down its speed following a distance formula written below [48]:

$$S = \alpha + \beta V + \gamma V^2 \tag{3.2}$$

In above equation, S is the next distance, V is the speed of present vehicle, α is vehicle length, β is reaction time, and γ is reciprocal of twice the highest mean deceleration of the next vehicle.

- The vehicle slows down its speed in front of intersection if it is red light or just a stop sign, and continues accelerating further if it is green light.

- If vehicle needs to take left or right turn after intersection, it slows down its speed in an appropriate way first, and it accelerates again traveling onto destined segment of road [25].

3.7.2 Inter-segment mobility

This model describes the mobility of vehicles at various intersections of the road which can be executed in two stages. Firstly, the mobility model has to implement admission control at every intersection. Secondly, there should be space for vehicle just after the intersection of the road [25].

3.8 User mobility parameters

The study of user mobility has attracted many researchers since recent years. During modelling of user mobility, its parameters are considered as major factors which are analysed step by step in various environments. The user mobility parameters can be categorised into three basic dimensions as described below [49].

- **Spatial dimension**: It describes the users' activities in the physical space such as distance covered by user.
- **Temporal dimension**: It characterizes the user mobility in a time-varying domain such as time expended by users in definite locations.
- **Social dimension**: It describes the regular activities of users which connects them to each other.

In this thesis work, mainly spatial and temporal dimensions of user mobility are focused. Among them, the major mobility parameters that are going to be investigated later are defined further below.

- **Speed:** Users follow different speed limits while travelling various places in their specific time stamps and it obeys a log-normal distribution [43]. Let's say, a walker's arrival time is t_i at location I_i with position (x_i ,y_i). When walker reaches new location I_{i+1} from I_i at time interval t_{i+1}, then its speed can be calculated as below [42]:

$$s_i = \frac{d_i}{e_i} \tag{3.3}$$

Here, s_i is speed of user, d_i is the Euclidean distance and e_i is the elapsed time.

- **Pause time:** Walkers often change direction while travelling from one location to another and the pause time is defined as the time interval of a walker expending in every end point before its directional variations [32]. It also typically exhibits a log-normal distribution and the path of mobility of users intimately resembles the path of streets and walkaways. The pause time of users can be expressed in terms of speed, distance and duration of travel while travelling from one location point to another in specific time intervals which is explained below [43].

$$p_i = e_i - q_i = e_i - \frac{d_i}{s_i'} \tag{3.4}$$

Here, p_i is the pause time, e_i is the elapsed time, q_i is the duration of travel, d_i is the Euclidean distance and s_i' is the average speed of the user. Moreover, in [43], the author has explained that the speed of users appeared to be slower due to presence of pause times and it was higher due to short elapsed time, e_i in the traces.

- **Probability of users' returning to the same point:** It is a probability of users' returning to a specific position where it was visited already. Users often like to travel same place from time to time and the frequency of visits can depend on the popularity of location. It is possible that users can return to same place number of times and in [31], it has been mentioned that users returning tendency applies within 24 hours or 48 hours of their first visit.

- **Travel time and travel time densities:** The travel time is defined as time taken by users to travel from one location to another without any pause times. Whenever walker travels through congested areas such as crowd, markets, campus or any highly populated regions, then they display higher pause times and lower travel times. On the contrary, if there is less crowd in the surrounding of users, then their travel time is higher and pause time is lower [49]. In case of vehicles' mobility, the travel times often get affected by the high traffic presented in the roads [27]. Hence, if there are more congestions in the roads, more will be its travel time and densities.

- **Longest and shortest route:** A route is the path followed by moving entity during traveling from one location to another. Their travelling route can be longer if there is more congestion on the way with higher travelling time and it can be shorter route when there is lower travelling time. In case of vehicles mobility, the length of route during traveling also may vary depending on the driving skills, knowledge about routes and behaviour of the driver [6].

- **Step:** When a walker moves from one location to another it may stop for a definite time interval before changing to its new position which is defined as pause time as mentioned earlier. The step, hence, can be defined as total distance travelled between two consecutive pauses of users moved in a straight line between any two locations [32].

- **Probability of occupancy:** The occupancy of taxicabs signifies two major characteristics: firstly, passengers' movement of getting in and out from the taxis in definite time stamps and location points and other is taxis' take-up/take-off activities. Moreover, occupancy of passengers consists of two values: 0 or 1 where 0 implies the taxicab is vacant and 1 means it's rented by the passengers. The probability of occupancy of taxicabs describes how often it was moving with or without passengers. Its occupancy information is attained by installing weight sensors or implanting GPS tools into its meters [3]. In this thesis, we analyse this parameter only in case of San Francisco's dataset as there is no such information available in Rome City's case.

- **Spatial mean centres:** The movements of taxicabs or any vehicles can often be stationary at some specific places e.g. parking places, railways stations, hotels, traffic jams, cross points at road segments and so on. The spatial mean centres of taxi driver's trips are the static points of their GPS traces which are related with the temporal as well as thematic characteristics such as standing duration, initial and stopping timestamps. The spatial mean centre formula is defined later in eq. (6.1).

- **Hotspot:** Hotspots are frequently visited places in city regions where mostly take-up and take-off processes of users occurs. The users' movements observed in hotspot areas can depict the spatial mobility model of the entire city [1].

4 VEHICULAR COMMUNICATION

This chapter mainly focuses on concept of vehicular communication, its applications and their routing protocols. Further, it describes the existing issues and contributions outlined by various researchers related with this thesis work. This thesis work is dealing with the extraction of various outdoor mobility patterns of taxicabs in an urban platform that engraves mainly the VANETs and its areas.

4.1 Vehicular Ad-Hoc Network

Along with the rapidly increasing number of vehicles on the streets, there is a demand of efficient and powerful technologies to support passengers providing crucial applications such as critical weather awareness, controlled driving assistance etc. Ad-Hoc Network can be defined as a wireless communication network formed by group of mobile nodes and wireless transcribing devices in a decentralised fashion. Further, those networks which are totally itinerant in nature and need less or no any infrastructure are termed as MANETs (Mobile Ad-Hoc Networks). As mentioned earlier, VANET (Vehicular Ad-Hoc Network) is a subset of MANET that incorporates communications between adjacent vehicles (V2V) and vehicles to roadside (VRC) or vehicle to infrastructure (V2I). VANETs contain some unique features than MANETs such as limitations of road pattern, bigger network size and energy source, dynamic topology and mobility models, localization techniques and so on. In VANETs, every vehicle is equipped with a device which can accept and transmit information among wireless networks e.g. traffic data, road blocking status, parking, tracing of locations, fuel stations and weather reports [50].

4.2 Applications

VANETs have various important applications such as traffic management, safety applications, toll services, location awareness applications and so on. Among them, few popular applications of VANETs are described briefly below.

4.2.1 Intelligent Transportation System (ITS)

ITS exhibits basically traffic management based applications such as global positioning system, traffic monitoring, traffic collision detection, routes information exchange etc. Whenever a vehicle is travelling on the streets, it exchanges information of traffic status on its way to the central authority which monitors and control traffic flow in order to provide optimum traffic signals to all vehicles nearby in VANETs. Further, in case of accidents on roads, the vehicles often share this data via traffic sensors to other oncoming vehicles or any emergency units like nearby hotels or gas stations. This way, vehicles' information sharing in

VANETs can help to avoid future accidents or any traffic collisions, to maintain best road routes and its route table updates and many more by means of geocast or broadcast routing systems [50].

4.2.2 Comfort Applications

This kind of applications provide comfortable features to passengers or any clients on the web using unicast routing protocols in VANETs. For instance, passengers in the vehicles can play games or listen to music integrating VANETs to the web. But, there can be more technologies adding to VANETs to ease V2V communications such as introduction of newer VANETs can enhance the range and qualities in network signals in rural areas [50].

4.2.3 Collision Avoidance

The regular communication between vehicles and roadside systems can save many accidents on the streets. Whenever a vehicle detects an accident on a specific place on the roads, it immediately sends information about it to the oncoming vehicles. Hence, all previous vehicles can be alerted controlling the speed and sending information to emergency units or nearby vehicles using VANETs. By means of broadcast of information about accidents on the roads, vehicles can be alerted to make better route decisions and can foretell future accidents or save many lives [50].

4.2.4 Cooperative Driving

This application is also related with safety of drivers and passengers on the roads. It provides many services such as turn conflict warning, violation warning, curve warning etc. If the drivers follow all these warnings honestly then the rate of accidents on the roads can be highly reduced. The name of this application itself describes about cooperation between the driver and the services provided by it which can play very important role to avoid road accidents [50].

4.2.5 Traffic Improvement

This application is based on congestion control of vehicles on the roads. Whenever a vehicle senses too many vehicles nearby it or any fluctuations of their speed, then it transfers information about it to all oncoming or surrounding vehicles. Hence, all the vehicles approaching to that congested location can be alerted and make better route decision avoiding the congestion on the roads [50].

4.2.6 Payment Services

It is one of the popular applications of VANETs based on E-payments such as toll collection, parking payment, gas payments etc. This application is quite helpful for saving time avoiding to stay in big waiting line and deceleration of vehicles speed [50].

4.2.7 Location-based Services

This application is based on GPS system which can easily inform drivers or passengers in VANETs about neighbouring hotels, restaurants, cafes, gas stations etc. tracking the locations on real time [50].

4.3 Routing Protocols

The routing protocols of vehicles can be classified based on their broadcasting technologies as below [51].

- **Topological based routing protocols:** They share links information from the VANETs forwarding the packets which can be either proactive or reactive in nature such as FSR, TORA etc.

- **Position based routing protocols:** These protocols use the geographic positioning systems to transfer the data of location to its nearby vehicle from one hop to another. It doesn’t require any knowledge about map to share information about their relative locations. This routing protocol can further be categorised as position based greedy V2V protocol and Delay tolerant protocol.

- **Broadcasting routing protocols:** Whenever vehicles need to transmit messages far from its range, they use broadcasting routing protocols. These protocols use flooding techniques to send messages among vehicles which results in waste of network bandwidth and replications of data. There are many broadcasting protocols such as BROADCOMM, V-TRADE etc.

- **Geocast routing protocols:** These routing protocols are used when vehicles need to transmit their location information from one to many vehicles inside a particular geographical area. This protocols use multicast techniques for packets delivery.

4.4 Challenges

In this section, we primarily describe some of the research issues and its outcomes outlined in the field of urban computing and mobility modelling of vehicles which is similar to this thesis work.

Tang et al. [52] employs a model called Support Vector Machine (SVM) to predict the traffic situations in specific regions of Hangzhou, a city in China, from various transportation and weather characteristics. First, the author analyses various traffic characteristics such as speed, pick-ups and drop-offs of passengers along with their temporal changes in given regions and later in the whole city. Then, the correlation behaviour is tested among those attributes and finally SVM model is applied. The author concludes that data statistics, visualization and machine learning tools can be fruitful to display the city's transportation scenarios by processing various types of urban data.

Cunha et al. [53] presents a study of interactions of vehicles in VANETs using two datasets: San Francisco and Rome. Further, they extract properties and behaviours from the mobility traces using statistical methods, graph theory and network analysis. From the detailed analysis, they examine that taxi passengers have similar routines to visit various regions and exhibit common interests and destinations building communities on the graph. Similarly, Veloso et al. [54] analyses 177169 taxi trips collected from Lisbon, Portugal to find the relationships between pick-up and drop-off locations and the effect of area types in taxi services. Further, they observed that each taxi trips were relatively random and with given information of time of day, weather situations, areas along with current pick-up location, only 5% of all trips were predictable. Zheng et al. [3] detects flawed urban planning with the GPS traces of 30,000 taxicabs in Beijing, China which results mainly two things: 1) regions with salient traffic problems and 2) the linking formats and correlation of them. Besides this, their results can help to appraise the planning of city roads and subway lines to improve any flaws in future.

Li et al. [1] proposes an improved prediction method called ARIMA (Auto-regressive integrated moving average) to estimate the spatial-temporal patterns of passengers in a hotspot. This method helps taxi drivers to find their next passengers by decreasing the time cost up to 37.1% and length of vacant driving distance by 6.4%. Similarly, Ge et al. [4] presents a method to analyse sequence of pick-up points as well as possible parking locations to taxi drivers which helps them to save waiting times for their next passengers. Schäfer et al. [55] computed real-time traffic information by using GPS-equipped vehicles in few European cities. The author here deliberates congested roads having vehicles' speed less than 10km/hr and reveals further that display of traffic conditions in a city can help to spot congested as well as jammed road segments. Yuan et al. [6] presents a method of building a graph where nodes are landmarks. Those landmarks are road segments mostly travelled by taxicabs. Further, they propose a technique to split a day into series of time segments related with the variance and entropy of travel times between those landmarks. This allows to spot the distributions of travel times among all landmarks. Moreover, Yang et al. [56] and Wong et al. [57] have performed analyses to enhance the taxi services in case of congested areas on the roads.

All of the works and accomplishments mentioned above represent mostly the methods for examining traffic conditions but not essentially modelling or predicting the mobility patterns of vehicular devices in urban scenarios. In our thesis work, we focus on modelling the vehicles' mobility patterns, by using various statistical parameters such as speed, pause time, travel time etc. and draw a conclusion with the approximate model of taxicabs mobility.

