COUNTRYWIDE ANOMALOUS INCIDENT DETECTION FROM AGGREGATED MOBILE PHONE DATA

by

Didem Gündoğdu

B.S., Computer Engineering, Yıldız Technical University, 1998

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Science

Boğaziçi University

2014

ACKNOWLEDGEMENTS

I am dedicating my thesis to my beloved brother, cousin Tahir Çanakcı, who kept believing me in every step of my prolonged study. As his saying, *"physical being, is just an illusion"*, I can not find such words, full of wisdom, but I do try my best...

Ozlem Durmaz Incel is not only my thesis supervisor, but also a friend, who helped me to get used to academic world. Without her patience, I would get lost in that unknown place. If I can manage to finish my PhD. one day, it will be because of her encouragement and mentorship. Albert Ali Salah, my thesis supervisor, I am really grateful for his endless support for improving my writing skills. If you can read this thesis, it is because of him. He introduced "The Element of Style" from William Strunk with me, and teach me the importance of expressing thoughts, in written form. I do not know how to thank Lale Akarun. She is the person who accepted me to Bogazici University 16 years ago. I will be grateful to her all my life. Ali Taylan Cemgil, magician, I will probably spend rest of my life to understand the world of probability, as he showed us.

To my lovely parents Nese and Gurol Gundogdu, without your endless support I would not finish that long run. I am proud to be your daughter, as you make me believe in me, what ever it takes.

Cigdem Patlak, I am glad that you motivate me to start my master degree back again, thank you so much. I am grateful to PI Lab and Net Lab, as they share their room, knowledge and joy with me. Binnur Görer and Okan Aşık, my great friends. I am so lucky that I met you, I know that your bright souls will bring you bright academic successes, even the robots have not won the Robocup yet... :) Çağıl and Ahter Uluşahin, I am grateful your companion, energy and hope you bring. Onur Güngör, Umut Şimşekli, Barış Evrim Demiröz, Ahmet Alp Kındıroğlu, Barış Kurt, Deniz Akyildiz, you make Machine Learning and all the mathematics behind it more comprehensible to me. Without your support, my rusty brain would not understand the chaotic

probability world, now I slightly got the idea. Bilgin Koşucu, yoga master, life guru; thanks for the small talks we have, as if we understand the world in some sense. Hande Alemdar thanks for your mentorship, and typographical assistance. Huseyin Demirtas, thanks for helping to access, Tosun-1 and Tosun-2, to run my scripts. MustafaTuğrul Özşahin, you are the best course analyst ever, you advice me to take Monte Carlo, and drop just after my enrollment. Although I am still trying to understand the content of the course, it is a milestone for me. I am glad that you persuade me. Finally all my friends, who do not give up being my friends, although I neglect you for a while. I am so lucky that I have you.

ABSTRACT

COUNTRYWIDE ANOMALOUS INCIDENT DETECTION FROM AGGREGATED MOBILE PHONE DATA

Mobile phones are extensively used both in developed and developing countries. Richer countries tend to use it as a computer, whereas in developing countries it is a kind of replacement of inadequacy of infrastructure, or it is used to cover more crucial needs. This work is an explorative analysis of the nationwide Call Detail Record (CDR) in Cote d'Ivoire. Our aim is to detect anomalous incidents and to explore the possibility of early detection of severe incidents from mobile phone data. Beside this, this work is a kind of roadmap who would like to work on CDR data, from obtaining open data set to visualization tool selection. Data is collected and anonimized in Cote d'Ivoire by Orange Telecom, from real call data. It has been published as a part of a scientific challenge, Data for Development (D4D). We explored irregularity of phone usage by probabilistic Markov modulated Poisson process (MMPP) method. During the data collection period (from December 2011 to April 2012) Cote d'Ivoire was suffering from civil war, which caused hundreds of deaths and many injuries. Validation of the experiments has been done through United Nations Security Council's report covering these dates.

