Understanding Human Mobility Patterns Through Mobile Phone Records: A cross-cultural Study

by

Yan Ji

Submitted to the Department of Civil & Environmental Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Civil and Environmental Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

© 2011 Massachusetts Institute of Technology. All Rights Reserved.

Signature of Author.....
Department of Civil & Environmental Engineering
May 4, 2011

Certified by .....
Marta C. González
Assistant Professor of Civil & Environmental Engineering
Thesis Supervisor

Accepted by .....
Heidi M. Nepf
Chair, Departmental Committee for Graduate Students

Understanding Human Mobility Patterns Through Mobile Phone Records: A cross-cultural study

by

Yan Ji

Submitted to the Department of Civil and Environmental Engineering
on May 4, 2011, in partial fulfillment of the
requirements for the degree of
Master of Science in Civil and Environmental Engineering

Abstract

In this thesis, I present a cross-cultural study on human's trip length distribution and how it might be influenced by regional socio-economic factors, such as population density, income and unemployment rate. Mobile phone records contain very detailed calling information of the spatiotemporal localization of hundreds of thousands of users, which can be used as proxies for human trips. The traveling behaviors of 24 autonomous regions in San Francisco (5 regions), Dominican Republic (3 regions) and a European country (16 regions) are studied through these rich mobile phone data sets. We found that people in different regions have very heterogeneous aggregate traveling patterns (trip length distribution) which can be generally grouped into four distinct families. The result of Self-organizing map shows that the trip length distribution has a certain degree of correlation to population density, which sparks our interests to conduct a thorough research on factors such as population density and income that can potentially influence the trip length distribution and human's traveling behavior.

Using a double exponential function to fit the radius of gyration distribution (i.e. a proxy to the trip length distribution), we are able to characterize human's traveling behavior with four parameters. By applying principle component analysis, the parameter space is transformed orthogonally and two principal components which contribute most to the variance of sample set are extracted. We tempted to find the regression relationship between population density and each of the components. However, the R^2 is not enough high for estimation purposes.

With the extensive information source regarding household income, median age, unemployment rate, we were able to conduct a multiple regression analysis in San Francisco Bay area. Using radius of gyration as regressand, population density, income, age, and unemployment rate as regressors, we found the R^2 is over 30%, which is sufficiently good for cross-sectional data analysis. Additionally, the significant estimated coefficients indicate that people living in wealthier and unpopulated areas tend to travel more frequently and make long distance trips. Furthermore, descriptive comments are provided for the connection between parameters in the fitting function and

population density and income.

Thesis Supervisor: Marta C. González
Title: Assistant Professor

Acknowledgments

The two years academic experiences at MIT were challenging but colorful. I would like to express my sincere appreciation to Department of Civil and Environmental Engineering, without which I cannot take the first crucial step as a MIT graduate student. I still remembered the first class 1.200 taught by Professor Wilson, which enlightened my interests in the underlying microeconomic theory of transportation policies. This class has gotten me interested in economics and motivate me to pursue a Ph.D. in MIT's economics department. Profiting from the flexible coursework of the Master program in Civil Engineering, I got the opportunities to take economics and finance classes, which enables me to better understanding existing economic theories and truly pursue my interests.

At MIT, the most important person in my life is Professor Marta González, who is my research advisor. I would like to express the highest appreciation to Marta's careful guidance and kind help in my two-year study. Marta taught me the way to conduct high-level research, and most importantly how to figure out original and interesting ideas. I admire Marta's dedication and intuition in research as a top researcher in her area. Moreover, I am deeply grateful to her for allowing me to take so many finance classes and his strong recommendation letter, which plays a vital role in my economics Ph.D. application.

I also want to thank my second-year academic advisor Professor Odoni and Professor Manso of Finance, who gave me invaluable advices in terms of coursework. I also appreciate their recommendation letter and in particular, Professor Manso's financial support for the fourth semester.

Finally, my deepest gratitude goes to my family, my parents for their love, understanding, constant support and encouragment. This work is dedicated to them.

Contents

1 Introduction **15**

1.1 Motivation and Problem Statement 15

1.2 Thesis Outline 17

2 Mobility Measures **21**

2.1 Data 22

2.1.1 Mobile Phone Data 22

2.1.2 Population Data 25

2.1.3 Social-economic Data 27

2.2 Calculation of Population density 27

2.3 Pattern measures 28

2.3.1 Characterizing Individual Calling Activity 28

2.3.2 Observations at A Fixed Inter-event Time 29

2.3.3 The Radius of Gyration Distribution 31

2.3.4 The Frequency of Visiting Different Locations 35

2.3.5 Modeling Human Mobility Patterns Using Spatial Density Function 36

2.3.6 Mobility Characterization around Towers 38

3 Identify Inter-relationship from Kohonen Map **41**

3.1 Kohoen Map and Multidimensional Clustering 41

3.2 Identification of Relationship Between Radius of Gyration and Population Density from Kohoen Map 43

3.2.1 Training Sample and Vector Construction 43
3.2.2 SOM with Geographical Data 44
3.2.3 SOM without Geographical Data 47

4 Mobility Characterization and Data Mining **49**
4.1 Samples Construction and Parameterization of Mobility 49
4.1.1 Area Divisions 50
4.1.2 Heterogeneity Classification 51
4.1.3 Parameterization of Radius of Gyration 53
4.2 Principal Component Analysis 57
4.2.1 Introduction 57
4.2.2 Computational Procedures 59
4.2.3 PCA Results 62

5 Regression Analysis **67**
5.1 A Simple Linear Regression Model 67
5.2 Multivariate-Regression Model 69
5.3 Regression in San Francisco Bay Area: Influence of Demographic Information 70

6 Conclusions **81**

List of Figures

1-1 (A) Location resolution: Each circle represents a mobile-phone tower and the dashed lines correspond to a Voronoi diagram that roughly delimits the main reception zone of each tower, partitioning the space into individual cells. The blue and red solid lines show the trajectory of two mobile-phone users, illustrating how the call activity helps us to track individual motion. (B) Global analysis: Preliminary results showing the calling pattern between Ruanda and the rest of the world. 16

2-1 A snapshot of human movements at San Francisco Bay area 22

2-2 Location of towers in San Francisco Bay area 23

2-3 Population density data from LandScan, figures from left to right refer to San Francisco Bay area, Dominican Republic, and EU country, respectively. 27

2-4 Inter event time distribution $P(\Delta T)$ of calling activity. ΔT is the time elapsed between consecutive communication records for the same user. Different symbols indicate the measurements done over groups of users with different activity levels(num. of calls). Figure 2-4.A shows the unscaled version of Figure 2-4.B. 29

2-5 Probability density function $P(\Delta r)$ of travel distances obtained for the mobile phone records. The solid line indicates a truncated power law for which the parameters are provided in the text (see Eq. (2.3)). . . 30

2-6 Displacement distribution $P(\Delta r)$ for fixed inter event times ΔT_0 based on the mobile phone records. The cutoff of the distribution is set by the maximum distance users can travel for shorter interevent times, whereas for longer times the cutoff is given by the finite size of the studied area. 31

2-7 A. Radius of gyration $r_g(t)$ versus time for mobile phone users separated into two groups according to their final $r_g(T)$, where $T = 1$ month. B. The distribution $P(r_g)$ of the radius of gyration measured for the users, where $r_g(T)$ was measured after $T = 1$ month of observation. The solid line represents a similar truncated power-law fit (see Eq. (2.5)). 34

2-8 The Zipf plot showing the frequency of visiting different locations (loc.). The symbols correspond to users that have been observed to visit $n_L = 5, 10, 30$ and 50 different locations. Denoting with L, the rank of the location listed in the order of visit frequency, the data are well approximated by Eq. (2.6) (the solid line) 35

2-9 The probability density function $\Phi(x, y)$ of finding a mobile phone user in a location (x, y) in the user's intrinsic reference frame. B. After scaling each position with σ_x and σ_y, the resulting $\tilde{\Phi}(x/\sigma_x, y/\sigma_y)$ has approximately the same shape for each group. The two plots, from left to right were generated for 10000 users with: $0 \leq r_g \leq 3$, $20 \leq r_g \leq 30$. 37

2-10 The change in the shape of $\Phi(x, y)$ can be quantified by calculating the anisotropy ratio $S \equiv \sigma_y/\sigma_x$ as a function of r_g. Error bars represent the standard deviation. 38

2-11 A voronoi division of towers in part of San Francisco Bay area. The red points represent towers, the polygons that contain the red points are defined as areas around towers according to voronoi division. $R_g^i(t)$ is computed for each area by Eq. (2.7). 39

3-1 Schematic self organizing map. 42

3-2 SOM results, the feature vector consists of location of towers (Lat, Lon), radius of gyration (AvgRg), number of users (nbusers), and population density (popdensity). 44
3-3 SOM results, the feature vector consists radius of gyration (AvgRg), number of users (nbusers), and population density (popdensity). . . . 47
3-4 The histogram of Average R_g (left figures), Population density (middle figures) and Number of users (right figures) in Dominican republic (upper figures), San Francisco Bay area (middle figures), and the anonymous EU contry (lower figures). 48

4-1 Three representative distributions of radius of gyration. The blue dots are distribution of radius of gyration in log scale; The black line is the fitted function according to Eq. (4.1) 52
4-2 The composition and variance of 4 components. 64

5-1 P_1 versus Population Density. 68
5-2 Population density and Income distribution in San Francisco Bay area. The upper-left graph is population density distribution, at a resolution level of $/km^2$; The upper-right graph is median household income distribution; The lower-left graph is the distribution of mobile phone users (note: we use user's most visited location as proxy for his/her home location); The lower-right graph is the distribution of number of calls. 71
5-3 The estimated r_g v.s. the true r_g. 73
5-4 $P(r_g)$ in each group 76
5-5 $P(r_g)$ distribution in areas with mediate level of income and different population density 77
5-6 $P(r_g)$ distribution in areas with mediate population density and different levels of income 77

List of Tables

2.1 Raw data format of tower location 23

2.2 Raw data format of calling information 24

2.3 Modified data format of tower location 25

2.4 Modified data format of calling information 25

4.1 $P(r_g)$ curve fitting result of 24 autonomous regions. 55

4.2 Descriptive statistics for parameters alpha and beta. 63

4.3 Correlation Matrix for parameters alpha and beta. 63

4.4 Total Variance Explained by each component. 63

4.5 Composition of components. 64

4.6 PCA results. 65

5.1 Population density statistics for all the autonomous regions. 68

5.2 Regression result: regressing radius of gyration on income, population density, age and unemployment rate. 72

5.3 Number of towers in each group. Low population density: $0-2039/km^2$, Median population density: $2039-7285/km^2$, High population density: $\geq 7285/km^2$; Low income: $0-61100$\$, Median income: $61100-89440$\$, High income: ≥ 89440\$ 74

5.4 Number of users in each group. 75

5.5 Parameters value for each group. 76

5.6 The upper table is obtained by fixing population density at a certain level (i.e. Low, Mid and High correspond to the first, second and third number in the parenthesis respectively) and calculating the difference between values at different income levels (Mid-Low, High-Low) for each parameter. The lower table is obtained by fixing income at a certain level (i.e. Low, Mid and High correspond to the first, second and third number in the parenthesis respectively) and calculating the difference between values at different population densities (Mid-Low, High-Low) for each parameter. 78

5.7 A transformation of Table 5.5 by replacing a positive value in the table with "+", a negative value with "-" and zero with "-/+". 79

Chapter 1

Introduction

1.1 Motivation and Problem Statement

Mobile phones are becoming increasingly ubiquitous throughout large portions of the world. In industrialized countries mobile phone penetration is almost 100%, while in non industrialized countries they constitute a large emergent market which is receiving huge investments from mobile phone carriers [1, 2]. For billing purposes, each mobile phone provider regularly collects extensive data about the call volume, calling patterns, and the location of the cellular phones of their subscribers. In order for a mobile phone to place outgoing calls and receive incoming calls, it must periodically report its presence to nearby cell towers, thus registering its position in the geographical cell covered by the closest tower (Fig. 1A). In consequence, very detailed information on the spatiotemporal localization of billions of users is contained in the extensive call records of today's mobile phone carriers. These data constitute a huge opportunity to science; in particular, they provide information on human motion at a scale not available heretofore. Maps with statistical trajectories of large-scale human movements from different continents would have unprecedented applications in urban planning, traffic forecasting and epidemic prevention, as in any area involving human motion.

However, little research has aimed toward a derivation of human behavioral patterns from this collective data; it is our aim to build algorithms to gain insight into

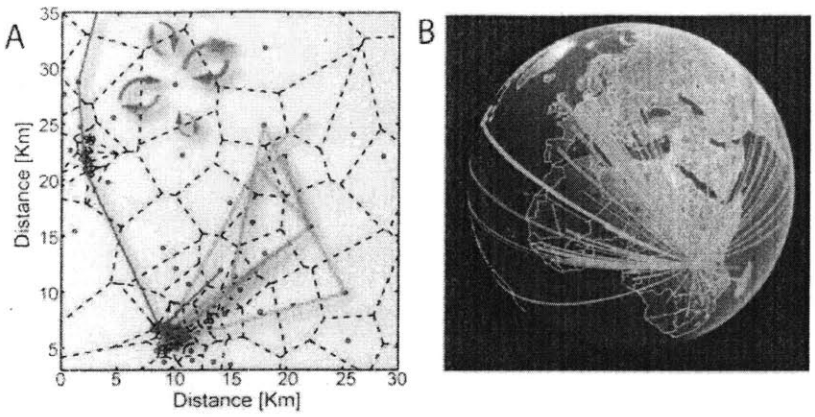


Figure 1-1: (A) Location resolution: Each circle represents a mobile-phone tower and the dashed lines correspond to a Voronoi diagram that roughly delimits the main reception zone of each tower, partitioning the space into individual cells. The blue and red solid lines show the trajectory of two mobile-phone users, illustrating how the call activity helps us to track individual motion. (B) Global analysis: Preliminary results showing the calling pattern between Ruanda and the rest of the world.

the interplay between outcomes of interest and movement patterns. By considering billions of data points of time and space from tens of millions of mobile phone subscribers in regions ranging from rural Dominican Republic to urban California, we can better understand the dynamics of these individuals, and the societies in which they live.

In a recent study [1], González showed that when analyzed with the appropriate techniques, mobile phone data offer the possibility of characterizing statistically human trajectories at a country scale. Those results established the basic elements for constructing realistic agent-based models in which the distribution of agents is proportional to the population density of a given region, and each agent has a characteristic trajectory size derived from an observed distribution. Such distributions must be derived from empirical data. Available studies have gathered data from industrialized countries [1 - 6] and have shown similarities in the calling patterns [4, 5] and travel distance distributions among them [1, 6, 9]. However, much less is known from the analysis of data coming from developing countries.

We expect that a thorough study of large data sets from developing countries will uncover important differences with respect to human mobility patterns, which may be rooted in particular economic constraints and, in some cases, in different cultural habits. Our main goal for this work is to make a cross-cultural study human mobility,

comparing the analysis of mobile phone data coming from industrialized countries and those coming from developing countries.

Our dataset includes the mobility and communication patterns extracted from cell phone logs of over 3,729,134 people around the world, spanning 24 autonomous regions in San Francisco bay area, a European country and Dominican Republic. In addition to characterize the human mobility patterns by a mathematical formula, we expect to observe how these patterns display differently across autonomous regions. The variations of human mobility patterns further imply the existence of essential underlying socio-economic factors that are of various levels and thereby contribute to dissimilar movement patterns. Finally, we concentrate on understanding the relationship between population density (plus income distribution for the available areas) and mobility patterns and expect to discover original results.

1.2 Thesis Outline

Chapter 2 introduces the general methods applied in processing raw data, characterizing calling activities and the calculation procedures of radius of gyration. The methodology in this chapter is identical to those in [1], but is applied to analyze another data set collected from San Francisco Bay area. Essentially distinctive mobility patterns are observed which cannot be fitted by a truncated power-law distribution. The differences in mobility patterns between San Francisco Bay area and the country analyzed in [1] motivate us to concentrate on searching for the underlying determinants.

To refine the comparative research on human movements across various regions, we divide the country into several autonomous areas. An autonomous area is an area of a country that has a degree of autonomy, or freedom from an external authority. Typically it is either geographically distinct from the country or is populated by a national minority or is an administrative region defined by the state or country. In general, it is arbitrary how to place the borders in the land to characterize trips, in this study we take the regions defined by each country for administrative purposes.

In Appendix A, several figures are presented, each of them represents the radius of gyration distribution of one autonomous region. By applying the same methods to the aggregated data sets recording subscribers' calling activities during one month in 24 autonomous regions, we observed that the quantified distribution are very different but can be essentially grouped into four distinct families. These curves are further analyzed in Chapter 5.

Additionally, we explored the role of some factors (such as population density, income, etc.) and how they potentially influence or even determine human movements. We start with investigating the inter-relationship between radius of gyration and population density, which is readily available from LandScan datasets [10]. The usage of Kohogen self organizing maps (SOM) certifies our conjecture - strong relationship exists between trip length distribution and population density distribution. The classification result from SOM is almost identical to the the classification according to geographical locations. The coincidence strongly implies that radius of gyration and population density are correlated with each other. Chaper 3 introduces the Kohogen map and the results we obtained for our data sets.

In Chapter 4, we present the method that are used to find the relationship between radius of gyration and population density. The distribution of radius of gyration can be approximated by a double exponential functions in log scale. Therefore, the trip length distributions are able to be characterized with four parameters. The extracted parameters provide us the opportunity to understand their relationship to population density. Principal Component Analysis (PCA) is used to disentangle correlations among the four factors, and further reduce the dimensionality into two geometrically orthogonal components that contribute most to the variance of our data sets.

In Chapter 5, regression analysis is presented with the goal to discover the relationship between these two parameters and population density. However, the result gives a very low R^2 if we use a simple regression model with population density and unit vector as regressors. To improve the performance of the model, we incorporate other factors including the normalized standard deviation of population density and another two explanatory variables derived from a trip length distribution model [11].

The enriched model gives a better R^2, but caution is needed when interpreting the relationship by using this model, because the number of parameters are not small enough compared to the number of sample points. Finally, a regression model with income data and population information as well as a sample collected at the resolution of tower level from San Francisco Bay area is analyzed. We find that people in rich and less populated areas tend to travel more frequently and make long distance trips. Finally, we grouped residents into nine category according to local population density and income level and calculated respectively their radius of gyration distribution. This enables us to analyze how parameters in the double exponential distribution are related to socio-economic factors, such as income and population density.

Chapter 6 concludes the thesis.

Chapter 2

Mobility Measures

The mobile phone records we analyzed include the one-month calling activities received by nearby communication towers in San Francisco Bay area, Dominican Republic and a European country (denoted as EU in what follows). These rich mobile data sets provide us unprecedented power to conduct a cross cultural analysis of human traveling behavior. We combine this information with the population density distribution at the resolution of 0.008333 decimal degrees. In particular, for San Francisco Bay area, the income distribution data is available at an equivalent spatial resolution, which enables us to analyze trip lengths with respect to population density and income distribution in Section 5.3. Section 2.1 presents the data formats and the methods we applied to obtain the refined data sets.

Understanding communication patterns from mobile phone records is important to characterize the humans movements. Communication patterns are known to be highly heterogeneous because some users rarely use the mobile phones while others make hundreds or even thousands of calls each month. Empirical and statistical methods are applied to investigate these data sets. Section 2.2 introduces how these methods are used in interpreting mobile phone records from San Francisco Bay area. At the end of this section, the measure of human mobility in small areas is presented. By extending the analysis to a micro-scale, namely, at the resolution level of communication towers, we expect to fully characterize the heterogeneity patterns existing among human's traveling behavior and expect to explore how such traveling behavior

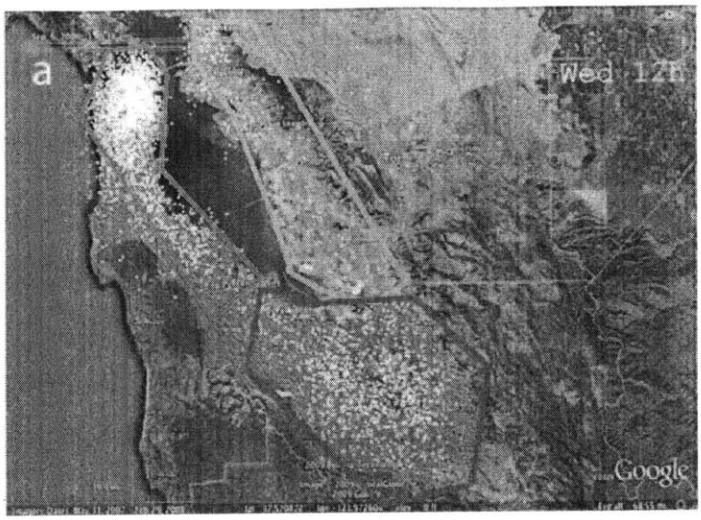


Figure 2-1: A snapshot of human movements at San Francisco Bay area

relates to regional socio-economic factors, such as income, population density and unemployment rate.

2.1 Data

This section introduces data sets in detail. The data sets we are analyzing include mobile phone records and population density information at $1km^2$ resolution from San Francisco Bay area, Dominican Republic, and EU. Moreover, rich information about household income, unemployment rate, median age are provided by Caliper company, which enable us to do finer analysis of San Francisco Bay area.

2.1.1 Mobile Phone Data

The data sets we analyzed are monthly calling activities received by phone service providers and the tower locations in San Francisco Bay area, Dominican Republic and EU. The one-month mobile phone data set includes 429,597 users and 374,221,753 phone calls which are recorded by 954 towers for San Francisco Bay area. 229,660 users, 37,636,984 calls are recorded by 184 towers in Dominican Republic and 3,069,877 users, 269,934,702 calls are recorded by 13,061 towers in EU.

Provided with rich mobile phone records, we can easily identify the time and

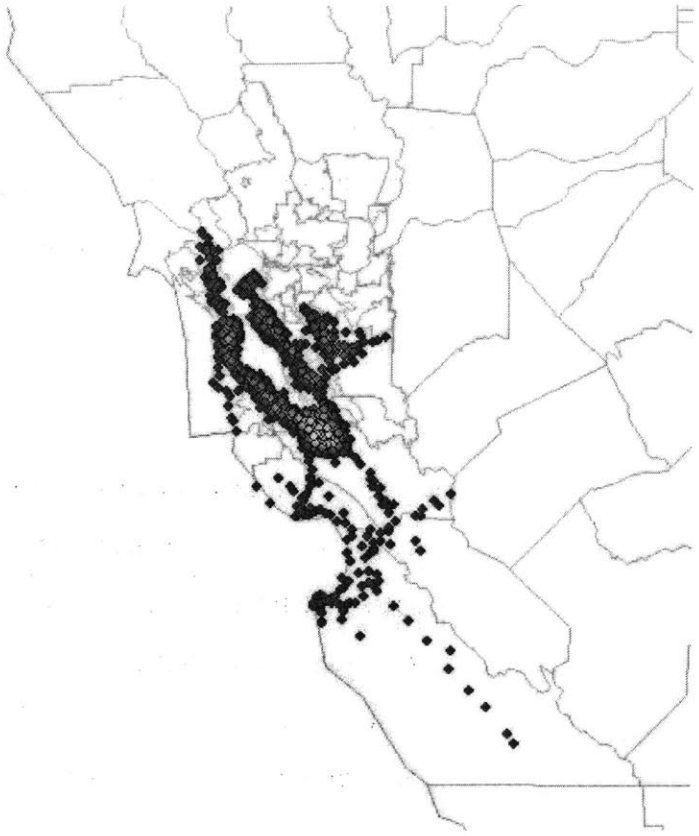


Figure 2-2: Location of towers in San Francisco Bay area

location of each calling activity and one step further, extract the characteristics or patterns of human movement. As an example, Fig. 2-1 and Fig. 2-2 displays a snapshot of calling activities and tower locations in San Francisco Bay area, respectively.

