Human Mobily Path and Prediction

Rahul Sarkar

SUNY Albany

Rsarkar3@albany.edu

Abstract

Individual mobility is the study that describes how people move within an area or network. In recent years several researches have been done for this purpose and there has been a surge in large data sets available in individual movements. Most of these data sets are collected from cellphone and/or GPS with variable accuracy depending on the distance from the tower. Large scale data such as mobile phone traces are very important source for urban modeling. The individual travel patterns collapse into a single probability distribution but despite the diversity of their travel history humans follow simple reproducible patterns. This similarity in travel pattern can help us in a very diverse areas of applications such as city planning, traffic engineering, spread of diseases and mobile viruses. The motive of this project is to show that by using a series of direct measurement that human trajectories do follow several high reproducible scaling patterns.

Keywords: Mobility, Dataset, GPS, Phone traces, Processing and utilizing data

Introduction

In recent years there has been a high increase in use of mobile devices all over the world weather it is a mobile phone or a GPS tracker. With the help of these devices several researches are going on finding the human movement patterns or group movement patterns. In some cases, they use these data to predict the next location of the user. By acquiring these patterns, we can solve a huge number of issues related to human movement such as city planning, traffic management, emergency responses, spread of mobile viruses and other diseases.

With the help of mobile phone trace data researchers examine human mobility behavior with lower collection cost, larger sample size as it can have huge human base, higher update frequency and broader spatial and temporal coverage. The mobile phone locations are regularly collected by Google maps and many more mobile applications as individual travel history; therefore, the datasets are available at no cost.

With so many advantages the mobile phone trace data also have several drawbacks for research. (1) Mobile phone users may not represent a whole population chosen from a random sample. The data should be analyzed before reaching into a result. (2) The datasets are generally not designed for analysis purposes so mostly not in an easy to use format, which again need some intensive processing of the raw data. As per recent

Global Attitude Survey there are more than 5 billion mobile devices across the world and half of them are smartphones. Most of the mobile applications running on these smart phones ask permission for location access, media access and a couple of different things. Though there is an option to allow or deny these permissions, the applications work more accurately when allowed access. For example, if we are using a food delivery application it needs to know the current location of the person and gives the food options accordingly. With these processes running in the background for the applications, the battery life decreases in an exponential way. If we can turn these processes off when not required will save the battery life to some extent.

This project represents the step towards building a methodology to utilize mobile phone data or tower location data for finding out the probability of the user's location based on their historical path. We will also use the GPS location data in this case to verify the accuracy of predicted location.

Related Work

There have been several approaches proposed to detect location tracking. Most of these approaches are generic and may produce false positive, i.e., wrong location detected by the system. However there has been surprisingly limited work done in the area of understanding the pattern of individual and group mobility. Recent developments in location-based technologies enables us to track individual movement and activity participation in urban space across time. Work has also been done on frequently-visit based location prediction in a quantitative manner.

Based on various empirical datasets, several human mobility models have been proposed that captures human mobility to a certain extent. There were quite a couple of researches where they used the cellphone data as GPS was not available on the mobile phones in early 20th century. So, researchers used the available tower data source. Due to the significant use of mobile phones, the study of human mobility has changed a lot. Current mobile phones utilize cell tower information and the GPS system for more accurate location tracking. With billions of people carrying their phones everyday provides a large amount of data on human movement. This data is being collected and made available in the form of opencellid or mozilla location service, open up new opportunities for modelling and predicting human mobility more accurately.

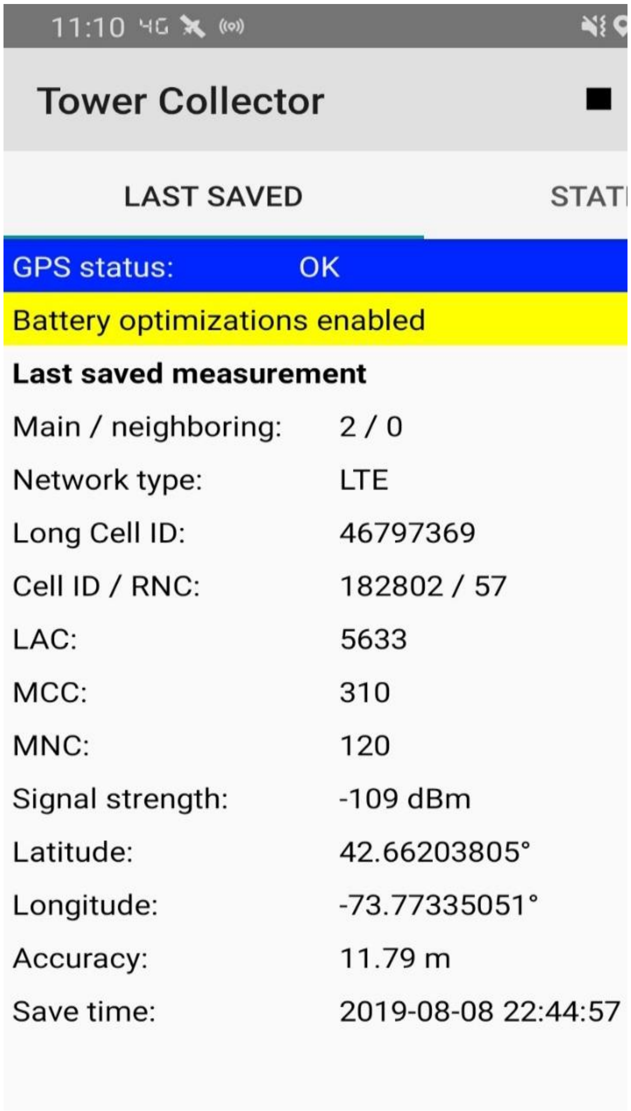
Mobility Data Sets

Several human mobility data have been gathered and published in the past. This mobility data contains rich knowledge about the locations and can help in addressing many challenges. For example, understanding the human mobility behavior inside a city can help forecasting of the traffic. Another example is that we can identify the locations by the means of the transition between these locations, e.g., people usually go to work in the morning and come back home after 4pm on weekdays and visit shopping centers after work or on weekends.

The main mobility datasets are recorded according to 1. relevant location with access points (Cellular tower, Wifi access points etc.), 2. GPS information by devices, 3. Aggregated GPS points recorded by vehicles such as taxis or buses. For this project work we collected the mobile data for a couple of months using several mobile applications such as Tower Collector, Network Cell info

Lite, GPS logger etc. We will cover them one by one below.

Tower Collector: This mobile application can run in background and collect the data. The collected data consists of measurements about the Tower, Network type, cell id, latitude and longitude of the tower etc. Once the data is collected, we can retrieve it using different file formats like CSV, GPX or JSON.



