Aalto University
School of Engineering
Degree Programme in Geomatics

Muneeba Raja

Exploring Human Mobility Patterns Based on Geotagged Flickr Photos

Master's Thesis
Espoo, July 8, 2015

Supervisor: Professor Kirsi Virrantaus
Advisors: Professor Jörg Ott, COMNET Aalto University
Esa Hyytiä, D.Sc. (Tech.), Docent, COMNET Aalto University

Aalto University
School of Engineering
Degree Programme in Geomatics

ABSTRACT OF
MASTER'S THESIS

Author:	Muneeba Raja		
Title:	Exploring Human Mobility Patterns Based on Geotagged Flickr Photos		
Date:	July 8, 2015	**Pages:**	v + 62
Major:	Geoinformation Technology	**Code:**	IA3002
Supervisor:	Professor Kirsi Virrantaus		
Advisors:	Professor Jörg Ott, COMNET Aalto University Esa Hyytiä, D.Sc. (Tech.), Docent, COMNET Aalto University		

Predicting human mobility behaviour has long been a topic of scientific interest. Such studies generally rely on tracking human movements through a range of data collection methodologies such as using GPS trackers, cellular network data etc. Some of this data may be confidential or hard to acquire. This thesis explores if existing publicly available data on online photo sharing platforms can be used to determine human mobility patterns with reasonable accuracy. We choose the Flickr website as the data collection medium as it has an extensive user base actively sharing photos many of which, have geo tags embedded in them which are preserved by Flickr. Our analysis reveals that while the data from Flickr is sparse and discontinuous making it unsuitable for reliable mobility prediction, typical human mobility trends based on time of day, day of week and month of the year can still be extracted. Such interesting patterns could be potentially used in traffic engineering domains or for user profiling purposes.

More specifically, we describe how to obtain a subset of frequent active users and their information from Flickr, and the sliding window mechanism to filter the active periods of the users. Later we explain the various statistical methods applied on the filtered subset of data to identify the categories in which users could be classified, mainly short distance travellers and long distance travellers. The short distance travellers are considered for mobility trends prediction.

Keywords:	Flickr, Movement Patterns, Micro-mobility, Spatial-temporal Analysis, Geotagged images, Pattern Mining.
Language:	English

Acknowledgements

First and foremost I would like to thank Allah for his countless blessings upon me and prayers of my parents without which this thesis wouldn't have been possible.

I would like to thank my supervisor, Prof. Kirsi Virrantaus for her guidance and continuous support. My sincere gratitude to my advisor, Prof. Jörg Ott for believing in me and guiding me throughout, with his immense knowledge and expertise in the field. He is not only the most supportive mentor but also the kindest human being I have ever met. I truly wish to get an opportunity to work with him in future.

Besides, I would really wish to thank my instructor Esa Hyytiä for encouraging me and always being ready to guide me with his tremendous experience in the field. I owe my eternal gratitude to Saba Ahmed, a colleague and a great friend for her support and valuable brainstorming sessions.

Last but not the least,I would dedicate this thesis to my husband, Syed Safi Ali Shah, without whose technical and moral support I could not have written a single word of this thesis. My deepest gratitude for his continuous encouragement and unconditional love. I am truly blessed to have him in my life. And my son Arham for his patience and compromises on his time to make it happen.

Espoo, July 8, 2015

Muneeba Raja

Abbreviations and Acronyms

OAuth	OAuth is the authentication protocol that is an industry standard for applications to log in to Flickr and other accounts. This guarantees the secure transfer of user's information. We can grant permissions like read and write to allow the application to communicate with Flickr on our behalf.
API	Application Programming Interface
EXIF	Exchangeable Image File Format
GPS	Global Positioning System
ROI	Regions of Interest
POI	Points of Interest
HTTP	Hypertext Transfer Protocol
JSON	Javascript Object Notation
URL	Uniform Resource Locator
XML	Extensible Markup Language
ECDF	Empirical Cumulative Distribution Function
km	Kilometers
DSLR	Digital Single Lens Reflex

Contents

Chapter 1

Introduction

Analyzing the human mobility patterns has been a well-studied area of research for years (e.g. [1–3]). Researchers often rely on tracking human movements and collecting data by using GPS devices attached to mobile individuals. Such data is suitable for studying fine grained micro-mobility aspects owing to its regular and continuous nature i.e. there are no long silent periods in the data. However, such data is rarely available to public and collecting it can be costly and laborious work [2, 4].

But then, with the rising popularity of online social platforms, more and more users are sharing information online. Photo sharing websites such as Flickr[1], Instagram[2] and Panoramio[3] are a huge source of photos, which can be utilized for the discovery of underlying patterns in the users' behaviour. In this thesis, we aim to find out the extent and accuracy, to which the data from such photo sharing platforms is able to help predicting human mobility patterns. Note that there are various other sources of location data as well, being used to understand mobility and some of them are highlighted in Chapter 2.

1.1 Photo Sharing Websites

Flickr and Panoramio are quite similar in their functionality and data providing services (APIs). They practice geotagging functionality for annotating the geographical information to the photos. They do so by extracting the information from the EXIF (Exchangeable Image File Format) tags, if available, or, also allow the users to mark their geographical coordinates by choos-

[1]https://www.flickr.com/
[2]https://instagram.com/
[3]http://www.panoramio.com/

ing their location on the map. Flickr's auto geotagging facility automatically extracts the geo location and timestamps from the photos and attach the corresponding geo tags to the photos. The timestamps information enables us to retrieve the upload time as well as the time of the photos taken.

Instagram is a tremendously growing photo sharing website [5] which also allows the users to associate their location information with the photos while uploading. However, it does so by providing the users with the list of nearby locations, only if the location detection functionality is enabled on the device while taking photos. Otherwise, the suggestions are based on the current locations at the time of upload. For example if the user captured photos somewhere in U.S with the location detection functionality disabled on the device at that time, and then tries to upload photos from Europe, the suggestion list for locations will be from Europe although the photo was taken in U.S. Instagram does not keep the EXIF tags, hence no GPS and timestamp information through EXIF could be extracted as it strips off these tags from the images while uploading. The location information relies entirely on user supplied locations [6, 7].

We choose Flickr as a data collection medium for our study. Flickr is a huge resource of geotagged images, containing over 5 billion photos [8]. Due to its free membership and fast photo uploading, it has been used widely around the globe for years. Millions of users upload their photos every day, and for each photo, the spatial-temporal metadata also gets uploaded. About 3-4% [9] of the total photos uploaded on Flickr are geotagged due to the modern camera devices storing location information in EXIF tags which include the longitude and latitude of the photo, the time it was taken and the time it was uploaded. In addition, it contains detailed information about the camera model, focal length and many other factors. Flickr provides easy access to the public EXIF tags of the photos, hence a lot of interesting information from the tags could be inferred such as what locations the users have been visiting, along with the time differences between the pictures, the frequency of movements and so on.

1.2 Problem Statement

The objective of this study is to utilize the publicly available data from the photo sharing website Flickr for understanding and analysing the mobility behaviours of the users from different parts of the world. The ultimate goal is to reveal the mobility patterns among the users from their travelling routines and covered distances. And furthermore, we want to assess the data and know that to what extent can it be utilized for generating mobility and

micro-mobility patterns.

1.3 Main Results

The purpose of this study is to use the Flickr geotagged photos' metadata, to find out how the user travels around at short distances, say 300 km, and if there are certain common movement patterns among the users. To achieve our goals, we sample Flickr users and filter the active users i.e. users who travel and capture photos regularly over a period of time. At the same time, the count of these users' photos should be large enough for drawing accurate mobility patterns. We find the following interesting facts about the usability of Flickr data and its users as a source for mobility data:

1. The geotagged photos on Flickr can be used as a standalone source for predicting mobility patterns at a coarse level. However, the details available are insufficient to infer fine-grained micro mobility patterns; to obtain those, the density of photos per user and thus the number uploaded (not necessarily taken) per user with EXIF data shared publicly would need to be much higher.

2. The active users can be categorized into four main classes based on the distance they travel and their travel frequency. Two broad categories of the users are the long distance and short distance travellers. And each of these categories can further be classified into frequent and infrequent travellers. We cluster the users into these categories by defining the threshold values for the amount of distance travelled and the number of journeys in a period of time.

3. Most of the users on Flickr are active and mobile only during the specific months of the year. This means that the maximum amount of travelling done and pictures taken by the users are only during a couple of months. And these months are common for most of the users in our data subset.

4. We spot the similarities and differences in travelling routines of the users during the weekdays and over the weekends. And we find that more than half of the travelling done during the week is only on the specific days.Weekends are found to be more suitable days for people to make excursion trips which depict the natural behaviour. Also the similarities in the activity behaviour of the people during the weekdays tell about their work routine by travelling the same amount of distance every day.

5. Most of the users have the same active times of the day and the exact busy hours. The pattern for the busy hours however are not static throughout the week. The busy hours for the weekdays are different from those during the weekends. The early morning hours are the busiest during the weekdays while over the weekends the activity level is extremely low during those hours.

We perform various statistical analysis techniques over the data to extract the meaningful information. The results of these analyses depict the similarities between the users' mobility behaviours.

Given the patterns revealed about the user's mobility, Flickr proves to be a useful source for successfully predicting mobility trends at a medium level of granularity. However, it is not possible to calculate the accurate velocities and flight times necessary for microscopic mobility models [1] using this data. The revealed patterns can be of potential interests for traffic engineering purposes, ensuring reliable and safe traffic flow of traffic by identifying busy times, but also to feed macroscopic mobility models. Moreover, using Flickr data, the targeted profile advertising is also possible.

1.4 Structure of the Thesis

Chapter 2 is the literature review of the related work done in human mobility utilizing different kinds of data sources analysis. Chapter 3 focuses on the photo sharing websites and research done using Flickr and other social media websites. We provide a detailed introduction to the Flickr API in Chapter 4, covering its advantages and limitations for downloading Flickr photos and their associated metadata. Chapter 5 tells about the data we obtained from Flickr and our data collection strategy. Chapter 6 covers the preprocessing steps and filtering methods applied on the data to make it suitable for mobility analysis. Chapter 7 presents the statistical analysis and our findings, focusing on the classification of active Flickr users and the mobility trends explored in the data by applying various statistical techniques. We conclude the thesis in Chapter 8 along with suggestions for future work in this dimension.

Chapter 2

Human Mobility Models

This chapter gives a high level literature review of the previous experiments and studies regarding human mobility using different data sources. First we explain some empirical data sources that have been used in different ways to draw human mobility patterns. GPS being the most popular data source, we describe it in detail in the later half of the chapter.

2.1 Data Sources for Human Mobility Patterns

Understanding the human mobility patterns has a great importance in a variety of areas ranging from urban planning, traffic engineering, controlling epidemic disease spread or disaster management. Predicting accurate mobility models is challenging because usually humans' movements data is not readily available. The amount of travel between locations can be quantified with either *direct measurement* method, which includes, e.g. tracking of human movements using the location data from their phones or more traditionally traffic surveys. Other way is the *indirect measurement* method, in which models are used to estimate human movements given the observed (public) data as input [10]. Both direct and indirect methodologies use different kinds of empirical data are explained next and summarized in Table 2.1.

2.1.1 Mobile Phone Calls Data

The straightforward way for drawing human mobility patterns is to use location data from the mobile phones and track movements directly. The data is collected from mobile operators using e.g. their billing system. The location of the nearest tower(radio base station) to the phone at the time its used is

tracked. The data is high quality and in abundance which resolves the resolution issues but the drawback of using this type of data is that the privacy of the people may be violated (note that the privacy issue arises also with the other sources discussed next) [10].

Palchykov et al. [10] aims to infer human mobility using only aggregated mobile phone calls data. The reason for using this data is the ubiquitous nature of mobile phones. This leads to the influence of strength of social ties between locations on travel patterns. Their simple model using frequency of mobile phone calls between two locations and geographical distance between them can well predict human mobility. They use variants of models which take location specific call frequencies as input. The use of only aggregate data mitigates significantly potential privacy issues that are involved in tracking individual data [10]. However, assuming the high penetration of mobile phone use in the societies ignores the individuals not contributing to this kind of data. For example, mobile phone call data cannot be obtained for individuals without mobile phone activities.

2.1.2 GPS Data

GPS enabled devices such as mobile phones or hand held devices are capable of tracking the position of the individuals to a level of high accuracy within 4 meters. Movement information is tracked in the form of longitude, latitude and time stamp. The indoor mobility, however, is difficult to track with GPS devices due to blockage of signals within the building or undergrounds. Also the weak signal strength or interference across tall buildings prevent tracking movements for longer periods [11]. Nevertheless, it has been used tremendously in understanding human mobility patterns, thus we discuss its role in a separate section.

Data Source	Resolution(typical)	Scale(typical)	Availability
Mobile Call Data	Km	metropolitan area	not public
GPS Data	m	global	limited-open sources

Table 2.1: Comparison of data sources used for predicting human mobility patterns.

2.2 Role of GPS Data in Understanding Human Mobility Patterns

In this section we take a closer look at the role of GPS data in predicting human behaviours. The research in human mobility patterns has grown over time with the advent of GPS technology. Gonzalez et al. analyses about 100,000 GPS based human trajectories taken anonymously during a time period of six months in [3]. The discovered human mobility patterns show regularities in the spatial and temporal aspects which contradicts the random walk patterns [12]. They further find out the the humans travel back and forth to their favourite locations and the travelled distance and time are short for trajectories [3].

According to Gonzalez et al. in [3], human mobility is 93% predictable, but the individual human mobility prediction depends on the entropy of ones' patterns. The amount of travel per individual doesn't affect the predictability of the mobility patterns. For example, the mobility patterns do not deviate much between a user who travels less than 10km and one who travels 100km on regular basis. The above mentioned studies suggest the data mining algorithms that could be used for understanding human mobility but do not as such produce actual mobility patterns.

The GPS trajectory data is also used for revealing the most visited locations and their sequence of visit using the frequency of visits and interest of individuals for visiting certain locations [13]. The authors claim it to be useful for listing popular landmarks and typical routes to facilitate the tourists to organize their travel plans. However, the aggregated data might not be a perfect source for location recommendation in the user's home city on every day basis. Data mining methods for discovering association between the locations are identified by Zheng et al. in [14], which can be used for location recommendations derived from collaborative filtering.

Location based services rely on the location and movements on the individual basis but the studies focus on the human mobility patterns in general with only small details of individual mobility patter prediction. They do not take the temporal dimension into account for the mobility prediction. This raises the question that can we incorporate time in understanding the mobility behaviour of the individuals? Can we draw the mobility patterns from the weekly or daily routine of individuals or for the large population and what could be inferred from that.

2.2.1 Daily and Weekly based Patterns

The daily and weekly based human mobility patterns are explored by [15] using 17,621 GPS trajectories from 178 users. They find out the most of the travelling for each user is within a few destinations (see Figure 2.1). The popular destinations and the daily commute trajectories can act as a base point for other locations to cover the overall travelling of a user. They also find out that people tend to have a fixed schedule of activities while travelling from one location to another. It can be interpreted as daily work routine during the weekdays and favourite hobbies or activities during the weekends as shown in Figure 2.2. In this figure, we can observe the variation among the trips during each day of the week. Solid black line represents the weekday average of the trips and the grey line depicts the average weekend trips. There is little or no traffic from midnight to 7am. The time period from 7am to 8am seems to be the busiest hour during the weekdays, as expected, while the traffic is also dense from 5pm to 9pm. The weekends seem to be quite different as the traffic is highly reduced throughout the day with only slight increase before 5pm. Individuals spending weekdays at work and involved in leisure activities during the weekends are interpreted from this graph by Herder and Siehndel in [15].

2.3 Conclusions

This chapter highlights the popular data sources that have been used for studying human mobility patterns. Various forms of users data could be taken as input for these methods e.g. census or cell phone data. Census data, e.g., is often publicly available but at the same time it is mostly incomplete, whereas cell phone data is in abundance and up to date. We furthermore focus on the role of GPS technology and its data being used for understanding human mobility. GPS data collected by using tracking devices attached to mobile individuals is highly favourable for conducting studies on micro-mobility patterns. Reason being that such GPS data is regular and continuous, i.e. there are no long silent periods in the data (except when indoors). However such data is rarely available in public. This would thus require considerable efforts for data collection which involves manual labour such as finding the right subjects to track and attaching GPS trackers with them, and running data collection exercises over a period of time with these subjects [2] [4].