5 MEASUREMENT DATA

This chapter deals with the description of data used in this thesis work. Further, the types of data parameters, their map plots showing the traces for few illustrative users, and finally the procedure of data handling in required format are explained.

5.1 Dataset Description

There are two datasets used in this thesis:

1) First dataset contains mobility traces of taxi cabs in Rome, Italy
2) Second dataset consists of mobility traces of taxi cabs in San Francisco, USA which is mentioned in [12] [13].

Both these datasets are imported from CRAWDAD (Community Resource for Archiving Wireless Data At Dartmouth) which is responsible for sharing data from real networks and real mobile users across the various research communities around the world. CRAWDAD is conducted by Dartmouth College under a grant from the National Science Foundation, USA. Additionally, they have certain terms and conditions to follow for utilising data for the research activities which are described below [58]:

- They provide a nonexclusive, non-transferable license to utilize the data for commercial, educational, and research activities only.

- It’s not allowed to turn around the anonymization procedure to recognise particular MAC addresses, IP addresses, telephone numbers, or other attributes, or to recognise their real locations. But it is allowed to use the header information in the packet traces.

- It’s mandatory to acknowledge the origin of data in any publications stating on Licensee’s usage of it.

- Dartmouth explicitly holds the right to utilise the Data by its faculty, staff and researchers, for educational and research activities. Also, they have the right to offer Data providers with statistical info about licensee’s approach to and usage of the Provider’s Data and the Licensee’s name and address.

- Similarly, they deliver the Data ‘’AS IS,’’ with no any warranties or technical supports, and they are not responsible for any damages whatsoever caused by the usage of Data.

- Dartmouth doesn’t make any warranties, directly or indirectly related to the Data, counting any warranties of merchantability or fitness for a specific reason.

Both of the datasets used in our thesis work are described further below in order to shed some light on the types of data parameters, values, ranges, definition and so on.

5.1.1 Rome Dataset

This dataset consists of GPS coordinates (latitude and longitude only) of 316 taxicabs collected over 30 days in Rome, Italy and it is approximately 392 MB of compressed data. All those taxi drivers work in the centre of Rome in Italy and these traces describe their positions gathered in every 7 seconds. Moreover, each taxi driver was equipped with a tablet which intermittently fetched the GPS locations and transmitted this information to its central server. The trace is a txt file organised as: DRIVER_ID; TIMESTAMP; POSITION where DRIVER_ID is an integer, Timestamp contains date and time, Position is organised as POINT (latitude, longitude). The trace is sorted on the timestamp. The parameters stored in the traces of this dataset are further described below [12].

- **Taxi Driver's ID**: First parameter in this dataset is ID of all taxicabs' drivers which signifies the unique identity among them.
- **Date and Time:** The second parameter in the dataset is Date and time where each value is in the format of Year-Month-Day and Hours: Minutes: Seconds together in a combined form. The starting point of date and time in this dataset is 2014-02-01 00:00:07.39166 and the last one is 2014-03-02 23:59:589.43143. For the easy processing, we convert these values into one single entity as time in hours.
- **Latitude:** It is in decimal degrees' unit and has been extracted from Position field of the dataset.
- **Longitude:** It is also in decimal degree's unit and processed as similar like Latitudes of the dataset.

5.1.2 San Francisco Dataset

This dataset consists of GPS coordinates (again, latitude and longitude only) of 536 taxicabs gathered over 30 days in the San Francisco Bay Area. This dataset is also in text file format containing approximately 91 MB of compressed data. These mobility traces are presented by the Exploratorium- the museum of science, art and user perception [59] via the cab spotting project [60] which is often doing investigations about the mobility nature of taxicabs and hence supporting analysis of economic, social, political and cultural matters discovered by their traces.

In each taxicab, there has been a GPS receiver equipped into it to reveal the location data and transmit it to the central server in about every 10 seconds. Moreover, this system dispatches information about the cab call number, location and its occupancy status. This dataset comprises the list of all taxicabs and for each taxicab its mobility trace is in a distinct ASCII file, e.g. 'new_abboip.txt'.

The structure of each mobility trace file contains several lines of data where each line resembles as [latitude, longitude, occupancy, time], e.g.: [37.75134 -122.39488 0 1213084687], where latitude and longitude are in decimal degrees, occupancy exposes if a taxicab possesses a fare (1 = occupied, 0 = free) and finally time is in UNIX epoch system. The starting point of date and time in this dataset is 2008-05-17 and the last one is 2008-06-10 [13].

5.2 How we processed the data?

As explained earlier, first of all, we imported two datasets of Rome and San Francisco cities from [12] [13] originally in txt file formats. The main steps taken for processing those two trace files are given below.

5.2.1 Processing of Rome dataset file

In case of Rome dataset, the size of the file appears to be huge to process at once so it has been broken down into many smaller txt files. Then, all the txt files were imported into MATLAB, which is our main platform for analysing datasets for this thesis work, and converted further into suitable formats using few functions. Moreover, the data and time fields were converted from Year-Month-Day hh:mm:ss into single unit called time in hours, Latitudes and Longitudes are suitably arranged first in proper formats and then converted from decimal degrees into xyz (Cartesian) coordinates in meters. Finally, all the taxicab drivers' data were stored in separate containers called cells in MATLAB on the basis of their unique value of ids.

5.2.2 Processing of San Francisco dataset file

In case of San Francisco dataset, the trace files are already provided in the format of multiple text files. Each text file here represents the unique taxicab driver's mobility traces consisting latitudes, longitudes, occupancy and time as described earlier. Besides this, there is one text file containing information about each taxicab driver's ids. There are few differences between Rome and San Francisco datasets such as formats of Date and time fields, Latitudes and Longitudes fields, taxi drivers' ids and occupancy.

In case of San Francisco dataset, the Latitudes and Longitudes are already in proper formats with degree as units. Moreover, there is one unique field here called 'occupancy' which was absent in Rome dataset. It contains binary values, that means, either 1's or 0's as integers showing the status of taxicab's occupancy by passengers. Finally, the last field in this dataset is time which is in the format of epoch time (also known as POSIX time or Unix time). The epoch time is a format for explaining instants in time, termed as the number of seconds which have passed since 00:00:00 Coordinated Universal Time (UTC), Thursday, 1 January 1970. It doesn't count leap seconds. Again, we convert this field into time in hours' format as in Rome dataset. Even though there are variations in data fields in Rome and San Francisco datasets such as date and time formats, occupancy, Taxi driver's ids etc., we have applied same techniques for processing and analysing both the datasets using MALTAB software because the final extracted parameters are the same: (x,y) local coordinates of each trace point and the corresponding time t.

5.3 Illustration of user traces

Figure 5-1 and Figure 5-2 display the traces of any five taxi drivers from Rome and San Francisco datasets, respectively, tracked for about 30 days. Different vehicles routes are shown in different colours. Many are overlapping, as the probability of visiting the same hotspots (points of high frequency) of several users is high.

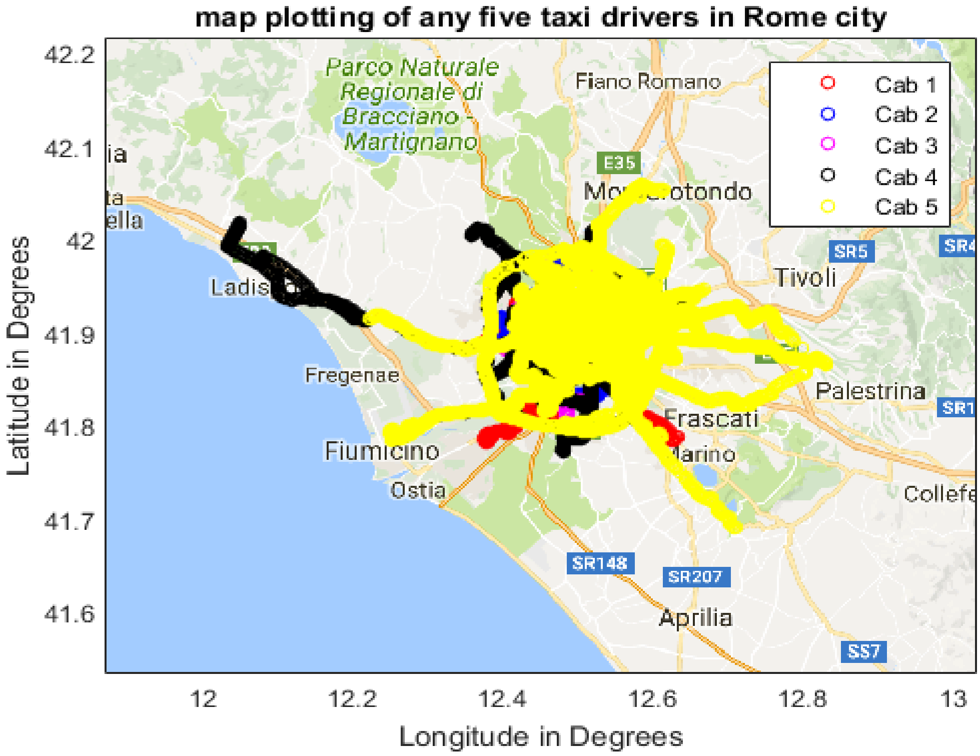

Figure 5-1 Map plotting of five taxi drivers' data in Rome dataset

Since, it is unclear to visualize all the users' traces in one map file, we present only five users' data here. The Latitudes and Longitudes in both Figure 5-1 and Figure 5-2 are presented in degrees. The map files in these figures are imported first from google online and plotted their traces in suitable formats into MATLAB.

Besides this, Figure 5-1 and Figure 5-2 describe show that users are travelling in various locations in a random way.

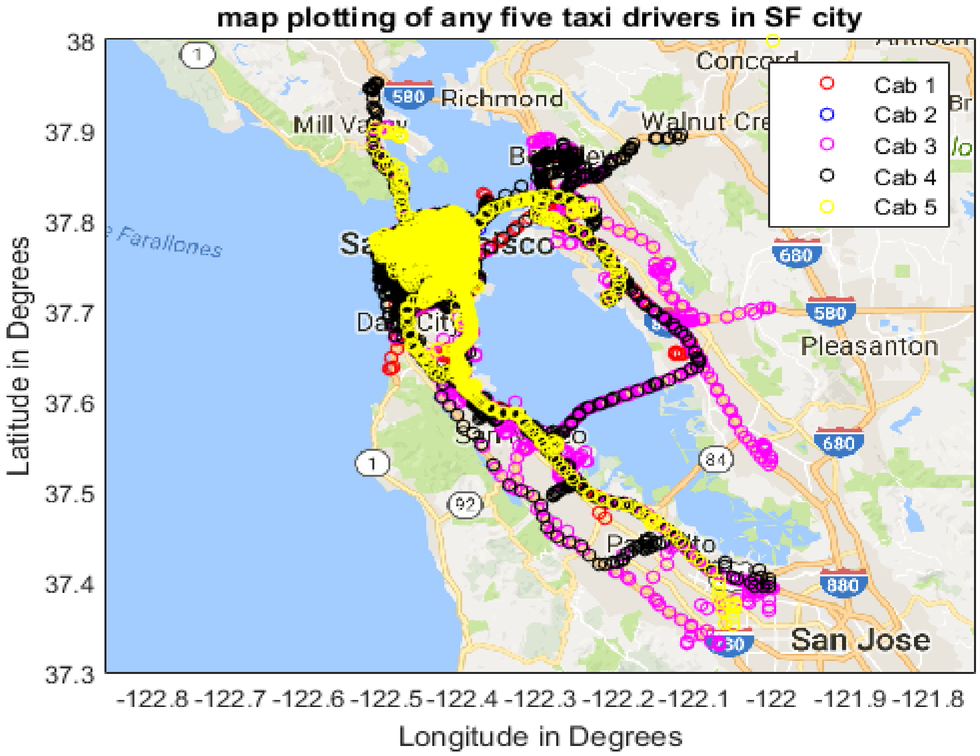

Figure 5-2 Map plotting of five taxi drivers' data in San Francisco dataset

Figure 5-3 and Figure 5-4 show few examples of the distance travelled (in Km) from the initial location over the duration of the measurements for Rome and San Francisco data, respectively. Three taxicabs are shown in each case; the curves have different lengths as the number of measured trajectory points per taxicab differs. The travelled distances in SF case are much longer (about 3 times longer on average) than those in Rome case, no doubt due to a larger served area by the taxis in San Francisco area (compared with Rome area).

