Project Report (Major Project- II)

on

Collecting Mobility Pattern of Humans from Mobile Phone Data

Submitted in partial fulfillment of the requirements

for the award of the degree of

Master of Technology

in

Software Technology

By

Rinki Dahiya

Roll No.: - 2K16/SWT/515

Under the guidance of

Dr. Ruchika Malhotra

Associate Professor

Department of Computer Science & Engineering

Delhi Technological University

(Formerly Delhi College of Engineering)

Bawana Road, Delhi 110042

2019

Delhi Technological University

(Formerly Delhi College of Engineering)

Bawana Road, New Delhi-42

DECLARATION

I hereby declare that the thesis entitled "**Collecting Mobility Pattern of Humans from Mobile Phone Data**" which is being submitted to the Delhi Technological University, in partial fulfillment of the requirements for the award of the degree of Master of Technology in Software Technology is an authentic work carried out by me. The material contained in this thesis has not been submitted to any university or institution for the award of any degree.

DATE:

SIGNATURE:

RINKI DAHIYA

2K16/SWT/515

CERTIFICATE

Delhi Technological University

(Formerly Delhi College of Engineering)

Bawana Road, New Delhi-42

This is to certify that project report entitled "**Collecting Mobility Pattern of Humans from Mobile Phone Data**" done by me for the Major Project 2 for the award of degree of Master of Technology Degree in Software Technology in the Department of Computer Science & Engineering, Delhi Technological University, New Delhi is an authentic work carried out by me.

Signature:

Student Name

Rinki Dahiya

2K16/SWT/515

Above Statement given by Student is Correct.

Project Guide:

Dr. Ruchika Malhotra

Associate Professor

Department of Computer Science & Engineering

Delhi Technological University, Delhi

Acknowledgement

No volume of words is enough to express my gratitude towards my guide **Dr. Ruchika Malhotra**, Department of Computer Science & Engineering, Delhi Technological University, Delhi, who has been very concerned and has aided for all the materials essentials for the preparation of this project report. She has helped me to explore this vast topic in an organized manner and provided me all the ideas on how to work towards a research-oriented venture.

I am also thankful to **Dr. Rajni Jindal**, HoD of Computer Science & Engineering Department and **Dr. Ruchika Malhotra** , Coordinator , for the motivation and inspiration that triggered me for the project work.

I would also like to thank the staff members and my colleagues who were always there at the need of hour and provided with all the help and facilities, which I required, for the completion of my project work.

Most importantly, I would like to thank my parents and the almighty for showing me the right direction, to help me stay calm in the oddest of the times and keep moving even at times when there was no hope.

Rinki Dahiya

(2K16/SWT/515)

ABSTRACT

In recent days, there has been a rapid increase in cell phone networks. Call Record Data which contains information for a number of users. These mobile data details are used in many important studies such as analyzing mobility pattern, daily activities (physical activities and sleep), size of social groups, call and text messages, etc. Studying human activities has always been a major focus of many researchers. As the society is evolving and more and more technologies are introducing, it is getting quite difficult and complex to understand the cities as compared to before. Pattern of mobility depends on the time a person spends on any particular location and the frequency with which that location is visited. In this paper we have classified individuals in 6 particular categories and finding out the population of a particular area based on the dataset.

Table of Contents

List of Figures

CHAPTER 1

INTRODUCTION

With the increase in the usage of mobile phones, the way people communicate has changed a lot. Mobile phone coverage has increased from 12 % in the year 2000 to 96% in the year 2014 and is increasing till now [1][2]. Since 2005, there has been a decrease in the usage of landline phones and it has connected people present in the remote areas of almost every developed and developing nation.

Previously mobile data records were used just to generate billing information but now a days with the change in technology and various data mining tools and methods these records can be used to generate various other information which could help us in analyzing mobility pattern, daily activities, size of social groups, call and text messages, etc. Wireless connectivity has changed the way people used to talk, play and work, phone data can be utilized to find the spatio temporal information of unidentified phone user whereabouts for finding their mobility patterns [3][4][5][6].

In this paper, our aim is to analyze the data to determine the mobility pattern of human beings based on the call data records in a particular area. The dataset must contain three important values:

1. Point of origin and destination (longitude and latitude)
2. Time for which a person is staying at a particular location
3. Frequency of the person’s movement from the territory.

Finally, we divide the dataset in 6 different groups for better analysis. Then, we present an approach to find the weekly pattern of mobility. We propose 6 weekly patterns observed by the mobility and compare this data with national surveys and census in order to validate our results.

1.1 Existing System

The use of cellular information other than client's billing information was to find location of the user and to find the estimate about the road traffic so that proper infrastructures on road can be made. The comparatively easy access to cellular data attracted many other studies, which were using the location of the user. So, in different ways this cellular network knowledge has been used by different researchers. Amongst them we tend to notice; studies of mobility pattern, is used on a large scale. Recent technologies work on finding and studying using different approaches so that the limitation of every single model can be avoided.

1.2 Proposed System

Cellular network is a network for communication where the last connection in the network is wireless. The network is spreaded over different areas which are known as cells. Each cell is served by a transreceiver or base station which provide network coverage which can be used to transmit call, message, data or any other type of content. These cells take form of hexagon. Each cell has a frequency which may range from f1 ~ f6. This frequency can be reused by other cells except the adjacent ones which may cause co-channel interference.

Various operators store call data in different formats. For mobile billing purpose these call data records are maintained so as to provide correct bill to the customer. Stored data includes the location, identification, duration of the call, type of call and other services used.

When any mobile device is used for any purpose for eg: call, message , data, etc) various operators store the information regarding that particular device. This information includes the identifying code, location, day, hour, type of services used by the user, duration of the call initiated or received, type of event on device, etc. These details are referred to as Call Data Records (CDR). We can say that each of the record contains a minimum of three basic info that are:

1. IMEI of the user, which would be unique to every user and help to identify the movement
2. Time and location of various services used by user which would help in knowing the location of user at any particular time.
3. Cell ID for the identification of network code

These data collectively will help to trace the location of any user at any particular time, which would then help us to identify the mobility pattern.

CHAPTER 2

USER PROFILING METHODOLOGY

2.1 User Profile and Profile Handling Algorithm

For our research, we have used a dataset which contains call data records of a particular area. We denote our area by S1. Now he dataset is divided into 6 categories based on this area S1.

Each Call Data Record in our research consists of below information:

< IMEI No, Incident Date, Day of Week, Month , Year, lat_rand, long_rand, Spec_Pop >

Now we define the various user profiles based on the dataset that we have selected. Users can be divided into following categories:

1. Commuters: These are the people who are working or studying in area S1 but are staying or living outside S1
2. One Single Transit: These type of people arrive in area S1 just once and are not seen more than a time period of 1 hour
3. Multiple Transit: These people are seen multiple times entering the area S1 but not at a regular interval of time. Their total time period within the area S1 exceeds 1 hour.
4. Residents working in zone: These are the people who are living, staying, sleeping, working in the area S1
5. Residents working out of zone: These type of people works out of the territory S1 during the daytime but sleeps and lives in area S1
6. Weekend: These type of people show their entry in the area S1 only in the weekends i.e Saturday and Sunday.

Each of these categories is important for our research because it will help us in knowing the actual population of our area S1 and can help us in knowing the actual residents of area S1. This can help in building infrastructures like schools, parks, malls, markets etc. for the people residing in the area S1.

