UNIVERSITY OF TARTU
Institute of Computer Science
Software Engineering Curriculum

Mubashar Shahzad

# Revisiting Group Mobility Modelling: A Systematic Evaluation

Master Thesis (30 EAP)

*Supervisor: Huber Flores*

TARTU, 2021

# Abstract

While human mobility modeling has been studied extensively over the years, there is still a partial understanding of the modeling of users moving together. In this thesis, we develop a method that can be used to build trajectories of users from mobile crowdsensed data. By using this method, we characterize different types of trajectories. We then perform rigorous experimental benchmarks to compare and analyze these trajectories in a systematic manner. Our results demonstrate that the optimal similarity score between trajectories is found when considering trajectories of the same type and a short length. In addition, our results also provide new insights into the level of (partial) similarity that can be found between users that move together.

**CERCS:** P170 Computer science, numerical analysis, systems, control

**Keywords:** group mobility, human mobility, mobile crowd-sensing, similar trajectory, mobility comparison

**Grupimobiilsuse modelleerimise ülevaatamine: süstemaatiline hindamine**

**kokkuvõte:** Inimeste mobiilsuse modelleerimist on läbi aastate põhjalikult uuritud, kuid siiski on arusaamine kasutajate koos liikumisest mittetäielik. Käesolevas lõputöös arendatakse välja meetod, millega saab mobiililt kogutud andmete põhjal kasutajate trajektoore moodustada. Meetodil põhinedes kirjeldame erinevaid trajektooride tüüpe. Seejärel teostame eksperimentaalseid mõõtlusi, et süsteemselt neid trajektoore võrrelda ja analüüsida. Meie tulemused näitavad, et optimaalseima trajektooride sarnasusskoori annab sama

tüüpi ja lühikese pikkusega trajektooride võrdlus. Lisaks pakuvad meie tulemused uut arusaama sellest, kuivõrd on võimalik (osalist) sarnasust leida kasutatajate puhul, kes liiguvad koos.

**CERCS:** P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine

**Märksõnad:** rühmaliikuvus, inimeste liikuvus, rahva mobiliseerimine teabe kogumiseks (ingl k crowdsensing), sarnane trajektoor, liikuvuse võrdlus

# Contents

# List of Figures

## LIST OF FIGURES

# List of Tables

## LIST OF TABLES

# 1

# Introduction

Human mobility has become a fundamental domain to understand people's behaviour, social relations, and interactions (1). By understanding the principles that govern human mobility, it is possible to design better digital systems and applications that blend into everyday activities. Domains that benefit from human mobility research include intelligent transportation systems (2), capacity planning in mobile networks (3) and autonomous vehicles (4, 5). More recently, fundamental principles of human mobility have been utilized to evaluate the spread of infections in urban areas, such that it is possible to design counter measurements against it (6), e.g. COVID-19. Likewise, vaccination planning also has been aided by human mobility patterns as a way to accelerate the reduction of infection rates in large populations (7).

Existing works on understanding human mobility have focused on modelling trajectories through users' personal data (8), e.g., surveillance cameras (9). Thanks to the increasing adoption of smart, IoT (Internet of Things), and wearable devices (10), it is possible to obtain enough spatial and temporal data that can be used to model the behaviour of users. Indeed, several techniques that exploit data collected from devices and applications can be used for this purpose. For instance, network probes from the mobile operator can capture the behaviour of application usage and visited locations from mobile subscribers (11). Another example, apps can be instrumented with mechanisms to triangulate position using WiFi connectivity (12). Similarly, urban locations can be fingerprinted based on their POI (Point of Interest) area distribution, and user mobility can be extrapolated by monitoring when users check-in in those areas. While

several methods can be utilized to analyze human mobility through trajectories, it is still challenging to identify when multiple users move together.

In this thesis, we revisit the modelling of human mobility to detect users that move together in a group. We first develop a generic modelling method in which trajectories are built by looking at samples collected in different locations. We rely on samples collected by a mobile operator network in Shanghai, China[1]. We next implement a DTW algorithm that can be used to calculate a similarity score between different trajectories. With this information, a systematic evaluation was conducted in which different types of trajectories were compared. Our results indicate the groups of users that move (or stay) together (partially) are more likely to be found when the trajectories are short and similar in length. Our work paves the way towards new methods to analyze human mobility as a group.

## 1.1 Contributions

The following sums up the contributions presented in this work:

- **Novel insights:** We develop a grid-based method that can be used to built trajectories from mobile crowdsensed data of users. By using this method over a mobile operator dataset, we found that it is possible to find users that move together by looking at trajectories that are short in length.
- **Improved performance:** We perform rigorous benchmarks demonstrating that our method can be applied over a large variety of situations and experimental conditions. The source code used in our benchmarks is available as open-source in GitHub [2].

## 1.2 Outline

This thesis is structured as follows:

- Chapter 2 reviews the state-of-the-art about human mobility and it's applications, mobile crowdsensing and metrics to analyze the similarly of trajectories.

[1] Partially released in Applens workshop 2019
[2] https://github.com/mobile-cloud-computing/GroupMobilityRevisited

- Chapter 3 describes the information about our raw dataset, and the processes used for cleaning and preparing the data for analysis.
- Chapter 4 describes the method for building trajectories from our crowdsensed dataset.
- Chapter 5 describes the experimental setup for analyzing similarity between trajectories.
- Chapter 6 presents the findings our analysis.
- Chapter 7 discusses the implications and limitations of our work.
- Chapter 8 presents the summary and conclusion of our work.

# 2

# State of the Art

The primary focus of this Chapter is to explain the concepts of human mobility, user trajectory modelling and similarity of trajectories. In the following, we explain these concepts in detail and review existing work in these fields.

## 2.1 Human mobility

**Background:** Human mobility patterns have been studied widely in recent years (13). Human mobility refers to the analysis of the locomotion behaviour of users in general - including transportation systems. As shown in Figure 2.1, thanks to the proliferation of smart devices and smart infrastructure, it is very easy to collect data that can be used to understand human mobility behaviours in urban environments (14, 15). For instance, bikes in a city can be used to optimize routes and touristic activities of visitors (16). Human mobility is mostly explored through the modelling of trajectories. However, other studies have taken different perspectives by analyzing the dynamics of crowd estimation and formation. For instance, amount of people available in a particular location (14, 17).

Interestingly, human mobility has shown to be predictable (1). As a result, prediction models have been studied to anticipate the in-flux and out-flux of people in specific urban locations (18). Moreover, as the current (COVID-19) pandemic has demonstrated, these prediction models are particularly important to regulate the interactions and encounters of people in locations, such that it is possible to reduce infection and apply better vaccinations policies. Similarly, other applications of these models include

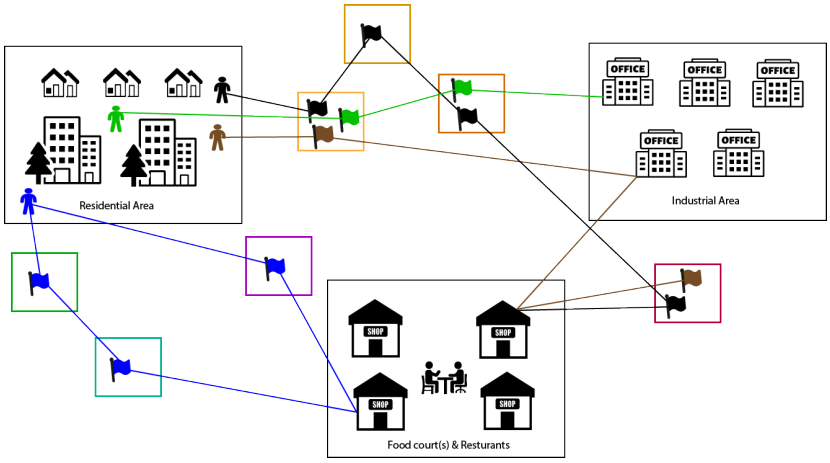



**Figure 2.1:** Human mobility in urban areas by several individual users

capacity estimation of users for the deployment of services, and smart city designs, among others (19).

**Existing work:** Several works have investigated the analysis of human mobility patterns. For instance, a framework (DRoF) has been developed to discover regions of different topics using point of interests in that region along with human mobility. Likewise, distribution of mobility patterns and identifying the intensity of distribution among different locations helps in urban planning, advertisement and business location identification (15). Different models also have been developed to find the migratory flows, traffic forecasting and urban planning using human mobility (individual and population) concerning short-range and long-range mobility types (20).

Additionally, other works have also investigated further relations between urban regions and mobility patterns of different communities. This interrelationship between spatial distribution and mobility patterns helps to figure out the density and evolution of communities over time (21). Besides this, human mobility research also has been applied to optimize resource management of services (22) and support the design and deployment of smart buildings (23) for city-scale planning. *In contrast to previous work, we study human mobility patterns in urban areas that are captured through mobile operator networks. We model trajectories of users with this data, such that it is possible to find users moving together.*

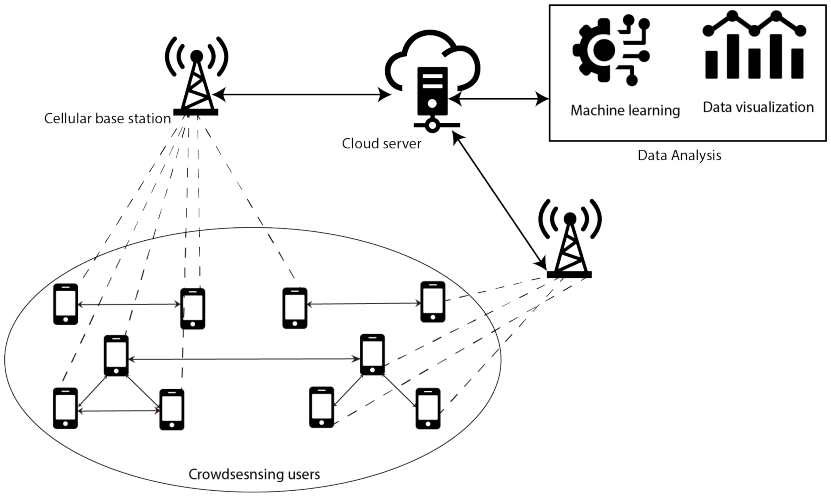



**Figure 2.2:** Information flow of mobile crowdsensing system. Smartphones conduct sensing tasks and subsequently transmit that data to a data collecting server over cellular networks.

## 2.2 Mobile crowdsensing

**Background:** Mobile crowdsensing (MCS) is a technique that can be used to reduce the complexity of sensing tasks into (easy) sub-tasks that can be distributed between multiple devices (24). As shown in Figure 2.2, various devices connect opportunistically to base stations (in range), sending (collected) samples to remote central locations (servers), in which data is aggregated and analyzed. Data analytics over the crowdsensed data involve digging and extracting human aspects and behavioural patterns that can be used to improve features of systems and applications (25), e.g., environmental, infrastructure, and socialization applications cases (24, 26). In addition, strategic implementation of data collected by sensing devices over time, e.g., traffic sensors, can help in continuous monitoring of users in large scale areas, such as cities (27). The main limitations of mobile crowdsensing are low quality of data, and high spatial and temporal data characteristics, which can make data hard to analyze.

**Existing work:** Mobile crowdsensing (MCS) has been applied not just for modelling human mobility but also for a large diversity of other cases. Approaches for sensing crowdsensed data (28) helped in developing health monitoring systems with a positive impact on quality and considerably lower cost (29). Several works investigated various

sensing techniques from the devices having multiple sensing capabilities (30). For instance, the construction of complex indoor structures is built with the help of mobile crowdsensing (31) that powered indoor navigation methods (32). Likewise, other works rely on MCS to update digital maps with the incorporation of AI (Artificial Intelligence) techniques (33). Similarly, MCS has been widely used in many Internet of Things(IoT) applications (34, 35).

In terms of human behaviour, several studies have investigated the use of MCS techniques for providing incentives to users, e.g., perks and benefits like free tickets to increase tourism in groups and individual tours (36, 37). Architectures for mobile crowdsensing also have been studied in the literature (34), including methods for data collection, communication media, data aggregation/fusion, as well as feature extraction and classification.

Since a key challenge for mobile crowdsensing is related to data sparsity, several works have proposed methods to improve sparsity and the quality of collected data. However, the sparsity challenge is a rather complex matter that comes with the cost of losing valuable information (38) when applying algorithms for augmenting quality properties of the data (39). Likewise, other studies suggested the use of imputation of missing values with the mean (40). Few other imputation techniques are the part of researches that includes data deletion, median imputation, KNN (k-nearest neighbours) imputation, and mean imputations (41). Imputation methods are more effective than deletion, and KNN imputations perform better than other techniques. However, research showed the use of machine learning could save time, thus lowering the cost of analysis (42).

*Unlike existing works in mobile crowdsensing, we rely on a crowd sensed dataset to model human mobility through the characterization of trajectories. With this information, we then analyze the optimal factors that need to be considered in trajectories to identify users that move together.*

## 2.3 Trajectory similarity analysis

**Background:** A trajectory is the user's footprints while moving from one point to another. One user can make multiple trajectories based on various day hours depicting

the mobility needed in every hour. A similarity score refers to the quantification of differences between two individual trajectories - depicted as signals or time-series, among others. A similarity score is used to establish similarity between trajectories. Several basic methods are available to calculate a similarity score, e.g., Euclidean distance. More sophisticated algorithms and techniques also have been developed to quantify such similarities. One such method is Dynamic Time Warping (DTW) (43). Primarily it was introduced for comparing time series (44). However, it was altered as per the demands of the problem, such as speech recognition (45) and software development (46). Furthermore, modified DTW included efficient searching to improve warping window distortion (47), optimization for distance calculations (48) and fault detection in nuclear power plants (49). DTW proved to be profitable in various domains like medical (50) (51), machine learning (52), data mining (53), intelligent transport system (54), speaker recognition (55) and urban planning (56). As a result, DTW has suffered many optimizations over the years, and recent improvements have been developed following its principles, e.g., ULTRAFASTWWSEARCH algorithm (57). Other methods and metrics that can be used to calculate similarity on trajectories include TrajectoryMatch Algorithm, EditDistance and LongestCommonSubsequence approaches (58). Unlike other static algorithms, we used DTW in our work as DTW allows dynamic comparison of trajectories with different spatial and temporal characteristics.

**Existing work:** Previous researches have demonstrated the drastic growth of data in volumes and velocity in recent years (59, 60). Opportunistic data collected from different sources can be used to model and capture human behaviour in terms of trajectories and behavioural patterns (61, 62, 63). Trajectory predictions caught attention of researchers to predict based on previous mobility patterns (64, 65). Trajectories possess temporal and spatial behaviours, and researches were conducted on both properties for various purposes. Spatial properties have been studied for multiple contributions (66). Moreover, researchers analyzed trips for individual users, as shown in Figure 2.1, for predictions (13). These pattern recognition systems rely not only on spatial characteristics but also on temporal behaviour (67) to study time divisions of the day, cost (68) and the tasks. However, studies showed the effect of factors like the weather could readily affect the temporal behaviour (69). In addition to the studies on predicting mobilities and analysis based on trajectories, group mobilities was also that part of studies on the

network level. Researches showed various group mobility studies on network infrastructures like Wireless Ad-hoc Network (WANET)[1] or Mobile Ad-hoc Network (MANET). Based on the network, few models like Reference Point Group Mobility (RPGM) (70) and Reference Velocity Group Mobility Model (RVGM) (71) were developed that took other factors like distance, velocity, and acceleration for predictions. Such methods were studied and compared in various researches (72).

*Unlike previous studies, we model trajectories of human mobility using a grid-like method that combines the deployment of bases stations. We then calculated the similarity scores of these trajectories using FastDTW (73). By arranging and sorting the trajectories using the similarity metrics, we then identify users that move together in our dataset.*

## 2.4 Summary

We performed a literature review on the domains of human mobility, mobile crowdsensing and similarity metrics of trajectories. We described and discussed these topics in depth and explored the existing works conducted in each of these domains. Lastly, we also stated the difference of our proposed work when compared with existing research.

In the next Chapter, we explain the dataset and the process applied for data cleaning and preparation prior to the estimation of trajectories using crowd sensed data.

[1] https://www.tech-term.in/2018/03/08/wanet/

# 3

# Dataset Description and Preparation

This Chapter provides an overview of the dataset used in our analysis. The significant aspects and preparation of the dataset are depicted, followed by in-depth insight based on selected criteria.

## 3.1 Dataset overview

We have used a pre-existing raw dataset from a mobile operator. The data consists of a significant number of records captured through the cellular network. This data was collected from the users connected to the internet on 21 August 2017 from the Shanghai district in China. The dataset covers the record for the entire day (24 hours) whenever a user is connected through mobile data. Multiple entries are recorded for each user during the use of multiple applications. Each of the applications has a unique identifier that enables us to distinguish between various mobile applications. Furthermore, the number of applications used by different users can also be retrieved from the dataset.

Every single record in the collected dataset contains the following information:
**Anonymized User Id** – an identifier for anonymous users
**Timestamp** – the time when specific application was connected to the internet
**Base Station Id** – an identifier for the base stations in Shanghai
**Traffic Volume** – the amount of data transferred during the session

**Table 3.1:** Dataset summary

| Dataset | Records | Stations | Users | Apps |
|---|---|---|---|---|
| Raw | 12310705 | 7663 | 998 | 1543 |
| After Data Cleaning | 7675606 | 3720 | 812 | 1444 |
| Interested Area X | 4236931 | 2314 | 570 | 1322 |
| After validation of X | 3160179 | 2271 | 529 | 515 |

**App Id** – an identifier for apps

**App Name** – application name associated with the App Id

**Latitude** – latitude associated with base station id

**Longitude** - longitude associated with base station id

### 3.1.1 Base stations verification

The raw dataset contains a total of 12310705 data points from 998 different users. Total applications used in the dataset are 1543 which are connected with 7663 base stations. The collected information consists of the different times during the day when the user connects to the tower (base station), hence providing footprints of the usage continuously.

In a developing city (like Shanghai), it is to be noted that the location of the base stations is not constant over time and is prone to distinct migrations due to network optimization and better coverage deployment. In order to correspond it with the current year, it was essential to verify the correctness of base stations locations for analyzing the data correctly. Hence before the start of the analysis, we have verified the base stations from OpenCelliD [1]. This allows us to verify that our data is representative for capturing human mobility using cellular network deployment.

