Mining Location and User Information from Users’ Trajectories

by

Marcello Tomasini

Laurea Magistrale in Ingegneria Informatica
Università di Modena e Reggio Emilia
Italy, 2013

Laurea in Ingegneria Informatica
Università di Modena e Reggio Emilia
Italy, 2010

A dissertation
submitted to Florida Institute of Technology
in partial fulfillment of the requirements
for the degree of

Doctorate of Philosophy
in
Computer Science

Melbourne, Florida
December, 2017
We the undersigned committee
hereby approve the attached dissertation

Mining Location and User Information from Users’ Trajectories
by
Marcello Tomasini

________________________
Ronaldo Menezes, Ph.D.
Professor
School of Computing
Committee Chair

________________________
Hector Gutierrez, Ph.D.
Professor
Mechanical Aerospace Engineering
Outside Committee Member

________________________
Eraldo Ribeiro, Ph.D.
Associate Professor
School of Computing
Committee Member

________________________
Marius Silaghi, Ph.D.
Associate Professor
School of Computing
Committee Member

________________________
Philip Bernhard, Ph.D.
Associate Professor and Director
School of Computing
ABSTRACT

Title:
Mining Location and User Information from Users’ Trajectories

Author:
Marcello Tomasini

Major Advisor:
Ronaldo Menezes, Ph.D.

The understanding of human mobility is integral to the advancement of many fields including public health, city planning, economic forecasting, and it attracts the interest of researchers from a broad number of fields. The work of such researchers would not have been possible without the availability of large localized datasets such as mobile phone Call Detail Records (CDRs) provided by telco operators and Location Based Social Networks (LBSNs) check-ins provided by popular applications such as Twitter, Facebook and Foursquare. However, such datasets are often incomplete, anonymized, and/or inaccurate. Hence there is a need to partially reconstruct or extract more information from such data. In this dissertation, we propose a framework to analyze and augment spatio-temporal data related to human trajectories that could be used to further the understanding of regularities in human mobility. In particular, the aim is to be able to augment the data to make it possible to classify users and places.
Table of Contents

Abstract iii

List of Figures vii

List of Tables ix

Acknowledgments x

Dedication xi

1 Introduction 1

2 Literature Review 6

2.1 Population Models ........................................... 7

2.1.1 The Origin-Destination Matrix ............................ 8

2.1.2 The Role of Distance ...................................... 9

2.2 Individual Mobility .......................................... 12

2.2.1 Trip Distance and Waiting Time Distribution .......... 13

2.2.2 Radius of Gyration ...................................... 15

2.2.3 Preferential Return ...................................... 16

2.2.4 Frequency vs Recency of Visited Locations .......... 18

2.2.5 Trajectory Entropy and Predictability ................. 19
A.1 POI Identification ............................................. 84
A.1.1 k-Means ..................................................... 85
A.1.2 Mean Shift .................................................. 85
A.1.3 HDBSCAN ............................................... 86
A.2 Extraction of Events and Activities ......................... 87
A.3 User Clustering .............................................. 88

B Evaluation of Location Filtering Criteria for Classification 91

C List of Publications ........................................... 96
C.1 Publications Related to This Dissertation .................... 96
C.2 Other Ph.D. Publications .................................... 97
List of Figures

3.1 Clustered GPS Points Projected on a Map ................. 29
3.2 Clustering Evaluation on Ground Truth Data .................. 30
3.3 Popular-times Histogram .......................................... 31
3.4 Popular-days Histogram ........................................... 32
3.5 Popular Times and Days by Category ....................... 32
3.6 Google Places API Query Example .............................. 34
3.7 Data Filtering for Classification ................................. 35
3.8 Confusion Matrix of Propagated Labels ....................... 38
3.9 Spatial Characterization of Events and Activities .......... 42
3.10 Temporal Characterization of Events ....................... 43
3.11 Temporal Characterization of Activities ................... 44
4.1 Filtering Steps to Prepare User Trajectories for Analysis .... 47
4.2 Hourly and Daily Entropy ......................................... 50
4.3 Entropy and Max Predictability Distribution of Users ........ 51
4.4 Relationship Between Radius of Gyration and Estimated Entropy . 52
4.5 Average Popular-times and Popular-days Feature Vectors per Cluster of Users ........................................... 55
4.6 Boxplot For Users Clustered Using Popular-times and Popular-days 56
## List of Tables

3.1 Overview of the Mobility Datasets ........................................ 27
3.2 Foursquare Labels Classification Performance ........................ 37
3.3 Google Labels Classification Performance ............................. 37
3.4 Foursquare Labels Distribution ........................................... 38
3.5 Topics Associated with the Feature Space Clusters ..................... 41
4.1 Overview of the Mobility Datasets for User Characterization ........ 48
4.2 Foursquare Check-ins Embedded in the Tweets .......................... 54
Acknowledgements

I would like to thank Dr. Ronaldo Menezes, my advisor, for providing me the opportunity to become a Ph.D student under his supervision while I was a visiting scholar for my master thesis. I will always be grateful for such opportunity and for all the guidance and support even through the rough times.

I would like to express my gratitude and appreciation for the committee members: Dr. Eraldo Ribeiro, Dr. Marius Silaghi, Dr. Hector Gutierrez. I had the pleasure to be both a student and a colleague to them. I learned a lot.

I am also grateful to the department staff, especially Karen Brown and Rosalyn Edwards, who had to deal with all my requests and questions.

I want to thank all the BioComplex Lab members for the interesting discussions and reciprocal exchange of knowledge; also, for being friends and walking with me this journey. Thanks to my house-mates over the years: Rahul, Ali, Tapas, Marie, and Gavin; we had a lot of fun. Special thanks to my family, especially my brother and sister, who always has my back. Thanks to my Italian childhood friends for their lifelong friendship.

Finally, I would like to thank my lovely girlfriend Nancy, aka Alexa, aka Awesomeness. She is my cheerleader, she is always right there when I need her.
Dedication

To my mom who made realizing my dreams possible.

To my dad who gifted me the seed of science.
Chapter 1

Introduction

The understanding of human mobility patterns is integral to the advancement of many fields including, but not limited to, public health [51], city planning [113], and economic forecasting [31]. With the recent availability of cell phone Call Detail Records (CDRs) and large localized datasets from social networking sites, the study of human mobility has experienced a massive growth. These datasets provide large amounts of mobility data thus allowing the study of many aspects of human dynamics.

While the study of people mobility is actually not new and started back in the 19th century with the study of population movement, it is with the seminal work of Brockmann et al. [17] in 2006 that it saw an explosive growth. Brockmann et al. revolutionized the modeling of human mobility by using bank notes as gateway of the individual’s mobility. When large phone Call Detail Records (CDRs) became available, researchers were able to identify and model several patterns of human mobility such as regular schedules and the fact that there are people who travels more than others [34], the tendency of people to go back to previously visited
locations [92] and that people movements are highly predictable [93], the tendency of people to go back to recently visited locations [9], and the connection between mobility and individuals’ social networks [42, 100]. Furthermore, part of the human mobility research tried to classify people in groups based on the characteristics of their movement. For example the characteristic displacement [69] or on the similarity of the user’s trajectories [109].

Besides CDRs, there is another source of mobility data that has been very useful for the study of human mobility. The data that come from Location Based Social Networks (LBSNs) have the advantage, compared to CDRs, to have contextual information associated with geographical positions, and therefore allows the extraction of more information related to the users. Scientists have been able to identify user’s information such as relevant locations to individual users [54] and give recommendations on future locations to visit based on location history [120, 111]. Furthermore, the trajectories augmented with information related to the activity performed or the purpose of movement allowed researchers to get better modeling of intra-urban mobility [108] and to identify urban mobility patterns and anomalies [32]. LBSNs data are also able to reveal information that is not related just to mobility, for instance each check-in is linked to a location type and possibly additional data. This in turn allowed to find connections between different types of locations [87], urban characteristics [103] and city neighborhoods [116]. LBSNs also include data on the social network of the users and thus they allow to explore how social ties and human mobility are intertwined [21, 82].

It is clear that CDRs and LBSNs data complement each other and current literature on human mobility put a strong emphasis on using a multitude of sources to obtain the best models and applications. Yet, when looking at geo-tagged
datasets, we are faced with several challenges. CDR data is usually limited to cellphone tower identifiers and timestamps, therefore lacking spatial resolution and any contextual information, while data from Location Based Social Networks (LBSNs) are sparse. Also, both sources are often not publicly available and it is even more rare to have both type of data available at the same time during the analysis. Furthermore, GPS traces show an unexpected problem regarding the high resolution of the data. The GPS information collected from several users creates a cloud of points representing the movements between specific locations (e.g. a bar, a coffee shop), however many different GPS coordinates exist within what we would call a single location. Hence, there is a need for a mechanism to define and exactly pinpoint the locations from this high-resolution cloud of GPS points. From these simple observations, we can infer several unwanted properties of the data in CDRs and LBSNs datasets:

- **Sparse**: there are few data records in a time period
- **Incomplete**: there are missing record data
- **Inaccurate**: the record data are noisy
- **Anonymized**: the record data have been hidden for e.g. privacy reasons
- **Not publicly available**: the data are not available through free or paid services

While we cannot do much regarding the availability and the sparseness as they are strongly dependent on specific dataset, we can improve an available mobility dataset by either extracting contextual information or inferring missing data.
Question 1: Can we extract contextual information at a fine spatial granularity (e.g. identify and characterize the locations or points of interest (POIs) visited by the people) from the GPS trajectories?

Gap 1. Most works do not attempt any fine grained characterization.

Description. CDR data does not allow for high resolution mapping. The availability of GPS traces and LBSNs data allows to analyze trajectories with a high degree of accuracy. However, different GPS sources have different degree of sparsity and offer more or less additional data (such as text or check-ins) making it difficult to have fine grained spatial characterization. It is important to identify a methodology and a set of features which are robust against sparsity or lack of additional data.

Gap 2. Most methods are application and dataset specific.

Description. Most of the literature employs ad-hoc data mining and machine learning techniques that rely on the availability of additional data than the trajectories. However, the highly recurrent and predictable human behavior provides evidence that it should be possible to characterize locations solely from GPS trajectories. Therefore, it is necessary to build methods and techniques to extract information from mobility datasets which can be applied generally and consistently to multiple sources.

Question 2: Can mobility patterns be disaggregated to reveal specific user’s trajectory properties and patterns?

Gap Human mobility studies mainly focus on aggregated user’s mobility scaling relationships.

Description There is a growing evidence that the observed scaling properties of
human mobility are actually the result of the convolution of different distributions originating from the user’s heterogeneity. It is important to develop methods and techniques that are able to profile users trajectories.

In this work, we introduce a framework to augment mobility related data by relying solely on geographic coordinates of social media posts (not exclusively the explicit check-ins) and associated timestamps. Spatio-temporal data in the form of GPS coordinates and associated timestamps can be used to extract more structured information from the users’ trajectories such as the position and the type of locations visited by the users, events and activities, types of users, patterns and anomalies. Our approach is based on the observation that distinct places have different hourly/daily patterns of visitation, and subsequent locations in a trajectory are temporally and semantically correlated. By annotating the user’s trajectories with the location information we can then characterize the users based on the spatio-temporal properties of their trajectories. Our findings are built upon the analysis of several datasets from Foursquare check-in data, Google Places API, and Twitter Stream API.

The remainder of the dissertation is organized as follows: Chapter 2 introduces the theoretical background on human mobility and LBSNs analysis; Chapter 3 introduces the techniques to characterize the locations visited by the users; Chapter 4 introduces the techniques to characterize the users based on their movements; Chapter 5 discusses and concludes the dissertation.
Chapter 2

Literature Review

Initial contributions to theories of human movement remained almost exclusively in the Demography and Social Psychology fields [68, 95] and were strongly influenced by Ravenstein’s seminal work “Laws of Migration” [75]. It was only half a century after Ravenstein’s work that the American Sociologist Stouffer proposed the intervening opportunities theory, a mathematical framework to attempt to determine the relationship between distance and migration [16, 96]. With the advent of mobile phone technology in the last decades and its massive diffusion, researchers were able to collect large amounts of data at the individual level. Since then, the field has seen an exponential growth producing an extensive literature on the subject. Such studies come mainly from Physics and Computer Science research, and most of them describe the presence of power-law relationships in the dynamics of both human mobility and social networks [17, 34, 35, 92, 88].