ÖZET

CEP TELEFONU VERİLERİNDEN OLAĞAN DIŞI OLAYLARIN ÜLKE ÇAPINDA TESPİTİ

Günümüzde cep telefonları gelişmiş ve gelişmekte olan ülkelerde yaygın olarak kullanılmaktadır. Gelir seviyesi yüksek olan ülkelerde telefon; bilgisayar işlevselliği, gelişmekte olan ülkelerde ise, daha çok alt yapıdan eksikleri yada daha hayati ihtiyaçları karşılamak amacıyla kullanılmaktadır. Bu çalışmada Fildişi Sahilleri'nin ülke genelinde toplanmış olan, cep telefonu konuşma kayıtları analiz edilip kullanılmıştır. Olağan dışı önemli olayların ülke boyutunda tespit edilebilmesi ve olay gerçekleşmeden tespit edilebilirliği nin, cep telefonu verisi kullanılarak mümkün olup olmadığı hedeflenmiştir. Bunun yanı sıra, bu çalışmanın cep telefonu verisi kullanarak analiz yapacak diğer çalışmalara, veri seti temininden görselleştirmeye kadar, bir yol haritası olması hedeflenmiştir. Bu çalışmada kullanılan veri seti, Orange Telekomunikasyon tarafından, Fildişi Sahillerinde, Aralık 2011-Nisan 2012 tarihleri arasında toplanılmış ve anonimize edilmiştir, aynı zamanda bu veri Data for Development (D4D) akademik yarışmasının bir parçası olarak paylaşıldı. Çalışmamızda Markov tabanlı Poisson süreci (MMPP) methodu kullanıldı. Verinin toplanıldığı sırada ülke, yüzlerce ölü ve yaralı ile sonuçlanan bir iç savaş içerisindeydi. Bu kargaşadan kaynaklanan olayları Birleşmiş Milletlerin Güvenlik Konseyi raporlarından elde ederek, çalışmamızda tespit edilen olayların başarısı değerlendirildi.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

A	Antenna Set
a_j	Antenna number in time t_j
a^E	Gamma Shape Hyperparameter, for event probability
b^E	Gamma Scale Hyperparameter, for event probability
N	Total number of users in the dataset
$N(t)$	Number of call
$N_0(t)$	Normal day call volume
$N_E(t)$	Anomaly day call volume
U	Set of mobile phone users
u_i	User i
t_j	Time
s	State
X	Random variables
$z(t)$	Transition Probability
$\delta_{d(t)}$	Week of day effect
$\eta_{d(t),h(t)}$	Hour of day effect for given day
λ	Poisson distribution rate parameter
λ_0	Poisson distribution initial rate parameter
$\mu(\rho, t)$	Mobility, number of distinct antenna in ρ hours, for time t
Ω	Longitude
ψ	Mobility diameter
ϕ	Latitude
$\tau(\rho, t)$	Talkativity, number of outgoing calls in ρ hours, for time t
τ	Threshold value

LIST OF ACRONYMS/ABBREVIATIONS

3D	Three Dimensional
ATM	Author Topic Model
AUC	Area Under the ROC Curve
CIV	Côte d'Ivoire
CDR	Call Detail Record
CSV	Comma Separated Values
CF	Collaborative Filtering
CLAF	Collaborative Location and Activity Filtering
D4D	Data For Development
DAG	Directed Acyclic Graph
DBSCAN	Density Based Spatial Clustering of Application with Noise
EM	Expectation Maximization
HMM	Hidden Markov Models
IFAD	International Fund for Agricultural Development
GIS	Geographic Information Systems
GPS	Global Positioning System
KL	Kullback Leibler Divergence
LDA	Latent Dirichlet Allocation
LSI	Latent Semantic Indexing
MCMC	Markov Chain Monte Carlo
MMPP	Markov Modulated Poisson Process
ML	Maximum Likelihood
NMC	Nokia Mobile Challenge
OCHA	United Nations Office for the Coordination of Humanitarian Affairs
OD	Origin Destination
OPHI	Oxford Poverty and Human Development Initiative
PCA	Principal Component Analysis
pLSI	Probabilistic Latent Semantic Indexing

POI	Point of Interest
RM	Reality Mining Dataset
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
SMS	Short Messaging Services
UN	United Nations
UNHCR	United Nations Refugee Agency
UPGMA	Unweighted Pair Group Method Using Arithmetic Averages
WIPER	Wireless Phone Based Emergency Response

1. INTRODUCTION

Mobile phones have become extensions of our daily lives. 1990's huge, clumsy, heavy phones evolved into smart, tiny, fashionable phones, with high computational capabilities. They are enhanced with many sensor capabilities carried on at all times. With these mentioned abilities, these tools make it easier for scientists who aim to collect and analyze data to understand human behavior, ranging from location data to context data. There are vast amounts of application areas existing with the use of mobile phone data, from epidemiology to urban planning. In this work we specifically focus on event detection from mobile phone usage data collected in Côte d'Ivoire[1] [1]. The reason behind our motivation on detection of events is that, during the period of data collection, there was a civil war in Côte d'Ivoire and many violent events were reported [2].