The OpenCelliD consists of information about cellular data and is the world's largest repository for the GPS positions of towers. OpenCelliD provides a plethora of information regarding different network types like GSM, CDMA, or LTE running in a specific

[1] https://www.opencellid.org/

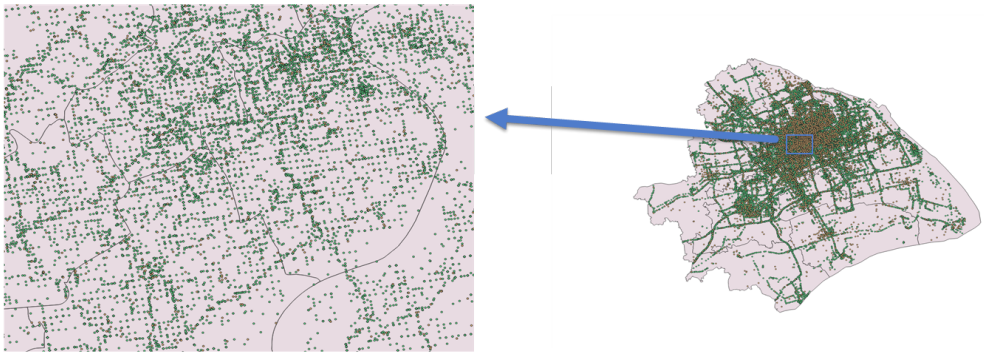


**Figure 3.1:** Overlapping view of coordinates of OpenCelliD and data

region like the USA, Germany, or any other countries. This project is open source, and redistribution and usage of data are granted with the reference.

For verification of data from OpenCelliD, we matched the base stations from OpenCelliD with the dataset's base station, but the results were surprisingly not sufficient as only a few of the stations matched precisely with the same station id. Furthermore, we utilized the GPS coordinates and mapped both datasets over the shanghai district geographically for visually inspecting the data. It can be seen from the Figure 3.1 that there is minimal overlap between the cells from the datasets (OpenCelliD and the raw data) and depicts that it cannot be considered for the analysis without manipulations.

Hence, in order to mitigate the above-mentioned issue, we draw radial distances of different sizes around the geolocated stations to verify our collected data from the OpenCelliD dataset. We have considered a 500-meter radial distance for verification of data. Other radial distances and verified station count are mentioned in the Table 3.2.

After cleaning the base stations, we have filtered 4236931 records from 3720 base stations. The total number of filtered users is 812, and the unique number of applications used is 1444. The filtered records are around 35 percent of the actual records of data.

## 3.2 Spatio-temporal characteristics

After verifying the base stations as described in the previous section, we have analyzed the spatial and temporal characteristics of the data. We validated the spatial characteristics by drawing the land and water use over the Shanghai district. The dataset used

**Table 3.2:** Radial distances considered for verification and verified base stations inside the distance

| Distance (in meters) | Verified Stations Count |
|---|---|
| **500** | **3720** |
| 700 | 3898 |
| 1000 | 4055 |
| 1500 | 4193 |
| 2000 | 4259 |
| 2500 | 4316 |
| 3000 | 4347 |
| 3500 | 4367 |
| 4000 | 4381 |

for mapping the land use over the map was published in July 2021 [1].

The map file for land use/land cover (LULC) is for the year 2020 with a high resolution of 10m. The interesting fact about the map data is that it is global and up to date. European Space Agency (ESA) Sentinel-2 satellite imagery developed the dataset by classifying different land covers using machine learning models and is scaled using Microsoft Azure Batch (74).

Figure 3.2 shows that most of the developed part of Shanghai is on the north side, and it covers a significant portion of the network area. This visualization assures sparsity issues in the data and reveals that the distribution of stations is not uniform over the Shanghai district.

To verify the users' sparsity, we took all verified stations and plotted them over the district. Figure 3.3a shows all verified stations, and it can be visualized that few areas are denser than others giving a solid argument for unequal distribution of network coverage.

The same pattern of network development can be understood from the Figure 3.2. Figure 3.2 shows the satellite data of the shanghai district, which focuses on the development of urban areas of the district. The satellite data shows more developed regions to the north, similar to what we observed from LULC data.

[1] https://www.arcgis.com/home/item.html?id=d6642f8a4f6d4685a24ae2dc0c73d4ac

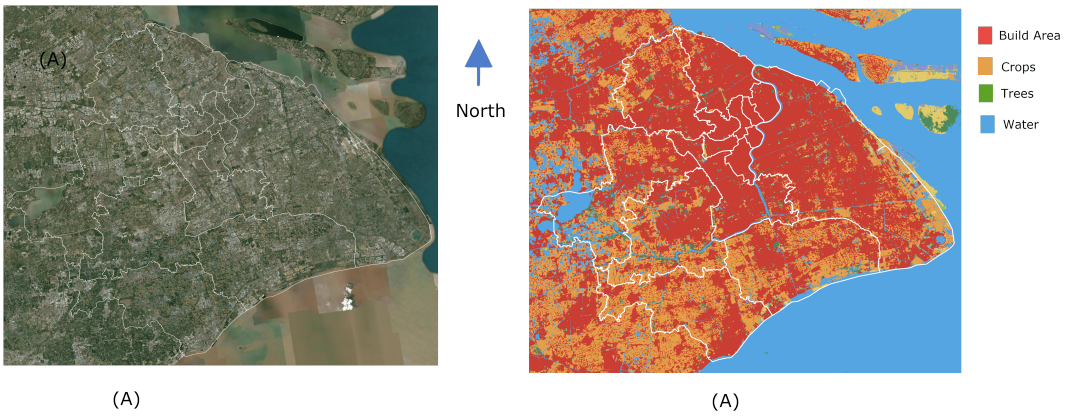



**Figure 3.2:** Shanghai geographical data a) satellite view b) land use/land cover (LULC)

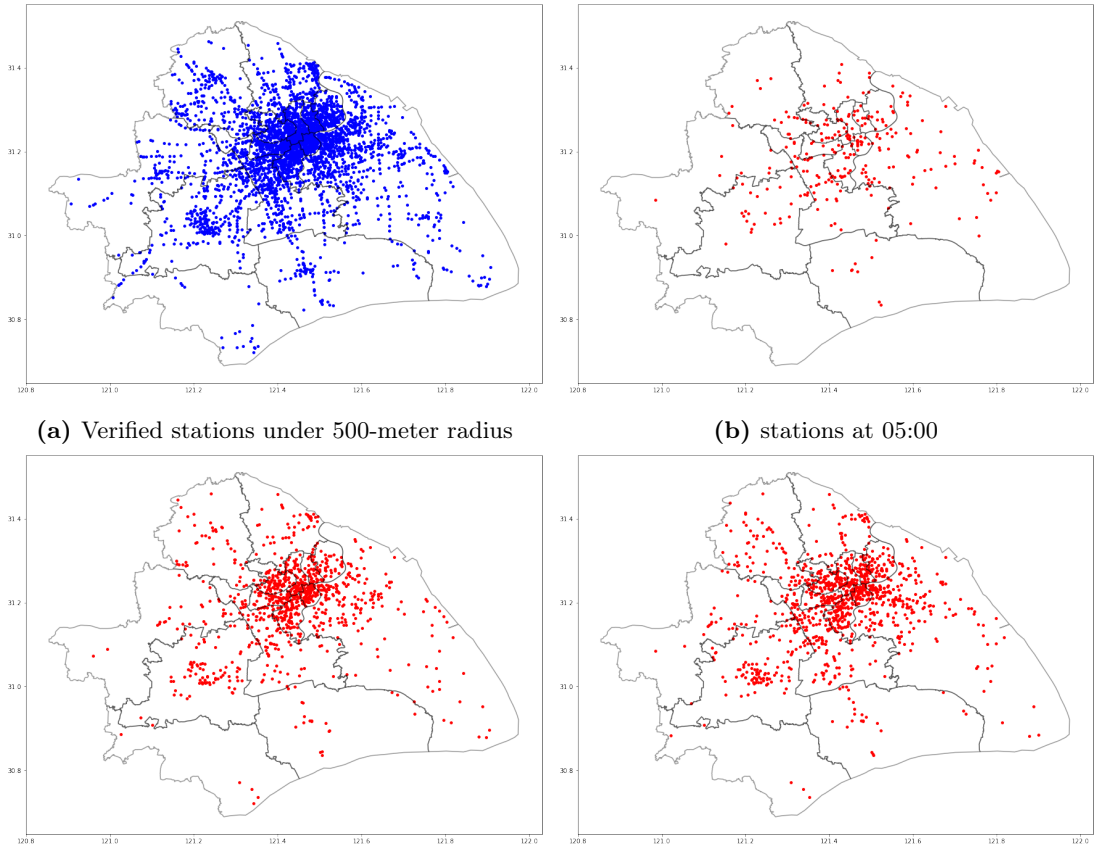
**Table 3.3:** Base stations and users count at different hours of the day

| Time | Stations Count | Connected Users | Apps Count |
|---|---|---|---|
| 5 (05:00:00 - 05:59:59) | 269 | 183 | 327 |
| 10 (10:00:00 - 10:59:59) | 1007 | 583 | 895 |
| 14 (14:00:00 - 14:59:59) | 1113 | 612 | 923 |

The collected records are from different time periods during the day. Hence, we have subdivided the samples into three different time routines during the day to reflect the temporal behaviours of the users. Table 3.2 shows the number of connected stations, users, and the number of applications used at each time routine and connected stations.

As in Table 3.3, samples collected at hours 5 are 75989 from only 269 stations. It is the time when only a few users are connected to the internet hence making only a few records of data. Collected samples at time 10 show that more users are connected at the given time. Ten is considered as the time when most people are rushing to the office. As a result of which the number of connected stations increases to 1007. To make the analysis more interesting, we collected samples at 14, and the number of connected stations jumped to 1113, which shows a minor change of stations between 10 and 14. Records collected at 14 are significantly greater than records collected at other times of the day. This is due to the fact that these times are considered as work routine time in urban areas.

As shown in Figure 3.3, base stations at different times vary, and most importantly, the dense area is the same region we visualized when we were checking the spatial



**(a)** Verified stations under 500-meter radius

**(b)** stations at 05:00

**(c)** stations at 10:00

**(d)** stations at 14:00

**Figure 3.3:** Verified stations over Shanghai

behaviour of data.

## 3.3 Data modelling

We have discussed the spatial and temporal characteristics of the dataset by mapping stations over the Shanghai district and checking usage at different times of the day. The results from the previous section illustrate that data is not uniformly distributed over the area and is denser in some regions. This irregular distribution makes it difficult to jump towards analysis. Therefore, before proceeding towards the analysis section, the lush area is filtered and chosen for further consideration.

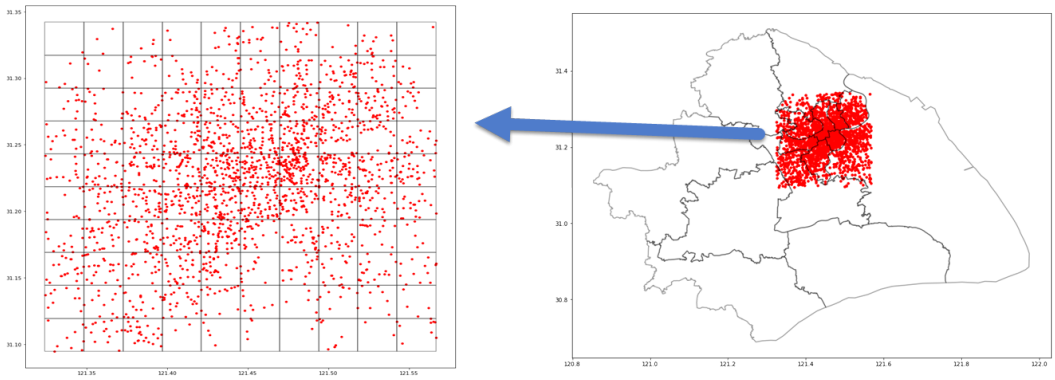


**Figure 3.4:** Verified stations over the grid

### 3.3.1 Interested area

For the purpose of in-depth analysis, we have considered a dense area with most land use and maximum possible stations. The interested area X is taken under consideration is also chosen based on density and land use to get group mobility patterns. Interested area X is captured from the whole dataset by making a grid of 10x10 over the dense area, as shown in Figure 3.4. After making the grid and considering only the stations inside the grid, we got 4236931 valid records from 2314 base stations. Filtered records contain data from 570 unique users that are using 1322 mobile applications during the connectivity. It can be seen from Figure 3.4 that interested area X is the area with the maximum number of valid stations making it more reasonable for such a choice. A dense area means more connected users hence helpful to extract mobility patterns.

### 3.3.2 Data pre-processing

The data we filtered in the previous section is still not valid because it contains records with outliers and other anomalies. We cleaned data in more detail to get clear and accurate results. Considering data from the only interesting area, the following cleaning process is considered in two significant steps.

In the first step, we have associated the number of applications with the number of users for getting useful records. Results showed that a total of 1322 apps are in the dataset, with multiple users using them at different times, but there are few applications

with just a few users. To get streamlined data, we set out the threshold values. For considering the threshold value, we made a percentile graph of the number of applications with the users' count.

After checking the percentile graph in-depth, we removed applications with the $60^{\text{th}}$ percentile, or in other words, we have removed applications with less than six users. This step further filtered the dataset concluding the 515 remaining applications for further consideration. This removes around 39 percent of applications from the filtered data. After clearing the applications, we have filtered the records from the dataset resulting in 3160412 good records for the subsequent analysis. The resultant dataset now contains 515 unique applications left out of 1322 total applications. The number of users is now 560 connected from 2279 base stations.

During the second step in pre-processing, we have grouped the users with the number of generated records. The total number of users is 560, making samples from a very low to a very high number of records. Few users are producing only one record, while few are making more than 40000 records. We made percentiles to make threshold values for filtering the dataset. As shown in Figure 3.4, many users produce fewer records that can affect the results. Hence, we removed the records of users with less than 20 samples that were in the $5^{\text{th}}$ percentile of the records. Once we filter the users with a higher number of records, we have filtered the significant dataset based on the result from this step. We have got 3160179 records that we are considering for the final analysis process.

After processing the data with the filtration processes, the data under consideration for the analysis contains 3160179 total records produced by 529 users. Considered data contains data from 2271 stations and 515 uniquely used mobile applications. The filtered base stations are shown in Figure 3.4 (data is from the dense area).

### 3.3.3 Validation of temporal characteristics

We have seen spatial and temporal characteristics of the data in Section 3.2. To validate data after filtration, we have made hourly bins at different times of the day before working with the interested area. After validating records for the interested area, we have to check the existence of the temporal characteristics in the selected data.

Firstly, we have made the histogram with all records at different times of the day. As suspected, the behaviour of data remains the same. From the Figure 3.5, it can be seen that number of records starting from the day start at 0 hrs. (from 00:00:00 to

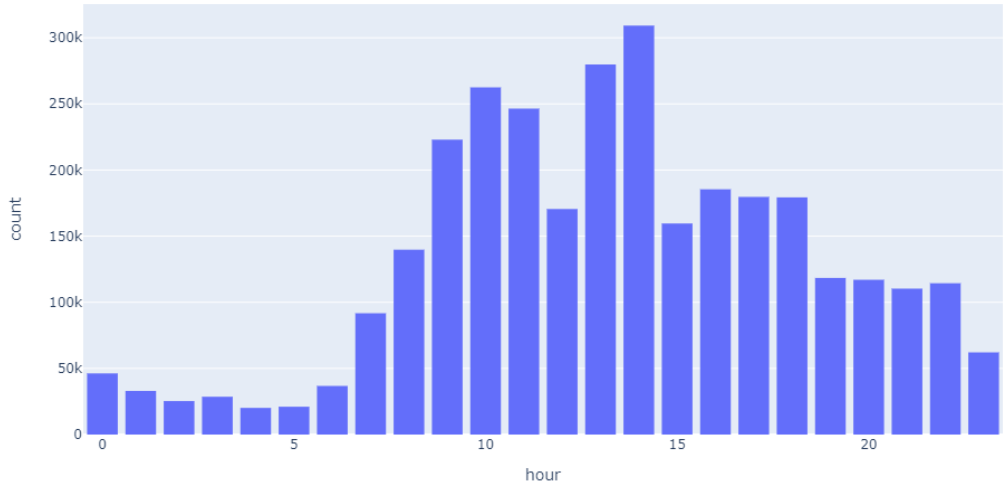



**Figure 3.5:** Hourly bins for number of records in filtered data

0:59:59) till five hours remains very less as compared to the working hours during the day. Records started to rise from 8 and 9 as business hours started, and people moved to offices, schools, and other business places. Connectivity increase with the time till 12. At 12 (from 12:00:00 to 12:59:59), the activity decreases compared to 10 and 11. As time passed till late afternoon, the number of records decreased. Such activities are demonstrated by several researchers (14) that people sleep during the night hence producing a lower number of records with less connectivity to the internet.

Firstly, we have made the histogram with the total number of records at different times of the day. As suspected, the behaviour of data remains the same. From the Figure 3.5, it can be seen that number of records starting from the day start at 0 hrs. (from 00:00:00 to 0:59:59) till five hr. remains very few as compared to the working time during the day. Records started to rise from 8 and 9 as business hours started, and people moved to offices, schools, and other business places. Connectivity increase with time to 12. At 12 (from 12:00:00 to 12:59:59) the activity decreases as compared to 10 and 11. As time passed till late afternoon, records started to decrease. People sleep during the night hence producing a lower number of records with less connectivity to the internet.

From the histogram, we have validated the temporal behaviour of data and illus-

trated the regular daily routine followed at urban places in the interested area.

## 3.4 Summary

In this Chapter, we provided an overview of our crowd sensed dataset. We also verified its validity and presented the processes that were used for cleaning and preparation before the modelling of trajectories. We verified the base stations from OpenCelliD within a radius of 500 meters. After that, we visualized the sparsity of data using its spatial-temporal characteristics. Data is not uniformly distributed, and we removed this sparsity issue by considering data from the interested region where data is dense and creating a grid overlay to check the distribution of data among the grid. After considering interesting areas and filtering records, we filtered further by considering users with a more significant number of samples and applications with a greater number of users. After all pre-processing steps, we made a histogram to validate the temporal property of data for the interested area. The filtered records count to 3160179, which is around 75 percent of the actual data from the grid, and only approximately 25 percent of data is left for analysis from the whole dataset.

The next Chapter will be focused on utilizing the processed data and finding users' mobility as a trajectory over the grid.

# 4

# Human Mobility Modelling

This Chapter discusses the methodology of modelling trajectories for users to aid the process of finding mobility patterns. Different mobility patterns are analyzed between users to determine the mobility and explore patterns.

## 4.1 Mobility modelling

In the previous Chapter, various techniques have been applied for data cleaning and preparation. Next, we proceed to model trajectories over the denser area over the map that is considered. The modelling of trajectories is achieved by overlapping a grid structure over the area of interest. Trajectories then are formed by describing changes between cells as the user traverse the area. In this section, we describe our grid method and tools used to model the trajectories with crowd sensed data.

### 4.1.1 Constructing grid

To overlap a descriptive grid over the area, we rely on QGIS, which is an open-source GIS tool [1]. QGIS has been used for grid assembly and modelling. The desktop version of QGIS has been utilized to extract the urban interested area of Shanghai from the map. Data were imported to the QGIS project for featuring the locations over the map. After careful consideration, a compact space with a higher volume of base stations was selected. Grid overlay was built over the selected area, and the selection of base stations

[1] https://qgis.org/en/site/

was performed using the Selection by Location feature from the QGIS tool. This selection aided in filtering valuable records. Figure 4.1 demonstrated the selected stations from the overall dataset, and further processing entails only these chosen stations.