In this chapter we provide a short survey on the human mobility modeling literature and the main quantities and relationships of interest. The interested reader is invited to read our extensive survey in [10]. The mobility models presented
in this chapter have been selected because of their relevance to the field and their accuracy in reproducing large-scale patterns of human trajectories. We also cover several studies which analyze human trajectories using data generated by online social services with the aim of extracting information that is beyond the user’s movements.

2.1 Population Models

The studies carried out by researchers in humanities, such as Social Scientists and Social Psychologists, focused on data at the population level. They were more concerned about human migrations and movement of people from one city to another and also lacked spatio-temporal data at the individual level.

The first look at human trajectories was done by Ravenstein in 1885, who proposed the *Laws of Migration* with the intent of explaining and predicting migration patterns both within and between nations [75]. He originally proposed seven laws and then added another two in 1889 [76]. The laws are quite intuitive and capture the fact that most people travel only short distances and in the direction of commerce or industry centers given that they offer the most work opportunities. These centers are rapidly-growing towns, so people move from more rural areas to the urban centers; the rural areas are then filled by migrants from other regions thus propagating the migratory flow. Obviously, people living in towns are less likely to migrate. It was also noted that females are more migratory than males, and the infrastructure plays an important role in the migratory flows. While Ravenstein’s findings were qualitative in nature, he correctly identified several socio-economic aspects of migrations and stimulated an enormous volume of work.
2.1.1 The Origin-Destination Matrix

The Origin-Destination (OD) matrix is the standard tool used to model aggregated mobility flows and it provides a way to represent the number of individuals traveling between the locations in the area under study over a period of time. Due to its characteristics, it is of importance for population level human mobility models such as the intervening opportunities and the gravity models.

A simple way to define an OD matrix $T$ is to divide the area of interest in $n \times m$ regions, where $n$ represents the “origins” and $m$ the “destinations” and usually $n = m$. If we define $i = 1, \ldots, n$ and $j = 1, \ldots, m$ then $T_{ij}$ are the elements of the OD matrix representing the number of people moving from $i$ to $j$. Such matrix defines a directed and weighted network, and in the general case is time-dependent

$$T(t) = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1m} \\ T_{21} & T_{22} & \cdots & T_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ T_{n1} & T_{n2} & \cdots & T_{nm} \end{bmatrix}. \quad (2.1)$$

The size of the OD matrix depends on the size and spatial scale/resolution of the data on the area of interest. Usually, the regions are administrative units, like census and electoral units, up to entire municipalities or states.

The OD matrix can be inferred empirically, but it is a hard task that has been subject of numerous studies [2, 19]. In fact, while studying aggregated flows there are several factors that could play an important role and that should be accounted for:

- Transportation modes: distinct modes of transportation offer different speed and cost [33, 46].
• Scale and discretization: it is important to use a discretization which is consistent with the mobility processes. It can be seen on the trip length distribution, which seems to display peaked laws at the urban scale and a broad distribution at national scales [34, 36, 66].

• Nature of the trip and heterogeneity of users. These factors are well known in transportation research [44, 107], and are necessary to distinguish different trips (commuting, travel, etc.) [114]. We also need to distinguish between different populations with different behaviors [48, 69].

2.1.2 The Role of Distance

In 1940, the American Sociologist Stouffer [96] made one of the first attempts to provide a conceptual and formal model of human mobility. In Stouffer’s model migrations are proportional to the number of opportunities and inversely proportional to the cumulative number of intervening opportunities between two locations. Therefore, human displacements are driven by the spatial distribution of opportunities rather than by distance.

Although several variants of the model exist, it is usually expressed in the Origin-Destination matrix form as given by Schneider [84]:

\[
T_{ij} = T_i e^{-LV_{ij-1}} - e^{-LV_{ij}} \frac{1 - e^{-LV_{in}}}{1 - e^{-LV_{in}}},
\]

where \( T_{ij} \) is the flow from the origin location \( i \) to the destination \( j \). \( T_i \) is the total number of trips originating from \( i \) and the fraction represents the probability of one of those trips to end in \( j \); \( 1 - e^{-LV_{in}} \) is a normalizing factor such that \( \sum_j T_{ij} = T_i \). \( V_{ij} \) is the cumulative number of opportunities up to the \( j \)-th location by travel cost.
$n$ is the total number of locations in the region considered, and $L$ is a constant adjusted to fit the data representing the probability of accepting an opportunity destination.

The usefulness of Stouffer’s theory has been corroborated in later studies. Bright and Thomas [16] tested the model using a larger dataset of interstate migration in the United States. Their findings suggested that the patterns of interstate migration in the US were closely in agreement with the intervening opportunities model. Furthermore, Noulas et al. [66] carried out an analysis on a global-scale dataset with more than 900,000 users of Foursquare, a location-based social network service. Their results suggested that the physical distance is not sufficient to fully determine universal rules of human mobility.

In 1946, George K. Zipf proposed the gravity model, a mobility model inspired by the Newton’s law of gravity [122]. In his work, Zipf reaffirms the importance of the distance for human migration patterns by stating that the magnitude of a migratory flow between two communities is proportional to the populations and inversely proportional to the distance. A generalization of the gravity model to account for more complex influence of the distance and weighting the population importance [6] was later defined in [107] as

$$T_{ij} = C \frac{N_i^\alpha N_j^\gamma}{f(r_{ij})},$$

(2.3)

where $T_{ij}$ is the commuting flow between two regions $i$ and $j$, whose populations are respectively $N_i$ and $N_j$. $C$ is a proportionality constant. $\alpha$ and $\gamma$ are parametric exponents, whereas $f(r_{ij})$ is a distance-dependent deterrence function. Parameters $\alpha$, $\gamma$, and the function $f(r_{ij})$ are chosen to best fit empirical data.
Even though a considerable amount of literature supports the goodness of fit provided by this model under certain conditions and scenarios [7, 43, 107], in recent years gravity models have been steadily challenged [55, 88]. For instance, Simini et al. [88] summarized a considerable number of limitations present on gravity models, and one of the most emblematic is an analytic inconsistency in the model. In fact, in Equation (2.3) when the destination population \( N_j \) grows, the number of migrants \( T_{ij} \) also grows, potentially exceeding the origin's population \( N_i \). Moreover, gravity models may require the estimation of a considerable number of parameters to fit the empirical data [41, 88].

The radiation model was introduced by Simini et al. [88] in an attempt to provide a unified mobility framework capable of reproducing regularities ranging from local commuting patterns to long-term migration. In the radiation model, the average commuting flow \( T_{ij} \) between two locations \( i \) and \( j \) is defined as

\[
T_{ij} = T_i \frac{N_i N_j}{(N_i + s_{ij})(N_i + N_j + s_{ij})},
\]

(2.4)

where \( N_i \) and \( N_j \) are the populations of locations \( i \) and \( j \) respectively, and \( T_i = \sum_{i \neq j} T_{ij} \) denotes the total number of migrants originating from \( i \). \( s_{ij} \) is defined as the total population within a circle centered at \( i \) whose radius \( r_{ij} \) is the Euclidean distance from \( i \) to \( j \) minus the populations of the two locations \( i \) and \( j \). It is worth noting that Equation (2.4) does not have any adjustable context-specific parameters such in gravity model (see Equation (2.3)). The authors argue that due to its parameter-free nature, the model does not suffer any of the limitations of the gravity model, such as its strong dependence on previous traffic data.

In a later work, Simini et al. [89] introduced a variation of the model called
radiation model with selection where authors sacrificed the parameter-free nature of the original radiation model in favor of realism and representation power. This model can account for socio-economic factors under the assumption that “each individual travels to the nearest location where she/he can improve her/his current working conditions (benefits)”.

Furthermore, Simini et al. [89] proved the gravity model and the intervening opportunities model are a particular case of the radiation model with selection.

### 2.2 Individual Mobility

The ubiquity of mobile phones lead to the availability of large datasets capturing spatio-temporal traces of human trajectories. The mathematical framework of Random Walks (RW) has been extremely successful at modeling and reproducing some of the statistical properties of movements observed in these datasets. The usefulness of such models is that they show that a simple process—such as a symmetric random walk—can be a good approximation for trajectories of complex processes such as prices in the stock market [29] or movements of animals [13].

The scientific literature has a wealth of applications of such models across a wide range of fields such as biology [13, 25], chemistry [5], economics [29], and physics [37].

The mechanism at the core of random walks consists of a series of successive alternating displacements (i.e. movements) and changes of direction [81]. Random-walk models can vary in many aspects such as the number of dimensions, the step-length distribution, and the wait time between jumps. One simple example is a symmetric random walk with constant jump length; such model can be used to
approximate the Brownian motion [81]. While simple random walks are not an appropriate model of human mobility, they engender a number of more complex mobility models such as the Continuous-time Random Walk and the Lévy Flight. Of the many random walk models, Lévy flights and Continuos Random Walks (CTRW) were found to be of high relevancy for human mobility modeling.

Lévy flights are stochastic processes whose step length follows a heavy-tailed distribution (e.g. power law) [44], that is a Lévy flight consists of many short jumps (i.e. daily routine) and occasional long jumps (i.e. vacation). One limitation of Lévy flight is that the time intervals between two successive steps are constant. Montroll et al. [57, 58] proposed the Continuous-Time Random Walk (CTRW) model consisting of a succession of random displacements and waiting times [44]. Both the displacements and waiting times are drawn from two probability density functions which in the context of individual mobility are fat tailed like in Lévy flights. We leave out further mathematical details of the model, the interested reader should refer to [11, 44, 57].

2.2.1 Trip Distance and Waiting Time Distribution

A seminal work in the study of individual trajectories has been done by Brockmann et al. [17] who carried out an extensive analysis on bank notes circulation in the United States as a proxy for human movements. They computed the displacement $\Delta r = |x_2 - x_1|$ between each pair of locations whenever a bill was reported on a website\(^1\) and the elapsed time between two consecutive reports. They found that human trajectories were well approximated by the CTRW model, because the

\(^1\)https://www.wheresgeorge.com/
step-lengths and waiting time distributions have been found to follow a power-law distribution:

\[ P(\Delta r) \sim \Delta r^{-(1+\alpha)}, \quad (2.5) \]
\[ P(\Delta t) \sim \Delta t^{-(1+\beta)}. \quad (2.6) \]

One of the main criticisms to the work of Brockmann et al. is that it does not describe human movement directly, but an aggregation of several trajectories segments. In fact, a bank note trajectory likely combines trajectories of several individuals and hence it is not representative of a single individual mobility.

With the availability of large phone Call Detail Records (CDRs), Gonzalez et al. [34] found that human trajectories are better approximated by a truncated Lévy flight, i.e. a Lévy flight whose displacements distribution follows a truncated power law:

\[ P(\Delta r) \simeq (\Delta r + \Delta r_0)^{-1-\beta} \exp \left( \frac{-\Delta r}{k} \right), \quad (2.7) \]

where \( \Delta r_0 \) is a lower limit to the space resolution of the datasets they used, and \( k \) is a constant in kilometers. The truncated power-law distribution behaves exactly the same as a regular power-law up to an upper limit \( k \), then the probability of finding values greater than \( k \) decays exponentially. The exponential cutoff were explained as a result of population heterogeneity (see Section 2.2.2). Similar results were found by Song et al. [92] who confirmed that CTRW are a good approximation of human mobility, but they noticed that the diffusion speed and the location visitation patterns were not consistent with the CTRW model (see Section 2.2.3).
2.2.2 Radius of Gyration

Gonzalez et al. [34] hypothesized that the distribution in Equation (2.7) was a convolution of Lévy walk trajectories and a population-level heterogeneity. Such hypothesis was confirmed by studying the distribution of the radius of gyration. In human mobility, the radius of gyration refers to the distribution of the user’s visited locations around the center of mass of the user’s trajectory and can be interpreted as the characteristic distance traveled by a user:

\[ r_g = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (r_i - r_{cm})^2}, \]  

(2.8)

where \( r_i \) represents the \( i = 1, \ldots, N \) positions recorded for the user and \( r_{cm} = 1/N \sum_{i=1}^{N} r_i \) is the center of mass of the user’s trajectory.