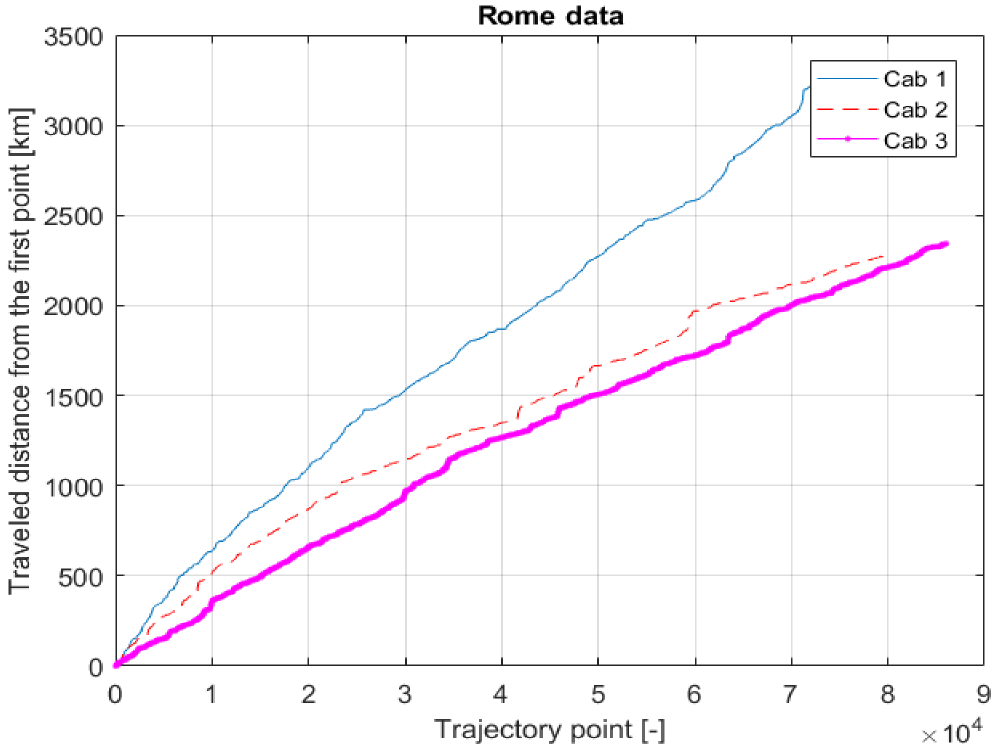

Figure 5-3 Illustration of the distance travelled from the initial point for 3 taxicabs, Rome data.

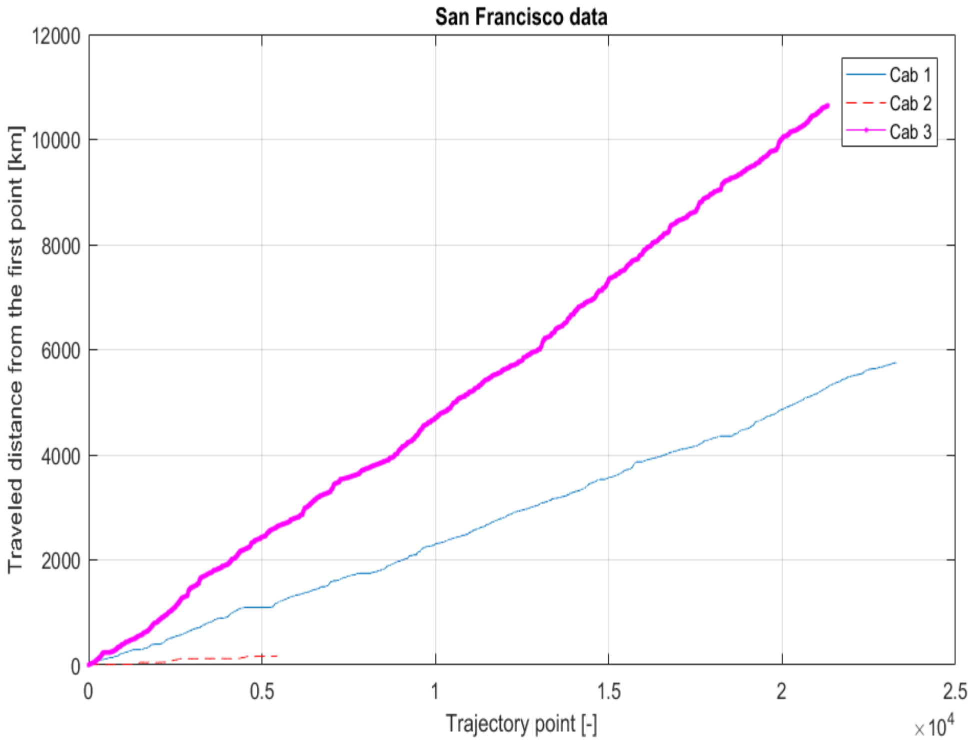

Figure 5-4 Illustration of the distance travelled from the initial point for 3 taxicabs, SF data.

6 DATASET ANALYSIS AND RESULTS

This chapter analyses, describes and compares the various mobility parameters computed from two datasets of Rome as well as San Francisco used in this thesis. The mobility parameters measured in this thesis work such as speed, route, step, travel time etc. are analysed from both datasets: Rome and San Francisco and compared. Besides this, the statistical results of all measured mobility parameters are described as a summary later in this chapter. Finally, the measured mobility parameters are compared with various theoretical distributions and fitted into suitable outdoor mobility model.

6.1 Speed

The histograms of the average speeds taken over all the taxicab drivers from Rome and San Francisco datasets, respectively are described below in Figure 6-1.

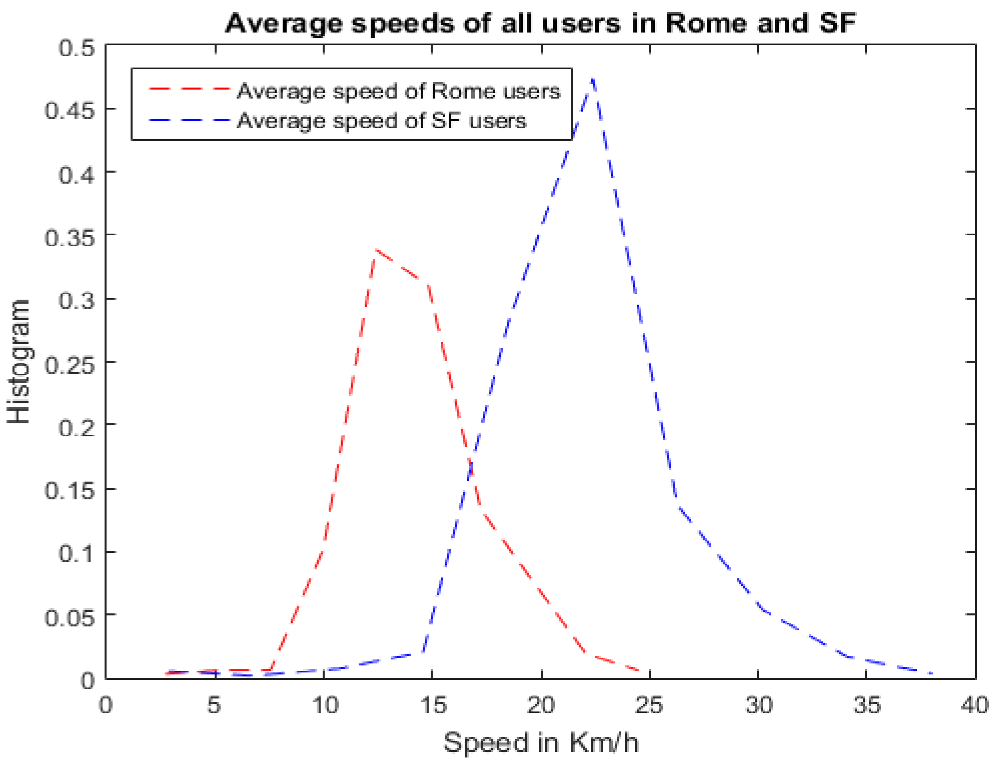

Figure 6-1 Average speeds (km/hr) of all taxicab drivers in Rome and San Francisco

In above figure, the x-axis represents the average speeds of all taxicab drivers from Rome and San Francisco datasets in Km/hour. Similarly, y-axis signifies the probability distribution function of their average speeds which is simply computed by dividing the measured values by its sum itself. The Figure 6-1 describes that the average speed of all users in Rome is lower compared to San Francisco's data. The fluctuations in taxicab's speed observed in above figure can be due to many reasons such as severe weather conditions e.g. heavy rain or snow, that reduces speed of vehicles resulting in smaller traffic volume. Furthermore, the street orientation is also another major factor for vehicles' speed variations. That means, the average speed in downtown regions can be smaller compared to freeways due to congestion of traffic. Besides this, the speed of taxicab drivers depends on different times of a day. For instance, there is heavy traffic in the office hours of weekdays, especially in morning time and evening time causing slower speeds. On the other hand, it is higher speed in weekends and other odd times of day or midnight.

The output patterns portrayed in above Figure 6-1 can be interpreted in three phases: congested phase, critical phase and free flow phase. As described earlier, the speed has decreased due to congestion of traffic first, then it reaches a peak point and finally it breaks down to higher speed range in less congested regions of city. The average speeds of all users in Rome dataset are measured as 14.19 Km/hour per day whereas it is 22.14 Km/hour per day in case of San Francisco dataset.

6.2 Pause Time

The pause times are computed as the difference of any two times over which the taxicab speed is close to zero (a threshold of 1km/h was chosen to allow for uncertainties in the measured GPS coordinates). The average pause times of all taxicab drivers from Rome and San Francisco datasets are described in Figure 6-2,where horizontal axis represents the pause times of users in hours and the vertical axis is the probability distribution function of their average pause times. Figure 6-2 further explains that the output patterns in case of Rome dataset look similar to an exponential distribution whereas San Francisco dataset look similar to a Gaussian distribution. However, a more thorough discussion about the distribution fitting are shown in Section 6.7.

As mentioned earlier, taxicabs often have to take pauses between source and destination depending on many scenarios. For example, a taxicab can be travelling on the street for mainly 3 reasons: i) picking up passengers to destination ii) travelling to taxi stand and iii) roaming randomly until called by passengers. For above first two cases, taxicabs often show pauses between source to destination but while roaming around in the street it can possess various travel sequences with smooth shifts.

As shown in Figure 6-2, the average pause time in Rome is 0.0023 hours (8 seconds).

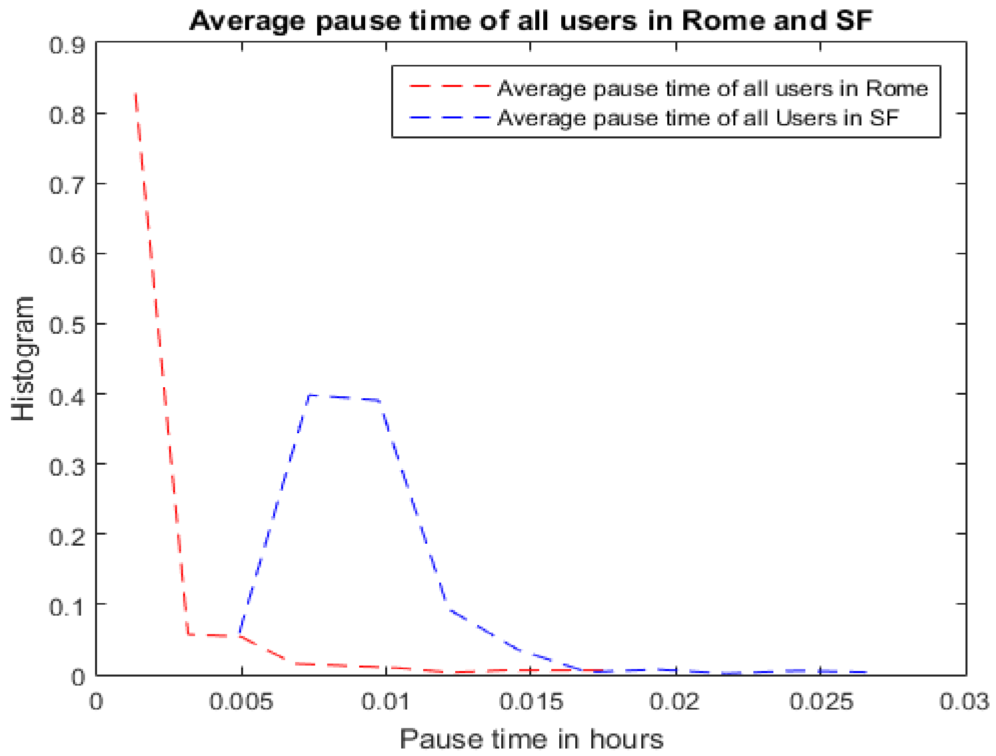

Figure 6-2 Average pause times (hours) of all taxicab drivers in Rome and San Francisco

Similarly, in San Francisco dataset, the average pause time of all users is 0.0091 hours (32.76 seconds). If we compare both datasets value in figure above, the pause times in case of San Francisco are higher.

6.3 Travel time

The average travel times (hours) of all taxicab drivers from Rome and San Francisco datasets are illustrated in Figure 6-3 . The x-axis in the figure below represents average travel times (hours) of all users and y-axis is probability distribution function for both datasets. The average travel times for all taxicab users from Rome and San Francisco datasets are 0.1288 hours and 0.3574 hours, respectively. That means, the average travel times in case of San Francisco is higher than Rome. Travel times of taxicabs while travelling from source to destination looks to be related to their pause times which can be observed by comparing Figure 6-2 with Figure 6-3.