The format of tower location data is as shown in Table 2.1.The data sets of calling information has the format as in Table. 2.2. In Table 2.2, the item "caller" denotes the user who makes the call or sends the text message, "callee" stands for the one who receives it. To make sure that these private information are protected, all the

Tower ID	Latitude	Longitude
69	37.552	-122.049
70	37.554	-121.982
71	37.824	-122.233
⋮	⋮	⋮

Table 2.1: Raw data format of tower location

Caller	Callee	Day	Time	Tower	Mode
4082000002	4082000012	1	15:30:12	3041	1
4082000002	4082000015	1	8:22:00	3121	2
4082000003	4082000002	2	11:54:08	167	1
⋮	⋮	⋮	⋮	⋮	⋮

Table 2.2: Raw data format of calling information

users are translated into hash formats. The items "day" and "time" record the exact time of the phone activity, while "tower" is the ID of wireless tower that is serving the call or text message, usually it is also the nearest tower to the caller. The item "Mode" is simply used to distinguish the calls and text messages.

These raw data sets contain sufficient information about human activities if analyzed with appropriate techniques. The raw data should be transformed first because these calling information only gives for each call the caller and callee and the time it happens, but to understand human activities and extract movement trajectories we need to know the different coordinates visited at different calls. To analyze these data appropriately, caution is needed as:

1. The purpose of our research is to analyze human mobility patterns, therefore we need information about calling activities of each user for at least one month, as we stated before. The transformed data sets should contain a list of users and the associated calling activities with the corresponding locations sorted in time.

2. To facilitate our analysis, day and time should be in conjunction converted into a single number in seconds.

3. We are interested in the trip length distributions of people which are recorded by the mobile phone records. These records are serving as proxies for our analysis of traveling behavior, therefore we don't need to distinguish the type of activities and the item "mode" is not taken into account.

4. Latitude and Longitude are not the best coordinates for our analysis, because the sphere shaped coordinates may make the results inaccurate in Euclidian Distance calculation between towers. In the modified tower files, unprojected Greenwich data

Tower ID	X-coordinate	Y-coordinate
69	6.160676	4.921061
70	12.059370	5.158100
71	-10.071578	35.125419
⋮	⋮	⋮

Table 2.3: Modified data format of tower location

User1	Number of calls
4082000002	18
Time of call	Tower ID
398224	3041
398318	3119
489017	3272
⋮	⋮
User2	Number of calls
4082000003	63
327344	165
⋮	⋮

Table 2.4: Modified data format of calling information

are transformed to a projected Cartesian coordinate system.

The processed formats of raw data sets are shown in Table. 2.3 and 2.4.

2.1.2 Population Data

In this section we described LandScan, our density of population data sources. Most of the descriptions presented here can also be found in the LandScan manual book [10]. Using an innovative approach with Geographic Information System and Remote Sensing, ORNL's(Oak Ridge National Laboratory) LandScanTM is the community standard for global population distribution [10]. At approximately $1km$ resolution($30'' \times 30''$), LandScan is the finest resolution global population distribution data available and represents an ambient population (average over 24 hours). LandScan population distribution models are tailored to match the data conditions and geographical nature of each individual country and region.

Format and extent:

The data is distributed in both an ESRI grid format and an ESRI binary raster format. The dataset has 20,880 rows and 43,200 columns covering North 84 degrees to South 90 degrees and West 180 degrees to East 180 degrees.

Data values:

The values of the cells are integer population counts representing an average, or ambient, population distribution. An ambient population integrates diurnal movements and collective travel habits into a single measure. Since natural or man made emergencies may occur at any time of the day, the goal of the LandScan model is to develop a population distribution surface in totality, not just the locations of where people sleep. Because of this ambient nature, care should be taken with direct comparisons of LandScan data with other population distribution surfaces.

Resolution and Coordinate System:

The dataset has a spatial resolution of 30 arc-seconds and is output in a geographical coordinate system - World Geodetic System (WGS) 84 datum. The 30 arc-second cell, or 0.008333333 decimal degrees, represents approximately $1km^2$ near the equator.

Since the data is in a spherical coordinate system, cell width decreases in a relationship that varies with the cosine of the latitude of the cell. Thus a cell at 60 degrees latitude would have a width that is half that of a cell at the equator (cos60 = 0.5). The height of the cells does not vary.

The values of the cells are integer population counts, not population density, since the cells vary in size. Population counts are normalized to sum to each sub-national administrative unit estimate. Also prior to all spatial analysis, we should ensure that extents are set to an exact multiple of the cell size (for example 35.50, 35.0) to avoid "shifting" of the dataset.

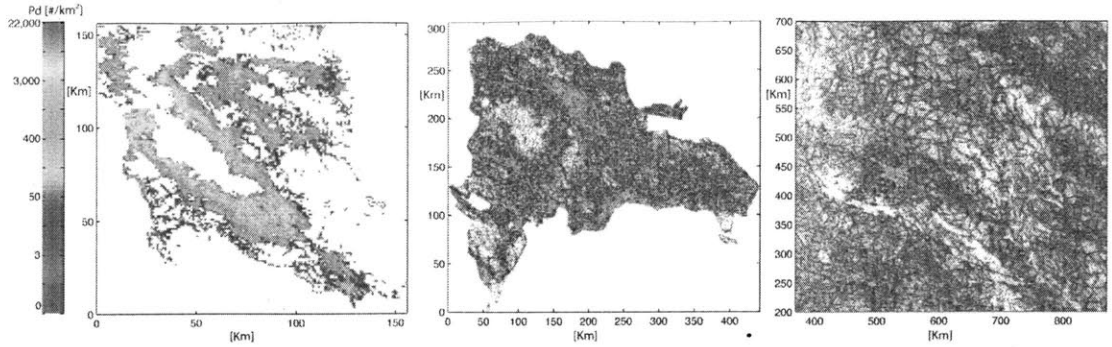


Figure 2-3: Population density data from LandScan, figures from left to right refer to San Francisco Bay area, Dominican Republic, and EU country, respectively.

2.1.3 Social-economic Data

The median family income data, unemployment rate and median age are provided by TransCAD software package [41] for San Francisco Bay area. In TransCAD, San Francisco Bay area is divided into several polygons, for each polygon, the associated income distribution of 2007 is available hence the median family income can be computed. Moreover, the rich data source for San Francisco Bay area also contains other useful information, such as traveling mode quantile distribution, male/female ratio, education level, etc.

2.2 Calculation of Population density

With the permission of ORNL for educational research, 2009 population counts file and global GRID-formatted file are downloaded from the LandScan [10].

Using ESRI ArcGIS with the Spatial Analyst extension, the database file containing population density distribution can be read, displayed and analyzed. The data are referenced by latitude/longitude (WGS84) coordinates, so the selected area should firstly be projected into WGS84 plain in order to be analyzed.

It is worth emphasizing that since the data is in a spherical coordinate system, cell width decreases in a relationship that varies with the cosine of the latitude of the cell. Therefore, projecting the data in a raster format to a different coordinate system will result in a re-sampling of the data and the integrity of normalized population

counts will be compromised.

The computation of population density for each cell is very simple, using Eq. (2.1):

$$D_i = C_i / s_i \tag{2.1}$$

Where D_i denotes the population density of cell i, C_i represents the population counts of cell i while s_i is its area. Once we have the population density of all the cells, we are able to calculate the distribution of the population for a particular area, the average population density as well as other quantities of interest.

2.3 Pattern measures

2.3.1 Characterizing Individual Calling Activity

Communication patterns are known to be highly heterogeneous: some users rarely use mobile phone while others make hundreds or even thousands of calls each month. To characterize the dynamics of individual communication activity, we grouped users based on their total number of calls. For each user we measured the probability that the time interval between consecutive calls is ΔT. Fig. 2-4.A shows that users with less activities tend to have longer waiting times between consecutive calls. By rescaling the axis with the average inter event time $\Delta T_a = 8.2$ hours as $\Delta T_a P(\Delta T)$ and $\Delta T / \Delta T_a$, we get Fig. 2-4.B. Hence the measured inter-event time distribution can be approximated by the expression $P(\Delta T) = 1/\Delta T_a F(\Delta T/\Delta T_a)$, where $F(x)$ is independent of the average activity level of the population. In addition, By aggregating individual calling activities in five groups with different total number of calls, we obtain the aggregated $P(\Delta T)$ curve, and this curve can be fitted by Eq. (2.2) with a R^2 0.9936.

$$P(\Delta T) = (\Delta T)^{-\alpha} exp(-\Delta T/\tau_c) \tag{2.2}$$

Where the power-law exponent $\alpha = 1.405 \pm 0.008$ (with 95% confidence bounds)

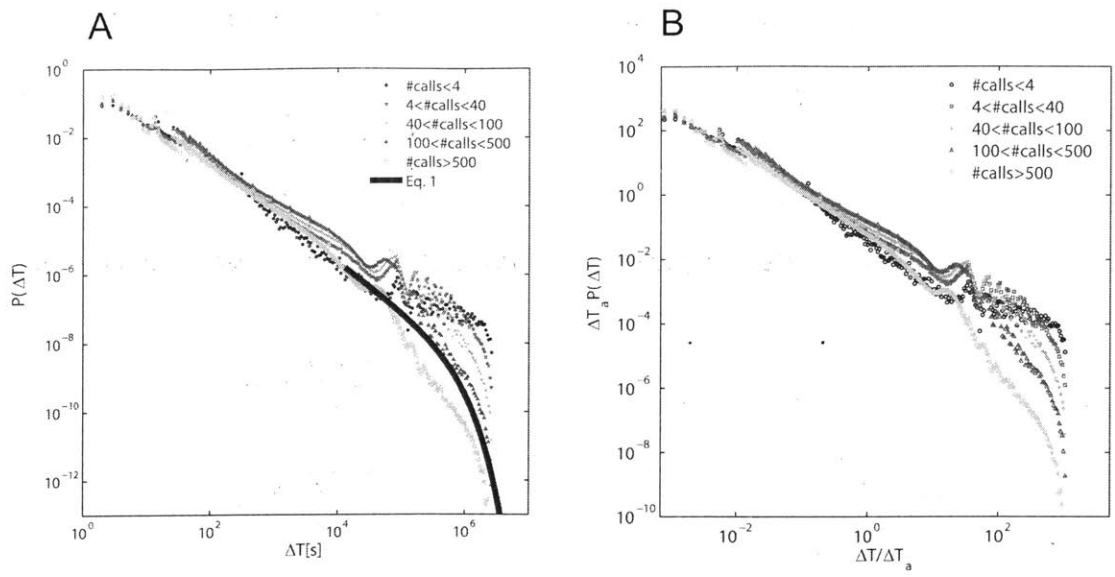


Figure 2-4: Inter event time distribution $P(\Delta T)$ of calling activity. ΔT is the time elapsed between consecutive communication records for the same user. Different symbols indicate the measurements done over groups of users with different activity levels(num. of calls). Figure 2-4.A shows the unscaled version of Figure 2-4.B.

is followed by an exponential cutoff of $\tau_c = 113$ hours . Eq. (2.2) is shown as a solid line in Fig. 2-4.

In Fig. 2-4.A. we observe that although Eq. (2.2) best fits the aggregated $P(\Delta T)$ curve, it performs poorly when applied to groups of users with different total number of calls. In contrast to the result from [3], where all the curves associated with different groups converge into a single curve denoted by Eq. (2.2). The divergent patterns of $P(\Delta T)$ for different groups of users imply the heterogeneity in users' calling frequencies.

2.3.2 Observations at A Fixed Inter-event Time

To explore the statistical properties of the population's displacement distributions, we measured the distance between user's positions at consecutive calls, capturing 373,792,156 displacements for the mobile phone records (See Fig. 2-5). We found that the distribution of displacements over all users is well approximated by a truncated

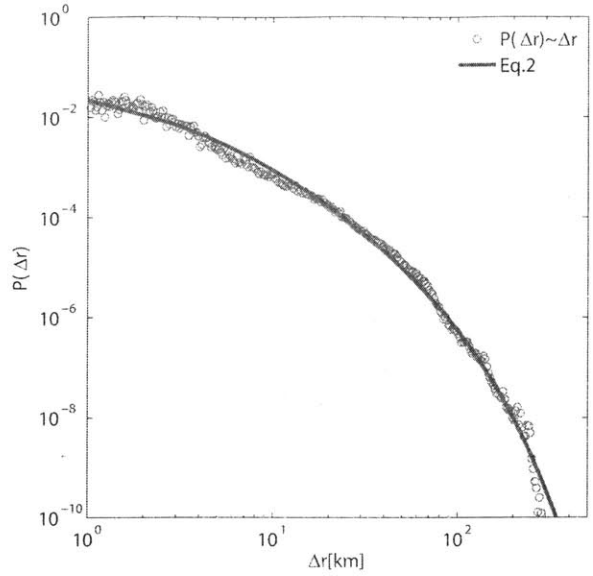


Figure 2-5: Probability density function $P(\Delta r)$ of travel distances obtained for the mobile phone records. The solid line indicates a truncated power law for which the parameters are provided in the text (see Eq. (2.3)).

power-law Eq. (2.3):

$$P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} exp(-\Delta r/\kappa) \tag{2.3}$$

with exponent $\beta = 2.618 \pm 0.027$ (with 95% confidence bounds), $\Delta r_0 = 3.536$ km and cutoff values $\kappa = 44.11$ km . The associated R^2 value is 0.9947.

Given the widely varying distribution of the inter event times between two calls (an therefore the localization data), we need to investigate if the observed displacement statistics are affected by this sampling heterogeneity. Using the mobile phone records, we calculated the displacement distribution $P(\Delta r)$ for consecutive calls separated by a time $\Delta T \pm 0.05\Delta T_0$, where ΔT_0 ranged from 20 minutes to one day. For $\Delta T_0 \leq 2$ hours , the observed displacements are bounded by the maximum distance that users can travel in the ΔT_0 time interval. For $\Delta T_0 \geq 4$ hours we already observe $\Delta r_{max} = 300$ km, which corresponds to the largest displacement we could possible observe given the area under study. We observe that the resulting $P(\Delta r)$ distributions for different ΔT_0 cannot be approximated by a truncated power-law with the same value of exponent $\beta = 2.618$ (See Fig. 2-6), suggesting that the sampling heterogeneity does influence

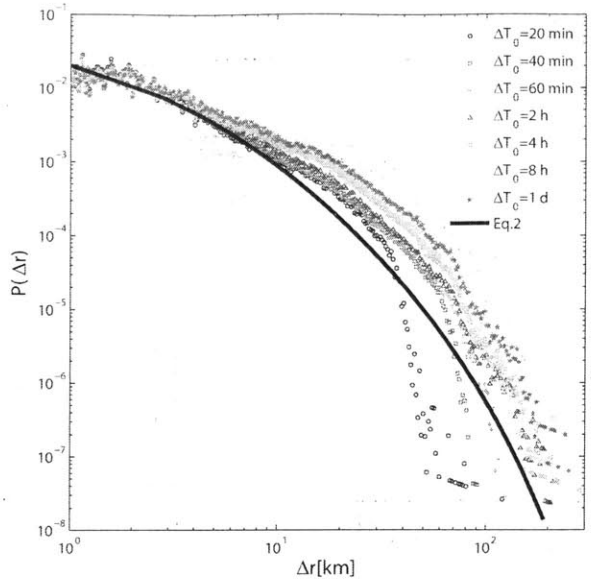


Figure 2-6: Displacement distribution $P(\Delta r)$ for fixed inter event times ΔT_0 based on the mobile phone records. The cutoff of the distribution is set by the maximum distance users can travel for shorter interevent times, whereas for longer times the cutoff is given by the finite size of the studied area.

the observed displacement statistics. The resulted divergent displacement patterns due to varying distribution of the inter event times between two calls reflect the heterogeneous mobility patterns in San Francisco Bay area.

2.3.3 The Radius of Gyration Distribution

To compare different users' trajectories we need to study them in a common reference frame. Inspired by the mechanics of rigid bodies, we assign each user to an intrinsic reference frame calculated a posteriori from a user's trajectory. We can think of the number times a user visited a given location as the mass associated with that particular position.

The intrinsic reference frame for individual trajectories can be calculated as below: Denoting a user's trajectory with a set of locations

$$(x_1, y_1), (x_2, y_2), ..., (x_{n_c}, y_{n_c})$$

where n_c is the number of positions available for the user. An object's moment of inertia is given by the average spread of an object's mass from a given axis. A two

dimensional object can be characterized by a 2×2 matrix known as the tensor of inertia

$$I = \begin{pmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{pmatrix}$$

We can calculate the inertia tensor for user's trajectory by using the standard physical formulas

$$I_{xx} = \sum_{i=1}^{n_c} y_i^2$$

$$I_{yy} = \sum_{i=1}^{n_c} x_i^2$$

$$I_{xy} = I_{yx} = -\sum_{i=1}^{n_c} x_i y_i$$

Since the tensor I is symmetric, it is possible to find a set of coordinates in which I will be diagonal. These coordinates are known as the tensor's principal axes $(\hat{e}_1, \hat{e}_2)$. In this set of coordinates I takes the form

$$I_D = \begin{pmatrix} I_1 & \\ & I_2 \end{pmatrix}$$

Where I_1 and I_2 are the principal moments of inertia. They also correspond to the eigenvalues of I and can be calculated from the original set of points as

$$I_1 = \frac{1}{2}(I_{xx} + I_{yy}) - \frac{1}{2}\mu$$

$$I_2 = \frac{1}{2}(I_{xx} + I_{yy}) + \frac{1}{2}\mu$$

with

$$\mu = \sqrt{4I_{xy}I_{yx} + I_{xx}^2 - 2I_{xx}I_{yy} + I_{yy}^2}$$

The corresponding eigenvectors determine the principal axes ($\hat{e}_1$ and $\hat{e}_2$), representing the symmetry axes of a given trajectory.

Since different users' principal axes $\hat{e}_1$ and $\hat{e}_2$ are different, to make a better evaluation, we transform each user's principal axes $(\hat{e}_1, \hat{e}_2)$ to a common intrinsic

reference frame $(\hat{e}_x, \hat{e}_y)$ calculating the angle between the axes $\hat{e}_x$ and $\hat{e}_1$, as

$$cos(\theta) = \frac{\vec{v}_1 \cdot \hat{e}_x}{|\vec{v}_1|}$$

Where $\vec{v}_1$, is the eigenvector associated with eigenvalue I_1

$$\vec{v}_1 = \begin{bmatrix} -\frac{I_{xy}}{1/2I_{xx}-1/2I_{yy}+1/2\mu} \\ 1 \end{bmatrix}$$

resulting in

$$cos(\theta) = -I_{xy}(1/2I_{xx} - 1/2I_{yy} + 1/2\mu)^{-1} \frac{1}{\sqrt{1 + \frac{I_{xy}^2}{(1/2I_{xx}-1/2I_{yy}+1/2\mu)^2}}}$$

After rotation by θ, we impose a conditional rotation of 180° such that the most frequent position lays always in $x > 0$. Using these transformed data, the radius of gyration is defined as

$$r_g^a(t) = \sqrt{\frac{1}{n_c^a(t)} \sum_{i=1}^{n_c^a} (\vec{r}_i^a - \vec{r}_{cm}^a)^2} \tag{2.4}$$

to characterize the linear size occupied by each user's trajectory up to time t. Where $\vec{r}_i^a$ represents the $i = 1, ..., n_c^a(t)$ positions recorded for user a and $\vec{r}_{cm}^a = 1/n_c^a(t) \sum_{i=1}^{n_c^a} \vec{r}_i^a$ is the center of mass of the trajectory.

To show that the observed distribution in Fig. 2-5 can be explained as a population-based heterogeneity, corresponding to the inherent differences between individuals, according to Eq. (2.4) we calculated the radius of gyration $r_g^a(t)$ for each user, interpreted as the characteristic distance traveled by user a when observed up to time t(Fig. 2-7.A). We measured the time dependence of the radius of gyration for users whose radius of gyration would be considered small ($r_g(T) \leq 3$ km), or large (10 km$\leq r_g(T) \leq 20$ km) at the end of our observation period $T = 1$ month. The result indicates that the time dependence of the average radius of gyration of mobile phone users display periodical fluctuation with a cycle of about 24 hours, which can be interpreted as the human's daily schedule pattern. The curvature also indicates

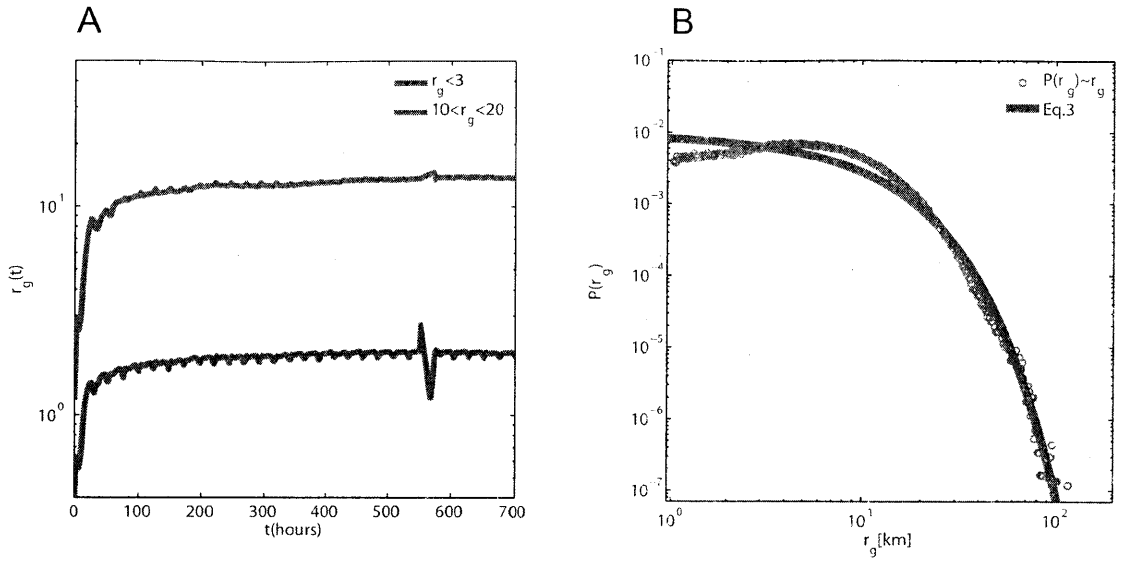


Figure 2-7: A. Radius of gyration $r_g(t)$ versus time for mobile phone users separated into two groups according to their final $r_g(T)$, where $T = 1$ month. B. The distribution $P(r_g)$ of the radius of gyration measured for the users, where $r_g(T)$ was measured after $T = 1$ month of observation. The solid line represents a similar truncated power-law fit (see Eq. (2.5)).

that the radius of gyration $r_g(t)$ versus time for mobile phone users is a saturation process. In addition, it is worth noticing that the prominent fluctuation at $t = 550$ hours is due to the absence of mobile phone data.

Moreover, we determined the radius of gyration distribution $P(r_g)$ by calculating r_g for all users in the record. Using the truncated power-law to fit the distribution, we obtained (Fig. 2-8.B):

$$P(r_g) = (r_g + r_g^0)^{-\beta} exp(-r_g/\kappa) \tag{2.5}$$

with $r_g = 67.43$ km, $\beta_r = 1.112 \pm 0.218$(with 95% confidence bounds), $\kappa = 9.452$ km.

In Fig. 2-7.B, we observe that $P(r_g)$ exhibits a peak around $r_g = 9km$, which cannot be fitted by truncated power-law. This pattern coincides with the fact that people living in the bay area often work in urban San Francisco. The R-square of this curve-fitting using Eq. (2.5) is below 90% which indicates truncated power-law cannot well describe the distribution of radius of gyration in San Francisco Bay area. This motivates our study of finding other functions that can best approximate the $P(r_g)$. I will state later in this thesis, a double exponential decaying function can be

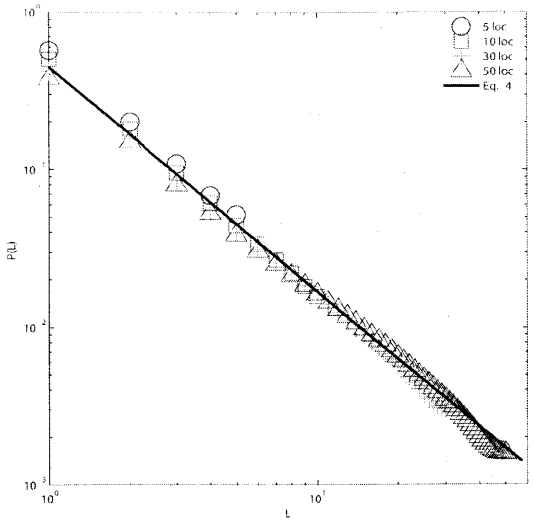


Figure 2-8: The Zipf plot showing the frequency of visiting different locations (loc.). The symbols correspond to users that have been observed to visit $n_L = 5, 10, 30$ and 50 different locations. Denoting with L, the rank of the location listed in the order of visit frequency, the data are well approximated by Eq. (2.6) (the solid line)

used to approximate $P(r_g)$ in log-scale.