Gonzalez et al. found that the radius of gyration distribution follows a truncated power law:

\[ P(r_g) = (r_g + r_g^0)^{-\beta_r} \exp \left( \frac{-r_g}{k_r} \right). \]  

(2.9)

The authors explained the mechanism at the base of \( r_g \) stabilization by suggesting that the visitation frequency of a location \( L \) is inversely proportional to its rank, following a Zipf’s law \( P(L) \sim 1/L \) independently of the number of locations visited by the user—that is, people tend to spend most of their movements towards few locations. Therefore, the logarithmic saturation of \( r_g \) finds its root in the high degree of regularity of daily travel patterns (see Section 2.2.3 and Section 2.2.5).

Pappalardo et al. [69] further observed that the high heterogeneity of users was seemingly in discrepancy with the high regularity and predictability of their trajectories. To explain such results, the authors introduced the \( k \)-radius of gyration.
as the radius of gyration computed over the $k$-th most frequented locations:

$$r_g^{(k)} = \sqrt{\frac{1}{N_k} \sum_{i=1}^{N_k} \left( r_i - \overline{r}^{(k)} \right)^2}.$$  

(2.10)

The authors studied $r_g^{(k)}$ for $k = 2, 4, 8$ and found two classes of users they defined as $k$-returners and $k$-explorers. $k$-returners are those individuals whose total $r_g$ is similar to $r_g^{(k)}$, as their characteristic traveled distance is mostly explained by their $k$-th most frequented locations. $k$-explorers are those individuals whose $r_g^{(k)}$ is only a small fraction of the total $r_g$ indicating that their characteristic distance is explained by more locations than just the $k$-th most popular ones.

### 2.2.3 Preferential Return

One criticism of random walk models is that human trajectories are not really random [77]. Song et al. [92] observed the following shortcomings:

I. The number of distinct locations $S(t)$ visited by a user after a time $t$ follows

$$S(t) \sim t^\mu$$  

(2.11)

with $\mu$ smaller than what predicted by Lévy Flights and CTRW. This implies that people visit new locations much more slowly than what is predicted by random walk models.

II. The visitation frequency $f_k$ of the $k$-th most visited locations follows a Zipf’s law as found in [34]:

$$f_k \sim k^{-\zeta},$$  

(2.12)
i.e. the visitation frequency of a location is inversely proportional to its rank. However, both Lévy Flight and CTRW models predict that the visitation frequency of a location is asymptotically uniform ($f_k \sim \text{const.}$).

III The Mean Square Displacement (MSD) follow a slower than logarithmic growth:

$$\langle \Delta x^2(t) \rangle^{\alpha/2} \sim \log \left( \frac{1 - S^{1-\zeta}}{\zeta - 1} \right) + c,$$

while diffusive processes as CTRW asymptotically follows a relation of the kind $\langle \Delta x^2(t) \rangle \sim t^{\nu}$. Such ultra-slow diffusion is linked to the tendency of humans to visit few locations on a daily basis (e.g. home and workplace), and therefore their trajectories have a high degree of recurrence and temporal periodicity.

Song et al. [92] proposed and incorporated in CTRW two key mechanisms, unique to human trajectories, that were missing from existing models: Exploration and Preferential Return. An individual has two choices for each next step:

**Exploration** As indicated by the scaling law in Equation (2.11), the tendency to visit unexplored locations decreases with time. In fact, the longer we observe a person’s trajectory, the more likely that most of nearby locations have been visited. Then, the individual moves to a new location situated at a distance $\Delta r$ (drawn from Equation (2.6)) with probability

$$P_{\text{new}} = \rho S^{-\gamma}$$

where $S$ is the number of previously visited locations.

**Preferential Return** Humans have a high propensity to return to previously-
visited locations, which is somewhat related to the cumulative advantage [74] from the social sciences, or the preferential attachment [8] of network science. Then, an individual returns to one of the previously visited locations with complementary probability

\[ P_{ret} = 1 - \rho S^{-\gamma}. \quad (2.15) \]

The location is selected with a probability that is proportional to the number of times the user has previously visited that location, \textit{i.e.} according to Equation (2.12).

The two parameters \( \rho \in (0, 1] \) and \( \gamma > 0 \) control the balance between the exploration phase and preferential return and are estimated by fitting the data.

### 2.2.4 Frequency vs Recency of Visited Locations

The preferential return assumptions (see Section 2.2.3) lead to discrepancies with what is observed in the real world. According to the cumulative advantage, the earlier a location is discovered, the more visits it is going to receive, which implies that early visited locations will also be the most visited ones. However, if that would be the case, people would never change their preferences, which is clearly not true.

Barbosa \textit{et al.} [9] introduced a model in which the return phase considers both frequency and \textit{recency} (\textit{i.e.} how long ago a location was visited) of locations. More specifically, they define a recency-based rank \( K_s \) and a frequency-based rank \( K_f \). Then, the return phase happens with probability \( P_{ret} \) that follows Equation (2.15), but the return jump selects the next location from the most frequently visited ones.
with probability $\alpha$ and with the complimentary probability $1 - \alpha$ from the most recently visited ones. That is:

$$P_{ret}^s = (1 - \alpha)P_{ret} \propto K_s^{-\nu},$$  \hspace{1cm} (2.16)$$

$$P_{ret}^f = \alpha P_{ret} \propto K_f^{-1-\gamma}.$$  

Note that when $\alpha = 1$ the model is identical to the preferential return model.

### 2.2.5 Trajectory Entropy and Predictability

In Section 2.2.2, Section 2.2.3, and Section 2.2.4 we covered how human mobility differs from random walks and instead exhibits a high degree of regularity and hence predictability. Such properties can be captured and quantified by the entropy, a measure from information theory [28].

Song et al. [93] computed three types of entropies. The random entropy $S^{\text{rnd}}$ when every location as the same probability to be visited, the temporal-uncorrelated entropy $S^{\text{unc}}$, and an estimate of the true entropy $S^{\text{est}}$ based on Lempel-Ziv compression [47] of user trajectories from CDRs:

$$S^{\text{rnd}} = \log_2 N,$$  \hspace{1cm} (2.17)$$

$$S^{\text{unc}} = -\sum_{i=1}^{N} p(i) \log_2 p(i),$$  \hspace{1cm} (2.18)$$

$$S^{\text{est}} = \left( \frac{1}{N} \sum_{i} \Lambda_i \right)^{-1} \ln N,$$  \hspace{1cm} (2.19)$$

where $N$ is the number of distinct locations visited by a user, $p(i)$ is the probability of visiting location $i$, and $\Lambda_i$ is the length of the shortest substring starting at position $i$ which does not appear in positions 1 to $i - 1$. Note that $S^{\text{est}}$ converges
to the true entropy $S$ when $N$ goes to infinity and $0 \leq S^{\text{est}} \leq S^{\text{unc}} \leq S^{\text{rnd}} < \infty$.

Song et al. studied the entropy distribution and found that it is peaked at 0.8 for the $S^{\text{est}}$, therefore the actual uncertainty of user visited locations is less than 2 locations. Song et al. also came up with a way to estimate the maximum predictability $\Pi^{\text{max}}$ of a user trajectory from the estimated entropy as

$$S = -[\Pi^{\text{max}} \log_2 \Pi^{\text{max}} + (1 - \Pi^{\text{max}}) \log_2 (1 - \Pi^{\text{max}})] + (1 - \Pi^{\text{max}}) \log_2 (N - 1)$$

and found out that $\Pi^{\text{max}}$ is peaked at 0.93, which it means that the next location a user will visit can be guessed correctly 93% of the time.

### 2.2.6 Social Aspect of Mobility

Human trajectories can be studied over three main properties: spatial, temporal, and social [42]. The previous sections covered only spatial and temporal aspects. However, humans are social entities and it seems only logical to take in account social aspects that influence human mobility. For instance, we might visit a place just because our friends are there. Also, it is intuitive to think that similar users should exhibit similar mobility patterns.

Toole et al. [100] measured the *mobility similarity* of user’s trajectories using the cosine similarity

$$\cos \theta = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|},$$

where $\mathbf{v}_i, \mathbf{v}_j$ are vectors representing the frequency of visited locations. They computed the similarities over the course of a typical weekday and weekend and then clustered them. They discovered three types of social relationships: acquaintances
have uniformly low levels of similarity, co-workers have high similarity on weekdays working hours and low similarity on nights and weekends, and family/friends have high similarity on nights and weekends.

With these results in mind, Tool et al. introduced the GeoSim model which updates the preferential return by choosing the next location to be visited with respect to where similar users are:

\[ P_{ret}^f = (1 - \alpha)P_{ret} \quad \Pi \propto f_k, \quad (2.22) \]
\[ P_{ret}^s = \alpha P_{ret} \quad P(v) \propto \cos(\theta_{u,v}), \quad (2.23) \]

where \( P_{ret} \) has the same meaning of Equation (2.15), \( P_{ret}^f \) is the probability of a frequency based return, while \( P_{ret}^s \) is the probability of a social based return; \( P(v) \) is the probability of selecting the user’s \( v \) location for the return jump. The model also applies the social aspect for the exploration jumps.

\section*{2.3 Annotation of Trajectories}

The research on human mobility is not limited to finding the fundamental laws of movement. With most of the scaling properties of human mobility already identified, researchers started focusing on how to extract additional information from user’s trajectories. The increasing use of smart phones and social media have created a large pool of data that scientists can study. The data that come from Location Based Social Networks (LBSNs) have the advantage, when compared to CDRs, to have contextual information associated with geographical position, and therefore allows the extraction of more information related to the users such as the
identification of activities, trip purpose and user’s social connections.

The identification of relevant places for the users (Points of Interest) and associated activities is of high relevance. Knowing the points of interest (POIs) forms the basis to better understand the mobility patterns in an area. Also, being able to extract and analyze mobility patterns enable location-based recommendation systems, which are services targeting people visiting specific areas [111, 120].

There is quite an extensive literature on the topic. Mamei et al. [54] provided an algorithm to automatically identify relevant points of interest to the users by clustering CDRs using a distance metric that considers both space and time. Zeng et al. [119] introduced a clustering algorithm based on stay points (point where a user spent some time and that are at a minimum distance apart) from user’s trajectories to build a collaborative location and activity recommendation system. Shaw et al. [86] used millions of Foursquare check-ins and relative metadata to develop a search algorithm to associate a noisy GPS user location to a point of interest. Lian et al. [49] extracted location names from the current location, time and check-in histories of the users. Liu et al. [50] used the tweet content and the user’s history information to infer the location type. Ye et al. [114] used a combination of explicit patterns of individual places, such as the number of visitors and distribution of check-ins over time, and the implicit relationships among similar places to annotate places (i.e. add descriptive tags to them).

The knowledge of POIs and their characterization in term of location type and activity can also be used to further the understanding of mobility. Xiao et al. [109] used trajectories with a semantic location history (i.e. adding the type of the locations) to aggregate users based on their trajectory similarity with the assumption that similar users have similar interests and social ties. Gabrielli et
al. [32] used the information related to the activity performed in each place as the purpose for the movement to augment the ability to detect urban mobility patterns and anomalies. Wu et al. [108] worked on modeling the intra-urban mobility of people by using an activity-based framework. Silva et al. [87] used Twitter data linked to Foursquare and user movement to find connections between different types of locations. Wakamiya et al. [103] were able to classify urban characteristics using geo-tagged tweets and their timestamps. Yuan et al. [116] used human mobility and points of interest to classify different sections of Beijing.

In all these works, some machine learning techniques (such as clustering and classification) are used to identify and characterize the locations of interest, the activities, and the users. However, they suffer of three major limitations:

i Studies that use CDRs data have intrinsically limited spatial resolution.

ii Studies that extract POIs directly from check-ins data have limited or no applicability to data where check-ins availability is sparse.

iii Studies that combine data from multiple sources lose generality as those sources might not be available everywhere and/or with the same accuracy.