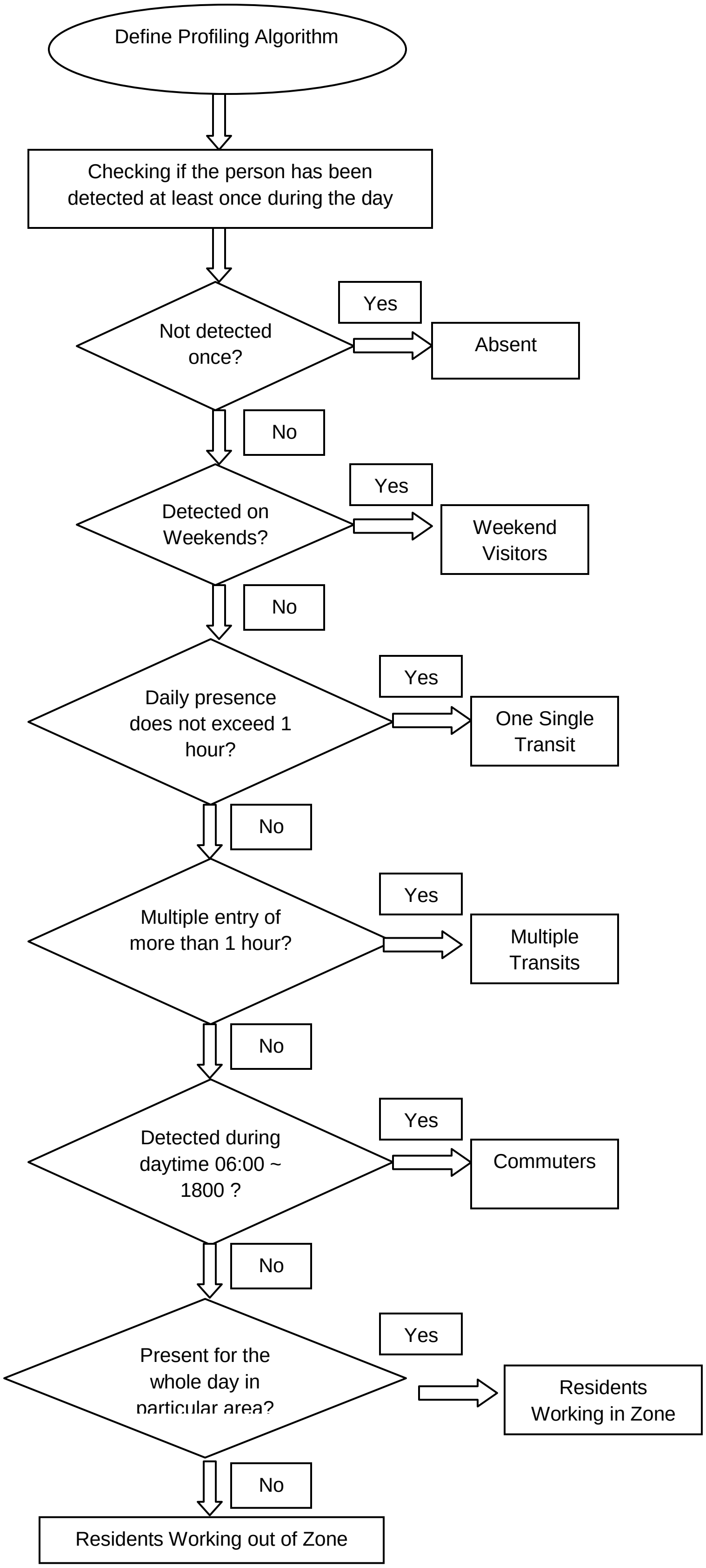

Define Profiling Algorithm
Checking if the person has been detected at least once during the day
Not detected once?
Yes
Absent
No
Detected on Weekends?
Yes
Weekend Visitors
No
Daily presence does not exceed 1 hour?
Yes
One Single Transit
No
Multiple entry of more than 1 hour?
Yes
Multiple Transits
No
Detected during daytime 06:00 ~ 1800 ?
Yes
Commuters
No
Present for the whole day in particular area?
Yes
Residents Working in Zone
No
Residents Working out of Zone

Figure 2.1 Algorithm for Profiling

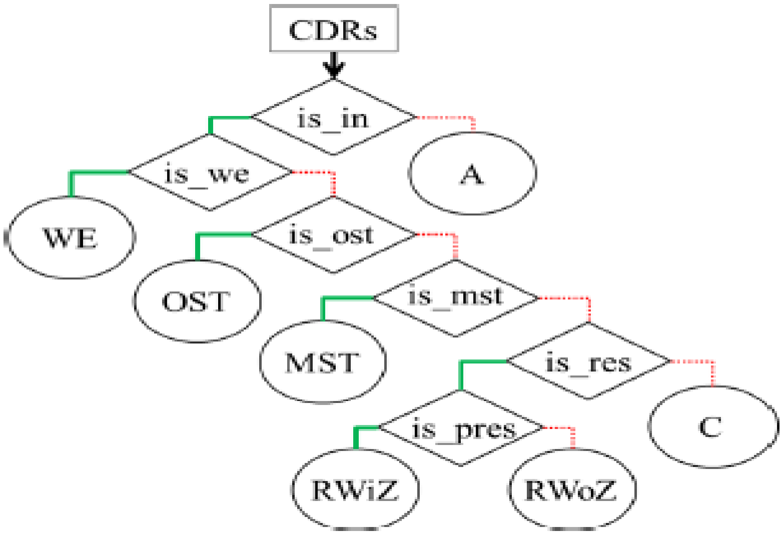


Figure 2.2: The Profiling Algorithm

Profiling algorithm as described in Figure 2.2 is used on every IMEI that is present in Call Data Record. Figure 2.1 shows our categorization process. Each diamond represents a Yes/No Condition. If the condition is satisfied then the flow will follow the bold line else it will follow the dotted line.

We track the events done by a particular IMEI in a week and record them in an order which would help us tracking the particular IMEI in a particular category. There is a possibility where in a particular IMEI gets detected in one week and then it is not present, for such cases the dataset should be deleted from record and must not be included.

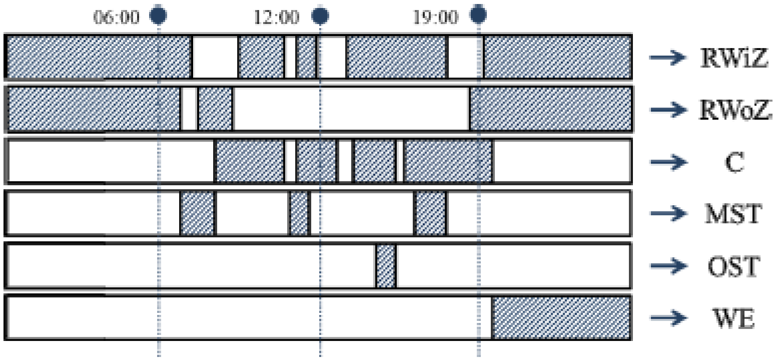


Figure 2.3 :Examples of profiling

2.2 Challenges of using Call Data Records

Call Data Records are highly dependent on users who are using their mobile device. If the device is not used frequently by the user then the output predicted will not be useful for us [7][8]. If the resident of the particular location is not using mobile device then it is not able to find the mobility of that person and the calculation in finding the population estimate of that particular area may go incorrect [9][10][11].

Many alternatives have been provided for handling these challenges. Despite of such disadvantages, still Call Data Records can be used to predict the mobility pattern since using mobile data network do not require extra hardware requirements unlike other technologies like Bluetooth sensors, Road Cameras, Road magnetic sensors, etc.

2.3 Event categories

A Call Data Record not only presents the location i.e the point of origination and destination, time for which a person is staying at a particular location but also presents the interaction between the operator (person is using) and the handset or the IMEI. There can be various types of interaction between the operator and the IMEI; some of them are as mentioned below:

- Transmitting or receiving a phone call
- Transmitting or receiving as text message
- Handovers during call
- Area or location of information receiving
- Events induced after every 3 minutes of call
- Events induced after every 3 hours of call

CHAPTER 3

SYSTEM DESIGN AND ARCHITECTURE

3.1 Introduction

Designing system is a way to explain the various modules, interfaces, components of the system so that once requirements are satisfied. Proper document is made so as to note the designing of the system and so that what is actually required is implemented and matches the requirements.