2.3.4 The Frequency of Visiting Different Locations

We ranked each location according to the number of times an individual was recorded in its vicinity to explore the probability of individuals return to the same location. For example, $L = 1$ represents the most-visited location for the selected individual, $L = 2$ represents the second-most-visited location, and so on. We divide individuals into four group on the basis of the total number of location visited and for every group calculated the average visiting probability associated with each location. We find that the probability of finding a user at a location with a given rank L is well approximated by Eq. (2.6):

$$P(L) = \lambda L^{-\mu} \tag{2.6}$$

where $\mu = 1.442 \pm 0.006$, $\lambda = 0.45$, independent of the number of locations visited by the user, as the solid line shown in the Fig. 2-8.

We can clearly see that about 40% of the time individuals are found at their first

two preferred locations. Therefore, individuals tend to devote most of time to a few locations with high ranks while spending the remaining time in other locations with diminished regularities according to the rank. The conformity of the frequency of visiting different locations among four groups supports the conclusion that individuals' daily travel patterns reflect high degree of regularity.

2.3.5 Modeling Human Mobility Patterns Using Spatial Density Function

Individuals live and travel in different areas, yet as shown in Fig. 2-8, each user can be assigned to a well defined area, where she or he can be found most of the time. We quantitatively modeled human mobility pattern in light of the spatial density function $\Phi_a(x, y)$, which provides the probability of finding an individual a in a given position (x, y). By diagonalizing each trajectory's inertia tensor using formula mentioned in section. 2.3.3, we can compare the trajectories of different users in the user's intrinsic reference frame. The probability of finding a user in a given position is plotted in a contour graph (Fig. 2-9). As shown in the figure, the spatial anisotropy is prominent for the $\Phi(x, y)$ function in this intrinsic reference frame. Notice that the left contour graph in Fig. 2-9 is generated with $0 \leq r_g \leq 3$, while the right one is generated with $20 \leq r_g \leq 30$, it is clear that the larger an individual's r_g, the more pronounced is this anisotropy.

To quantify the degree of anisotropy associated with individuals with different r_g, we defined the anisotropy ratio $S \equiv \sigma_y/\sigma_x$, where

$$\sigma_x = \sqrt{\frac{1}{n_c}\sum_{i=1}^{n_c}(x_i - x_{cm})^2}$$

$$\sigma_y = \sqrt{\frac{1}{n_c}\sum_{i=1}^{n_c}(y_i - y_{cm})^2}$$

represent the standard deviation of the trajectory measured in the user's intrinsic reference frame. By calculating σ_y/σ_x according to a serial values of r_g, we found

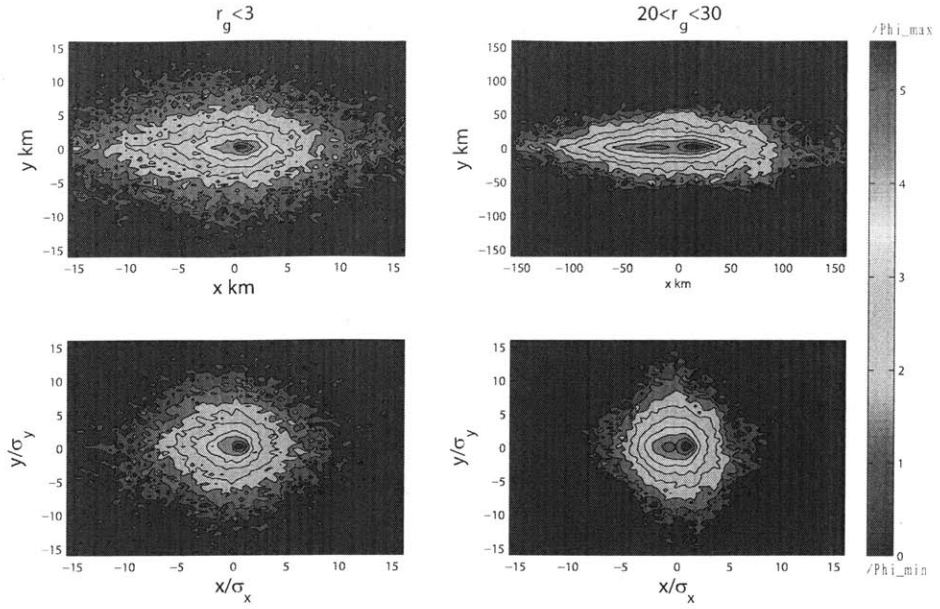


Figure 2-9: The probability density function $\Phi(x, y)$ of finding a mobile phone user in a location (x, y) in the user's intrinsic reference frame. B. After scaling each position with σ_x and σ_y, the resulting $\widetilde{\Phi}(x/\sigma_x, y/\sigma_y)$ has approximately the same shape for each group. The two plots, from left to right were generated for 10000 users with: $0 \leq r_g \leq 3$, $20 \leq r_g \leq 30$.

that S does not decrease monotonically with r_g (See Fig. 2-10).

By rescaling each user's trajectory with its respective σ_x and σ_y, we can better compare the trajectories of different users with individual anisotropy removed. As shown in the bottom figures in Fig. 2-9, the recalled $\widetilde{\Phi}(x/\sigma_x, y/\sigma_y)$ distribution is similar for the two groups with considerably different r_g. Therefore, in the absence of the dependence exists between the anisotropy and r_g, all individuals seem to follow the same universal $\widetilde{\Phi}(x, y)$ spatial probability distribution. Using the predicted anisotropic rescaling, combined with the density function $\widetilde{\Phi}(x, y)$, we can obtain the likelihood of finding a user in any locations.

It has been shown that human trajectories exhibit a high degree of temporal and spatial regularity, each individual being characterized by a time independent characteristic length scale and a significant probability to return to a few highly frequented locations. After correcting for differences in travel distances and the inherent anisotropy of each trajectory, the individual travel patterns collapse into a single spatial probability distribution, indicating that despite the diversity of their travel history, humans follow simple reproducible patterns.

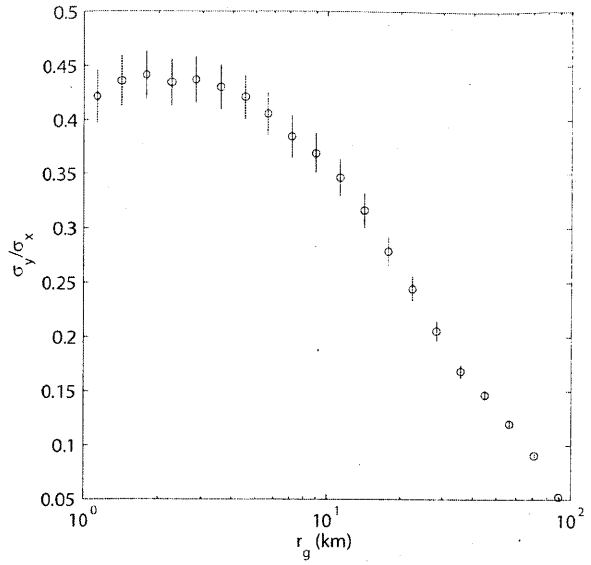


Figure 2-10: The change in the shape of $\Phi(x, y)$ can be quantified by calculating the anisotropy ratio $S \equiv \sigma_y/\sigma_x$ as a function of r_g. Error bars represent the standard deviation.

2.3.6 Mobility Characterization around Towers

Section 2.3.3 introduces the concept of individual radius of gyration. By applying the method detailed in that section, we are able to obtain the radius of gyration of each individual up to time t. Hence, the trip lengths distribution of an interested area can be characterized as the aggregated distribution of the radius of gyration of all the individuals making at least one call in that area (As Fig. 2-7.B shows). However, we are also interested in the individual mobility patterns from a given tower area, since resolution at the tower-level would provide us with more elaborate user mobility information (See Fig. 2-11).

Let $R_g^i(t)$ be the radius of gyration of tower i up to time t, then $R_i(t)$ is calculated according to Eq. (2.7).

$$R_g^i(t) = \frac{1}{|A(i,t)|} \sum_{a \in A(i,t)} r_g^a(t) \tag{2.7}$$

where $A(i, t)$ denotes the set including all the users who most frequently make calls around tower i up to time t, that is tower i recorded most calling activities before t for user i. $r_g^a(t)$ denotes the radius of gyration of user i, computed by Eq. (2.4). $|A(i, t)|$

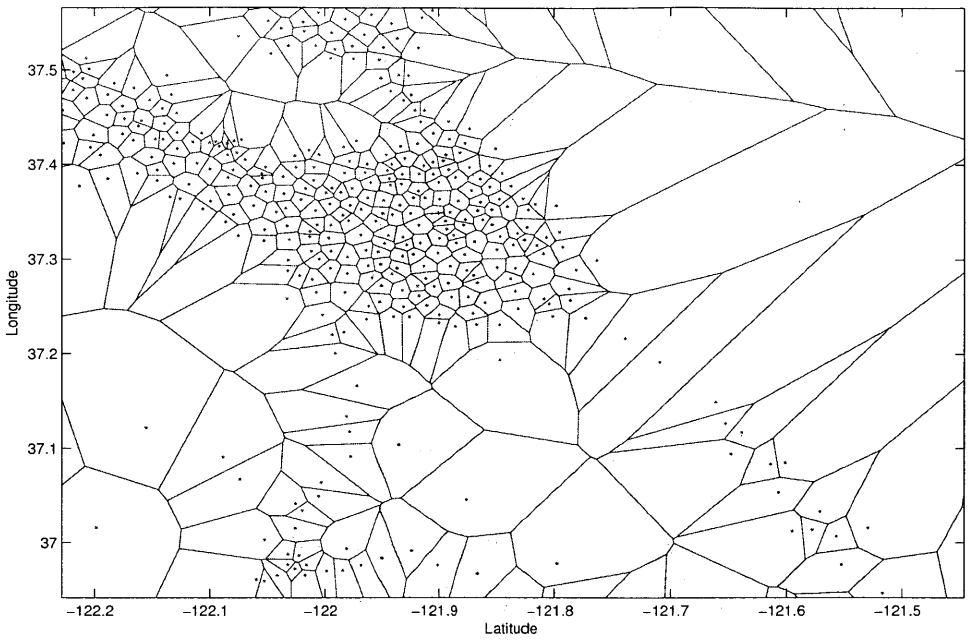


Figure 2-11: A voronoi division of towers in part of San Francisco Bay area. The red points represent towers, the polygons that contain the red points are defined as areas around towers according to voronoi division. $R_g^i(t)$ is computed for each area by Eq. (2.7).

denotes the number of elements in set $A(i,t)$. Introducing the measure of 2.7 enables us to characterize traveling behavior at a micro-scale, namely, around communication towers. Based on the micro-characterization at the tower level, we are interested in analyzing people's heterogeneous traveling behaviors and how they might potentially relate to regional income, unemployment rate and other socio-economic factors.

Chapter 3

Identify Inter-relationship from Kohonen Map

Kohonen map is widely known to identify correlations among multiple factors. In this section, we introduce briefly Kohoen map and how it can be applied to analyze our data. The result of Kohonen map clearly visualize the inter-relationship between radius of gyration and population density distributions.

3.1 Kohoen Map and Multidimensional Clustering

The Kohen Map, also named Self Organizing Map is type of artificial neural network that is trained under unsupervised learning to produce a low-dimensional and discretized representation (usually called a map) of the input sample space. Kohen Map preserves the topological properties of the input space which makes it useful for visualizing low-dimensional views of high-dimensional sequence. (Note, what follows in this section is copied from Wikipedia with slight modifications.)

A self-organizing map consists of components called nodes or neurons, which formulates an $m \times p$ bi-dimensional surface. Each node has a hexagonal shape surrounding by six neighbors (Fig. 3-1). Basically, the algorithm defines the nodes number as $m \times p = \sqrt{k}$, where k is the population to be clustered. Associated with each node is a weight vector of the same dimension as the input data vectors and a position in the

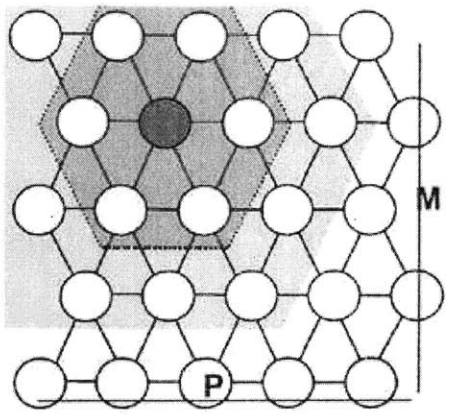


Figure 3-1: Schematic self organizing map.

map space. The self-organizing map describes a mapping from a higher dimensional input space to a lower dimensional map space. The procedure for placing a vector from data space onto the map is to find the node with the closest weight vector to the vector taken from data space and to assign the map coordinates of this node to our vector.

The training utilizes competitive learning. When a training example is fed to the network, its Euclidean distance to all weight vectors is computed. The neuron with weight vector most similar to the input is called the winner. The weights of the winner and neurons close to it in the SOM lattice are adjusted towards the input vector. The magnitude of the change decreases with time and with distance from the winner. The update formula for a neuron with weight vector $Wv(t)$ is Equ. (3.1)

$$Wv(t+1) = Wv(t) + \Theta(v,t)\alpha(t)(D(t) - Wv(t)) \tag{3.1}$$

where $\alpha(t)$ is a monotonically decreasing learning coefficient and $D(t)$ is the input vector. The neighborhood function $\Theta(v,t)$ depends on the lattice distance between the winner and neuron v. In the simplest form it is one for all neurons close enough to winner and zero for others, but a gaussian function is a common choice, too. Regardless of the functional form, the neighborhood function shrinks with time. At the beginning when the neighborhood is broad, the self-organizing takes place on the global scale. When the neighborhood has shrunk to just a couple of neurons the

weights are converging to local estimates.

This process is repeated for each input vector, the network winds up associating output nodes with groups or patterns in the input data set.

It is also common to use the U-Matrix to characterize the distance between neighbor nodes. The U-Matrix value of a particular node is the average distance between the node and its closest neighbors. In a hexagonal grid, we might consider six neighbors. To some extent, the visualization of distance between neighbor nodes can give us a roughly classification.

3.2 Identification of Relationship Between Radius of Gyration and Population Density from Kohoen Map

Intuitively, there must exist at least vague inter-relationship between radius of gyration and population density, since people travel differently in those areas with dense population versus areas with lower population density. Radius of gyration is a good indicator of human's traveling behavior in terms of characterizing trip length distributions [3], hence what we are interested here is the inter-relationship between radius of gyration and population density.

The results in the following sections, where we applied SOM learning algorithm to visualize mobile phone records and population data, show a strong connection between these two factors. The perfect match between the classification obtained from Kohoen map and the actual geological separation confirms our intuition.

3.2.1 Training Sample and Vector Construction

A typical issue associated with a statistical learning algorithm is the number of observations in the interested sample. At the resolution level of autonomous areas, only 24 observations are available, which is far from enough to give us a reliable result on the statistical relationship between radius of gyration and population information. To

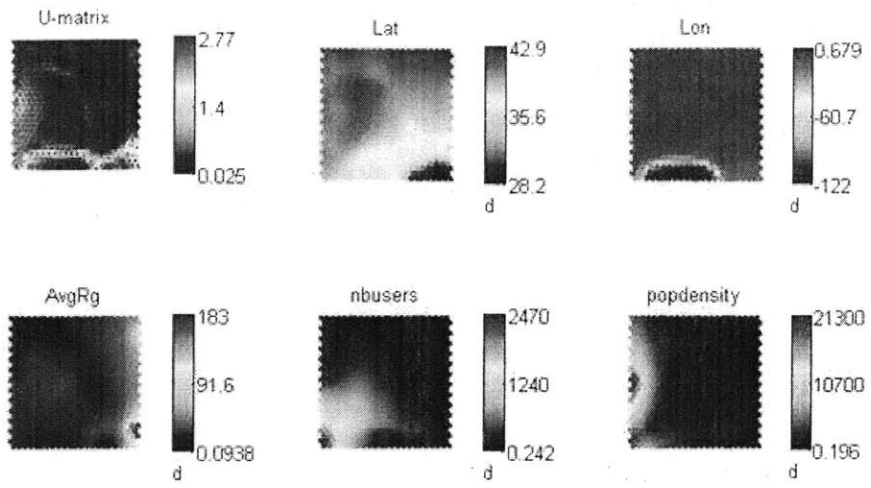


Figure 3-2: SOM results, the feature vector consists of location of towers (Lat, Lon), radius of gyration (AvgRg), number of users (nbusers), and population density (popdensity).

improve the effectiveness of SOM algorithm, the sample points should be sufficiently large. Therefore, we use a higher resolution at the tower levels, where 14199 towers (sample points) are ready to be analyzed. Moreover, at this resolution, every tower is associated with 262.6 users on average, which is enough to reflect the aggregated mobility patterns and ensure the accuracy of each point in the sample.

In SOM, each node is associated with a weight vector $V = (v(1), v(2), ..., v(n))$. $v(i)_{i=1}^{n}$ are features in the input data. In our experiment, candidates for these features include: The location of towers (latitude and longitude), radius of gyration (as defined in Eq. (2.7)), number of users, and population density.

Notice that the radius of gyration is the average radius of gyration of all the users around the tower, as defined in Eq. (2.7)). The population density is the average density in a polygon containing the interested tower, in which the polygons are obtained according to voronoi division.

3.2.2 SOM with Geographical Data

In this experiment, we include all the candidates in the vector, that is, the vector contains location of towers, radius of gyration, number of users, and population density. The result is shown in Fig. 3-2:

At the first glance of the U-matrix in Fig. 3.2, we observe that a green curve at the bottom divides the whole map into three parts (clusters). Coincidentally we have three different regions which are Dominican Republic, Bay area and Spain, thereby we need to explore whether the three parts classified by U-matrix actually represent the regions. Namely, is the U-matrix classification reasonable or just the outcome of a data mining method without any worthful meaning?

Obviously, if we separate the points completely according to the location of towers, then since the towers in different regions are not overlapped with each other, the result should be grouping all the towers belonged to a same region into one cluster. This classification is meaningful because it coincides with the geographical separation among these three regions. Therefore, if we can prove that the classification by U-matrix is the same as the classification by the location of towers, then we can contend that U-matrix classification is reasonable.

Take a look at the "Lat" and "Lon" plots in Fig. 3-2. In the construction of vector, we know these two features should rely on information about the location of towers. In "Lon", we observe that a shape contrast between the bottom-left part with blue color and the rest. According to the color bar, blue color denotes the points with longitude around $-122°$, which coincides with the longitude range of Bay area $(-122.6°, -120.9°)$. The longitude range of Dominican Republic is $(-71.7°, -68.4°)$, and the longitude range of Spain is $(-18.0°, 4.3°)$, which cannot be visualized as blue color. In "Lat", the bottom-right part is blue, representing the nodes with latitude around 28.2. However, the latitudes of towers range from 18.2° to 19.8° for Dominican Republic, from 36.0° to 38.2° for Bay area, and from 27.7° to 43.7° for Spain, no such regions coincides with the latitude level shown in the plot. The reason is that in SOM, the features of every sample point fed to the network are broadcasted to the neighborhood. Although the effects are diminishing the longer the Euclidian distance is, if the surrounding sample points' features are strong enough (here, by "strong" we mean the features are shapely different to the neighborhoods) and if the number of sample points surrounding the interested area are sufficiently large, the features' values of interested area would be changed but still maintaining

the topological pattern. This is exactly the case in "Lat": the bottom-right part is surrounded by a number of nodes with high latitude (about 38°, yellow color), which broadcast their features and increase the values associated with the nodes in the bottom-right part. Since Dominican Republic has the smallest latitude, we can assert that the bottom-right part with blue color should represent this region.

Therefore, by according to the "Lat" and "Lon" plots, we are able to identify the three regions, with the bottom-right be Dominican Republic, the bottom-left be Bay area, and the rest be Spain. Moreover, taking into consideration the number of towers, we can get the same results: Spain has the largest number of towers, followed by Bay area, and Dominican Republic has the smallest number of towers. Hence, Spain should occupy most of the nodes in the Kohoen map and Dominican Republic should be associated with least nodes, which are in accordance with our classification.

Coincidentally, the classification according to U-matrix has a perfect match with the classification according to the combination of "Lat" and "Lon". This proves that the classification by U-matrix is meaningful and coincides with the geographical separation. However, since by purely relying on the location of towers, we can obtain a classification that is in accordance with the reality, we are wondering that if the weights of the location of towers are much more than the weights associated with the other three features, then the classification by U-matrix gives us no information about the relationship between radius of gyration and population density, but only a geographical partition.

Therefore, by now we are able to propose three hypotheses:

1. U-matrix is dominated by the location of towers.

2. U-matrix is dominated by the relationship between radius of gyration and population density.

3. U-matrix is generated by the combination of all the features.

We expect to exclude the possibility that the location of towers plays a predominant role in the U-matrix classification, and finally reach hypothesis 2.

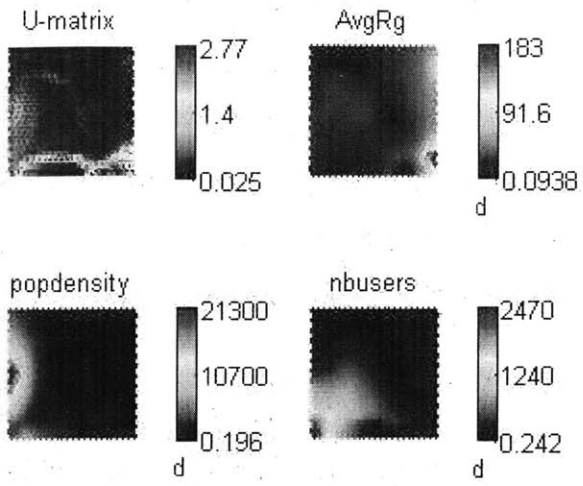


Figure 3-3: SOM results, the feature vector consists radius of gyration (AvgRg), number of users (nbusers), and population density (popdensity).

3.2.3 SOM without Geographical Data

To further test the hypotheses, we exclude geographical data (i.e. latitude and longitude), and construct feature vector only from radius of gyration, number of users, and population density. Fig. 3-3 gives the outcome of SOM

We observe that there is only slightly difference between the U-matrix in Fig. 3-2 and Fig. 3-1, which implies that hypothesis 2 is acceptable, and that by purely relying upon radius of gyration and population information, the towers can be classified in accordance with the real geographical locations.

Therefore, there should exist a relationship between radius of gyration distribution and population density. The relationship may vary across countries because some other factors that are not discovered may influence it, however, at least the relationship exists and can be used to separate towers in different areas.

The hypothesis is weak because at this stage we can do nothing but assert the existence of a relationship between radius of gyration and population. We expect to apply data mining method to parameterize this kind of relationship and eventually find an exact function which ties one to the other.