Therefore, there is the need to address the aforementioned issues. In particular, (i) can be overcome by using mobile phone GPS trajectories which offer high spatial resolution data. In this work we address issues (ii) and (iii) by introducing a framework that relies only on GPS points and associated timestamps of users’ trajectories to extract and characterize visited locations and types of users.
Chapter 3

Location Characterization

The analysis of user’s trajectories is dependent on the concept of location as we need to know from where each movement originated and ended to be able to properly reconstruct the user’s trajectory. Additionally, the locations have a semantic meaning linked to the cause of the movement; being able to characterize user’s locations can further the understanding of human mobility and it also finds many applications in location aware services and recommendation systems.

Mobility datasets usually provide an identifier of user’s position every time a user’s movement is recorded and a time recording when that movement happened. In CDRs the location identifier is usually provided in the form of the cell phone tower which does not pinpoint user’s location accurately enough for a fine grained spatial analysis (but it is still useful to extract information on the user’s movement patterns). In LBSNs and online social networks user’s position is usually provided as a user’s check-in or a geolocated record which provide the GPS coordinates of the device from which it was generated (usually, the user’s mobile phone).

The proposed framework is designed to work with data that provide a reason-
ably accurate temporal and spatial location of the user (i.e. GPS coordinates and a timestamp with at least hour accuracy). The choice is justified by recent evidence that mobility models and other human mobility research results are strongly influenced by the spatial resolution of the dataset under study [1]. Therefore, we focused our analysis on the type of data that is generated by LBSNs and several other online social networks such as Twitter and Facebook. Furthermore, we restricted ourselves to work with a minimum amount of information and data that are as generic as possible to extend the applicability of the framework to many data sources. For example, we do not require the data of the user check-ins such as the studies in [49, 111, 114, 116]. Finally, we should note that while it is expected data from online social networks are biased towards social activities and younger population [90], it does not impact the validity of our work because we did not make any assumption related to a specific activity or subset of the population. Ideally, the framework should not be limited to any specific region or activity, but rather relies on the fundamental patterns of human mobility.

3.1 Datasets

3.1.1 Twitter Dataset

The datasets we used in this chapter contains one month of geo-tagged tweets (from August 28th to September 29th, 2014) that were collected using the Twitter stream API from the New York City Manhattan area. Only tweets from inside the Manhattan area were considered because of the high amount of activity which makes the analysis more reliable. We collected 1,217,229 tweets from 109,305 users. We applied a filtering step to remove non-human Twitter’s users. Users
with a travel speed between two consecutive locations faster than 80m/s were removed. 80m/s has been considered a relative safe threshold for users’ movement max speed. We also removed users with a suspiciously high activity. All users with more than 500 tweets in a month were removed. Although we understand that it may be possible for a user to tweet frequently, we did not want to run the risk of including “robots” in our dataset. The final dataset, contains 1,022,286 tweets from 108,341 users. We call this dataset $D_1$.

3.1.2 Foursquare Dataset

We identify this dataset as $D_2$. In order to provide the labels used to train the classifier we used the New York City Foursquare dataset provided by Yang et al. [112]. $D_2$ is a collection of 227,428 Foursquare check-ins from 1,083 users lasting for about 10 months (from 12 April 2012 to 16 February 2013) and covering 38,333 locations (of which only 5,420 located in Manhattan area). Venues in Foursquare are classified in root categories and sub-categories\(^1\); 400 sub-categories are represented in this dataset. The dataset provided in Yang et al. [112] has been filtered by the authors by selecting only the users who have performed at least one check-in per week (these users are regarded as active users), and removing fake check-ins originated by users who performed unusually sudden-moves (consecutive check-ins with a speed faster than 1200 km/h).

\(^1\)http://developer.foursquare.com/docs/venues/categories
Table 3.1: Overview of the Mobility Datasets

<table>
<thead>
<tr>
<th></th>
<th>$D_1$</th>
<th>$D_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users</td>
<td>108,341</td>
<td>1,083</td>
</tr>
<tr>
<td>Total number of records</td>
<td>1,022,286</td>
<td>227,428</td>
</tr>
<tr>
<td>Unique locations</td>
<td>N/A</td>
<td>38,333</td>
</tr>
</tbody>
</table>

3.2 Location Classification

3.2.1 Identification of Points of Interest (POI)

A *Point of Interest (POI)* is a specific location that someone may find useful or interesting; in our case, the POIs are the locations a user visits while moving around and are of importance because they represent the meaning behind the movement itself. The user’s trajectories are provided by the dataset $D_1$ as a set of GPS points associated with the user’s tweets. Since we do not rely on the user’s check-ins, we do not know from which location a specific point in the user’s trajectory was originated. Hence, we need to infer the POIs from the GPS traces left by people movements.

The GPS coordinates lie on a non-Euclidean space, which is a problem for some clustering algorithms, therefore we transformed the geographical coordinates from spherical to planar by using the Mercator projection [91]. The Mercator projection is known to deform the area proportionally to higher latitudes, yet it conserves relative distances as long as these distances are suitably small (less than one degree latitude/longitude). For the Manhattan bounding box, the width ($\epsilon_w$), height ($\epsilon_h$), and diagonal distance ($\epsilon_d$) errors were found to be $\epsilon_w \simeq 6.09\%$, $\epsilon_h \simeq 6.23\%$, $\epsilon_d \simeq 6.09\%$ respectively. This distance errors were computed by comparing the haversine distance [78] between the real-world coordinates to the euclidean
distance between the projected points.

The projected points were assigned to the locations (POIs) using clustering techniques; each inferred location position is the average geographical coordinates of the cluster of points representing that location. We compared several clustering algorithms: $k$-means, mean shift [26], and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [20]. We tested $k$-means with $k \in [5,000,50,000]$ with a step increment of 5,000. Mean shift with a bandwidth value $b$ from $\approx 1m$ to $\approx 110m$ and noise classification both enabled and disabled. HDBSCAN was tested with the minimum cluster size $m$ from 2 to 45.

Several methods were used to evaluate the quality of the clustering algorithms as a function of the specific parameters of the clustering algorithms. For instance, the Bayesian Information Criterion (BIC) [30], the silhouette score [80], the elbow rule [45], the number of clusters found, the number of unclustered points, as well as visual inspection (Figure 3.1).

The optimal parameters according to the aforementioned metrics were found to be $k = 25,000$ for $k$-means, $b \in [22.2, 44.5]$ for the mean shift, and $m \in [7, 10]$ for HDBSCAN (for more information, see Appendix A.1). Comparing the output of several clustering algorithms is required to obtain a robust clustering of the GPS points. In fact, a clustering algorithm will always return a set of clusters, but in order to assess its validity we can compare its consistency with the result of other algorithms: if multiple algorithms agree, we can infer that the clustering we obtained is fairly correct. For instance, a comparison of the results showed that the optimal number of clusters from the $k$-means analysis is in agreement with mean shift and HDBSCAN results. Specifically, for mean shift a bandwidth $b = 44.5$ resulted in 28,663 clusters, while HDBSCAN with a minimum cluster sizes of $m = 7$. 

28
Figure 3.1: Small sample of GPS points from user’s trajectories which has been clustered using HDBSCAN. The clustering process shows a high sensitivity and spatial resolution as the clusters closely follow the distribution of buildings and businesses.

yielded to 26,093 clusters. Furthermore, we should note that synthetic metrics do not have any real-world meaning, which it means that choosing a clustering parameter solely to optimize a metric might lead to a poor clustering against ground truth data. Therefore, we also used a set of 7,586 Foursquare check-ins extracted from the text of the tweets. Then, we selected those POIs with at least 3 Foursquare check-ins and computed how many check-ins were associated to unique locations for each cluster (Figure 3.2). We decided to use the clustering performed by HDBSCAN because it greatly outperforms the other clustering algorithms and due to the advantages of this algorithm over the others (it allows for non-round clusters, clusters with different density, and noise classification).
Figure 3.2: The clustering algorithms might perform poorly when comparing the results to the ground truth data even when the clustering parameters are chosen optimally. Only HDBSCAN appropriately separate distinct locations in different clusters, while k-means and mean shift aggregate 2 or more distinct locations in the same cluster 30% of the time.

3.2.2 Feature Extraction for Location Classification

The HDBSCAN clustering identified 26,093 locations/POIs. In order to classify these locations into their business categories, we extracted feature vectors associated with the temporal aspect of visitation by the users. The features we used are the combination of the popular-times histogram (Figure 3.3) and popular-days histogram (Figure 3.4). Since the clustering of GPS points was performed us-
ing HDBSCAN with a minimum cluster size of seven, some of the popular-times
histograms can potentially have only seven points of data. Two solutions were
implemented to combat the feature sparsity. First, we applied a one dimensional
Gaussian filter [18] to smooth and fill areas of the histogram for which there is po-
tentially missing data (Figure 3.3). Second, we improved the discriminative power
of the feature vectors by adding the popular-days histogram which aggregate the
points over only 7 dimensions (the days of the week). The extracted 31-dimensional
feature vectors are then normalized before being passed as input for the classifier.

Figure 3.3: Example of histogram of popular times of the day. We applied a one
dimension Gaussian filter to smooth the histogram as an attempt to combat the
sparsity of the data for some of the locations.

The expressiveness and power of such feature vectors is showed in Figure 3.5.
Very characteristic temporal patterns emerge and can be associated with the type
of the locations. For instance, we can observe a work-days/weekend behavior,
where office locations are mostly active during the working days while food is
prominent in the weekends. Also, each category of locations has specific hours at
which the user’s activity is at its top, revealing how the time has a strong influence
on the type of the locations visited by the people.
Figure 3.4: Example of popular days of the week histogram for a place that is frequented especially over the weekend. The popularity of a location during the days of the week is combined with the popularity of the location during the time of the day. The resulting feature vectors have 31 (24+7) components.

Figure 3.5: The normalized feature vectors, averaged by location category, show that each category has a specific hourly and daily pattern. For instance, office is very popular during common office hours and work-days (gray background), but not at night and in the weekends.
3.2.3 POI Classification

In order to perform classification, which is a supervised task, we need to assign a label representing the location category to each POI feature vector. We labeled the feature vectors using Foursquare data provided by the dataset $D_2$. For each POI generated by the clustering, we assigned as the label for the feature vector associated with the POI the location category of the closest Foursquare check-in within 20 meters radius in the dataset $D_2$. The process generated 3,959 labeled locations from 174 categories. We used twenty meters as the given radius because even though many locations remained unlabeled, expanding the radius would lead to incorrect labeling and the introduction of noise into the data used to train the classifier.

The small amount of labeled feature vectors obtained using the Foursquare dataset prompted us to test the robustness of our approach by using a second dataset of labels gathered querying Google Places API\(^2\). We queried Google Places API for each of the POIs we identified (Figure 3.6) to find the nearest locations within 20 meters radius which lead to 13,595 labeled feature vectors. Each query returns a maximum of ten locations and each location has a list of categories from the most specific to the least specific (\textit{e.g.} bakery, store, food or furniture store, clothing store, home goods store, store). This means that depending on whether the most or least specific label is used, some categories are sub or super sets of the others (\textit{e.g.} bakery and clothing store are both a subset of store). To reduce the number of categories and automatically group similar categories, we decided to select the most general labels returned by the Google Places API.

\(^2\)https://developers.google.com/places
Figure 3.6: Google Places API can return up to 10 locations in a radius of 20 meters in Manhattan (NYC) area due to the high density of POIs. Furthermore, each POI is assigned a list of labels representing the categories of the POI from the most specific to the least specific.

Given the large amount of categories, the relatively small amount of feature vectors to train the classifier, and the lack of ground truth data that could lead to mislabeling (introducing noise in the labels), we used several filtering steps to improve the labeling accuracy and help the training of the classifier (see Figure 3.7). The idea was to improve accuracy by training and testing only robust locations, i.e. those that have enough data, produced by several different users, but not too much data as there is the possibility that multiple locations were clustered into one. In fact, the classifier cannot learn properly if the labels are noisy (not correctly assigned) or the feature vectors have sparse features.

We combined similar categories (e.g. Food and Drink Shop, Mexican Restaurant, American Restaurant, Bar) into broader categories and removed underrepresented categories (< 2% of total dataset, so that each category has several feature
Figure 3.7: In order to effectively train the classifier the data need to be filtered (gray boxes) to obtain the best feature vectors. This multi-step filtering process is required because of the noisy training data. Labels are filtered by aggregating similar categories and removing those with not enough feature vectors. Locations are filtered by removing possible private residences (imposing a minimum number of users), those that do not have enough data points (imposing a minimum number of points), or are possibly multiple locations clustered into one (imposing a maximum number of points).