3.2 System Design Document

3.2.1 Overview

This document shows various requirements of the system, the operating system, files, database, input, output, various interfaces, logic implemented, internal interface and external interface.

3.2.2 INTRODUCTION

Introduction part describes the following things:

1. Purpose and scope of the implementation
2. The summary of the project to be implemented from management view
3. The overview of the system using both technical and non technical terms. The flowchart and layout of the project with appropriate diagrams.
4. Various constraints that are seen while implementing the project and any imaginations/exceptions made by the team while developing.
5. Future problems that may arise after developing the system
6. Various stakeholders involved while implementing. These can be the Project Manager, QA Manager, Security, Configuration, Organization, etc.
7. Various references that have been used while implementing the project
8. A summary of abbreviations and short forms that have been used.

3.2.3 SYSTEM ARCHITECTURE

This describes the architecture of the system that has been used. The architecture can be of the hardware or the software level. Various architectures that are described are as follows:

1. System Hardware Architecture which includes the complete architecture of system including various components.
2. System Software Architecture which described the software part, language, functions, tools, classes, Object-oriented diagrams, etc.
3. Internal Communication Architecture which describes the communication between the various components and modules of the system.

3.2.4 FILE AND DATABASE DESIGN

Along with the hardware and software part, the next important component is the file and database design. This includes the interaction of the Database Administrator with the various files (Both DBMS and non-DBMS files) which are related to the development of the project. It also describes how the data is stored in the DBMS and what are the various schemas, sub-schemas, tables, records, sets, etc used while storing the data in the DBMS. Various methods to access the tables and records and the size of the database and tables described.

Non- DBMS files includes the description of the files with proper input and output. This also includes methods of how to access the files in the database, that is, the keys and indexes or any other reference data that has been used, ways to access the files that have been stored in the DBMS.

3.2.5 HUMAN-MACHINE INTERFACE

Along with the hardware, software and communication with the database, the hardware must interact with humans as well. This section explains the interaction of the hardware component with humans. From the initial start of the project, the person using the project should be able to understand the flow of the program and the next step involved in the program. The interaction between machine and human should be smooth enough for the proper flow. Various parts involved in this interface are as described below:

- Inputs

- Outputs

3.2.6 DETAILED DESIGN

This part describes the detailed designing of all the components involved which are hardware, software, interaction between various modules, etc. Depending upon the implementation all the details are required, some are as described below:

- Details of various hardware components used
- Various connectors / cables required
- Power requirement for input
- Memory requirement / Storage Space requirement
- Processor Speed and functionality
- Switches and cables used
- GUI of all the components used in the hardware
- Functions , algorithms and interaction of various modules used in the implementation
- Data Entry Methods and various ways to access the records and files structures.
- Various elements used in saving the data to a particular storage area.
- Communication ways used to transfer the data
- Topology of the cables and connections
- Number of Clients and server used in maintaining the connection

3.2.7 SYSTEM INTEGRITY CONTROLS

This includes ways to recover the data in case of complete failure, unauthorized access or misuse of the implemented system. The system must be accessible and must be recoverable in such cases and there should be proper security measures to recover the system in all such cases. There should be proper review and audit system that must occur on a particular stage of the implementation. Security must be of concern so that the implementation is not openly available to all and must be controlled by limiting the number of users. All the data entered must be verified with proper sources.

CHAPTER 4

Cluster Analysis

There are various types of clustering algorithms for eg: Partitioning methods, fuzzy, Density based, model based clustering, etc. The most popular clustering algorithm is k-means clustering [12][13]. K-means clustering is based on unsupervised learning algorithms.

For each cluster, there is a centroid; a centroid is a central data point which is at euclidean distance from all the data points. We try to keep the centroids far away so that a lesser variation is observed [14][15]. After this, each data point is assigned nearest centroid. We use the same clustering process in our research methodology [16][17].

4.1 Clustering Algorithm

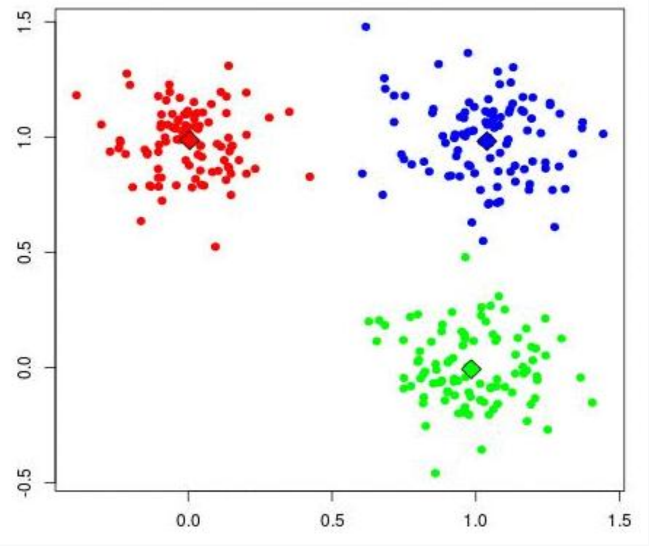


Figure 4.1: Clustering Algorithm

Figure 4.1 explains the clustering algorithm for a dataset. There are 3 different clusters and for each cluster there is a centroid which is present at Euclidean distance from different data points [18][19]. Euclidean distance is calculated as :

Euclidean Distance = $\sqrt{}(V_H- V_1)^2 + (V_W-W_1)^2$

Where:

V_H= Variable Height Value

V_1= Variable Height of Centroid value of Cluster 1

V_W= Variable Weight Value

W_1= Variable Weight of Centroid value of Cluster 1

There can be a number of clusters in a given dataset, so our first step is to reduce the number of clusters . For our research purpose, we are limiting the number of clusters to the count that represent more than 1% of the total individuals. This would limit our cluster count to 6 [20][21].

Now the difficulty lies in selecting the value of k. In our study we are taking K as 2,3 and 5 and analyzing the values. Keeping the values as high as 20,30 and 50 would give very few elements in cluster and cluster count would be very high. This would result in difficulty in studying our output.

The following three different clusters each in Figure 4.2, Figure 4.3 and Figure 4.4 represents the count of people lying in different categories of defined datasets. Based on these clusters we can filter the type of data that we want to see and analyze.