Fig. 3-4 shows the distribution or average radius of gyration, population density and number of users in these areas. Although the relationship between radius of

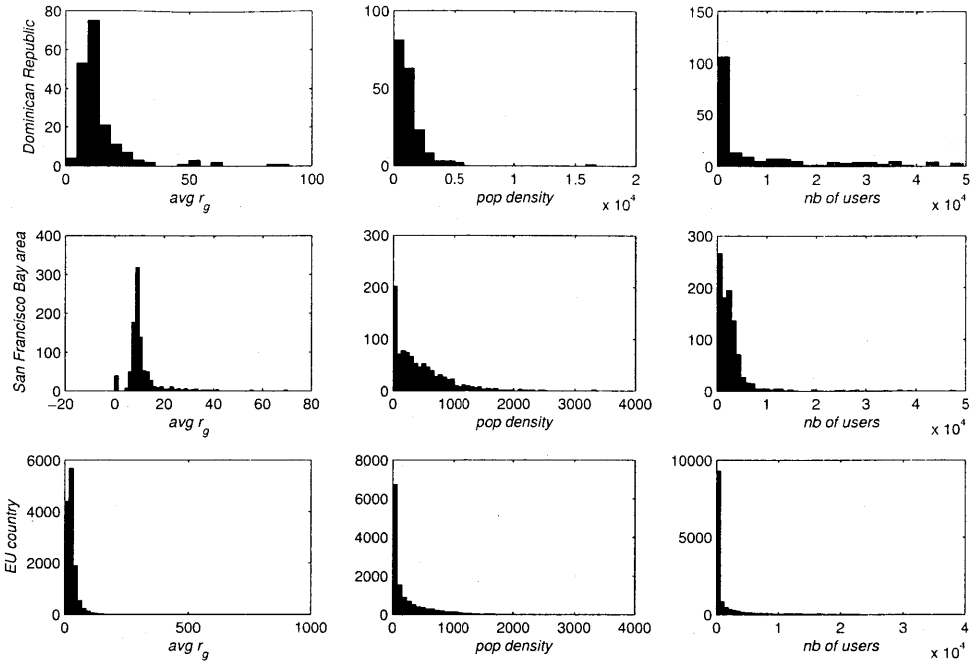


Figure 3-4: The histogram of Average R_g (left figures), Population density (middle figures) and Number of users (right figures) in Dominican republic (upper figures), San Francisco Bay area (middle figures), and the anonymous EU contry (lower figures).

gyration and population density is not as clear as what SOM shows, we can indeed observe the differences existing among three countries. The variance of population density distribution in EU is the biggest among the three countries and San Francisco has smallest variance. If we treat population density as a random variable, we can roughly say that population density distribution of San Francisco Bay area first order stochastically dominates Dominican Republic, which in turn dominates EU. Moreover, the variance and skewness of average R_g distributions are significantly different by observation. Intuitively, all such differences can imply the existence of the relationship between radius of gyration distribution and population density distribution, which is clearly illustrated by SOM.

Chapter 4

Mobility Characterization and Data Mining

4.1 Samples Construction and Parameterization of Mobility

The analysis of trip length distributions requires a framework for reconciling different sets of mobility data and will include methodologies for bias measurement and correction, as well as advances in geographic-based reconciliation. Although methodologies in Chapter 2 produce a set of characterizations of trip length distributions, without further polishment and modification, these measures can hardly reflect the underlying nature of human movement. The framework we need to establish will enable us to characterize how these cross-sectional data co-vary with each other. Relied upon data mining and principle component analysis we hope to develop an understanding of how outcomes of interest, such as population density (and income for San Francisco Bay area), can influence these mobility patterns. We aim to develop an additional mobility index that has significant explanatory power on variables such as average income and education levels.

The existing metrics that can be extracted from the raw call data records includes travel distances, anisotropy and radius of gyration. Among these metrics, the radius

of gyration should be the predominant one because it properly measures the linear size occupied by each user's trajectory up to a specified time. The radius of gyration represents the vital diameter within which the user is most likely to be found in the observation period. We will experiment with quantifying these individual movements as a "regional radius of gyration" to quantify the bounds on the individual's mobility. These regional metrics can be used to characterize a neighborhood or other urban region with known outcomes of interest.

With the probability distributions of radius of gyration for each region, we can search for salient features that are hidden behind them. Finally, combining with the population information data we hope to find and fundamentally explain the inter-relationship between them and furthermore, try to characterize trip length distributions in various regions.

4.1.1 Area Divisions

Combining mobility data sets from a wide variety of countries enables us to make a comparison of human travels at a scale never obtained before. The data sets provided include detailed calling activities in three countries: San Francisco bay area (representing part of U.S.), Dominican Republic, the anonymous EU country. Although these data sets are extensive but they cannot be used straightforwardly for data mining at the statistical level, due to the following reasons:

1. The sampling data represents trip length distributions in three regions with completely different density of population (For example, the population density of San Francisco bay area, Dominican Republic and the EU country are $500 - 2000/km^2$, $210/km^2$ and $96/km^2$, respectively).

2. For most data mining method, at least 20 different samples are required to ensure the reliability and generality of the outcome. In this case, more elaborate area divisions are necessary to better understand the traveling behavior.

To ensure that all the four countries are divided into several sub-regions within which the factor values are alike, the country is segmented according to autonomous communities (the first-level political division of the country).

For example, Dominican Republic is segmented into 3 sub-regions: Santo Domingo, Santiago and La Romana. San Francisco bay area includes 5 sub-regions: Oakland-Alameda-Fermont metropolitan area, SanFrancisco-Redwood City-San Mateo metropolitan area, San Jose- SunnyValey- Santa Clara metropolitan area, San Rafael - Novato-Sausalito metropolitan area and Santa Cruz-Watsonville-Monterey-Salinas metropolitan area. And the EU country is divided into 16 sub regions. Hence, totally 24 regions are available for analysis after segmentation.

4.1.2 Heterogeneity Classification

As shown in section 2.1.4, The radius of gyration of one user characterizes the linear size occupied by his/her trajectory up to a specified time. By aggregating the radius of gyration of all the users in a limited space and constructing its probability density distribution, the radius of gyration distribution for a particular autonomous region ($P(r_g)$) is obtained. These regional metrics can be used to characterize a neighborhood or other urban region with known outcomes of interest.

We observed that the distributions of raidus of gyration can be generally grouped into four families as shown in Fig. 4-1:

A. The $P(r_g)$ curve with single peak in the middle.

B. The $P(r_g)$ curve with single peak near the very beginning.

C. r_g (in logscale) distributes as a double-peak curve: achieving locally-maximum probability density at two points.

D. The $P(r_g)$ curve with no peaks, decreasing monotonically from the beginning.

Since Santa Cruz, three autonomous regions in the EU country have less than 2000 users, their $P(r_g)$ distributions are not considered here. We found that all the areas in San Francisco Bay areas have $P(r_g)$ like A. Santa Domingo, some island region and the capital of the EU country have $P(r_g)$ like C. Santiago and La Romata have $P(r_g)$ like D. Moreover, all the other areas in the EU country have $P(r_g)$ like B.

Individuals living in areas belonged to Group A have a tendency to travel a relatively longer distance (Corresponding to the peak in the middle). Since the $P(r_g)$ of all the areas in the San Francisco Bay areas have this kind of shape, we provide two

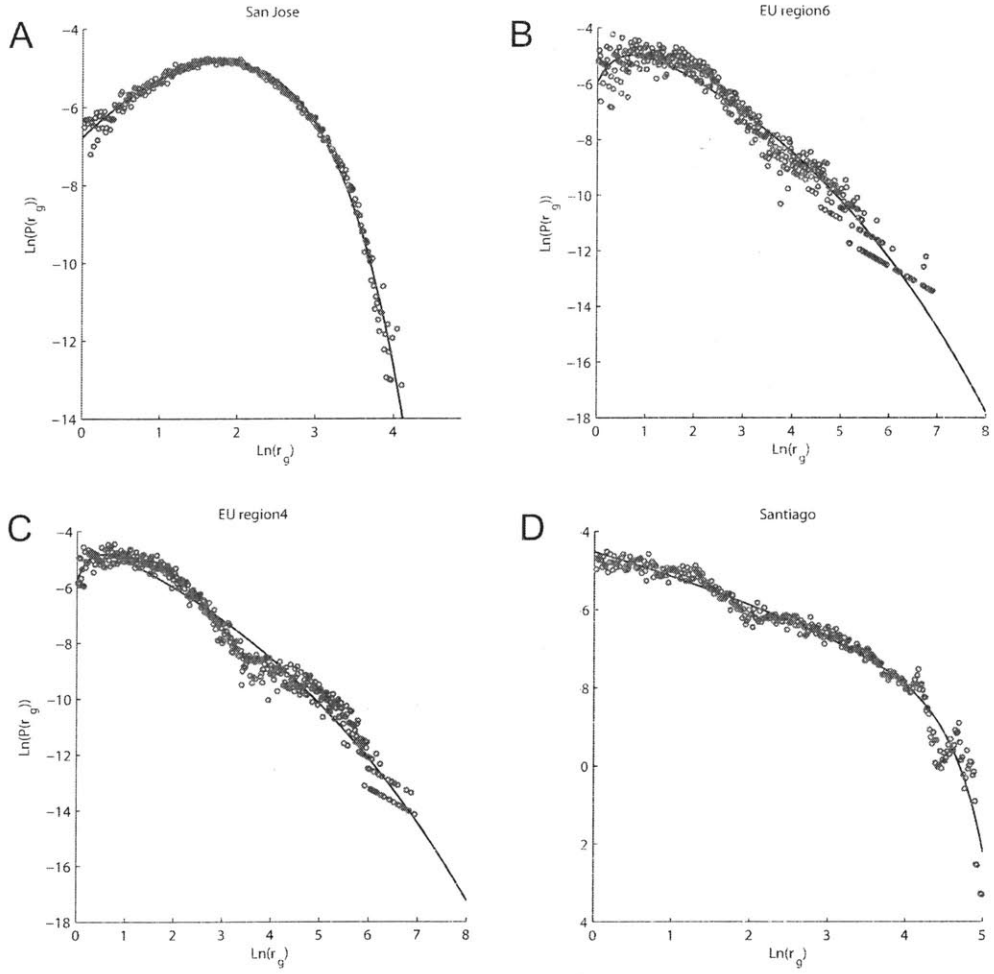


Figure 4-1: Three representative distributions of radius of gyration. The blue dots are distribution of radius of gyration in log scale; The black line is the fitted function according to Eq. (4.1)

possible explanations: Firstly, because of the higher prices of housing and renting in downtown areas, people tend to live in surburbs which resulted in average long distance of daily commuting. Secondly, San Francisco Bay area is best known as the heart of venture capital and high-tech firms of United States, the vital economic activities and frequent business traveling may cause the $P(r_g)$ has single peak in the middle.

Most of the areas belonged to the EU country fall into Group B. Since the not too large EU country is further divided into 16 smaller autonomous regions, the single peak near the beginning of $P(r_g)$ curve implies that facilities within each autonomous region are well developed thereby people have few demand for outside traveling. This coincides with reality in that EU country.

$P(r_g)$ in group C is quite special, which has two peaks. As we observed, only three areas have this shape. One is a island in the EU country, which have numerous excellent harbours and is extremely fertile in all produce, except wine and olive oil. The peak with lower r_g is resulting from people's movements within the islands, and the peak with higher r_g corresponding to transportation to the mainland of the country for commercial purposes. The other two are both capitals of the country, one of them is Santa Domingo, which is the capital of Dominican Republic.

Regions in Group D display curves with monotonic property. Since only two regions of Dominican Republic are belonged to this category, we think the shape might be determined by the country's specific factors. It seems that people living in both regions have monotonic traveling behavior, namely, their traveling preferences are negatively correlated with the distance to destination.

4.1.3 Parameterization of Radius of Gyration

In order to facilitate parameters analysis, we need to search for one function that can best fit all the $P(r_g)$ curves. The function should be simple enough for mathematical analysis and better using least variables to eliminate the possibility of over fitting. With Matlab curve fitting tools we tried several commonly used functions including polynomial function, exponential function, logarithmic function, log normal function

and log logistic function. Finally, we found that the double exponential decaying function is the one that achieving the largest R^2 for all the regions on average. The mathematical representation of this function is:

$$Y = \alpha_1 e^{\beta_1 x} + \alpha_2 e^{\beta_2 x} \tag{4.1}$$

where α_1, α_2, β_1 and β_2 are fitting parameters and x is the variable. Particularly, in this case, $Y = ln(P(r_g))$ and $x = ln(r_g)$.

The following three properties can be found by observation:

Claim 1:

The function value Y is the sum of two simple exponential functions $\alpha e^{\beta x}$, the two elements beside the plus sign are symmetrical and have identical influence on the value of Y.

Claim 2:

If $x \to 0$, the function value Y is mainly determined by $\alpha\beta$, that is Y can be approximated by $g(x) = C + \alpha\beta x$, where C is a constant.

Proof:

Apply Taylor series expansion to Eq. (4.1):

$$Y = \alpha_1 e^{\beta_1 x} + \alpha_2 e^{\beta_2 x} = \alpha_1(1 + \beta_1 x + \frac{\beta_1^2}{2}x^2 + o(x^2)) + \alpha_2(1 + \beta_2 x + \frac{\beta_2^2}{2}x^2 + o(x^2))$$

when $x \to 0$, terms of x with power ≥ 2 can be omitted:

$$Y = \alpha_1(1 + \beta_1 x) + \alpha_2(1 + \beta_2 x) = \alpha_1 + \alpha_2 + \alpha_1\beta_1 x + \alpha_2\beta_2 x$$

It is worth emphasizing that, if $\frac{\alpha_1\beta_1}{\alpha_2\beta_2} \gg 1$, Y can be almost entirely determined by the term $\alpha_1\beta_1 x$.

Claim 3:

If $x \gg 0$, influence by parameter α can be omitted, Y is mainly determined by parameter β.

Proof:

	α_1	β_1	α_2	β_2	$\frac{\alpha_1\beta_1}{\alpha_1\beta_1}$	β_1/β_2	R^2
La Romana	-4.7820	-0.2498	-0.8583	0.5291	-2.6304	-0.4721	0.8591
Santa Domingo	-3.8830	0.3451	0.1440	0.7934	-11.7289	0.4350	0.9643
Santiago	-4.5070	0.1603	0.0000	4.9680	1.757e+10	0.0323	0.9546
Oakland	-4.6430	0.0827	-0.0004	2.8460	326.5193	0.0291	0.9927
San Jose	-4.9240	-0.2010	-0.3480	0.9258	-3.0720	-0.2171	0.982
San Rafael	-4.8620	-0.0546	-0.0240	2.0660	-5.3650	-0.0264	0.9886
San Francisco	-4.3530	0.1477	0.0000	5.8190	1.4806e+6	0.0254	0.9763
Santa Cruz	-4.6610	0.0728	-0.0205	1.3510	12.2289	0.0539	0.9812
EU region1	-4.4270	0.1414	-0.0001	2.3250	1.8736e+3	0.0608	0.9925
EU region2	-4.6490	0.0242	-0.0942	1.0530	1.1330	0.0230	0.9901
EU region3	-4.4390	0.1432	-0.0001	2.2030	3.9290e+3	0.0650	0.9971
EU region4	-4.7110	0.0307	-0.0385	1.3700	2.7361	0.0224	0.9912
EU region5	-4.1500	0.2289	-0.0001	2.5750	4,144.5545	0.0889	0.9729
EU region6	-4.6050	0.1147	0.0000	3.5110	2.6296e+4	0.0327	0.9603
EU region7	-4.6150	0.1144	0.0000	3.6360	9.0865e+4	0.0315	0.9641
EU region8	-4.5710	0.1134	-0.0046	1.5110	73.8063	0.0750	0.9893
EU region9	-4.3350	0.1576	-0.0001	2.3730	2.6732e+3	0.0664	0.9962
EU region10	-4.2750	0.1681	-0.0006	1.9360	616.4954	0.0868	0.9951
EU region11	-4.5450	0.1089	-0.0017	2.0440	145.9602	0.0533	0.9923
EU region12	-4.6050	0.1147	0.0000	3.5110	2.6296e+4	0.0327	0.9902
EU region13	-4.6370	0.0215	-0.0518	1.4080	1.3663	0.0153	0.9962
EU region14	-4.6230	0.1061	-0.0007	1.9500	377.3457	0.0544	0.9906
EU region15	-4.3340	0.1576	-0.0002	1.9250	1.4459e+3	0.0819	0.9955
EU region16	-2.9130	0.3307	-2.0940	-1.0160	-0.4528	-0.3255	0.9817

Table 4.1: $P(r_g)$ curve fitting result of 24 autonomous regions.

Intuitively, β is an exponential coefficient while α is a linear coefficient, the sensitivity of Y with respect of β is more pronounced than that of α if $x \gg 0$. Moreover, if $\frac{\beta_1}{\beta_2} \gg 1$, Y is almost entirely determined by $e^{\beta_1 x}$.

Applying formula 4.1 to each $P(r_g)$ we have, the curve fitting result is shown in Table 4.1:

In the last column, all the R-square are above 90% (except for La Romana due to limited sample size), this proves that the double exponential decaying function can to some extent fulfill our requirement for curve fitting of $P(r_g)$.

To better understand the characteristics of this function, the associated $P(r_g)$ curves of 24 regions in San Francisco Bay area, Dominican Republic and the EU country are displayed in Appendix A.

In Appendix A, we observe:

1. All α_1's are negative and around $4.00-5.00$, and its value determines the shape of curve when x is small (the left side of the curve). Furthermore, if $\frac{\alpha_1\beta_1}{\alpha_2\beta_2}$ is large, one of the exponential function with parameters α_1 and β_1 plays a dominant role in shaping the left side of the curve. Plots of Santiago, Oakland and San Francisco are such cases. If $\frac{\alpha_1\beta_1}{\alpha_2\beta_2} \ll 1$, the shape of the curve to some extent is also influenced by the other exponential function with parameters α_2 and β_2.

2. If β_1 is positive, the curve steadily moves downward from the beginning point (the left-most point). However, if β_2 is positive, the curve moves upward from the beginning point up to some point after which moves downward steadily (because all β_2's are positive and $\beta_2 > \beta_1$, so the left tail is determined by β_2, the shape should moves downward steadily after a threshold of). Moreover, the larger the absolute value of x, the more pronounced the upward tendency would be. (See the plots of La Romana, San Jose, San Rafael)

3. If α_2 has a negative value, the second exponential element would strengthen the decaying speed of the curve, because in this case, both of the two exponential functions have negative αs which in turn have identical influences on the decaying tendency of the curve. Additionally, since β_2 is greater than β_1, the larger the value of x, the more the curve is influenced by the second exponential element. It worth emphasizing that α_2 is positive for Santa Domingo, which is opposite to the sign of α_1. Hence, the decaying tendency of first exponential element $\alpha_1 e^{\beta_1 x}$ is more than offset by the influence of the second exponential element with positive exponent α_2, and eventually the curve would move upward.

4. All β_2s are positive and greater than β_1s, hence the left tail of the curve is mainly determined by the value of β_2. Since the greater the β_2, the faster the decaying speed is. For example, the tails in plots of Santiago, Oakland and San Francisco are steeper, which are consistent with their larger β_2s. Furthermore, a larger value of β_2/β_1 implies that the tail of the curve is mostly determined by the second exponential function.

4.2 Principal Component Analysis

Because of the limited number of data points, it is impossible to characterize the relationship between each parameter in Eq. (4.1) and population density. By applying PCA we are able to extract the first two principle components of the four parameters, which contribute most to the variance of the sample data set. As a first step of data mining, we expect the PCA can provide us with the possibility to formulate the relationship at large.

4.2.1 Introduction

Principal component analysis (PCA) involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. PCA is theoretically the optimum transform for given data in least square terms.

By applying curve fitting to the probability density distribution of radius of gyration in each autonomous community, we are provided with four important parameters that can almost characterize the mobility patterns (regional radius of gyration) in each region. However, these four parameters may correlated with one another, the resulted redundant information would increase the complexity of our analysis of the inter-relationship between these factors and socio economic conditions thus lead to inaccurate conclusions. The four parameters could be regarded as a combining vector, hence PCA can be utilized to disentangle the cross correlation among them.

The mathematical foundation behind PCA is very simple:

Let X be a d-dimensional random vector expressed as column vector. Without

loss of generality, assume X has zero mean. We want to find a $d \times d$ orthonormal transformation matrix P such that,

$$Y = P^T X$$

s.t. $cov(Y)$ is a diagonal matrix, $P^{-1} = P^T$

By substitution, and matrix algebra, we obtain:

$$cov(Y) = E[YY^T] = E[(P^T X)(P^T X)^T] = E[(P^T X)(X^T P)] = P^T E[XX^T]P = P^T cov(X)P$$

We have,

$$Pcov(Y) = PP^T cov(X)P = cov(X)P$$

Rewrite P as $d \times 1$ column vectors, so

$$P = [P_1, P_2, ..., P_d]$$

and $cov(Y)$ as

$$\begin{pmatrix} \lambda_1 & \dots & 0 \\ \\ 0 & \dots & \lambda_d \end{pmatrix}$$

Substituting into equation above, we obtain:

$$[\lambda_1 P_1, \lambda_2 P_2, ..., \lambda_d P_d] = [cov(X)P_1, cov(X)P_2, ..., cov(X)P_3]$$

Notice that in $\lambda_i P_i = cov(X)P_i$, P_i is an eigenvector of the covariance matrix of X. Therefore, by finding the eigenvectors of the covariance matrix of , we find a projection matrix P that satisfies the original constraints.

4.2.2 Computational Procedures

Following is a detailed description of PCA using the covariance method. The goal is to transform a given data set X of dimension M to an alternative data set Y of smaller dimension L. Equivalently, we are seeking to find the matrix Y, where Y is the Karhunen-Love transform (KLT) of matrix X:

$$Y = KLT\{X\}$$

Organization of data set

Suppose you have data comprising a set of observations of M variables, and you want to reduce the data so that each observation can be described with only L variables, $L < M$. Suppose further, that the data are arranged as a set of N data vectors $X_1, ..., X_N$ with each X_i representing a single grouped observation of the M variables.

1. Write $X_1, ..., X_N$ as column vectors, each of which has M rows.
2. Place the column vectors into a single matrix X of dimensions $M \times N$.

Calculation of the empirical mean

1. Find the empirical mean along each dimension $m = 1, ..., M$.
2. Place the calculated mean values into an empirical mean vector u of dimensions $M \times 1$.

$$u[m] = \frac{1}{N} \sum_{i=1}^{N} X[m, i]$$

Calculation of the deviation from the mean

Mean subtraction is an integral part of the solution towards finding a principal component basis that minimizes the mean square error of approximating the data. Hence we proceed by centering the data as follows:

1. Subtract the empirical mean vector u from each column of the data matrix X.
2. Store mean-subtracted data in the $M \times N$ matrix B.

$$B = X - uh$$

where h is a $1 \times N$ row vector of all 1s:

$$h[n] = 1, for\ n = 1, ..., N$$

Calculation of the covariance matrix

Find the $M \times M$ empirical covariance matrix C from the outer product of matrix with itself:

$$C = E[B \oplus B] = E[B \cdot B^*] = \frac{1}{N} \sum B \cdot B^*$$

where $\oplus$ is the outer product operator, and $*$ is the conjugate transpose operator. Note that if B consists entirely of real numbers, which is the case in many applications, the "conjugate transpose" is the same as the regular transpose.

Calculation of the eigenvectors and eigenvalues of the covariance matrix

1. Compute the matrix V of eigenvectors which diagonalizes the covariance matrix C:

$$V^{-1}CV = D$$

where D is the diagonal matrix of eigenvalues of C. This step will typically involve the use of a computer-based algorithm for computing eigenvectors and eigenvalues. These algorithms are readily available as sub-components of most matrix algebra systems, such as MATLAB.

2. Matrix D will take the form of an $M \times M$ diagonal matrix, where

$$D[p, q] = \lambda_m, for\ p = q = m$$

is the mth eigenvalue of the covariance matrix C, and

$$D[p, q] = 0, for\ p \neq q$$

3. Matrix V, also of dimension $M \times M$, contains M column vectors, each of length M, which represent the M eigenvectors of the covariance matrix C.

4. The eigenvalues and eigenvectors are ordered and paired. The mth eigenvalue corresponds to the mth eigenvector.

Rearrangement of the eigenvectors and eigenvalues

1. Sort the columns of the eigenvector matrix V and eigenvalue matrix D in order of decreasing eigenvalue.

2. Make sure to maintain the correct pairings between the columns in each matrix.

Computation of the cumulative energy content for each eigenvector

The eigenvalues represent the distribution of the source data's energy among each of the eigenvectors, where the eigenvectors form a basis for the data. The cumulative energy content g for the mth eigenvector is the sum of the energy content across all of the eigenvalues from 1 through m:

$$g[m] = \sum_{q=1}^{m} D[q,q], for\ m = 1, ..., M$$

Set a subset of the eigenvectors as basis vectors

1. Save the first L columns of V as the $M \times L$ matrix W:

$$W[p,q] = V[p,q], for\ p = 1, ..., M,\ q = 1, ..., L$$

where $1 \leq L \leq M$.