11:10 4G

Tower Collector

LAST SAVED STATI

GPS status: OK

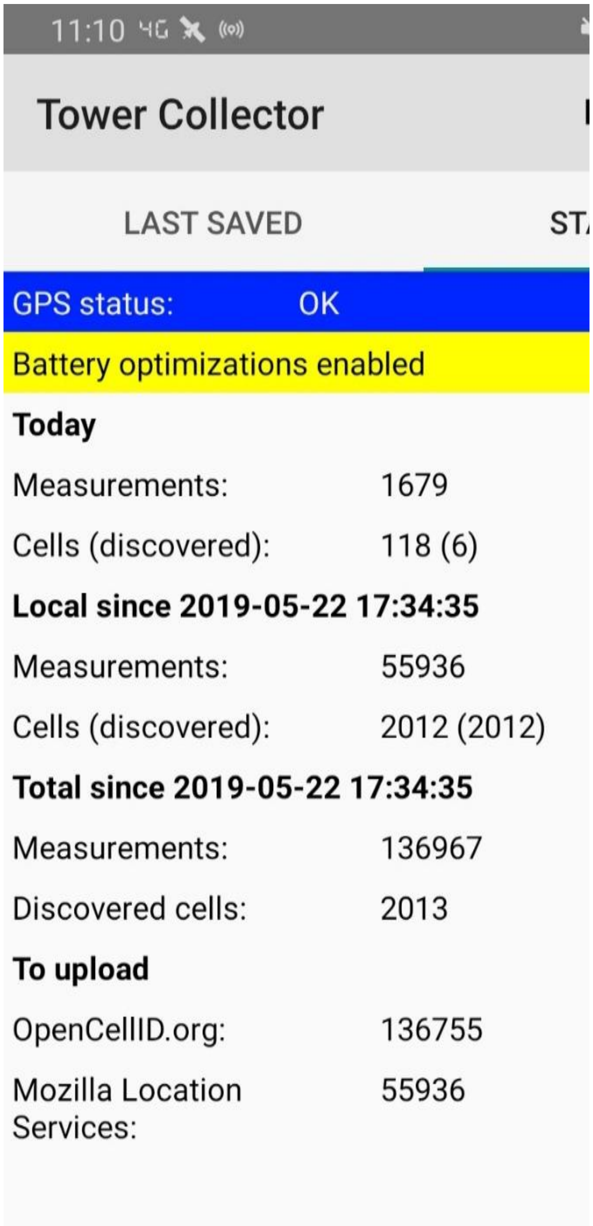
Battery optimizations enabled

Last saved measurement

Main / neighboring:	2 / 0
Network type:	LTE
Long Cell ID:	46797369
Cell ID / RNC:	182802 / 57
LAC:	5633
MCC:	310
MNC:	120
Signal strength:	-109 dBm
Latitude:	42.66203805°
Longitude:	-73.77335051°
Accuracy:	11.79 m
Save time:	2019-08-08 22:44:57

Fig 1: Current tower information in Tower Collector application

.



11:10 4G

Tower Collector

LAST SAVED ST

GPS status: OK

Battery optimizations enabled

Today

Measurements:	1679
Cells (discovered):	118 (6)

Local since 2019-05-22 17:34:35

Measurements:	55936
Cells (discovered):	2012 (2012)

Total since 2019-05-22 17:34:35

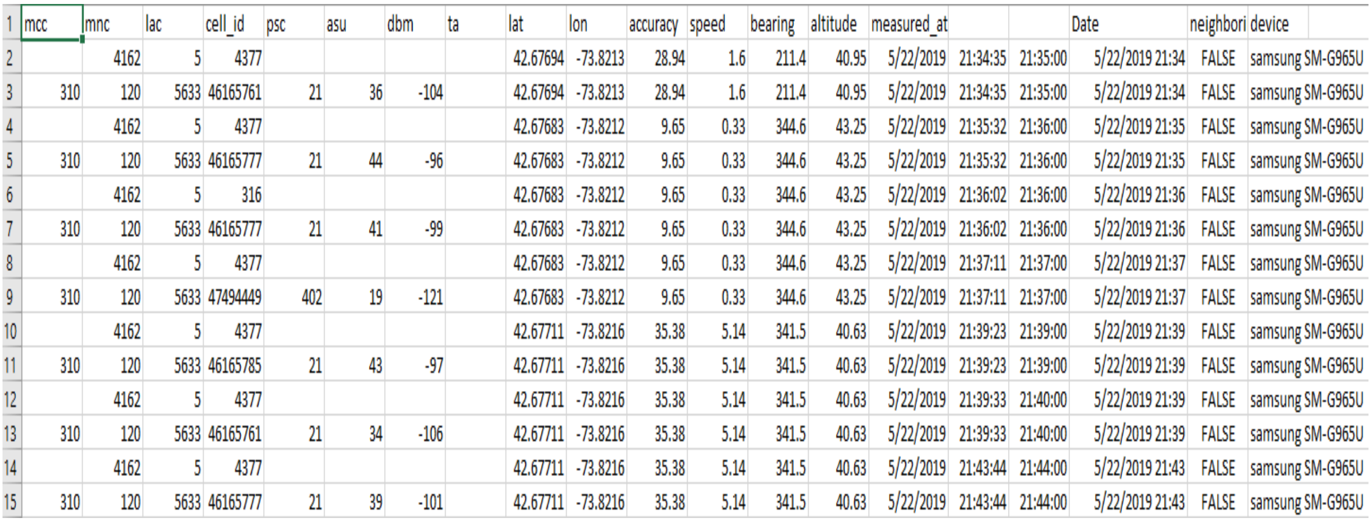
Measurements:	136967
Discovered cells:	2013

To upload

OpenCellID.org:	136755
Mozilla Location Services:	55936

Fig 2: Information about all the collected tower data

The advantage about the Tower Collector application is that it collects all the required information and stores them cumulatively without losing and old data from the data sets.