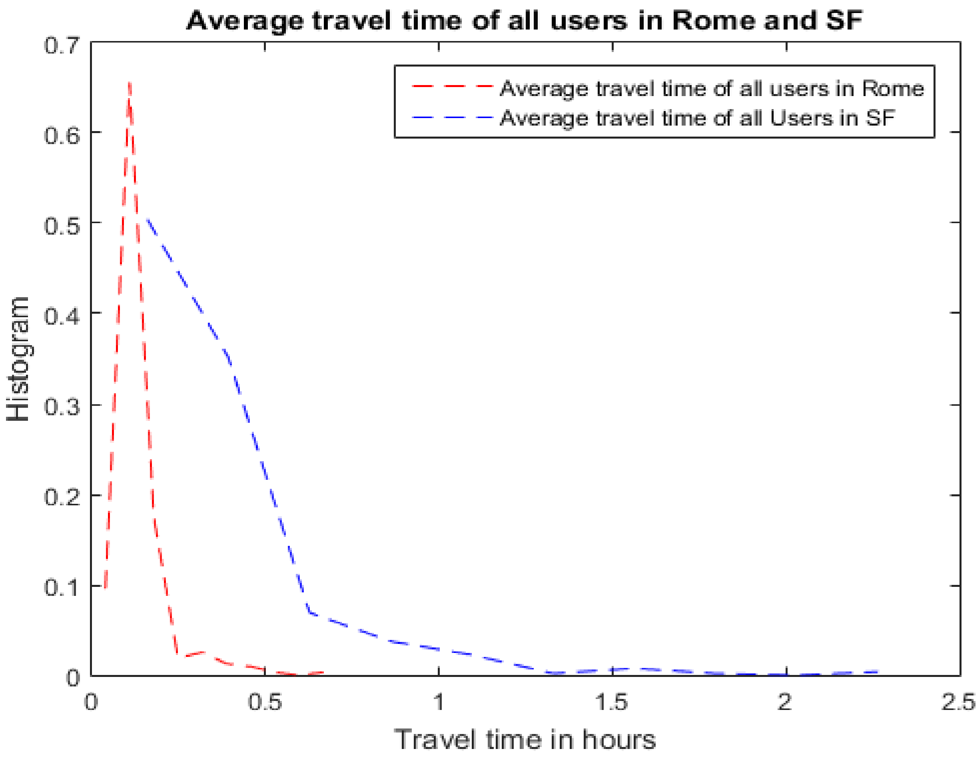

Figure 6-3 Average travel times (hours) of all taxicab drivers in Rome and San Francisco

6.4 Route

The steps (Km) of all taxicab drivers from Rome and San Francisco datasets are shown in Figure 6-4 below. The x-axis in the Figure 6-4 denotes the average steps (Km) of all taxicab drivers whereas y-axis is probability distribution function for both datasets. From the Figure 6-4, it can be observed that the patterns of average steps from both datasets look like obeying a Gaussian distribution. That means, paths taken by taxicab drivers in both Rome and San Francisco can vary randomly depending on the street layouts, weather condition, driver's skills or time of day or night. The average routes of all users per day from Rome dataset is 398.5448 Km and it is 429.3634 Km in case of San Francisco dataset.

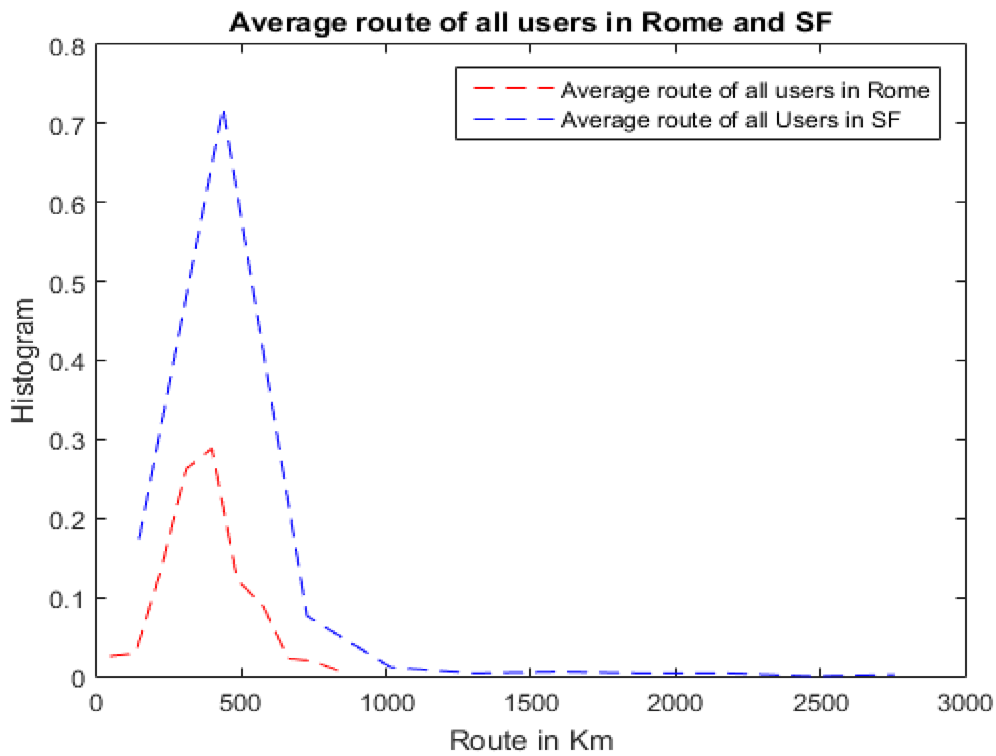

Figure 6-4 Average routes (km) of all taxicab drivers in Rome and San Francisco

6.5 Step

As described earlier, steps are the distances travelled by users between two consecutive pauses. In Figure 6-5, the average steps (Km) of all taxicab drivers from Rome and San Francisco datasets are described. The horizontal axis in Figure 6-5 represents the average steps (Km) of all taxicabs of both datasets whereas the vertical axis is their probability density function. Besides this, the output patterns in Figure 6-5 explain the average steps of all users from both datasets seem to obey an exponential distribution. The average steps of all taxicabs per day in Rome is 0.2448 Km, whereas it is 1.6222 Km in the case of San Francisco dataset. Hence, the average steps of all users are higher in case of San Francisco dataset. Furthermore, the steps are dependent on the pause times, travel times, direction of travelling and the distance covered by the users between source and destination.

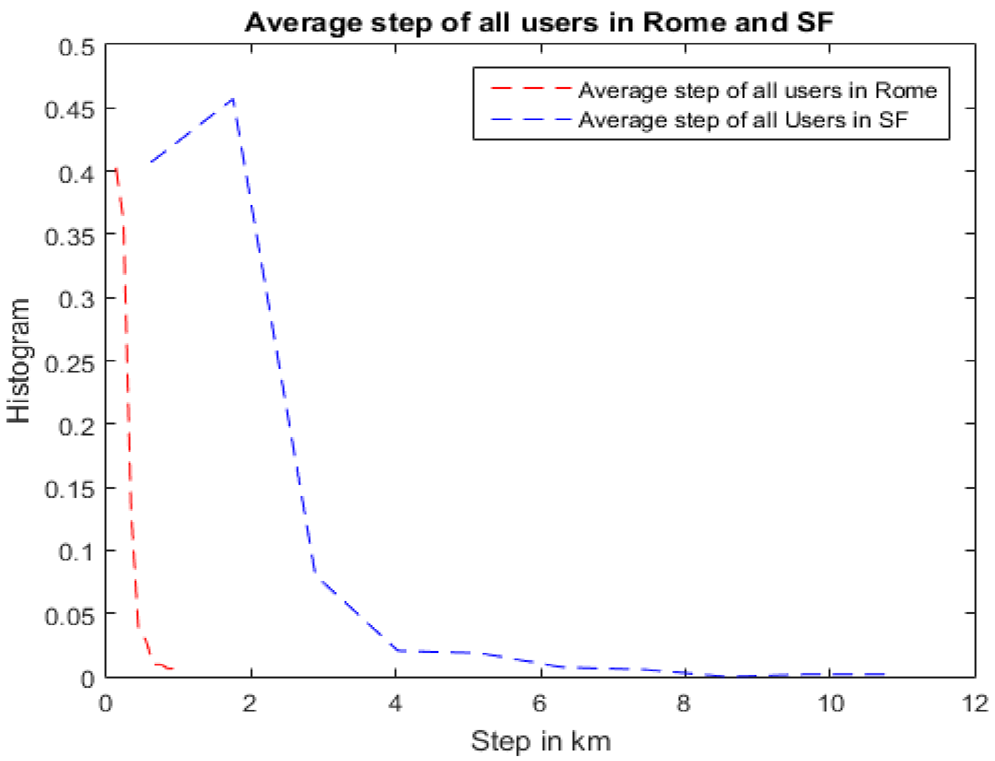

Figure 6-5 Average steps (km) of all taxicab drivers in Rome and San Francisco

6.6 Probability of returning to the same point

Generally, most of the users often like to return to the previously visited places within 24 hours to 72 hours of their travel time. Two examples of how far from their spatial means a user is likely to travel are shown in Figure 6-6 and in Figure 6-7 for Rome and San Francisco data respectively. In Figure 6-8, the probabilities of returning to the same point of all taxicab drivers from Rome and San Francisco dataset are described. It is common that often taxicab drivers pick up and drop off passengers from one location to another in hotspots such as hospitals, shopping malls, colleges, schools, tourist areas, bus or railway stations etc.

In this thesis work, we have computed users' probability of return to same point from Rome and San Francisco datasets by using their spatial centres. The spatial mean centres are the stationery spots of users in frequently visited areas. The spatial mean centres can be calculated as given below [61].

Here, if taxi trip $t = ((x_1, y_1), \ldots, (x_n, y_n))$, then its spatial mean centre can be defined as:

$$mc_{trip}\,(t) = \left(\frac{1}{n} \sum_{i=1,..n} x_i \;,\; \frac{1}{n} \sum_{i=1,..n} y_i\right) \qquad (6.1)$$

Again, if taxi trips $T = (t_1, \ldots\, t_m)$ in a single day, then its spatial mean centre is given as:

$$mc_{trip\_day} = \frac{1}{m} \sum_{t \in T} mc_{\mathrm{trip}}(t) \qquad (6.2)$$

In above equations, t is denoted as taxi trip, x and y are coordinates of trip position, mc_{trip} is the mean center and n is the total number of coordinates of trip position.

First, the mean spatial centres of all users from both datasets were calculated. Then, the distance between each spatial centre and overall mean of them were computed. Finally, the output was plotted together with their probability distribution function as shown in Figure 6-8 below. The patterns of probability of return of users from both datasets in Figure 6-8, explains that it follows the exponential distribution function.

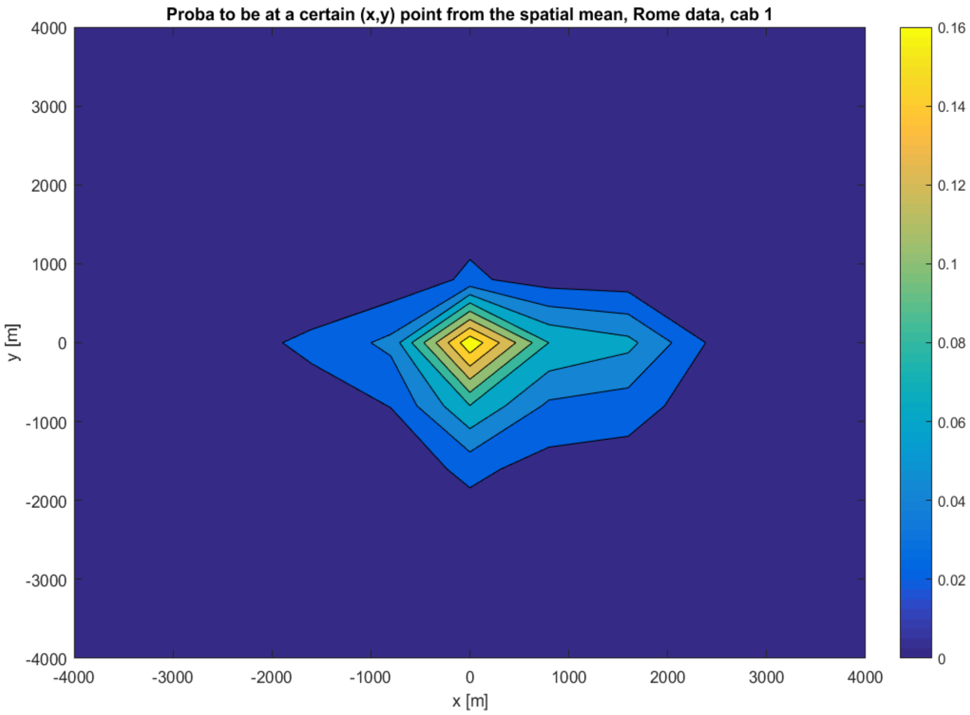

Figure 6-6 Probability to be at certain point from the spatial mean of one's trajectory, Rome data