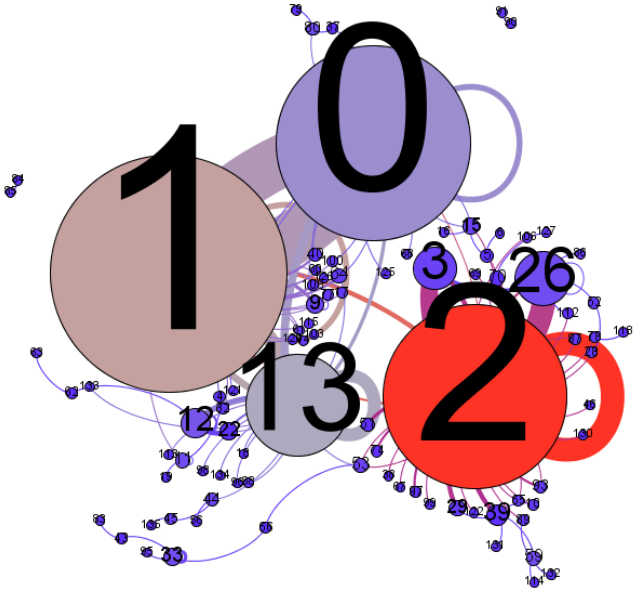

Figure 2.1: Connections between user's locations - Size and color of the node describes the number of times the user has travelled to and from that location. e.g. Red and large means frequent, blue and small means seldom. The width and color of the edge describes the number of times user travelled between these two locations [15].

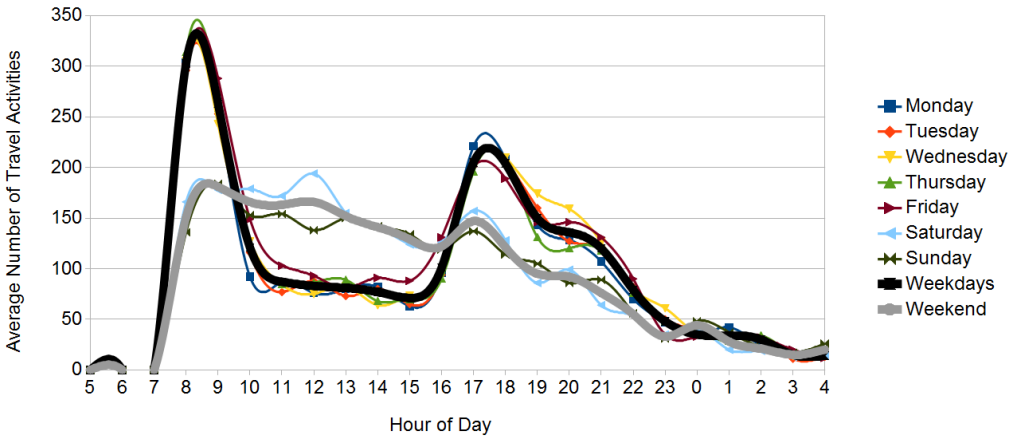

Figure 2.2: Daily travel activity during the weekdays and on weekends[15].

Chapter 3

Online Social Network and Data Mining

Our study aims to lower the barrier for conducting human mobility patterns study by making use of already existing and publicly accessible data on online photo sharing platforms. The photos sharing websites such as Flickr and Panoramio, has millions of users who are uploading pictures every day from all over the world, which contains the semantic data about the pictures in their headers. This data includes the user's information, the location information and the timestamps besides many other things. This kind of data has been used in the recent past for several purposes. However, there has been little or no work regarding the mobility patterns using photos sharing websites specifically. In this chapter we briefly list the contexts for which researchers have utilized the social networking sites data.

3.1 Analyses with Flickr

Flickr data has been used in a lot of useful research communities in the past for various purposes [6] [16] [17] [18] [19] [20]. The most popular use of Flickr data corresponds to either classification of photos on the basis of social tagging [18] [21] [19], Regions-of-Interest (ROI) or Points-of-Interest (POI) mining [6] [16] [17] [20]. Each of the concepts for which the Flickr data has been used are described briefly below.

3.1.1 Tag-based Content Analysis

Tagging is the feature which enables the users to add textual labels to the content on social media. They are freely chosen keywords that user associates

with either photos, website or blogs. They usually result in an unstructured knowledge, but they greatly reflect the properties of associated data due to the flexibility and variation in their nature [21]. The key contribution of Kennedy et al. in [18] include the tags, locations and image based content analysis for organizing the photos automatically which are taken in different parts of the world. This is done by structuring the disorganized public media content available from Flickr. Location and tags driven are the approaches used in their research. The location based approach defines the meaningful tags for the arbitrary locations in the world. Tag driven approach reveals the semantic information about location and event from the Flickr tags using the tag patterns. They also suggest the use of visual algorithms with the extracted patterns to fetch the geographical images from the Flickr dataset with a better accuracy. They succeed in filtering erroneous parts out of the noisy Flickr dataset and thus transforming haphazard Flickr data into useful information.

Rattenbury, Good and Naaman suggest the methods for extracting semantics regarding places and events using the concept of tag-based analysis in [19, 21]. They map the tags to either event or place which enables powerful and accurate image search and other photos relevant operations. They also evaluate the techniques used for knowledge extraction and propose that the 'scale-structure identification' in [21] performs better for this purpose.

3.1.2 Regions-of-Interest(ROI) Mining and Travel Patterns

Zheng et al. perform route analysis using the regions of interest or attractions mining in [20]. Regions of interest are the locations with high tourist visiting percentages. The number of tourists at these locations depict the popularity of the attraction. They have also analysed the travel route characteristics by various tourists. They do so by observing the traffic flow through the ROI's and clustering the routes in a sequential manner gives the topological characteristics of the routes. They perform the analysis on four major cities to understand the popular travel patterns.

Kisilevich et al. introduce modification in the standard DBSCAN clustering algorithm for mining points of interests(POI) using a huge collection of geotagged images on Flickr- which is very adaptive yet limited to POI identification [16]. The concept of density threshold and adaptive density in [16] are defined in their research. The threshold is set according to the count of people taking photos in the vicinity. The adaptive density is for optimizing the high density areas search and rapid convergence of the algorithm into

high density clusters.

In addition, Crandall et al. investigate the methods to organize a huge collection of geotagged photos on Flickr [22]. The proposed approach is a combination of context and structural analysis which identifies the attractive places. These locations are predicted using the visual, textual and temporal characteristics in their classification methodology. The resulting automatic tag suggestions tend to reduce the work load of adding manual geo locations from Flickr. Lee et al. find points of interests and the positive associative patterns among them by utilizing the geotagged Flickr photos in [17]. Clustering and associative pattern mining are the techniques used for data mining in their research. These associative relationships among the POIs help understand the travel patterns of the travellers. Most of the above mentioned studies aim to identify the most visited places or find collaborative patterns from tag-based analysis using the Flickr geotagged photos.

3.1.3 Travel Planners

Recently, the Flickr data has been used for route and itinerary recommendations by corresponding the geotagged photos to the sequence of visited locations and computing the travel patterns and sequences [23][24][25]20].

Okuyama and Yanai use Flickr images as sequence of locations for the travel route recommendation system in [25]. They use hierarchial clustering technique for finding popular landmarks and trip models for route generation. A new approach of personalized landmark recommendations by integrating both the user-based land mark preferences and category-based landmark similarity is introduced by Shi et al. in [26]. Their technique follows category based factorization approach and they prove with their results that the category based approach performs better than the typical popularity based landmark approach. Furthermore, study by Lu et al. [24] introduces the online customized automatic route planner by allowing users to specify their preferences for travelling by taking the geotagged images from Panormaio for the travel plan formulation. Their travel plan includes the must visit destinations, order of visits, time spending suggestions for each destination and the travel path to follow within the tourist attraction. Their approach is interactive and allows users to specify their preferences for locations, time, season etc. While Kurashima and his colleagues provide travel route recommendation by integrating the topic model and Markov model in their study [23]. They take photographers' travel route histories for the travel route recommendation method. Their model is probabilistic which takes the photographer's behaviour into account for making recommendations. Their model is composed of two sub models which are based on user's preferences

and the typical routes, together they ensure the prediction accuracy of the travel routes.

3.1.4 Positional Accuracy- Flickr and Panoramio

Flickr and Panoramio geotagged photos have been used together in studies for either finding positional accuracy of the photos [9] or defining digital tourist survey field [6]. Several accuracy specific questions have been addressed in [27]. For example the reliability of the user generated tags for describing the place and how to overcome the problems regarding the position, bias and accuracy. The second issue is the defining of the city center and urban space using tags. Moreover, they experimented the use of geo tags to gather the collective knowledge about the region. The Flickr images at Hyde park London are analysed to match the georeferenced and geographical positions. It is found that 86% tagged images for Hyde park covers both the internal and external premises(outer road side) of the park.The two dimensional positional accuracy in [9] for different image types has been tested by comparing the geotagged images locations and the manually corrected camera positions. The street related features are found to be most accurately positioned while the bridges tend to have lowest accuracy values. The vast amount of information provided by the people on different social media sites have been listed in [6] by characterizing and detecting the communities on these sites based on the user's behaviours.

3.2 Analyses with Instagram

Instagram, due to its rapidly growing number of users and photos has been in a lime light for the research community since 2010. Various analyses have been performed recently using Instagram data [5, 28–30]. The domain of studies primarily relate to applying computer vision techniques for examining the photo content [29]. They perform the quantitative and qualitative analysis of the photo content available at Instagram and cluster the active users based on their activities. Their findings include different categories of posted photos, users and the followers. The study [28] intends to reveal the social and cultural characteristics in the expanding visual information available at Instagram. In order to infer various insights from the visual information, they propose visualization framework for the analysis of location-based visual information flows. The visual material from Instagram has also been used for the detection of information like age of the user by utilizing textual and facial recognition methods and performing comparative study of teens

and adults [5]. They also infer the trends in teens and adults and their way of posting and commenting on photos and the major difference between the two age groups is that the teens post fewer photos but associate a lot of tags to attain more likes and comments. In addition, Bakhshi et al. conducts the study for understanding of the engagement and interaction of people with their content on social media in [30]. According to their study, photos of the faces of people is the most common type of photos shared on Instagram. They also analyse the impacts of faces, age and gender on the engagement activities and find out that photos having faces of the people receive 38% more likes and 32% more comments than the other photo types.

3.3 Conclusions

In this chapter, we have discussed several interesting studies that are related to our work. However, those research studies do not cover the mobility patterns of users at a fine grained level of distance and time, where data is collected from online photo sharing websites. They do not track the personal day to day routine of the people to make some inferences about their movement and activity behaviour. Our study is unique in a way that we focus on the individual user behaviours and compute the distances they cover by taking their sequenced geotagged photos as our data source. And taking into account the distances covered per day by all the users, we extract knowledgeable patterns in their daily routines. From our analysis, we also highlight the activity levels of the users on Flickr leading to the identification of pros and cons of using its data for mobility and micro-mobility patterns analysis.

Chapter 4

Flickr API

A prerequisite for this thesis is an in-depth understanding of the Flickr API, which is required especially in the data collection phase. Therefore it is necessary to get an overview of the basic structure of Flickr API and the mandatory steps required by Flickr API to obtain the data. The reason for choosing Flickr for this study is not only because of its popularity as a photo sharing website, but also due to its comprehensive and well documented API Explorer [1] that enables using the API simply through a web browser.

4.1 Introduction to Flickr API

Web API is basically a web development style to allow the data access and services. There are specific ways by which we can utilize API for searching and downloading data. We need to have a conceptual knowledge for searching the required data and interpreting the results obtained from API. Flickr API is based on HTTP request and response mechanism. The user or application sends an HTTP request to Flickr Web server, the server queries the database for the required data, gets data from the database and sends back an HTTP response to the user which contains data formatted in either JSON or XML. The general work flow of the webAPI is shown in Figure 4.1 and the steps that are required for our data collection process are described briefly below.

4.1.1 Obtaining a Flickr API Key

In order to use the services of Flickr API, we are required to sign up for an access key. This key is a number/integer string passed in an HTTP request

[1] https://www.flickr.com/services/api/

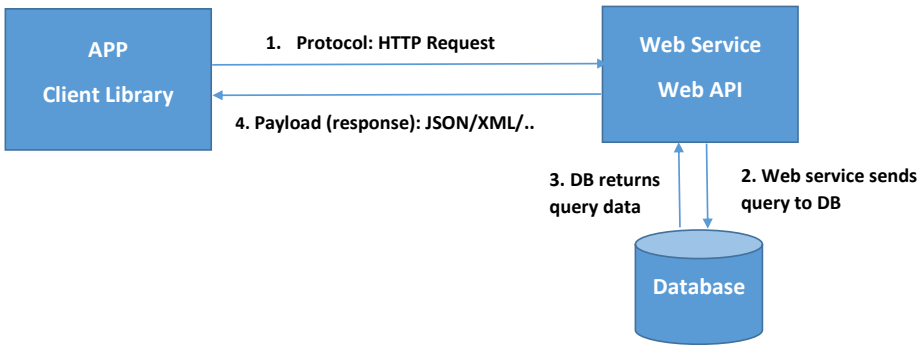

Figure 4.1: Client-Webserver request-response flow through an API.

to certify the source to the API. We need to register our application with Flickr to get this key. We have to follow the registration method[2] to get a key every time and we can see the obtained keys on our personal page[3] after signing in [31].

The use of key is a typical process for many web application APIs and they are a source of introduction and verification of the user to the API provider as it can be limited to at least the email address of the user. This is also a way for API provider to keep a record for the users about their API call number and access style. If the volume of calls exceed the granted limit, then it might apply the terms of use by blocking access to the API through that key [31].

Once we have a key, we can make a simple call to Flickr API database using the corresponding. The limit for making API calls is 3600 requests/hour for normal users. Flickr API also provides a commercial key, but as it does not facilitate the user much with enhanced bandwidth option, we use the basic research/student key for our study.

[2]http://www.flickr.com/services/api/keys/apply/
[3]http://www.flickr.com /services/api/keys/

4.1.2 Flickr API Methods Used

We have used multiple methods from Flickr API for our data search. All the methods used in a sequence are briefly described next.

- *flickr.people.findByUsername*: At first we randomly choose our root user having thousands of photos and hundreds of contacts and plenty of groups joined on Flickr website. Then we pass the username to this method to get the user id, NSID which is later used as a parameter for various methods.

- *flickr.photos.search*: This method takes parameters like userid, upload dates etc and returns an xml response containing number of photos, number of pages, description, photo id,owner, location information and date for each photo. We can specify the number of photos per page, which is 50 by default.

- *flickr.contacts.getPublicList*: This method takes the userid as input and gives all the contacts with their userids and names in response.

- *flickr.people.getGroups*: Takes userid as an input parameter and give all the groups joined by the user along with group id, group name, members count etc. in response. This method requires OAuth authentication with read permissions.

- *flickr.photos.getExif*: This is the method which outputs a list of EXIF, TIFF and GPS tags for a photo. It takes a photo id as an input parameter and requires the calling user to have a view permission for that photo. The example EXIF response can be seen in Figure 4.2

```
<photo id="4424" secret="06b8e43bc7" server="2">
  <exif tagspace="TIFF" tagspaceid="1" tag="271" label="Manufacturer">
    <raw>Canon</raw>
  </exif>
  <exif tagspace="EXIF" tagspaceid="0" tag="33437" label="Aperture">
    <raw>90/10</raw>
    <clean>f/9</clean>
  </exif>
  <exif tagspace="GPS" tagspaceid="3" tag="4" label="Longitude">
    <raw>64/1, 42/1, 4414/100</raw>
    <clean>64° 42' 44.14&quot;</clean>
  </exif>
</photo>
```

Figure 4.2: EXIF Response from Flickr API [32].

4.1.3 User Authentication-OAuth

Signing in to the Flickr website is mandatory for making calls to various methods of Flickr API. OAuth specification is an industry standard that enables the secure login for the Flickr users with all the supporting account types such as gmail, yahoo etc [33]. The authentication via OAuth requires three steps shown in Figure 4.3 and described as follows:

- *Request Token:* In the first step we obtain a temporary request token from the Flickr API. The token along with its secret authenticates the user to the application [33].
- *User Authorization:* The application should redirect the user to the authorization page after obtaining a request token. This page contains the permissions being requested by the application. After accepting the authorization, the user is redirected to the application again [33].
- *Access Token*: Once the application is authorized by the user, the request token is replaced by access token which can be stored in the application database for making authenticated requests to Flickr in future [33].