Furthermore, spatial join [1] was applied on the chosen data and the gird for unique identification of the grid cells. As a result, identifiers of each cell aided in tracking trajectories for further considerations.

### 4.1.2 Mapping trajectories over grid

Besides combining data with the grid, the unique identification of cells helped us explore the formation of trajectories. The grid structure has an arrangement of ten rows and ten columns, resulting in a total of 100 individual cells. Merging the cells with the stations provided the output of one grid cell accommodating multiple stations. We limited our analysis to this grid cell granularity as our main goal is to identify users moving together, and this configuration was good enough to capture human mobility. Further analysis of the cell granularity may provide more accurate results. However, we do not foresee significant changes in performance.

As shown in previous research, trajectories are counted as the movement between the cells on the grid, rather than using direct and oscillating connectivity to base stations. For instance, Figure 4.2, shows the trajectory of two possible scenarios. Figure 4.2a depicts the trajectories of two different users moving from point A to point B, making considerable footprints of their trajectories. In contrast, Figure 4.2b exhibits the trajectories of the other two users, which shows the trajectory but can not be considered as mobility samples due to the fact of limited mobility within the same cell. Both figures provides evidence for considering the mobility inter-cell from the grid.

## 4.2 Data extraction

Before taking the next steps, filtered data after grid construction resulted in 528 users with 3160179 samples distribution which is 25 percent of the overall dataset. Furthermore, screened data encompasses 2271 base stations. Further analysis was conducted to explore the sample's distribution of users. This step was performed to validate the

[1] https://geopandas.org/en/stable/docs/user_guide/mergingdata.html

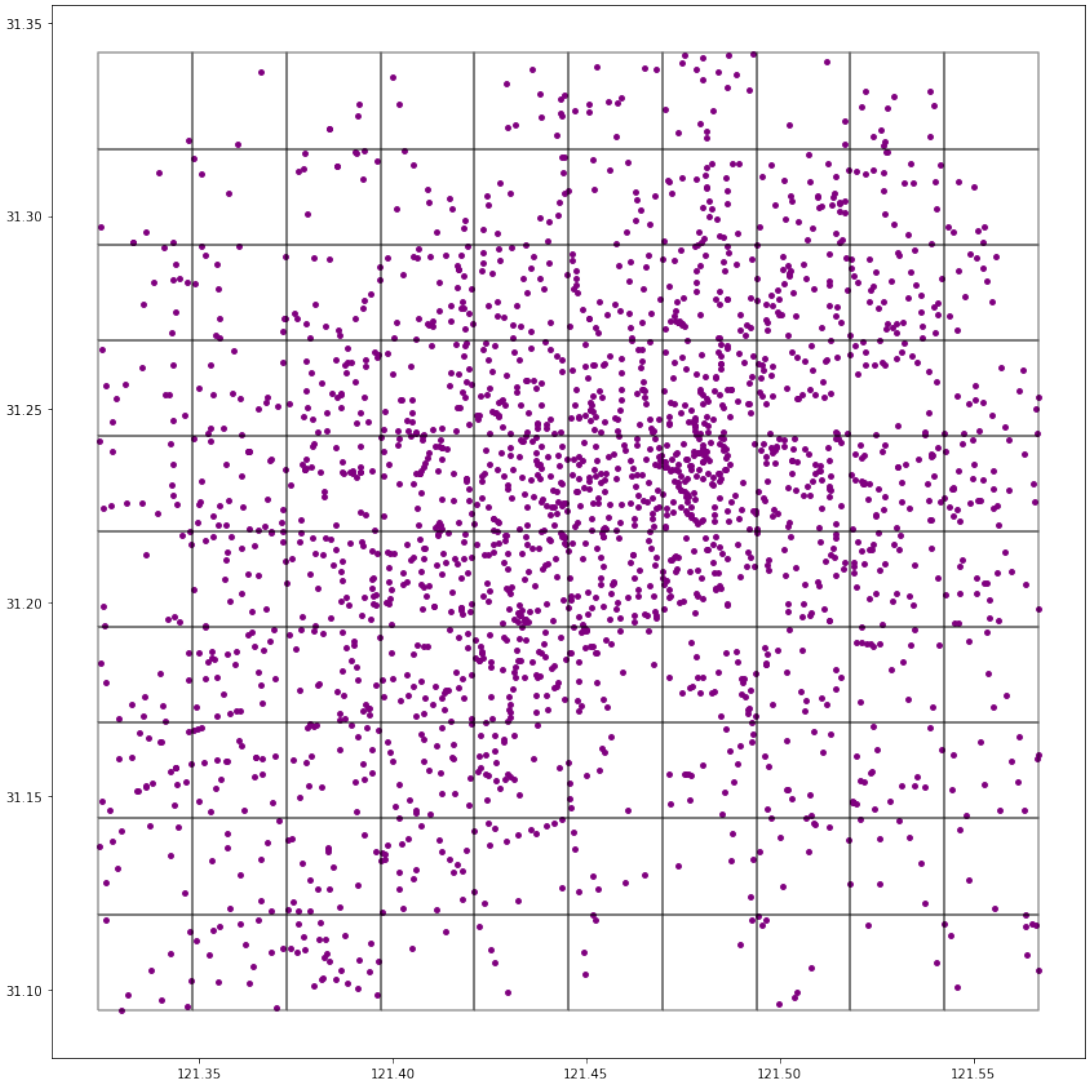
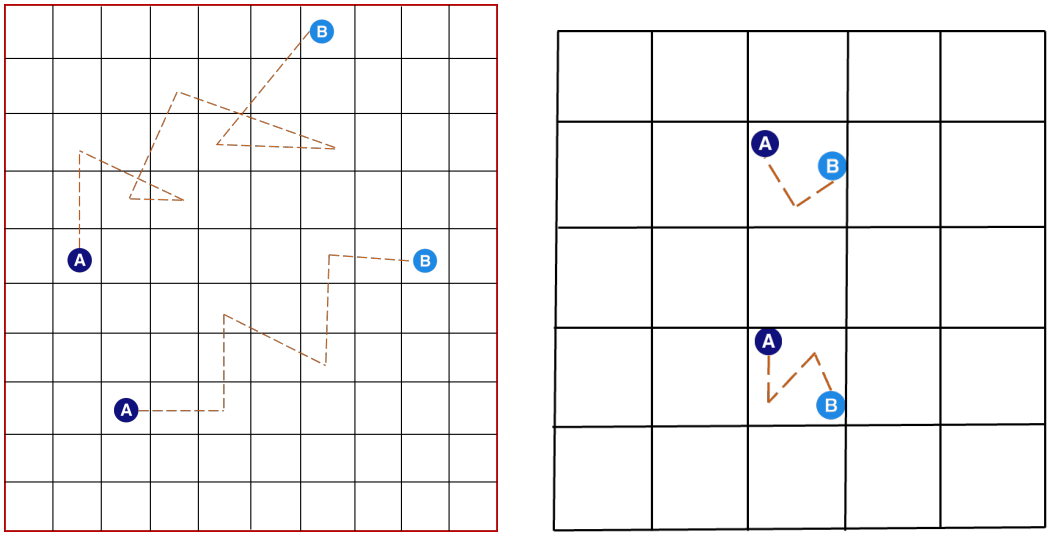



**Figure 4.1:** Grid with selected basestations




**(a)** multi-cell trajectory **(b)** same cell trajectory

**Figure 4.2:** Fictional users' trajectories for understanding mobility over the grid

samples associated with the users. Figure 4.3 exhibits the division of the overall samples concerning percentiles. To our expectations, a drastic difference has been explored between users and samples generated by them. For further clarification and better visualizations, MovingPandas [1], a python library created for visualizing trajectories based on geodata, was utilized. MovingPandas helped to understand the patterns and trajectories more conveniently.

### 4.2.1 Samples distribution

Figure 4.3 shows the samples produced by each user, which was the resultant of categorizing users with all the samples produced in the whole considered data. Figure 4.3 illustrates the sparsity of samples distribution. Fewer users generated excess records, i.e., more than 40K samples, while the median of the data samples is 5961 (around 6k samples per user).

This erratic behaviour was explored by the percentile table 4.1. The Table 4.1 shows a significant difference between the $90^{th}$ percentile and $100^{th}$ percentiles. Results show that only a few users produced more significant number of records as compared

[1] https://github.com/anitagraser/movingpandas

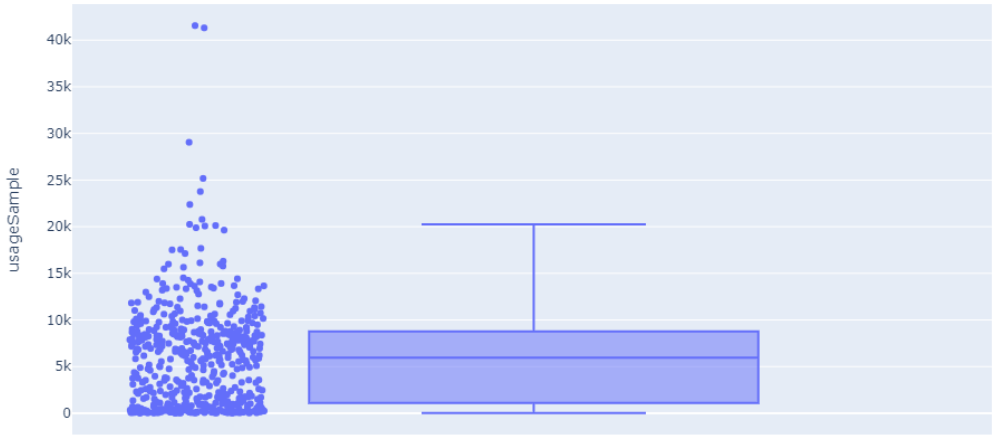



**Figure 4.3:** Samples distribution per user from inetrested area

**Table 4.1:** Percentile summary of samples distribution for interested area

| Percentile | Number of Samples |
|---|---|
| 10 | 136.0 |
| 20 | 658.0 |
| 30 | 1741.2 |
| 40 | 3608.6 |
| 50 | 5961.0 |
| 60 | 7417.0 |
| 70 | 8332.6 |
| 80 | 9599.0 |
| 90 | 11844.0 |
| 100 | 41558.0 |

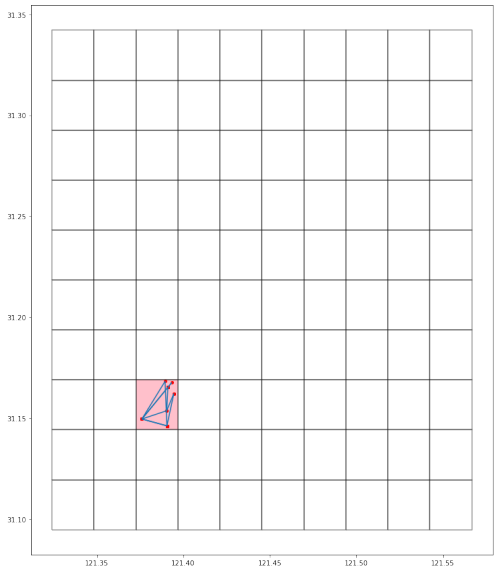
to others. But the association of trajectory is yet to be explored between generated samples and users.

Since our goal is to model the mobility of users, we then explore the relationship between the number of collected samples and mobility. To do this, different samples were taken from the low samples ($10^{\text{th}}$ percentile), medium samples ($50^{\text{th}}$ percentile), and high samples ($90^{\text{th}}$ percentiles) categories. Considering the hypothesis of a direct relation between trajectory and samples, the top four users are considered from each sample category for visualization and validation.

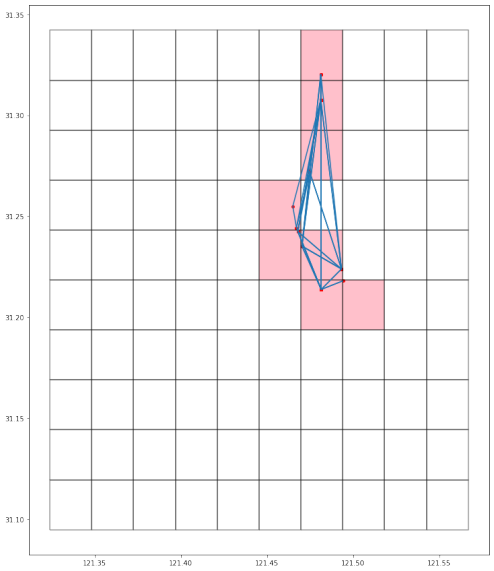
Figure 4.4 illustrates the trajectories of the uppermost four users by the filtration criterion of the sample's distribution with less than or equal to 136 samples that counted as category 1 with low samples which contains $10^{\text{th}}$ percentile of actual data. Results are surprisingly different from the expectations. The trajectory of user 1 (see Figure 4.4a) exhibits the lowest trajectory among all considered users keeping the mobility restricted to one grid cell location. In contrast, the remaining users showed more mobility by making footprints among multiple cells from the grid. As category 1 contains the users with low samples distribution, the results are not uniform as far as trajectories are concerned. Figure 4.4b, 4.4c and 4.4d unveil the trajectories of other considered users for the study from this category.

The next category under consideration is category 2, with the users of medium samples distribution. Users are filtered by implementing the filtration criterion of users having samples less than or equal to 5961, which constitutes the $50^{\text{th}}$ percentile of actual data. Category 2 contains more significant users as compared to category 1, yet the results exhibit similar behaviour as in category 1. For instance, few users, as in Figure 4.5b and Figure 4.5c, showed lesser mobility by restricting the mobility to a couple of cells from the grid despite having greater samples in the data. In contrast, few users, as in Figure 4.5a and Figure 4.5d, made greater trajectory patterns despite having similar samples distribution from category 2. Figure 4.5 illustrates erratic behaviour of trajectories in category 2.

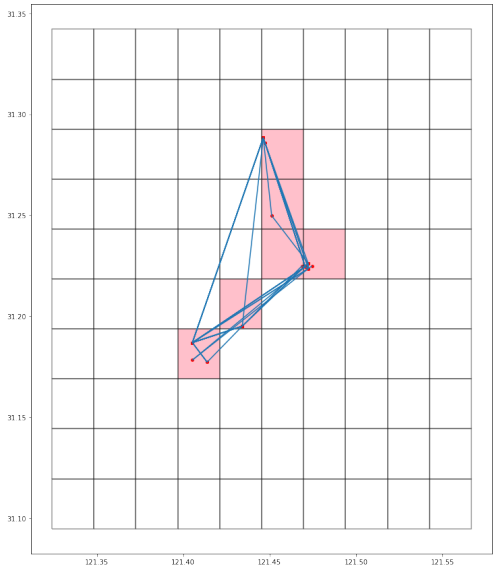
Category 3 was constructed based on users with the higher samples distribution with the selection criterion of samples lesser than or equal to 11844, which constitutes the $90^{\text{th}}$ percentile of the actual data. In accordance with the hypothesis made earlier, the result should exhibit higher trajectories as the users possess higher samples distribution. In contrast to the provided statement, the results are unexpectedly variant. Figure 4.6



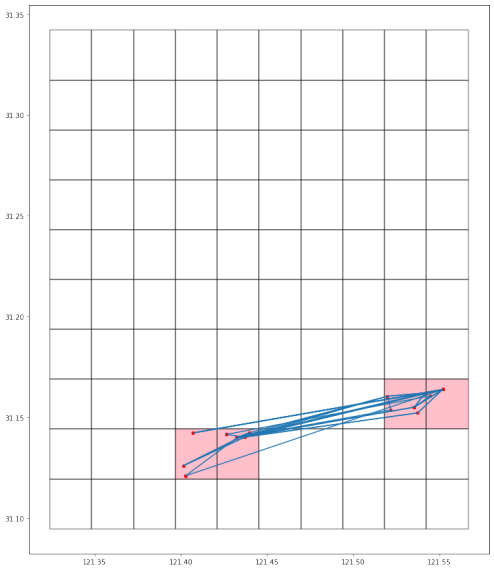

**(a)** Trajectory of user 1




**(b)** Trajectory of user 2

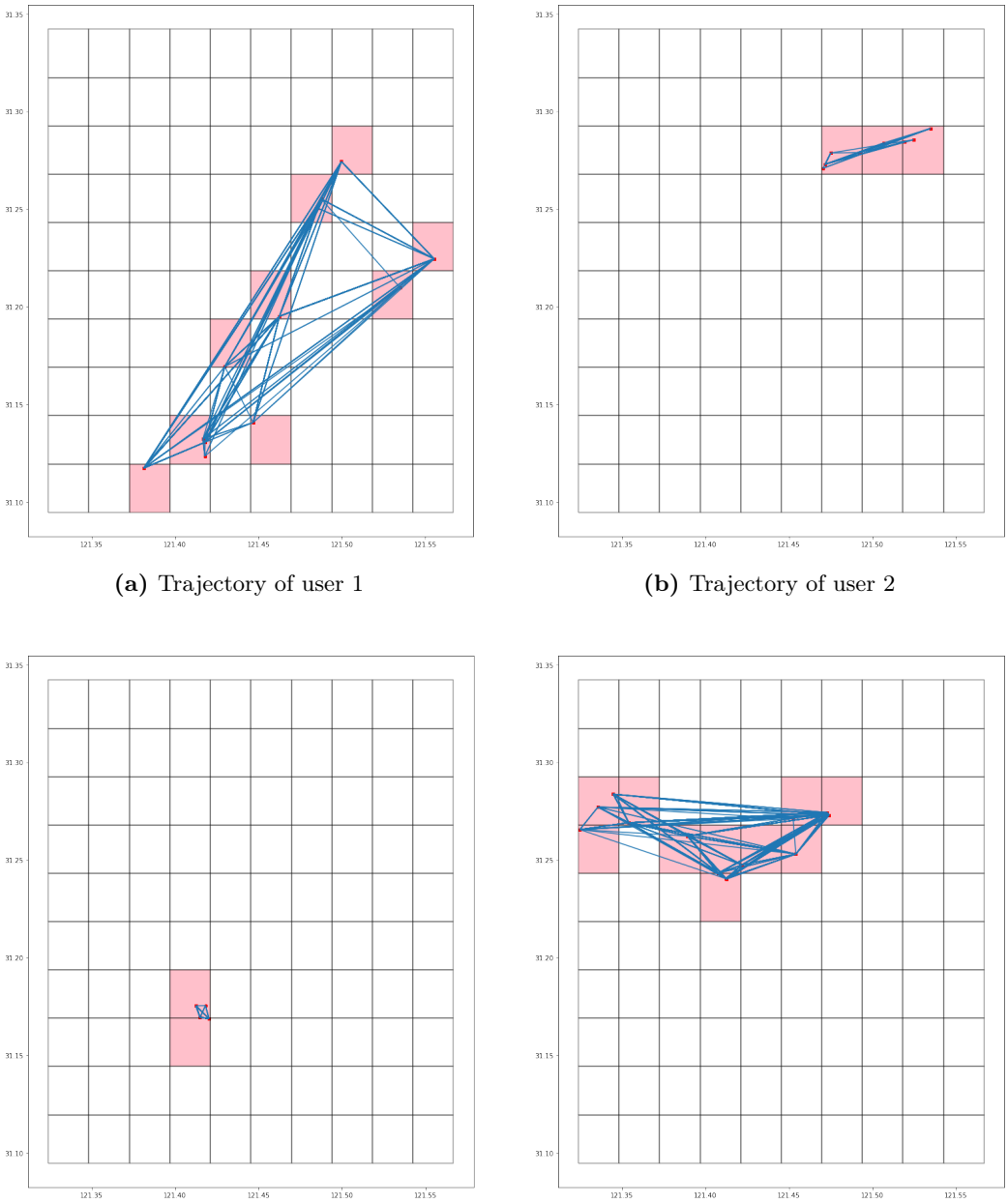



**(c)** Trajectory of user 3




**(d)** Trajectory of user 4

**Figure 4.4:** Top 4 users in category 1: low samples ($10^{\text{th}}$ percentile)




**(a)** Trajectory of user 1

**(b)** Trajectory of user 2

**(c)** Trajectory of user 3

**(d)** Trajectory of user 4

**Figure 4.5:** Top 4 users in category 2: medium samples ($50^{\text{th}}$ percentile)

exhibits the top four considered users from category 3. One from the selected four users showed more significant displacement as seen in Figure 4.6b. However, remaining users, as in Figure 4.6a, Figure 4.6c and Figure 4.6d, the trajectories are lower, thus giving the least mobility even with more outstanding samples production.