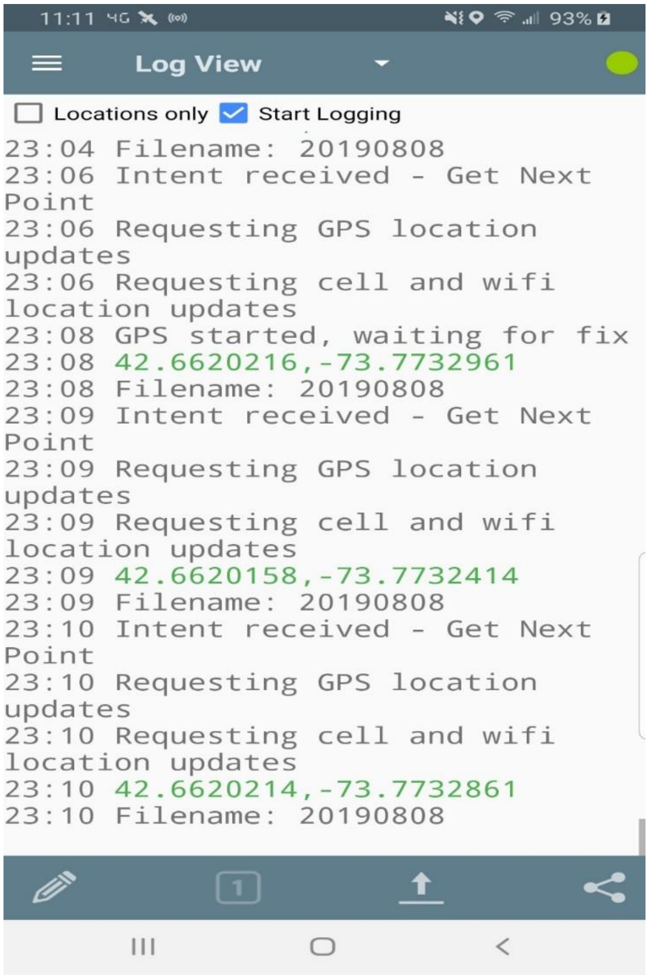
1	mcc	mnc	lac	cell_id	psc	asu	dbm	ta	lat	lon	accuracy	speed	bearing	altitude	measured_at			Date	neighbori	device
2		4162	5	4377					42.67694	-73.8213	28.94	1.6	211.4	40.95	5/22/2019	21:34:35	21:35:00	5/22/2019 21:34	FALSE	samsung SM-G965U
3	310	120	5633	46165761	21	36	-104		42.67694	-73.8213	28.94	1.6	211.4	40.95	5/22/2019	21:34:35	21:35:00	5/22/2019 21:34	FALSE	samsung SM-G965U
4		4162	5	4377					42.67683	-73.8212	9.65	0.33	344.6	43.25	5/22/2019	21:35:32	21:36:00	5/22/2019 21:35	FALSE	samsung SM-G965U
5	310	120	5633	46165777	21	44	-96		42.67683	-73.8212	9.65	0.33	344.6	43.25	5/22/2019	21:35:32	21:36:00	5/22/2019 21:35	FALSE	samsung SM-G965U
6		4162	5	316					42.67683	-73.8212	9.65	0.33	344.6	43.25	5/22/2019	21:36:02	21:36:00	5/22/2019 21:36	FALSE	samsung SM-G965U
7	310	120	5633	46165777	21	41	-99		42.67683	-73.8212	9.65	0.33	344.6	43.25	5/22/2019	21:36:02	21:36:00	5/22/2019 21:36	FALSE	samsung SM-G965U
8		4162	5	4377					42.67683	-73.8212	9.65	0.33	344.6	43.25	5/22/2019	21:37:11	21:37:00	5/22/2019 21:37	FALSE	samsung SM-G965U
9	310	120	5633	47494449	402	19	-121		42.67683	-73.8212	9.65	0.33	344.6	43.25	5/22/2019	21:37:11	21:37:00	5/22/2019 21:37	FALSE	samsung SM-G965U
10		4162	5	4377					42.67711	-73.8216	35.38	5.14	341.5	40.63	5/22/2019	21:39:23	21:39:00	5/22/2019 21:39	FALSE	samsung SM-G965U
11	310	120	5633	46165785	21	43	-97		42.67711	-73.8216	35.38	5.14	341.5	40.63	5/22/2019	21:39:23	21:39:00	5/22/2019 21:39	FALSE	samsung SM-G965U
12		4162	5	4377					42.67711	-73.8216	35.38	5.14	341.5	40.63	5/22/2019	21:39:33	21:40:00	5/22/2019 21:39	FALSE	samsung SM-G965U
13	310	120	5633	46165761	21	34	-106		42.67711	-73.8216	35.38	5.14	341.5	40.63	5/22/2019	21:39:33	21:40:00	5/22/2019 21:39	FALSE	samsung SM-G965U
14		4162	5	4377					42.67711	-73.8216	35.38	5.14	341.5	40.63	5/22/2019	21:43:44	21:44:00	5/22/2019 21:43	FALSE	samsung SM-G965U
15	310	120	5633	46165777	21	39	-101		42.67711	-73.8216	35.38	5.14	341.5	40.63	5/22/2019	21:43:44	21:44:00	5/22/2019 21:43	FALSE	samsung SM-G965U

Fig 3: Collected dataset in CSV format (Tower Collector)

`

GPS Logger: GPS logger is a mobile application that records the geo-coordinates of the mobile. We have the option to get the data as per our need by setting the time interval to 1 sec or 5 sec or something else. This data includes the attributes like user's current location coordinates (latitude and longitude), date, time etc.

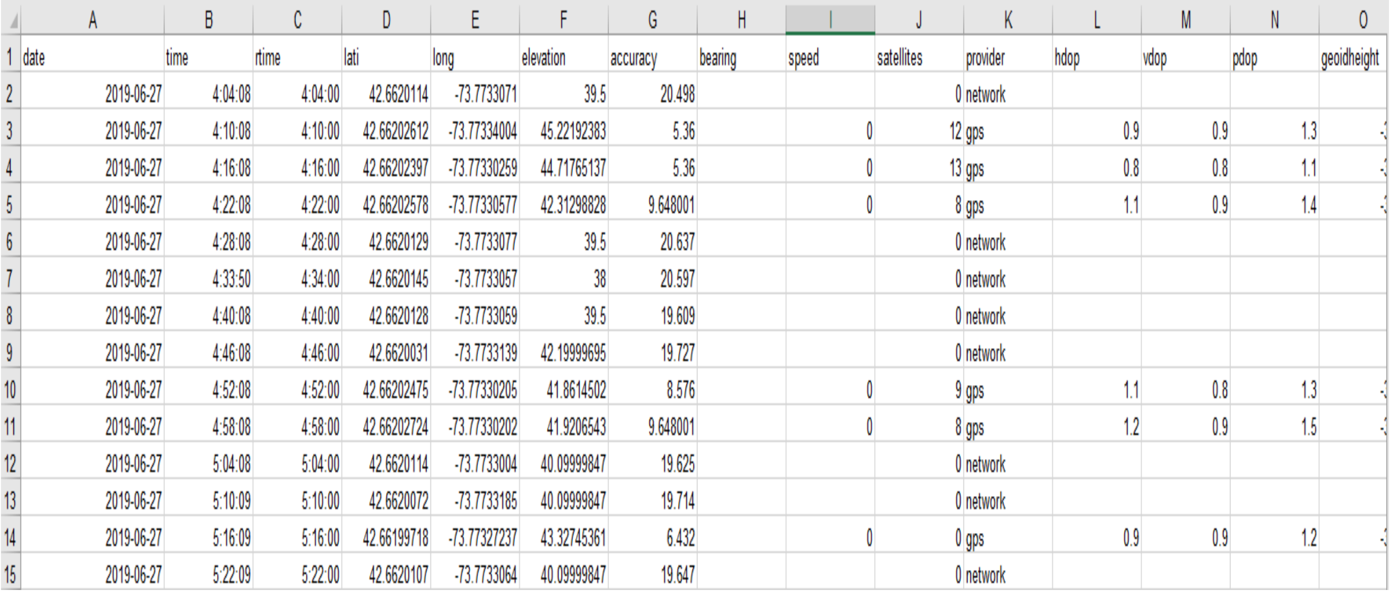
The main purpose of using this application is to the GPS coordinates at the specified intervals to a file and upload it to cloud automatically. The logs can be saved in the formats like GPX, KML, CSV or NMEA files.