4.2 Limitations of Flickr API

Flickr API being an easy and handy tool for accessing the web data, also has some limitations that prevents us from extensive research options. Perhaps the biggest limitation that we encountered in our study is the unavailability of a method which can retrieve all users's public photos containing EXIF tags.

The method *flickr.photos.search* does not take any parameter, where we could specify the access to EXIF tagged photos. The only option is to first retrieve all the photos of a user and then call the method *flickr.photos.getExif* to check if a single photo contains EXIF tags and whether those EXIF tags describe the GPS information or not. Consequently, one easily exceeds the API limit of 3600 requests per hour [34] when accessing thousands of photos for thousands of users. There seems to be no other method to work around this problem. The only possible way is to get the Flickr geotagged photos (lead over to next chapter) and perform an analysis relying on the geodata obtained for the photos from this method.

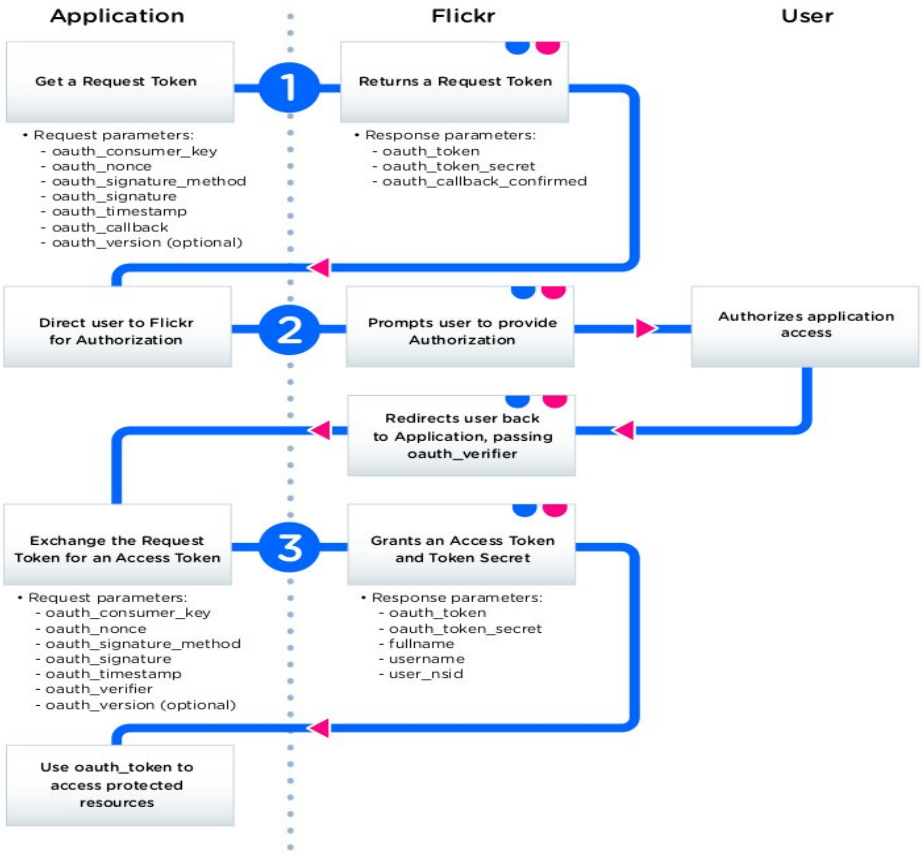

Figure 4.3: Flow Diagram-Flickr OAuth Authorization [33].

Chapter 5

Flickr Location Data: Collection

The first step in our study is to collect the data from the Flickr website. The data collection phase is one of the most crucial phases of the study. The end results depend entirely on the amount and type of data taken as input. In order to get an appropriate set, we first list the requirements for data that we need for our analysis. This procedure comprises multiple stages of fetching and filtering until we finally achieve the appropriate subset of data for the mobility analysis. In this chapter, we describe the data collection strategy defined and used in our study, and then we discuss also some implementation details for fetching data using the FlickrAPI[1].

5.1 Data Requirements

In order to analyse the mobility patterns of the users on Flickr, we should have abundant amount of users to get realistic patterns. Furthermore, each of the user must be active on Flickr, meaning that he regularly takes and uploads his photos on the website. In an ideal case, we aim to get micro-mobility (see Appendix A.4) trails of the users. To this end, we constrain ourselves to users who are taking more than 1000 pictures in a day. This means that a user takes on average at least 40 pictures in an hour or 1 picture per minute. However, since not all the users take pictures every minute, so the required total amount of pictures (i.e. 1000) could be taken at any time during the day. For example, the user who takes more than 1000 pictures, but only during the mornings is also a valuable user for drawing a micro-mobility trail and interesting information can be extracted from his activity times. Having this kind of data, we can easily draw accurate trajectories for

[1] https://www.flickr.com/services/api/

the users and combining the trajectories can reveal detailed patterns from the movement behaviours.

5.2 Search Strategy

After defining the requirements for the data, we construct a search strategy to collect the users and their photos meta data from the Flickr website.The most effective and logical way is to take a random subset of users and their photos for the analysis. In previous studies, many different approaches have been applied for data collection from Flickr, see, e.g. [35, 36]. The data collection methodology used by [35] is such that the crawl begins with a randomly chosen user on Flickr following the friends links with a breadth first search in a forward direction. The contact list for users is public on Flickr and they use FlickrAPI to download all the data. Our data collection strategy is similar to [36] and will be discussed next.

5.2.1 Breadth First Search

In this section, we discuss our data collection strategy which resembles to the one proposed in [36] but is slightly different from the one used in [35]. In order to get users from different regions, we incorporate the groups in our search criteria. So the root user having a bunch of contacts is chosen randomly from Flickr and then all his contacts are stored, then for each contact the groups that they belong to are obtained (see Figure 5.1). Utilizing the group information from FlickrAPI methods, we retrieve all the members of the group to build up their dataset. Here we ensure that all the users obtained are having greater than 1000 pictures. The time span chosen for our study is Jan 2013 to Jan 2015 [2]. The breadth first search could be performed at different depth levels to gather the required amount of data. Alternatively, we could also pick another root user from a totally different domain and perform this search. The breadth first search of root user $\rightarrow$ contacts $\rightarrow$ groups $\rightarrow$ group members make it one recursion, and we performed this search up to depth level 1, which already gave us an abundant amount of data altogether, sufficient for the study and analysis purpose. This tells that the users are well connected and closely grouped on Flickr.

[2]The reason for choosing the time period of Jan 2013 to Jan 2015, is to obtain the most recent data. This ensures the presence of EXIF feature in the devices. Moreover, the users have become more educated about the geotagging facility on Flickr in the recent years, so there are less chances of random tagging errors. As the study began from Feb 2015, the most recent data available at that time is gathered

5.3 Flickr API Methods for Data Retrieval

Our particular interest relates to the photos meta data. To reduce the memory constraints, we only store the meta data of each photo that is necessary for our analysis purposes. After having all the user ids stored in our database, we call Flickr API methods to get information of the corresponding photos for each user. We explored two different techniques in order to get the attributes like location, timestamp and photo id of each photo. Figure 5.1 depicts the high level flow diagram for data collection using the FlickrAPI.

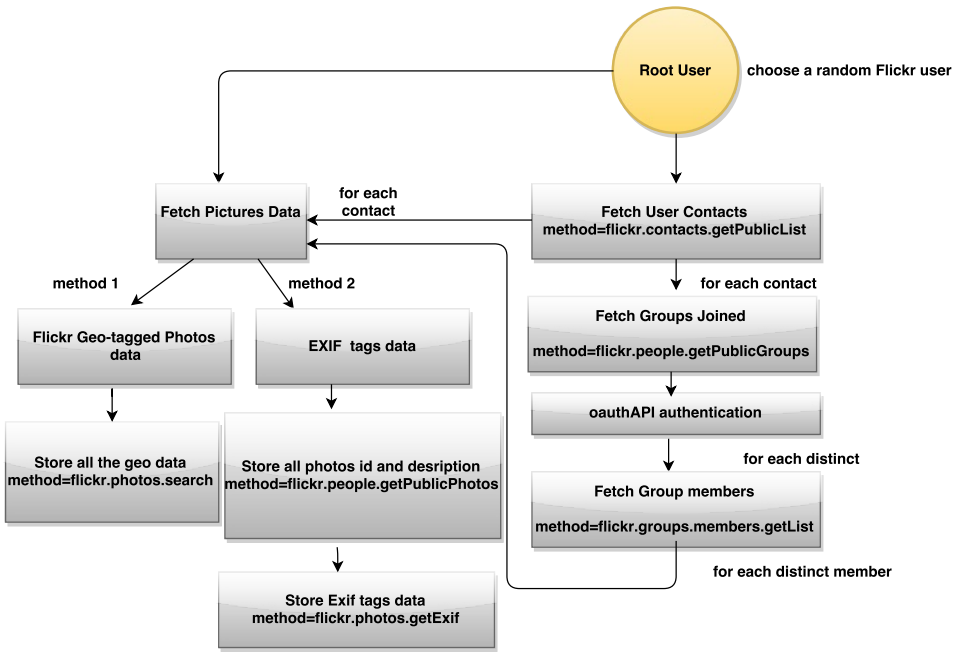

Figure 5.1: Data Search Methodology for Recursion level 1.

5.3.1 EXIF Tags Method

The approach which initially we use to get the photos meta data is through EXIF tags. The Flickr website keeps the EXIF tags separate from the photos. So, in order to get the EXIF data from the FlickrAPI, we first store all the photo ids for each user in a local file by calling the API method *flickr.people.getPublicPhotos*. Then, for each photo we call the API method

flickr.photos.getExif. If an EXIF tag exists for the photo, and if it contains both GPS and timestamp information, which is public, then we store the metadata in our database. This is done by creating separate tables or records for each user and storing photo details for each photo of that user. The example of storing files with EXIF tags information is shown in Figure 5.2.

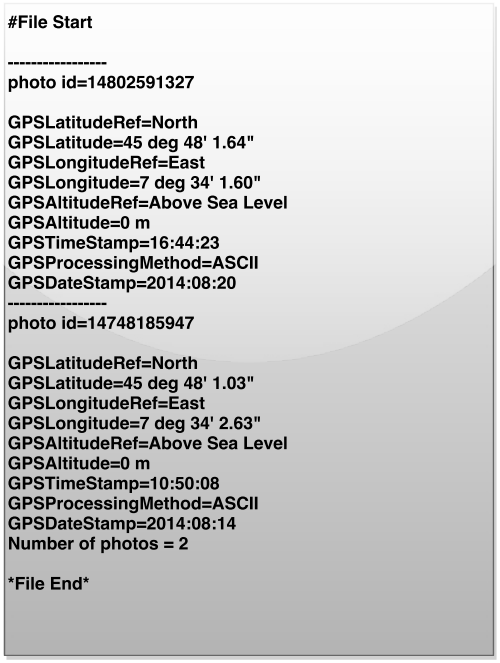

Figure 5.2: Example of User's Photos Metadata from EXIF Tags.

5.3.2 Geotagged Photos Method

The other way in which information of photos can be retrieved is using the method *flickr.photos.search* and giving parameters userid, hasgeo=1, extra==geo, date_taken, date_upload, description. This gives all photos details having the geo location information and the timestamps. This geo location information is added up by Flickr itself, and it is extracted from the EXIF tags of the photos. The Flickr users can, however, set these parameters manually. Figure 5.3 shows an example of how we store the metadata of photos for each user by using the geotagged photos information.

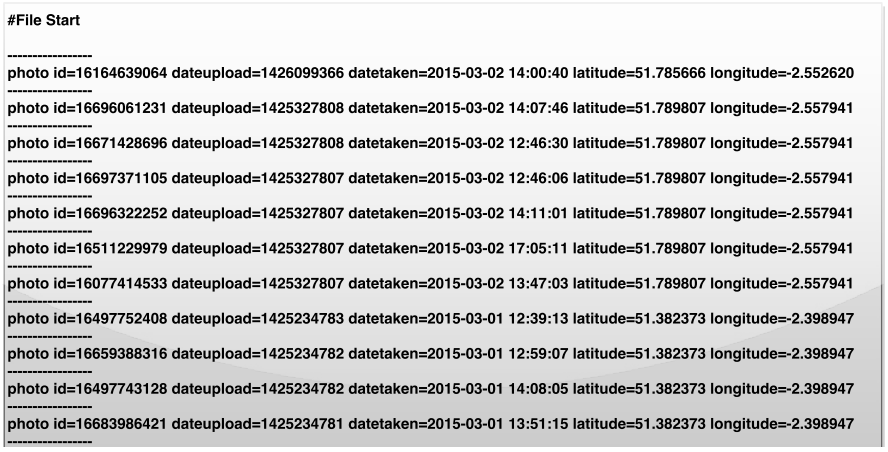

Figure 5.3: User's Photos Metadata from Geo Tags.

5.3.3 Comparison of Flickr API Methods

The second approach is much more efficient than the first one. The *getExif* method of FlickrAPI is extremely slow because we need to call the API method for each photo separately. So, if a single user has 10,000 photos, then the getExif method is called 10,000 times even if none of the photos contains an EXIF tag. In particular, it seems that there is no short cut to get this information such that we only call this method if the photo contains EXIF tag with location and time information or not.

Mostly EXIF geo data and the Flickr geo data are the same, as most geo data is simply extracted from the EXIF tags by Flickr. However, as Flickr allows users to manually enter their location, so there can be photos with user marked location and without EXIF. We may not have permissions to view the EXIF or Flickr geo data depending on how a user has set his privacy. In our case, the total userids stored are 150,000 and out of those, less than 1000 users contain more than 1000 photos with EXIF tags. However, the users with more than 1000 geotagged photos are about 8000. The EXIF data collection procedure took more than a month to traverse all the users while the geotagged data was obtained in couple of days. We take the geotagged metadata in our study for the analysis due to its abundant and fast retrieval [3]. We also get this information from this data collection phase that out of 150000, only 8000 users have geotagged photos. This leads to the

[3]The query running time for EXIF data collection would be much faster if, e.g. Flickr runs the same query internally

inference that about 5% of the photos on the Flickr website are geotagged which conforms to the statistics made by Zielstra and Hochmair in [9].

5.4 Facts About the Data Collected

The data collected from this strategy had 40% of the users belonging to North and South America, 35% were from United Kingdom and the rest of 25% belonged to different parts of Europe and Asia (Figure A.6). We believe that the large percentage of users from these two countries is a bias introduced by the choice of our root user, who resided in US. This is because there is a higher probability that the root user has contacts belonging to the same region. Also the groups joined by the contacts tend to be similar because of the same language medium and common interests among the contacts, hence the obtained users belonged to similar regions. There were quite a few outliers as well, about whom we can say that they were most likely either immigrants or visitors in these countries.

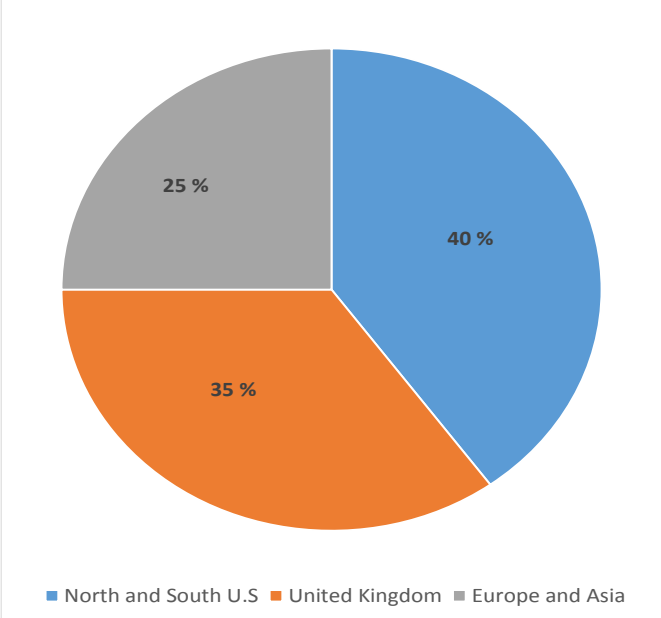

Figure 5.4: Location Distribution of Collected Users.