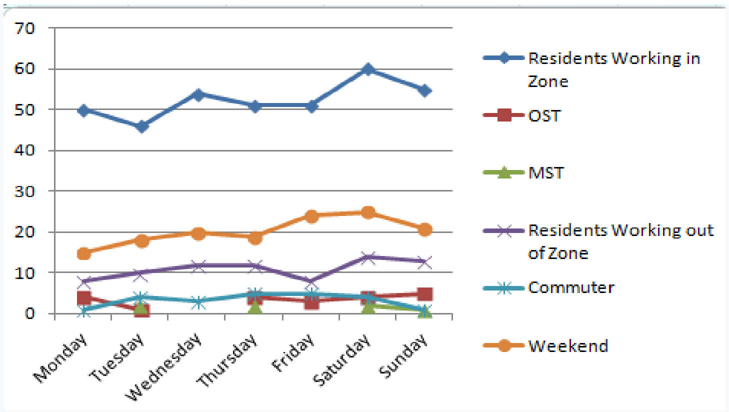


Figure 4.2: Cluster 1

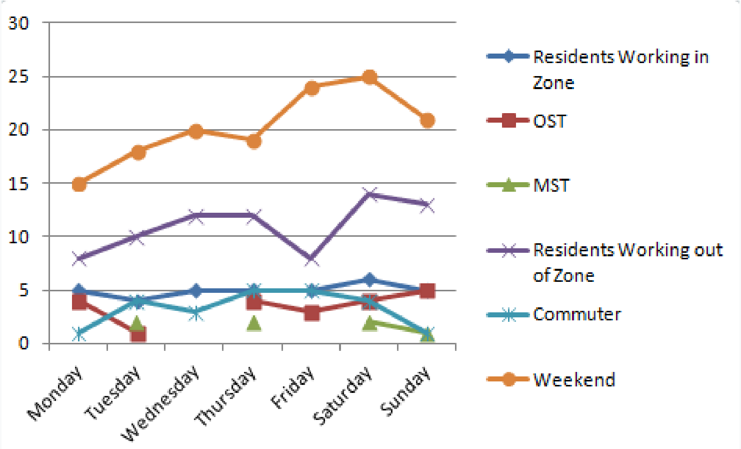


Figure 4.3 : Cluster 2

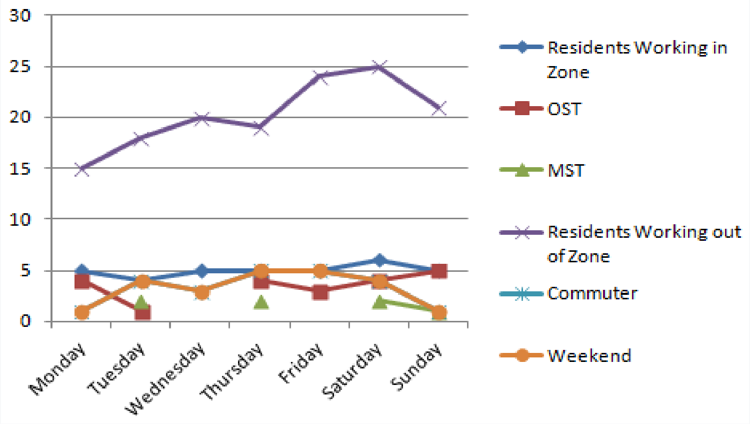


Figure 4.4: Cluster 3

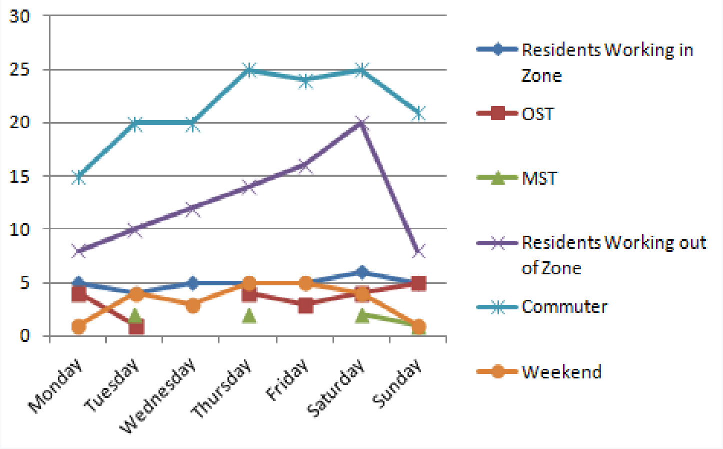


Figure 4.5 :Cluster 4

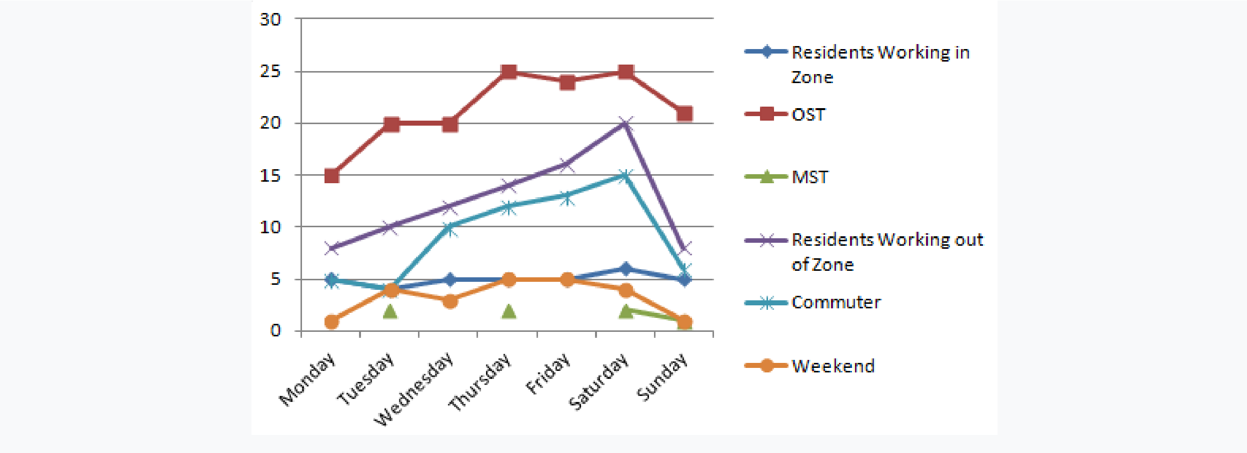


Figure 4.6 :Cluster 5

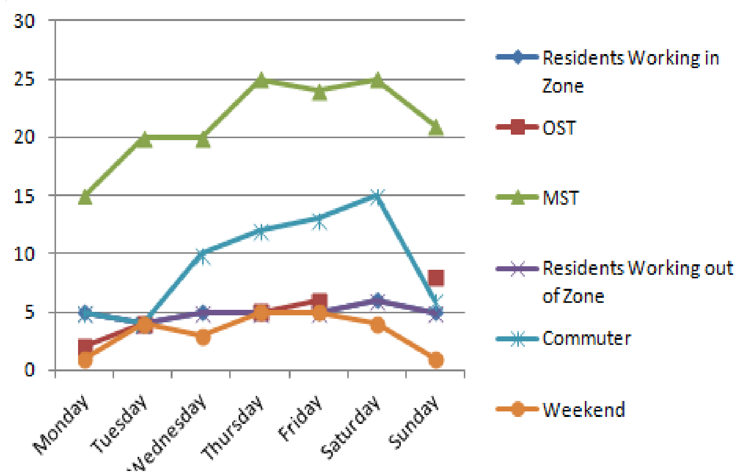


Figure 4.7 : Cluster 6

Now we analyze the output of these 6 clusters as represented in Figure 4.5,Figure 4.6, Figure 4.7. Each of these clusters depict a particular analysis. Cluster 1 shows the individuals working in zone with a high count. Similarly cluster 2 shows the individuals who are visible only on weekends . Cluster 3 shows the people who are out of area S1 and are coming just for work in S1. Cluster 4 represents the people who doesnot belong to area S1, they are just working or studying.

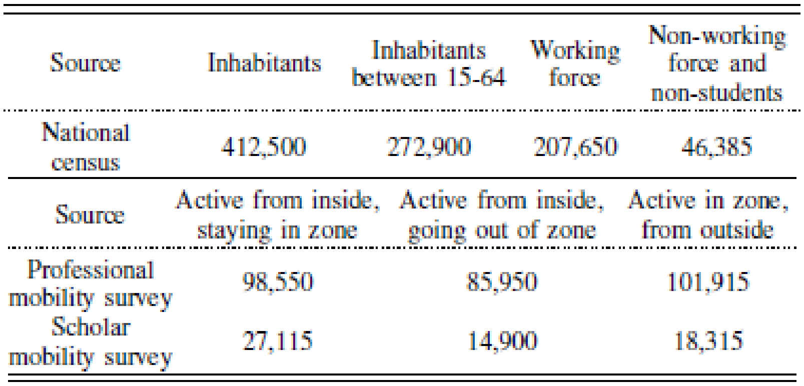
Cluster 5 shows majority of people whose appearance was seen just once and were not in area S1 for more than an hour. Lastly cluster 6 shows the people who made multiple entries in area S1 and were present for more than an hour but doesnot belong to area S1 [22][23][24].