2. Use the vector g as a guide in choosing an appropriate value for L. The goal is to choose a value of L as small as possible while achieving a reasonably high value of g on a percentage basis. For example, choose L so that the cumulative energy g is above a certain threshold, like 90%. In this case, choose the smallest value of L such that

$$g[m = L] \geq 90\%$$

Conversion of the source data to z-scores

1. Create an $M \times 1$ empirical standard deviation vector s from the square root of each element along the main diagonal of the covariance matrix C:

$$s = s[m] = \sqrt{C[p,q]}, for\ p = q = m = 1, ..., M$$

2. Calculate the $M \times N$ z-score matrix:

$$Z = \frac{B}{sh}$$

Projection of the z-scores of the data onto the new basis

1. The projected vectors are the columns of the matrix

$$Y = W^* Z = KLT\{X\}$$

2. W^* is the conjugate transpose of the eigenvector matrix. 3. The columns of matrix y represent the Karhunen-Loeve transforms (KLT) of the data vectors in the columns of matrix X.

4.2.3 PCA Results

In Table 4.1 we present the curve fitting parameters for the distribution of radius of gyration in each autonomous region. Here we apply the PCA method to transform these four vectors into another four but orthogonal vectors and extract the principal components from them.

The average value and standard deviation of each parameter is computed and presented in Table 4.2. Totally the sample space is composed of 24 data points. and the standard deviation of $\alpha_1, \beta_1, \alpha_2, \beta_2$ are $0.400, 0.133, 0.457, 1.421$, respectively. Notice that the standard deviations are close to each other, that is, no single parameter can contribute most of the variability existing among different data points. This motivates us to use PCA to orthogonally transform the sample space. The associated correlation matrix is presented in Table 4.3, as shown clearly the matrix is far from

	Mean	Std. Deviation	Sample Num
α_1	-4.456	0.400	24.000
β_1	0.100	0.133	24.000
α_2	-0.141	0.457	24.000
β_2	2.128	1.421	24.000

Table 4.2: Descriptive statistics for parameters alpha and beta.

	α_1	β_1	α_2	β_1
α_1	1.000	0.760	-0.631	-0.304
β_1	0.760	1.000	-0.018	0.124
α_1	-0.631	-0.018	1.000	0.561
β_2	-0.304	0.124	0.561	1.000

Table 4.3: Correlation Matrix for parameters alpha and beta.

diagonal.

The eigenvalues and eigenvectors of the covariance matrix are computed according to information in Table 4.3. Four components are extracted and the top two principal components can explain 97.03% variance in the data sets, which guarantees the efficiency of using PCA to transform parameters' space. The two components are further analyzed in conjunction with the population density in Section 5.1. (See Table 4.4, 4.5 and Fig. 4-2)

As Table 4.5 indicates, the two principal components can be calculated using the formula:

$$P_1 = 0.098\alpha_1 - 0.009\beta_1 - 0.193\alpha_2 - 0.976\beta_2 \tag{4.2}$$

	Initial Eigenvalues		
Components	Total	% of Variance	Cumulative %
1	2.020	87.65%	87.65%
2	0.216	9.38%	97.03%
3	0.068	2.93%	99.96%
4	0.001	0.04%	100.00%

Table 4.4: Total Variance Explained by each component.

	Component			
	1	2	3	4
α_1	0.098	0.718	0.577	0.377
β_1	-0.009	0.166	0.388	-0.907
α_2	-0.193	-0.646	0.714	0.189
β_2	-0.976	0.198	-0.087	0.009

Table 4.5: Composition of components.

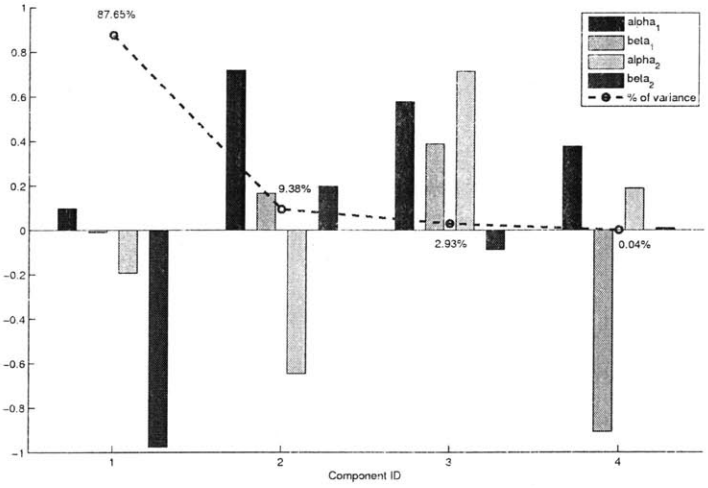


Figure 4-2: The composition and variance of 4 components.

	α_1	β_1	α_2	β_2	P_1	P_2
La Romana	-4.782	-0.2498	-0.8583	0.5291	1.6703	1.6703
Santa Domingo	-3.883	0.3451	0.144	0.7934	1.3013	0.0031
Santiago	-4.507	0.1603	0	4.968	-2.8058	0.4447
Oakland	-4.643	0.0827	-0.0004	2.846	-0.7466	-0.0861
San Jose	-4.924	-0.201	-0.348	0.9258	1.1701	-0.4908
San Rafael	-4.862	-0.0546	-0.024	2.066	-0.0008	-0.4054
San Francisco	-4.353	0.1477	0	5.819	-3.6214	0.7219
Santa Cruz	-4.661	0.0728	-0.0205	1.351	0.7151	-0.3839
EU region1	-4.427	0.1414	-0.0001	2.325	-0.2175	-0.0247
EU region2	-4.649	0.0242	-0.0942	1.053	1.0218	-0.3948
EU region3	-4.439	0.1432	-0.0001	2.203	-0.0996	-0.0572
EU region4	-4.711	0.0307	-0.0385	1.37	0.6955	-0.4114
EU region5	-4.15	0.2289	-0.0001	2.575	-0.4352	0.2383
EU region6	-4.605	0.1147	0	3.511	-1.3925	0.0780
EU region7	-4.615	0.1144	0	3.636	-1.5155	0.0956
EU region8	-4.571	0.1134	-0.0046	1.511	0.5643	-0.2911
EU region9	-4.335	0.1576	-0.0001	2.373	-0.2555	0.0536
EU region10	-4.275	0.1681	-0.0006	1.936	0.1770	0.0121
EU region11	-4.545	0.1089	-0.0017	2.044	0.0459	-0.1695
EU region12	-4.494	0.1308	0	2.962	-0.8458	0.0516
EU region13	-4.637	0.0215	-0.0518	1.408	0.6683	-0.3437
EU region14	-4.623	0.1061	-0.0007	1.95	0.1299	-0.2452
EU region15	-4.334	0.1576	-0.0002	1.925	0.1820	-0.0344
EU region16	-2.913	0.3307	-2.094	-1.016	3.5948	1.7849

Table 4.6: PCA results.

$$P_2 = 0.718\alpha_1 + 0.166\beta_1 - 0.646\alpha_2 + 0.198\beta_2 \tag{4.3}$$

where P_1 and P_2 denote the first two principal components, Table 4.6 exhibits values of the first two components for each area. Since the two components captures most of the variability of the sample points, we expect to search for their relationship to population density. This work is presented in the next section.

Chapter 5

Regression Analysis

5.1 A Simple Linear Regression Model

We expect to find the relationship between radius of gyration and population for autonomous areas. With the assistance of ORNL LandScan data sets, we are able to collect the 2008 population density information accurately for every interested area.The statistics presented in Table 4.6 are in number of persons per square kilometer.

To find the correlation property, one trivial way is to regress the first component P_1 on the population density we collected from Landscan data. Since we already know that P_1 contributes 87.65% variance of the four parameters, thereby if we can identify the relationship of P_1 and population density, then we could roughly estimate the 4 original parameters $\alpha_1, \alpha_2, \beta_1, \beta_2$ by holding the other three components constant or with small random variations (which will not affect the estimation greatly because the total variance of the remaining components are about 12.35%).

However, as we see in Fig. 5-1, the 24 points are almost randomly distributed on the plane and it is hardly possibly to find a clear relationship. Since by observation, there is no clear relationship between population information and radius of gyration, we expect to use linear regression model to discover this relationship.

Let P_d represent the population density data, P_1 be the most significant component that reflect human's radius of gyration.

Population density		Mean	Std. dev.
Dominican Republic	La Romata	119.454	1100.535
	Santa Domingo	164.250	1205.920
	Santiago	151.955	1153.863
San Francisco Bay Area	Oakland	763.633	1553.614
	San Jose	524.755	1237.002
	San Rafael	174.431	550.223
	San Francisco	1158.813	3059.732
	Santa Cruz	257.717	782.182
Spain	EU region1	88.892	619.176
	EU region2	98.638	800.856
	EU region3	25.913	272.903
	EU region4	98.523	508.866
	EU region5	292.965	1657.708
	EU region6	47.972	426.564
	EU region7	55.199	466.438
	EU region8	25.022	341.798
	EU region9	192.292	1179.188
	EU region10	173.497	950.976
	EU region11	101.810	541.583
	EU region12	21.798	185.247
	EU region13	666.767	2481.453
	EU region14	25.158	187.652
	EU region15	80.117	636.881
	EU region16	158.510	589.263

Table 5.1: Population density statistics for all the autonomous regions.

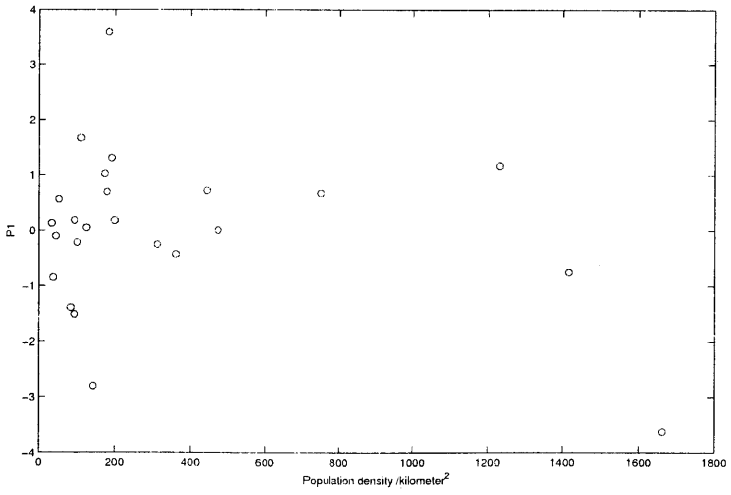


Figure 5-1: P_1 versus Population Density.

Consider the simplistic linear regression model:

$$P_1 = \gamma_0 + P_d\gamma_1 + \epsilon \tag{5.1}$$

Suppose the cross-sectional data we have satisfies the Gauss-Markov conditions, which guarantees the validity of OLS linear regression:

GM1: $Y = X\beta + \epsilon, \beta \in R^k$, linearity

GM2: rank(X)=k, identification

GM3: $E_\theta[\epsilon|X] = 0\ \forall\theta \in \Theta$, orthogonality, correct specification, exogeneity

GM4: $E_\theta[\epsilon\epsilon^{'}|X] = \sigma^2 I_{n\times n}$

Where, in our model, Y is P_1, X is $[1, P_d]$.

The regression result by Eq. (5.1) is not appealing, the associated R^2 is 0.010, which imply nothing but a random walk. Hence, simply using population density as the regressor is not sufficient to provide us an acceptable result. As we have shown in Chapter 3, there do exist a relationship between population information and radius of gyration. Here by regressing the regressand P_1 on regressors the unit vector e_1 and population density P_d, the regression performance is very bad. To improve the regression, we generally come up with two choices:

1. Include other statistics of population density, such as its standard deviation.
2. Include other socio-economic factors, such as income, unempolyment rate, etc.

In the next section, we conduct a further regression analysis for each of the above propositions.

5.2 Multivariate-Regression Model

The mean of population density reflects the average value of the population, but fails to reflect the variations of population distribution in an interested area. As a natural way to enrich the simple linear regression model in Eq. (5.1), we incorporate the normalized standard deviation $\sigma(P_d)/P_d$.

Additionally, the multivariate-regression model we propose in this chapter contains another two explanatory variables c_β and σ^2 derived from a Multiplicative Spatial

Models of Supply and Demand [11].

So, we use the model Eq. (5.3)

$$P_1 = \gamma_0 + P_d\gamma_1 + \frac{\sigma(P_d)}{P_d}\gamma_2 + c_\beta\gamma_3 + \sigma^2\gamma_4 + \epsilon \tag{5.2}$$

The estimators given by OLS are:

$$\hat{\gamma_0} = -3.5524,\ \hat{\gamma_1} = 1.0776,\ \hat{\gamma_2} = 6.6929,\ \hat{\gamma_3} = -0.0019,\ \hat{\gamma_4} = -0.5024,\ R^2 = 0.2879$$

Although, a 0.2879 R^2 is far from enough to give accurate estimation, at least it gives us prediction power with population information given by the census data. However, it is worth noticing that in Eq. (5.3), we are using 4 regressors which are two much for a sample containing 24 points. In this case, a better approximation of the data points (implied by a larger R^2) doesn't always mean a better estimation of the true regression parameters' values according to econometrics theory. The standard error of our estimator gamma is given by Eq. (5.4):

$$std.error = \frac{s^2}{\sqrt{n}} = \frac{\hat{e}^T\hat{e}}{n-k} \tag{5.3}$$

where $\hat{e}$ is the residual, k is the number of regressors, as Eq. (5.3) implies, if k increases, the standard error of *beta* may also increase, this effect is particularly significant for a small value of n. Therefore, the improved R^2 obtained from Eq. (5.3) doesn't imply that the value of β will be more reliable.

5.3 Regression in San Francisco Bay Area: Influence of Demographic Information

The rich information available in San Francisco Bay area enables us to conduct a specific research. The population density, median family income, unemployment rate and median age, etc. at a resolution level of the transmission distance of towers are obtained from TransCAD (See Fig. 5-2). The radius of gyration associated with each

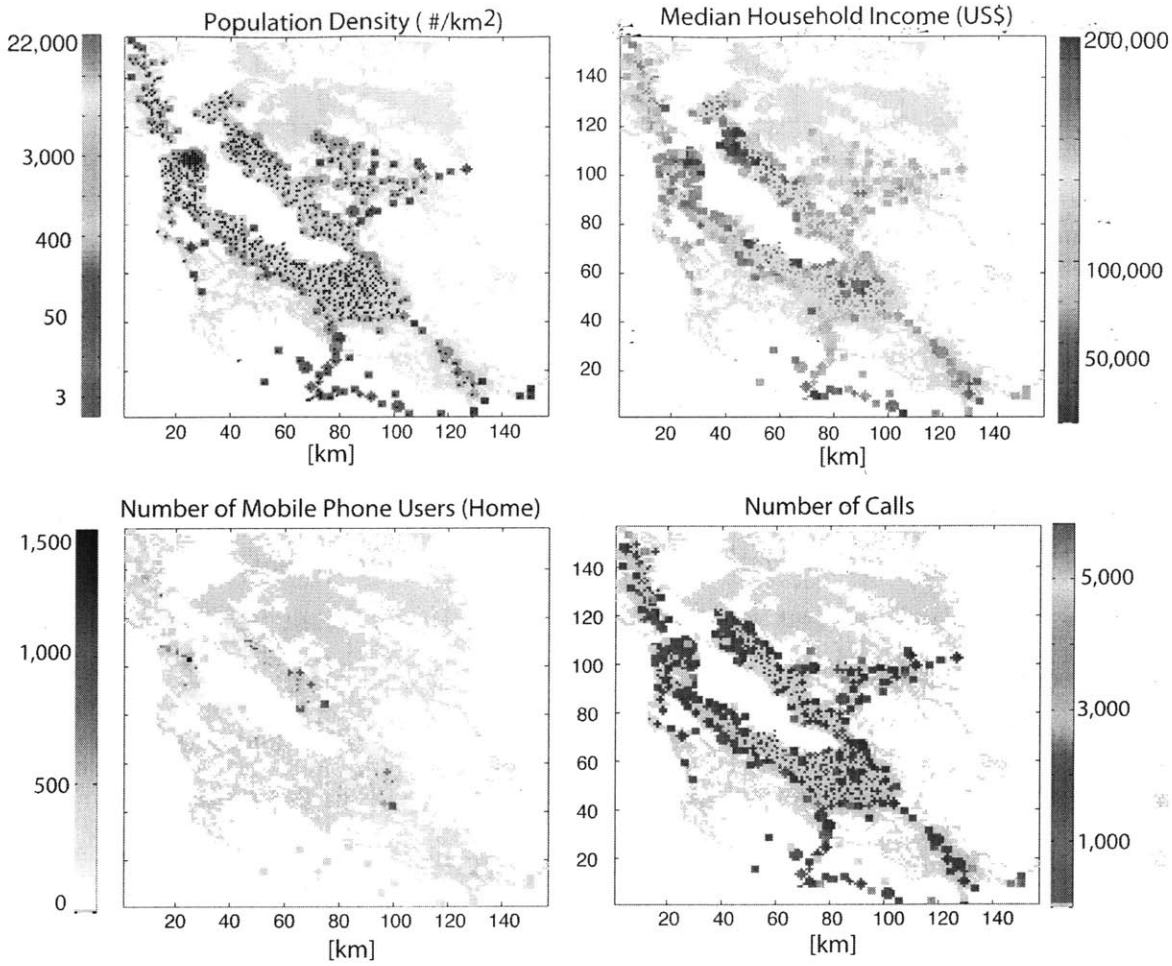


Figure 5-2: Population density and Income distribution in San Francisco Bay area. The upper-left graph is population density distribution, at a resolution level of $/km^2$; The upper-right graph is median household income distribution; The lower-left graph is the distribution of mobile phone users (note: we use user's most visited location as proxy for his/her home location); The lower-right graph is the distribution of number of calls.

tower is defined as the mean value of radius of gyration of users who most frequently visit the tower.

Notice that among the 954 towers, 401 towers are distributed in the areas with very low population or with less than 100 mobile phone users around(according to our one-month mobile phone record). Hence these data points are eliminated and 553 data points are available for regression analysis.

Let I be the median family income, P_d be the population density, A be the median age, U be the unemployment rate for people who are at least 16 years old and in the labor force, Eq. (5.4) is used as the regression model:

	$\hat{\gamma_0}$	$\hat{\gamma_1}$	$\hat{\gamma_2}$	$\hat{\gamma_3}$	$\hat{\gamma_4}$
Value	10.9862	1.1927	-0.6398	-1.5870	-0.1354
$\hat{S.E.}$	0.5049	0.2162	0.0544	0.4792	0.0954
t-stat	21.7600	5.5158	-11.7501	-3.3117	-1.4196

Table 5.2: Regression result: regressing radius of gyration on income, population density, age and unemployment rate.

$$R_g = \gamma_0 + I\gamma_1 + P_d\gamma_2 + A\gamma_3 + U\gamma_4 + \epsilon \tag{5.4}$$

We normalized the data of median family income, population density, median age, and unemployment rate by dividing the value of each data point by the mean. The estimators and test statistics given by Ordinary Least Squares regression are presented in Table 5.2:

We obtain adjusted R-square $R^2 = 0.3188$ and mean squared error $MSE = 2.2667$, which is sufficient for the cross-sectional data. Fig. 5-3 presents the estimated r_g using Eq. (5.4) and the true r_g obtained from mobile phone records. As shown in the figure, the dots lie around the 45 degree line, implying that the estimated r_gs approximate the true r_gs well.

A positive $\hat{\gamma_1}$ implies that the radius of gyration is larger in wealthier areas, which is very reasonable since wealthier people may tend to travel frequently and can afford long distance trips. $\hat{\gamma_2}$ is negative implies that the radius of gyration is smaller in populated areas. The low population density area is usually rural and surburb area, which means people are more likely to travel by cars between rural (their home) area to downtown (high population, work, office) area. The high attraction of the downtown will cause people living far away to visit it, thus increase the value of r_g which reflects the average trip length of people in that area. On the other hand, the high population density areas are caused by a huge amount of people living in the downtown, who commute in a short distance by means of walking or using public transportation every day. A negative $\hat{\gamma_3}$ implies that the elder people travel less which is fairly reasonable. A negative sign of $\hat{\gamma_4}$ implies that in the area with a high unemployment rate, people traveling less, this substantiates the analysis above,

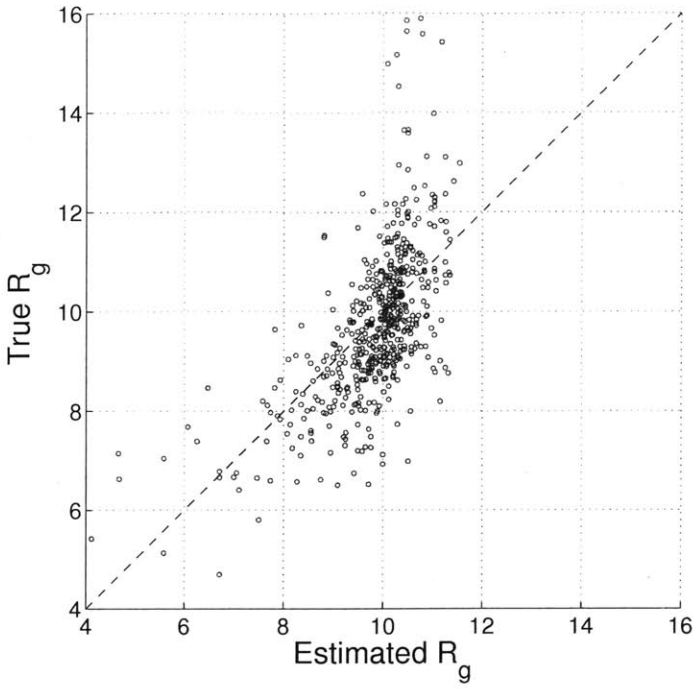


Figure 5-3: The estimated r_g v.s. the true r_g.

where we state that people living populated areas travel less because they undertake pressures from the possibility of losing jobs. In fact, the t-stat of $\hat{\gamma}_4$ is only -1.4196, which is not sufficiently significant.

We are particularly interested in how the distribution of radius of gyration depends on demographic information, namely, how the parameters in Eq. (4.1) related to population density, income, age, and unemployment rate, etc. However, due to the lackness of available data, it is impossible to precisely identify such relationships. Therefore, we expect to use some general methods to provide intuitions.

Since the information about age and unemployment rate have little influences on r_g (if we run a regression of r_g on age and unemployment rate, the resulting R^2 is only about 0.52875), here we only consider population density and family median income information.

We separated the 951 towers in San Francisco Bay area into 9 group (3 towers are eliminated due to lack of income and population density information):

1). towers in low population and low income area

num. of towers	Low Population	Mid Population	High Population	Total
Low Income	85	73	159	317
Mid Income	81	119	117	317
High Income	151	125	41	317
Total	317	317	317	951

Table 5.3: Number of towers in each group. Low population density: $0-2039/km^2$**, Median population density:** $2039-7285/km^2$**, High population density:** $\geq 7285/km^2$**; Low income:** $0-61100$\$**, Median income:** $61100-89440$\$**, High income:** ≥ 89440\$

2). towers in low population and mediate income area

3). towers in low population and high income area

4). towers in mediate population and low income area

5). towers in mediate population and mediate income area

6). towers in mediate population and high income area

7). towers in high population and low income area

8). towers in high population and mediate income area

9). towers in high population and high income area

As shown in Table. 5.3 and Table. 5.4, each group of towers have sufficient number of towers and enough number of users. Notice that for the high population area, there is a larger percentage of people (about $62640/114301 = 54.8\%$ users) with low income than those areas with median or low population area. This fact is reasonable and substantiates the validness of our mobile phone data.