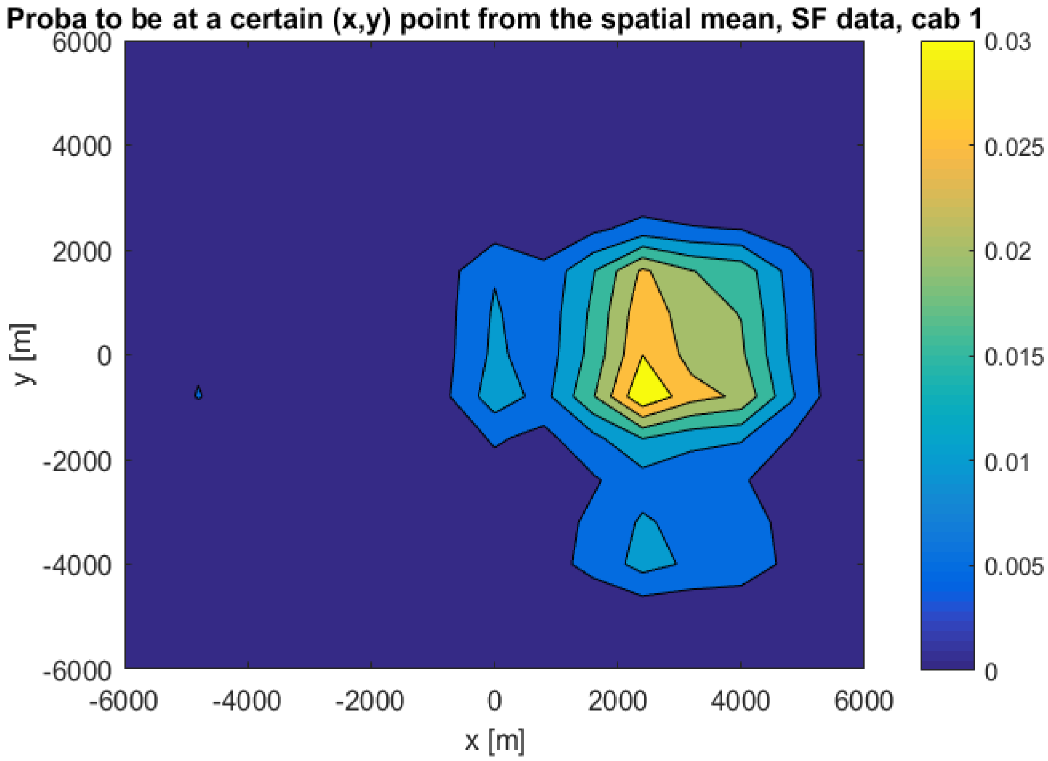

Figure 6-7 Probability to be at certain point from the spatial mean of one's trajectory, Rome dataset

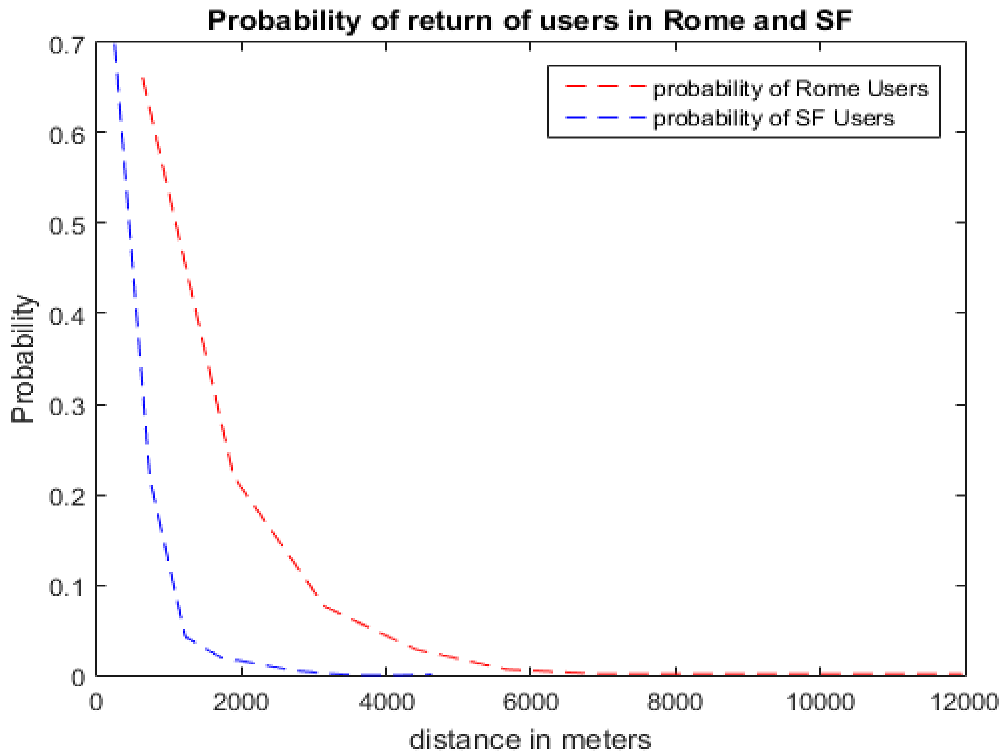

Figure 6-8 Probability of return of all taxicabs drivers in Rome and San Francisco

The probability of return of all users in case of San Francisco is found to be 0.7 (70 %) as highest probability within the range of nearly 200 meters of travelling. Whereas, it is 0.65 (65 %) as highest probability within the range of approximately 500 meters. The probability rate is being decreased as the distance of users from San Francisco dataset is increasing in exponential distribution ranging from nearly 200 meters to 4000 meters (4 Km). Meanwhile, in case of San Francisco dataset, the probability rate is being decreased exponentially ranging from approximately 500 meters to 12 Km.

In Table 6.1, the statistical summary of measured mobility parameters from both the datasets Rome and San Francisco are outlined.

Table 6.1 Summary of statistics achieved from datasets of Rome (316 users) and San Francisco (536 users) in a per day basis

Parameters	Rome	San Francisco
% outliers	2.8659 %	0.6895 %
Longest Route (Km)	8860.7601	15701.7904
Shortest Route (km)	0.7193	1.9835
Average Route (Km)	398.5448	429.3634
Minimum Pause Time (Seconds)	3.7440	4.32
Maximum Pause Time (Hours)	2.2980	0.4892
Average Pause Time (Hours)	0.0023	0.0091
Minimum Travel Time (Seconds)	3.1004	10.0008
Maximum Travel Time (Hours)	78.1830	28.5318
Average Travel Time (Hours)	0.1288	0.3574
Minimum Speed (Km/hour)	0	0
Maximum Speed (Km/hour)	139.4947	139.0866
Average Speed (Km/hour)	14.1909	22.1442
Minimum Step (Km)	0.0027	0.2463
Maximum Step (Km)	23.8885	61.0475
Average Step (Km)	0.2448	1.6222
Average probability of occupancy of all users in San Francisco		44.5 %

6.7 Mobility Model Validation

In this section, the suitable outdoor mobility models are analysed for our given trace data of both Rome and San Francisco datasets used in this thesis work. To accomplish it, we pursued the following steps:

- Firstly, the histograms of the measured mobility parameters were computed.
- Then, various theoretical distributions were compared against histograms of measured parameters.

- Finally, Kullback Leibler (KL) divergence criteria [62] was applied in order to detect the best fitting of measured parameters compared to existing theoretical distributions.

In Figure 6-9 below, the histogram of angle change (degrees) of first user from Rome dataset is described. In the horizontal axis of Figure 6-9 , the change of angles of first user are plotted in degrees and the vertical axis represents some probability distribution functions such as uniform distribution and Gaussian distribution.

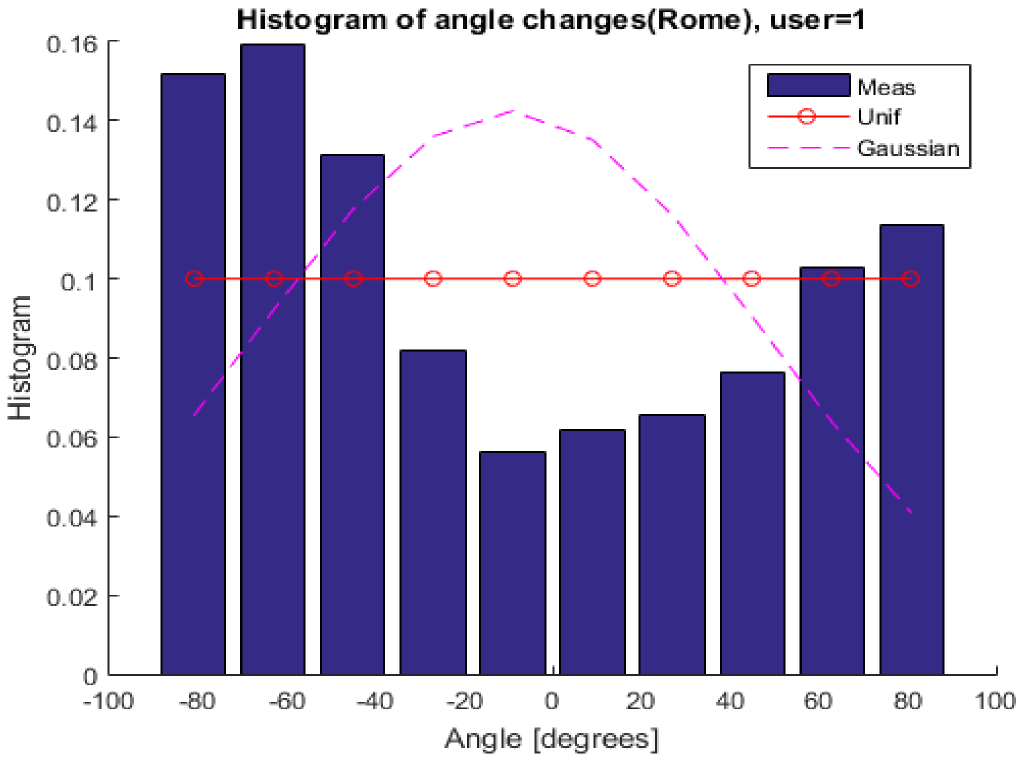

Figure 6-9 Histogram of angle change (degrees) of first user in Rome

Similarly, in Figure 6-10 below, we have applied the KL divergence criterion for change of angles of all users in Rome dataset against some possible theoretical distributions like uniform distribution and Gaussian distribution. In KL divergence technique, a lowest KL value obtained among all the possible theoretical distributions and measured parameters, signifies the best fitting of mobility models. From Figure 6-9 and Figure 6-10, it is observed that the lowest KL values for angle data seemed to fit for uniform distribution. That means, uniform distribution of angle data here points to be fitted with random walk model. Moreover, there are few outliners in both the Figure 6-9 and Figure 6-10 which is due to some GPS measurement errors in the datasets.

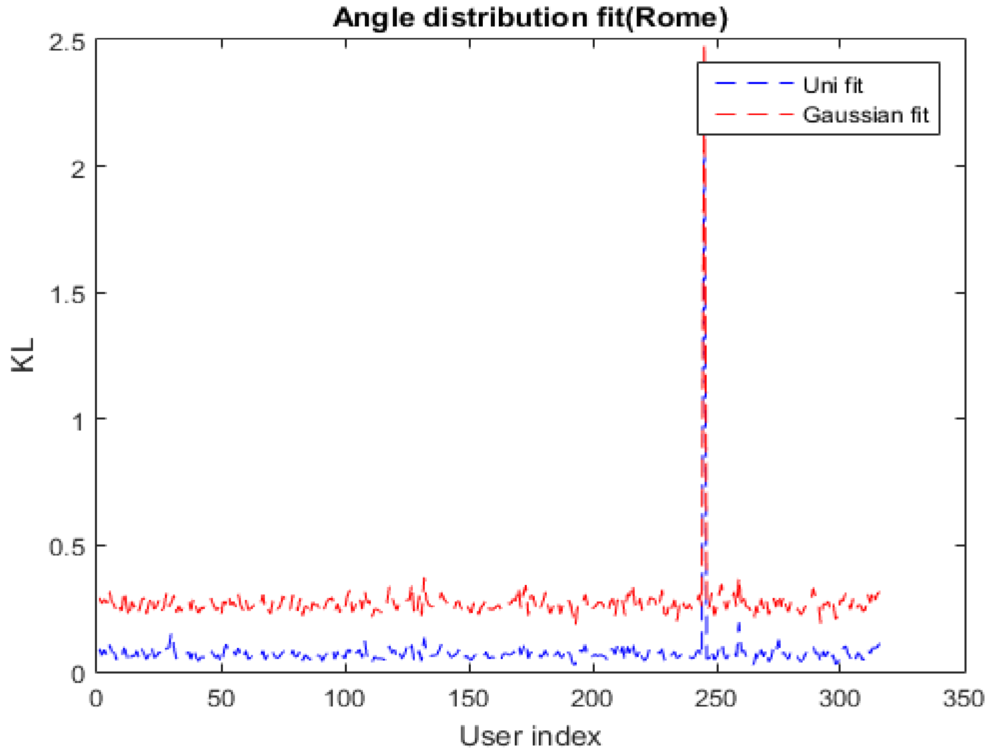

Figure 6-10 KL divergence analysis of angle change for Rome

Similarly, Figure 6-11 and Figure 6-12 show the histogram of speed of first user and KL divergence plot for speed distribution of all users in Rome dataset compared against three different distributions such as uniform, Gaussian and exponential distributions. As a result, both the exponential and Gaussian distribution are appeared to fit with the speed data of Rome dataset. On the other hand, from Figure 6-1 mentioned above, we can observe that the average speeds of all users are displaying the random (Gaussian) distribution. That means, users' speed is changing in random manner in different time intervals along with some data measurement errors spotted in their mobility trace. As a result, it is meaningful to validate average values among all the users' instantaneous speeds that yields more accurate data. Hence, we can conclude that the speed distribution of all users in Rome better fit towards the random (Gaussian) distribution.