Chapter 6

Flickr Location Data: Preprocessing

In this chapter, we explain the preprocessing steps performed on our data including the mechanism to filter the most active periods of the active users sampled from the data collection phase. The data stored from the data collection phase is not ready to be used as such for analysis. The criterion of having 1000 photos over 2 years span is not sufficient to make accurate predictions about the mobility. There are possibilities of having many different variations in the users activities, some upload a bunch of photos only few times in a year, say during the holiday, and remain dormant rest of the time, while others upload the photos regularly over the observed period of time.

In order to make our mobility estimates more accurate, we are only interested in the days when users have taken their photos. We should filter out the days with no activity. There can be multiple ways in which we can filter out the sampled data to get the active periods, such as;

- *Photo Count*: Count the number of photos per day. If the count is greater than a defined threshold, than consider it as an active time and store it, otherwise ignore that day.

- *Distance Count*: Measure the distance travelled per day. If the distance covered is greater than a defined threshold, say X km, then consider it as an active day and store, otherwise ignore that day.

- *Sliding Window*: This is basically a rolling or sliding window concept, in which we take snapshot of the data over a fixed time period. We refer to this as a window. Then we slide the window over the entire data set, each time sliding it forward by a fixed time interval. We can analyse the data within the window at every step. If the count of photos

inside a window is greater than a predefined threshold than consider it as active time. We design sliding window mechanism according to our data requirements in order to detect the active days, the details of which are explained below.

6.1 Sliding Window Mechanism

The Sliding Window Mechanism is based on the following algorithm: *If the N hours window contains at least P number of geotagged photos, then that window is an active window and that period is an active time period of the user. The window progresses every h hours. If the user has at least W active windows in a search time span, then he is an active user.* See Figure 6.1 for details.

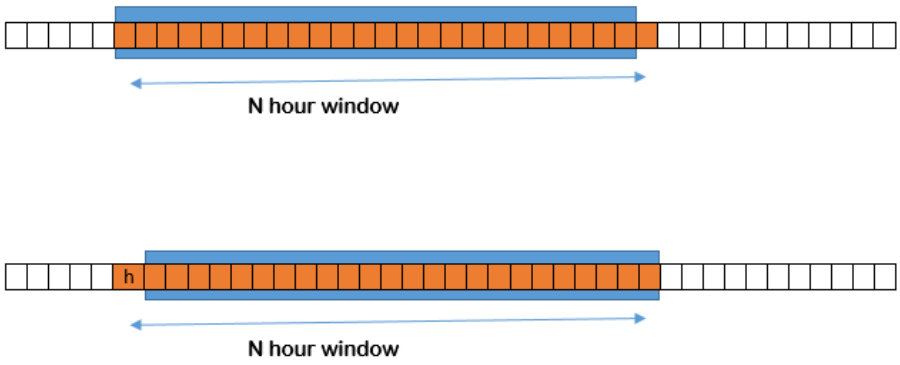

Figure 6.1: Sliding windows mechanism. N hours window moving forward every h hours.

We make an initial hypothesis that if the users on the Flickr upload about 100 pictures or more in a day, then it's a good count to track their mobility trail. In order to achieve that, we use the concept of sliding window mechanism, where the active period is the time window of 24 hours ($N = 24$), when user has uploaded photos above a defined threshold P. The 24 hour sliding window starts from 01-Jan-2013, progresses every hour ($h = 1$) up to 01-Jan-2015. All the non-overlapping active windows detected during this

time period are stored for each user, along with the locations and timestamps of each photo. The empirical cumulative distribution function graphs are a good source for evaluating the data set and observing the distribution trends. Analyzing the ECDF of the total number of pictures per day, it was found out that the most active users on the Flickr website upload 30 or less photos in a 24 hours period. This can be seen by an ECDF of number of pictures in a day (see Figure 6.2). The final threshold to classify an active window is set to $P = 20$ photos per day and users having at least 15 ($W = 15$) active days in the total time span of 2 years are considered active and useful. Applying the sliding window mechanism with this threshold gave the count of users be 440 out of 8000. These users are called as interesting or active users, who enable predicting meaningful mobility trends. The overall summary of data collection can be seen from Table 6.1.

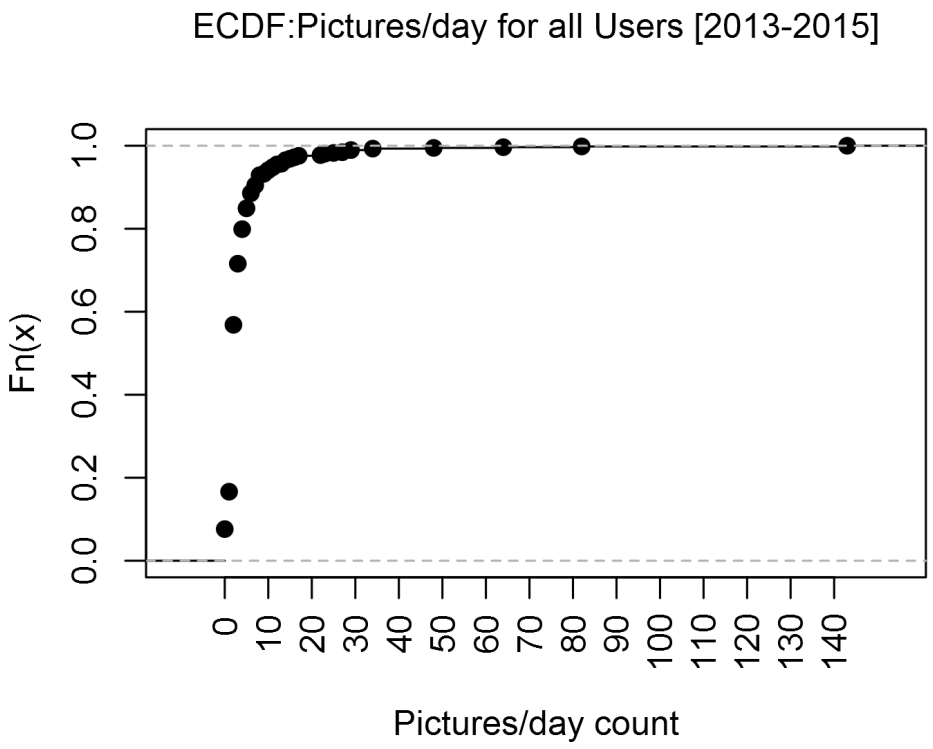

Figure 6.2: ECDF for count of photos/day for all users.

Our initial hypothesis of getting users with more than 1000 pictures/day becomes false with this ECDF and we did not get the expected results for

the data. This implies that the idea of understanding micro-mobility seems vague at this point with this kind of data set. Despite of this, we would still want to know upto what level correctly can we understand the behaviour of the people. To this end, we continue with the active time periods of the users and perform further analyses. The overall data collection statistics can be seen in Table 6.1.

6.2 Distance Computation

Once having all the required filtered data with us, we consider doing some mathematical operations and scripting to study different aspects of mobility. One aspect is taking the number of pictures at defined intervals and try to reveal the interesting information related to the picture count. This can relate to finding the favourite spots of the users and ranking them to detect most popular to least popular destinations.

The other aspect to uncover the mobility patterns can be studied by taking the distance into account. We have information like longitude, latitude, datetaken and dateupload for each photo which we use for further preprocessing to reach the analysis phase. Therefore the next step in the study is to find out the distances covered by each user per day using his photos timestamps. For that, we compute the distance between every two consecutive photos that reside in an active window, and that is followed for all the active periods sequentially. The distance between the two consecutive photos using their longitude and latitude can be computed using Haversine Formula[1][37], which finds the great circle distance between the longitude and latitude values and it can also be stated as shortest crow flight between two points over the earth's surface (for details of longitude and latitude see Appendix A.2).

$$
\begin{aligned}
a &= \sin^2(\Delta\phi/2) + \cos\phi_1 \times \cos\phi_2 \times \sin^2(\Delta\Lambda/2) \\
c &= 2 \times \arctan 2(\sqrt{a}, \sqrt{1-a}) \\
d &= R \times c
\end{aligned}
$$

" Where ϕ is latitude, Λ is longitude, R is the radius of the earth (mean radius = 6,371km), c is the angular distance in radians, and a is the square of half the chord length between the points " [37].

These computed distances initially tell us how much the user has been moving. For example, if the distance between two consecutive photos is zero, then the corresponding photos have been taken in the same place. To make it more meaningful, we define the static state or pause time when the user takes

[1]http://www.movable-type.co.uk/scripts/latlong.html

Year	2013-2015
Total users crawled	150000
Users with geotagged photos	8000
Users with active windows	440
Total photos per user	>1000

Table 6.1: Flickr data collected for mobility trend analysis

photos within 20m distance. Figure 6.3 shows an example of a file stored for a user with various computations performed for consecutive photos. long1 and latd1 are the longitude and latitude of the first photo, long2 and latd2 are the longitude and latitude of the next consecutive photo, distance(m) is the computed distance using Haversine Formula, timeDifference(s) is the time difference between two photos taken, velocity is computed using the formula v=d/t and pauseTime is set to true if the distance is less than 20m, and false otherwise. After the distance calculations for each user, the cumulative distances are calculated in order to visualize them better on the graphs and for better interpretations.

long1,latd1,long2,latd2,d1,d2,distance,TimeDiff,velocity,pauseTime
-99.085972,40.672495,-90.368728,41.586589,"2005-06-19 11:23:59","2005-06-19 14:14:24",736788.74635911,10225,72.057579106025,False
-90.368728,41.586589,-104.992263,39.740009,"2005-06-19 14:14:24","2005-06-19 17:17:01",1248790.2369026,10957,113.97191173702,False
-104.992263,39.740009,-104.987711,39.744096,"2005-06-19 17:17:01","2005-06-19 17:34:52",598.30708809216,1071,0.55864340624852,False
-104.987711,39.744096,-104.987711,39.744096,"2005-06-19 17:34:52","2005-06-19 17:46:02",0,670,0,True
-104.987711,39.744096,-104.987711,39.744096,"2005-06-19 17:46:02","2005-06-19 17:47:10",0,68,0,True
-104.987711,39.744096,-104.987711,39.744096,"2005-06-19 17:47:10","2005-06-19 17:50:57",0,227,0,True
-104.987711,39.744096,-104.987884,39.744189,"2005-06-19 17:50:57","2005-06-19 17:50:57",18.046828745753,0,0,False
-104.987884,39.744189,-104.987711,39.744096,"2005-06-19 17:50:57","2005-06-19 17:54:32",18.046828745753,215,0.083938738352339,True
-104.987711,39.744096,-104.987884,39.744189,"2005-06-19 17:54:32","2005-06-19 17:54:32",18.046828745753,0,0,False
-104.987884,39.744189,-104.987884,39.744189,"2005-06-19 17:54:32","2005-06-19 17:55:37",0,65,0,True
-104.987884,39.744189,-104.987863,39.744123,"2005-06-19 17:55:37","2005-06-19 17:58:52",7.5551980368109,195,0.038744605316979,True
-104.987863,39.744123,-104.987711,39.744096,"2005-06-19 17:58:52","2005-06-19 18:06:37",13.33700749859,465,0.028681736556108,True
-104.987711,39.744096,-104.987711,39.744096,"2005-06-19 18:06:37","2005-06-19 18:07:32",0,55,0,True
-104.987711,39.744096,-104.990158,39.744075,"2005-06-19 18:07:32","2005-06-19 18:25:29",209.21815559221,1077,0.19426012589806,False
-104.990158,39.744075,-104.990158,39.744075,"2005-06-19 18:25:29","2005-06-19 18:27:21",0.094930730546311,112,0.00084759580844921,True
-104.990158,39.744075,-104.990158,39.744075,"2005-06-19 18:27:21","2005-06-19 18:27:48",0.094930730546311,27,0.0035159529831967,True
-104.990158,39.744075,-104.990158,39.744075,"2005-06-19 18:27:48","2005-06-19 18:28:11",0.094930730546311,23,0.0041274230672309,True
-104.990158,39.744075,-104.990158,39.744075,"2005-06-19 18:28:11","2005-06-19 18:28:42",0.094930730546311,31,0.0030622816305262,True
-104.990158,39.744075,-104.990158,39.744075,"2005-06-19 18:28:42","2005-06-19 18:28:42",0.094930730546311,0,0,False

Figure 6.3: Distance Between Every Two Photos Lying in Active Windows.

6.3 Conclusions

The data preprocessing is an important and non-trivial task. When data points are distributed across fixed window boundaries, one window may not

```
totDistance,totTime,totalPhotos,startTime,endTime,normalizedDistance
0,23113,122,"2013-06-01 10:13:32","2013-06-02 10:13:32",0
1339.8351005195,5173,44,"2013-06-08 10:13:32","2013-06-09 10:13:32",30.450797739079
803.5611804436,2091,24,"2013-06-09 10:13:32","2013-06-10 10:13:32",33.481715851817
0,4038,71,"2013-06-15 10:13:32","2013-06-16 10:13:32",0
56526.721806572,41973,21,"2013-07-19 10:13:32","2013-07-20 10:13:32",2691.7486574558
1518.696437396,4315,112,"2013-08-29 10:13:32","2013-08-30 10:13:32",13.559789619607
1350.2025522109,13292,267,"2013-08-30 10:13:32","2013-08-31 10:13:32",5.0569383977937
3618.6827919166,10117,63,"2013-10-03 10:13:32","2013-10-04 10:13:32",57.439409395502
4462.0331061972,5752,43,"2013-10-04 10:13:32","2013-10-05 10:13:32",103.76821177203
15475.511277232,13937,59,"2013-10-05 10:13:32","2013-10-06 10:13:32",262.29680130901
680.58402547867,9347,41,"2013-10-12 10:13:32","2013-10-13 10:13:32",16.599610377529
505.33488123641,1390,44,"2013-10-26 10:13:32","2013-10-27 10:13:32",11.484883664464
71618.654272259,155946,39,"2013-11-25 10:13:32","2013-11-26 10:13:32",1836.3757505708
104529.25228672,82667,22,"2013-11-26 10:13:32","2013-11-27 10:13:32",4751.3296493961
4492.9023151479,5470,61,"2014-02-21 10:13:32","2014-02-22 10:13:32",73.6541363139
13.195371545937,10223,139,"2014-03-15 10:13:32","2014-03-16 10:13:32",0.094930730546311
4570.6446812776,14510,22,"2014-05-17 10:13:32","2014-05-18 10:13:32",207.75657642171
4513.8907626971,70381,74,"2014-09-29 10:13:32","2014-09-30 10:13:32",60.998523820231
2833.9408128935,5236,41,"2014-10-01 10:13:32","2014-10-02 10:13:32",69.120507631548
6530.8660608271,7838,22,"2014-10-02 10:13:32","2014-10-03 10:13:32",296.85754821941
6263.672288367,15673,20,"2014-11-16 10:13:32","2014-11-17 10:13:32",313.18361441835
6056.1279947347,5064,53,"2014-11-24 10:13:32","2014-11-25 10:13:32",114.26656593839
```

Figure 6.4: Commulative Distances Covered by All Photos Taken in a 24 Hours Time.

be able to capture all of the data points, and this may lead us to miss a potential active period. Sliding window ascertains that we do not have window boundaries at fixed times, rather we slide gradually over all data points and capture active periods more effectively. Figure 6.5 explains how sliding window captures better results than fixed windows. Green data points included in first fixed window (light green), orange data points included in second fixed window (shows in yellow). Data points in these windows independently do not add up to the threshold. Finally the blue sliding window captures some green and some orange data points which add up to be greater than the threshold.

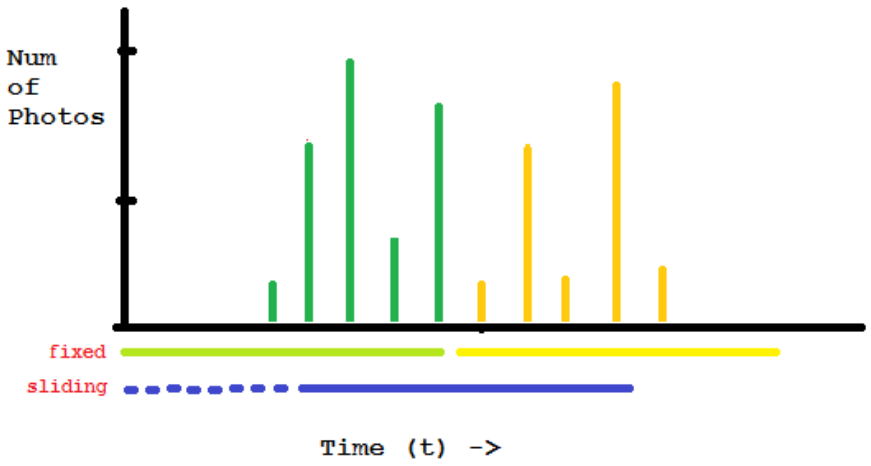

Figure 6.5: Sliding Window vs. Fixed Window.