Fig. 5-4 shows the $P(r_g)$ of each group, the arrangement order is in accordance with that in Table. 5.3. Fig. 5-5 shows the $P(r_g)$ distributions in areas with the same mediate level of income but different population densities. As it is clearly shown, the area with low population density has a higher probability to achieve larger r_g than the area with mediate or high population density. This is because areas with low population density are most likely to be suburbs, where people need to travel relatively long distances everyday from home to work places or do anything else. In contrast, areas with high population density are mostly downtown, where people have less tend to make long distance trips. Fig. 5-6 compares the $P(r_g)$ distributions in areas with the same population density but different levels of income. As shown in

num. of users	Low Population	Mid Population	High Population	Total
Low Income	11591	16960	62640	91191
Mid Income	6199	23564	40891	70654
High Income	11904	15579	10770	38253
Total	29694	56103	114301	200098

Table 5.4: Number of users in each group.

the figure: On the one hand, the area with low level of income has a higher probability to achieve larger r_g than the area with mediate or high level of income. This is very reasonable since areas with low level income are most likely to be less developed, which don't have sufficient facilities or public goods to satisfy the daily demand of individuals. Therefore, people need to go to other regions for shopping or whatever they need, which in turn increases the probability of larger r_gs. In contrast, areas with high level of income are economically healthier and well developed, individuals living in such areas find there demands easily to be satisfied within the area hence have no needs to make long distance trips. On the other hand, however, the area with low level of income also has a higher probability to achieve smaller r_g, this is clearly shown in the figure by taking a look at the cutoff point on y-axis. This phenomenon can be explained as individuals with low income tends to make less trips (corresponding to the cutoff, which is the probability of do not traveling.), because they cannot afford the cost from too much traveling. Or they prefer to make short distance trips, which may cost less. The higher probability of short distance trips and long distance trips are not in contradiction, because those long distance trips are necessary trips to satisfy their basic demands that all individuals cannot avoid.

We fitted the $P(r_g)$ curves in Fig. 5-4 by applying Eq. (4.1), and the associated parameters for each group are presented in Table. 5.4. The sufficiently large R^2s indicate that Eq. (4.1) indeed fit the $P(r_g)$ curve very well. The numbers in table. 6 provide some useful hints on determining values of parameters if one want to use Eq. (4.1) to characterize the radius of gyration distribution in another region. The ranges of $\alpha_1, \alpha_2, \beta_2, \beta_2$ are $(-5, -8)$, $(-0.4, -1.2)$, $(-0.3, -0.6)$, $(0.5, 0.8)$, respectively. As we can see, the range of α_1 and α_2, and the range of β_1 and β_2 are very different,

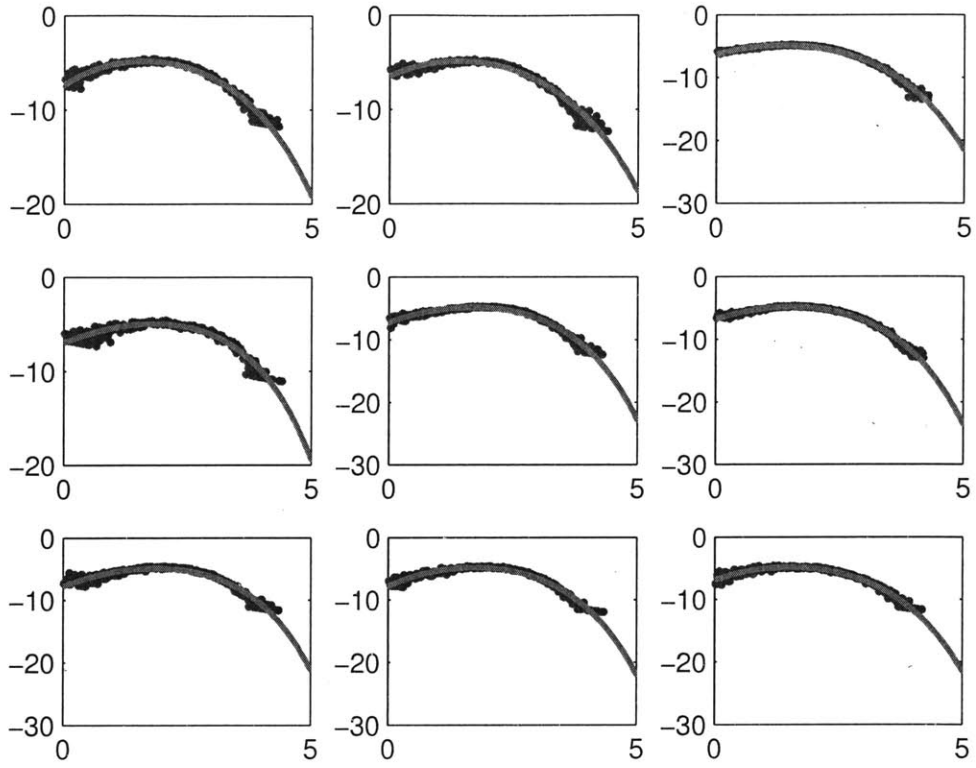


Figure 5-4: $P(r_g)$ **in each group**

	α_1	β_1	α_2	β_2	R^2
Low Population Low Income	-6.415	-0.613	-0.9243	0.60331	0.963
Low Population Mid Income	-6.543	-0.384	-0.4288	0.7524	0.9297
Low Population High Income	-7.323	-0.4884	-0.4765	0.7525	0.9395
Mid Population Low Income	-5.377	-0.5739	-1.127	0.5581	0.9677
Mid Population Mid Income	-6.719	-0.5015	-0.5613	0.7364	0.9667
Mid Population High Income	-7.192	-0.5674	-0.6176	0.7103	0.9568
High Population Low Income	-5.371	-0.5318	-0.9556	0.6196	0.989
High Population Mid Income	-6.117	-0.5125	-0.6899	0.7025	0.9842
High Population High Income	-6.06	-0.5599	-0.8342	0.6462	0.9676

Table 5.5: Parameters value for each group.

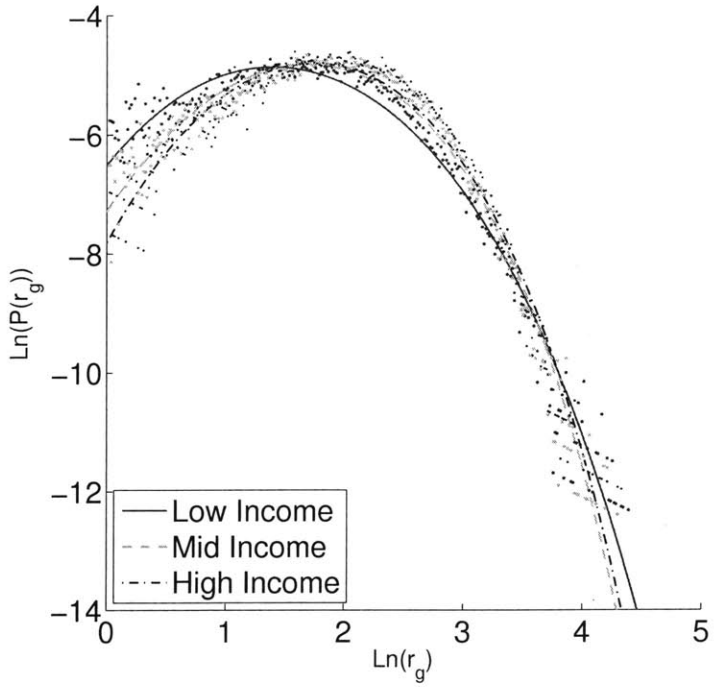


Figure 5-5: $P(r_g)$ **distribution in areas with mediate level of income and different population density**

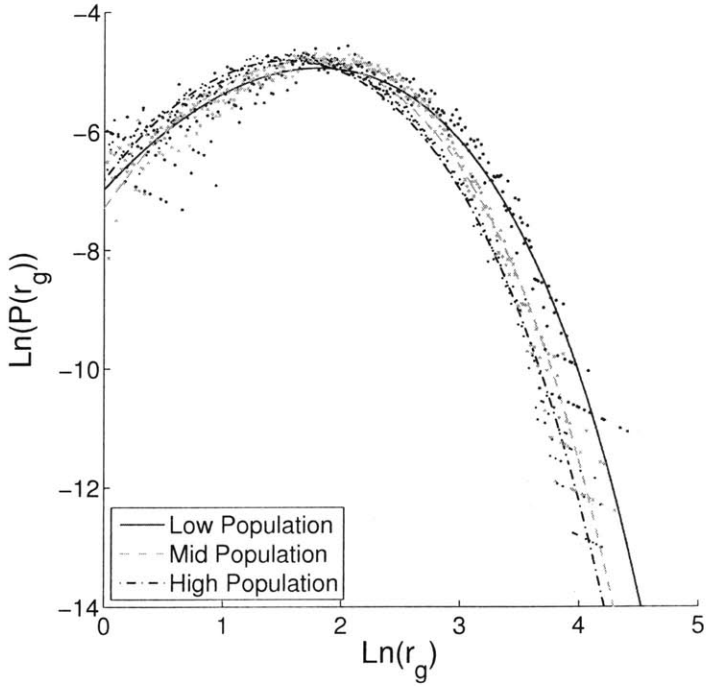


Figure 5-6: $P(r_g)$ **distribution in areas with mediate population density and different levels of income**

Fixed Population Density				
Income	α_1	β_1	α_2	β_2
$Mid - Low$	(-0.13, -1.34, -0.75)	(0.23, 0.07, 0.02)	(0.50, 0.57, 0.27)	(0.15, 0.18, 0.08)
$High - Mid$	(-0.78, -0.47, 0.06)	(-0.10, -0.07, -0.05)	(-0.05, -0.06, -0.14)	(0.00, -0.03, -0.06)
Fixed Income				
Population	α_1	β_1	α_2	β_2
$Mid - Low$	(1.04, -0.18, 0.13)	(0.04, -0.12, -0.08)	(-0.20, -0.13, -0.14)	(-0.05, -0.02, -0.04)
$High - Mid$	(0.01, 0.60, 1.13)	(0.04, -0.01, 0.01)	(0.17, -0.13, -0.22)	(0.06, -0.03, -0.06)

Table 5.6: The upper table is obtained by fixing population density at a certain level (i.e. Low, Mid and High correspond to the first, second and third number in the parenthesis respectively) and calculating the difference between values at different income levels (Mid-Low, High-Low) for each parameter. The lower table is obtained by fixing income at a certain level (i.e. Low, Mid and High correspond to the first, second and third number in the parenthesis respectively) and calculating the difference between values at different population densities (Mid-Low, High-Low) for each parameter.

which provides sufficient flexibility to characterize distributions of radius of gyration. A simple sensitivity test or mathematical analysis would show that:

1. the shape of the beginning part of Eq. (4.1) is mainly determined by the values of $\alpha_1 + \alpha_2$, which is also the cutoff on the y-axis. (Can be proved by Talor expansion at $r_g = 0$)

2. the shape of the tail is mainly determined by $\alpha_2 e^{\beta_2 ln(r_g)}$, i.e. the second exponential element. This is because β_1 is negative, so when r_g is large (corresponding to the tail part), the first exponential element decays to zero, and have little influence. However, β_2 is positive which increases the second element exponentially and in turn provides much stronger influence.

3. the middle part of the curve is determined by both exponential elements. Since the sign of β_1 and $beta_2$ are different, the two elements have offsetting effects on the shape which imply the models flexibility. (The offsetting effects can be showed mathematically from the first derivative of the function.)

Roughly known the range of the parameters value is not enough if we want to achieve a high level of accuracy. Therefore, we are interested in the relationship between each parameters and the demographic factors, namely population density and income. A simple calculation as shown in Table 5.5 provides a lot of insights.

Table. 5.6 is a transformation of Table 5.5 by replacing a positive value in the table with "+", a negative value with "-" and zero with "-/+".

As the signs in table. 5.6 clearly show, for fixed population density: α_1 decreases when income increases (except for areas with high population density, where α_1 firstly

Fixed Population Density				
Income	α_1	β_1	α_2	β_2
$Mid - Low$	(-, -, -)	(+, +, +)	(+, +, +)	(+, +, +)
$High - Mid$	(-, -, +)	(-, -, -)	(-, -, -)	(-/+, -, -)
Fixed Income				
Population	α_1	β_1	α_2	β_2
$Mid - Low$	(+, -, +)	(+, -, -)	(-, -, -)	(-, -, -)
$High - Mid$	(+, +, +)	(+, -, +)	(+, -, -)	(+, -, -)

Table 5.7: A transformation of Table 5.5 by replacing a positive value in the table with "+", a negative value with "-" and zero with "-/+".

decreases and then increases). α_2, β_1 and β_2 increases when income increases from low level to mediate level, but decreases above the mediate leve. For fixed income level: α_1 increases when population density increases (except for areas with mediate income level, where α_1 decreases first and then increases). β_1 increases when population density increases at the low level income areas, however β_1 decreases when population density decreases at the mediate level income areas. α_2 decreases when population density decreases (except for areas with low income level). β_2 decreases when population density decreases (except for areas with low income level).

Chapter 6

Conclusions

Mobile phone records provide us unprecedented access to the spatiotemporal localization of hundreds of millions of users. Analyzing the calling activities through mobile phone data in conjunction with population density data and income data from GIS files, we conducted a thorough research with the purpose of understanding heterogenous trip length distributions in various regions and how they might be correlated with socio-economic factors, such as population density and income.

The trip length distributions can be characterized by radius of gyration distribution, which measures the linear size occupied by individual's trajectory up to a specific time. Moreover, the radius of gyration can also be interpreted as the radius of a circle within which the individual can be probably found. By display the radius of gyration and population density in a single Self-organizing map, we visually identified the inter-relation between the two factors. The existence of such kind of inter-relationship between radius of gyration and population density motivates us to formally characterize them with a mathematical formula. Using a double exponential function, we are able to well fit all the 24 $P(r_g)$ curves, that is, each curve can be characterized by four parameters and an exact function. However, using population density (or additional socio-economic factors, which are not considered in this dissertation) as the independent variable to mathematically represent all the four parameters in a function form are impossible because of the limitation of available sample points. This leads us to conduct the PCA analysis which is widely-known as

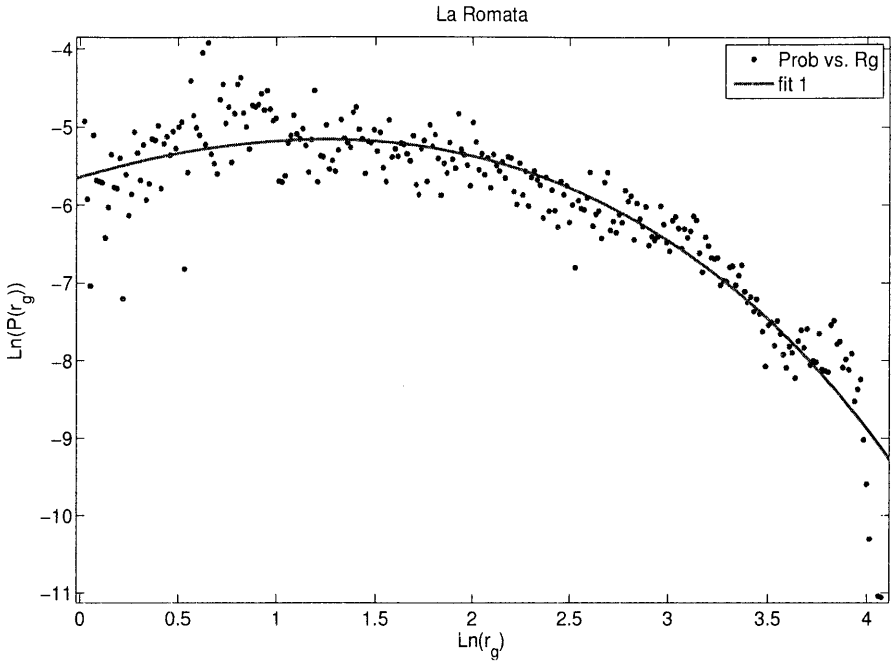
a powerful tool in terms of reducing the dimensionality of sample space. By applying PCA, we project the 4 parameters getting from fitted distribution of radius of gyration of each autonomous area into orthonormal vectors, and find the principal component that contributes most to the variation of data. A multi-variate linear regression associates the principal component with the population information in 24 autonomous region. The resulted R^2 is not enough good, which indicates that the regression function cannot be used for prediction purposes. The failure of using PCA to characterize the relationship between population density and radius of gyration maybe due to the fact that too few sample points are available to obtain a significant regression result. Furthermore, the resulted low R^2 excludes us from using a regression formula for prediction purposes even if the regression result might be significant. Generally, we think although there do exist correlation between population and density, the correlation is not strong enough to provide us an exact function relationship. It seems that other socio-economic factors should also be considered, and more complicated function form is necessary. Due to the limitation of data sources, we could not conduct further analysis.

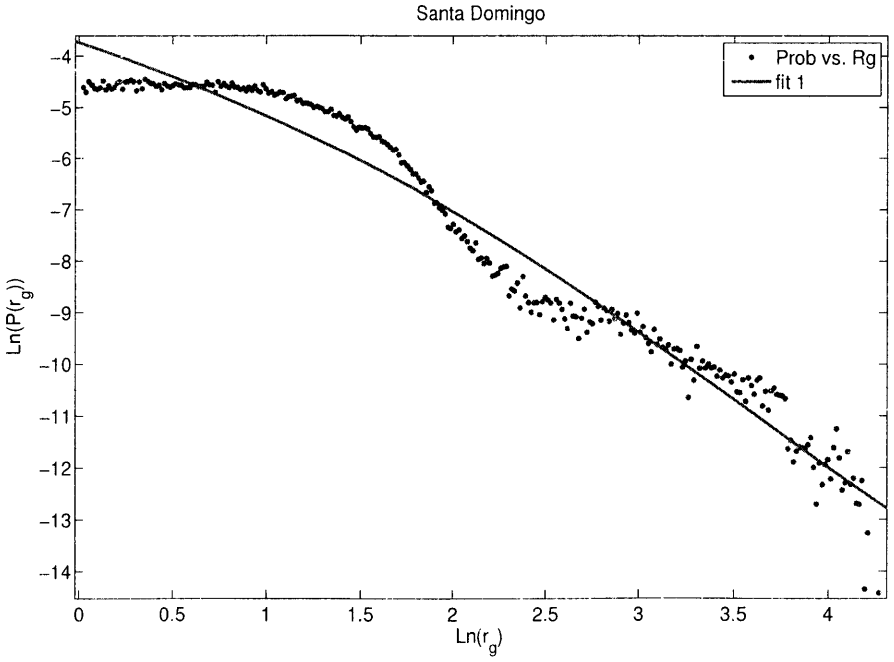
Finally, a finally as a case study we consider the San Francisco Bay is conducted in San Francisco Bay area. By dividing the whole Bay area into several polygons according to tower locations, we are able to analyze the data at a micro-scale (at the resolution of towers). Regressing radius of gyration on population density, median family income, unemployment rate and median age, we find a negative correlation between radius of gyration and population density and a positive correlation between radius of gyration and income. This implies that people living in wealthier and unpopulated areas tend to travel more frequently and make long distance trips. We are also particularly interested in how the distribution of radius of gyration depends on demographic information, namely, how the parameters in Eq. (4.1) related to population density, income, age, and unemployment rate, etc. We separated the 951 towers in San Francisco Bay area into 9 groups according to local income level and population density. We found that the area with low level of income has a higher probability to achieve larger r_g than the area with mediate or high level of income.

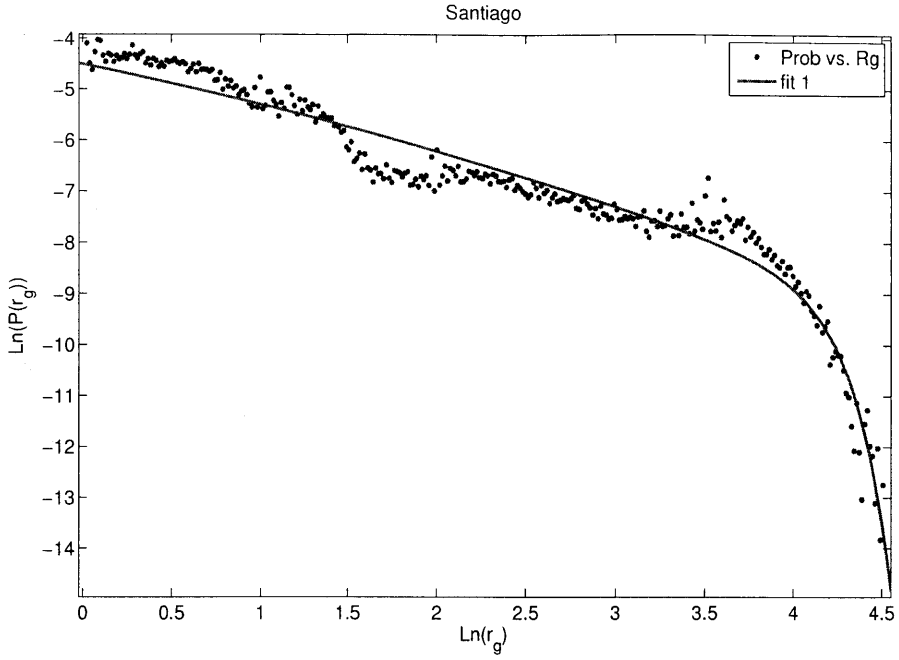
This is very reasonable since areas with low level income are most likely to be less developed, which don't have sufficient facilities or public goods to satisfy the daily demand of individuals. Therefore, people need to go to other regions for shopping or whatever they need, which in turn increases the probability of larger r_gs.

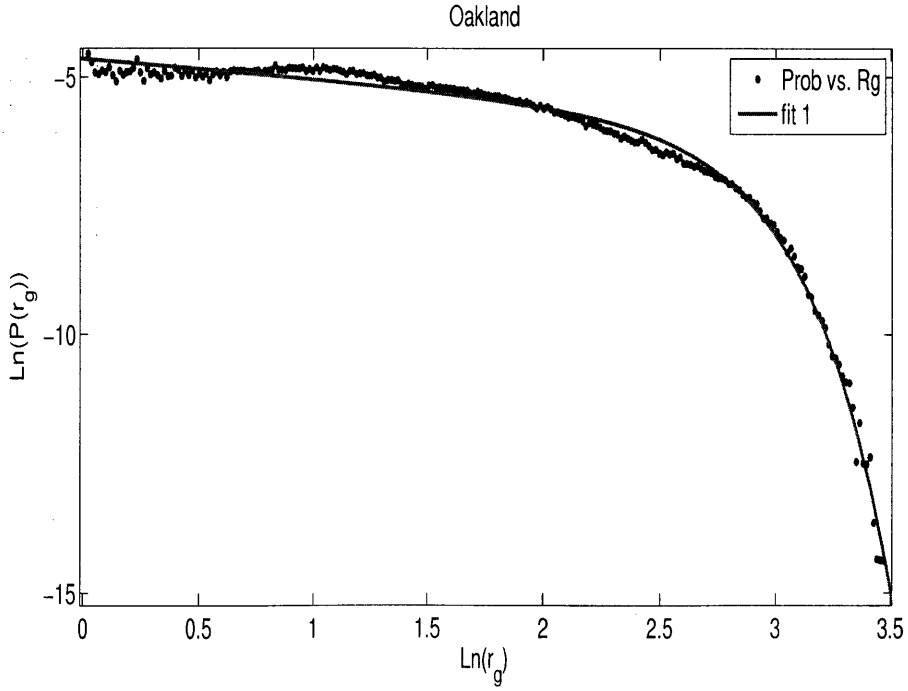
In this thesis, we presented a cross-cultural study of human's trip length distribution by considering billions of data points of time and space from tens of millions of mobile phone subscribers in regions ranging from rural Dominican Republic to urban California. Similarities of trip length distribution in three countries, i.e. United States, Dominican Republic, and a European Country are captured by a double exponential function. The research described in this thesis also open opportunities to use mobile phone data to detect commuting. Since traditional travel preference surveys are expensive and time-consuming, the analysis of mobile phone records presented in this thesis also provides an efficient method to collect trips among locations and estimate origin-destination matrix. Furthermore, this thesis is a step further in the contribution to numerically analyze and understand mobile phone records, which can be used for public services.