We also filtered the locations based on the number of users per location, maximum and minimum cluster size (in term of number of points in each cluster). We required a minimum number of users to remove private residences as there are no given data on them and therefore should not be included as POIs. We imposed a maximum cluster size to eliminate excessively densely packed areas. Because New York city has many skyscrapers, which have locations on top of each other by nature of having multiple floors, and it is very dense city in general, we may ended up clustering multiple locations into one. In fact, we cannot factor in that locations might be on top of each other using longitude and latitude points. There is some evidence that this might be the case as we found that the location size distribution follows a power-law and there are few locations that have up to 10,000 tweets. Furthermore, locations might overlap if they are within the GPS error...
range (about 10m). By removing POIs with too many associated GPS points, we aimed to filter out clusters that are actually several locations. Finally, we imposed a minimum level of activity by setting a minimum cluster size. Such criteria make sure that the histograms on which the feature vectors are built are less sparse. The ideal parameters for the Foursquare labeling dataset were found to be a minimum of 10 users, a maximum cluster size of 80 points, and a minimum cluster size of 24 points (for more information see Appendix B). The ideal parameters for the Google Place API data were found to be a minimum of 20 users per location and cluster sizes between 36 and 80 (for more information see Appendix B).

We studied the performance of a SVM classifier after applying the filters to the data and applying the labels either from Foursquare (Table 3.2) or Google (Table 3.3). While filtering improved classification performance, it should be noted that each filtering step significantly reduced the number of valid labeled feature vectors to use to train and test the classifier. Furthermore, there are two distinct issues that prevent the classifier performance from being optimal. The Foursquare data we used is sparse, thus leading to too few feature vectors per category to be able to learn effectively. Google data, while less sparse, suffer of noise in the labels. In fact, in many instances, up to 10 locations are returned when searching for locations within a 20 meter radius of the POI. This means that we were unable to select the correct location category to assign to the POI feature vector since we are in the range of GPS accuracy. In this case, it would be proper to develop a more complex heuristic than choosing just the closest place.

The amount of noise in the labels used to train the classifier and the limited amount of feature vectors available for training prevented the SVM classifier from being able to learn effectively. Therefore, to mitigate these issues, we took a new
Table 3.2: Foursquare SVM classifier performance after each filtering step.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Consolidate &amp; 2% Filter</th>
<th>&gt; 10 Users</th>
<th>&lt; 80 Points</th>
<th>&gt; 24 Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>40%</td>
<td>40%</td>
<td>45%</td>
<td>49%</td>
</tr>
<tr>
<td>Precision</td>
<td>49%</td>
<td>52%</td>
<td>53%</td>
<td>54%</td>
</tr>
<tr>
<td>Recall</td>
<td>40%</td>
<td>41%</td>
<td>43%</td>
<td>43%</td>
</tr>
<tr>
<td>Locations</td>
<td>1,619</td>
<td>803</td>
<td>755</td>
<td>326</td>
</tr>
<tr>
<td>Categories</td>
<td>9</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 3.3: Google SVM classifier performance after each filtering step.

<table>
<thead>
<tr>
<th>Metric</th>
<th>2% Filter</th>
<th>&gt; 20 Users</th>
<th>&lt; 80 Points</th>
<th>&gt; 36 Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>15%</td>
<td>20%</td>
<td>23%</td>
<td>26%</td>
</tr>
<tr>
<td>Precision</td>
<td>25%</td>
<td>38%</td>
<td>39%</td>
<td>44%</td>
</tr>
<tr>
<td>Recall</td>
<td>14%</td>
<td>21%</td>
<td>21%</td>
<td>26%</td>
</tr>
<tr>
<td>Locations</td>
<td>13,595</td>
<td>1,138</td>
<td>957</td>
<td>487</td>
</tr>
<tr>
<td>Categories</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

approach based on a semi-supervised learning algorithm. We used a small portion of the collected tweets which were verified to be posted from the Foursquare app (when the user links her Twitter profile in the Foursquare app, a check-in will result in a tweet containing the user’s text and the Foursquare check-in data) to extract the location category. As a result, we built a small amount of ground truth data representing 556 locations. The location categories have been found to have largely varying weights with the “food/bar” and “coffee” labels comprising over 70% of the ground truth labels (Table 3.4), in agreement with the intuition that people mostly check-in in places of leisure. This weights are taken in account by weighting the classes to remove any bias in the training of the classifier. Finally, we used the label propagation algorithm [15] to spread the ground truth labels to the rest of the unlabeled feature vectors (meeting the filtering criteria of a minimum of 10 users and 24 points, and a maximum of 80 points per POI) resulting in a total of 3,304 labeled feature vectors. The SVM classifier performance improved to 65% average accuracy, 67% precision and 66% recall (Figure 3.8).
Table 3.4: Breakdown of ground truth Foursquare labels. The most common categories are associated with common places of the user’s every day routine.

<table>
<thead>
<tr>
<th>Category</th>
<th>Size (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>food/bar</td>
<td>35.4</td>
</tr>
<tr>
<td>coffee</td>
<td>35.1</td>
</tr>
<tr>
<td>transportation</td>
<td>11.3</td>
</tr>
<tr>
<td>park</td>
<td>6.7</td>
</tr>
<tr>
<td>gym</td>
<td>5.8</td>
</tr>
<tr>
<td>museum</td>
<td>2.9</td>
</tr>
<tr>
<td>office</td>
<td>2.9</td>
</tr>
</tbody>
</table>

Figure 3.8: Confusion matrix for a SVM classifier trained on the propagated labels. Some categories are problematic to classify due to either lack of enough training data or because the types of places are often co-located.
3.3 Events and Activities Classification

The information on location position and location category helps defining the users’ movement. However, it does not tell us what the users do in those locations or the actual purpose of the trip. For this reason we introduce the concepts of event, which is something that occurs in a certain place during a particular interval of time, and activity, which is linked to a specific deed, action, function, or sphere of action. While events tell us what the user do in a specific location, activities are associated to the purpose of the trip.

3.3.1 Extraction of Events and Activities

Recall each point in the feature space is a feature vector representing a POI and has several tweets associated to it as part of the same cluster of tweets (the GPS points associated with the tweets were clustered with the process described in Section 3.2.1). We examined the structure of the feature space using spectral clustering with the idea to use spectral clustering as a way to learn the “labels” while solving a manifold learning problem on the feature vector space [102]. By using BIC and the silhouette score we identified the optimal number of clusters $k$ which is $k = 15$ (for more information see Appendix A.2).

Once the feature vectors were grouped, we run a textual analysis of the aggregated tweets in each cluster to extract the topic representing the category of the locations in the cluster. The text content of the tweets has been treated with the usual text processing techniques using NTLK [73]. Punctuation and capitalization were removed from the tweets along with special characters. Furthermore, we removed stop-words and words that lacked thematic significance because they were
very common in the dataset, namely the set \{'new', 'york', 'ny', 'newyork', 'city', 'nyc'\}. We then examined the frequency of specific words within each cluster to get an insight into the clustering.

The textual analysis revealed a surprisingly strong consistency and specificity of the clusters which enabled us to extract the topic representing the locations in the clusters. In fact, few keywords per cluster represented the majority of occurrences of those keywords over the total body of text from the tweets. For example, the “Jeter” cluster had the words ‘jeter’ and ‘re2pect’ representing more than 65% of those word occurrences. Similarly, the “Madison S.G.” cluster had the words ‘madison’ and ‘garden’ at more than 55% representation overall. “9/11” cluster was even stronger with words like ‘memorial’, ‘9/11’, and ‘wtc’ at over 75% of the total representation.

3.3.2 Spatio-temporal Characterization

In the framework we introduced, we assume to work a minimum set of data, therefore the text analysis cannot be used to identify events and activities even though it is useful to get an intuitive understanding. Instead, we need to characterize events and activities based solely on the properties of the POIs associated with them.

The first observation is that, by definition, events are associated with a specific area in space and therefore only few POIs, while activities could be spread over a wide area. Therefore, small clusters of feature vectors (i.e. those with few POIs) correspond to specific events, while larger clusters (i.e. those that contain multiple locations) correspond to more general types of locations usually associated with popular activities. Consistently with word frequency analysis, smaller clusters
The cluster size is the percentage of feature vectors in the feature space associated with a specific cluster. The topics associated with the clusters depend on the size of the cluster (number of locations in the cluster). Events, such as the 9/11 memorial or the Smorgasburg flea market in Brooklyn are associated with only few specific locations.

<table>
<thead>
<tr>
<th>Size (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Times Square</td>
</tr>
<tr>
<td>NYFW I</td>
</tr>
<tr>
<td>NYFW II</td>
</tr>
<tr>
<td>Museum</td>
</tr>
<tr>
<td>Madison Square Garden</td>
</tr>
<tr>
<td>Yankees</td>
</tr>
<tr>
<td>Unknown</td>
</tr>
<tr>
<td>Night Party</td>
</tr>
<tr>
<td>Class</td>
</tr>
<tr>
<td>Jeter</td>
</tr>
<tr>
<td>Climate March</td>
</tr>
<tr>
<td>9/11</td>
</tr>
<tr>
<td>Smorgasburg Flea</td>
</tr>
<tr>
<td>Soulcycle</td>
</tr>
<tr>
<td>Salsa</td>
</tr>
</tbody>
</table>

Clusters have clear keywords relating to the event (Table 3.5): there exists a “9/11” cluster (note that we gathered information in September of 2014), a “Derek Jeter” cluster (Jeter was an acclaimed Yankees baseball player that played his last game during our data collection period), a “People’s Climate March” cluster, and a “Smorgasbord Flea Market” cluster. Furthermore, the locations associated with an event must also be physically close. To capture this aspect, we computed the radius of gyration (see Section 2.2.2), a common metric used in human mobility to measure how locations visited by an user are spread apart, for each cluster of POIs. We found the radius of gyration to be small for events, and, conversely, activities comprise many locations which are distributed around the city, therefore the radius of gyration associated with activity clusters were higher (Figure 3.9).
The average radius of gyration for events (excluding the single outlier) is 1,979 meters, while for activities is sensibly larger at 2,696 meters. Note that, as we increase the area of observation to the whole city, the two distributions, which seems relatively close here, are expected to fall apart.

![Figure 3.9](image.png)

Figure 3.9: The radius of gyration distribution for the events and activities resulting from the clustering of the POI feature vectors. With the exception of a single outlier, events have a sensibly smaller radius of gyration than activities as only few physically close locations are associated with an event.

The second observation is that events happen during a specific interval of time which usually does not last longer than a day, therefore clusters associated to an event see a single large peak of activity in the day of the week corresponding to when the event took place (Figure 3.10). There were two other small clusters, “Soulcycle” and “Salsa”, which besides a specific day have also distinct temporal peaks (6 am and 3 am respectively) due to the events taking place over only few hours. On the other hand, each large cluster shows a distinct hourly signature (Figure 3.11). For instance, the “Night Party” category is, intuitively, far above
average popularity between midnight and 5 am, the “Museum” category is above average in the afternoon, and the “Class” is popular early in the morning. Unlike the smaller categories, all of the locations in these clusters are not actually specific to the most popular words. For example in “NYFW” (New York Fashion Week) clusters the word ‘nyfw’ is very popular simply because NYFW is a popular week-long annual event in New York. However, the size of the cluster can give information as to the typical temporal behavior of the locations within. That is, the labels we placed here are just to get an intuitive understanding, but we could reason on the temporal properties of each cluster without actually knowing what topic is associated with the locations.

In conclusion, we showed how POIs can be identified with a high spatial resolution from the GPS trajectories of multiple users (instead of treating the trajectories individually) by using standard clustering techniques. Once the POIs are identi-
Figure 3.11: The larger clusters in the feature space have distinct hourly patterns and an activity peak associated with the typical popular time of the respective activity.