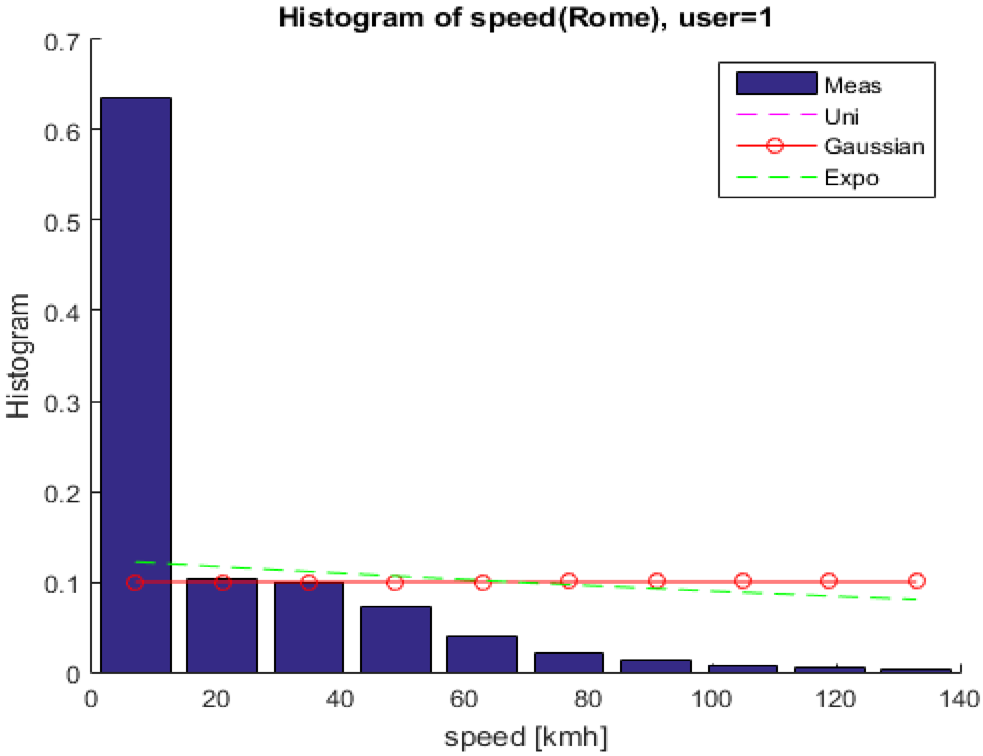

Figure 6-11 Histogram of speed (Km/h) of first user in Rome

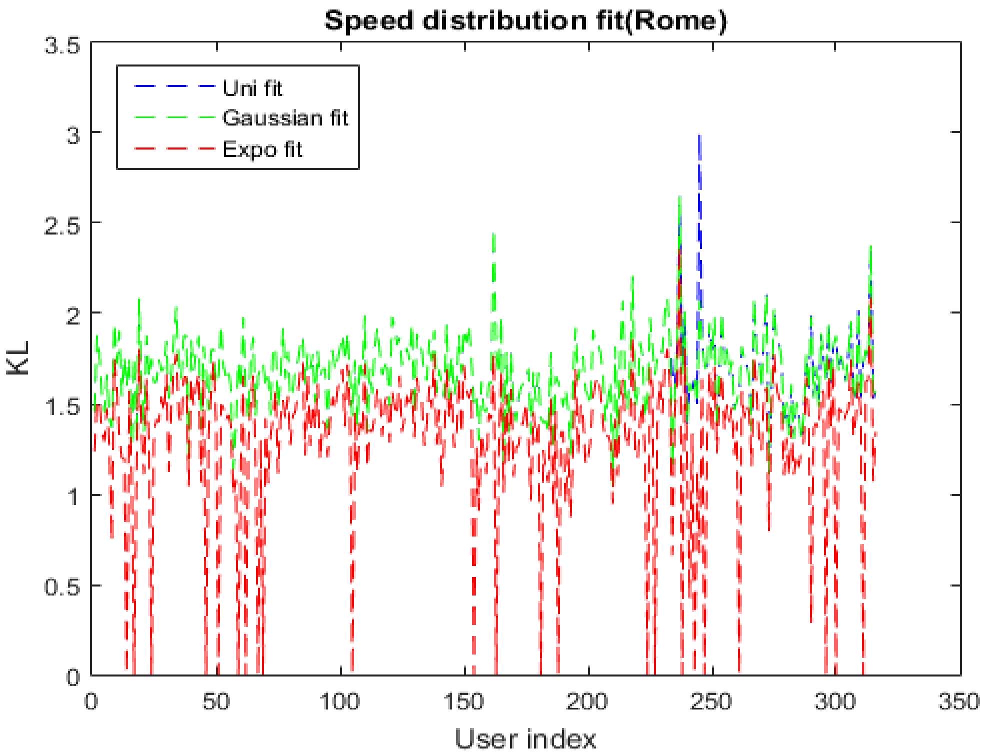

Figure 6-12 KL divergence analysis of speed for Rome

Similarly, in Figure 6-13 and Figure 6-14 , histogram of step of first user and KL divergence plot of steps of all users from Rome dataset are compared against uniform distribution, Gaussian distribution, and exponential distribution. From both Figure 6-13 and Figure 6-14 , it is observed that steps of users in Rome dataset follow the exponential distribution.

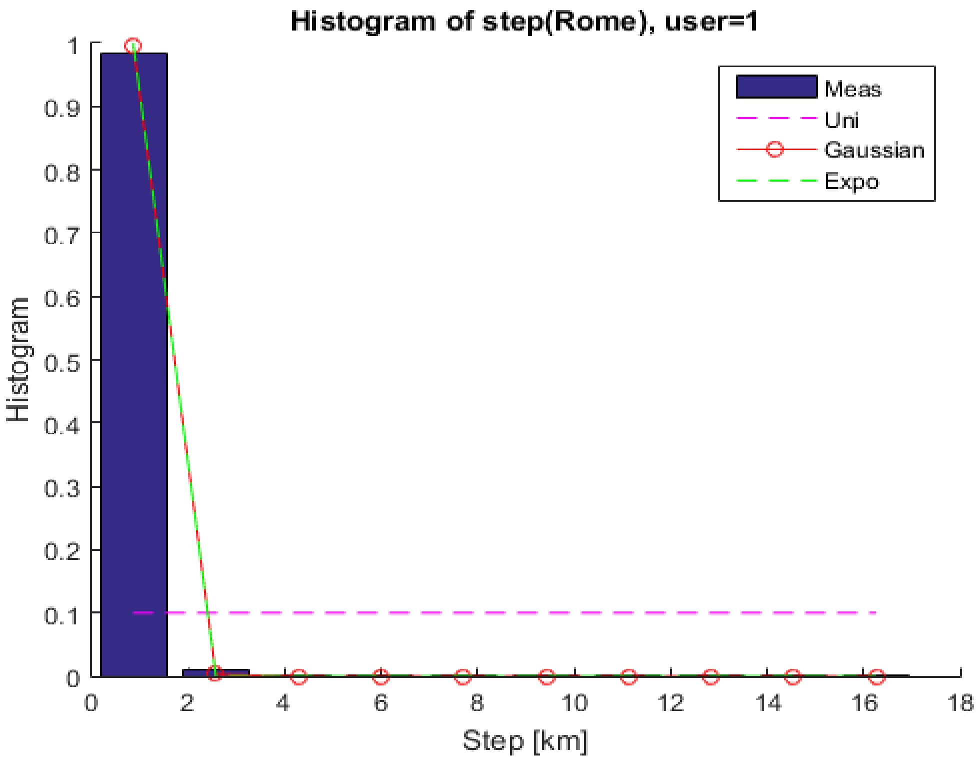

Figure 6-13 Histogram of step (Km) of first user of Rome

Furthermore, Figure 6-15 , Figure 6-17 and Figure 6-19 represent the histograms of angle data, speed and step of first user from San Francisco dataset respectively. Meanwhile, Figure 6-16 , Figure 6-18 and Figure 6-20 are the KL divergence plots of angle data, speed and step for all users in San Francisco dataset mentioned below. We have accomplished the histogram and KL divergence plot of angle, speed and step of users from San Francisco dataset using uniform, Gaussian and exponential distributions in a similar way to Rome dataset explained in earlier figures starting from Figure 6-9 to Figure 6-14. Consequently, we observed that angle data of users matches with uniform distribution, speed data matches with Gaussian distribution and finally step data are fitting with exponential distribution in San Francisco dataset. Hence, we observed that both the Rome and San Francisco dataset are fitting nearly into Random walk mobility model on the basis of angle change, speed and step parameters compared by histogram and KL divergence techniques.