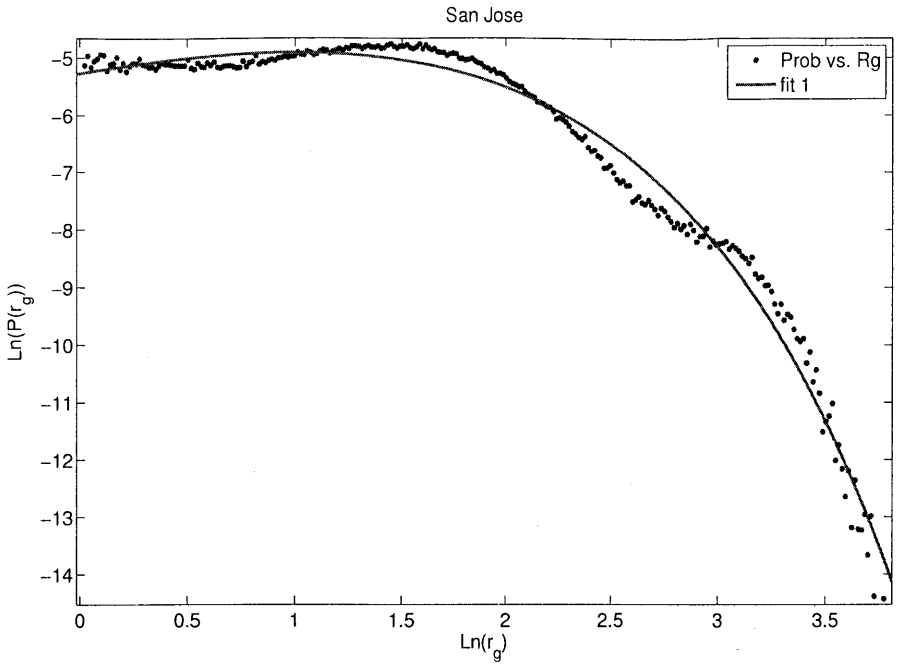
Appendix A

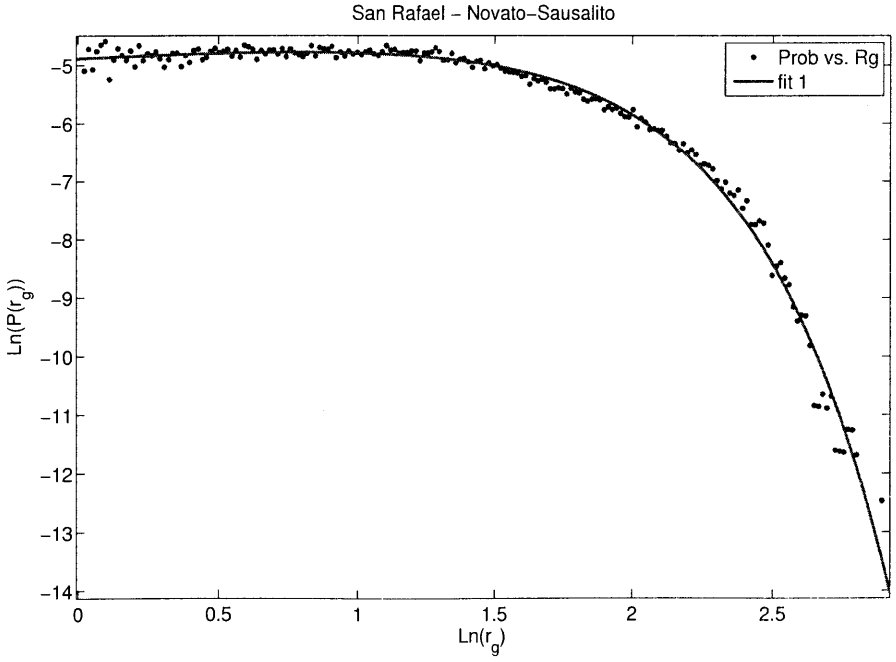

La Romata
Prob vs. Rg
fit 1
Ln(P(r_g))
Ln(r_g)
-4
-5
-6
-7
-8
-9
-10
-11
0
0.5
1
1.5
2
2.5
3
3.5
4

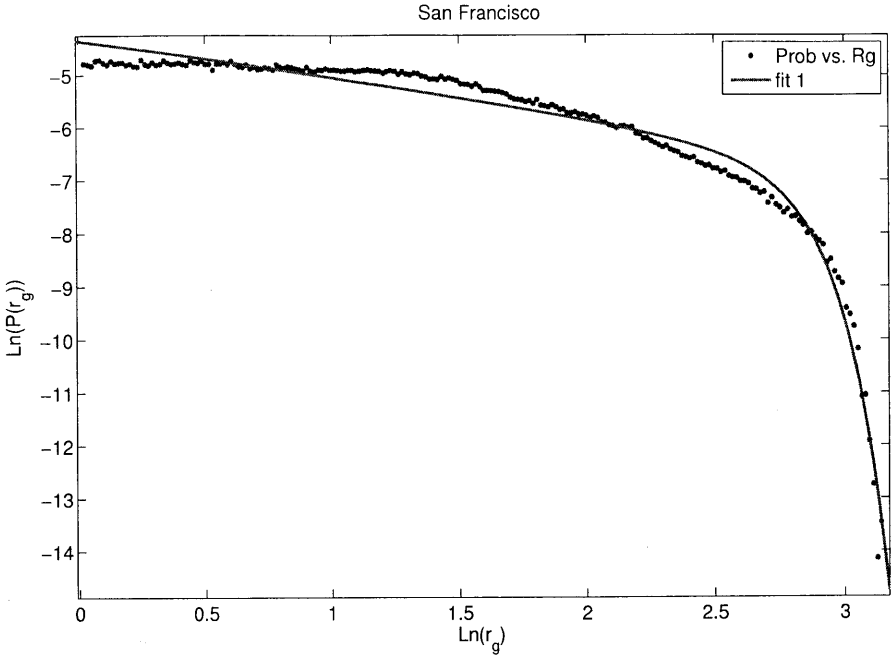

Santa Domingo
Prob vs. Rg
fit 1
Ln(P(r_g))
Ln(r_g)
-4
-5
-6
-7
-8
-9
-10
-11
-12
-13
-14
0
0.5
1
1.5
2
2.5
3
3.5
4

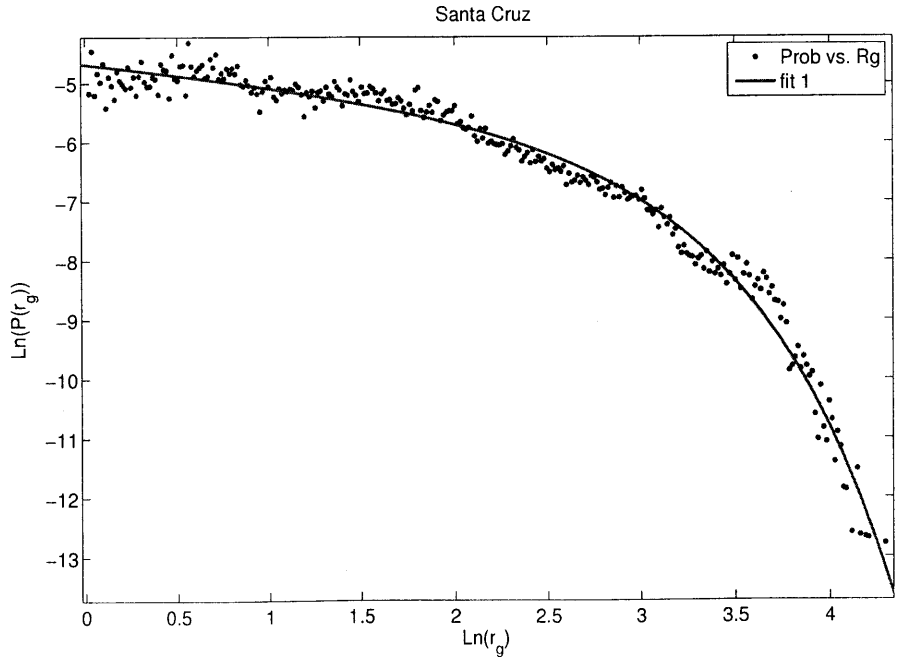

Santiago
Prob vs. Rg
fit 1
Ln(P(r_g))
Ln(r_g)
-4
-5
-6
-7
-8
-9
-10
-11
-12
-13
-14
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5

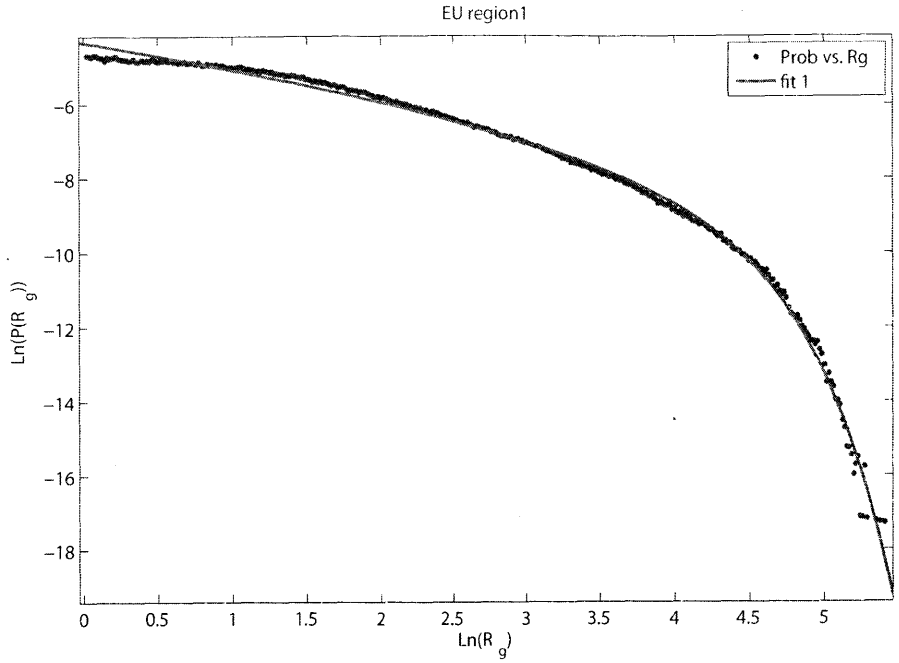

Oakland
Prob vs. Rg
fit 1
Ln(P(r_g))
Ln(r_g)
-5
-10
-15
0
0.5
1
1.5
2
2.5
3
3.5

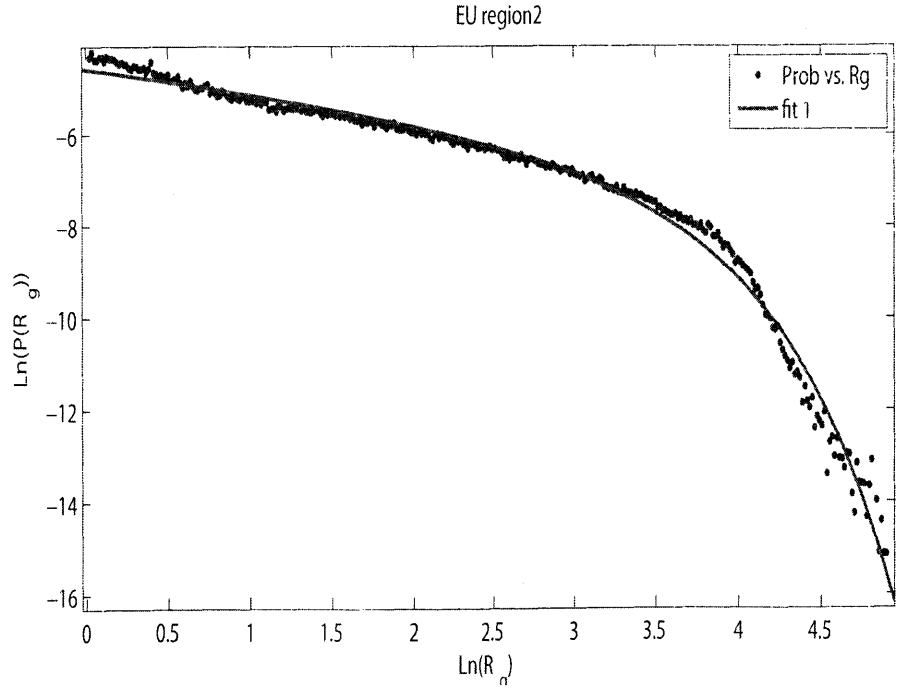

San Jose
Prob vs. Rg
fit 1
Ln(P(r_g))
Ln(r_g)
-5
-6
-7
-8
-9
-10
-11
-12
-13
-14
0
0.5
1
1.5
2
2.5
3
3.5

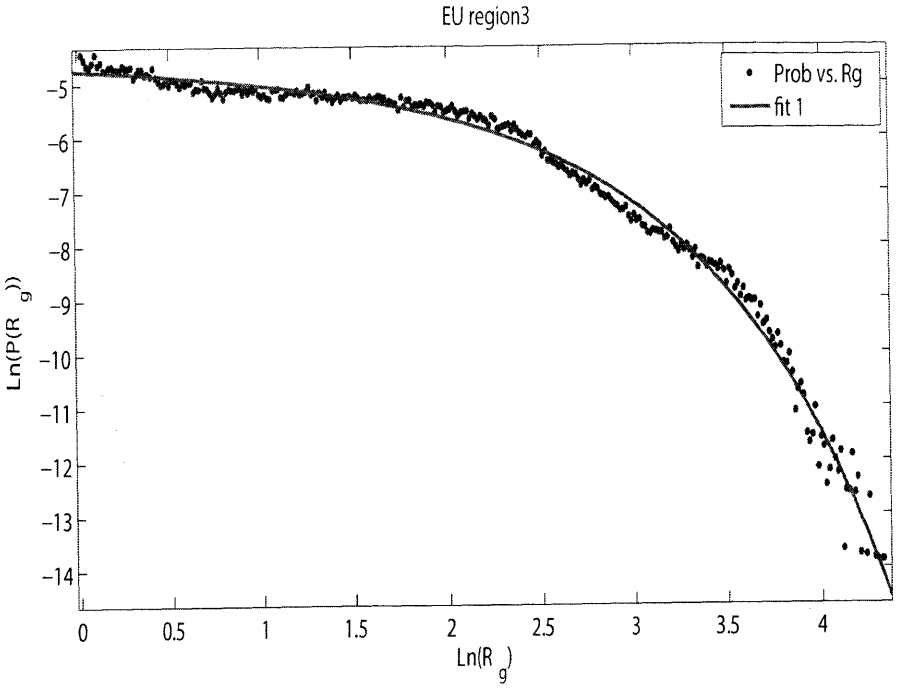

San Rafael – Novato–Sausalito
Prob vs. Rg
fit 1
Ln(P(r_g))
Ln(r_g)
-5
-6
-7
-8
-9
-10
-11
-12
-13
-14
0
0.5
1
1.5
2
2.5

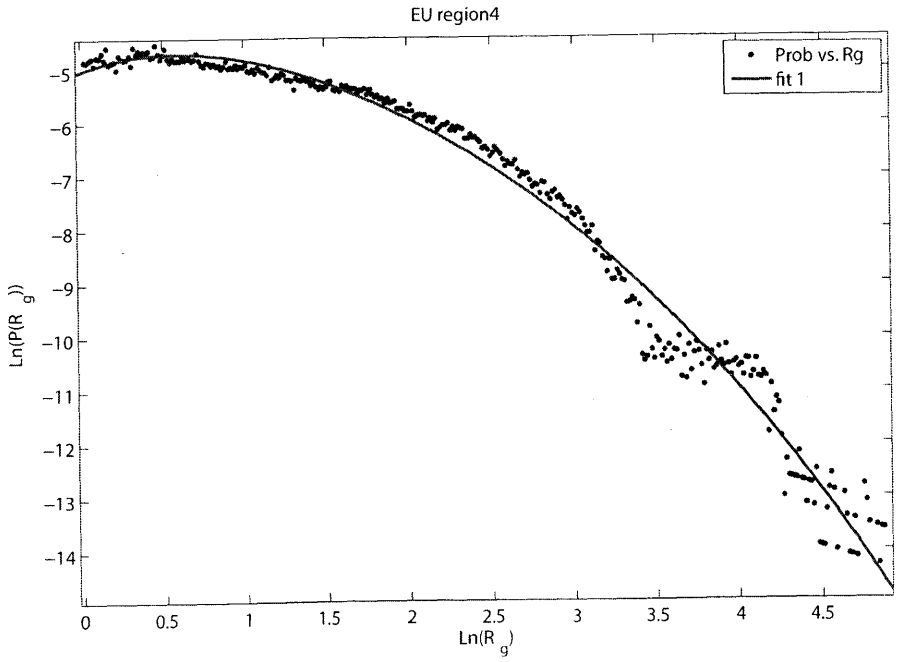

San Francisco
Prob vs. Rg
fit 1
Ln(P(r_g))
Ln(r_g)
-5
-6
-7
-8
-9
-10
-11
-12
-13
-14
0
0.5
1
1.5
2
2.5
3

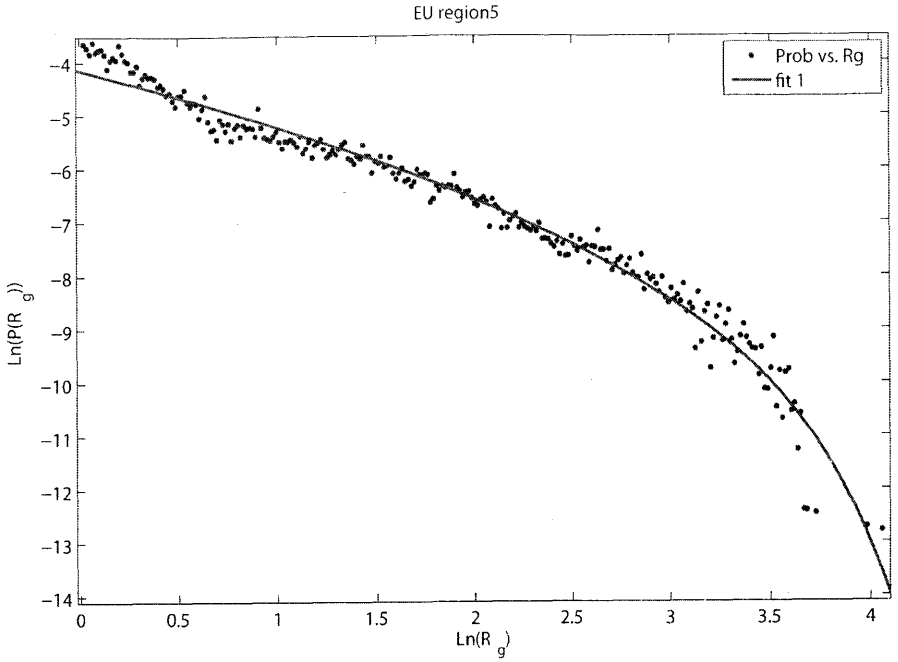

Santa Cruz
Prob vs. Rg
fit 1
Ln(P(r_g))
Ln(r_g)
-5
-6
-7
-8
-9
-10
-11
-12
-13
0
0.5
1
1.5
2
2.5
3
3.5
4

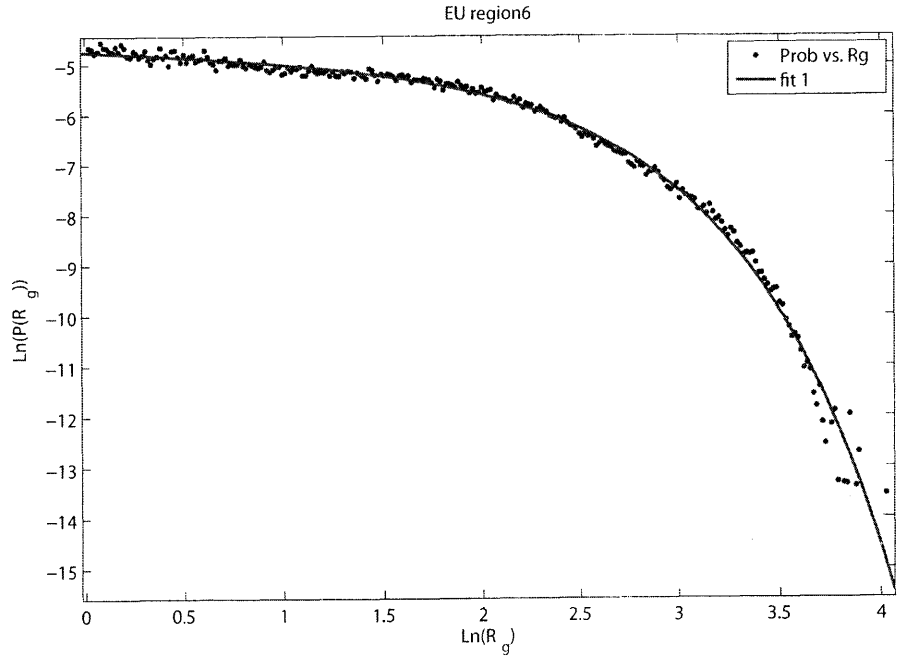

EU region1
Prob vs. Rg
fit 1
Ln(P(R g))
Ln(R g)
-6
-8
-10
-12
-14
-16
-18
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

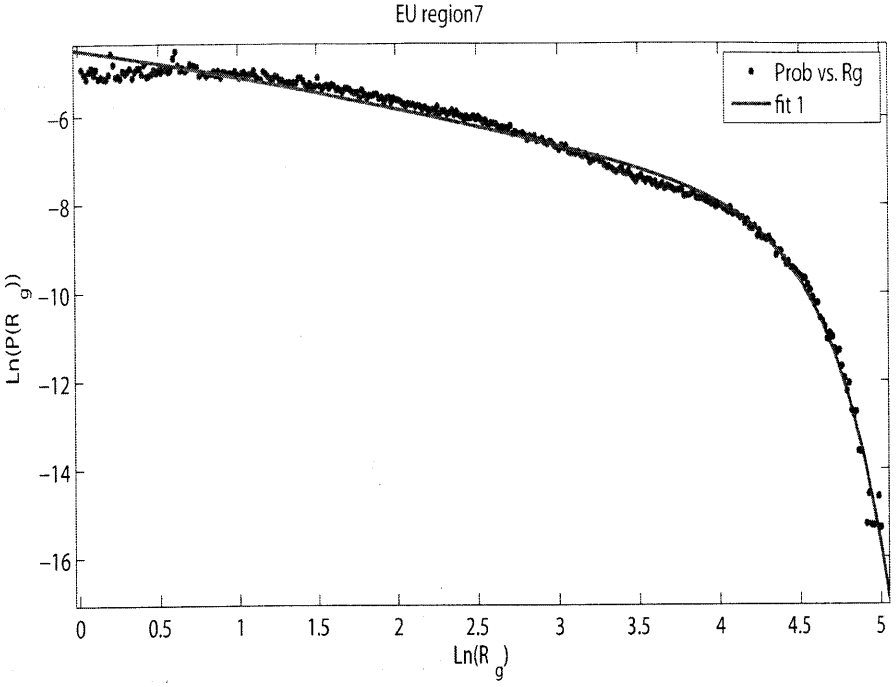

EU region2
Prob vs. Rg
fit 1
Ln(P(R g))
Ln(R g)
-6
-8
-10
-12
-14
-16
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5

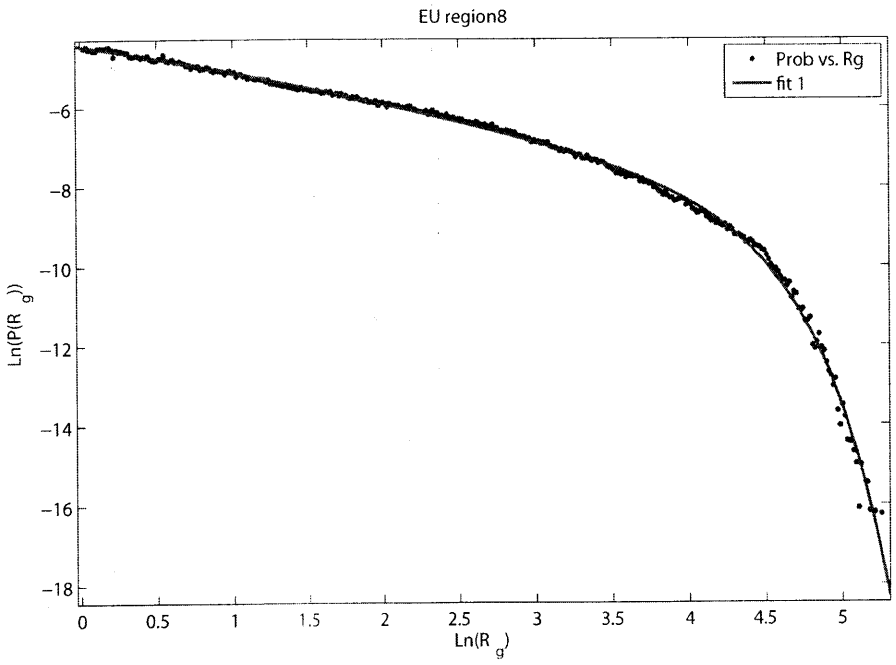

EU region3
Prob vs. Rg
fit 1
Ln(P(R g))
Ln(R g)
−5
−6
−7
−8
−9
−10
−11
−12
−13
−14
0
0.5
1
1.5
2
2.5
3
3.5
4