We identified how the temporal patterns extracted from the user’s trajectories are powerful features that can be used to identify the category of the locations. With a properly trained classifier it would be an easy task to label any dataset from which we can extract the popular-times and popular-days histograms. Finally, we could identify and characterize events and activities based solely on the spatio-temporal properties of the clusters of locations. Events are identified by clusters of few, physically-close locations, which exhibit a temporal peak of activity in a specific day of the week. Activities are identified by larger clusters of spread out locations which exhibit a characteristic hourly pattern, but no specific daily peak. The techniques we introduced can be useful to assign a semantic meaning not only to the POIs, but also to the purpose of the user’s trips.
Chapter 4

User Characterization

The study of the general scaling properties of human mobility is important to construct better models. However, such relationships hardly tell anything about single individuals. Furthermore, there is a growing evidence that the observed scaling relationships are actually the result of the convolution of multiple distributions originated by the heterogeneity of the user’s movement such as modes of transportation [110, 118] and classes of users [69]. At the same time, the ability to discern specific characteristics of the users is very important for applications such as location services and recommendation systems. Therefore, it is instrumental to build tools that are able to provide information to a granularity at the level of the specific individual or classes of individuals.
4.1 Methods

4.1.1 Datasets

We used 3 datasets of geo-located tweets collected over the Manhattan (NYC) area, henceforth called $D_1$ (introduced in Section 3.1.1), $D_2$, and $D_3$ (Table 4.1). $D_1$ was collected between August 28 and September 29, 2014 and contains 108,341 users. $D_2$ is comprised of 152,671 users collected between June 17 and November 4, 2016. $D_3$ is comprised of 147,942 users collected between April 5 and September 11, 2017. Note that for $D_2$ and $D_2$ the POIs were extracted by using the process introduced in Section 3.2.1 and clustering the tweets with HDBSCAN with a minimum cluster size $m = 7$ (for more details see Appendix A.3).

Additionally, when analyzing user’s trajectories we must carefully cure the data (see Figure 4.1). In fact, there are at least two sources of population-level heterogeneities in the data that can significantly impact the results: the Twitter activity and the mobility. We want to eliminate the effects of the specific source of data, hence we selected only those users with a level of activity which allows for a reliable analysis of the user’s trajectory properties. Specifically, we removed users with anomalous movement speed (over 80m/s) and an anomalous number of tweets. Then, we selected users that have tweeted at least in 2 distinct locations (e.g. home and work), so that the trajectory entropy is not zero in any case, and have an average activity of approximately one tweet per day. Furthermore, we resampled the data in 30 minutes intervals to mitigate the bursty dynamics of the tweeting behavior [61]. The resampling process also avoids artificially lowering the entropy and skewing the radius of gyration of the trajectory disproportionately towards those locations where the bursty behavior happened. In the rare event a
Figure 4.1: The data of the trajectories need to be processed to remove invalid users (e.g. robots) and select those trajectories with enough data to guarantee a reliable analysis. The resampling and interpolation step is also important to mitigate the bursty behavior of human dynamics. Once the data are curated, we can extract the feature vectors to cluster and characterize the users.

user tweeted from multiple locations within a 30 minutes bin, we selected the most popular location among them (or the first one in the case of a tie). We also did some basic interpolation by carrying over the location identifier of the location visited in the previous half hour if we did not have data for an immediately consequent interval. This step assumes it is very likely the user did not move to another location in such a short time span. Finally, we removed users who were present in more than one dataset, retaining only the trajectories from where the users appeared the first time; this step avoids relating trajectories that were collected temporally far apart. We selected 4,185 users in $\mathcal{D}_1$, 861 users in $\mathcal{D}_2$, and 933 users in $\mathcal{D}_3$, hence we conducted the analysis on a total of 5,979 unique users.
Table 4.1: Overview of the Mobility Datasets for User Characterization

<table>
<thead>
<tr>
<th></th>
<th>$D_1$</th>
<th>$D_2$</th>
<th>$D_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td># of users</td>
<td>108,341</td>
<td>152,671</td>
<td>147,942</td>
</tr>
<tr>
<td>Total number of records</td>
<td>1,022,286</td>
<td>1,068,939</td>
<td>1,179,389</td>
</tr>
<tr>
<td># of users after filtering</td>
<td>4,185</td>
<td>861</td>
<td>933</td>
</tr>
<tr>
<td>Date range</td>
<td>Aug 28–Sep 29, 2014</td>
<td>Jun 17–Nov 04, 2016</td>
<td>Apr 05–Sep 11, 2017</td>
</tr>
</tbody>
</table>

4.1.2 Feature Extraction and User Clustering

Unlike location classification (cf. Section 3.2.3), we do not know the possible user classes a priori. For this reason it is not possible to assign labels to the users for a supervised learning process. Instead, we rely on similarity metrics to discover and group users in alike classes. The main methodology is to apply the spectral clustering algorithm on the users’ feature vectors extracted from the GPS trajectories. We used the Bayesian Information Criteria (BIC) and silhouette score to identify the optimal number of clusters. Furthermore, we retain only those clusters that contain at least 5% of the users in order to be able to draw reliable conclusions as a small sample of users would be too sensitive to noise in the data.

One aspect that needs to be defined is how to represent the users through some descriptors (i.e. the feature vectors) that are able to discriminate different types of users. We decided to focus our attention on two types of features: popular-times and popular-days histograms and hourly and daily entropies.

The popular-times and popular-days histograms are similar to the features we built in Section 3.2.2 for the locations. The two histograms are concatenated to form a 31-dimensional feature vector for each user. A one-dimensional Gaussian filter is applied over the histograms to smooth the distribution and fill in missing data. Finally, the feature vectors are normalized by applying standardization. The analysis of spectral clustering revealed an optimal number of clusters $k = 15$ (for
The assumption is that people have strong habits \cite{62} and such behavior property implies that the effect cumulates over time reinforcing the hours and days at which a person is mostly active. For example, someone that is very active on the weekends and late hours might represent a younger person profile than someone that follows a 9–5 hours routine; consequently, even the type of locations and the movement patterns should differ substantially.

The hourly entropy and the daily entropy (Figure 4.2) are similar to the popular-times and popular-days histograms, but the quantity computed per hour or per day is the estimated entropy $S_{est}$ of the sequence of the visited locations (Equation (2.19)). The hourly entropy is computed as the entropy of the locations visited by the user when the user’s trajectory is aggregated per hour of the day over the period of observation. The daily entropy is computed as the entropy of the locations visited by the user when the user’s trajectory is aggregated per day of the week over the period of observation. The analysis of the spectral clustering of hourly and daily entropy feature vectors revealed an optimal number of clusters $k = 9$ (for more information see Appendix A.3). This type of features is especially powerful because it captures at the same time both the spatio-temporal activity of the user’s movement per unit of aggregation and the longitudinal recurrent patterns. For instance, locals might exhibit very low entropy during the work hours and days and a sudden increase of entropy during the weekends; on the other hand tourists should show a relatively high and stable entropy independently of the day of the week. Also, if we observe for example that a user exhibit every Monday a sequence of the type Home → Coffee shop → Work his entropy on Monday will be lower than somebody that does not show recurrent habits (both in the same day and across the day of the week over the period of observation). Finally, it should be
Figure 4.2: The hourly and daily entropy feature vectors are computed using the Lempel-Ziv entropy estimate to capture long-term correlations in the sequence of visited locations. The hourly entropy is computed on all the locations visited by the user, in order of appearance, per hour bin over the whole dataset. The daily entropy is computed similarly, but aggregating per day of the week.

noted that we do not need to know the exact type of the locations, just that they are the same locations (as per the clustering applied to identify the POIs), and therefore the methodology offers great generality.

4.1.3 User Type Characterization

Since the labels assigned by a clustering algorithm do not have any real-world equivalent, we need to characterize those classes indirectly. For this purpose we study several properties for each class of users both from a statistical and descriptive point of view. There are two metrics that are especially important in human mobility: the radius of gyration (see Section 2.2.2), which considers only the spatial aspect of the trajectory, and the entropy (see Section 2.2.5), which
captures temporal correlations. The radius of gyration has been found to follow a truncated power law distribution (see Equation (2.9)). Such broad distribution is the result of the heterogeneity of user’s mobility and it is a strong indicator of multiple classes of users being aggregated together. Therefore it should be possible, at least in principle, to separate these classes of users. The entropy of a user trajectory can be estimated using Equation (2.19). The lower the entropy, the higher the predictability of a user (Figure 4.3). At the same time, the larger the radius of gyration, the more likely a user has visited more distinct locations and consequently the higher the entropy (Figure 4.4). It should be noted that $S_{est}$ does not depend on the trajectory length (as long as it is long enough to provide a good estimate of the entropy), but only on the sequence itself, and therefore it is possible to compare sequences with a different length [3, 117].

Figure 4.3: The entropy and max predictability distributions. $S_{rnd}$ is the random entropy, $S_{unc}$ is the Shannon entropy, and $S_{est}$ is the Lempel-Ziv entropy estimate and they are associated respectively with the max predictabilities $\Pi_{rnd}$, $\Pi_{unc}$, $\Pi_{max}$. The true entropy estimate $S_{est}$ is peaked between 1 and 2 meaning user’s trajectories are well approximated by about 4 distinct locations and therefore the predictability is very high for most users ($\geq 60\%$).
Figure 4.4: The radius of gyration and the entropy estimate show a strong correlation. Even though $S_{est}$ is mostly independent of the trajectory length, a high radius of gyration is associated with the user exploring locations far apart, thus the more complex trajectory reflect in a higher entropy.

We tested the radius of gyration, the entropy, and the predictability distributions of different classes in two ways. First, by using the two sample Kolmogorov-Smirnov (KS) test [115] with the underlying hypothesis that different classes should exhibit different distributions for the considered metrics. The KS statistic can be used to test whether two underlying one-dimensional probability distributions differ, i.e. it checks the null hypothesis $H_0$ that the two data samples come from the same distribution. Given two samples of size $n$ and $m$ respectively, the two sample KS statistic is defined as

$$ D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| , $$

(4.1)

where $F_{1,n}$ and $F_{2,m}$ are the empirical distribution functions of the first and the
second sample. The null hypotheses \( H_0 \) is rejected at a confidence \( p \) if

\[
D_{n,m} > c(p) \sqrt{\frac{n + m}{nm}},
\]

(4.2)

where \( c(p) = \sqrt{-\frac{1}{2} \ln(p/2)} \) equals to 1.36 for a 95% confidence. To study each cluster we run this test for all the possible combinations of cluster labels (which for \( n \) clusters, they are simply the combinations of two elements of the first \( n \) integers. The number of such combinations is \( \binom{n}{2} \)). Second, such procedure can introduce a bias on the \( p \)-values that should be adjusted to account for the multiplicity of tests. In order to ensure that the results from the KS tests are actually robust we further test that the distributions of the quantities under study are effectively not the same by using the \( k \)-sample Anderson–Darling (AD) test \([85]\). It tests the null hypothesis \( H_0 \) that \( k \) samples are drawn from the same population without having to specify the distribution function of that population.

Then we characterized the classes using the content of the tweets posted by the users to associate a qualitative description, namely a label, to specific spatio-temporal properties of the user’s trajectories, similarly to the process described in Section 3.3 for the locations. A simple word frequency analysis using Rapid Automatic Keyword Extraction (RAKE) algorithm \([79]\) did not reveal significant patterns, therefore we focused on the study of the profile of locations visited by the users inside each cluster, that is the distribution of the type of locations.

The location categories for the POIs were obtained by parsing the text of the tweets for a Foursquare check-in from which we extract the type of the location. Foursquare and Twitter actively prevents crawling of their website, therefore we extracted location only from a subsample representing 20% of the tweets from
each class of users. We extracted a total of 1,987 check-ins. We also consolidated the name of similar categories following the hierarchy used by Foursquare\(^1\). The final categories we considered were: Restaurant & Food, Nightlife Spot (\textit{e.g.} bars, pubs, nightclubs), Coffee Shop (originally it was listed under restaurant & food, but we kept it separated as it is popular among all the clusters and less specific), Arts & Entertainment (\textit{e.g.} events, concerts, museums, monuments), Outdoors & Recreation (\textit{e.g.} gyms, parks, scenic views), Shop & Service (\textit{e.g.} stores, markets), Travel & Transport (\textit{e.g.} hotel, motel, bus/train/metro stations), Professional & Other Places (\textit{e.g.} offices, schools, hospitals, private homes). The distribution of the categories extracted from the check-ins is presented in Table 4.2, where the restaurant and food category were the most popular as expected.