After carefully considering the selection criterion of making categories based on samples distribution and comparing the trajectories of a few users from each selected category, the hypothesis proved to be invalid as trajectories are unevenly distributed. Results showed no direct relation of trajectories with the samples.

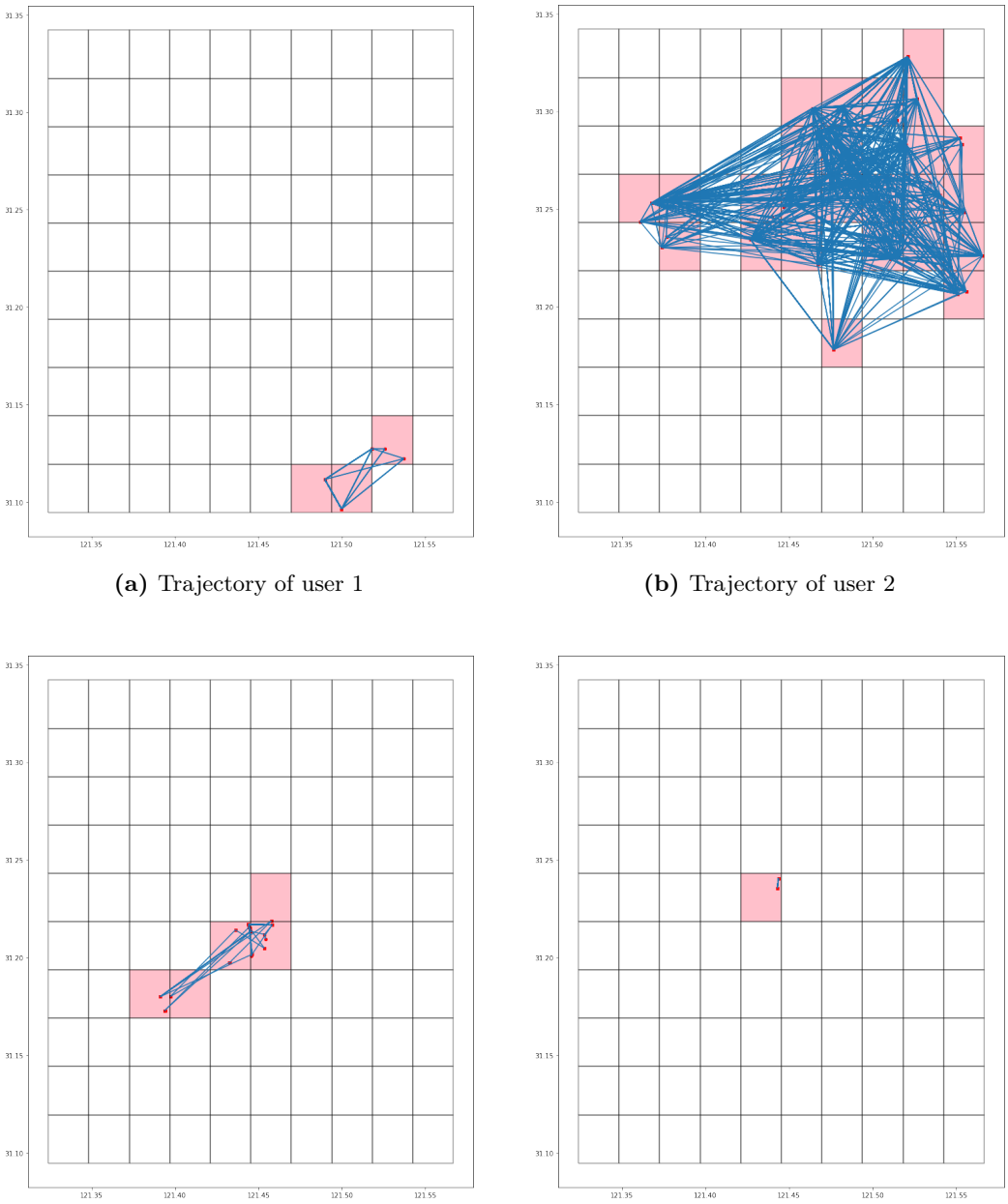
## 4.3 Trajectory extraction

In the sub-section 4.2.1, the outcome illustrated the indirect relation of samples and trajectories distribution. Therefore, the following steps are performed to change the criteria for making categorization for further consideration and quantifying similarities between trajectories.

A trajectory is the user's footprints while moving from one point to another. One user can make multiple trajectories based on various day hours depicting the mobility needed in every hour. Therefore, the first step was to extract the user's trajectories, clustering users at every hour was performed. This step was taken to aid the process of exploring trajectories at different hours of the day for a specific user. This procedure facilitated the calculations by reducing the recorded samples when user connection status was to a single point. Thus reducing the overhead of computations performed.

For instance, as seen in Figure 4.6d, a single user was considered to validate the process of selection based on hours. This user was specifically under consideration for this process as it showed the least trajectory but with a higher samples distribution, i.e., 11732 samples. The calculated hours included 9, 10, 11, 12, 13, 14, 15, 16, 19, 20, and 21, with samples produced in each hour. This shows the lumpy characteristics as the user stays connected most of the day while no recorded trajectory. It depicts the uneven conduct of hourly distribution. Furthermore, the trajectories construction based on hours and samples proved to be insufficient to extract the mobility patterns.

In addition to the previous considerations, the trajectory calculations were enriched by exploiting the grid cell locations during the whole day. Measuring the location of the cells gave the trajectory movement for each user. For instance, a user moved from




**(a)** Trajectory of user 1

**(b)** Trajectory of user 2

**(c)** Trajectory of user 3

**(d)** Trajectory of user 4

**Figure 4.6:** Top 4 users in category 3: high samples ($90^{\text{th}}$ percentile)

point A to point B, cloaking three cells over the grid. Thus the trajectory includes the location of each cell, i.e., 22, 42, and 36. Figure 4.6a illustrated the trajectory of a user within three cells over the grid. Hence the mobility score for the given user is three.

Similarly, the same approach was implemented to extract the trajectory scores for each user from the data. The results showed scores ranging from one (least trajectory length) to 28 (most considerable trajectory length). Table 5.3 shows randomly selected users with their trajectories.

## 4.4 Summary

This Chapter describes the methodology applied over data for exploring mobility as trajectories. The first step included filtering the crowd sensed data by making a grid over the map and extracting base stations inside the grid. The next step included the usage of tools and libraries like QGIS and *movingpandas* to visualize the trajectories. Initial exploration included constructing categories based on samples distribution: low samples, medium samples, and high samples and studying uppermost users from each category for analyzing mobility. However, the categories creation criteria based on samples proved to be non-favourable to conclude considerable results. Afterwards, trajectories are further studied based on criteria for varying cells over the grid. As a result, trajectory scores were extracted from which further categories criteria were explored. This later approach demonstrated to be suitable for analysis, and it is used for analyzing the characteristics of different types of trajectories.

The next Chapter describes our experimental setup, which includes the procedure for similarity score calculations and comparison between different types of trajectories. These estimated scores are then used to identify users moving together.

# 5

# Experimental Setup

This Chapter describes the experimental setup used to explore the similarity between trajectories and identify users that move together. In the following, we state the selection criteria and scope of our analysis.

## 5.1 Characterized groups

Our work is based on samples collected from users when those connect to base stations. By combining these incremental samples, it is then possible to construct trajectories that capture human mobility in urban areas. Naturally, different users contribute with varying amounts of data. Thus, trajectories are modelled based on different spatial and temporal data characteristics.

We built four groups that depict different types of trajectory lengths that can be found in our dataset. We choose to divide the resulting trajectories from our data into four groups, such that each group is balanced in terms of users and the number of samples. Table 5.1 shows the resulting division.

**Table 5.1:** Users group formation criteria and trajectories distribution

| Group | Percentile (%) | Trajectory length | Active users | Average samples distribution |
|---|---|---|---|---|
| Short trajectory (Group 1) | 25 | <= 3 | 135 | 5739 |
| Short-medium trajectory (Group 2) | 50 | > 3 And <= 8 | 132 | 5037 |
| Medium-large trajectory (Group 3) | 75 | > 8 And <= 16 | 129 | 5838 |
| Large trajectory (Group 4) | 100 | > 16 | 132 | 7316 |

**Table 5.2:** Selected user from each trajectory group

| Group | Trajectory length | Users |
|---|---|---|
| Short trajectory (Group 1) | 3 | 6 |
| Short-medium trajectory (Group 2) | >= 6 | 10 |
| Medium-large trajectory (Group 3) | >= 10 | 8 |
| Large trajectory (Group 4) | >= 28 | 10 |

## 5.2 Experimental setup

**Goal:** To verify whether a group of users moving together can be identified by analysing the similarity score between the mobility trajectories of users.

**Overall users:** After careful pre-processing, a total of 528 users are extracted from the dataset.

**Selected users:** Selected users that contain the densest and representative mobility trajectories are analysed further in our experiments. These users are selected as they provide rich mobility patterns and accurate estimated trajectories that support our experimental goal. Table 5.2 describes the users selected per each group.

**Procedure:** As per our grid modelling, we calculate the (24 hrs) trajectory of each user in our dataset. We then estimate the similarity score between all users using a similarity metric based on the DTW algorithm (See details in the next subsection). In our analysis, we first consider individual users, and then we analyse selected/overall groups of users.

## 5.3 Similarity metric

The similarity score between trajectories of users was extracted by implementing FastDTW [1], an algorithm to find the optimal match between two provided arrays or temporal sequences of variant lengths. We used DTW in our work as DTW allows dynamic comparison of trajectories with different spatial and temporal characteristics. To minimize the distance between two sequences, one-to-many and many-to-one relationships are built. For utilizing the DTW algorithm, the grid cells were counted by grouping the

[1] https://pypi.org/project/fastdtw/

**Table 5.3:** Random users trajectory length from selected area

| User | Trajectory | Trajectory length |
|---|---|---|
| 918576 | [22, 32, 51, 61, 62, 72, 73, 74, 75, 85, 95] | 11 |
| 1881449 | [24, 34] | 2 |
| 940509 | [49] | 1 |
| 1785365 | [11, 21, 22, 34, 45, 56, 57, 58, 66, 76, 85, 95] | 12 |
| 1533650 | [24, 32, 54, 62, 97] | 5 |

trajectories' locations. The result ranged from one to 28. One is the lowest trajectory restricting a user inside a single grid cell, while 28 is the most considerable trajectory covering major grid cells.

Table 5.3 shows some examples of data trajectories and related similarity scores. For instance, user 918576 showed 11 grid points for mobility showing larger displacement. However, in contrast, user 1881449 exhibits the trajectory for two cells. Therefore, calculating the similarity score between these two users would have unfavourable results. Random samples selection exhibits the irregular dispersion of trajectories. For this reason, users are grouped based on the trajectory distribution.

Dynamic time warping (DTW) has been utilized to measure the similarity in the trajectories of users within each selected group. The DTW algorithm gave us the similarity score, also referred to as distance in the following chapters. These scores helped differentiate between users' mobility and find similarity based on calculated scores. For instance, the mobility of two users is identical if the score is zero. If the score increases to one, then both users have similar mobility, but their trajectories differ with one cell on the gird. The same calculation technique has been implemented on every user group to find similarities within the group and across groups.

## 5.4 Summary

This Chapter explains the procedure used for performing the experiments. Group formation selection criteria were discussed deeply, with the proper systematic reasoning for analyzing differences between trajectory types. Furthermore, homogeneous and heterogeneous trajectories similarity selection was explained for examining in detail the optimal type of trajectories that can be used to identify users moving together.

The next Chapter presents the results of our experiments and highlights our main findings.

# 6

# Analysis and Results

This Chapter presents the results of our experiments. After conducting rigorous experimental benchmarks, the main results of our study can be summarized as follows:

- Partial similarity between trajectories is found for a significant amount of users. The complete similarity between trajectories only occurs when users remain in the same location, meaning that users do not move but just share the exact location over time.
- Similarity between trajectories is optimal for trajectories that are short in length. This suggests that when analyzing human mobility is better to analyze small segments of the trajectories separately rather than the overall trajectory at once.
- Our results demonstrate that by calculating the similarity score between trajectories, it is possible to find (partial) users sharing the same trajectory patterns. However, we found that these groups are small in size.

## 6.1 Quantifying similarity between trajectories

We begin our analysis by demonstrating that comparison between trajectories using DTW can be used to calculate a similarity score to identify similar mobility patterns in users. To do this, we perform a systematic evaluation using the whole dataset where different group categories of users are evaluated (See Section 5.2 for a detailed description of the experimental setup). Each group contains trajectories that share similar mobility lengths (modelled in our grid as cell count displacement). The results are described in

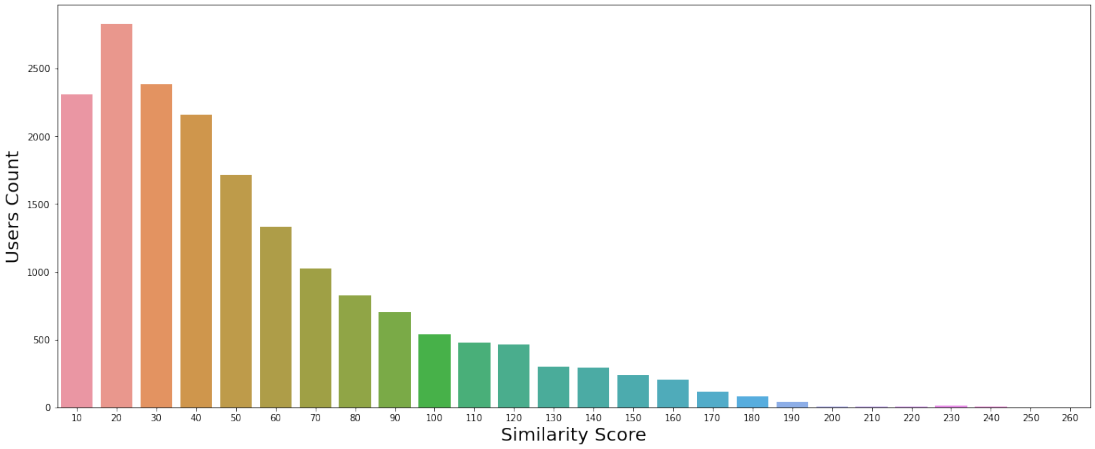



**Figure 6.1:** Similarity score of users with short trajectories

the following:

**Characterizing similarity in short trajectories:** We characterize the similarity of short trajectories provided by 135 users available in this mobility group (Group 1). Figure 6.1 shows the results. From the figure, we can observe the amount of similarity found between users when taking into account all possible combinations. Interestingly, we found several users that share similar mobility patterns. However, as the users in this group have shorter trajectories, we found that this similarity is mostly due to users not moving at all (between cells) and just sharing the same cell location, meaning the similarity score is zero. Likewise, we also found several users sharing at least partial cells. Thus, segments of their mobility trajectory overlap. We did not find complete trajectories overlapping between different users.

**Characterizing similarity in short-medium trajectories:** Next, we characterize the similarity of the users with short to medium trajectories (Group 2). A total of 132 users were considered, providing more than 17000 score samples. From our analysis, we found five users whose similarity score is zero, indicating (equal) matching patterns of mobility. Figure 6.2 illustrates the trajectories of 4 out of 5 users with zero scores. From the figure, we can observe that all trajectories matched the same trajectory described by the grid in terms of movements between cells. This indicates that DTW can indeed find users that share similar mobility patterns. Naturally, we can observe

that the similarity between trajectories is just shared between a small group of users ($< 2$ users). In addition, we also found trajectories that are partially similar, meaning that the similarity score is not zero but close to zero. For instance, Figure 6.3 shows users whose similarity score is equal to 1. From the figure, we can observe a partial matching between trajectories, which suggests that similarity of users can be found not in the whole trajectory but rather in segments of it. Moreover, Figure 6.4 shows the users with the most similar mobility patterns. We also summarize the distribution of similarity scores that can be found in this group in Figure 6.5.

**Characterizing similarity in medium-large trajectories:** We then proceed to analyze the medium to large trajectories of users (Group 3). A total of 129 users were considered in this analysis. Surprisingly, we did not find users whose mobility patterns matched exactly. Instead, we found that the trajectories of users matched partially, meaning that the similarity score is greater than zero. To illustrate this, Figure 6.6a and 6.6b exhibits the trajectories that scored a similarity of one. Likewise, Figure 6.6c and 6.6d illustrates the trajectory of users with the similarity score of two. As shown in the figure, no users in our analysis shared identical mobility patterns, similarity can be found only within overlapping segments. Figure 6.7 shows the users with the most similar scores in our analysis. Lastly, Figure 6.8 shows the distribution between all users in the medium to large trajectory group. We can observe from the figure that the number of users sharing similar mobility patterns are reduced when compared with short trajectories. This suggests that trajectories become less similar as their length increases. This is reasonable as individuals are expected to spend some but not the whole time together.

**Characterizing similarity in large trajectories:** We also analyze the trajectories with the larger lengths (Group 4). To do this, a total of 132 users were considered. Interestingly, we did not find users that share partial similarity with lower scores. In contrast to previous groups, we did not find users that have partial similarity below a similarity score of 3. Instead, most of the similarity scores of this group have a similarity score greater than 4. Figure 6.9 shows the results of trajectories with this similarity score. Our results suggest that the longer the trajectory, the more difficult to find similar users sharing the same mobility pattern.