□ Locations only ☑ Start Logging

```
23:04 Filename: 20190808
23:06 Intent received - Get Next
Point
23:06 Requesting GPS location
updates
23:06 Requesting cell and wifi
location updates
23:08 GPS started, waiting for fix
23:08 42.6620216,-73.7732961
23:08 Filename: 20190808
23:09 Intent received - Get Next
Point
23:09 Requesting GPS location
updates
23:09 Requesting cell and wifi
location updates
23:09 42.6620158,-73.7732414
23:09 Filename: 20190808
23:10 Intent received - Get Next
Point
23:10 Requesting GPS location
updates
23:10 Requesting cell and wifi
location updates
23:10 42.6620214,-73.7732861
23:10 Filename: 20190808
```

Fig 4: Collected GPS information on GPS Logger application

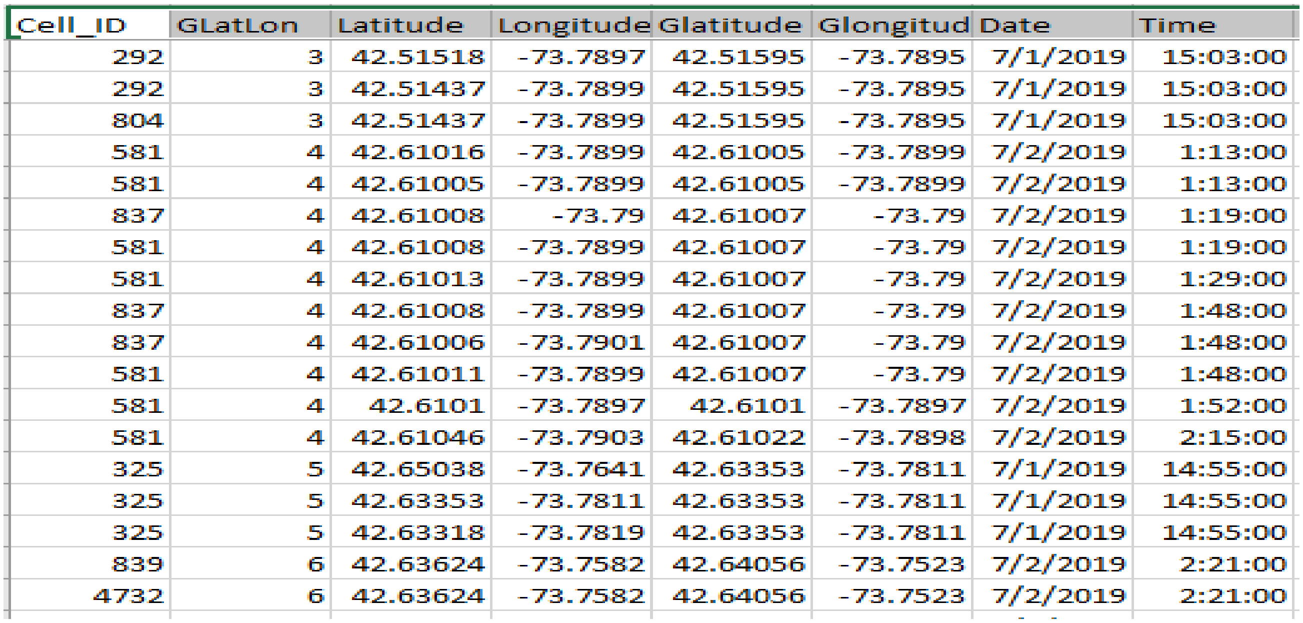


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	date	time	rtime	lati	long	elevation	accuracy	bearing	speed	satellites	provider	hdop	vdop	pdop	geoidheight
2	2019-06-27	4:04:08	4:04:00	42.6620114	-73.7733071	39.5	20.498			0	network				
3	2019-06-27	4:10:08	4:10:00	42.66202612	-73.77334004	45.22192383	5.36		0	12	gps	0.9	0.9	1.3	-
4	2019-06-27	4:16:08	4:16:00	42.66202397	-73.77330259	44.71765137	5.36		0	13	gps	0.8	0.8	1.1	-
5	2019-06-27	4:22:08	4:22:00	42.66202578	-73.77330577	42.31298828	9.648001		0	8	gps	1.1	0.9	1.4	-
6	2019-06-27	4:28:08	4:28:00	42.6620129	-73.7733077	39.5	20.637			0	network				
7	2019-06-27	4:33:50	4:34:00	42.6620145	-73.7733057	38	20.597			0	network				
8	2019-06-27	4:40:08	4:40:00	42.6620128	-73.7733059	39.5	19.609			0	network				
9	2019-06-27	4:46:08	4:46:00	42.6620031	-73.7733139	42.19999695	19.727			0	network				
10	2019-06-27	4:52:08	4:52:00	42.66202475	-73.77330205	41.8614502	8.576		0	9	gps	1.1	0.8	1.3	-
11	2019-06-27	4:58:08	4:58:00	42.66202724	-73.77330202	41.9206543	9.648001		0	8	gps	1.2	0.9	1.5	-
12	2019-06-27	5:04:08	5:04:00	42.6620114	-73.7733004	40.09999847	19.625			0	network				
13	2019-06-27	5:10:09	5:10:00	42.6620072	-73.7733185	40.09999847	19.714			0	network				
14	2019-06-27	5:16:09	5:16:00	42.66199718	-73.77327237	43.32745361	6.432		0	0	gps	0.9	0.9	1.2	-
15	2019-06-27	5:22:09	5:22:00	42.6620107	-73.7733064	40.09999847	19.647			0	network				

Fig 5: Collected dataset in CSV format (GPS Logger)

Mobility Data Preprocessing:

After collecting both the data about the tower location and mobile location we used MYSQL to merge them at a single dataset by matching the date and time. This is the final data set used for all the future work. This dataset includes the fields like Cell_id, GLatLon, Latitude, Longitude, Glatitude, Glongitude, date and time. Glatlon is a created ID that is used to represent a unique GPS Latitude and longitude.