Chapter 7

Flickr Location Data: Statistical Analysis

This chapter covers the analysis and results phase of our study. We analyse the pre-processed data by employing statistical methods and plotting multiple graphs to visualize interesting trends (see Appendix A.3 for a description of the relevant statistics). First, with the help of various graphs, we are able to classify Flickr users into different categories based on their activity levels. Then we infer mobility patterns for each category. The classification of users and detection of mobility trends are major steps in the data analysis, and we discuss them in detail below.

7.1 Classification of Users

Analysis of cumulative distance per day graphs (see Appendix A.3 for graphs) for the active users leads to the interpretation that there are similarities and differences in the travelling routines of the users. In order to cluster similar users to same groups, we need to define a classification strategy based on the suitably chosen parameters. The two main parameters used to classify the active users are:

$$D_i = cumulative\ distance\ travelled\ in\ a\ day$$
$$F_i = No.\ of\ travels\ per\ month$$

Definition 1 *Travel is defined as movement from one point to another by means of e.g. car, bus, aeroplane or by foot.*

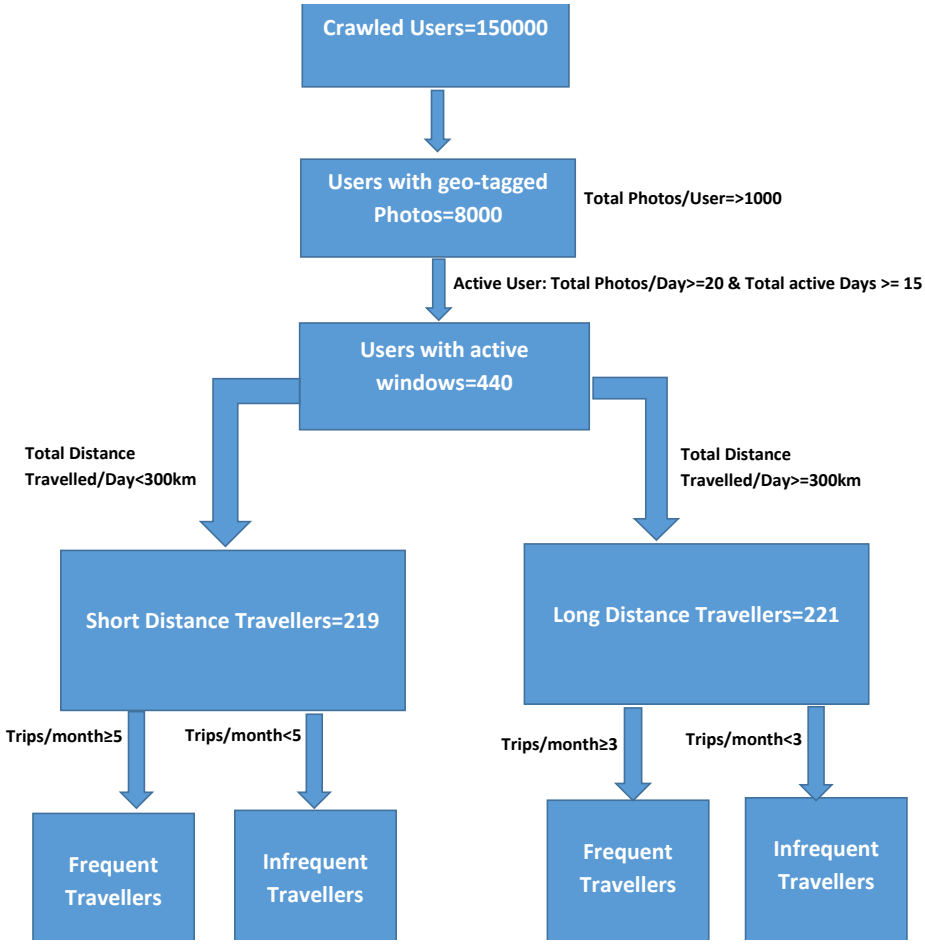

Figure 7.1: Classification of Users.

We categorize the active users further into sub-classes as listed below. We fine-tune the parameters used to classify users into these categories by applying trial and error method on our data. A high level view of users classification is shown in Figure 7.1, explained in detail next.

7.1.1 Short and Long Distance Travellers

Out of the 440 active users, about 219 users are the ones that travel at short distances (a distance of 300km or less) mostly, either frequently or infrequently. However they might travel at long distances no more than thrice in the 2 years span. These users are categorized as short distance travellers having threshold $D_i < 300$km. We can define the short distance

an long distance travellers as:

Definition 2 *A user is a* short distance traveller *if he travels less than 300km ($D_i = 300km$) in a day, and otherwise a* long distance traveller.

The distance vs. time graph of one of a frequent short distance traveller can be seen in Figure 7.2. The x-axis represents that time period when the user has been active, the scale increments with the months after detecting the time of the first active window. The y-axis represents the cumulative distances travelled on each active day. This distance is the sum of all the distances between every two consecutive photos in an active window. This graph depicts that the user was active from May-2013 to Nov-2014 and travelled frequently, at least 5 times a month, at short distances of about 20 km to 50km. The other interesting information which is inferred from this user

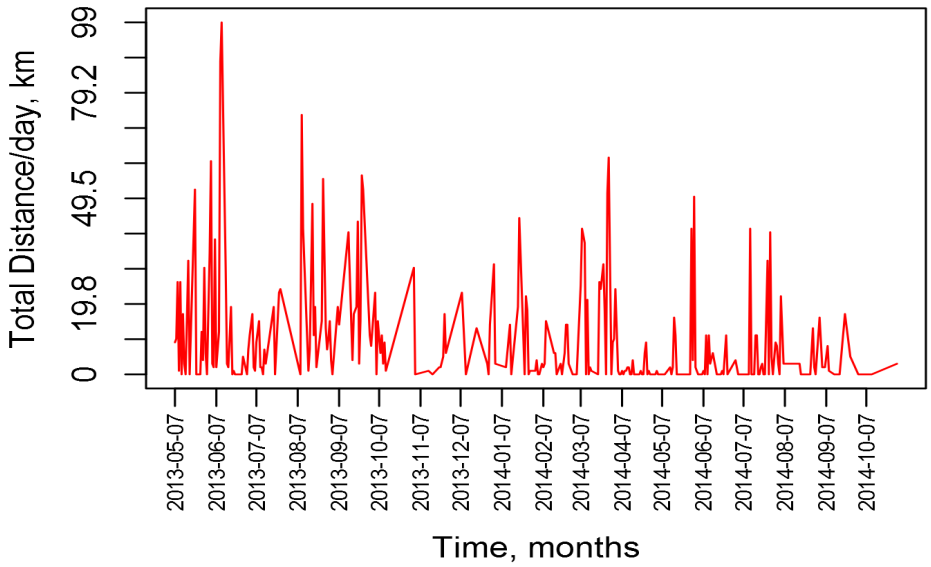

Figure 7.2: A Sample Short Distance Frequent Traveller.

is that he travels at short distances to the same locations. It is obvious in the Figure 7.3, that the user has a few favourite spots that he visits. This means that there is no randomness in his travelling routine. His expected travel destinations are the same, which are near London, Grays and Maid stone in United Kingdom (atleast when the user takes photos).

This kind of similarity is found among all the short distance travellers. In particular, these type of movement patterns can make the Flickr data a good candidate for user profiling purposes.

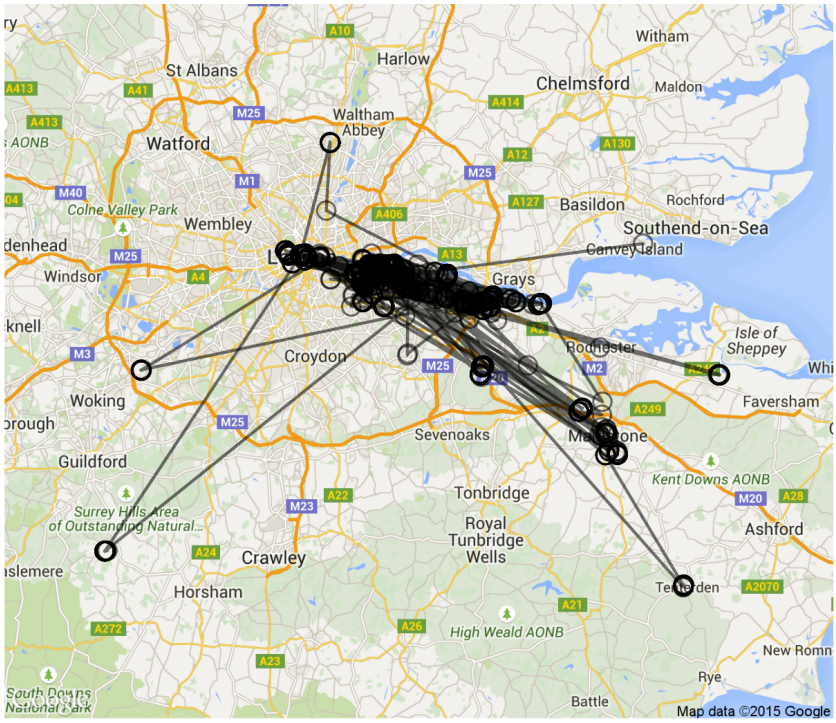

Figure 7.3: Short Distance Traveller (movement on the map).

The remaining 221 out of 440 users are the long distance travellers as defined in 2. Furthermore, we can also say that they take pictures only when they are on longer journeys. Figure 7.4 shows a long distance traveller who travelled mostly between 500 to 1000 km and his activity level was very high during this time.

7.1.2 Frequent and Infrequent Travelers

We further classify short distance travellers and the long distance travellers from the active users set into frequent and infrequent travellers. We do that for each of the short and the long distance traveller's category, by setting a threshold Fi, say, 5 journeys per month repeating almost every month, to fall in the short distance frequent travellers and 3 journeys per month to categorize as long distance frequent travellers. This threshold was chosen after manual analysis of the individual (Time vs. Total Distance/day) graphs. They can be formally defined as:

Definition 3 *A user is a* short distance frequent traveller *if he makes at least 5 short*$(D_i <= 300km)$ *distant trips in a month* $(F_i >= 5)$*, otherwise*

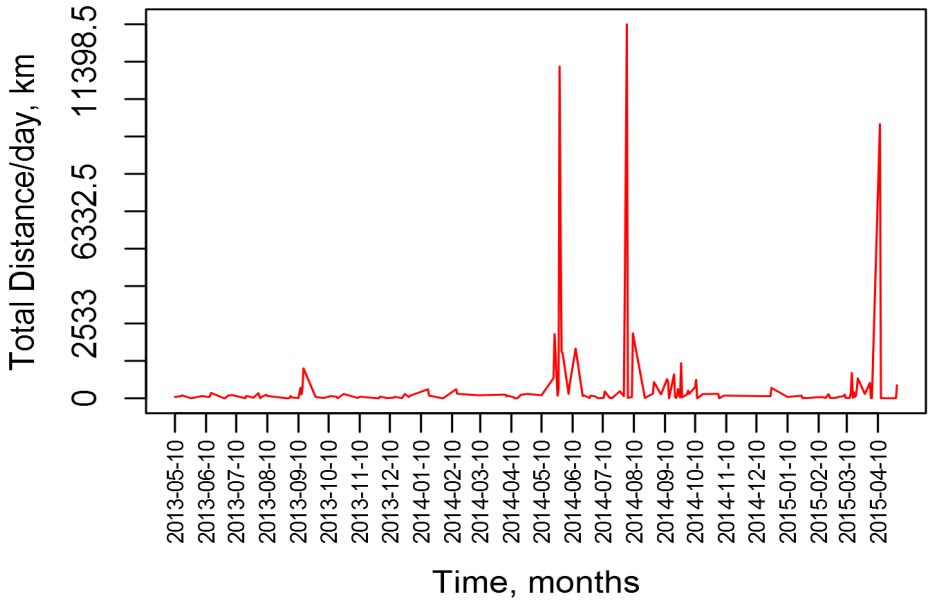

Figure 7.4: A Sample Long Distance Frequent Traveller.

he is an infrequent traveller.

The short distance infrequent traveller's category example is shown in Figure 7.5 along with the travel visited locations in Figure 7.6. The first Figure 7.5 shows that the user has no specific pattern or regularity in his movement. He prefers to travel short distances but he is definitely not a daily routine active member on Flickr. He seems to be active only during the month on July, 2014, but that behaviour is not repeating in 2013 or 2015. However, if we look at Figure 7.6, we get the information that this particular user has taken a lot of pictures on the visited spots. So, basically he was chosen as an active member by our sliding window mechanism due to the large number of photos. However, a closer inspection reveal that the user has made only a few trips during which he has taken a lot of photos there.

Definition 4 *A user is a* long distance frequent traveller *if he makes at least 3 long ($D_i > 300km$) distant trips in a month ($F_i >= 3$), otherwise he is an infrequent traveller.*

The example of the long distance infrequent travellers is shown in Figure 7.7. The Figure 7.7 tells that this user has barely made two long trips in a month. This excludes the user from the frequent traveller's category.

In order to focus on the mobility trends at urban level, we take the category of short distance travellers for further analysis.

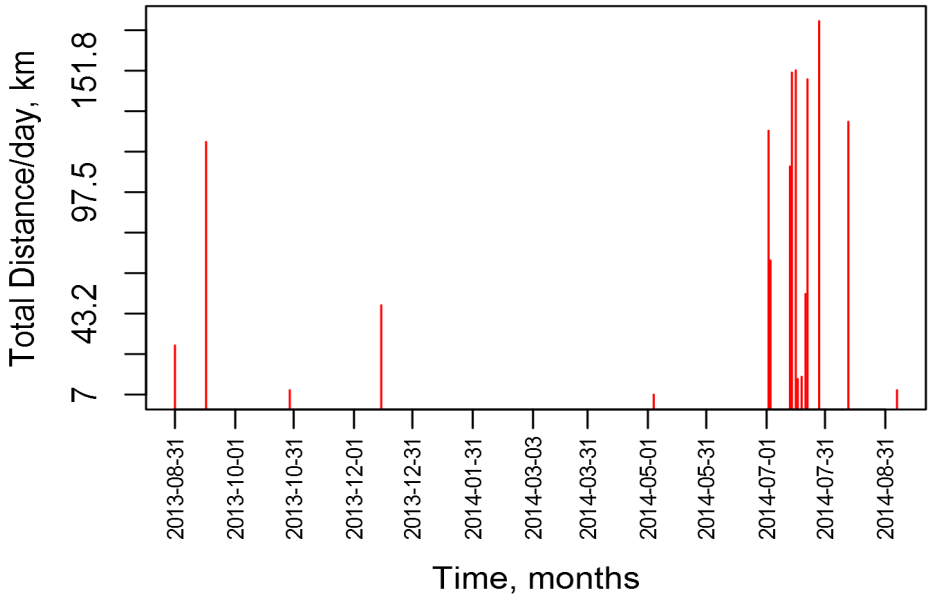

Figure 7.5: Short Distance Infrequent Traveller.

7.2 Mobility Trends

We analyse the computed cumulative distances for the short distance frequent traveller users in various different ways. We inspect different aspects and levels of time to detect if there are interesting patterns in the travelling attire of the users. The three main aspects; months of the year, days of the week and times of the day are scrutinized for predicting mobility trends among the urban users. The main reason for choosing short distance frequent travellers as our candidate data is to avoid the outliers that might occur due to time zone errors in the GPS timestamps of the photos while travelling in different time zones. The time zones of many camera devices are not changed every time with the travel, which may lead to wrong inferences in performing time dependent analysis. Each of the aspects for mobility trends are discussed in detail below.