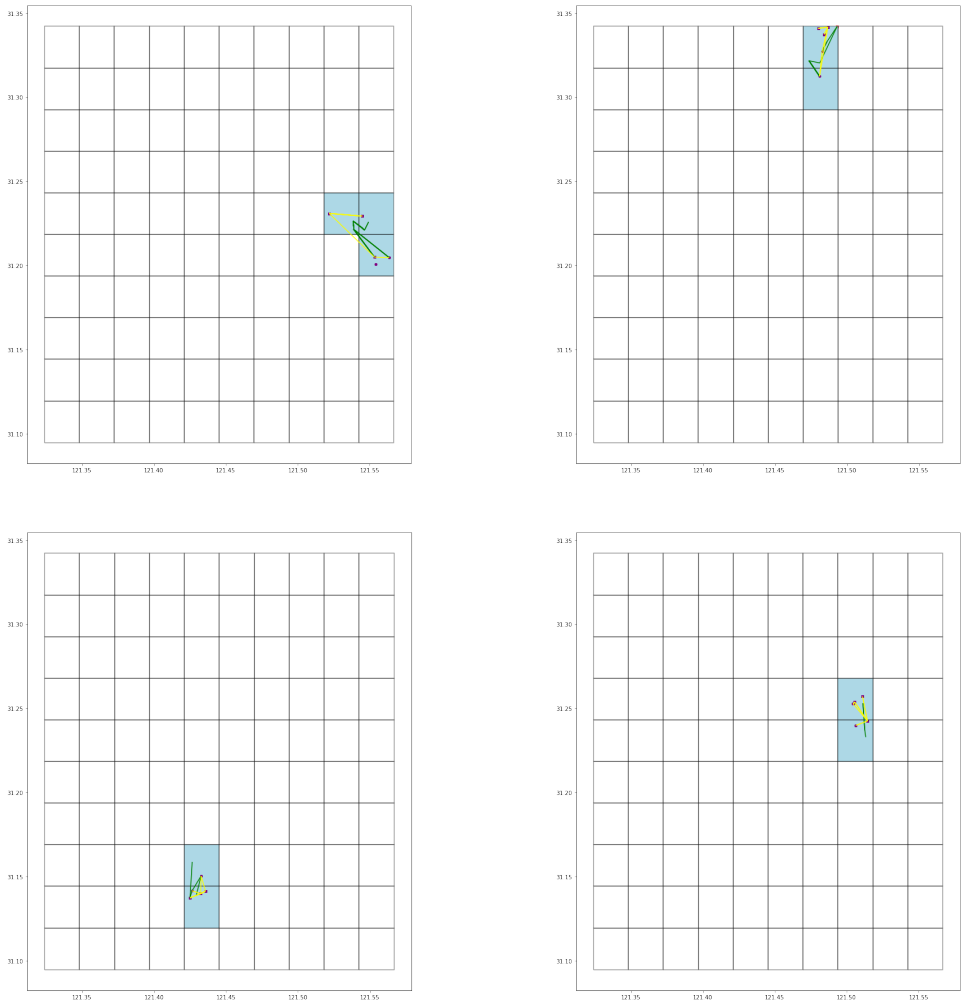


**Figure 6.2:** Users with similar trajectories in short-medium trajectories group with similarity score of zero

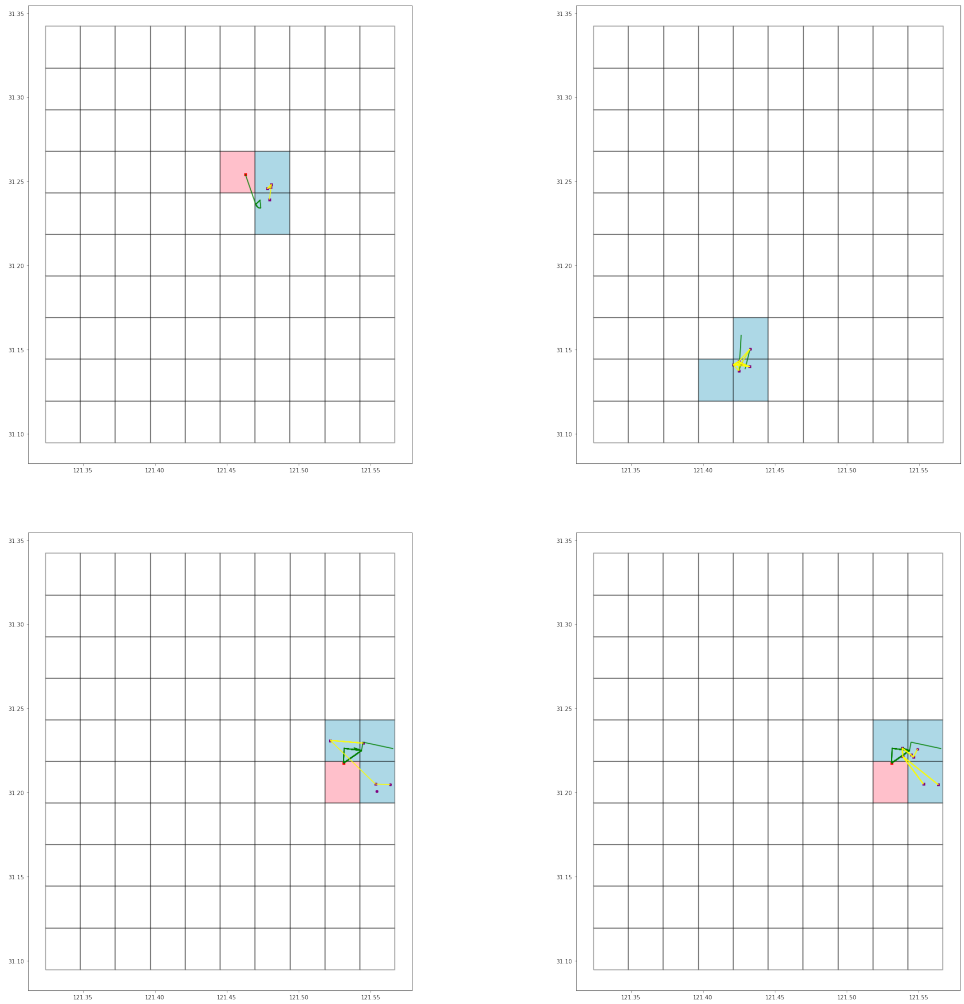
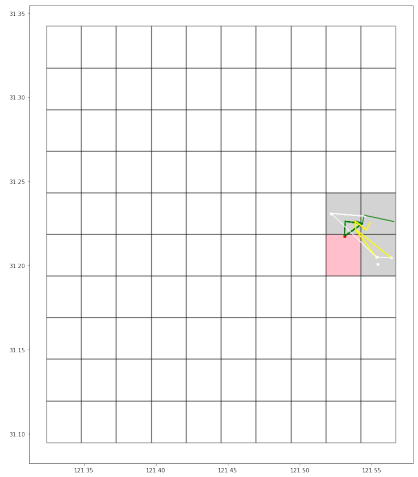
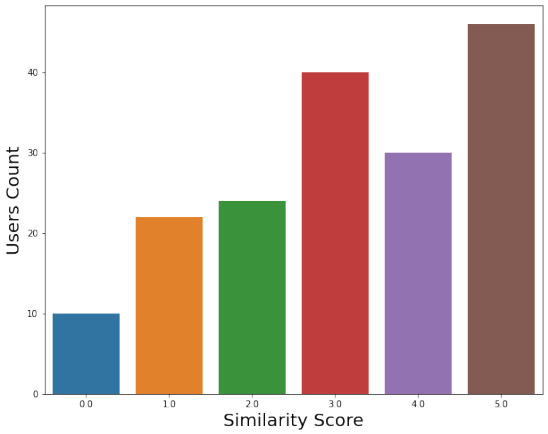


**Figure 6.3:** Users with similar trajectories in short-medium trajectories group with similarity score of one



**(a)** Users with similar mobility pattern




**(b)** Similarity score distribution for selected score range

**Figure 6.4:** Users sharing similar mobility patterns and similarity score distribution for selected score range from short-medium trajectories group

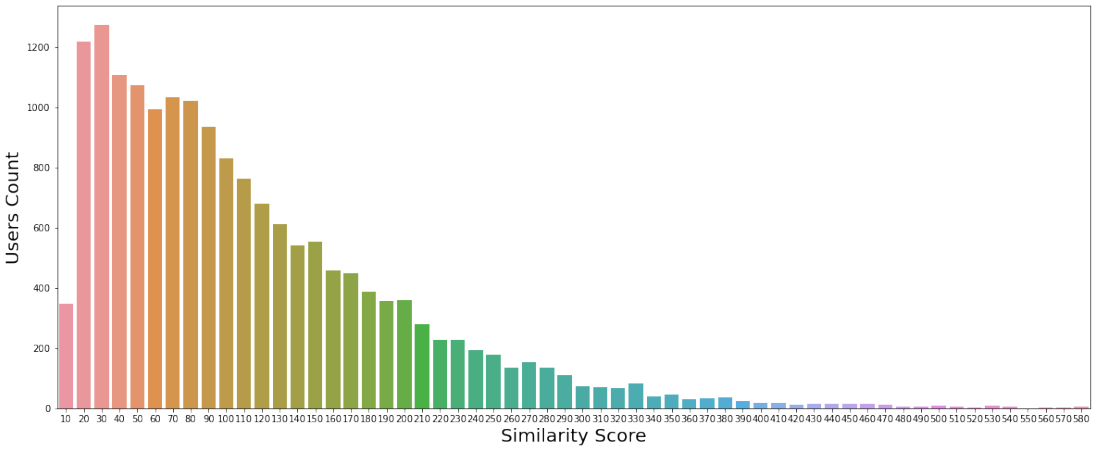
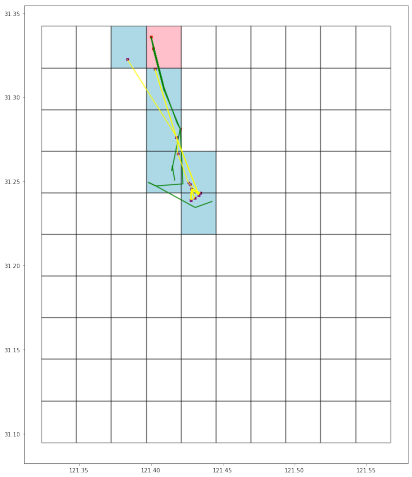
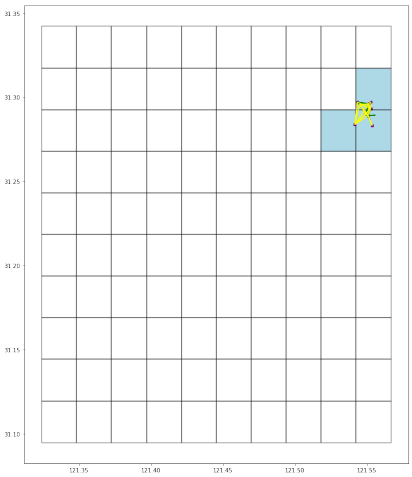


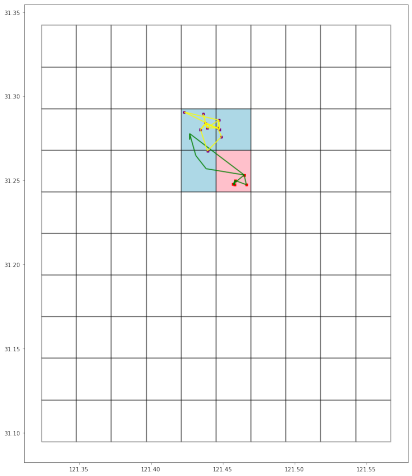

**Figure 6.5:** Similarity score of users with short-medium trajectories



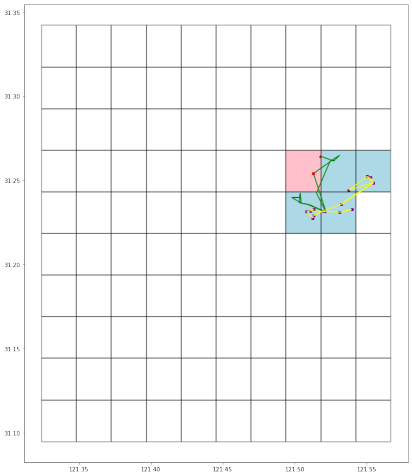
**(a)** User trajectories with similarity score of one



**(b)** User trajectories with similarity score of one

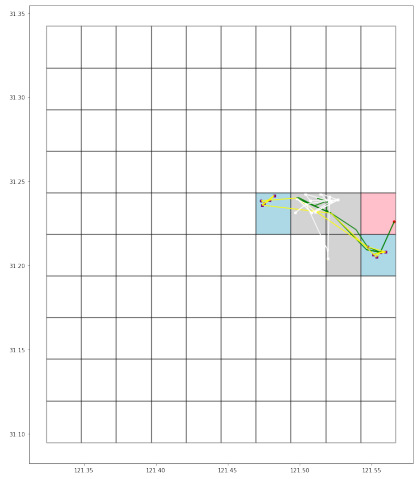


**(c)** User trajectories with similarity score of two

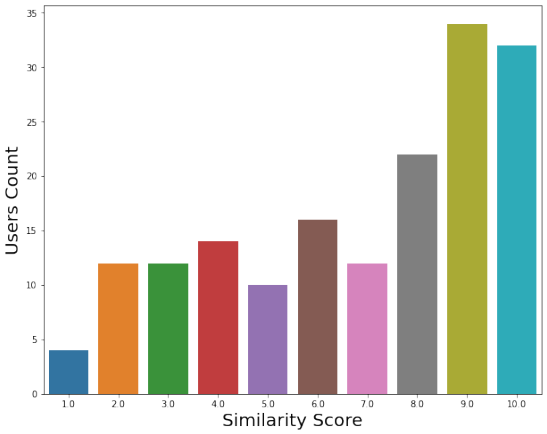


**(d)** User trajectories with similarity score of two

**Figure 6.6:** Users with similar trajectories in medium-large trajectories group



**(a)** Users with similar mobility pattern




**(b)** Similarity score distribution for selected score range

**Figure 6.7:** Users sharing similar mobility patterns and similarity score distribution for selected score range from medium-large trajectories group

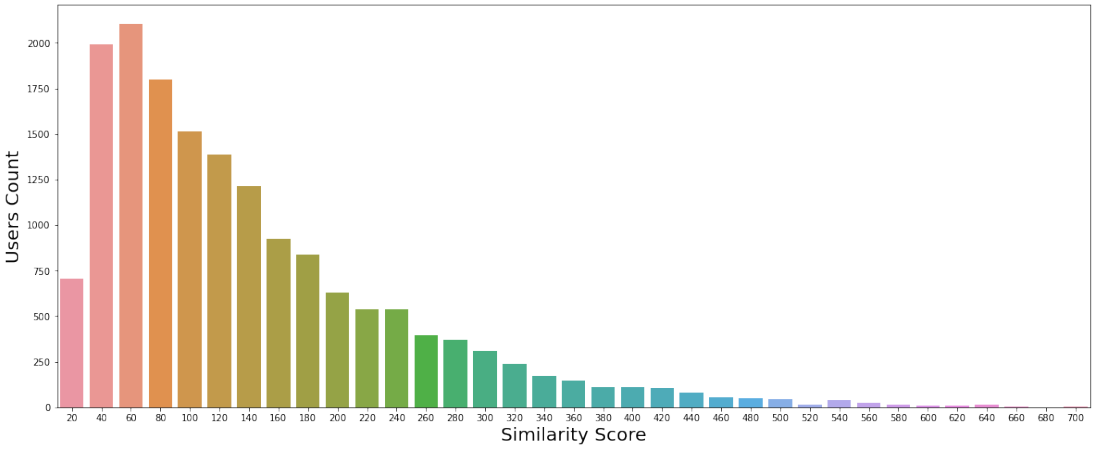



**Figure 6.8:** Similarity score of users with medium-large trajectories

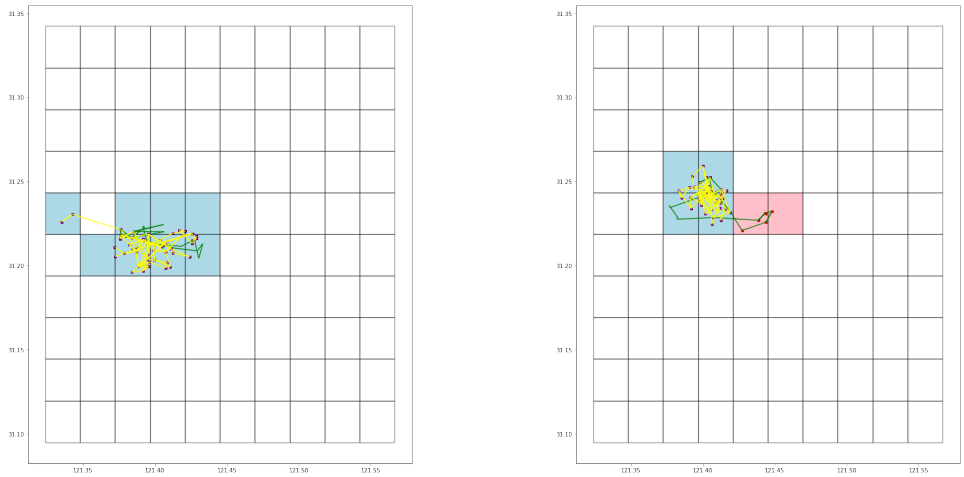
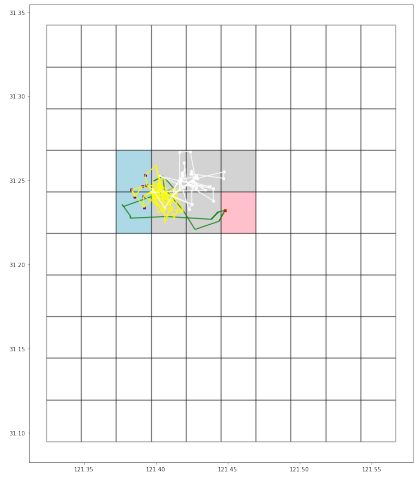


**Figure 6.9:** Users exhibiting least similarity scores in trajectories from large trajectories group

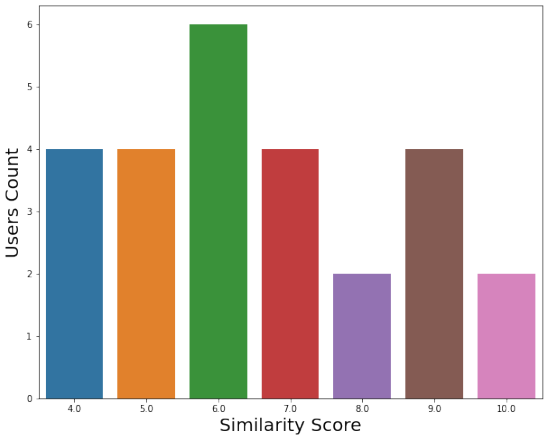
Additionally, we analyze further the partial similarity of long trajectories. Figure 6.10a shows the trajectories of users that scored a similarity between 4 and 6. From the Figure 6.10a, we can observe that shared mobility patterns occurred just between specific cells. Figure 6.10b illustrates further the most similar trajectories that were identified in this group. In addition, we also illustrate the overall distribution of similarity scores for all users in Figure 6.11. From the Figure, we can observe that a very small portion of users shares similar trajectories by overlapping at various locations over the grid. Interestingly, we observe that the trajectories that are the most similar are also the shortest that can be found in this group. All in all, our results suggest that it is easier to find similarities between short trajectories rather than larger ones. This thus implies that similarity analysis of user mobility should be the focus on small trajectory segments.