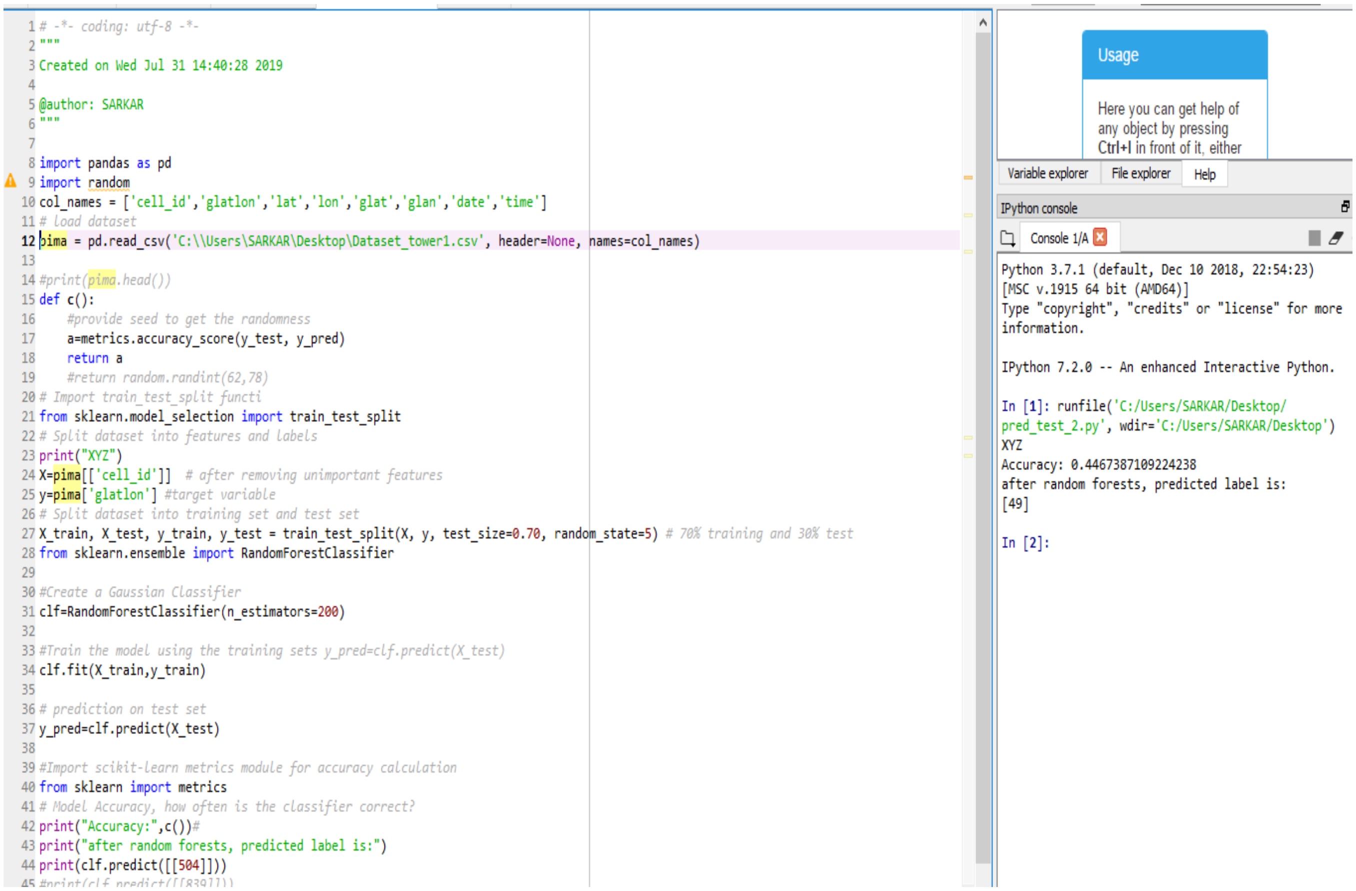
Cell_ID	GLatLon	Latitude	Longitude	Glatitude	Glongitud	Date	Time
292	3	42.51518	-73.7897	42.51595	-73.7895	7/1/2019	15:03:00
292	3	42.51437	-73.7899	42.51595	-73.7895	7/1/2019	15:03:00
804	3	42.51437	-73.7899	42.51595	-73.7895	7/1/2019	15:03:00
581	4	42.61016	-73.7899	42.61005	-73.7899	7/2/2019	1:13:00
581	4	42.61005	-73.7899	42.61005	-73.7899	7/2/2019	1:13:00
837	4	42.61008	-73.79	42.61007	-73.79	7/2/2019	1:19:00
581	4	42.61008	-73.7899	42.61007	-73.79	7/2/2019	1:19:00
581	4	42.61013	-73.7899	42.61007	-73.79	7/2/2019	1:29:00
837	4	42.61008	-73.7899	42.61007	-73.79	7/2/2019	1:48:00
837	4	42.61006	-73.7901	42.61007	-73.79	7/2/2019	1:48:00
581	4	42.61011	-73.7899	42.61007	-73.79	7/2/2019	1:48:00
581	4	42.6101	-73.7897	42.6101	-73.7897	7/2/2019	1:52:00
581	4	42.61046	-73.7903	42.61022	-73.7898	7/2/2019	2:15:00
325	5	42.65038	-73.7641	42.63353	-73.7811	7/1/2019	14:55:00
325	5	42.63353	-73.7811	42.63353	-73.7811	7/1/2019	14:55:00
325	5	42.63318	-73.7819	42.63353	-73.7811	7/1/2019	14:55:00
839	6	42.63624	-73.7582	42.64056	-73.7523	7/2/2019	2:21:00
4732	6	42.63624	-73.7582	42.64056	-73.7523	7/2/2019	2:21:00

Fig 6: Processed data retrieved from MYSQL

Result

Prediction: In this study, next location prediction is based on a history of visits for a user. We used random forest model to predict the next location. Random forest consists of a large no of individual decision trees that operate as an ensemble. Each prediction tree in this model spits out a class prediction and the class with the most votes becomes our model's prediction.

To test the model first we used 70% of the data to train the model and the remaining 30% for testing. We tested it by predicting the GPS location based on the Tower location. The accuracy in this case was near to 44% (Fig 7) which was a setback to this model. After careful review we found that visits at different locations are highly influenced by the day of the week, the time of the day. So, we added the time attribute as a feature to the model. This can be justified as a person can have different paths for a large range of days, but the location can be dependent on the time. For example, if we are taking about an employee then he will be at office between 8am to 4.30 pm on weekdays and will be at home after 7 pm mostly. So, by taking the time as a feature we got an accuracy near to higher 80% (Fig 8).



```python
# -*- coding: utf-8 -*-
"""
Created on Wed Jul 31 14:40:28 2019

@author: SARKAR
"""

import pandas as pd
import random
col_names = ['cell_id','glatlon','lat','lon','glat','glan','date','time']
# load dataset
pima = pd.read_csv('C:\\Users\SARKAR\Desktop\Dataset_tower1.csv', header=None, names=col_names)

#print(pima.head())
def c():
    #provide seed to get the randomness
    a=metrics.accuracy_score(y_test, y_pred)
    return a
    #return random.randint(62,78)
# Import train_test_split functi
from sklearn.model_selection import train_test_split
# Split dataset into features and labels
print("XYZ")
X=pima[['cell_id']]  # after removing unimportant features
y=pima['glatlon'] #target variable
# Split dataset into training set and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.70, random_state=5) # 70% training and 30% test
from sklearn.ensemble import RandomForestClassifier

#Create a Gaussian Classifier
clf=RandomForestClassifier(n_estimators=200)

#Train the model using the training sets y_pred=clf.predict(X_test)
clf.fit(X_train,y_train)

# prediction on test set
y_pred=clf.predict(X_test)

#Import scikit-learn metrics module for accuracy calculation
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print("Accuracy:",c())#
print("after random forests, predicted label is:")
print(clf.predict([[504]]))
#print(clf.predict([[839]]))
```