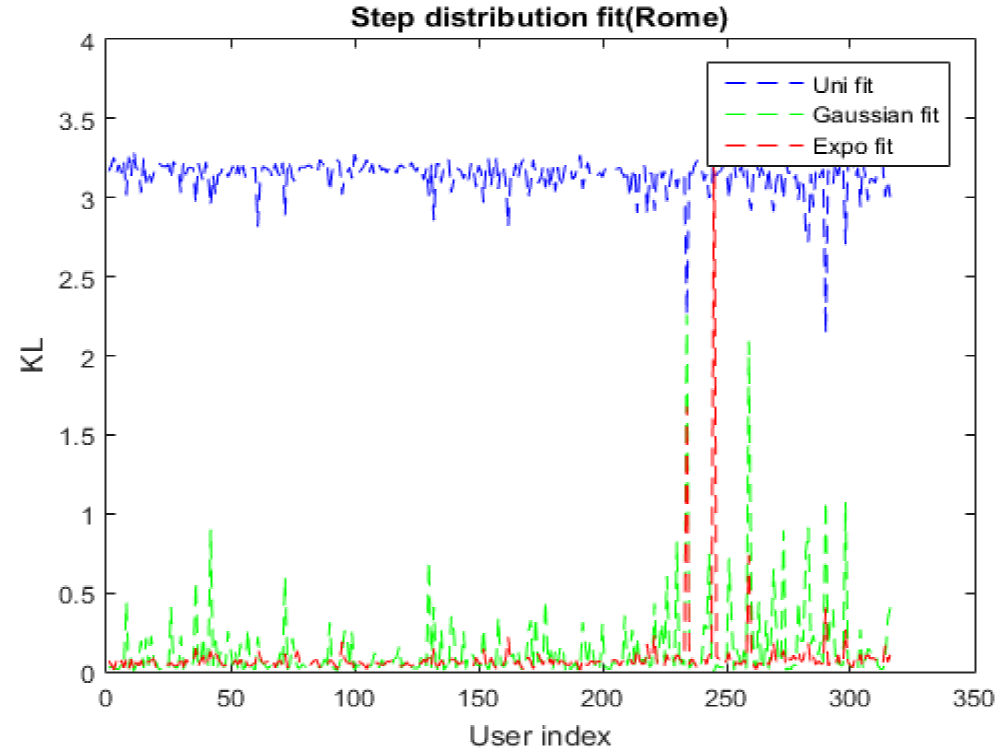

Figure 6-14 KL divergence analysis of step for Rome

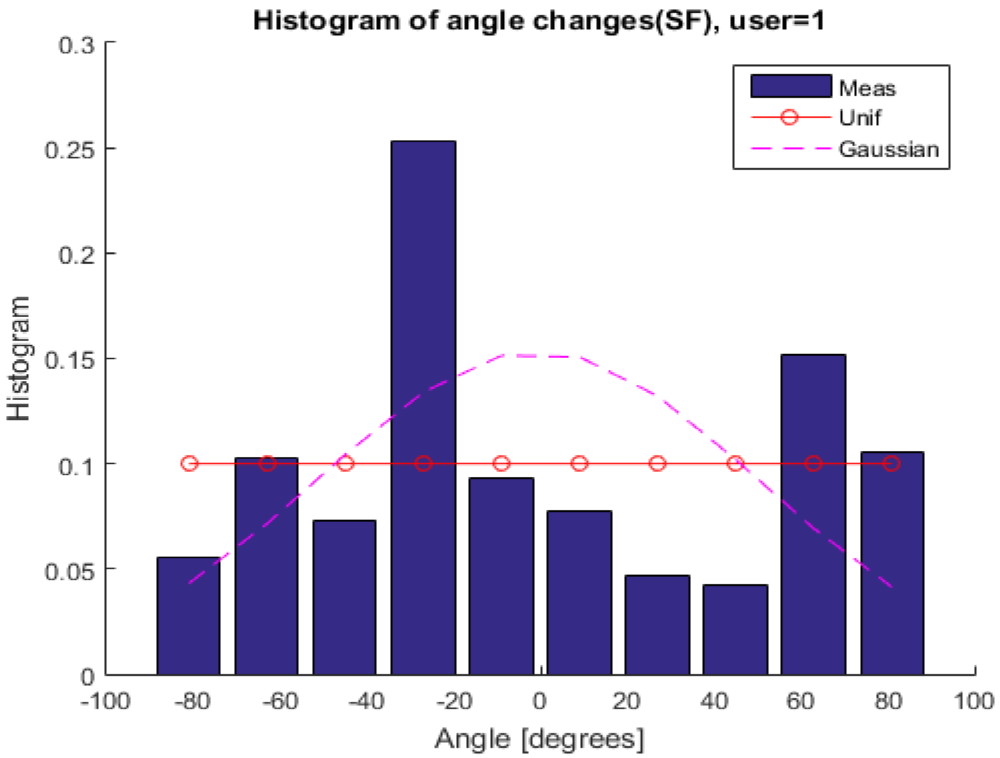

Figure 6-15 Histogram of angle change(degrees) of first user of San Francisco

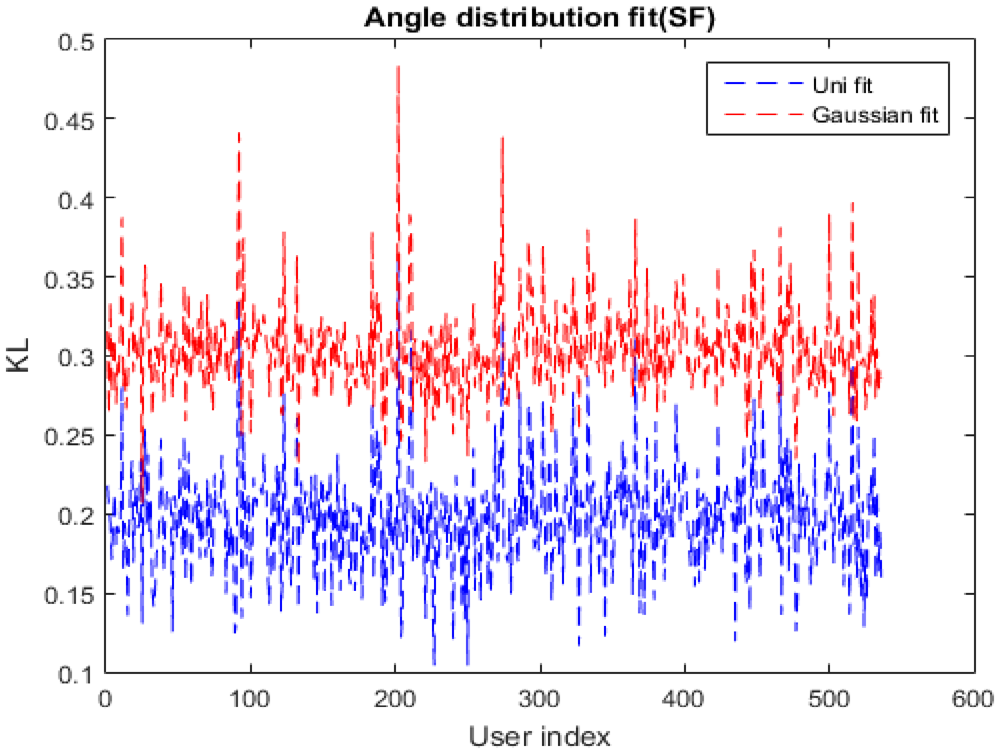

Figure 6-16 KL divergence analysis of angle change for San Francisco

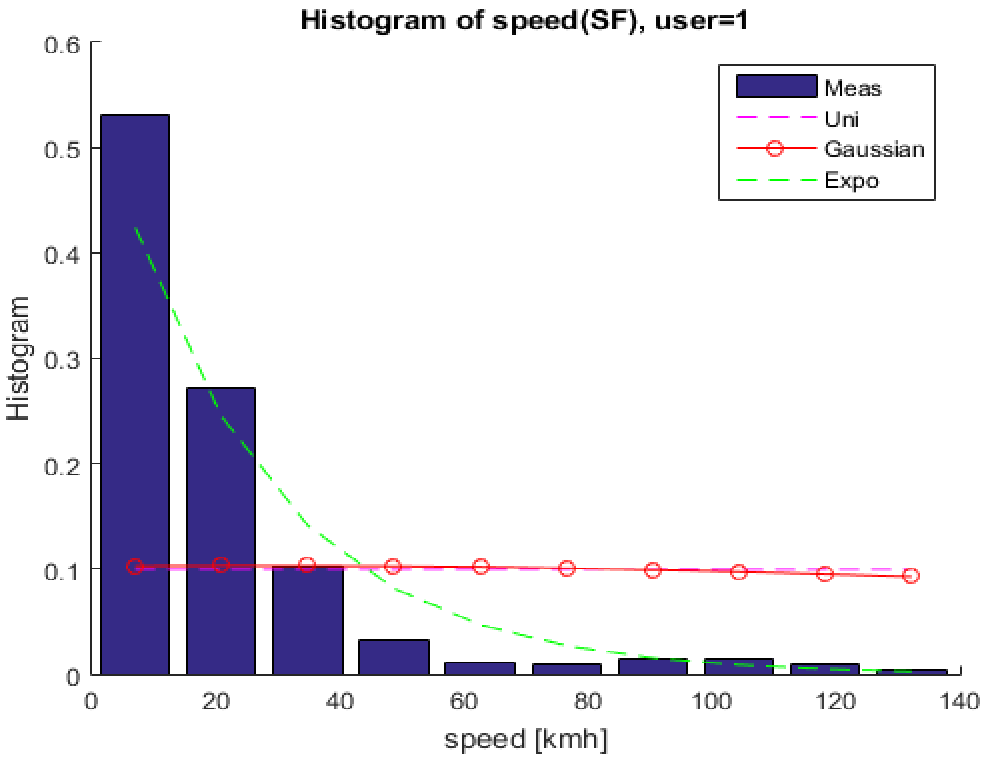

Figure 6-17 Histogram of speed (Km/h) of first user of San Francisco

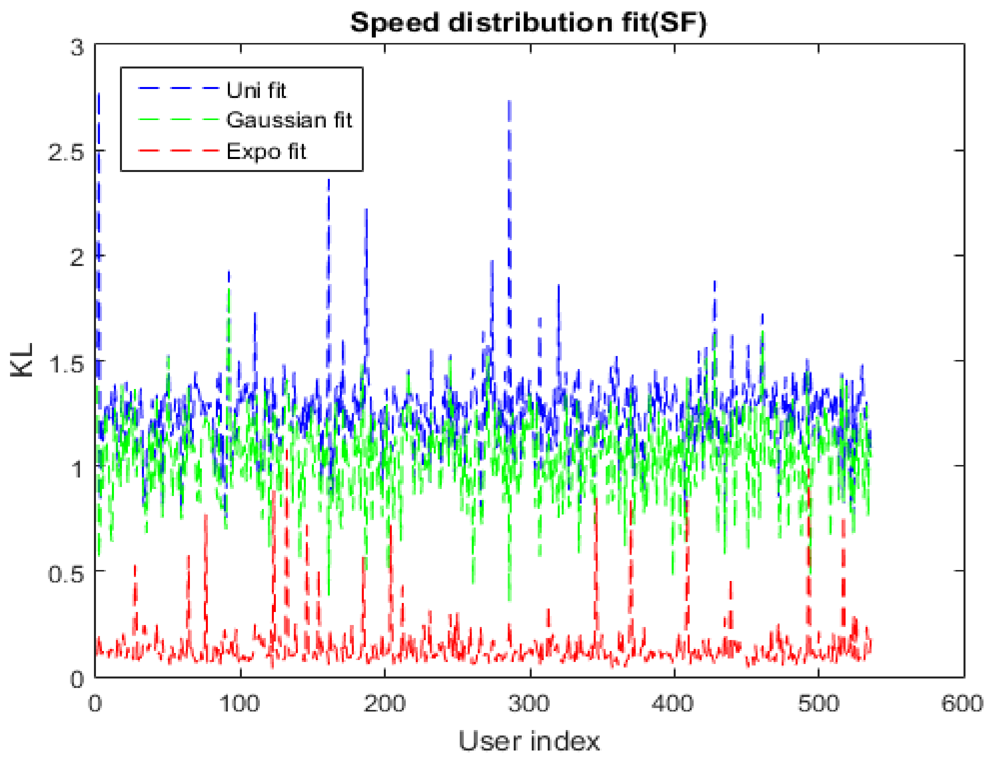

Figure 6-18 KL divergence analysis of speed for San Francisco

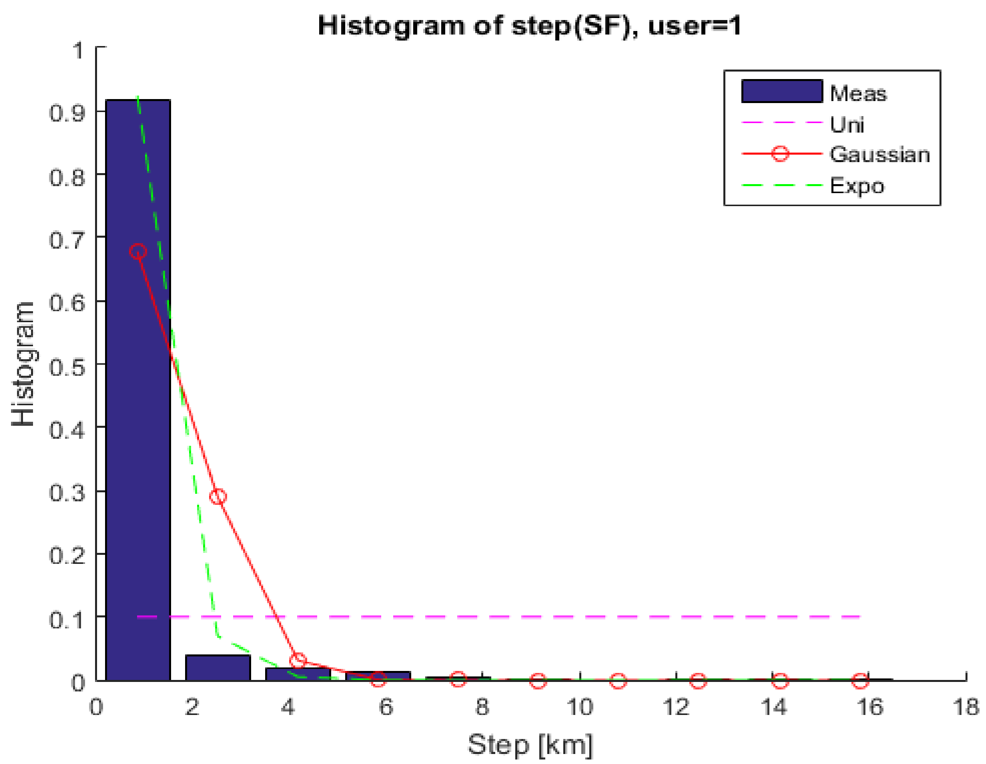

Figure 6-19 Histogram of step (Km) of first user of San Francisco

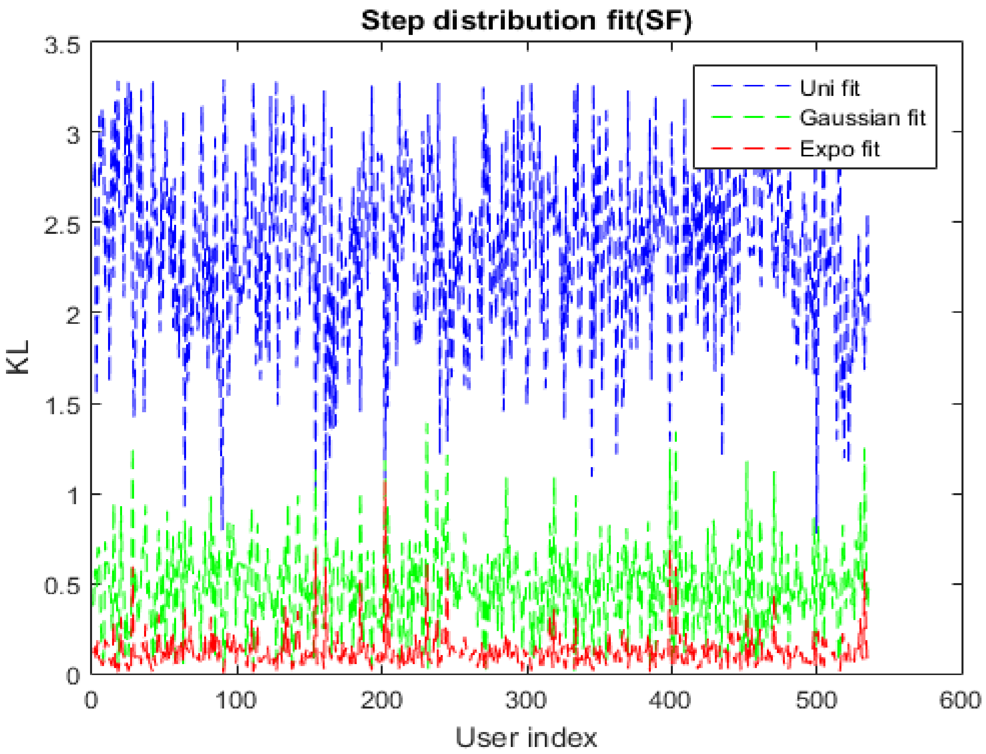

Figure 6-20 KL divergence analysis of step for San Francisco

In Table 6.2, the main findings of this thesis work obtained by few analyses between measured data and some theoretical distributions have been summarised.

Table 6.2 Summary of findings

Best distribution fit among the considered ones	Rome	San Francisco
Speed	Gaussian distribution	Gaussian distribution
Angle	Uniform distribution	Uniform distribution
Step	Exponential distribution	Exponential distribution

7 CONCLUSION AND FUTURE WORK

This chapter outlines the conclusions of this thesis work along with the future directions and the open issues.