## 6.2 Characterizing similarity of homogeneous trajectories

In the previous section, we demonstrated that DTW could be used to compute similarity scores between trajectories, such that it is possible to find users with similar mobility patterns. We also showed that trajectories with short lengths are preferable for identifying similarity rather than long length trajectories. As a result, in this section, we



**(a)** Users with similar mobility pattern




**(b)** Similarity score distribution for selected score range

**Figure 6.10:** Users sharing similar mobility patterns and similarity score distribution for selected score range from large trajectories group

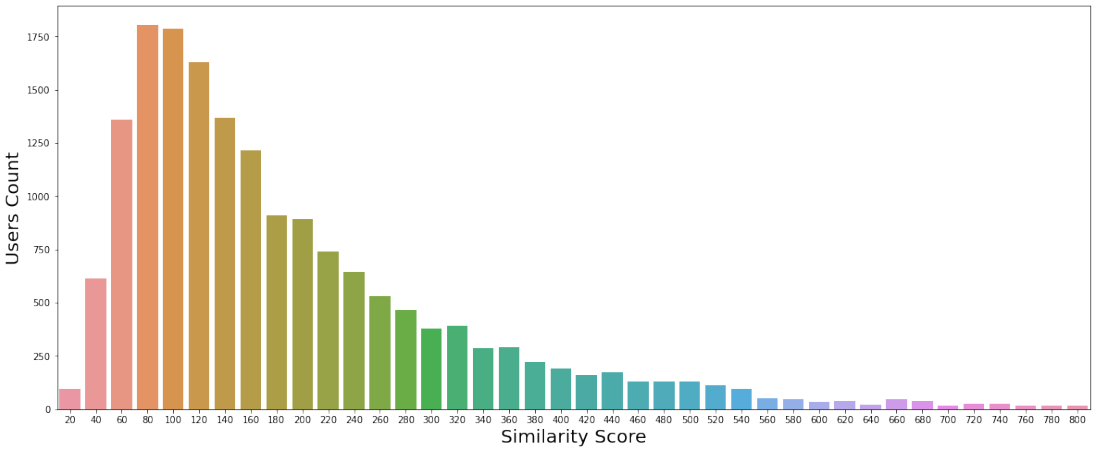
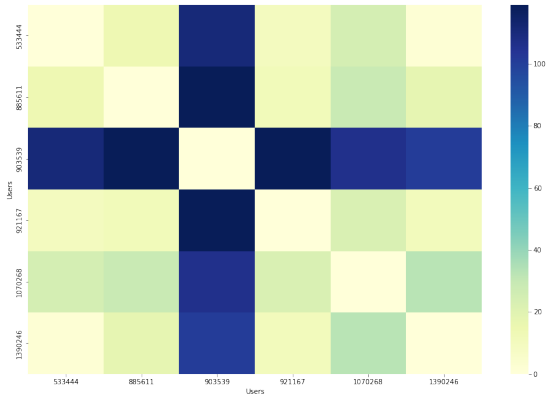



**Figure 6.11:** Similarity score of users with large trajectories

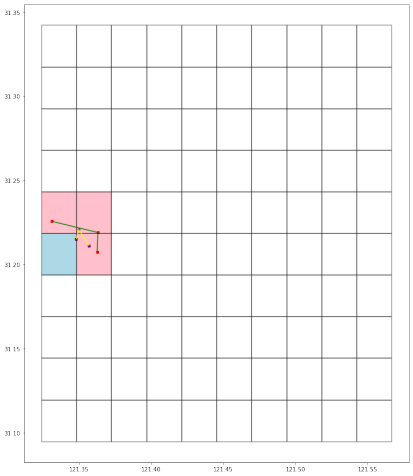
analyze further the individual characteristics of different types of trajectories. Thus, we first analyze the similarity in homogeneous trajectories.

**Group 1: Short trajectories:** We selected the top 6 users from this type of users to perform this analysis. Figure 6.12 from the results. From the figure, we can observe the similarity scores between trajectories of different users. We can observe values ranging between 2 to 119. Overall scores distributions of our analysis can be visualized as a heat-map in Figure 6.12a. Users whose trajectories are similar scored smaller values, while very different trajectories scored higher values. Figure 6.12b shows the similar mobility we have found is between users 533444 and 1390246 with a score of 2. Similarly, users 533444 and 921167 scored a similarity value of 10 as shown in Figure 6.12c. As the similarity score is low, we can observe that both trajectories highly overlap and also that those share similar spatial characteristics (meaning approximately the same location). Likewise, as shown further in the figure, similarity scores start increasing until reaching a score value of 119. User 903539 showed the maximum difference with the other users from the group. This difference in similarity can be appreciated further in Figure 6.12d. Pink-coloured cells depict the trajectory for user 903539, while blue coloured cells show the trajectory of 921167, demonstrating further the trajectories are highly different and do not share any overlapping.

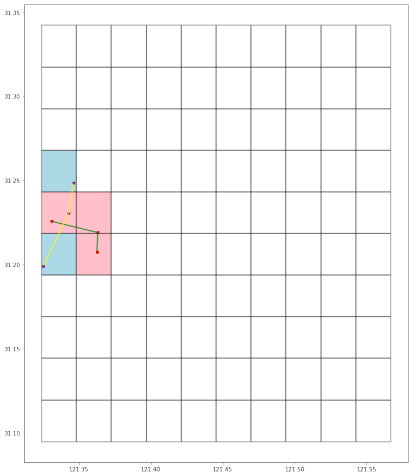
**Group 2: Short to medium trajectories:** We found a considerable amount of users with medium trajectory length (between 3 and 8) from the data. This means that we have more user distribution over urban areas. To analyze the spatial behaviour, we selected ten users from this group. Figure 6.13a shows the similarity score between trajectories of selected users showing similar trajectories with less score and variant trajectories with higher scores. Scores values range between 15 to 409. Higher scores mean a significant difference in the trajectories. The overall score distribution of our analysis can be visualized as a heatmap in Figure 6.13a. The least similarity score came out to be 15, which depicts the sparsity in mobility distribution. Figure 6.13b shows the mobility of users 383866 and 1718360 with the least score from the group. Similarly, Figure 6.13c shows the mobility of users 80020 and 723840 being the second least calculated score. Due to higher scores, we can not find similar mobility. However, we observed overlapping in a few locations. The highest similarity score was calculated



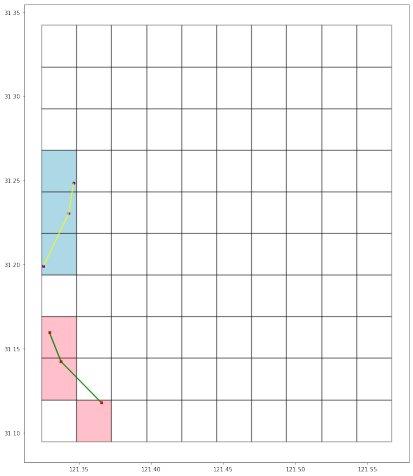

**(a)** Similarity score distribution



**(b)** Trajectory of users with the similarity score of two



**(c)** Users trajectories with the second least calculated similarity score



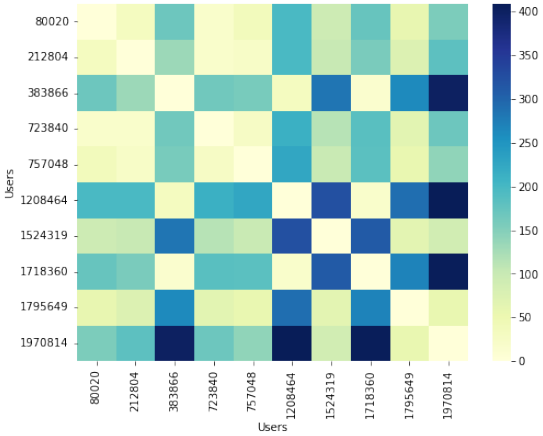
**(d)** Users trajectories with highest similarity score(most distant users)

**Figure 6.12:** Users with similar trajectories from short trajectories group

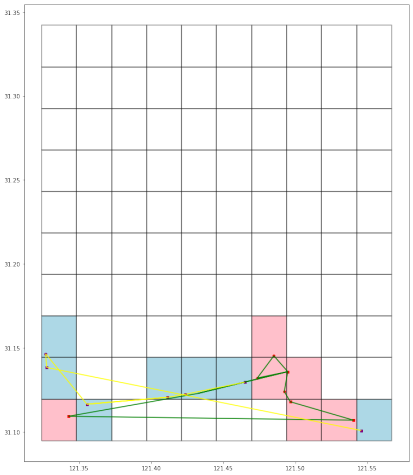
as 409, which shows no similarity among the users. Figure 6.13d shows the mobility of users with the most different mobility demonstrating no overlapping or similarity at any point.

**Group 3: Medium to large trajectories:** We performed a systematic evaluation on the user group organized by the users from medium to large trajectories. We quantified the similarity in trajectories on selected eight users. As trajectories are not confined to fewer urban areas, we have more dispersed mobility. Figure 6.14a depicts the distribution of the scores for selected users. Furthermore, the scores scale exhibits the range between 34 to 604 visualized as a heatmap. We got a higher similarity score, meaning more differences between trajectories. The lowest calculated score is 34 disclosing the difference in trajectories. Figure 6.14b and 6.14c shows the mobility of the top two interactions, which validates the differences in the trajectories. Likewise, as shown in Figure 6.14d, users with the highest score showed off the extreme difference in the trajectories over the urban area. Trajectories cover around 20% of the overall area of the grid. Besides covering a greater area of the gird, trajectories don't show any similarity yet overlapping at certain urban locations. Furthermore, Figure 6.14d exhibits the trajectories of the user with the highest similarity score covering more than 20% of the area. Overall quantification showed very few users with the score range between 30 to 40. Hence demonstrating the least similarity of users within the group.

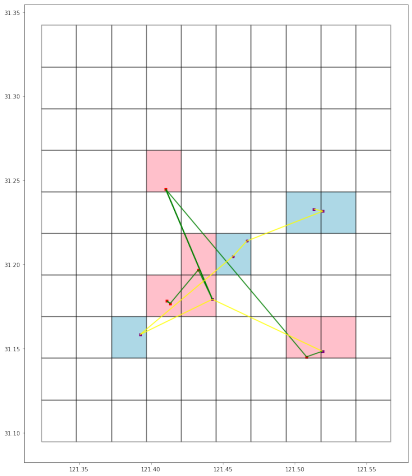
**Group 4: Large trajectories:** The users with the largest trajectories fall in group 4. Despite having 132 users in the group, we screened out the top ten users from the group for matching trajectories. The group holds the users with the significant trajectories means more spread movements. Figure 6.15a shows the similarity scores in the form of a heatmap. The least value of the similarity score is 45, while 453 is the highest calculated score. Both upper and lower boundaries of scores are outside considerable values for similar trajectories. Moreover, we got relatively smallish users lying in the score range of 40 to 50. In that case, we visualized the trajectories of the top two scores (being 45) as seen in Figure 6.15b and Figure 6.15c. Trajectories cover around 50% of the overall area of the grid. Besides covering a larger area of the gird, trajectories don't show any similarity yet overlapping at certain urban locations. Furthermore, Figure 6.15d exhibits the trajectories of the user with the highest similarity score covering



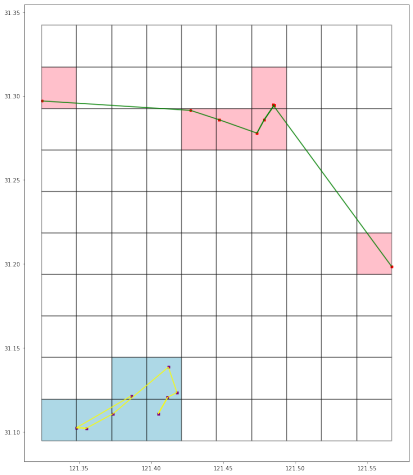

**(a)** Similarity score distribution



**(b)** Users trajectories with the similarity score of 15

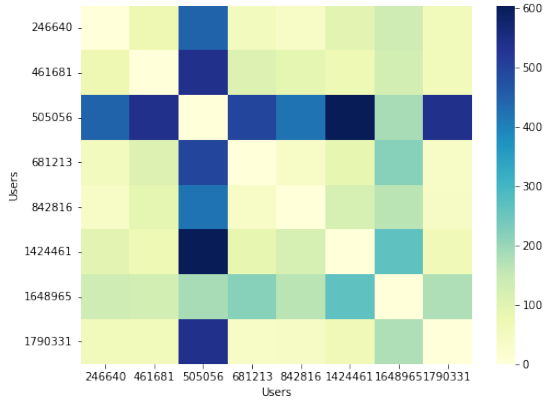


**(c)** Users trajectories with the similarity score of 18score(most distant users)

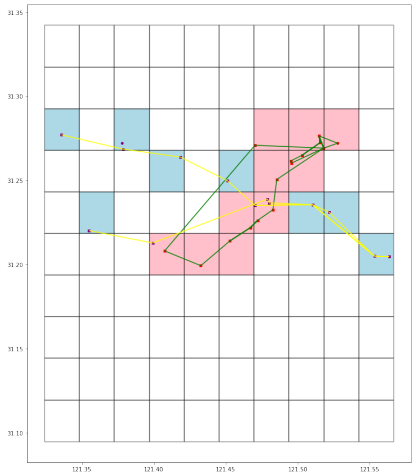


**(d)** Users trajectories with highest similarity

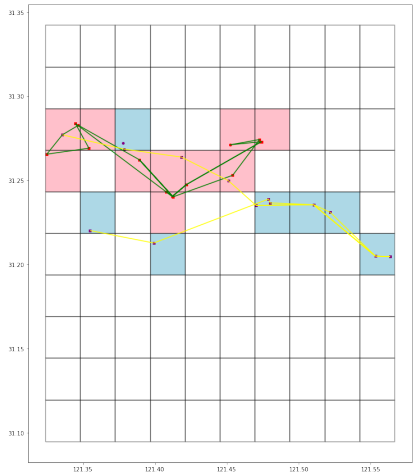
**Figure 6.13:** Users with similar trajectories from short-medium trajectories group



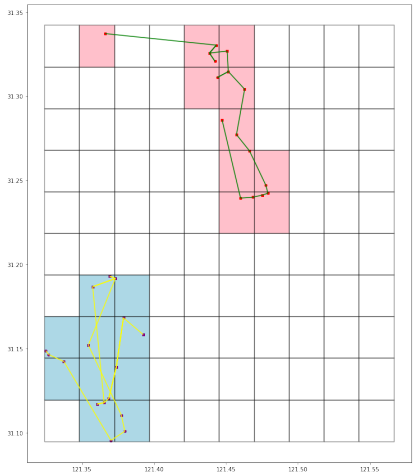

**(a)** Similarity score distribution



**(b)** Users trajectories with the least similarity score



**(c)** Users trajectories with second least similarity score



**(d)** Users trajectories with highest similarity score(most distant users)

**Figure 6.14:** Users with similar trajectories from medium-large trajectories group

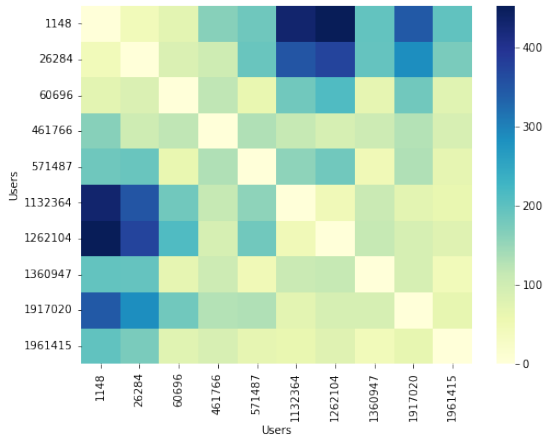
more than 50% of the area. Results demonstrated that in spite of having large trajectories, we found less promising results as compared to all other groups under consideration.

## 6.3 Characterizing similarity of heterogeneous trajectories

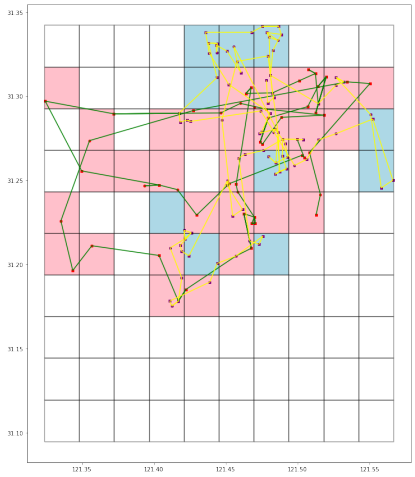
We proceeded next to to analyze the similarity between heterogeneous trajectories, meaning we compare trajectories of different lengths that capture a wide variety of mobility patterns.

**Characterizing similarity between Group 1 and Group 3:** We first characterize the similarity between users from short trajectories (group 1) and medium-large trajectories (group 3). Selective users were considered from both groups to explore similarities across supposed groups. Upon calculating the similarity score, a total of 48 unique scores were extracted. Surprisingly, scores ranged from 30 to around 600. Figure 6.16a shows the heatmap of scores distribution between all considered users, which illustrates the diversity in similarity scores. Furthermore, Figure 6.16b illustrates the trajectories of four users, two from each group. Blue and green trajectory indicates the users with a short trajectory (length of three), while black and red represent trajectories of users from a medium-large group. Likewise, Figure 6.16c exhibits the trajectories of users with average scores from the considered users. Green trajectory exposes the user from the short trajectory group while others show group 3. Group 3 users matched the trajectories of group 1 users but did not produce identical characterization. In contrast, Figure 6.16d demonstrates users' trajectories with the highest calculated scores, which shows the difference in trajectories of group 3 and group 1 user. This characterization helped to find more similarity in trajectories despite having more outstanding similarity scores. However, it is due to the fact of having a significant difference in trajectories length.