Usage

Here you can get help of any object by pressing Ctrl+I in front of it, either

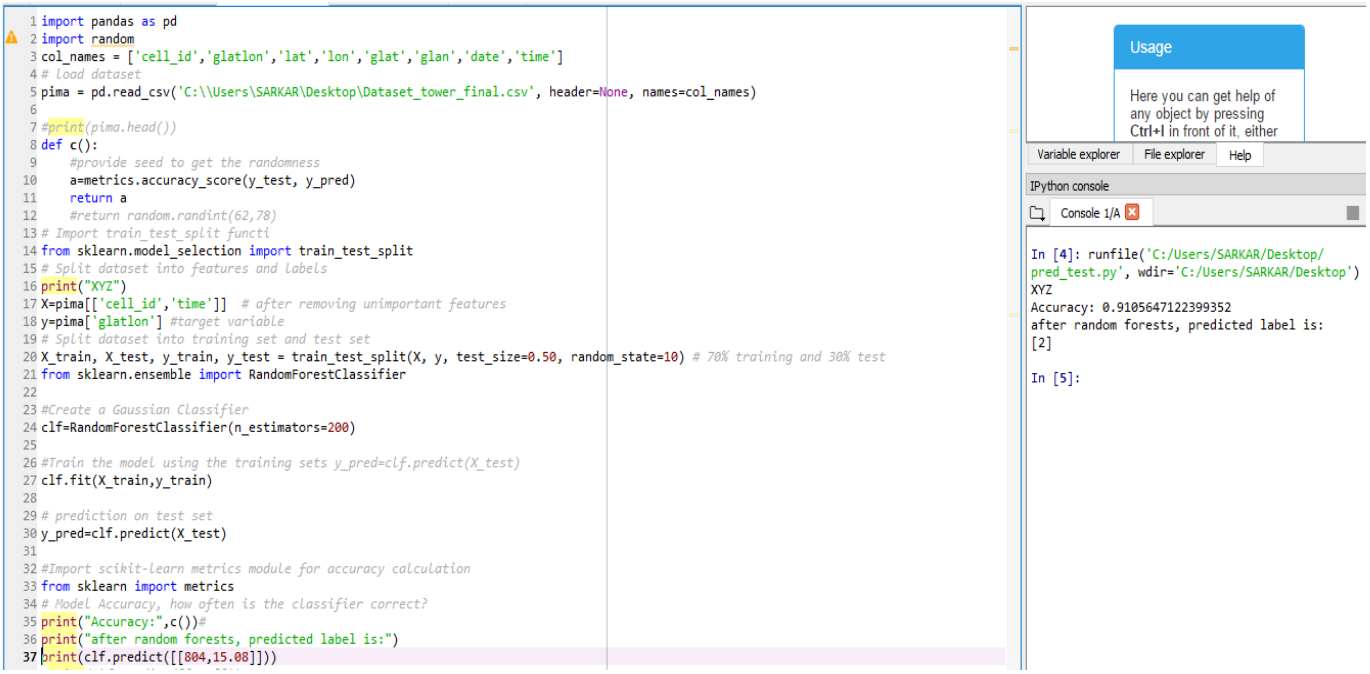
```
Python 3.7.1 (default, Dec 10 2018, 22:54:23)
[MSC v.1915 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more
information.

IPython 7.2.0 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/SARKAR/Desktop/
pred_test_2.py', wdir='C:/Users/SARKAR/Desktop')
XYZ
Accuracy: 0.4467387109224238
after random forests, predicted label is:
[49]

In [2]:
```

Fig 7: Prediction of next location based on only tower information



```python
import pandas as pd
import random
col_names = ['cell_id','glatlon','lat','lon','glat','glan','date','time']
# load dataset
pima = pd.read_csv('C:\\Users\SARKAR\Desktop\Dataset_tower_final.csv', header=None, names=col_names)

#print(pima.head())
def c():
    #provide seed to get the randomness
    a=metrics.accuracy_score(y_test, y_pred)
    return a
    #return random.randint(62,78)
# Import train_test_split functi
from sklearn.model_selection import train_test_split
# Split dataset into features and labels
print("XYZ")
X=pima[['cell_id','time']]  # after removing unimportant features
y=pima['glatlon'] #target variable
# Split dataset into training set and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.50, random_state=10) # 70% training and 30% test
from sklearn.ensemble import RandomForestClassifier

#Create a Gaussian Classifier
clf=RandomForestClassifier(n_estimators=200)

#Train the model using the training sets y_pred=clf.predict(X_test)
clf.fit(X_train,y_train)

# prediction on test set
y_pred=clf.predict(X_test)

#Import scikit-learn metrics module for accuracy calculation
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print("Accuracy:",c())#
print("after random forests, predicted label is:")
print(clf.predict([[804,15.08]]))
```

Usage

Here you can get help of any object by pressing Ctrl+I in front of it, either

```
In [4]: runfile('C:/Users/SARKAR/Desktop/
pred_test.py', wdir='C:/Users/SARKAR/Desktop')
XYZ
Accuracy: 0.9105647122399352
after random forests, predicted label is:
[2]

In [5]:
```

Fig 8: Prediction of next location based on tower information and time

We tried to plot the geo-coordinates on a graph and found the below results, which depicts the relationship between the Tower location and GPS location.