There are different types of data that can be collected through mobile phones. Some of these data types are, the location of the user (via GPS or by logging call antennae), proximity of users (Bluetooth), social behavior of the user (indirectly, via internet and application usage) and even physical activities (by collecting accelerometer and gyroscope information over time). The variety in these data sources, and the information that can be gathered by combining these sources make it possible to predict the behavior of the users to a certain extent [3].

The collection of personal data creates privacy issues, and even if data are available, researchers cannot always access and use these sources. To promote explorative data analysis and to research what the potential of such applications are, researchers work with toy data sets of synthetic data, or seek real data collected under carefully controlled conditions. The later is tackled via data campaigns, for which ethical concerns and privacy issues are carefully monitored, and strict guidelines are observed. In these campaigns data are collected and made available to researchers. In many cases

[1] Côte d'Ivoire is adopted as the official name of the country, and the use of Ivory Coast is discouraged.

data are collected in university campuses, with students as subjects. Ethical committee approvals are obtained for collecting the data.

We use Data for Development Challenge (D4D) data set in our research, which consists of real life mobile phone data of Côte d'Ivoire collected from December 2011 to April 2012. This challenge is organized by Orange Telecommunication company, Université Catholique de Louvain (Belgium) and Massachusetts Institute of Technology (USA) to contribute novel solutions to develop better life conditions for Côte d'Ivoire. The challenge is categorized in several branches: *Social and Economical Development, Data Mining, Mobility/Transport, Health and Epidemics*, respectively.

This data set is one of the biggest open data sets, consisting of 500,000 users' call detail records (CDRs), from over 5 million users with 2.5 billion calls and short messaging services (SMS), for a duration of 3,600 hours. The data is divided into four subsets. The first subset Set_1, composed of hourly aggregated numbers of calls per antenna, is used for event detection. Two subsets Set_2 and Set_3 of D4D data set that we work on, have similar characteristics; 50,000 user with antenna location precision, 500,000 users with sub-prefecture precision. We use these two subsets for mobility and location based analysis and experiments. To have an insight into the daily life of Côte d'Ivoire, we classify calls, within the time period that the data set was collected. The locations correspond to daytime taken as office, night time period taken as home location. Under these assumptions only 30% of the population have a commuter life. The reason behind this unexpected figure is explained by socio-political chaos in Côte d'Ivoire during the data collection period.

We first analyze human movements, among the borders of the country to understand irregular patterns such as immigrants paths. These transition patterns are important in terms of identifying smuggling, and tracking infectious disease penetration.

In our study, we faced challenges in the analysis of data. By nature, real life CDR data rarely include ground truth, in order to ensure the privacy of the users. We

have to correlate our findings with other data sources, depending on the social aspects we wish to analyze. We have primarily used Oxford Poverty and Human Development Initiative (OPHI) or United Nation (UN) sources. UN Security Council's reports that cover the duration of the challenge have been taken as ground truth for evaluating our incident detection model [2]. The data collection period also includes several social events like the African Cup[2] and Christmas.

In the data analysis part, we use Markov modulated Poisson process (MMPP), because we model the number of occurrences of count data in a temporal sequence. Compared to existing studies [4, 5], our approach is novel in the sense that it is an adaptive methodology, applied to anomalous event, detection on country-wide scale CDR data [6].

Our contributions can be listed as follows;

- We analyze and visualize call data of an entire country over a large time window in Chapter 5 and 6.
- We generate violent and social event ground truth data from multiple data sources in Appendix C and Appendix D.
- We implement a Markov modulated Poisson process to model a real CDR data, as multiple correlated time-series in Chapter 5.
- We automatically detect anomalous and social events in Chapter 5.
- We visualize probable events on a country-wide scale spatial temporal video given as additional electronic material.

This thesis organized as follows: In Chapter 2 we first present our preliminary work, which motivated us to detect events through a mobile phone data set. Background information regarding geographic mobile data terminology, related works, major open data sets and information about the Côte d'Ivoire D4D Challenge are summarized. Application areas of mobile phone data set usage are also presented in Chapter 2. In Chapter 3, topic model implementation to understand movement patterns ob-

[2]http://en.wikipedia.org/wiki/2012_Africa_Cup_of_Nations

tained from CDR is visualized over maps. This chapter represents our initial work with the same data set, and can be treated independently from the following chapters. In Chapter 4 the methods for analysis, processing, and visualization of big mobile phone data sets are presented. Chapter 5 describes calculations of Markov Modulated Poisson Process parameter estimations and evaluation of spatio-temporal data. In Chapter 6, our experimental results, discussion and concluding remarks are given.