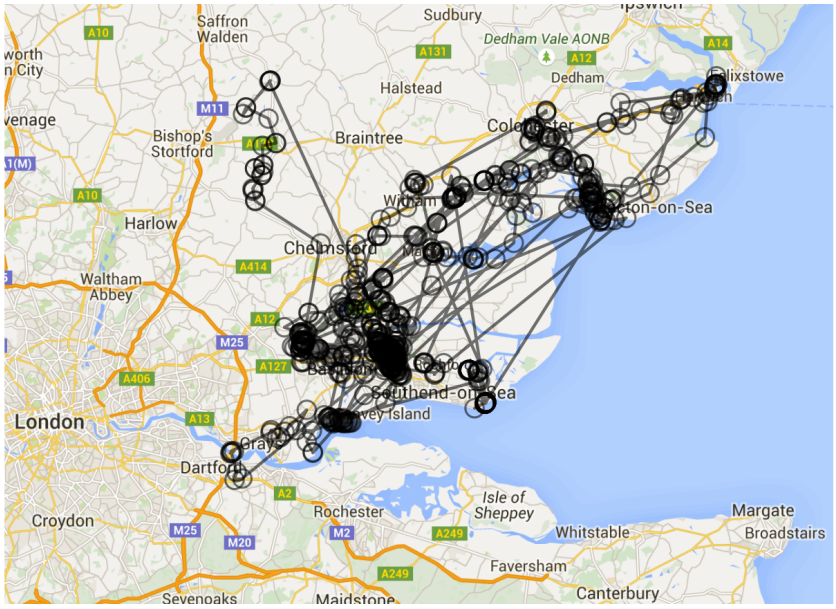

Figure 7.6: Short Distance Infrequent Traveller (movement on the map).

7.2.1 Active Months of the Year

The months of the year when Flickr users prefer to travel can be quite interesting piece of information in understanding mobility of the users. Figure 7.8 presents the box plot of distances travelled by users per month. The first glance of the graph shows that July and August are the most popular months among the chosen users in terms of high activity on Flickr. This is quite a reasonable prediction which is confirmed by the Flickr data, as most of the people around the world take summer vacations during the month of July and August, and it makes a perfect sense for them to take more pictures during that time. This suggests that the mobility patterns are more likely to be reliable during that period of time.

7.2.2 Active Days (Weekdays vs. Weekends)

In order to get an in depth analysis of mobility patterns, we separate the distances covered on each day of the week and aim to figure out whether the users' activity depends on the specific days. Figure 7.9 is the graphical representation of the percentages of the distances covered on all days. The first glance at it tells that the activity level of users stay uniform during the weekdays, however it begins to rise on Friday and almost doubles on

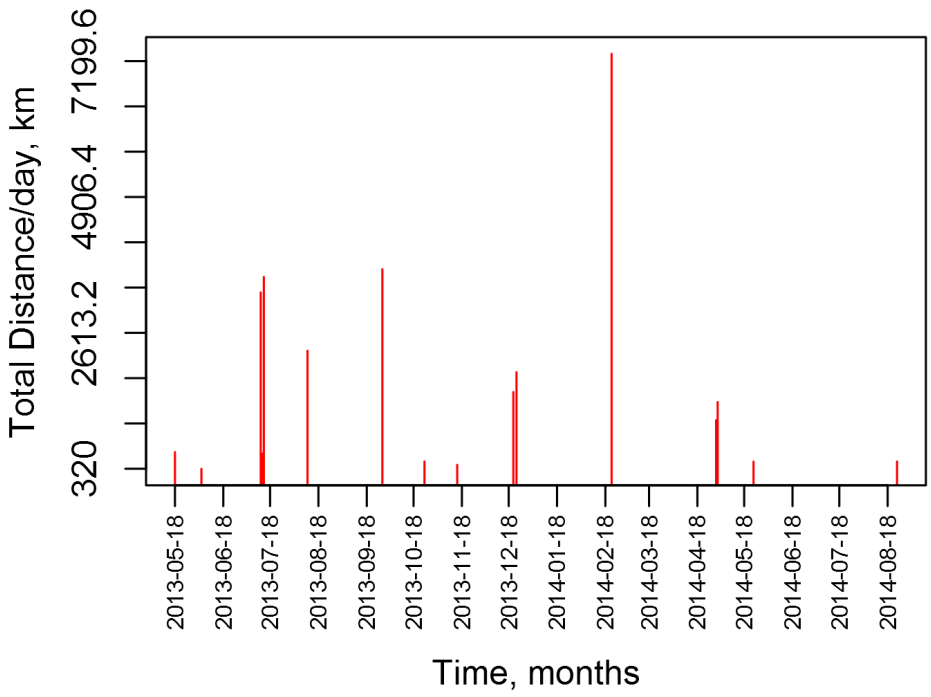

Figure 7.7: Long Distance Infrequent Traveller.

Saturday and finally comes back to normal on Sunday. We can say that the cumulative distance covered during the 5 working days of the week is slightly less than the cumulative distance covered over the weekends. We can infer that on average people take more photos on the weekend and are more mobile then. This is quite predictable as the people generally have more leisure time during the weekends and they are not much interested in taking photos during the workdays.

7.2.3 Active Times of the Day

We investigate another aspect of time in our study by analysing the active times of the day for the users. It can be quite useful information to know the times during the day when the people are travelling the most. For example, if there lies a similarity among the times of the day when people are mobile. For that purpose, we divide the day into 5 broad times; morning, noon, evening, night and other [38]. The classification according to times can be seen in Table 7.1. Finding the cumulative distances for each of the times, enables us to extract an informative pattern among the users' routines. According

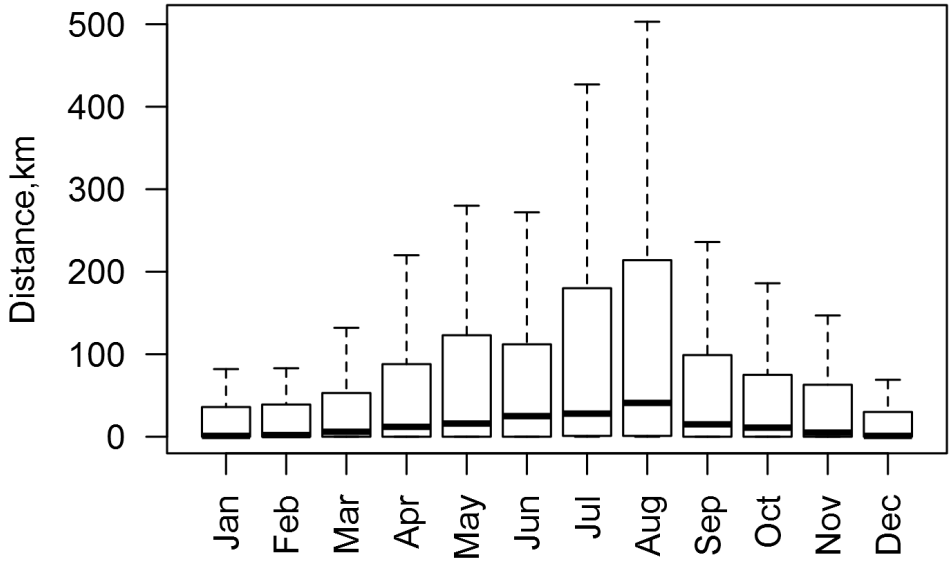

Figure 7.8: Active Months of the Year (Distance vs. Months).

Morning	Noon	Evening	Night	Other
7:00-10:59	11:00-14:59	15:00-18:59	19:00-22:59	23:00-6:59

Table 7.1: Times of the Day Classification.

to the analysis (see Figure 7.10), the mobility ramps up during the day time and evening, and ramps down in the night and finally the stationary state is achieved after mid night. The analysis from Flickr conforms the natural phenomena and work behaviour. The mornings and nights are less active in terms of mobility as the travellers might not stop to take the photos on the way to work and they might prefer to relax during the night time. Evenings are the most mobile and active times of the day, which depicts that this time is the leisure time for the users, when they travel for excursion and are most interested in capturing photos.

7.2.4 Active Hours of the Day

Extracting a pattern in the times of the day analysis prompts us to further get the in depth detail of the hours of the day. We can explore the patterns in certain hours in terms of mobility. For that we calculate the cumulative

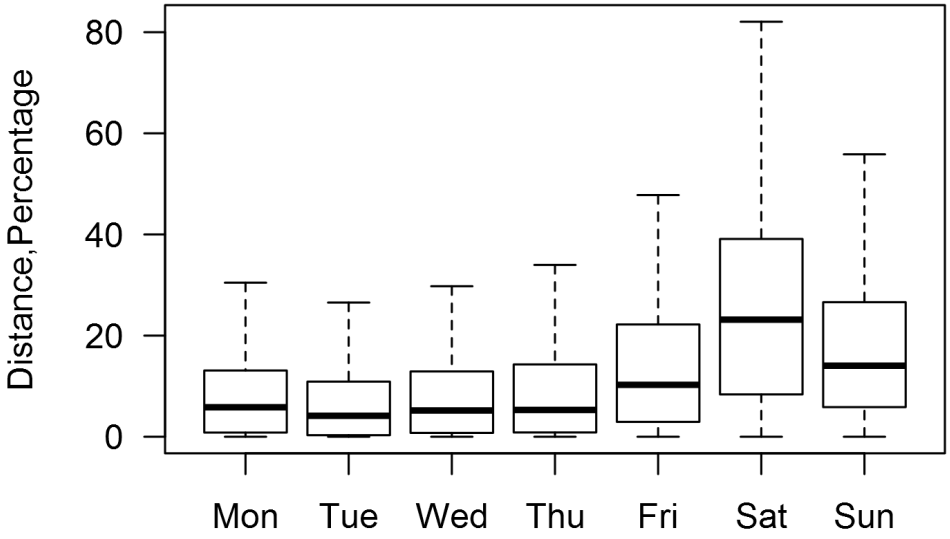

Figure 7.9: Active Days of the Week (Weekend vs. Weekdays).

distances for every hour in the day. This is further separated for weekdays and the weekends to figure out the dissimilarities among the patterns with the days. Figure 7.11 and Figure 7.12 depicts the hour of day distances for weekdays and weekends respectively. Comparing the two graphs for each hour, we can see a distinct spike difference for a few specific hours. For example the patterns in the hour 8-8:59 is extremely different for weekdays and weekends. The users cover a lot more distance in this hour during the weekdays while on the weekends this hour shows almost zero mobility. We infer from this difference that at this hour, the people tend to be resting in their homes during the weekends, and are bound to travel during the weekdays. We can see the obvious difference in the spikes of the hour 16:00 -16:59 interval, when users are very active during the day times and they do not have to follow this routine over the weekends. The active hours of the day analysis is very beneficial to predict the work routine mobility patterns for the users. This can potentially benefit the traffic engineering departments where they can spot the most active hours and ensure the safe traffic flow for the travellers.

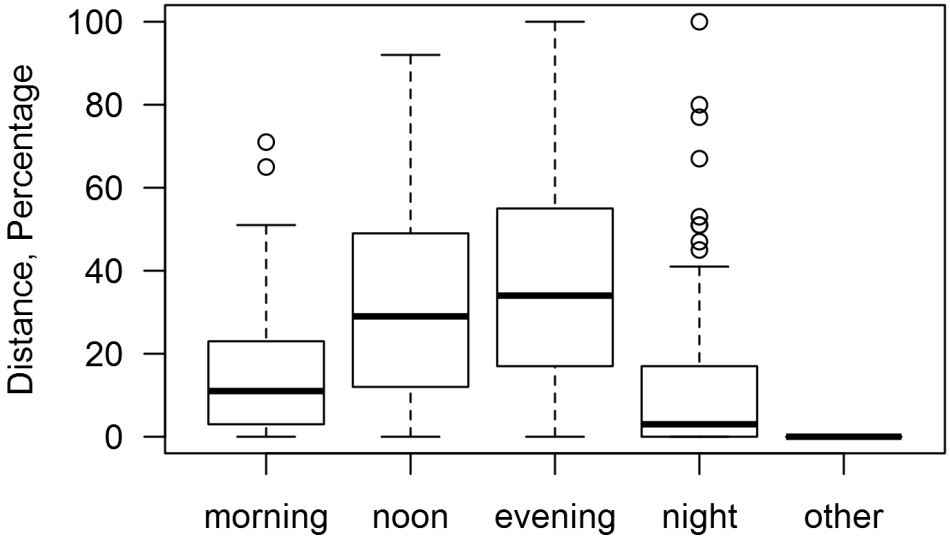

Figure 7.10: Times of the Day Distances (times vs. distance percentages).

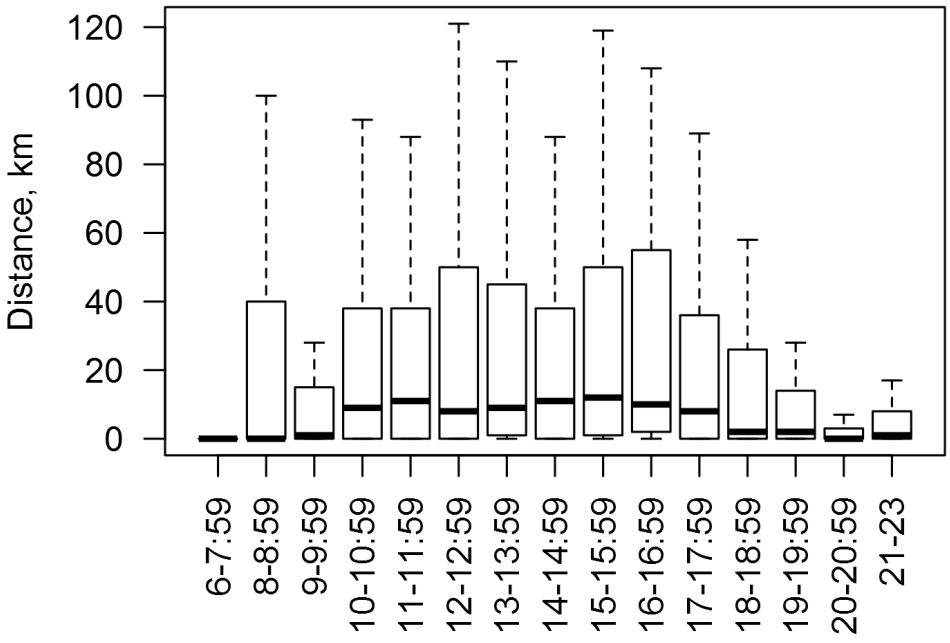

Figure 7.11: Hours of the Day Distances for Weekdays (hours vs. distances in km).

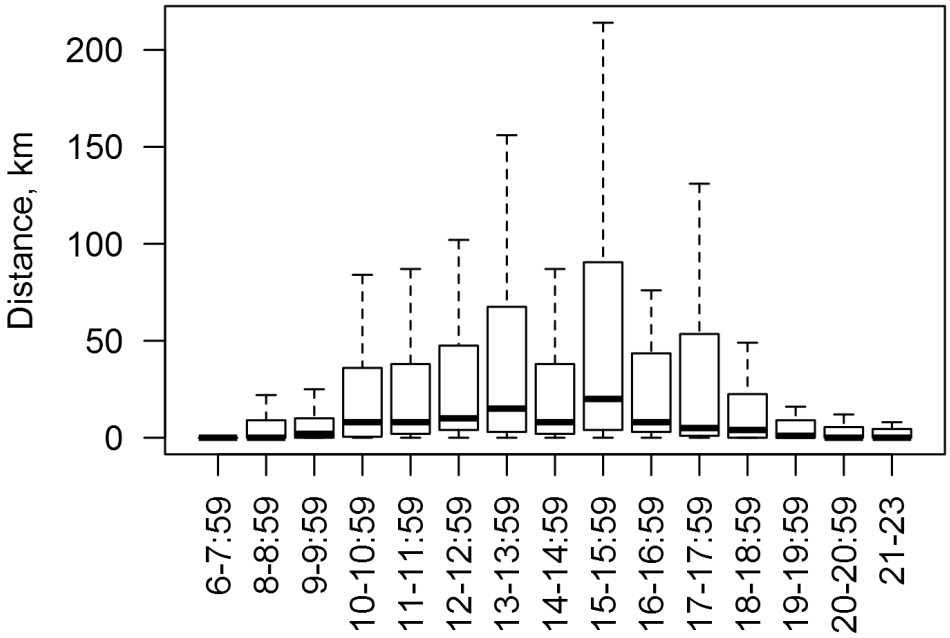

Figure 7.12: Hours of the Day Distances for Weekends (hours vs. distances in km).