4.2 Evaluation

For validation of the data represented by our study, we can make a comparative analysis between the data represented by national census and surveys.

From the national census there are four different variables:

1. Total number of residents
2. Number of residents between age 20 to 50
3. Number of registered workers
4. Number of registered non-workers i.e students,etc



Source	Inhabitants	Inhabitants between 15-64	Working force	Non-working force and non-students
National census	412,500	272,900	207,650	46,385

Source	Active from inside, staying in zone	Active from inside, going out of zone	Active in zone, from outside
Professional mobility survey	98,550	85,950	101,915
Scholar mobility survey	27,115	14,900	18,315

Figure 4.8 : National Data about area S1

Figure 4.8 represents the National Data of the zone under study. This depicts the number of residents, workers, students and others in the particular area.

CHAPTER 5

Research Background

5.1 Java

Java is both a high-level programming language and a platform. For implementation of this project Java language is used since Java is Simple, Object-Oriented, Portable, High performance, Dynamic, Integrated and Secure. Figure 5.1 explains the basic flow of a Java Program.

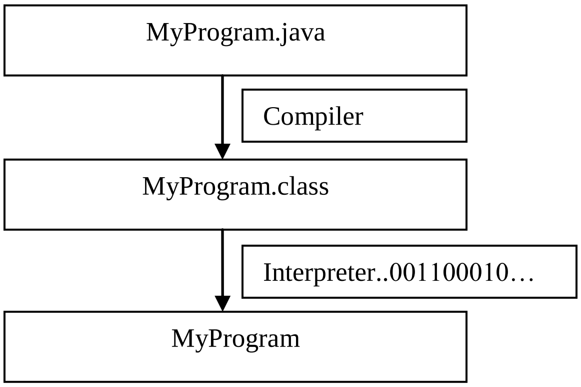


Figure 5.1 :Java Program Flowchart

5.2 Java Platform

Java platform consists of 2 parts as described in Figure 5.2:

- The Java Virtual Machine (Java VM)
- The Java Application Programming Interface (Java API)

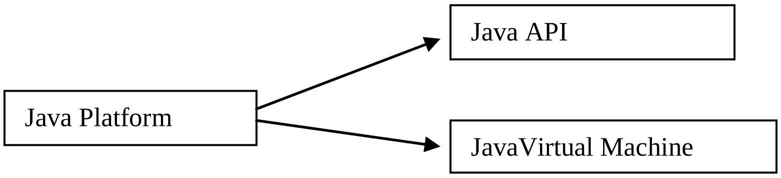


Figure 5.2: Java Platform

5.3 Tools Used

The experiments were performed on a Microsoft$^{©}$ Windows® 10 machine with 8GB RAM, intel Core i5 processor, 32MB cache memory. Code implementation was done with Java Language, apache-tomcat-6.0.20, eclipse-jee-mars-2-win32-x86_64, WampServer2.2b-x64 (used for creating web applications).

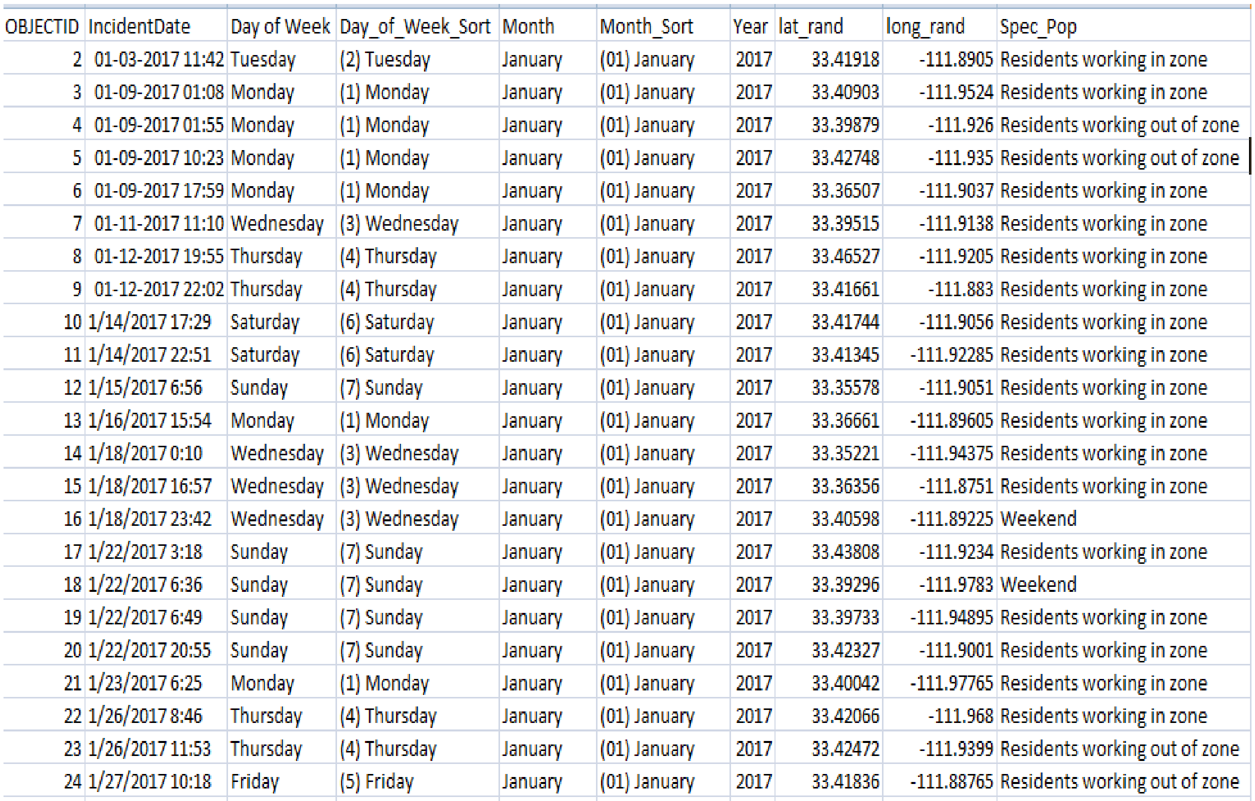
The first step for our implementation includes defining an area S1 and then collecting a database of various transactions/events performed by different users moving in and out of that area in the complete day. Some of the events that can occur are as described below:

- Transmitting or receiving a phone call
- Transmitting or receiving as text message
- Handovers during call
- Area or location of information receiving
- Events induced after every 3 minutes of call
- Events induced after every 3 hours of call

Database as shown in Figure 5.3 includes details like: Unique ID i.e the IMEI of the device, Incident Date, Week of the day, Month, Year, Longitude, Latitude and finally the categorization of users based on our defined categories which are:

1. Commuters
2. One Single Transit
3. Multiple Transit
4. Residents working in zone
5. Residents working out of zone
6. Weekend

Snippet of the database selected is as shown below:



OBJECTID	IncidentDate	Day of Week	Day_of_Week_Sort	Month	Month_Sort	Year	lat_rand	long_rand	Spec_Pop
2	01-03-2017 11:42	Tuesday	(2) Tuesday	January	(01) January	2017	33.41918	-111.8905	Residents working in zone
3	01-09-2017 01:08	Monday	(1) Monday	January	(01) January	2017	33.40903	-111.9524	Residents working in zone
4	01-09-2017 01:55	Monday	(1) Monday	January	(01) January	2017	33.39879	-111.926	Residents working out of zone
5	01-09-2017 10:23	Monday	(1) Monday	January	(01) January	2017	33.42748	-111.935	Residents working out of zone
6	01-09-2017 17:59	Monday	(1) Monday	January	(01) January	2017	33.36507	-111.9037	Residents working in zone
7	01-11-2017 11:10	Wednesday	(3) Wednesday	January	(01) January	2017	33.39515	-111.9138	Residents working in zone
8	01-12-2017 19:55	Thursday	(4) Thursday	January	(01) January	2017	33.46527	-111.9205	Residents working in zone
9	01-12-2017 22:02	Thursday	(4) Thursday	January	(01) January	2017	33.41661	-111.883	Residents working in zone
10	1/14/2017 17:29	Saturday	(6) Saturday	January	(01) January	2017	33.41744	-111.9056	Residents working in zone
11	1/14/2017 22:51	Saturday	(6) Saturday	January	(01) January	2017	33.41345	-111.92285	Residents working in zone
12	1/15/2017 6:56	Sunday	(7) Sunday	January	(01) January	2017	33.35578	-111.9051	Residents working in zone
13	1/16/2017 15:54	Monday	(1) Monday	January	(01) January	2017	33.36661	-111.89605	Residents working in zone
14	1/18/2017 0:10	Wednesday	(3) Wednesday	January	(01) January	2017	33.35221	-111.94375	Residents working in zone
15	1/18/2017 16:57	Wednesday	(3) Wednesday	January	(01) January	2017	33.36356	-111.8751	Residents working in zone
16	1/18/2017 23:42	Wednesday	(3) Wednesday	January	(01) January	2017	33.40598	-111.89225	Weekend
17	1/22/2017 3:18	Sunday	(7) Sunday	January	(01) January	2017	33.43808	-111.9234	Residents working in zone
18	1/22/2017 6:36	Sunday	(7) Sunday	January	(01) January	2017	33.39296	-111.9783	Weekend
19	1/22/2017 6:49	Sunday	(7) Sunday	January	(01) January	2017	33.39733	-111.94895	Residents working in zone
20	1/22/2017 20:55	Sunday	(7) Sunday	January	(01) January	2017	33.42327	-111.9001	Residents working in zone
21	1/23/2017 6:25	Monday	(1) Monday	January	(01) January	2017	33.40042	-111.97765	Residents working in zone
22	1/26/2017 8:46	Thursday	(4) Thursday	January	(01) January	2017	33.42066	-111.968	Residents working in zone
23	1/26/2017 11:53	Thursday	(4) Thursday	January	(01) January	2017	33.42472	-111.9399	Residents working out of zone
24	1/27/2017 10:18	Friday	(5) Friday	January	(01) January	2017	33.41836	-111.88765	Residents working out of zone

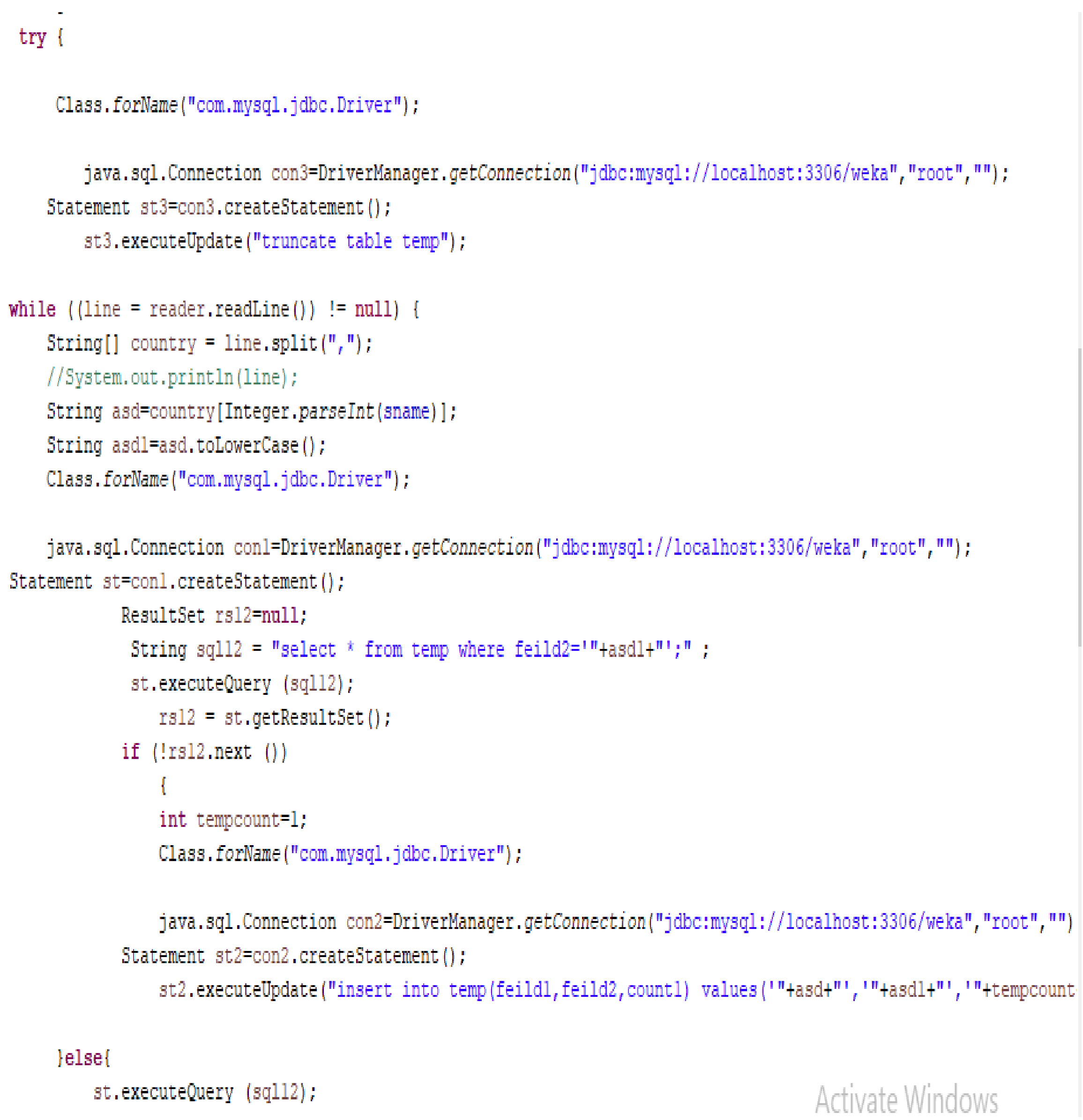
Figure 5.3: Snippet of Database Selected

The next step involves creating a training algorithm which can understand the input that is being provided and based on the input produces an output of our reference. Training algorithms are those which can provide a related and sensitive data from a set of database. There are 2 approaches that are followed while training a database:

- Supervised Learning Algorithms
- Unsupervised Learning Algorithms

Both algorithms differ in the way they are mapped to the output of the program. Supervised learning maps the input of the data to the output while unsupervised learning does not concern mapping the input to the output rather it studies the pattern in the database.