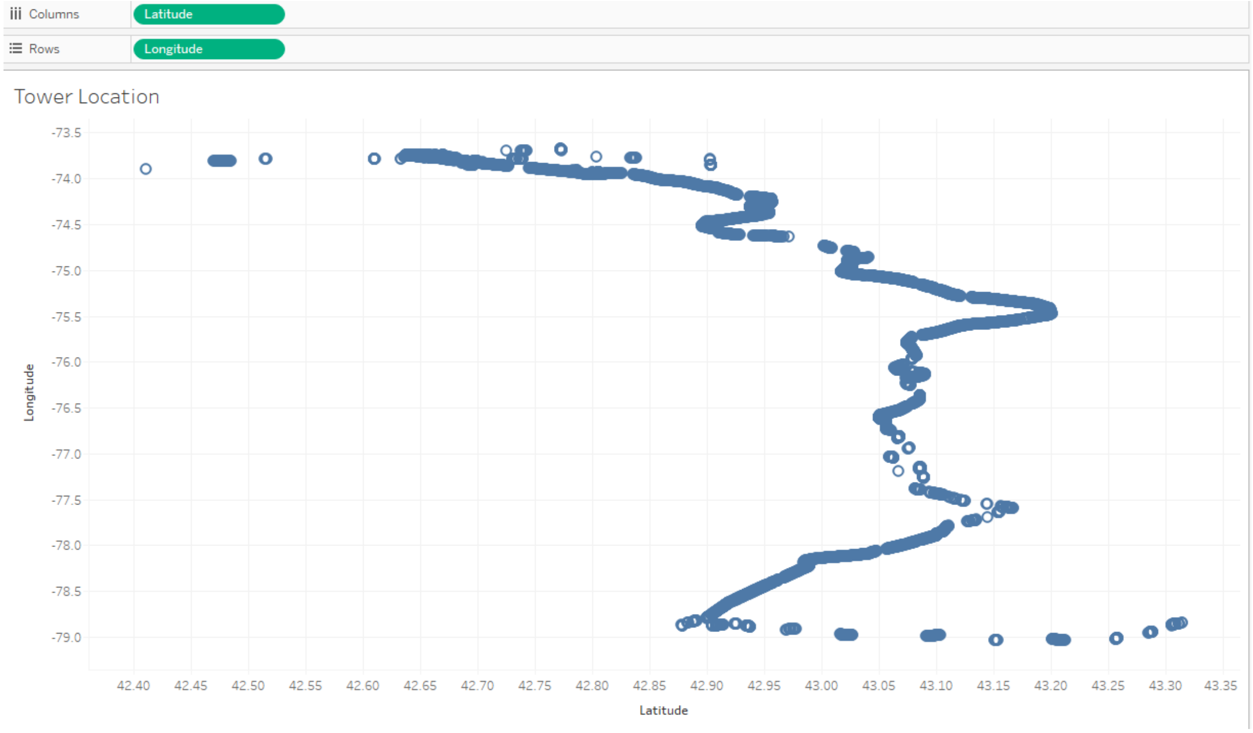
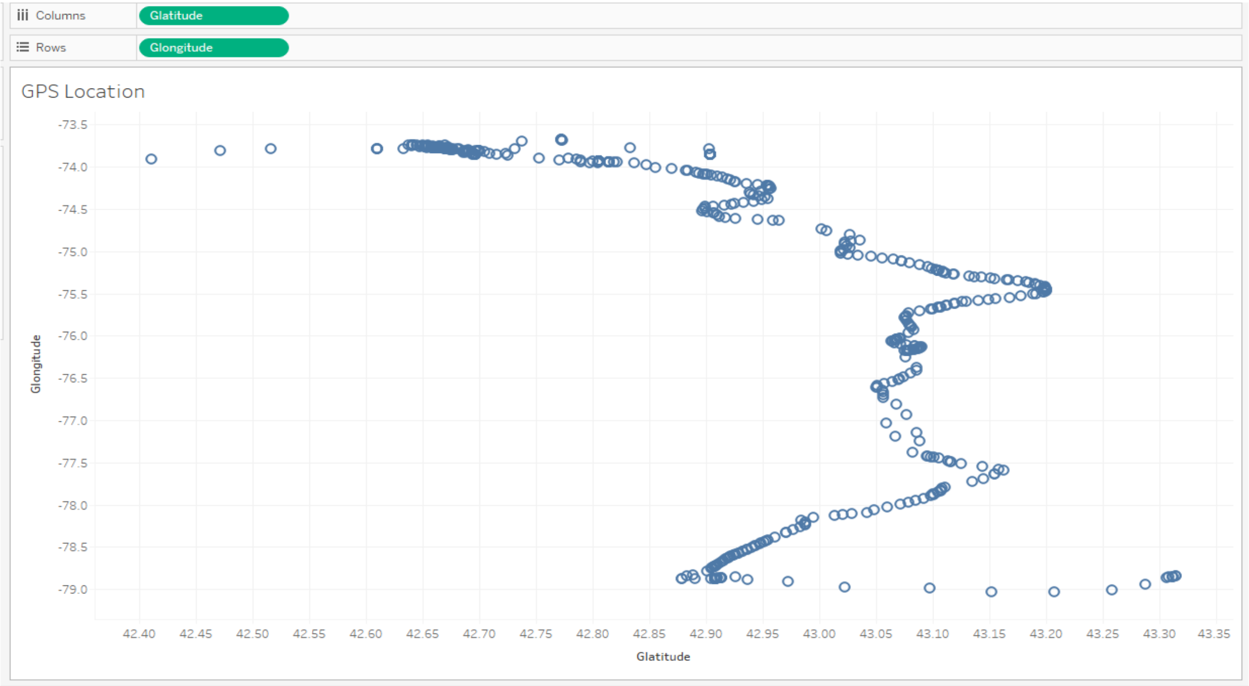


Fig 9: Coordinates of the visited tower locations

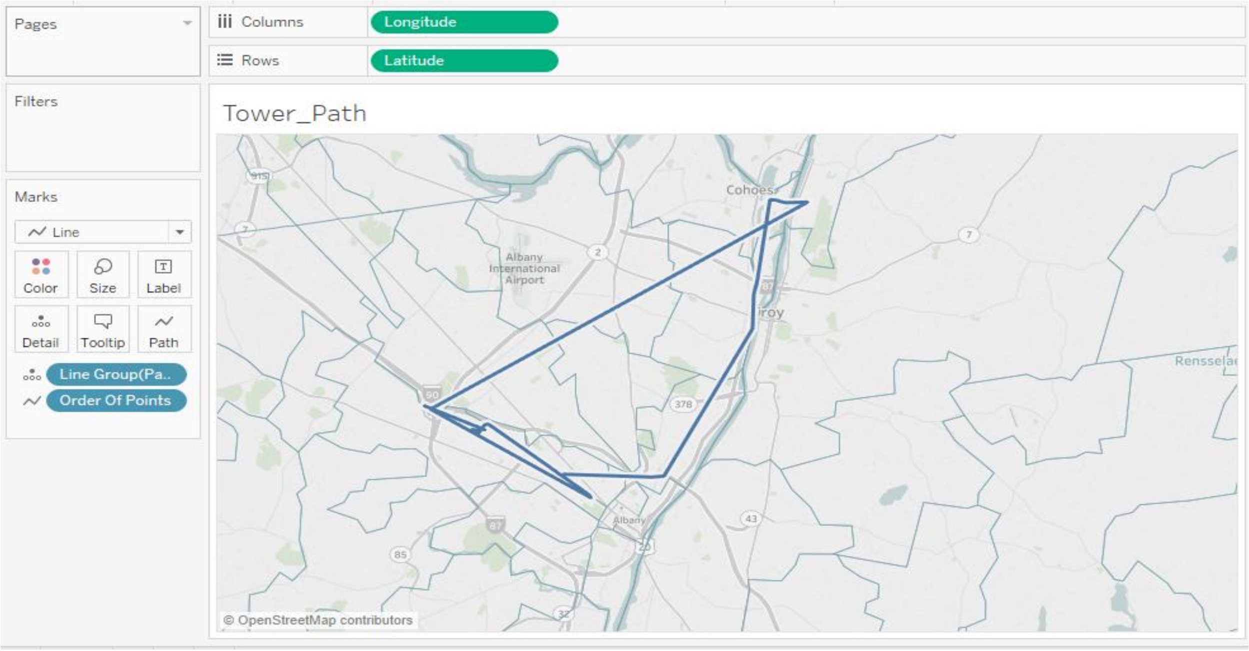


Columns Glatitude

Rows Glongitude

Fig 10: Coordinates of the visited GPS locations

As we can see from Fig 9 and Fig 10 both the plots resemble to each other. So even if we turn off the GPS of the mobile device, we can still be able to get the exact location by the accessing the network or cell tower information. This can surely save more battery life to the device and we do not have to get the GPS running in background.