Chapter 8

Conclusions

In this thesis, we used Flickr as an example to investigate if photo sharing platforms can be used to understand the mobility behaviour of people around the world. Perhaps, the main advantage is that we can work on a huge set of users' photos and picked dataset can be as global and random as we want. In other words, there is a lot of data readily available. The time range for study purposes can vary according to the research requirements. Above all, photos sharing data is open source and easily accessible. In contrast, data sets from mobile operators are usually difficult to obtain and data from GPS traces are limited in time, location, size of the dataset, mostly due to cost and privacy constraints.

During this study, we discovered that the data collected from Flickr is insufficient for analysing micro-mobility patterns, flight times and velocity prediction due to its sparse nature. Despite its shortcomings, we were able to observe some typical user mobility patterns in our Flickr data, which confirms that while not always accurate on fine-grained micro-mobility aspects, some level of user mobility can be inferred from this source. However, the rapid growth of social networking platforms and increasing awareness of location-based features among the users brings the hope that Flickr might become a standalone source for observing micro-mobility behaviours among the people around the globe in near future.

We classified active Flickr users into short distance travellers and long distance travellers and further subdivided into frequent and infrequent travellers. We observed that the short distance travellers visit similar locations back and forth, i.e. their travelling patterns are not random and they only prefer to travel to a few favourite locations. Among short distance travellers, July and August are the most active months in terms of mobility. Furthermore, the users are more mobile over the weekends because they cover more distances during the weekends. This could either indicate a user's preference

to take more photographs over the weekend giving a better mobility view or that the user is likely to take short weekend trips. The busy hour patterns for the users are distinguishably different between weekdays and weekends. The hours 8-8:59 and 16:00-16:59 on weekdays are most active whereas the activity level during these hours drops over the weekends, depicting, in our opinion, work based commuting trends amongst the users.

Obviously, Flickr is just one possible source of data. Our future work includes extending the same analysis to data collected from Panoramio and potentially other online social networks that allow fetching public data through their APIs. Comparisons could then be made between results obtained from Flickr and Panoramio. In addition, we consider learning more details about the users and their photo content to get further insight of the observed patterns. For example, examining users' habits like, which camera(s) they use (mobile phone/tablet vs. compact camera vs. DSLR). This might help us to categorize and qualify user behaviour further to differentiate between the commuter on the way to work, the professional at work, and the tourists. Photo content analysis might contribute here as well. We would also like to gain insight, e.g., on the number and visiting frequency of favourite spots for users of a certain class. This could assist in confirming and extending mobility models (for example, community mobility models) but also to infer behavioural context from types of places visited, and to classify users further. Finally, we are curious if the proliferation of better quality cameras in mobile phones and the continued popularity growth of photo sharing services (and maybe apps that foster virtually continuous uploading) will lead to denser data sets and thus, ultimately, allowing us to infer closer to microscopic mobility characteristics for the users after all.

Bibliography

[1] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless Communications & Moblie Computing (WCMC): Special Issue on Mobile Ad Hoc Networking: Research, Trends and Applications*, vol. 2, pp. 483–502, 2002.

[2] A. Gonga, O. Landsiedel, and M. Johansson, "MobiSense:Power-efficient micro-mobility in wireless sensor networks," in *Distributed Computing in Sensor Systems and Workshops (DCOSS)*, 2011.

[3] M. C. González1, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, 2008.

[4] V. A. Paz-Soldan, J. Robert C. Reiner, A. C. Morrison, S. T. Stoddard, and U. Kitron, "Strengths and weaknesses of global positioning system (GPS) data-loggers and semi-structured interviews for capturing fine-scale human mobility: findings from Iquitos, Peru," *PLoS Neglected Tropical Diseases*, vol. 8, p. 6, 2014.

[5] J. Y. Jang, K. Han, P. C. Shih, and D. Lee, "Generation like: comparative characteristics in Instagram," in *International Conference on Human Factors in Computing Systems (CHI)*, 2015.

[6] G. Chareyron, J. Da-Rugna, and T. Raimbault, "Big data: a new challenge for tourism," in *IEEE International Conference on Big Data*, 2014.

[7] J. Dong, Z. Ou, and A. Ylä-Jääski, "Utilizing internet photos for indoor mapping and localization -opportunities and challenges," in *SmartCity 2015 (Workshop of Infocom)*, 2015.

[8] Flickr Statistics. [Online]. Available: https://www.flickr.com/photos/franckmichel/6855169886

[9] D. Zielstra and H. H. Hochmair, "Positional accuracy analysis of flickr and panoramio images for selected world regions," *Spatial Science (Impact Factor: 0.41)*, vol. 58, p. 2, 2013.

[10] V. Palchykov, M. Mitrovic, H.-H. Jo, and J. S. andRaj Kumar Pan, "Inferring human mobility using communication patterns," Nature Publishing Group, Tech. Rep., 2012.

[11] Gps. [Online]. Available: http://www.pdvwireless.com/analyzing-the-benefits-of-gps-technology-within-the-commercial-environment/

[12] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, pp. 462–465, 2005.

[13] Y. Zheng, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *WWW 2009*. Association for Computing Machinery, Inc., 2009.

[14] ——, "Mining correlation between locations using human location history," in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2009.

[15] E. Herder and P. Siehndel, "Daily and Weekly Patterns in Human Mobility," 2012. [Online]. Available: http://ceur-ws.org/Vol-872/aum2012_paper_3.pdf

[16] S. Kisilevich, F. Mansmann, and D. Keim, "P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos," in *Proceeding COM.Geo '10 Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application*, 2010.

[17] I. Lee, G. Cai, and K. Lee, "Mining points-of-interest association rules from geo-tagged photos," in *46th Hawaii International Conference on System Sciences (HICSS)*, 2013, pp. 1580 – 1588.

[18] L. Kennedy, M. Naaman, S. Ahern, and R. Nair, "How flickr helps us make sense of the world: context and content in community-contributed media collections," in *MULTIMEDIA '07 Proceedings of the 15th international conference on Multimedia*, 2007.

[19] T. Rattenbury, N. Good, and M. Naaman, "Towards extracting flickr tag semantics." in *16th International Conference on World Wide Web*, 2007, pp. 1287–1288.

[20] Y.-T. Zheng, Z.-J. Zha, and T.-S. Chua, "Mining travel patterns from geotagged photos." *ACM Transactions on Intelligent Systems and Technology (TIST)* , vol. 3, no. 56, p. 3, 2012.

[21] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from flickr tags," in *SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 103–110.

[22] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proceeding WWW '09 Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 761–770.

[23] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura, "Travel route recommendation using geotags in photo sharing sites," in *Proceeding CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 579–588.

[24] X. Lu, C. Wang, J.-M. Yang, and Y. Pang, "Photo2Trip: generating travel routes from geo-tagged photos for trip planning," in *Proceeding MM '10 Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 143–152.

[25] K. Okuyama and K. Yanai, "A travel planning system based on travel trajectories extracted from a large number of geotagged photos on the web," in *Pacific-Rim Conference on Multimedia*, 2011.

[26] Y. Shi, P. Serdyukov, A. Hanjalic1, and M. Larson, "Personalized landmark recommendation based on geotags from photo sharing sites," in *ICWSM '11: the 5th AAAI Conference on Weblogs and Social Media*. AAAI, 2011, pp. 622–625.

[27] L. Hollenstein and R. S. Purves, "Exploring place through user-generated content: Using Flickr tags to describe city cores," *Journal Of Spatial Information Science*, vol. 1, pp. 21–48, 2010.

[28] N. Hochman and R. Schwartz, "Visualizing instagram:tracing cultural visual rhythms," The Workshop on Social Media Visualization (SocMedVis) in conjunction with The Sixth International AAAI Conference on Weblogs and Social Media (ICWSM-12), Tech. Rep., 2012.

[29] Y. Hu, L. Manikonda, and S. Kambhampati, "What we instagram: a first analysis of instagram photo content and user types," in *Eighth International AAAI Conference on Weblogs and Social Media*. AAAI Publications, 2014.

[30] S. Bakhshi, D. A. Shamma, and E. Gilbert, "Faces engage us: photos with faces attract more likes and comments on Instagram," in *Proceeding CHI '14 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 965–974.

[31] Flickr Guide. [Online]. Available: http://mashupguide.net/1.0a/858Xch06.pdf

[32] Flickr API. [Online]. Available: https://www.flickr.com/services/api/flickr.photos.getExif.html

[33] FlickrAPI. [Online]. Available: https://www.flickr.com/services/api/auth.oauth.html

[34] Flickr API query limit. [Online]. Available: https://www.flickr.com/services/developer/api/

[35] M. Cha, Alan, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the Flickr social network," in *Proceeding WWW '09 Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 721–730.

[36] S. Kisilevich, M. Krstajic, D. Keim, N. Andrienko, and G. Andrienko, "Event-based analysis of people's activities and behavior using Flickr and Panoramio geotagged photo collections," in *14th International Conference Information Visualisation*, 2010.

[37] Haversine Formula. [Online]. Available: http://en.wikipedia.org/wiki/Haversine_formula

[38] NIST, Physical Measurement Laboratory. [Online]. Available: http://www.nist.gov/pml/div688/times.cfm

[39] "Wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/Empirical_distribution_function

[40] "EcdfBlog." [Online]. Available: http://www.r-bloggers.com/

[41] Techtarget. http://whatis.techtarget.com/definition/latitude-and-longitude.

[42] Y. Dodge, *The Oxford Dictionary of Statistical Terms*. OUP, 2006.

[43] "Statistical Analysis." [Online]. Available: https://en.wikipedia.org/wiki/Statistics

[44] D. J. Rumsey, *Statistics For Dummies*, 2nd ed., 2011.

[45] I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong, "Human mobility patterns and their impact on routing in human-driven mobile networks," in *Sixth Workshop on Hot Topics in Networks (HotNets-VI)*, 2007.

Appendix A

First appendix

A.1 Empirical Distribution Function

ECDF (empirical distribution function) gives the empirical measure of the sample data in statistical analysis. " CDF (cumulative distribution function) is a step function that jumps up by 1/n at each of the n data points " [39]. An ECDF gives non-parametric estimate of the CDF of a random variable. This is done by assigning 1/n probability to each data point, by ordering the data in ascending order of their value, and finding the sum of probabilities assigned to each datum (data point) [40] (see Figure A.1). The ECDF has following advantages:

- For a large data set, it gives an accurate approximation of the true CDF which is helpful in statistical inferences [40].
- it can show how fast the CDF approaches to 1, by giving a visual display; the get and feel for the data are possible to attain by plotting key quantiles like the quartiles [40].
- the comparison of ECDF to well-known CDFs of commonly used distributions can help finding out if the data can be approximated by any of those common distributions [40].

A.2 Longitude and Latitude

A point on earth surface (or on sphere) can be defined in terms of longitude and latitude coordinates. In other words, the unique latitude/longitude combination is the global address assigned to each location on the sphere. The

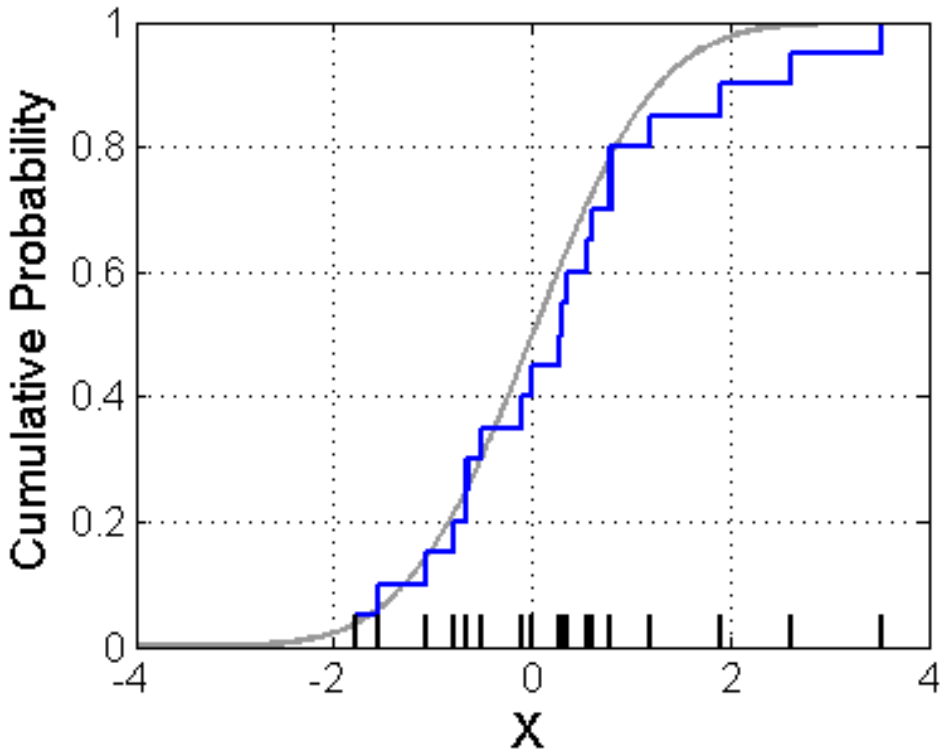

Figure A.1: The ECDF is represented by blue steps. Vertical black bars are the samples corresponding to the ecdf. Smooth gray curve represents the true CDF function [39].

numbering system in Lat/Lon differs from the usual street address numbering system in a way that it works on the grid numbering system and has intersecting horizontal and vertical lines. The address of a particular location is the numbers or coordinates (horizontal and vertical) which intersect at that point. However, longitude and latitude of earth are not exactly straight lines, but horizontal or vertical half circles encircling around the earth.

The distance calculation between two points lying on the earth surface or sphere need the longitude and latitude information, used by the Haversine formula in our thesis.

A.2.1 Latitude

In Figure A.2, the latitiude of a point P lying on the surface of the sphere is an angle formed by a straight line passing through the centre of the sphere

C and P, subtending with reference to the equatorial plane. The latitude is positive if the point P lies above the equatorial plane (towards North Pole), and it is negative if the P lies below the plane or towards the South Pole. The range of the Latitude angles goes from +90 to -90 degrees [41].

A.2.2 Longitude

Meridian, shown in Figure A.2, are the half circles encircling the sphere from pole to pole. Longitude is defined by choosing a reference meridian, called the prime meridian. The Longitude of a point P lying on the surface is an angle subtended by the meridian crossing P with respect to the prime meridian. Longitude of P is positive if it lies towards the East of the reference prime meridian and negative on its West side. The range of Longitude goes from +180 to -180 degrees.

A.3 Statistical Analysis

"Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data" [42]. The purpose is to discover the underlying patterns and trends in the data. Statistics can be applied to various domains such as science, industry, government or business in order to improve the previously built models or make more logical decisions based on the facts obtained from the data. Statistical population is the starting point of applying statistics where the statistical population can refer to different topics such as some specie of animals in a particular region or all the members of a multinational company etc. [42].

It is not always possible to gather all the data, therefore the statisticians obtain the data by survey sampling or specific experiments. The sampled data collected guarantees that the statistical decisions or inferences made about the data can be generalized to the whole or actual set of population. Statistics include experimental study and observational study. Taking actual measurements and manipulations in the system under observation are the basics of experimental study while observational study does not require the actual measurements and manipulations with the data.

A.3.1 Statistical Methodologies

The data analysis comprises of two broad statistical methodologies; descriptive statistics and inferential statistics. Both of them are defined below:

- *Descriptive Statistics:* This gives summary of data by using sample indexes; for example standard deviation or mean [43].

- *Inferential Statistics:* This summarizes data with random variations like observational errors and sampling variations [43].

A.3.2 Statistical Graphs

- *Bar Chart:* The bar chart represents quantitative data by displaying horizontal or vertical bars on a grid. It does so by breaking down the data on the basis of groups and the length of the bars tell about the amount of data in each group [44].

  Figure A.3 illustrates an example bar chart. This depicts information about the amount of money people spend on their transportation while going to work and coming back home. The data is categorized on the basis of house hold income groups marked on x-axis. It can be inferred from this graph that more the house hold income is higher the amount of money spent on transportation [44].

- *Histogram:* Histogram provides a bigger picture or shape of the statistical data. It is applied by breaking down the data into numerically ordered groups. It is different from bar chart in a way that the data is ordered and bars are joined, as opposed to the bar chart in which the bars are separated and represent categories without following any order. The histogram does not provide the actual values of the data, rather it just tells about which group the data value belongs to. The height of the bar tells about either the count or frequency of individuals in each group or the relative frequency, which is the percentage of individuals in each group.