2. BACKGROUND

In this thesis we analyze mobile phone data, particularly CDR data, which includes information about the locations of the base stations serving the users. Before deciding our research question, *"Anomalous Event Detection"*, we attempt to answer different questions with mobile phone data. In Section 2.1, we summarized the paths, we walked through, and why we decide to focus on anomalous event detection.

To familiarize the reader about the properties of CDR data, in Section 2.2 first we describe the geolocation data types and then elaborate on the features of the Côte d'Ivoire D4D data set, which is used in the thesis. Related studies on usage of mobile phone data, particularly for event detection, are also discussed together with a taxonomy comparing their performances.

2.1. Motivation

Our first intention was to predict mobile phone user's age, from the call patterns and restricted mobility data obtained through antennae locations, given via CDRs. Levinson classified different behavior trends while aging [7], details are given in Appendix E.1. Our motivation was to improve the outcome of the psychological researches, which are mainly based on questionnaires or observations, and are quite subjective. Nokia Mobile Data Challenge data set includes age information of the phone users, however accessing to the data set is prolonged, as the owner of the data set has changed during that time [8]. We learned that being able to *access a rich data set is as important as having the right research question.*

The D4D data set, which is used in this thesis, is obtained from Orange Telecom. The details its subsets, as well as similar resources in the literature, are found in Section 2.3.1. D4D is collected from real phone usage of a whole country, Côte d'Ivoire, preventing access to any information regarding the phone user for privacy issues. The number of researches on mobile phone data sets are limited, but diverse, in terms of

application domains. The main problem of the D4D challenge is that, it does not have a clear task with ground truth annotation, therefore it is not easy to formulate a machine learning problem on this data set.

Each time we make a call, we leave a trace in the CDR. Farrahi and Gatica-Perez applied a document classification method, *Topic Model*, to classify human behaviors [9–11]. This approach is based on word counts in documents and groups the documents into "topics" based on similarity of computed probability distribution of words. Inspired from that work, we applied *topic model* into locations where the call takes place for each subject. In our approach, call locations are similar to words, and topics are different types of mobility patterns, and documents are call records of subjects. During the data collection period there was a civil war in Côte d'Ivoire, and thousands of people displaced within the country and migrated to neighboring counties. We would like to answer, whether the immigrants or displaced people's mobility patterns differ from a commuter's pattern. As shown in experiments in Chapter 3 almost 99% of the subjects tend to go back and forth to the same location, possibly travelling between their significant locations such as *home* or *work*. Since the outlier class is not apparent in these experiments, it is not possible to draw strong conclusions. Related work on that topic are presented under topic model methodology in Section 2.5.2.1, implementation with experiments in Sections 3.1 and 3.2.

This mobility information, generated through CDR, is used for modeling infectious disease penetration, such as HIV and Malaria [12–17]. US Census Bureau's, Measure Demographic and Health Surveys(DHS) surveys [3] , US Agency for International Development (USAID) give prevalence rate for HIV in Côte d'Ivoire in sub-prefecture level. These two data sets' subjects are independent from each other, and no correlation is present. This abstraction prevents us to evaluate mobility pattern generated from CDR data and HIV prevalence rate.

These preliminary works led us to search for possible applications in this domain. Ihler *et al.* proposed an event detection approach that uses count data of cars pass-

[3]Demographic and Health Surveys and Multiple Indicator in Côte d'Ivoire (EDS-MICS) 2011-2012

ing by a stadium and a campus door entrance [5]. Pawling *et al.*, inspired by their study, detected emergency events from a Wireless Phone Based Emergency Response (WIPER) system [18].We applied Markov modulated Poisson process to real life data of a country, inspired from these two works.

Regarding anomalous events, we compare our results with United States Security Reports, news and important events like African Football Championship through that period. Unveiling social incidents, or security problems through CDR data, may reduce impacts of the event, with fast action. In chaotic, poor and underdeveloped countries such observations may save lives, especially when the data sources may not be reliable or reachable.