7.1 Conclusion

A study for investigating various outdoor mobility patterns of taxicabs in urban areas was accomplished in this thesis. Firstly, a concept of smart city and trace data were discussed. Further, the different methods of collecting trace data of a city from various sources such as mobile devices, vehicles, smart cards and floating sensors etc. were discussed. Then, the existing challenges and applications of a smart city were described. In chapter 3, some important vehicular mobility models were illustrated along with the distinct comparisons between MANET and VANET. Besides this, the vehicular mobility models were described further on the basis of numerous categories such as synthetic models, survey-based models, trace-based models etc. Additionally, various mobility parameters were described into three major dimensions such as spatial, temporal and social dimension. This thesis work focused on first two dimensions such as speed, pause time, travel time, step, route, probability of return of users in the same point and so on.

In chapter 4, firstly the concept of vehicular communication was briefly introduced. Then, the various significant applications of VANET were described such as intelligent transportation applications, comfort applications, collision avoidance etc. Additionally, it was also described that how the vehicles communicate to each other through various mediums in VANET unfolding various types of routing protocols. Also, the various research issues and contributions related with this thesis work were briefly described at the end of chapter 4. Furthermore, in chapter 5, the different types of data parameters from Rome and San Francisco datasets, which were analysed entirely during this thesis work, were briefly described. Besides this, it was also explained how the importing and processing of datasets were performed.

Finally, in chapter 6, the analysis of various mobility parameters obtained from both datasets Rome and San Francisco was accomplished. All the mobility parameters such as speed, angle, step, route etc. from both datasets were analysed and compared their distributions using a single histogram in MATLAB which were portrayed in figures 6-1 to 6-5 and 6-8. Besides this, the summary of statistics of mobility parameters computed from both datasets is outlined in Table 6.1. Subsequently, all the measured mobility parameters from both datasets were analysed and fitted with suitable outdoor mobility models using Kullback Leibler (KL) divergence criteria. Few major theoretical distributions were included in the analysis, namely the uniform distribution, Gaussian distribution, and exponential distribution as illustrated in figures 6-9 to 6-20. A thorough work for the continuation would be to expand such studies to much more distributions that might be better suited to model vehicular mobility. As a result, both the Rome and San Francisco datasets approximately fitted with the random walk model on the basis of mainly three parameters angle change, speed and step of users.

7.2 Future directions and open issues

This thesis work is just a small attempt towards the diversified fields of user mobility modelling in urban scenarios. There is lot of space which can be added to this thesis work and can be continued in further extents in the future. For example, the hotspot detection and spatial mean centres of taxicab drivers were not completely analysed due to time limitation in this thesis work. Besides this, the data of probability of occupancy of passengers in this thesis could be used further to detect the user's pick up and drop off locations in a city. The measured mobility parameters could be tested with more outdoor mobility models such as 3D random walk model, Levy walk and so on.

REFERENCES

[1] X. Li, G. Pan, Z. Wu, G. Qi, S. Li, D. Zhang, *et al.*, "Prediction of urban human mobility using large-scale taxi traces and its applications," *Frontiers of Computer Science,* vol. 6, pp. 111-121, 2012.

[2] T. Kindberg, M. Chalmers, and E. Paulos, "Guest editors' introduction: Urban computing," *IEEE Pervasive Computing,* vol. 6, pp. 18-20, 2007.

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

[4] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, "An energy-efficient mobile recommender system," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 899-908.

[5] A. Gühnemann, R.-P. Schäfer, K.-U. Thiessenhusen, and P. Wagner, "Monitoring traffic and emissions by floating car data," 2004.

[6] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, *et al.*, "T-drive: driving directions based on taxi trajectories," in *Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems*, 2010, pp. 99-108.

[7] B. D. Ziebart, A. L. Maas, A. K. Dey, and J. A. Bagnell, "Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior," in *Proceedings of the 10th international conference on Ubiquitous computing*, 2008, pp. 322-331.

[8] L. Liu, C. Andris, and C. Ratti, "Uncovering cabdrivers' behavior patterns from their digital traces," *Computers, Environment and Urban Systems,* vol. 34, pp. 541-548, 2010.

[9] D. Zhang, N. Li, Z.-H. Zhou, C. Chen, L. Sun, and S. Li, "iBAT: detecting anomalous taxi trajectories from GPS traces," in *Proceedings of the 13th international conference on Ubiquitous computing*, 2011, pp. 99-108.

[10] S. Phithakkitnukoon, M. Veloso, C. Bento, A. Biderman, and C. Ratti, "Taxi-aware map: Identifying and predicting vacant taxis in the city," in *International Joint Conference on Ambient Intelligence*, 2010, pp. 86-95.

[11] H.-W. Chang, Y.-C. Tai, and J. Y.-J. Hsu, "Context-aware taxi demand hotspots prediction," *International Journal of Business Intelligence and Data Mining,* vol. 5, pp. 3-18, 2009.

[12] L. Bracciale, M. Bonola, P. Loreti, G. Bianchi, R. Amici, and A. Rabuffi, "CRAWDAD data set roma/taxi (v. 2014-07-17)," ed: July, 2014.

[13] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser, "CRAWDAD data set epfl/mobility (v. 2009-02-24)," ed, 2009.

[14] V. Buscher, L. Doody, M. Webb, and C. Aoun, "Urban mobility in the smart city age," ed: London: Schneider Electric, ARUP, The Climate Group, 2014.

[15] G. Pan, G. Qi, W. Zhang, S. Li, Z. Wu, and L. T. Yang, "Trace analysis and mining for smart cities: issues, methods, and applications," *IEEE Communications Magazine,* vol. 51, pp. 120-126, 2013.

[16] C. Harrison, B. Eckman, R. Hamilton, P. Hartswick, J. Kalagnanam, J. Paraszczak, *et al.*, "Foundations for smarter cities," *IBM Journal of Research and Development,* vol. 54, pp. 1-16, 2010.

[17] A. Boustani, L. Girod, D. Offenhuber, R. Britter, M. I. Wolf, D. Lee, *et al.*, "Investigation of the waste-removal chain through pervasive computing," *IBM Journal of Research and Development*, vol. 55, pp. 11:1-11:11, 2011.

[18] N. A. Alrajeh, M. Bashir, and B. Shams, "Localization techniques in wireless sensor networks," *International Journal of Distributed Sensor Networks*, 2013.

[19] I. A. Halepoto, A. A. Sahito, M. A. Uqaili, B. S. Chowdhry, and T. Riaz, "Multi-criteria assessment of smart city transformation based on SWOT analysis," in *Information Technology: Towards New Smart World (NSITNSW), 2015 5th National Symposium on*, 2015, pp. 1-6.

[20] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1100-1108.

[21] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti, "Activity-aware map: Identifying human daily activity pattern using mobile phone data," in *International Workshop on Human Behavior Understanding*, 2010, pp. 14-25.

[22] M. M. Rathore, A. Ahmad, A. Paul, and S. Rho, "Urban planning and building smart cities based on the internet of things using big data analytics," *Computer Networks*, vol. 101, pp. 63-80, 2016.

[23] S. Yang, X. Yang, C. Zhang, and E. Spyrou, "Using social network theory for modeling human mobility," *IEEE Network*, vol. 24, pp. 6-13, 2010.

[24] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez, "Show me how you move and I will tell you who you are," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*, 2010, pp. 34-41.

[25] S. Debnath and A. Majumder, "A Survey on Mobility Models for Vehicular Ad hoc Networks," *IJRET: International Journal of Research in Engineering and Technology*.

[26] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless communications and mobile computing*, vol. 2, pp. 483-502, 2002.

[27] I. Stepanov, P. J. Marron, and K. Rothermel, "Mobility modeling of outdoor scenarios for manets," in *Proceedings of the 38th annual Symposium on Simulation*, 2005, pp. 312-322.

[28] E. Weisstein, "CRC Concise Encyclopedia of Mathematics, CRC Press," *Boca Raton, FL, USA*, 1998.

[29] J. Broch, D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva, "A performance comparison of multi-hop wireless ad hoc network routing protocols," in *Proceedings of the 4th annual ACM/IEEE international conference on Mobile computing and networking*, 1998, pp. 85-97.

[30] B. Pazand and C. McDonald, "A critique of mobility models for wireless network simulation," in *Computer and Information Science, 2007. ICIS 2007. 6th IEEE/ACIS International Conference on*, 2007, pp. 141-146.

[31] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779-782, 2008.

[32] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the levy-walk nature of human mobility," *IEEE/ACM transactions on networking (TON)*, vol. 19, pp. 630-643, 2011.

[33] R. Thomas, H. Gilbert, and G. Mazziotto, "Influence of the movement of the mobile station on the performance of a radio cellular network," in *Proc. 3rd Nordic Seminar*, 1988, p. 9.4.

[34] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE transactions on Vehicular Technology,* vol. 35, pp. 77-92, 1986.

[35] R. A. Guérin, "Channel occupancy time distribution in a cellular radio system," *IEEE transactions on Vehicular Technology,* vol. 36, pp. 89-99, 1987.

[36] H. Xie and D. J. Goodman, "Mobility models and biased sampling problem," in *Universal Personal Communications, 1993. Personal Communications: Gateway to the 21st Century. Conference Record., 2nd International Conference on,* 1993, pp. 803-807.

[37] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "SLAW: self-similar least-action human walk," *IEEE/ACM Transactions on Networking (TON),* vol. 20, pp. 515-529, 2012.

[38] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature,* vol. 439, pp. 462-465, 2006.

[39] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of Human Mobility on the Design of Opportunistic Forwarding Algorithms," in *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, 2006, pp. 1-13.

[40] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the Levy-Walk Nature of Human Mobility," *IEEE/ACM Transactions on Networking,* vol. 19, pp. 630-643, 2011.

[41] C. Tuduce and T. Gross, "A mobility model based on WLAN traces and its validation," in *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, 2005, pp. 664-674 vol. 1.

[42] J. Yoon, B. D. Noble, M. Liu, and M. Kim, "Building realistic mobility models from coarse-grained traces," in *Proceedings of the 4th international conference on Mobile systems, applications and services*, 2006, pp. 177-190.

[43] M. Kim, D. Kotz, and S. Kim, "Extracting a Mobility Model from Real User Traces," in *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, 2006, pp. 1-13.

[44] J.-K. Lee and J. C. Hou, "Modeling steady-state and transient behaviors of user mobility: formulation, analysis, and application," in *Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing*, 2006, pp. 85-96.

[45] V. Srinivasan, M. Motani, and W. T. Ooi, "Analysis and implications of student contact patterns derived from campus schedules," in *Proceedings of the 12th annual international conference on Mobile computing and networking*, 2006, pp. 86-97.

[46] J. Harri, F. Filali, and C. Bonnet, "Mobility models for vehicular ad hoc networks: a survey and taxonomy," *IEEE Communications Surveys & Tutorials,* vol. 11, pp. 19-41, 2009.

[47] A. Mahajan, N. Potnis, K. Gopalan, and A. Wang, "Urban mobility models for vanets," in *2nd IEEE International Workshop on Next Generation Wireless Networks*, 2006.

[48] D. Choffnes and F. E. Bustamante, "STRAW-An Integrated Mobility and Traffic Model for VANET," in *Proc. of the 10th International Command and Control Research and Technology Symposium (CCRTS)*, 2005.

[49] D. Karamshuk, C. Boldrini, M. Conti, and A. Passarella, "Human mobility models for opportunistic networks," *IEEE Communications Magazine,* vol. 49, pp. 157-165, 2011.

[50] S. Singh and S. Agrawal, "VANET routing protocols: Issues and challenges," in *2014 Recent Advances in Engineering and Computational Sciences (RAECS)*, 2014, pp. 1-5.

[51] Z. J. Haas, "A new routing protocol for the reconfigurable wireless networks," in *Proceedings of ICUPC 97 - 6th International Conference on Universal Personal Communications*, 1997, pp. 562-566 vol.2.

[52] H. Tang, "Urban Transportation Analysis Using Taxi Trajectory and Weather Data," Kent State University, 2016.

[53] F. D. Cunha, D. A. Alvarenga, A. C. Viana, R. A. Mini, and A. A. Loureiro, "Understanding interactions in vehicular networks through taxi mobility," in *Proceedings of the 12th ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks*, 2015, pp. 17-24.

[54] M. Veloso, S. Phithakkitnukoon, and C. Bento, "Sensing urban mobility with taxi flow," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, 2011, pp. 41-44.

[55] R.-P. Schäfer, K.-U. Thiessenhusen, E. Brockfeld, and P. Wagner, "A traffic information system by means of real-time floating-car data," 2002.

[56] H. Yang, M. Ye, W. H. Tang, and S. C. Wong, "Regulating taxi services in the presence of congestion externality," *Transportation Research Part A: Policy and Practice*, vol. 39, pp. 17-40, 2005.

[57] K. Wong and M. G. Bell, "The optimal dispatching of taxis under congestion: a rolling horizon approach," *Journal of advanced transportation*, vol. 40, pp. 203-220, 2006.

[58] "CRAWDAD Data Licence," [Online]. Available: http://crawdad.org/data-license-agreement.html.

[59] "Exploratorium," [Online]. Available: http://www.exploratorium.edu/about-us

[60] "Cabspotting Project," [Online]. Available: http://www.exploratorium.edu/id/cab.html

[61] L. Ding, H. Fan, and L. Meng, "Understanding taxi driving behaviors from movement data," in *AGILE 2015*, ed: Springer, 2015, pp. 219-234.

[62] "Kullback-Leibler divergence ," [Online]. Available: http://en.wikipedia.org/wiki/Kullback–Leibler_divergence