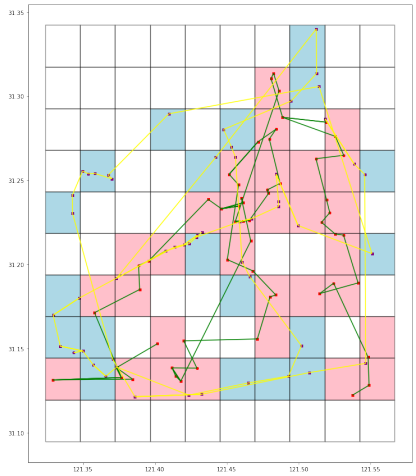
**Characterizing similarity between Group 1 and Group 4:** Next, we characterize differences between short (Group 1) and large trajectories (Group 4). Both groups contain nonidentical users; hence, the similarity score was more diverse than previous heterogeneous comparisons. Figure 6.17a exhibits the resultant similarity scores that



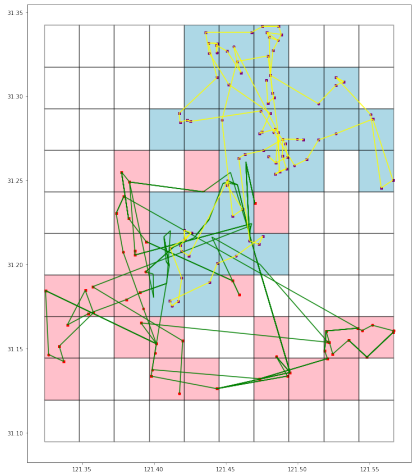

**(a)** Similarity score distribution



**(b)** Users trajectories with the least similarity score

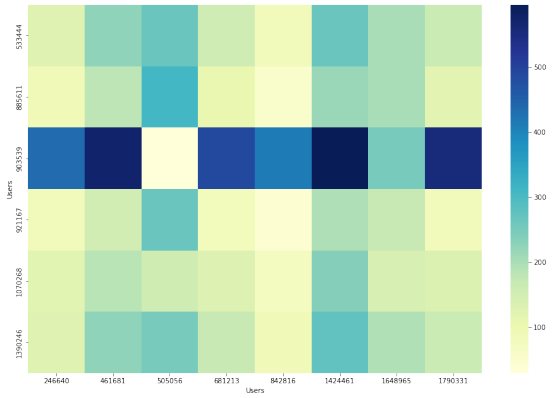


**(c)** Users trajectories with the second least similarity score

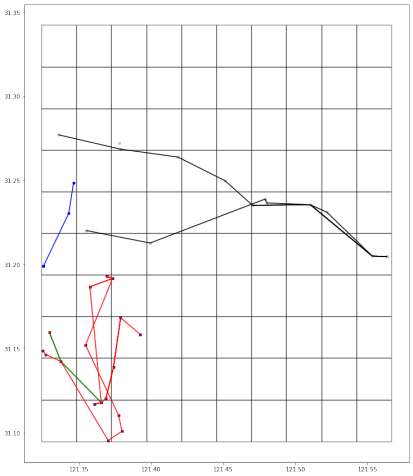


**(d)** Users trajectories with highest similarity score(most distant users)

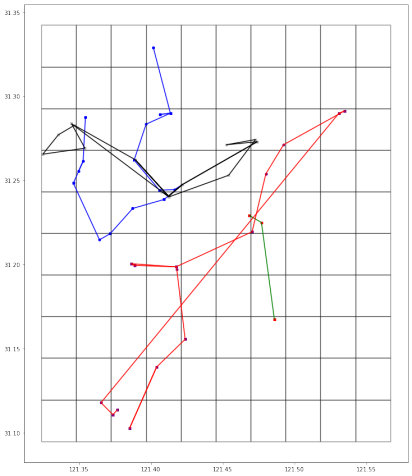
**Figure 6.15:** Users with similar trajectories from large trajectories group



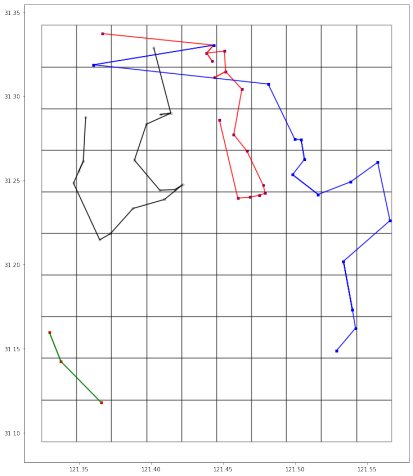

**(a)** Similarity score distribution




**(b)** Trajectories of users with lower similarity score




**(c)** Trajectories of users with medium similarity score




**(d)** Trajectories of users with higher similarity score

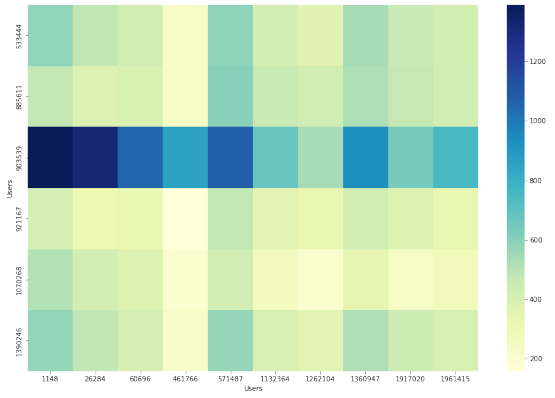
**Figure 6.16:** Characterizing similarity between Group 1(short trajectories) and Group 3(medium-large trajectories)

came out to be starting from 158, which is a number out of scope to be considered similar. However, visualization showed the large trajectories overlapping and containing the short trajectories. The maximum score is about 1400, which is found between users with the largest trajectory and users from the short trajectory group. Figures 6.17b, 6.17c, and 6.17d demonstrate the trajectories of users with low, medium, and high similarity scores, respectively. As in the Figure 6.17, we can observe that short trajectories are either contained by or overlapped with the large trajectories irrespective of score difference. This characterization helped to extract the heterogeneous behaviour of user groups which demonstrates that short trajectories and large trajectories users can not be grouped to extract group mobility. However, short trajectory users (even with higher similarity scores) might show similar trajectories to the large trajectory users overlapping at certain points.

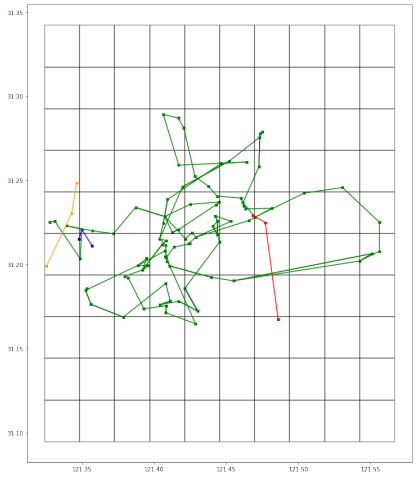
**Characterizing similarity between Group 2 and Group 4:** We then proceed to characterize differences between short-medium (Group 2) with large (Group 4) trajectories. Users from both groups possess more significant trajectories; hence the resulting scores are higher than homogeneous similarity scores. Scores ranged from 69 to 1604, which showed higher variance due to more differences in trajectories length. Figure 6.18a illustrates the distribution of the overall score among users. From the Figure 6.18a, we can observe that most similarity scores lie between 150 to 300, which elaborated that users from the upper bounds of group 2 and lower bounds of group 4 have more similarities. Visualization of trajectories of the low, medium, and high scores distribution are demonstrated in Figures 6.18b, 6.18c, and 6.18d, respectively. We observed that heterogeneous characterization would give favourable results if the trajectories' lengths are equivalent to each other. Moreover, comparing trajectories with considerable differences is not a good approach.

## 6.4 Identifying group mobility

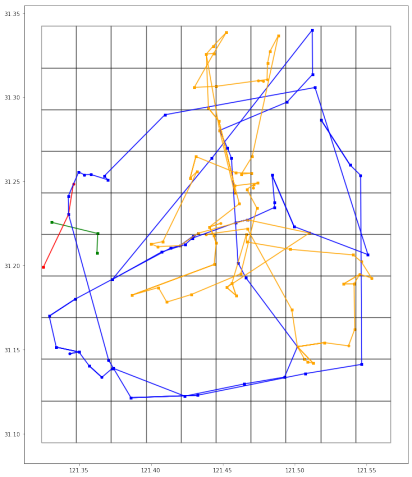
In our experiments, we by far have demonstrated that a similarity score between trajectories primarily indicates partial overlapping between mobility patterns of users. Indeed, the equal overlap between trajectories is possible but uncommon. We also showed that



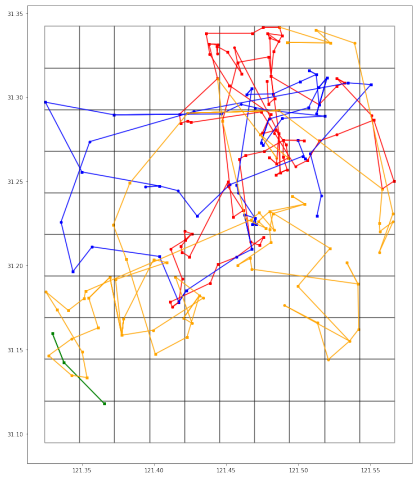

**(a)** Similarity score distribution




**(b)** Trajectories with lower similarity score

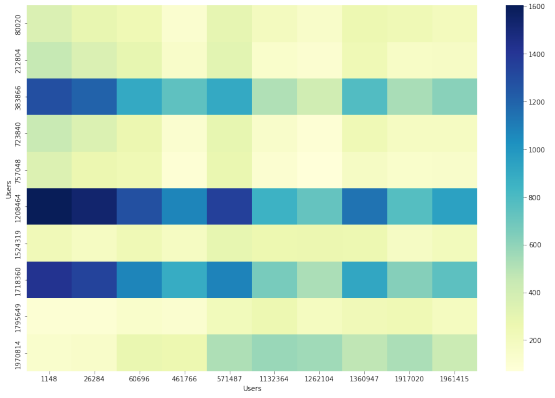



**(c)** Trajectories with medium similarity score

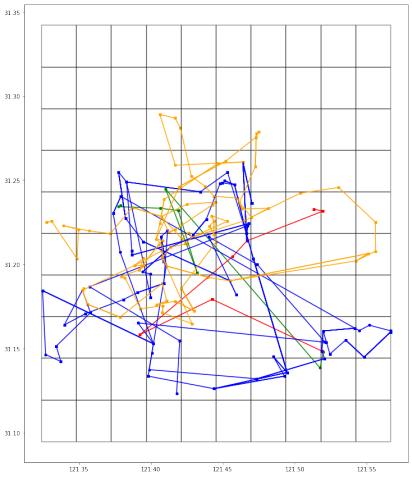



**(d)** Trajectories with higher similarity score

**Figure 6.17:** Characterizing similarity between Group 1(short trajectories) and Group 4(large trajectories)

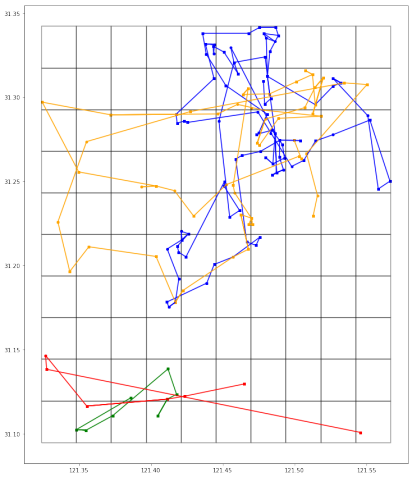



**(a)** Similarity score distribution



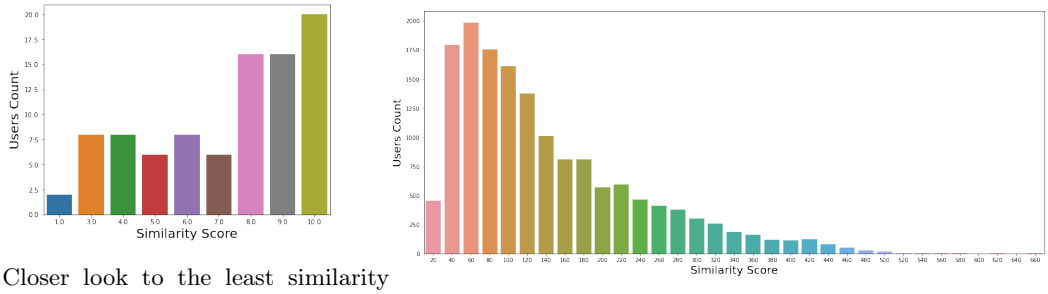
**(b)** Trajectories with lower similarity score

**(c)** Trajectories with medium similarity score



**(d)** Trajectories with higher similarity score

**Figure 6.18:** Characterizing similarity between Group 2(short-medium trajectories) and Group 4(large trajectories)




**(a)** Closer look to the least similarity scores

**(b)** Similarity score of all users

**Figure 6.19:** Similarity score distribution for identifying group mobility

when finding similarities between trajectories, the length of the trajectory is a key factor to consider to optimize the finding of similar trajectories. Thus, based on our experiments, short and same length trajectories provide the best results for finding users that move together. To validate this further, in this section, we then quantify to what extent a group of users move together based on the similarity of trajectories.

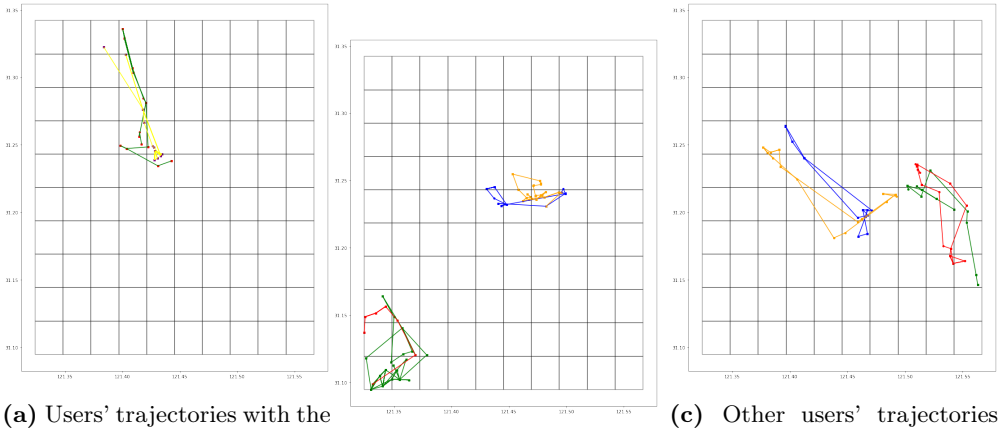
**Users moving together:** To perform this analysis, a group of users is produced, which contains users with homogeneous trajectories. As short length trajectories are preferable for finding users with similar mobility patterns (as demonstrated by our results), we selected an average trajectory length of six to create this group. Users with a trajectory length ranging between 5 and 8 (around our mean average of 6) are thus considered, producing a group with similar trajectory characteristics. Overall, 125 users fulfil these characteristics.

**Results:** As expected, meaningful results were obtained from this groups' selection criteria, meaning results showed more similarity than previously considered groups. For instance, we got a lower similarity score of below or equal to 10 from many users. Figure 6.19a illustrates the users' distribution over similarity score. From the Figure 6.19, we can observe the similarity found between users. Likewise, we evaluated the similarity score distribution for all users in the group. From Figure 6.19b, it can be observed that a substantial amount of users lies in more similar trajectories with a lower score. Moreover, most users belong in between a score of 40 to 80.

**Table 6.1:** Users with least simialrity score for group identification

| User 1 | User 2 | Score |
|---|---|---|
| 780684 | 767513 | 5 |
| 1085840 | 1556110 | 5 |
| 203763 | 793258 | 5 |
| 190146 | 767513 | 4 |
| 1158578 | 1085840 | 4 |
| 1396703 | 1689637 | 4 |
| 1344509 | 1062137 | 4 |
| 473134 | 542233 | 3 |
| 627128 | 941266 | 3 |
| 438845 | 2071437 | 3 |
| 1652160 | 1036343 | 3 |
| 1960133 | 849610 | 1 |

Table 6.1 presents further the users with the lowest similarity score. Since our goal is to find users moving together (group mobility), we then overlap the trajectories of these users to assess whether is possible to find users moving together just by looking at the similarity score of their trajectories. Figure 6.20a shows the users' trajectories with the fewer differences or, in other words, that are the more similar. Likewise, Figures 6.20b and 6.20c illustrates the trajectory of users with a similarity score of three. From the results, we can observe that while some trajectories of users overlap, the amount of users overlapping at the same time is small. Thus, it is possible to find small groups of users (partially) moving together. Figure 6.21 shows further the users' trajectories exhibiting similar scores and thus making a group formation. It is evident from these Figures( 6.20, 6.21) that few users exhibit almost identical trajectories while few share the trajectories at certain locations but differ at other places. Moreover, few users share similar starting locations, and few share similar ending locations while overlapping multiple times at different locations.



**(a)** Users' trajectories with the similarity score of one (identical group mobility)

**(b)** Users' trajectories with the similarity score of three

**(c)** Other users' trajectories with the same similarity score of three

**Figure 6.20:** Users' trajectories having least similarity score for exploring group mobility

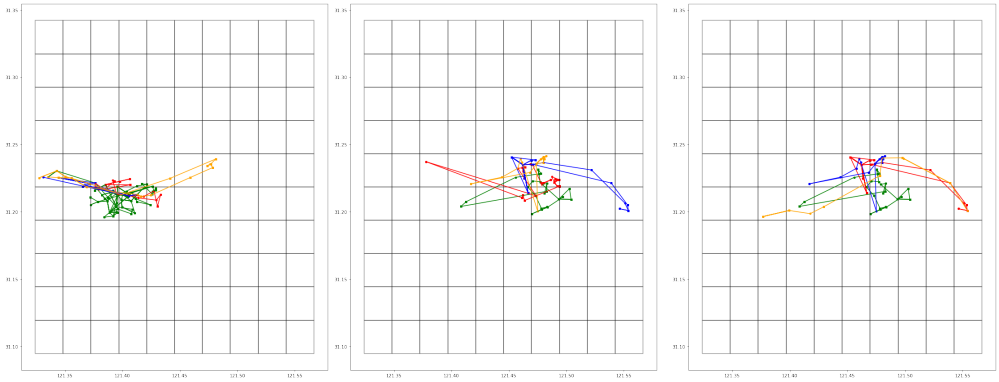


**Figure 6.21:** Users sharing similar mobility patterns (groups mobility)

## 6.5 Summary

This Chapter presented the results from performing a set of experiments to find the similarity between mobility patterns of users. Overall, characterizing the similarity of heterogeneous trajectories proved to be a promising approach for identifying users moving together. Short and same length trajectories are likely to provide better results when searching for users that move as a group, even if the groups that are found are small.

The next Chapter discusses the work's possible applications and limitations.

# 7

# Discussion

This Chapter discusses the implications and limitations of the work.

## 7.1 Room for improvement

We performed experiments in a controlled environment; however, our findings have room for improvement in multiple aspects that we would like to explore further. Some key improvements to consider:

**Directional mobility:** Our study analyzes trajectories irrespective of their direction in which the user was moving. Our approach can be enhanced by considering directions and finding similarities in the paths. This would help find users' flow during specified hours and the whole day.

**Time and data:** We have extracted a tiny portion of helpful information as we have one-day data only. The proposed approach can be applied to more volume of data to extract extensive valuable information. Moreover, we can check possible scenarios with hours distribution and daily routines when there is more data available.