Table 4.2: Foursquare Check-ins Category Distribution

<table>
<thead>
<tr>
<th>Category</th>
<th>Size (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restaurant &amp; Food</td>
<td>22.395571</td>
</tr>
<tr>
<td>Nightlife Spot</td>
<td>16.809260</td>
</tr>
<tr>
<td>Outdoors &amp; Recreation</td>
<td>16.758933</td>
</tr>
<tr>
<td>Shop &amp; Service</td>
<td>11.373931</td>
</tr>
<tr>
<td>Professional &amp; Other Places</td>
<td>10.317061</td>
</tr>
<tr>
<td>Travel &amp; Transport</td>
<td>7.498742</td>
</tr>
<tr>
<td>Arts &amp; Entertainment</td>
<td>6.894816</td>
</tr>
<tr>
<td>Coffee Shop</td>
<td>5.938601</td>
</tr>
</tbody>
</table>

The distribution of types of places is expected to vary significantly for example between locals, who have a typical home–work routine, and tourists who will visit mostly monuments and attractions. One caveat is that restaurants, bars, and coffee places are generally popular among every user category. While such charac-

\(^1\)https://developer.foursquare.com/docs/resources/categories
terization would not be possible on a dataset of only GPS points and associated
timestamps, in the context of this work the additional data extracted from the
text of the tweets are solely used to get a better understanding of the efficacy of
the methodology we introduced, without any loss of generality.

4.2 Popular-times and Popular-days

The spectral clustering of the popular-times and popular-days feature vectors re-
turned 8 clusters with at least a 5% user’s representation. In these 8 clusters we
can clearly observe typical hourly and daily patterns (Figure 4.5). For instance,
we have peaks early in the morning, at noon, at 8pm and at 1 am. Weekly activity
is more balanced with only few clusters showing a clear week-day versus week-end
patterns.

Figure 4.5: Average popular-times (left) and popular-days (right) feature vectors
for each cluster of users. We represented only those clusters with a size of at least
5% of the users in the dataset. The gray areas indicate the typical working hours
and work-days. Distinct hourly and daily patterns are visible. For example, users
in cluster 4 are active during working hours and work-days, while users in cluster
1 and 15 are active in the evening/night hours and the weekend.
The KS statistics confirm that most clusters of users exhibit significantly different distributions for the quantities of interest (compactly represented using box-plots in Figure 4.6). The AD statistic for the 8 samples from the clusters supports the KS statistics ($p \approx 10^{-10}$ for the radius of gyration, $p \approx 10^{-66}$ for the estimated entropy, and $p \approx 10^{-29}$ for the max predictability). However, there are pairwise exceptions for the KS statistic. The radius of gyration distribution does not seem to vary much between clusters, for example clusters 0 and 4 ($p \simeq 0.50$), clusters 4 and 12 ($p \simeq 0.60$), and clusters 7 and 15 ($p \simeq 0.45$). There are also exceptions for the entropy, such as clusters 1 and 12 ($p \simeq 0.13$) and 1 and 15 ($p \simeq 0.14$). Finally, clusters 4 and 12 ($p \simeq 0.52$) and clusters 7 and 12 ($p \simeq 0.64$) did not reject $H_0$ for the max predictability.

Figure 4.6: The distributions of the radius of gyration, entropy estimate and predictability of the trajectories among the clusters with a size representing at least 5% of the users. Between some of the clusters of users there are significant differences pointing to potentially fundamentally different movement patterns between types of users.

While at first these results seem in contrast (at least partially, as not every pair of clusters is showing differences) they actually give a new insight into human mobility. In fact, given two types of users we can end up into three possible configurations:
I The intuitive situation where different activity patterns should have different
distributions for the trajectory properties (*e.g.* clusters 12 and 15) and vice versa.

II Classes of users with similar times and days of activity can have wildly dif-
ferent trajectory properties (*e.g.* clusters 0 and 15).

III Classes of users with distinct times and days of activity might exhibit sur-
prisingly similar trajectory properties (*e.g.* clusters 1 and 15).

As an additional graphical example, let’s consider clusters 0, 4, and 15 in Figure 4.5. Cluster 4 clearly represents work-day activity having a peak of activity in the middle of the day and work-days. Clusters 0 and 15 on the other hand have a very different pattern of activity and show the maximum activity in the evening (at pretty much the same time) and during the weekends (see Figure 4.5). Although clearly distinct based on the activity patterns, cluster 0 and 4 have a much more similar entropy distribution and max predictability (even though still different) than cluster 15 which shows a much lower entropy. This simple example let us introduce another counter intuitive kind of observations enabled by this analysis method. We know that the entropy is affected by the complexity of the trajectory, therefore we would expect that the routinely work days are the ones with the lowest entropy, while the weekends and night activities is where people try new places and the entropy should be higher. In fact, it is the exact opposite, as shown in Figure 4.7. This type of observations enabled by the framework we introduced show how powerful of a tool it is in the study and characterization of human mobility.

We then conducted the analysis of the type of places visited by the users of each cluster using the Foursquare check-ins. First of all, we notice that even
Figure 4.7: Radius of Gyration, entropy and predictability distributions for the typical “work-day activity” users (cluster 4) and “dinner and nightlife activities” (clusters 0 and 15). While users in clusters 0 and 15 have more similar temporal patterns than users in cluster 4, clusters 0 and 4 have more similar distributions (although statistically different) for the trajectory properties than cluster 15. Moreover, Cluster 4 exhibit higher entropy than the cluster 15 even though work-day activity is supposed to be a routine activity.

though the sampling procedure is completely fair, different clusters have relatively different number of Foursquare check-ins. It is possible that users usually check-in more frequently in Foursquare during daytime activities [21] which represents a limitation of using Foursquare as a proxy of visited places. Keeping in mind this limitation, we can try to understand how categories break down given specific temporal patterns (Figure 4.8). We observed that every cluster with a late hourly activity peak has restaurant and nightlife spot as popular categories, with the exclusion of cluster 1, which has a peak for travel and transport and outdoor. The anomaly of cluster 1 can be explained by the limited number of check-ins and the fact that is possible that late at night people tweet on their way back home from metro stations and the like. Clusters with morning or afternoon activity (12 and 4) have a high activity of professional places and outdoor. These simple results confirm that hourly and daily peak activity are a good descriptor for the type of locations visited by a user.
Figure 4.8: Each of the plots represent the location categories breakdown (%) for the corresponding cluster (identified by the label in the top right corner). There are distinctive distributions of location types that characterize each cluster of users.

4.3 Hourly and Daily Entropy

The spectral clustering of the hourly and daily entropy feature vectors returned 4 clusters with at least a 5% user’s representation. In these 4 clusters we can observe distinctive hourly and daily patterns (Figure 4.9). Users in clusters 1 and 2 are associated with higher entropy in the weekends and late at night. Cluster 5 shows lower entropy during the weekends and a sudden increase of entropy around 10am. Clusters 6 resembles the typical average behavior (notice the circadian rhythm) of the population as a whole.

Once again, most of the pairwise KS statistics reject the null hypothesis with a 95% degree of confidence, with the sole exception of the pair of clusters 1 and 5 with regard to the radius of gyration. The 4 sample AD statistic is also strongly
in agreement ($p \approx 10^{-41}$ for the radius of gyration, $p \approx 0$ for the entropy, and $p \approx 10^{-161}$ for the predictability). A compact representation of the cluster distributions of the radius of gyration, the entropy, and the max predictability is represented in Figure 4.10. We observe cluster 1 has very low entropy with several outliers (resulting from a bimodal distribution as seen in Figure 4.11), and cluster 6 has a much higher entropy (lower predictability) than the other clusters.

We can compare pairs of clusters with the respective hourly and daily patterns to try to extract meaningful insights. Since the hourly and daily entropy feature vectors are based on the entropy, the differences in terms of entropy distribution between clusters of users are maximized (Figure 4.11). Interestingly however is the pair of clusters 1 and 5. Cluster 1 is interesting because exhibit a bimodal distribution. Note also the two peaks late at night and at around 9 am could be the cause of the bimodal entropy distribution (Figure 4.9). It is possible we are observing two classes of users: one which is active in the weekend and goes out...
Figure 4.10: The distributions of the radius of gyration, entropy estimate and predictability of the trajectories when the users are clustered based on their hourly and daily trajectory entropies. We represented only clusters with a sample size of at least 5% of the total number of users. Between clusters of users there are significant differences pointing to fundamentally different movement patterns between types of users.

late at night ("nightlife class"), and one that is active in the weekend mornings ("family class"). Cluster 5 exhibit an unusually high entropy with a large spike in the entropy at around 10 am and lower entropy in the weekend (Figure 4.9); it is possible that the large spike in the entropy is caused by people leaving their home to reach the workplace.

Figure 4.11: The radius of gyration, the entropy and the maximum predictability distributions when users are clustered according to their hourly and daily trajectory entropies. The distributions are significantly different and exhibit non trivial patterns (e.g. cluster 1) or strong divergence from the other clusters of users (e.g. cluster 6).
The analysis of the distribution of the types of locations inside of the clusters could shed some light on the observed patterns (Figure 4.12). Cluster 1 location representation can explain the bimodal distribution; a high representation of nightlife and restaurant (nocturnal activities) is combined with the art and entertainment, outdoors & recreation, and shop & service categories (diurnal activities). Cluster 2 is clearly linked to activities in the evening and night. Cluster 5 shows a high representation of outdoors activities, which could explain the unusually high entropy rate in the morning, along with shop & service and professional places, which could explain the low entropy in the weekends and night hours. Finally cluster 6 has no clear pattern being a mix, which would explain the “average” behavior.

Figure 4.12: Each of the plots represent the location categories breakdown (%) for the corresponding cluster (identified by the label in the top right corner). There are distinctive patterns that characterize each cluster of users. For instance, cluster 2 users mostly visit nightlife spots, while cluster 5 users do a lot of outdoors activities.
In conclusion, we have shown how the popular-times and popular-days along with the hourly and daily entropies are powerful features which can be built from simple GPS trajectories without any additional data. Such features made possible to identify classes of users that can then be compared using well understood human mobility metrics and non-parametric statistic tools. The classes of users showed significantly different spatio-temporal patterns associated with the time and day of activity. The efficacy of the methods were further verified by studying the distribution of the categories of the visited locations inside each cluster. One question that remains open is what are exactly these user types and what impact they have on the modeling of human mobility and human dynamics.
Chapter 5

Significance, Impact, and Conclusion

The framework and methods introduced and discussed in this work are an attempt to automatically discover and classify locations and users based on geo-temporal data without relying on contextual information. By taking advantage of the ubiquity of smartphones with GPS capability which allow to collect thousands of user’s trajectories with a high degree of accuracy, users locations do not need to be known a priori. Furthermore, users can be characterized by analyzing the spatio-temporal aspects of the location augmented data providing insights beyond simple scaling relationships. The benefit of such approach is its generality and applicability to datasets coming from different and heterogeneous sources. In essence, the main contributions of this work can be summarized as follows:

1. We showed how users’ POI can be identified with high spatial accuracy using clustering techniques and user’s trajectories.
We introduced and proved that popular-times and popular-days histograms are powerful features for POI classification.

We showed that locations can be associated to activities or events based on their temporal and spatial properties.

We introduced and proved the effectiveness of hourly and daily user’s activity and entropy as features to identify and characterize specific groups of users.

We introduced a methodology to compare group of users based on the spatio-temporal properties of their trajectories.

In the framework we introduced, we showed that the popular-times and popular-days histograms of POIs alongside with the characterization of the users based on the user’s popular-times and popular-days of activity and hourly and daily trajectory entropy improve the ability to analyze geo-temporal data by extracting more structured information. The framework enables us to get new insights that can be used to further the understanding of human mobility and activities. This is, in turn, of great importance to public health, city planning, and crime reduction. For instance, there is a lot more information on knowing that people tend to go from a restaurant to a nightclub, than knowing in a more general way that they went from point A to point B (with unclassified data). On the same line, knowing the movement pattern of specific classes of users could be especially important for monitoring the spread of diseases and for national security. The regularities in human movement based on classified information are a lot more powerful than regularities in unclassified data.