  Figure A.4 is an example of a histogram. The data is relates to the age of the actresses at the time they won the 'best actress award' in Oscars. The x-axis show the age groups of 5 years each increasing from left to right. The y-axis is the percentage or relative frequency of the actresses having certain age group. By enlarge, we can infer that about 27% of the actresses were at the age of 30 to 35 when they won the best actress award [44].

- *Line Graph:* It is also known as time chart, which depicts the trends of data appearing in a period of time. The x-axis represents the time data, such as, months, years or days etc. The y-axis represents the

values of the variables under consideration, such as population size or total sales etc [44].

Figure A.5 is an example of a line graph. The data is the same as that used for histogram in the description above. However, it is the time series data, with years on x-axis and age on y-axis. The up and down cyclic trends can be seen from left to right.

- *Pie Chart:* It is a circle in shape with slices, each represents a distinct group of the categorical data, showing the percentage of indiviuals that fall in each group. As the sum of the percentages must be 100, this means that one individual can only belong to a single group. Pie charts provides easy comparison and contrast of data [44].

  Figure A.6 is an example of a pie chart. The data is about the population in U.S above 65 years of age. The age groups are 65-69 years, 70-74 years and so on. One pie chart represents the percentage of people in each group calculated in 2010, while its projection for 2050 is shown in the 2nd pie chart. The later is calculated on the basis of multiple factors such as birth rate and death rate etc [44]. In this case, it is quite simple to make comparisons for each group in two graph and understanding the shift in trends from 2010 to 2050.

- *Box Plot:* " A boxplot can give you information regarding the shape, variability, and center (or median) of a statistical data set. It is particularly useful for displaying skewed data " [44].

  Box plot tells about the symmetry of the data, meaning both sides are similar if a box is cut down the middle. The data is symmetric if the median of the data lies in the middle of the box and skewed otherwise. Figure A.7 is an example of box plot representing skewness in data. The data is skewed right because the longer part of the box is on the right of the median. This shows that the ages of the actresses are skewed right. The part of the box plot towards the left of the median showing young actresses is short as compared to the part on the right of the median showing old actresses. This infers that the ages on left are closer to each other than ages on right [44].

  The left and right side of the median does not tell anything about the size of the data, rather the box plot is based upon the percentages of sample size and not the sample size itself [44].

A.4 Micro-mobility

Micro-mobility, in the context of our thesis, can be thought of as a fine scale movement measurements gathered to reveal how users move. It can be explained with an example study, performed by Rhee et al. [45]. In this study, they record the traces of human walks captured by GPS devices for the period of 1000 hours at five different locations. They introduce a Lévy walk mobility model that predicts the walking behaviour of humans in outdoor mobile network. This model can be useful, e.g., in mobile network simulations. In practice, the mobility is simulated by computing the flights, pause times, velocity and direction from the measured data. The flight is defined as the straight line path travelled by an individual from one point to another without pausing or changing direction. The pause time is the time period during which the individual has travelled less than a predefined threshold distance [45].

If we get such fine grained data from the online social websites, then we can replace the costly GPS data gathering procedure with the free of cost online websites data when analysing human mobility patterns.

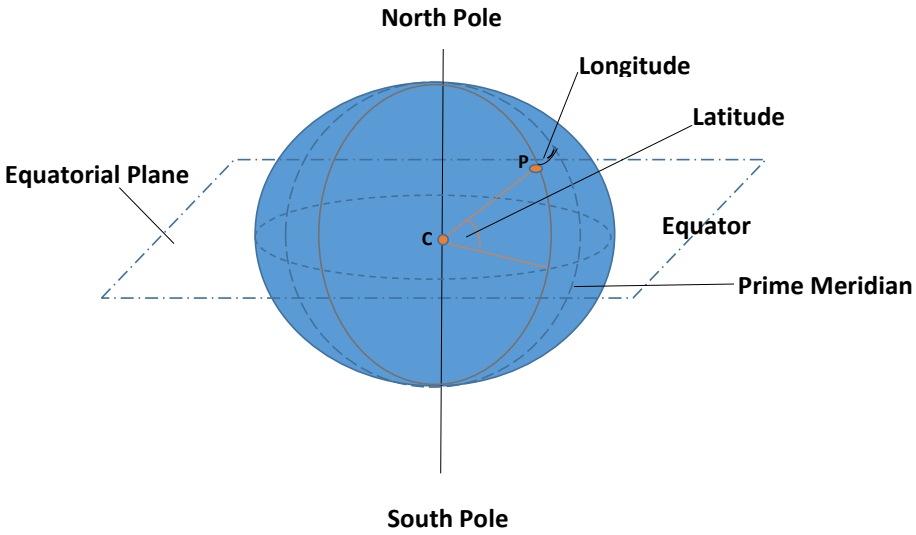

Figure A.2: Latitude and Longitude [41].

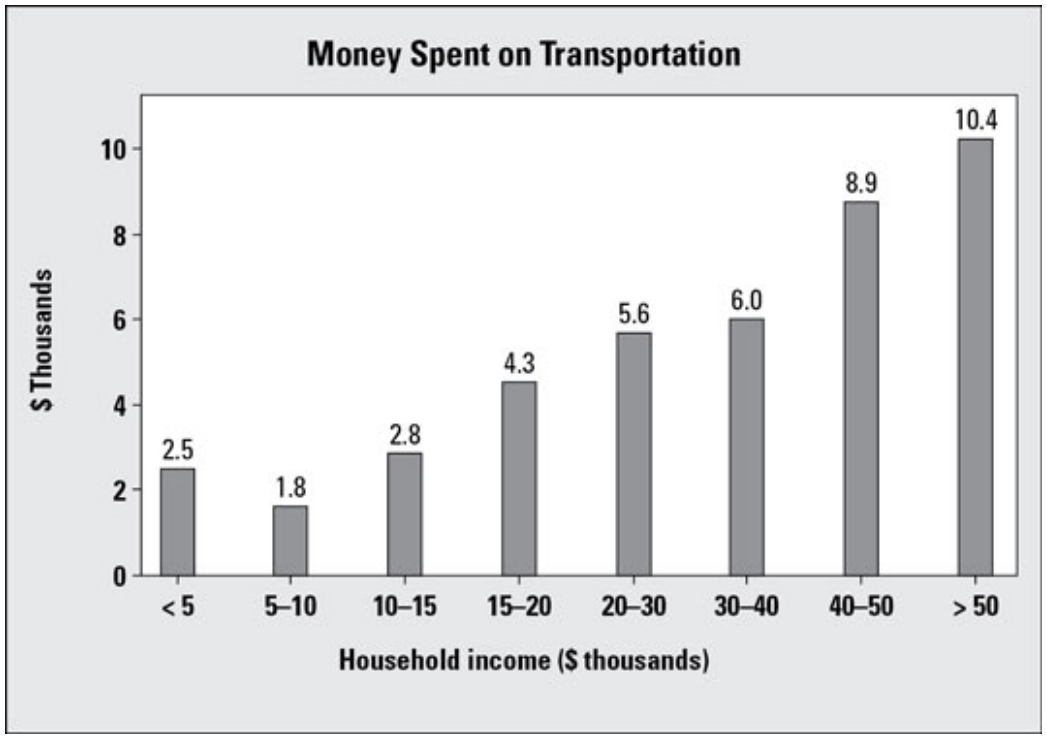

Figure A.3: Bar Chart Example [44].

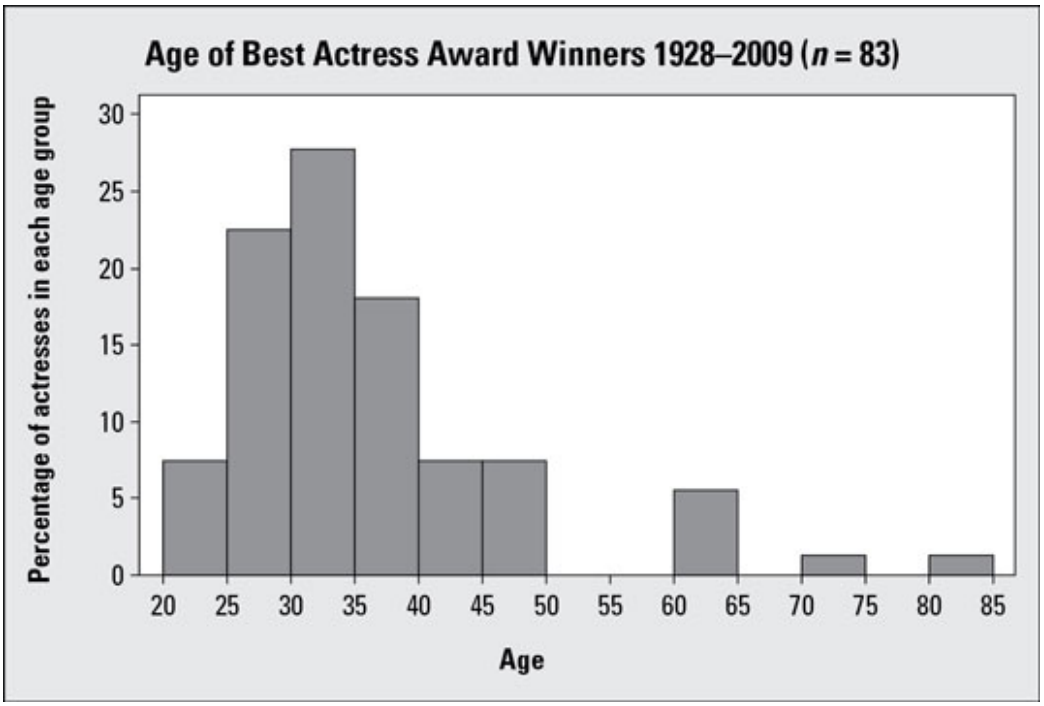

Figure A.4: Histogram Example [44].

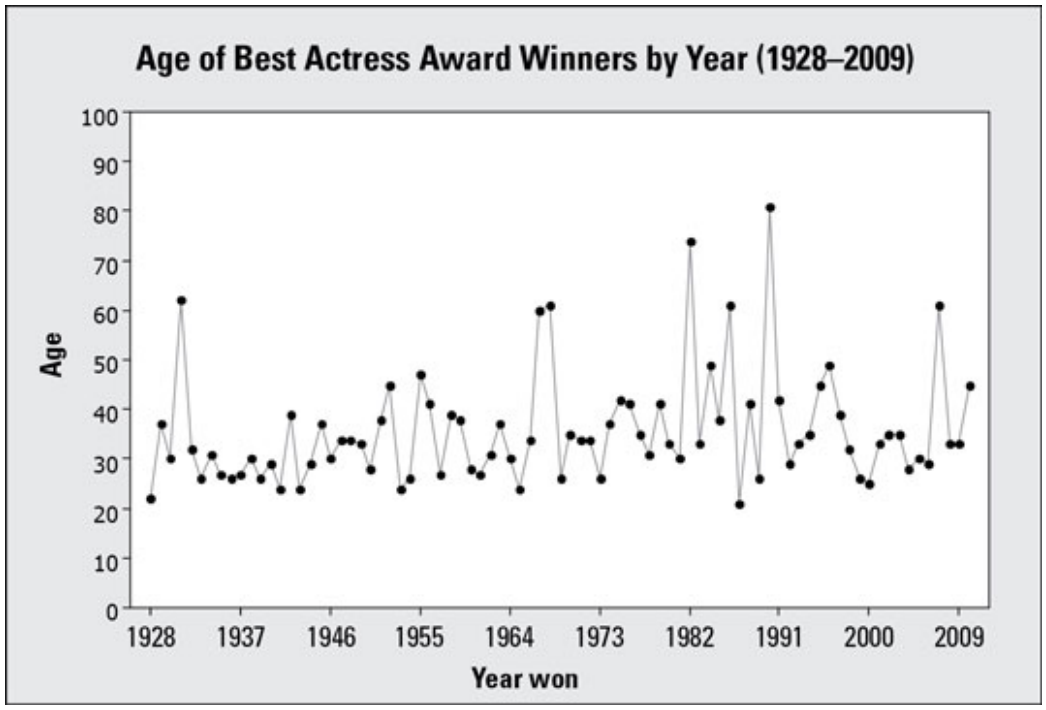

Figure A.5: Line Graph Example [44].

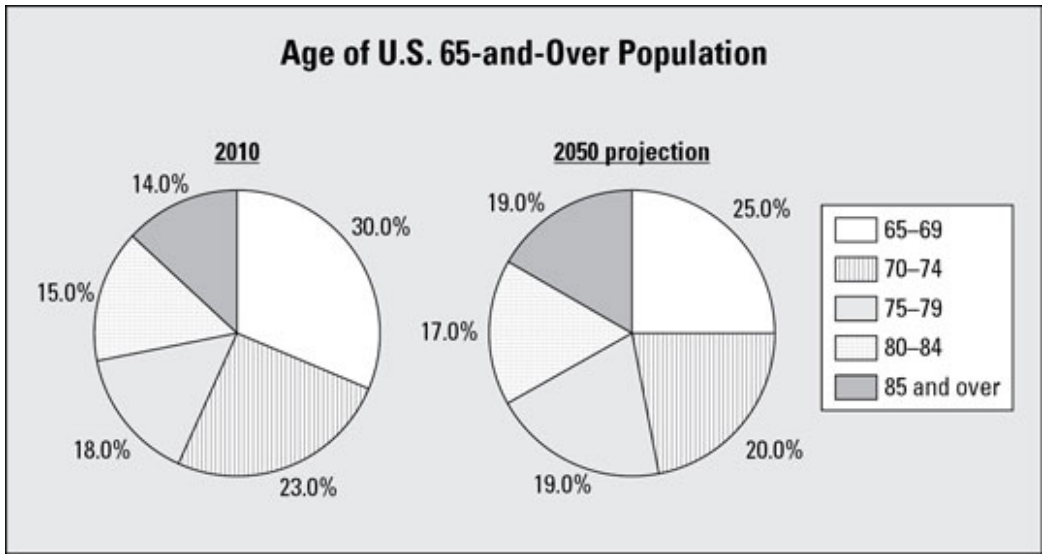

Figure A.6: Pie Chart Example [44].

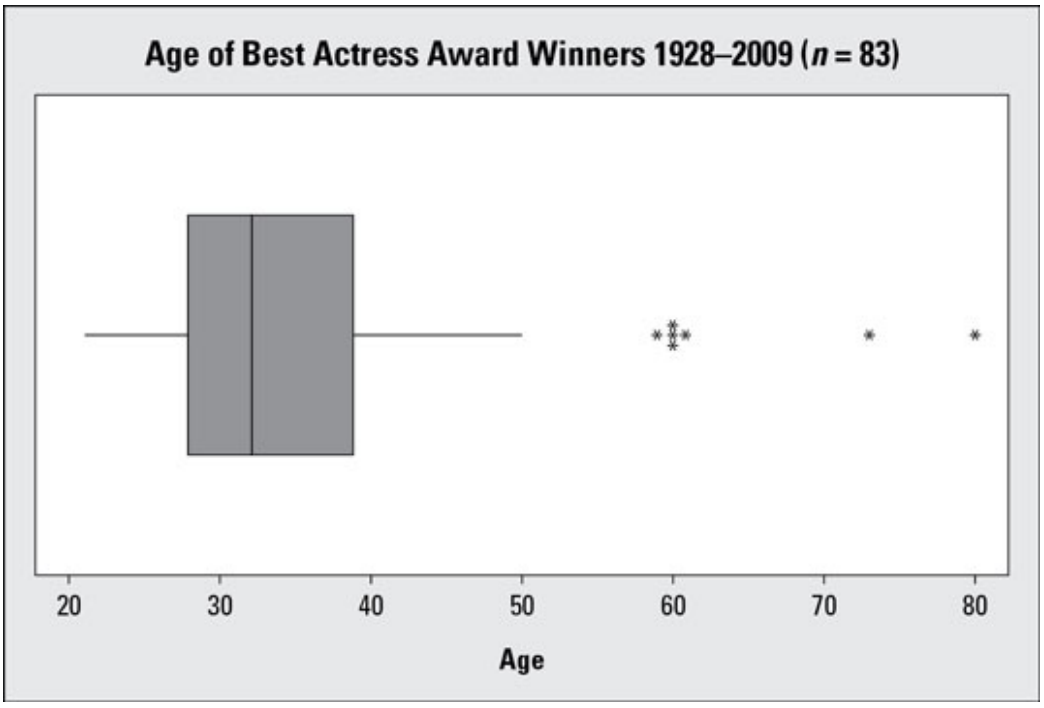

Figure A.7: Box Plot Example [44].