**Grid size trade-off:** Performed experiments and results are based on the grid with denser users' selection. In addition, our analysis had a fixed grid size. Changing grid area and cells' size may give us different valuable results. Current work included a fixed 100 cells (10x10) grid, and changing this to having 200 cells or any other more significant number within the same area would result in more significant trajectories. However, this will increase users' congestion in one cell and may have more empty cells than the current work. There would be a trade-off between cells and base station distribution.

**Data sparsity:** We observed a few users very distant from each other during our experiments while remaining inside a selected area. Users' distribution was random, and the data sparsity affected the results in a greater manner. Removal of the sparsity issue would provide valuable insights into mobility similarity.

**Grid with same point selection:** Our work was based on grid selection and mobility between cell locations. However, our approach can be extended to have the exact grid location of devices having the same start location. The same method can be implemented on the exact ending location to find the precise location of users' mobility points. Afterwards, those starting or ending locations can be compared to the intermediate points of other users that pass through that location. This would help to identify users' interactions at different times.

## 7.2 Implications

We have got valuable results that can be utilized in multiple ways. A few critical implications are as follows:

**Carpool:** Our results can be applied directly to carpooling applications. Finding similarities in the mobility would result in sharing cars for the users following the identical routes. Thus reducing carbon footprints and making the environment more eco-friendly.

**Route congestion:** We explored group mobility and individual users' mobility similarity. We have discussed critical points for overlapping of users in some points. Results from our approach can be utilized to find the locations where most of the users collide, making congestion. With these congestion results, urban planners would benefit by analyzing the user needs on that location by analyzing missing and present resources.

**Grid selection:** In addition to the relationship between data produced and mobility, grid selection is another important key factor in finding similarities. Our results suggested that users moving more between grids can give us more mobility samples. Filtering users that stay in the exact grid location and changing only the base station was proved to be the novel approach in removing devices with the least mobility. Thus making it easy to find similarities between users who change location more often. The same filtering technique would help researchers find effective results in a reduced time.

# 8

# Summary and Conclusion

In this thesis, we have performed a systematic evaluation for analyzing group mobility. We conducted our analysis on a crowd sensed dataset collected by a cellular operator. With this information, we have then built trajectories that depict users mobility, and we have created different categories and groups to analyze them. We have quantified the differences (similarity score) between trajectories using a DTW algorithm. After a rigorous experimental benchmark, our results indicate that the best way of finding users moving together is by looking at trajectories that are short in length. Our results also indicate that the number of users moving together as a group is small. All in all, our results highlight that for analyzing users moving together as a group is better to divide a trajectory into small segments. Lastly, we also discussed the implications and limitations of our work.

# Bibliography


[1] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779–782, 2008. 1, 5

[2] Dongyoun Shin, Daniel Aliaga, Bige Tunçer, Stefan Müller Arisona, Sungah Kim, Dani Zünd, and Gerhard Schmitt. Urban sensing: Using smartphones for transportation mode classification. *Computers, Environment and Urban Systems*, 53:76–86, 2015. 1

[3] Humberto T Marques-Neto, Faber HZ Xavier, Wender Z Xavier, Carlos Henrique S Malab, Artur Ziviani, Lucas M Silveira, and Jussara M Almeida. Understanding human mobility and workload dynamics due to different large-scale events using mobile phone data. *Journal of Network and Systems Management*, 26(4):1079–1100, 2018. 1

[4] Anne Goodchild and Jordan Toy. Delivery by drone: An evaluation of unmanned aerial vehicle technology in reducing co2 emissions in the delivery service industry. *Transportation Research Part D: Transport and Environment*, 61:58–67, 2018. 1

[5] Ludovic Apvrille, Tullio Tanzi, and Jean-Luc Dugelay. Autonomous drones for assisting rescue services within the context of natural disasters. In *2014 XXXIth URSI General Assembly and Scientific Symposium (URSI GASS)*, pages 1–4. IEEE, 2014. 1

[6] Gustavo Romanillos Arroyo, Juan Carlos García Palomares, Borja Moya Gómez, Javier Gutiérrez Puebla, Javier Torres, Mario López, Oliva G Cantú-Ros, and Ricardo Herranz. The city turned off: Urban dynamics during the covid-19 pandemic based on mobile phone data. 2021. 1

[7] Corentin Cot, Giacomo Cacciapaglia, Anna Sigridur Islind, María Óskarsdóttir, and Francesco Sannino. Impact of us vaccination strategy on covid-19 wave dynamics. *Scientific Reports*, 11(1):1–11, 2021. 1

[8] Yuren Zhou, Billy Pik Lik Lau, Chau Yuen, Bige Tunçer, and Erik Wilhelm. Understanding urban human mobility through crowdsensed data. *IEEE Communications Magazine*, 56(11):52–59, 2018. 1

[9] David Ryan, Simon Denman, Clinton Fookes, and Sridha Sridharan. Crowd counting using multiple local features. In *2009 Digital Image Computing: Techniques and Applications*, pages 81–88. IEEE, 2009. 1

[10] Yufeng Wang, Xueyu Jia, Qun Jin, and Jianhua Ma. Mobile crowdsourcing: framework, challenges, and solutions. *Concurrency and Computation: Practice and experience*, 29(3):e3789, 2017. 1

[11] Stefan Höffken and Bernd Streich. Mobile participation: Citizen engagement in urban planning via smartphones. In *Citizen E-Participation in urban governance: crowdsourcing and collaborative creativity*, pages 199–225. IGI Global, 2013. 1

[12] Martin W Traunmueller, Nicholas Johnson, Awais Malik, and Constantine E Kontokosta. Digital footprints: Using wifi probe and locational data to analyze human mobility trajectories in cities. *Computers, Environment and Urban Systems*, 72:4–12, 2018. 1

[13] Jiechao Zhang, Samiul Hasan, Kamol Chandra Roy, and Xuedong Yan. Predicting individual mobility behavior of ride-hailing service users considering heterogeneity of trip purposes. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3685–3690, 2021. 5, 9

[14] Huber Flores, Agustin Zuniga, Leonardo Tonetto, Tristan Braud, Pan Hui, Yong Li, Sasu Tarkoma, Mostafa Ammar, and Petteri Nurmi. Collaboration stability: Quantifying the success and failure of opportunistic collaboration. pages 1–8. 5, 19

[15] Jing Yuan, Yu Zheng, and Xing 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*, pages 186–194, 2012. 5, 6

[16] Wenwen Li, Shaohua Wang, Xiaoyi Zhang, Qingren Jia, and Yuanyuan Tian. Understanding intra-urban human mobility through an exploratory spatiotemporal analysis of bike-sharing trajectories. *International Journal of Geographical Information Science*, 34(12):2451–2474, 2020. 5

[17] Huber Flores, Agustin Zuniga, Farbod Faghihi, Xin Li, Samuli Hemminki, Sasu Tarkoma, Pan Hui, and Petteri Nurmi. Cosine: Collaborator selector for cooperative multi-device sensing and computing. In *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–10. IEEE, 2020. 5

[18] Katayoun Farrahi and Daniel Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):1–27, 2011. 5

[19] Francesco Calabrese, Giusy Di Lorenzo, and Carlo Ratti. Human mobility prediction based on individual and collective geographical preferences. In *13th International IEEE Conference on Intelligent Transportation Systems*, pages 312–317, 2010. 6

[20] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018. 6

[21] Li Shi, Guanghua Chi, Xi Liu, and Yu Liu. Human mobility patterns in different communities: a mobile phone data-based social network approach. *Annals of GIS*, 21(1):15–26, 2015. 6

[22] Artemis Psaltoglou and Eusebi Calle. Enhanced connectivity index–a new measure for identifying critical points in urban public transportation networks. *International Journal of Critical Infrastructure Protection*, 21:22–32, 2018. 6

[23] Zeinab Ebrahimpour, Wanggen Wan, José Luis Velázquez García, Ofelia Cervantes, and Li Hou. Analyzing social-geographic human mobility patterns using large-scale social media data. *ISPRS International Journal of Geo-Information*, 9(2):125, 2020. 6

[24] Raghu K Ganti, Fan Ye, and Hui Lei. Mobile crowdsensing: current state and future challenges. *IEEE communications Magazine*, 49(11):32–39, 2011. 7

[25] Huadong Ma, Dong Zhao, and Peiyan Yuan. Opportunities in mobile crowd sensing. *IEEE Communications Magazine*, 52(8):29–35, 2014. 7

[26] Jinwei Liu, Haiying Shen, Husnu S. Narman, Wingyan Chung, and Zongfang Lin. A survey of mobile crowdsensing techniques: A critical component for the internet of things. *ACM Trans. Cyber-Phys. Syst.*, 2(3), jun 2018. 7

[27] Arvind Thiagarajan, Lenin Ravindranath, Katrina LaCurts, Samuel Madden, Hari Balakrishnan, Sivan Toledo, and Jakob Eriksson. Vtrack: accurate, energy-aware road traffic delay estimation using mobile phones. In *Proceedings of the 7th ACM conference on embedded networked sensor systems*, pages 85–98, 2009. 7

[28] Jinwei Liu, Haiying Shen, and Xiang Zhang. A survey of mobile crowdsensing techniques: A critical component for the internet of things. In *2016 25th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–6, 2016. 7

[29] Gilles Virone, A Wood, Leo Selavo, Quihua Cao, Lei Fang, Thao Doan, Zhimin He, and J Stankovic. An advanced wireless sensor network for health monitoring. In *Transdisciplinary conference on distributed diagnosis and home healthcare (D2H2)*, pages 2–4. Citeseer, 2006. 7

[30] Raghu K. Ganti, Fan Ye, and Hui Lei. Mobile crowdsensing: current state and future challenges. *IEEE Communications Magazine*, 49(11):32–39, 2011. 8

[31] Ruipeng Gao, Mingmin Zhao, Tao Ye, Fan Ye, Yizhou Wang, Kaigui Bian, Tao Wang, and Xiaoming Li. Jigsaw: Indoor floor plan reconstruction via mobile crowdsensing. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*,

MobiCom '14, page 249–260, New York, NY, USA, 2014. Association for Computing Machinery. 8

[32] Chi Zhang, Kalyan P Subbu, Jun Luo, and Jianxin Wu. Groping: Geomagnetism and crowdsensing powered indoor navigation. *IEEE Transactions on Mobile Computing*, 14(2):387–400, 2014. 8

[33] Zhe Peng, Shang Gao, Bin Xiao, Songtao Guo, and Yuanyuan Yang. Crowdgis: Updating digital maps via mobile crowdsensing. *IEEE Transactions on Automation Science and Engineering*, 15(1):369–380, 2018. 8

[34] Khalid Abualsaud, Tarek M. Elfouly, Tamer Khattab, Elias Yaacoub, Loay Sabry Ismail, Mohamed Hossam Ahmed, and Mohsen Guizani. A survey on mobile crowd-sensing and its applications in the iot era. *IEEE Access*, 7:3855–3881, 2019. 8

[35] Abbas M Ali Al-muqarm and Furkan Rabee. Iot technologies for mobile crowd sensing in smart cities. *J. Commun.*, 14(8):745–757, 2019. 8

[36] Anastasios Zopiatis, Antonis L Theocharous, Petros C Kosmas, Craig Webster, and Yioula Melanthiou. Developing a country-wide tourist loyalty scheme: A barren landscape. *International Journal of Tourism Research*, 18(6):579–590, 2016. 8

[37] Yunchuan Sun, Houbing Song, Antonio J Jara, and Rongfang Bie. Internet of things and big data analytics for smart and connected communities. *IEEE access*, 4:766–773, 2016. 8

[38] Alan C Acock. Working with missing values. *Journal of Marriage and family*, 67(4):1012–1028, 2005. 8

[39] Martínez-Plumed Fernando, Ferri Cèsar, Nieves David, and Hernández-Orallo José. Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, 2021. 8

[40] A Rogier T Donders, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091, 2006. 8

[41] Edgar Acuna and Caroline Rodriguez. The treatment of missing values and its effect on classifier accuracy. In *Classification, clustering, and data mining applications*, pages 639–647. Springer, 2004. 8

[42] Jianglin Huang, Yan-Fu Li, and Min Xie. An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and software Technology*, 67:108–127, 2015. 8

[43] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:, 1994. 9

[44] Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 262–270, 2012. 9

[45] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007. 9

[46] Pavel Senin. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, 855(1-23):40, 2008. 9

[47] Chang Wei Tan, Matthieu Herrmann, Germain Forestier, Geoffrey I Webb, and Francois Petitjean. Efficient search of the best warping window for dynamic time warping. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 225–233. SIAM, 2018. 9

[48] Daiping Wei, Xiaofeng Liu, Bangxin Wang, Zhi Tang, and Lin Bo. Damage quantification of aluminum plates using sc-dtw method based on lamb waves. *Measurement Science and Technology*, 2021. 9

[49] Ibrahim Ahmed, Enrico Zio, and Gyunyoung Heo. Fault detection by signal reconstruction in nuclear power plants. 2021. 9

[50] Dongdong Li, Kohei Kaminishi, Ryosuke Chiba, Kaoru Takakusaki, Masahiko Mukaino, and Jun Ota. Evaluation of postural sway in post-stroke patients by dynamic time warping clustering. *Frontiers in Human Neuroscience*, 15, 2021. 9

[51] Milad Jabbari, Rami N. Khushaba, and Kianoush Nazarpour. Combined dynamic time warping and spatiotemporal attention for myoelectric control. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 5940–5943, 2021. 9

[52] Raihan Rafif, Sandiaga Swahyu Kusuma, Siti Saringatin, Giara Iman Nanda, Pramaditya Wicaksono, and Sanjiwana Arjasakusuma. Crop intensity mapping using dynamic time warping and machine learning from multi-temporal planetscope data. *Land*, 10(12), 2021. 9

[53] Shaghayegh Gharghabi, Shima Imani, Anthony Bagnall, Amirali Darvishzadeh, and Eamonn Keogh. An ultra-fast time series distance measure to allow data mining in more complex real-world deployments. *Data Mining and Knowledge Discovery*, 34:1104–1135, 2020. 9

[54] Xinmin Tang, Xiaoqi Ji, and Jinan Liu. Predicting aircraft taxiing estimated time of arrival by cluster analysis. *IET Intelligent Transport Systems*, 2021. 9

[55] Sana Boujnah, Xianfang Sun, David Marshall, Paul L Rosin, and Mohamed Lassaad Ammari. A novel approach for speaker recognition in degraded conditions. In *Advanced Methods for Human Biometrics*, pages 139–146. Springer, 2021. 9

[56] Yining Qiu, Jiale Ding, Mengxiao Wang, Linshu Hu, and Feng Zhang. Understanding the urban life pattern of young people from delivery data. *Computational Urban Science*, 1(1):1–16, 2021. 9

[57] Chang Wei Tan, Matthieu Herrmann, and Geoffrey I Webb. Ultra fast warping window optimization for dynamic time warping. *IEEE International Conference on Data Mining (ICDM 2021)*, 2021. 9

[58] Nehal Magdy, Mahmoud A Sakr, Tamer Mostafa, and Khaled El-Bahnasy. Review on trajectory similarity measures. In *2015 IEEE seventh international conference on Intelligent Computing and Information Systems (ICICIS)*, pages 613–619. IEEE, 2015. 9

[59] John Koetsier. There are now 8.9 million mobile apps, and china is 40% of mobile app spending, Jun 2021. 9

[60] Mike Hazas, Janine Morley, Oliver Bates, and Adrian Friday. Are there limits to growth in data traffic? on time use, data generation and speed. In *Proceedings of the second workshop on computing within limits*, pages 1–5, 2016. 9

[61] Feng Xia, Jinzhong Wang, Xiangjie Kong, Zhibo Wang, Jianxin Li, and Chengfei Liu. Exploring human mobility patterns in urban scenarios: A trajectory data perspective. *IEEE Communications Magazine*, 56(3):142–149, 2018. 9

[62] Irena Pawlyszyn, Halyna Ryzhkova, et al. Methodical aspects of planning sustainable urban mobility. *European Research Studies Journal*, 24(Special 5):344–365, 2021. 9

[63] Theo Demessance, Chongke Bi, Sonia Djebali, and Guillaume Guérard. Hidden markov model to predict tourists visited places. In *2021 22nd IEEE International Conference on Mobile Data Management (MDM)*, pages 209–216, 2021. 9

[64] Azalden Alsger, Ahmad Tavassoli, Mahmoud Mesbah, Luis Ferreira, and Mark Hickman. Public transport trip purpose inference using smart card fare data. *Transportation Research Part C: Emerging Technologies*, 87:123–137, 2018. 9

[65] Brynn L Hudgins, Stephanie P Kurti, Elizabeth S Edwards, and Trent A Hargens. The impact of the covid-19 pandemic on physical activity habits at a residential university. *Journal of American College Health*, pages 1–6, 2021. 9

[66] Mark Birkin. Spatial data analytics of mobility with consumer data. *Journal of Transport Geography*, 76:245–253, 2019. 9

[67] Chen Zhao, An Zeng, and Chi Ho Yeung. Characteristics of human mobility patterns revealed by high-frequency cell-phone position data. *EPJ Data Science*, 10(1):5, 12 2021. 9

[68] Andres Fielbaum, Rafał Kucharski, Oded Cats, and Javier Alonso-Mora. How to split the costs and charge the travellers sharing a ride? aligning system's optimum with users' equilibrium. *European Journal of Operational Research*, 2021. 9

[69] Mojdeh Sharafi. *Station and city-level modelling of bike-sharing system for Montreal*. PhD thesis, 2021. 9

[70] Xiaoyan Hong, Mario Gerla, Guangyu Pei, and Ching-Chuan Chiang. A group mobility model for ad hoc wireless networks. In *Proceedings of the 2nd ACM international workshop on Modeling, analysis and simulation of wireless and mobile systems*, pages 53–60, 1999. 10

[71] Wen-Tsuen Chen and Po-Yu Chen. Group mobility management in wireless ad hoc networks. In *2003 IEEE 58th Vehicular Technology Conference. VTC 2003-Fall (IEEE Cat. No. 03CH37484)*, volume 4, pages 2202–2206. IEEE, 2003. 10

[72] Cherry Ye Aung, Boon Chong Seet, Mingyang Zhang, Ling Fu Xie, and Peter Han Joo Chong. A review of group mobility models for mobile ad hoc networks. *Wireless Personal Communications*, 85(3):1317–1331, 2015. 10

[73] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007. 10

[74] Krishna Karra, Caitlin Kontgis, Zoe Statman-Weil, Joseph C Mazzariello, Mark Mathis, and Steven P Brumby. Global land use/land cover with sentinel 2 and deep learning. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 4704–4707. IEEE, 2021. 14

# 9

# Appendix

## 9.1 Licence

**Non-exclusive licence to reproduce thesis and make thesis public**

**I, Mubashar Shahzad,**

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   **Revisiting Group Mobility Modelling: A Systematic Evaluation**,

   supervised by Huber Flores.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

## 9. APPENDIX

Mubashar Shahzad

**06/01/2022**