The geo-mapping of locations and the user characterization also find many applications in the industry such as location-based services and recommendation
systems (e.g. Yelp, Trip Advisor). Currently these companies rely on expensive local inspection, by convincing businesses to subscribe to their paid services, or by relying on user reporting. It is quite common for users to be asked to select a specific location they are in, if they want to “check-in” using their location-based services. Our classification and characterization can make this task irrelevant given that the location can be classified later, that is, users will not need to select where they are because this can be automated. Our approach also hints to a more unified approach for location classification; while there has been some data sharing between companies, such data has been an increasing competitive advantage, and therefore is considered extremely valuable. A general method as the one presented has the potential to drastically reduce the cost of this services while simultaneously improving the quality of the service.

Currently, there are two outstanding limitations of the proposed framework which could be addresses in future works. First, location classification still requires to properly train a classifier but there is a lack of availability of an appropriate dataset. Furthermore, we have to consider that temporal patterns may be subject to regional, seasonal, and cultural differences which have to be taken in account. For example, the dinner time in USA is usually much earlier than in Spain. Second, the user characterization tells us only that there are different classes of users and which patterns they exhibit, but it cannot tell us what those classes actually are and why different classes exist. In order to get a greater insight, we need additional data and further studies.
Bibliography


method? answers via model-based cluster analysis. The computer journal,

[31] Xavier Gabaix, Parameswaran Gopikrishnan, Vasiliki Plerou, and H Eugene

[32] Lorenzo Gabrielli, Salvatore Rinzivillo, Francesco Ronzano, and Daniel Villa-
toro. From tweets to semantic trajectories: mining anomalous urban mobility

[33] Riccardo Gallotti and Marc Barthelemy. Anatomy and efficiency of urban

[34] Marta C González, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding


[36] Samiul Hasan, Christian M Schneider, Satish V Ukkusuri, and Marta C
González. Spatiotemporal patterns of urban human mobility. Journal of Sta-

[37] Eugene Helfand. Theory of inhomogeneous polymers: Fundamentals of the
gaussian random-walk model. The Journal of Chemical Physics, 62(3):999–
1005, 1975.


[40] Sage Jenson, Majerle Reeves, Marcello Tomasini, and Ronaldo Menezes. Mining location information from users’ spatio-temporal data. IEEE.


Appendix A

Evaluation of Clustering Algorithms

Several methods were used to evaluate the quality of the clustering algorithms, including the number of clusters found, the Bayesian information criterion (BIC) [30], the number of unclustered points, and the silhouette score [80] as a function of the specific parameters of the clustering algorithms. The BIC is a criterion to choose a specific model from a finite set of models. The BIC weighs the complexity of the model with how well the model fits the relevant data and we look to maximize the BIC. The BIC is defined as

\[
\text{BIC} = \ln(n)k - 2 \ln(\hat{L}) \approx -2 \ln \hat{L} + k(\ln(n) - \ln(2\pi)), \quad (A.1)
\]

where \( \hat{L} \) is the maximized value of the likelihood function of the model, \( n \) is the number of data points, and \( k \) the number of free parameters to be estimated. The silhouette score is a measure of intra and inter cluster distances, or how dense
clusters are compared to how far apart different clusters are. Because of this, we did not include noise points when calculating the silhouette score for both HDBSCAN and mean shift. The silhouette score is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

(A.2)

where $a(i)$ is the average dissimilarity of point $i$ to a cluster $c$ computed as the average of the distance from $i$ to all points in $c$ and $b(i)$ is the lowest average dissimilarity of $i$ to any other cluster of which $i$ is not a member. An optimal silhouette score is close to 1. The average $s(i)$ overall data of the entire dataset is a measure of how appropriately the data have been clustered.

The BIC and silhouette scores suffer of some shortcomings: the clusterings produced by HDBSCAN and mean shift have variable numbers of noise points which are not considered; also, they are a centroid-based criterion, thus may not work well with oblong clusters. This is because these two metrics examine the standard deviation for a cluster around a centroid, which may not be an accurate representation of the quality of a density-based clustering. Therefore the most important metrics for HDBSCAN are the number of clusters and the number of points left unclustered. We measure unclustered points because we don’t want to incorrectly classify noise points; either by classifying location points as noise or noise points as locations.

A.1 POI Identification

In Section 3.2.1 we used clustering to identify the points of interest visited by the users. Since clustering is an unsupervised learning method, we need to carefully
evaluate the results. In this specific case, the low dimensionality of data helps clustering algorithms and metrics to perform the best as well as allowing us to visually inspect the result of the clustering.

A.1.1  

**k- Means**

*K*-means is the simplest algorithm used because it is strictly centroid-based. There is a decisive maximum for the silhouette score (Figure A.1b). The silhouette maximum (after which silhouette barely increases) coincides with the elbow of the BIC (Figure A.1a). This leads us to conclude that the optimal value for *k*-means is *k* = 25,000.

![Figure A.1: Clustering metrics for k-means. *k* is the number of clusters.](image)

A.1.2  **Mean Shift**

Mean shift was tested with and without noise classification. We are interested in finding a bandwidth parameter that produces a satisfactory clustering. The elbow of the number of clusters (Figure A.2a) and the elbow for the BIC (Figure A.2d) suggest a bandwidth within [22.2, 44.5]. The number of noise points (Figure A.2b)
remains relatively constant within this range and the silhouette score (Figure A.2c) decreases relatively linearly with respect to the bandwidth parameter.

Figure A.2: Clustering metrics for mean shift with noise classification ($m_n$) and without noise classification ($m_p$), i.e. it partitions the data. $b$ is the bandwidth parameter expressed in meters.

### A.1.3 HDBSCAN

The number of clusters (Figure A.3a) levels out around a minimum cluster size between 5 and 10. This is also the point where the BIC becomes linear (Figure A.3d). For HDBSCAN the silhouette score (Figure A.3c) does not tell us as much as for the centroid-based algorithms. The number of noise points (Figure A.3b)
has a clear elbow in the aforementioned range that agrees with the elbow of the BIC.

![Graphs](image)

(a) Number of Clusters  
(b) Number of Noise Points  
(c) Silhouette Score  
(d) BIC

Figure A.3: Clustering metrics for HDBSCAN. $m$ is the minimum number of points to be in a cluster (i.e. cluster size).

### A.2 Extraction of Events and Activities

In Section 3.3 we used spectral clustering to cluster the feature vectors representing the locations. Since the feature vectors have a high dimensionality, the property of spectral clustering of learning a manifold in a lower dimensional space is optimal to reduce the effect of the “curse of dimensionality”. Furthermore, unlike several
other clustering algorithms (such as $k$-means), spectral clustering does not assume that a cluster follows a normal distribution around a centroid [102]. We attempted several other clustering algorithms, including K-means and HDBSCAN. However, $k$-means clustered based on popular days of the week, and HDBSCAN classified nearly every point as noise.

To prepare our feature space for clustering (i.e. to remove sparse clusters and incorrectly found clusters), we limited the analysis to those locations that have over 24 data points. Additionally, to limit esoteric clusters such as homes, we imposed a minimum number of users at each location. We fixed the minimum cluster size, while varying the minimum number of users in a cluster as well as the number of clusters. Among all of the variable filtering parameters, the BIC has an elbow around 15 clusters (Figure A.4a). The silhouette score is not very informative (Figure A.4b) but together with the BIC as a function of the number of users (Figure A.4a left), it reinforces the selection of number of cluster as $k = 15$.

A.3 User Clustering

In Chapter 4 we employ clustering in two occasions. Once for clustering the GPS point of the datasets $D_2$ and $D_3$, similarly to what we did in Section 3.2.1 for the dataset $D_1$. It is of no surprise that HDBSCAN parameter, the minimum number of points in a cluster $m = 7$, is consistent among the datasets as they are all from Manhattan area and are relatively similar in the number of points. We show just the number of clusters as function of $m$ (Figure A.5) as the analysis is very similar to what presented in Appendix A.1.
Figure A.4: Metrics for spectral clustering on filtered location data. $k$ is the number of clusters. (a) The normalized BIC as function of the minimum cluster size (right) and the minimum number of users (left). (b) The silhouette score $s$. The color refers to the filtering of the data, specifically the minimum number of users per cluster.

We also clustered two types of feature vectors using spectral clustering. Here we try to identify the correct number of clusters $k$. The clustering scores for the feature vectors built out of popular-times and popular-days histograms show a strong elbow at $k = 15$ (Figure A.6) while the clustering scores for the feature vectors built out of hourly and daily entropies show a strong elbow at $k = 9$ (Figure A.7).
Figure A.5: The number of clusters drops sharply and has an elbow when the minimum cluster size is $m = 7$.

Figure A.6: Silhouette $s$ and BIC score as function of the number of clusters $k$ when clustering the popular-times and popular-days user’s feature vectors.

Figure A.7: Silhouette $s$ and BIC score as function of the number of clusters $k$ when clustering the hourly and daily entropies user’s feature vectors.
Appendix B

Evaluation of Location Filtering Criteria for Classification

In Section 3.2.3 we used several filtering criteria to improve classification accuracy of the locations we labeled using Foursquare. We grouped similar categories and removed underrepresented categories (categories represented by less than two percent of the locations, so that we have enough feature vectors for each category to effectively train the classifier). We also filtered the locations based on the amount of activity received (i.e. number of tweets associated with the cluster representing the location) and popularity of the location among users (i.e. the number of users that tweeted from a specific location). The filtering step is crucial to obtain a set of locations with robust statistics and maximize the learning ability of the classifier.

A low number of users associated with a location could imply a private residence and because there are no given data on private residences, locations representing residences should be removed. Therefore we impose a minimum number of users per location (Figure B.1a). We found that clusters with more than 10 users pro-
vided optimal results.

We also checked to filter the maximum cluster size, i.e. the number of tweets per cluster, because New York has a high density of buildings and skyscrapers. By using longitude and latitude points, we cannot factor in the fact that locations might be on top of each other by nature of the locations being on different floors. There is some evidence that this might be the case as we found that the location size distribution follows a power-law and there are few locations that have up to 10,000 tweets. Furthermore, very close locations might be within the GPS error range. By filtering out too dense locations we are hoping to filter out clusters that are actually several locations. We found the best improvement by limiting maximum cluster size to 80 tweets (Figure B.1b).

Finally, we set a minimum level of activity by imposing a minimum cluster size (Figure B.1c). Such criteria make sure that the histograms on which the feature vectors are built are less sparse. We found that clusters with more than 24 tweets provided optimal results.

We conducted a similar analysis for locations labeled using the Google Places API, although relatively less sensitive to parameters change, and we found out the minimum number of users to be 20 (Figure B.2a), a maximum cluster size of 80 tweets (Figure B.2b), and a minimum cluster size of 36 tweets (Figure B.2c).

As we varied the number of users and the number of tweets per cluster, we tested a SVM classifier using five random-split cross-validation and studied the performance of the classifier in term of accuracy, precision and recall. These metrics
are defined as:

\[
\text{Precision} = \frac{t_p}{t_p + f_p}, \quad (B.1)
\]
\[
\text{Recall} = \frac{t_p}{t_p + f_n}, \quad (B.2)
\]
\[
\text{Accuracy} = \frac{t_p + t_n}{t_p + t_n + f_p + f_n}, \quad (B.3)
\]

where \(t_p, t_n, f_p, f_n\) stands for true positive (correct classification), true negative (correct rejection), false positive (misclassification, type I error), and false negative respectively (miss, type II error) for a specific class; the metrics are computed for all the classes from the confusion matrix and then averaged to obtain the overall model values.
Figure B.1: Filtering parameters analysis for locations labeled using Foursquare.
Figure B.2: Filtering parameters analysis for locations labeled using Google.
Appendix C

List of Publications

C.1 Publications Related to This Dissertation

Marcello Tomasini and Ronaldo Menezes. “Characterization of Users by Using Hourly and Daily Spatio-temporal Patterns Extracted from GPS Trajectories.” [UNDER PREPARATION]


C.2 Other Ph.D. Publications