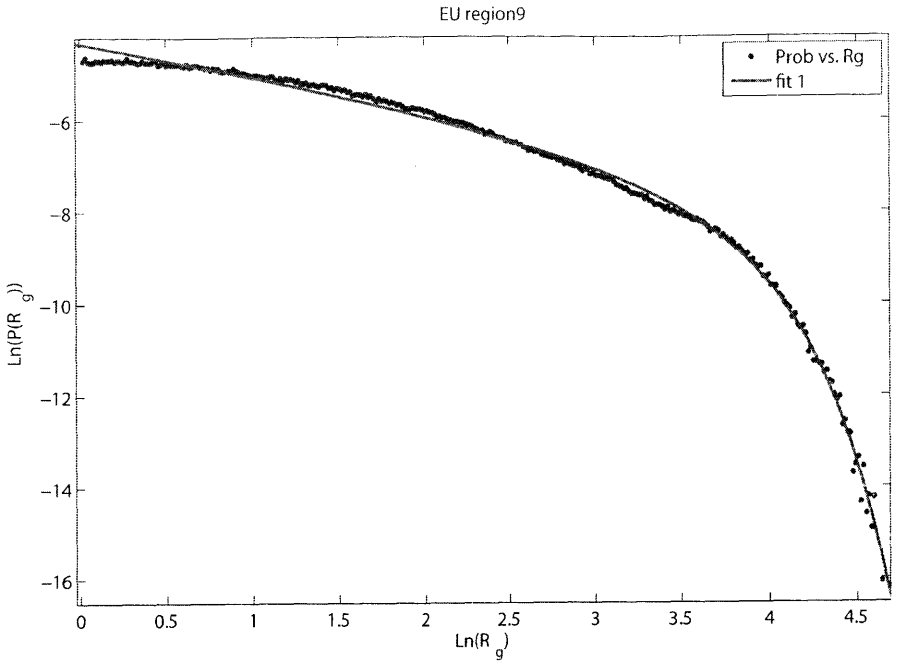

EU region4
Prob vs. Rg
fit 1
Ln(P(R g))
Ln(R g)
−5
−6
−7
−8
−9
−10
−11
−12
−13
−14
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5

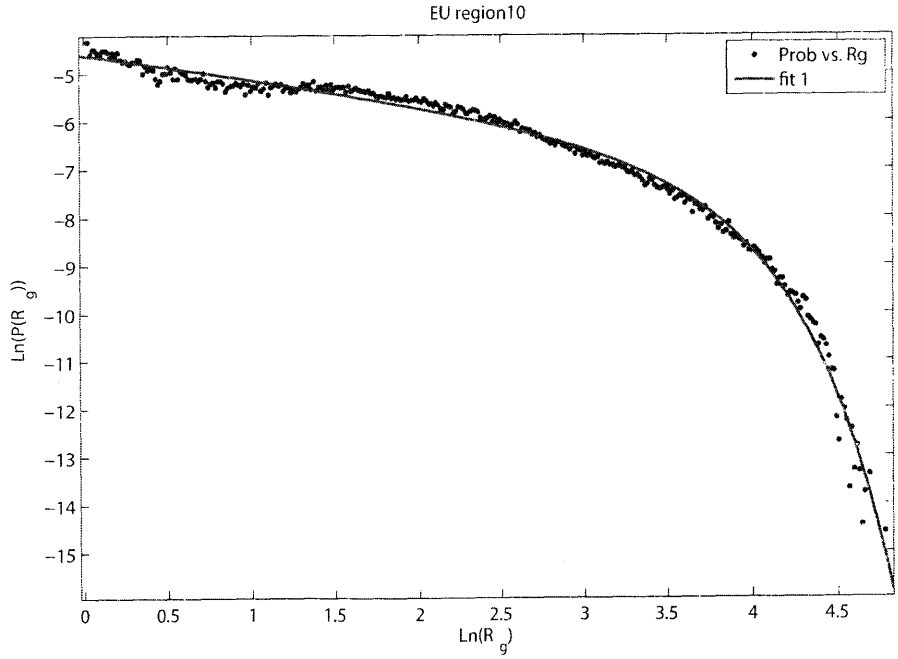

EU region5
Prob vs. Rg
fit 1
Ln(P(R g))
Ln(R g)
−4
−5
−6
−7
−8
−9
−10
−11
−12
−13
−14
0
0.5
1
1.5
2
2.5
3
3.5
4

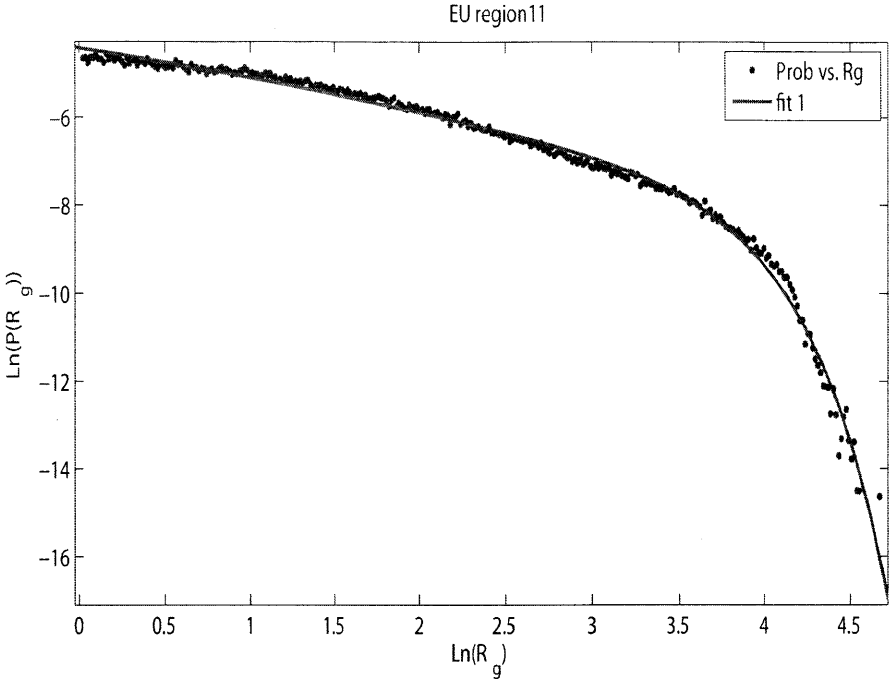

EU region6
Prob vs. Rg
fit 1
Ln(P(R g))
Ln(R g)
−5
−6
−7
−8
−9
−10
−11
−12
−13
−14
−15
0
0.5
1
1.5
2
2.5
3
3.5
4

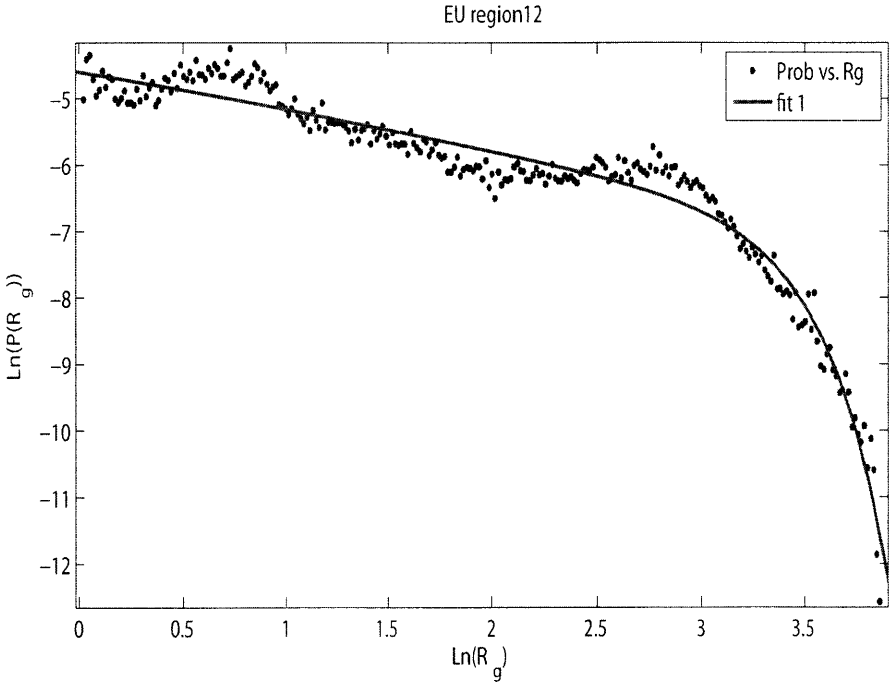

EU region7
Prob vs. Rg
fit 1
Ln(P(R g))
Ln(R g)
-6
-8
-10
-12
-14
-16
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

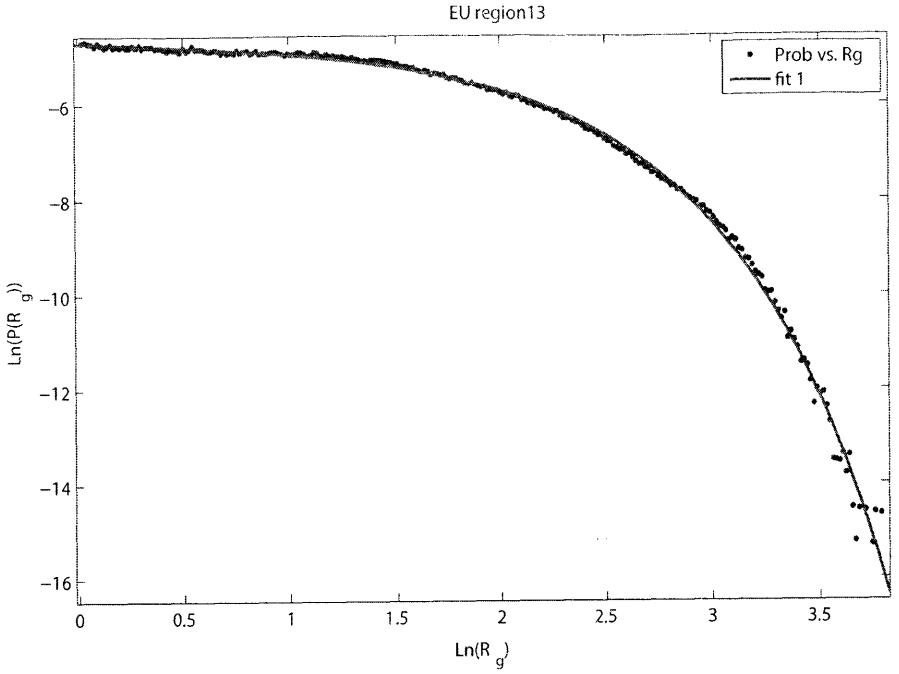

EU region8
Prob vs. Rg
fit 1
Ln(P(R g))
Ln(R g)
-6
-8
-10
-12
-14
-16
-18
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

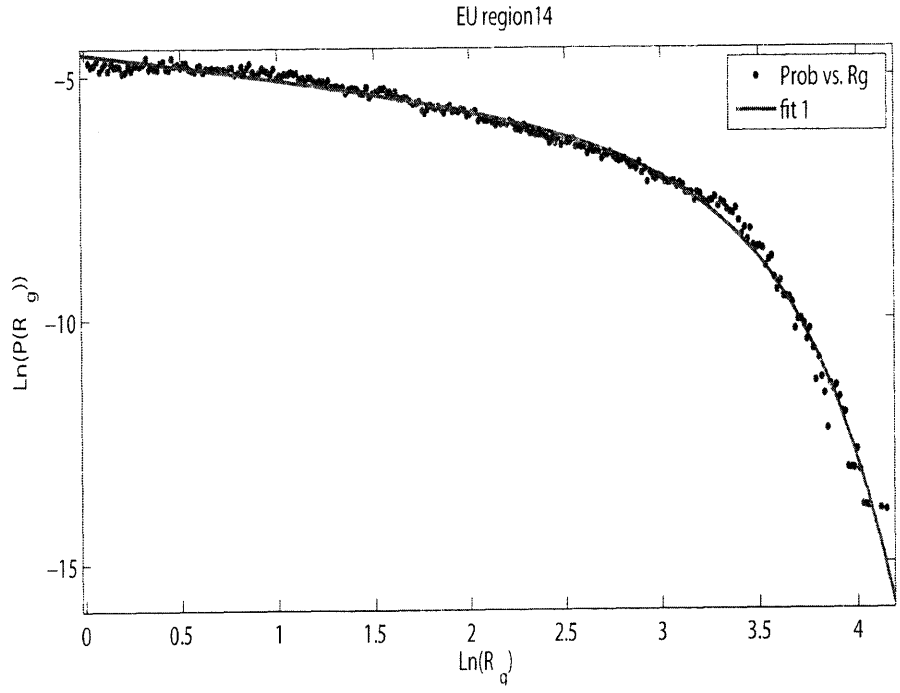

EU region9
Prob vs. Rg
fit 1
Ln(P(R g))
Ln(R g)
-6
-8
-10
-12
-14
-16
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5

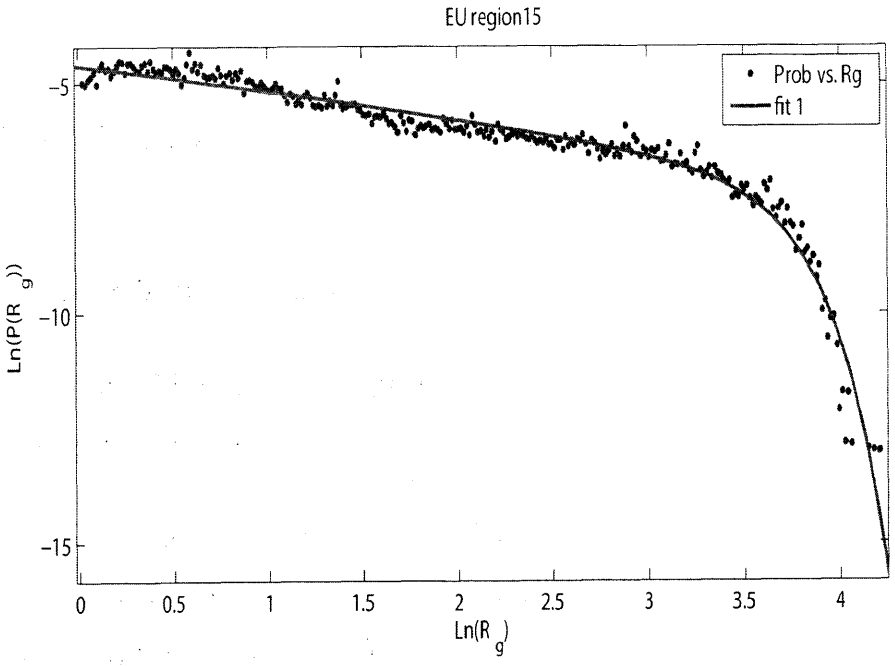

EU region10
Prob vs. Rg
fit 1
Ln(P(R g))
Ln(R g)
-5
-6
-7
-8
-9
-10
-11
-12
-13
-14
-15
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5

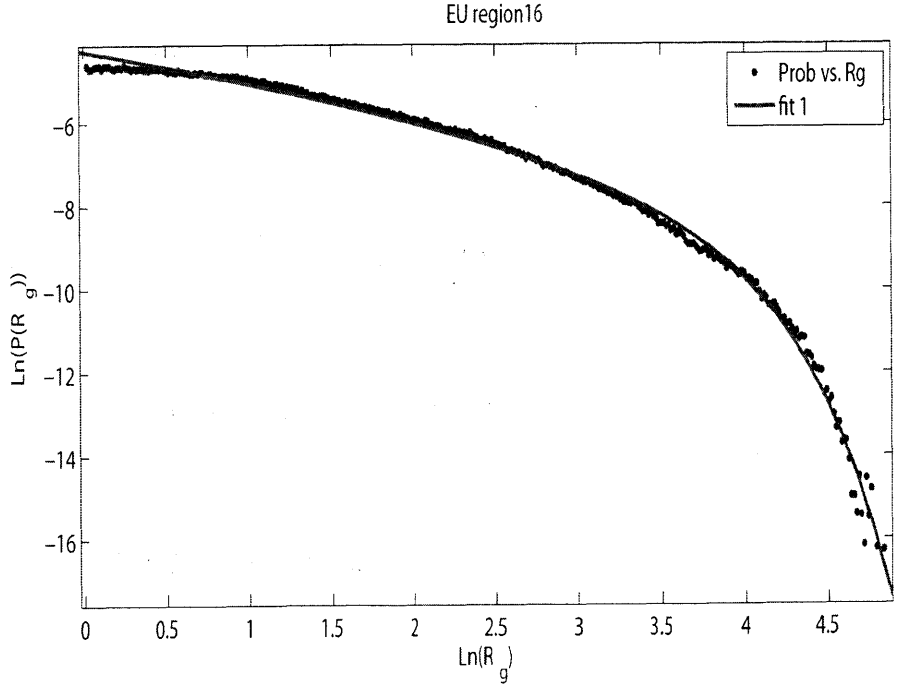

EU region11
Prob vs. Rg
fit 1
Ln(P(R g))
Ln(R g)
-6
-8
-10
-12
-14
-16
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5


EU region12
Prob vs. Rg
fit 1
Ln(P(R g))
Ln(R g)
-5
-6
-7
-8
-9
-10
-11
-12
0
0.5
1
1.5
2
2.5
3
3.5


EU region13
Prob vs. Rg
fit 1
Ln(P(R g))
Ln(R g)
-6
-8
-10
-12
-14
-16
0
0.5
1
1.5
2
2.5
3
3.5


EU region14
Prob vs. Rg
fit 1
Ln(P(R g))
Ln(R g)
-5
-10
-15
0
0.5
1
1.5
2
2.5
3
3.5
4


EU region15
Prob vs. Rg
fit 1
-5
-10
-15
Ln(P(R g))
0
0.5
1
1.5
2
2.5
3
3.5
4
Ln(R g)


EU region16
Prob vs. Rg
fit 1
-6
-8
-10
-12
-14
-16
Ln(P(R g))
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
Ln(R g)

Bibliography

[1] J. Candia, M.C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási, Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical, 41, 224015.

[2] http://eprom.mit.edu/whyafrica.html

[3] M.C. González, C.A. Hidalgo A.-L. Barabási, Understanding individual human mobility patterns. Nature, 453, 479-482 (2008).

[4] R. Lambiotte, V. D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren, Geographical dispersal of mobile communication networks, Physica A, 387, 5317-5325 (2008).

[5] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A.-L. Barabási, Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Science, 104, 18, 7332-7336 (2007).

[6] Pan Hui and Jon Crowcroft, Human Mobility Models and Opportunistic Communication System Design, Royal Society Philosophical Transactions B, 366, 1872 (2008).

[7] G. M. Viswanathan, V. Afanasyev, S. V. Buldyrev, E. J. Murphy, P. A. Prince, H. E. Stanley, Levy Flight Search Patterns of Wandering Albatrosses. Nature, 381, 413-415 (1996).

[8] Deer. A. M. Edwards, R. A. Phillips, et. al, Revisiting Lévy Flight Search Patterns of Wandering Albatrosses, Bumblebees, and Deer. Nature, 449, 1044-1048 (2007).

[9] D. Brockmann, L. Hufnagel, T. Geisel, The scaling laws of human travel. Nature, 439, 462-465 (2006).

[10] LandScan website: http://www.ornl.gov/sci/landscan/index.shtml

[11] M. C. Gonzalez, Trip Length Distribution Under Multiplicative Spatial Models of Supply and Demand: Theory and Sensitivity Analysis.

[12] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. A. de Menezes, K. Kaski, A.-L. Barabási, J. Kertész, Analysis of a large-scale weighted network of one-to-one human communication. New Journal of Physics, 9, 179 (2007).

[13] G. Szabó, A.-L. Barabási, Network effects in service usage. preprint physics/0611177 (2006).

[14] C. Ratti, R. M. Pulselli, S. Williams, D. Frenchman, Mobile Landscapes: using location data from cell phones for urban analysis. Environment and Planning B: Planning and Design, 33, 727-748 (2006).

[15] C. Ratti, A. Sevtsuk, S. Huang, R. Pailer, Location based services and telecartography section V, p 433. Berlin: Springer.

[16] G. Palla, A.-L. Barabási, T. Vicsek, Quantifying social group evolution. Nature, 446, 663-667 (2007).

[17] G. Palla, A.-L. Barabási, T. Vicsek, Community dynamics in social networks. Fluctuation and Noise Letters, 7, 3, 273-287 (2007).

[18] M. W. Horner and M. E. O Kelly, Embedding economies of scale concepts for hub networks design. Journal of Transport Geography, 9, 255C265 (2001).

[19] R. Kitamura, C. Chen, R.M. Pendyala, and R. Narayaran, Micro-simulation of daily activity-travelpatterns fortraveldemandforecasting, Transportation 27, 25C51 (2000).

[20] V. Colizza, A. Barrat, M. Barthélémy, A. J. Valleron, and A. Vespignani, Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions, PLoS Medicine 4, 95C110 (2007).

[21] S. Eubank, et al, Controlling epidemics in realistic urban social networks, Nature 429, 180C184 (2004).

[22] L. Hufnagel, D. Brockmann, T. Geisel, Forecast and control of epidemics in a globalized world, Proceedings of the National Academy of Sciences, USA 101, 15124C15129 (2004).

[23] J. Kleinberg, The wireless epidemic, Nature 449, 287C288 (2007).

[24] D. D. Brockmann, L. Hufnagel and T. Geisel, The scaling laws of human travel, Nature 439, 462C465 (2006).

[25] S. Havlin, D. Ben-Avraham, Diffusion in disordered media, Advances in Physics, 51, 187C292 (2002).

[26] G. Ramos-Fernandez, Lévy walk patterns in the foraging movements of spider monkeys (Ateles geoffroyi), Behav. Ecol. Sociobiol. 273, 1743C1750 (2004).

[27] D. W. Sims, Scaling laws of marine predator search behaviour. Nature 451, 1098C1102 (2008).

[28] J.Klafter, M. F. Shlesinger and G. Zumofen, Beyond brownian motion. Phys. Today 49, 33C39 (1996).

[29] R. N. Mantegna, H. E. Stanley, Stochastic process with ultraslow convergence to a gaussian: the truncated Levy flight. Phys. Rev. Lett. 73, 2946C2949 (1994).

[30] T. Sohn, et al. in Proc. 8th Int. Conf. UbiComp 2006 212C224 (Springer, Berlin, 2006).

[31] S. Redner, A Guide to First-Passage Processes (Cambridge Univ. Press, Cambridge, UK, 2001).

[32] S. Condamin, Bénichou, O., V. Tejedor, J. Klafter, First-passage times in complex scale-invariant media. Nature 450, 77C80 (2007).

[33] R. Schlich, K. W. Axhausen, Habitual travel behaviour: evidence from a six-week travel diary. Transportation 30, 13C36 (2003).

[34] N. Eagle, A. Pentland, Eigenbehaviours: identifying structure in routine. Behav. Ecol. Sociobiol. (in the press).

[35] S.-H. Yook, H. Jeong, A. L. Barabási, Modeling the Internet's large-scale topology. Proc. Natl Acad. Sci. USA 99, 13382C13386 (2002).

[36] G. Caldarelli, Scale-Free Networks: Complex Webs in Nature and Technology. (Oxford Univ. Press, New York, 2007).

[37] S. N. Dorogovtsev, J. F. F. Mendes, Evolution of Networks: From Biological Nets to the Internet and WWW (Oxford Univ. Press, New York, 2003).

[38] A.-L. Barabási, The origin of bursts and heavy tails in human dynamics. Nature, 435, 207-211 (2005).

[39] C. Hidalgo, A.L. Barabási, Inter-event time of uncorrelated and seasonal systems. Physica A, 369, 877-883 (2007).

[40] H. Goldstein, Classical Mechanics. Addison-Wesley (1959).

[41] http://www.caliper.com/tcovu.htm