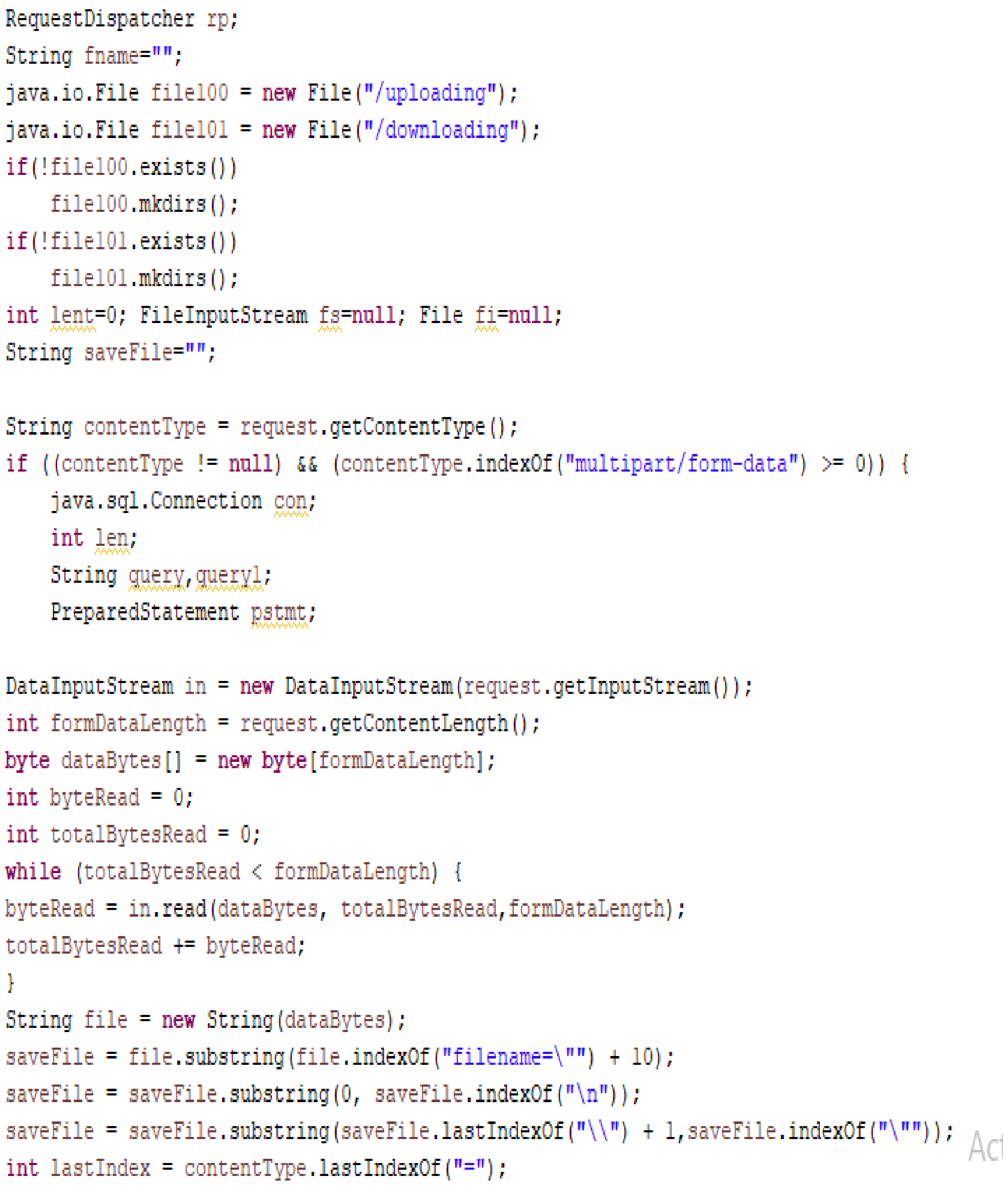
The following Figure 5.4 shows the implementation of k-means:



```
        -
  try {

      Class.forName("com.mysql.jdbc.Driver");

         java.sql.Connection con3=DriverManager.getConnection("jdbc:mysql://localhost:3306/weka","root","");
     Statement st3=con3.createStatement();
         st3.executeUpdate("truncate table temp");

 while ((line = reader.readLine()) != null) {
     String[] country = line.split(",");
     //System.out.println(line);
     String asd=country[Integer.parseInt(sname)];
     String asd1=asd.toLowerCase();
     Class.forName("com.mysql.jdbc.Driver");

     java.sql.Connection con1=DriverManager.getConnection("jdbc:mysql://localhost:3306/weka","root","");
 Statement st=con1.createStatement();
             ResultSet rsl2=null;
              String sql12 = "select * from temp where feild2='"+asd1+"';" ;
              st.executeQuery (sql12);
                 rsl2 = st.getResultSet();
             if (!rsl2.next ())
                 {
                 int tempcount=1;
                 Class.forName("com.mysql.jdbc.Driver");

                 java.sql.Connection con2=DriverManager.getConnection("jdbc:mysql://localhost:3306/weka","root","")
             Statement st2=con2.createStatement();
                 st2.executeUpdate("insert into temp(feild1,feild2,count1) values('"+asd+"','"+asd1+"','"+tempcount

      }else{
          st.executeQuery (sql12);
```

Figure 5.4 : Implementation of K-Means

The final output will display the count of users lying in each category as defined in our input data. The total count will be displayed for every day of the week. Also a comparison of various applied algorithms has been done which depicts the count of residents in area S1 on each day. Some of the applied algorithms are Random Forest, Logistic Regression, k-Means and Support Vector Machine.



```
RequestDispatcher rp;
String fname="";
java.io.File file100 = new File("/uploading");
java.io.File file101 = new File("/downloading");
if(!file100.exists())
    file100.mkdirs();
if(!file101.exists())
    file101.mkdirs();
int lent=0; FileInputStream fs=null; File fi=null;
String saveFile="";

String contentType = request.getContentType();
if ((contentType != null) && (contentType.indexOf("multipart/form-data") >= 0)) {
    java.sql.Connection con;
    int len;
    String query,query1;
    PreparedStatement pstmt;

DataInputStream in = new DataInputStream(request.getInputStream());
int formDataLength = request.getContentLength();
byte dataBytes[] = new byte[formDataLength];
int byteRead = 0;
int totalBytesRead = 0;
while (totalBytesRead < formDataLength) {
byteRead = in.read(dataBytes, totalBytesRead,formDataLength);
totalBytesRead += byteRead;
}
String file = new String(dataBytes);
saveFile = file.substring(file.indexOf("filename=\"") + 10);
saveFile = saveFile.substring(0, saveFile.indexOf("\n"));
saveFile = saveFile.substring(saveFile.lastIndexOf("\\") + 1,saveFile.indexOf("\""));
int lastIndex = contentType.lastIndexOf("=");
```

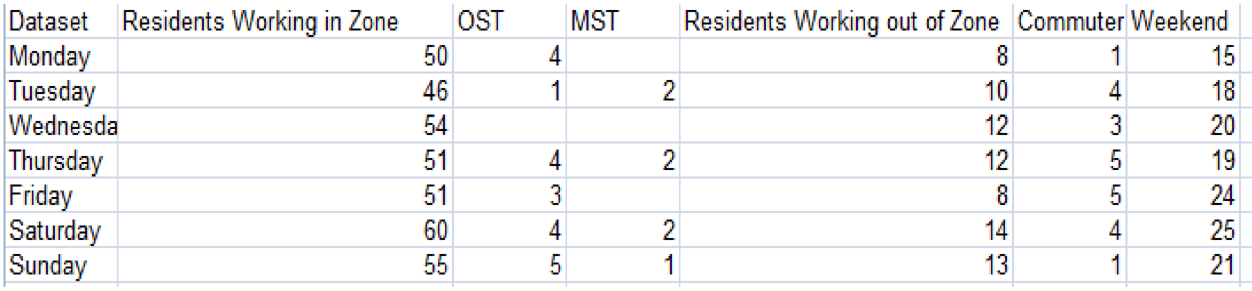
Figure 5.5 :Generating Output File of K-means implementation

CHAPTER 6

RESULTS

After training numerous models we were able to achieve an overwhelming best validation accuracy of 70% with all the profile curves. As a result it is possible that an IMEI is detected in one week and in the other week it is not present. The Profile curves also show that many of the users are Absent which does not mean that the algorithm is not able to set a profile but those individuals are not present in the area chosen for our test.