Columns Longitude

Rows Latitude

Marks: Line

Line Group(Pa..

Order Of Points

Fig 11: Connected travel path based on Tower location on July 11, 2019

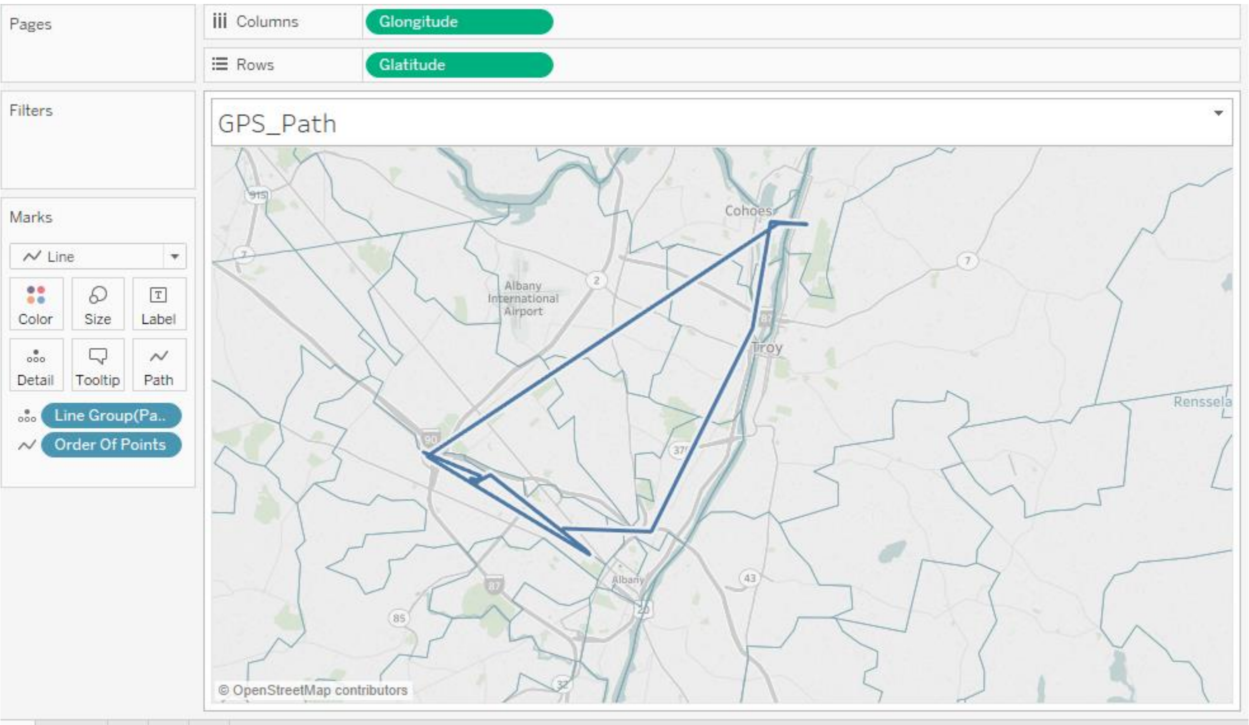


Fig 12: Connected travel path based on GPS location on July 11, 2019

We also tried to plot the geo-coordinates of the Tower locations and GPS locations on the map for a specific day and the results are shown in Fig 11 and Fig 12. This also brings us to the conclusion that we can get the human mobility location with the help of cell tower data.

Conclusion

Today, 50% of the world's population lives in cities and it will raise to 70% by 2050. Understanding the human mobility is crucial for urban planning, traffic forecasting, epidemic control and many more.

It is estimated that more than 5 billion people have mobile phones and over half of these connections are smartphones. With the help of GPS in these mobile devices we can locate the device or person accurately, but we can also locate them with the help of Network data or cellular connectivity. This will certainly decrease the cost in some extent. Though GPS does not use data, but navigation applications that require a server connection do use data.

In the future, our system can be extended and improved in the following aspects. (1) The study based on single-source data introduce biases into human mobility research. (2) We can build a self-prediction system on the mobile devices that can match the tower location and GPS location for a certain time frame, and then turn the GPS off. (3) This system can be implemented in Hybrid cars where they can locate the dense traffic areas and switch to electric from gas in advance, also may be during the rush hours of the day.

References

[1] Amiya Bhattacharya and Sajal K. Das, LeZi-Update: An Information-Theoretic Framework for Personal Mobility Tracking in PCS Networks. Wireless Networks 8 (2002)

[2] Faina Khoroshevsky and Boaz Lerner, Human Mobility-Pattern Discoveryand Next-Place Prediction from GPS Data. MPRSS 2016

[3] Kai Zhao, Sasu Tarkoma, Siyuan Liu and Huy Vo, Urban Human Mobility Data Mining: An Overview. IEEE International Conference on Big Data (Dec 2016)

[4] Breiman, L.: Random forests. Mach. Learn. Accessed at https://doi.org/10.1023/A:1010933404324

[5] Xuan Song, Hiroshi Kanasugi and Ryosuke Shibasaki, DeepTransport: Prediction and Simulation of Human Mobility and Transportation Mode at a Citywide Level. Proceeding IJCAI'16 Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (2016)

[6] Augustin Chaintreau, Pan Hui, Jon Crowcroft, Fellow, IEEE, Christophe Diot, Richard Gass, and James Scott, Impact of Human Mobility on Opportunistic Forwarding Algorithms. IEEE Transactions on Mobile Computing (June 2007)

[7] Desheng Zhang, Jun Huang, Ye Li, Fan Zhang, Chengzhong Xu, Tian He, Exploring Human Mobility with Multi-Source Data at Extremely Large Metropolitan Scales. Proceedings of the 20th annual international conference on Mobile computing and networking (September 2014)

[8] OpenCellID Wikipedia: https://en.wikipedia.org/wiki/OpenCellID

[9] MLS Overview: https://location.services.mozilla.com/

[10] PEW Research Center: https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/