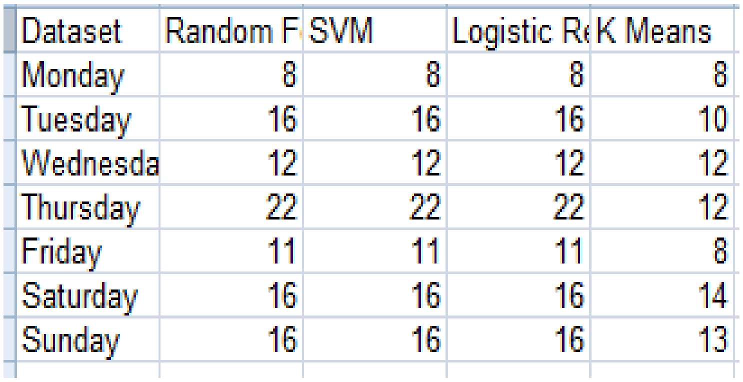
After training the model we were able to predict the population of the area and were able to categorize people in different categories as shown in Figure 6.1.



Dataset	Residents Working in Zone	OST	MST	Residents Working out of Zone	Commuter	Weekend
Monday	50	4		8	1	15
Tuesday	46	1	2	10	4	18
Wednesda	54			12	3	20
Thursday	51	4	2	12	5	19
Friday	51	3		8	5	24
Saturday	60	4	2	14	4	25
Sunday	55	5	1	13	1	21

Figure 6.1 : Results obtained after training model

After running various algorithms as shown in Figure 6.2 we were able to compare the output produced and the population detected by all algorithms in the tested area. These numbers need to be adjusted as per the population of the studied area and the data may also vary depending upon the data collected by different operators in that particular area/country.



Dataset	Random F	SVM	Logistic Re	K Means
Monday	8	8	8	8
Tuesday	16	16	16	10
Wednesda	12	12	12	12
Thursday	22	22	22	12
Friday	11	11	11	8
Saturday	16	16	16	14
Sunday	16	16	16	13

Figure 6.2 : Results obtained after running various algorithms

CHAPTER 7

CONCLUSION AND FUTURE WORK

The cluster represents the number of individuals those are tracked by the mobile network operator based on the Call Data Record. These numbers must be managed accordingly based on the total population of the particular area. The outcome of our research is in consistent with the national survey data. This also has various drawbacks, some of which are as mentioned below:

1. It is impossible to track the movement of individual from one location to another and then mapping the individual into a particular category
2. There can be many privacy issues since we are tracking the location of individual at every point of time
3. Too much dependency on user for tracking the data, if an individual is not using mobile device then it is not possible to count that individual in that particular area.

Future Work

Our future works concerns on finding the similarity between the various clusters formed and defining the threshold of the similarity. We would focus on another region and try to find the similarity between the two regions studied and therefore finding a way to generalize the study so that it can be implemented anywhere anytime.

REFERENCES:

[1] J. H. Kang, W. Welbourne, B. Stewart, and G. Borriello, “Extracting Places from Traces of Locations,” in *Proceedings of the 2nd ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots*, ser. WMASH 04. Philadelphia, PA, USA: ACM, 2004, pp. 110–118.

[2] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, “Cellular Census: Explorations in Urban Data Collection,” *IEEE Pervasive Computing*, vol. 6, no. 3, pp. 30–38, Jul. 2007.

[3] G. Rose, “Mobile Phones as Traffic Probes: Practices, Prospects and Issues,” *Transport Reviews*, vol. 26, no. 3, May 2006.

[4] C. Iovan, A.-M. Olteanu, T. Couronn´e, and Z. Smoreda, “Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies,” in *Geographic Information Science at the Heart of Europe*, ser. Lecture Notes in Geoinformation and Cartography, D. Vandenbroucke, B. Bucher, and J. Crompvoets, Eds. Springer International Publishing, May 2013, pp. 247–265.

[5] F. Calabrese, L. Ferrari, and V. D. Blondel, “Urban Sensing Using Mobile Phone Network Data: A Survey of Research,” *ACM Comput. Surv.*, vol. 47, no. 2, pp. 25:1–25:20, Nov. 2014

[6] M. C. Gonz´alez, C. A. Hidalgo, and A.-L. Barab´asi, “Understanding individual human mobility patterns,” *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[7] C. Song, Z. Qu, N. Blumm, and A.-L. Barab´asi, “Limits of Predictability in Human Mobility,” *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb.2010.

[8] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, “Real- Time Urban Monitoring Using Cell Phones: A Case Study in Roma,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 141–151, Mar. 2011.

[9] F. Alhasoun, A. Almaatouq, K. Greco, R. Campari, A. Alfaris, and C. Ratti, “The City Browser: Utilizing Massive Call Data to Infer City Mobility Dynamics,” in *The 3rd International Workshop on Urban Computing* (UrbComp 2014), New York, NY, Aug. 2014.

[10] R. Becker, R. C´aceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, “Human Mobility Characterization from Cellular Network Data,” *Commun. ACM*, vol. 56, no. 1, pp. 74–82, Jan. 2013.

[11] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira Jr., and C. Ratti, “Understanding individual mobility patterns from urban sensing data: A mobile phone trace example,” *Transportation Research Part C: Emerging Technologies*, vol. 26, pp. 301–313, Jan. 2013.

[12] C. Joumaa, A. Caminada, and S. Lamrous, “Mask Based Mobility Model A new mobility model with smooth trajectories,” in *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks and Workshops*, 2007. WiOpt 2007. 5th International Symposium on, 2007, pp. 1–6.

[13] K. A. Ali, M. Lalam, L. Moalic, and O. Baala, “V-MBMM: Vehicular Mask-Based Mobility Model,” in Networks (ICN), 2010 Ninth International Conference on, 2010, pp. 243–248.

[14] S. Isaacman, R. Becker, R. C´aceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger, “Human Mobility Modeling at Metropolitan Scales,” in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys ’12. Low Wood Bay, Lake District, UK: ACM, 2012, pp. 239–252.

[15] D. J. Mir, S. Isaacman, R. C´aceres, M. Martonosi, and R. N. Wright, “DP-WHERE: Differentially private modeling of human mobility,” in 2013 *IEEE International Conference on Big Data*, Oct. 2013, pp. 580–588.

[16] J. Yuan, Y. Zheng, and X. Xie, “Discovering Regions of Different Functions in a City Using Human Mobility and POIs,” in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’12. New York, NY, USA: ACM, 2012, pp. 186–194.

[17] J. L. Toole, M. Ulm, M. C. Gonz´alez, and D. Bauer, “Inferring Land Use from Mobile Phone Activity,” in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, ser. UrbComp ’12. New York, NY, USA: ACM, 2012, pp. 1–8.

[18] A. Wesolowski, N. Eagle, A. M. Noor, R. W. Snow, and C. O. Buckee, “The impact of biases in mobile phone ownership on estimates of human mobility,” *Journal of The Royal Society Interface*, vol. 10, no. 81, p. 20120986, Apr. 2013.

[19] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, “Exploring Human Mobility with Multi-source Data at Extremely Large Metropolitan Scales,” in *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom ’14. New York,NY, USA: ACM, 2014, pp. 201–212.

[20] D. Zhang, J. Zhao, F. Zhang, and T. He, “coMobile: Real-time Human Mobility Modeling at Urban Scale Using Multi-view Learning,” in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL ’15. New York, NY, USA: ACM, 2015, pp. 40:1–40:10.

[21] M. A. Bayir, M. Demirbas, and N. Eagle, “Discovering spatiotemporal mobility profiles of cellphone users,” in 2009 *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks Workshops*, Jun. 2009, pp. 1–9.

[22] R. Fergus, A. Zisserman, and P. Perona, “Sampling methods for unsupervised learning,” in Advances in Neural Information Processing Systems 17. Neural information processing systems foundation, 2005.

[23] P. J. Rousseeuw, “Silhouettes - A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, no. 0, pp. 53–65, 1987.

[24] R. W. Hamming, “Error Detecting and Error Correcting Codes,” Bell System Technical Journal, vol. 29, no. 2, pp. 147–160, Apr. 1950. INSEE, “Definition of working place.” “Recensement de la population - La pr´ecision des resultants du recensement,” INSEE, Tech. Rep., Aug. 2009.