2.2. Geolocation Data

Geolocation data identifies the coordinates of an object with latitude, longitude and altitude values. In geography, latitude is a geographic coordinate that specifies the north-south position of a point on the Earth's surface, whereas longitude is a coordinate that specifies the east-west position of a point on the Earth's surface. Altitude is the height from the sea level. We use only latitude and longitude values to identify a location in this work, because the D4D data set does not include this information, and a two dimensional representation is adequate for the type of analysis we conduct.

Global Positioning System (GPS) is a common sensor in smartphones, which provides the location information. However, the main drawback of this technology is the high energy consumption. This is one of the key challenges for data collection projects, or applications which process location data for inference. In our case, the data are obtained from the CDRs of a telecommunication operator. Location information of a user is represented as latitude and longitude values of the antenna, which services the call. The antenna locations in our data set are not very accurate. They are given with some added noise, and the exact locations of the antennae are considered to be a commercial secret of the company. The data are discrete, also have some missing and noisy features. GPS data accuracy may change according to environmental factors,

such as ionospheric conditions.

2.3. Dataset

Accessibility to mobile data sets are strictly controlled, because of data privacy issues. There are few open data sets for research purposes, as elaborated in Section 2.3.1. Usually research institutes initiate such data collection campaigns with a limited number of people, in the orders of hundreds. The data set that is utilized in this thesis, namely D4D, is one of the largest data sets, and it is collected from a large number of participants, in the order of thousands [1].

2.3.1. Mobile Data Collection Challenges

Before we describe D4D, we first give the details of some of the most widely used public data sets.

One of the most commonly used data sets is the *Reality Mining* data set [19]. It is the first data collection challenge in the literature which and was carried out in 2004. It contains Bluetooth data, call logs, cell tower identifiers, application usage, and phone status (such as charging and idle) information, collected from 100 users over 350,000 hours (approx. 40 years). Most participants are university students, who spend most of their time in the campus. This is one of the limitations of this data set, as it is not straightforward to generalize observations obtained from university students who spend most of their time in specific locations (library, classrooms, dorms, etc.) to a general public, with different occupations, working hours, etc.. Nonetheless, research on the Reality Mining data set proved that it is possible to classify people into meaningful groups (such as undergraduates vs. graduates, engineering students vs. management students) just by looking and comparing mobile phone usage data [20].

The Nokia Mobile Challenge (NMC) [8], is another open data set, which was collected in Lausanne (Switzerland) between 2009-2011, with 170 participants. It includes not only location and call logs, but also application usage, and behavioral data, gath-

ered with surveys for demographic attributes of the participants and manual tagging for frequently and infrequently visited places. Although data are anonymized, backward tracing GPS data may expose the identity of the subject. For the privacy of the subjects, NMC gives the flexibility to the subjects to exclude information belonging to themselves before, publicly available to researchers.

Social fMRI is a large data collection campaign, organized with the help of 130 adults, and their families, to understand the social interaction of individuals [21]. In this dataset, the overall observation period is one year, and in addition to the data coming from different sensors, surveys capture the psychological, economical and physical activity states of the participants. Accelerometer, GPS, Bluetooth and WIFI scans, call and SMS logs, application logs, and power states are the main data collected. This data set is partially available to the public [21].

Airsage is a company which offers anonymized mobility data, in various scales of granularity [22]. Data are provided by different types; geographic location from ZIP code to State level (USA), aggregated or averaged, including demographic, residence class (i.e. resident, visitor, commuter), trip types (i.e. car, bike etc.). Although they support academic researches with 20% discount, it is very expensive and the smallest package starts from $10,000.

One of the important questions in data collection campaigns is to decide the duration of the campaign. Depending on the application, the amount of data that needs to be collected differs in order to make reliable decisions. In the work of Altshuler *et al.* [23], the researchers focus on the amount of data required to make reliable analysis by using an open-ended approach, and collect data while there is improvement in the estimation of the target behavior. Once the estimation accuracy curve stabilizes, the data collection is stopped in their work. In this approach, the analysis is run in parallel to data collection.

2.3.2. Data for Development Dataset (D4D)

D4D is an open data challenge to encourage research teams to identify some of the existing problems of Côte d'Ivoire and to develop new solutions in order to improve the life conditions there [1].

There are around 20 million people living in Côte d'Ivoire, 5 million of that population is residing in Abidjan, former capital of the country. Yamoussoukro is the capital city which has around 800,000 inhabitants. 56% of the population is capable of reading and writing from CIA Factbook.[4] The distribution of the population can be viewed in Figure 2.1.

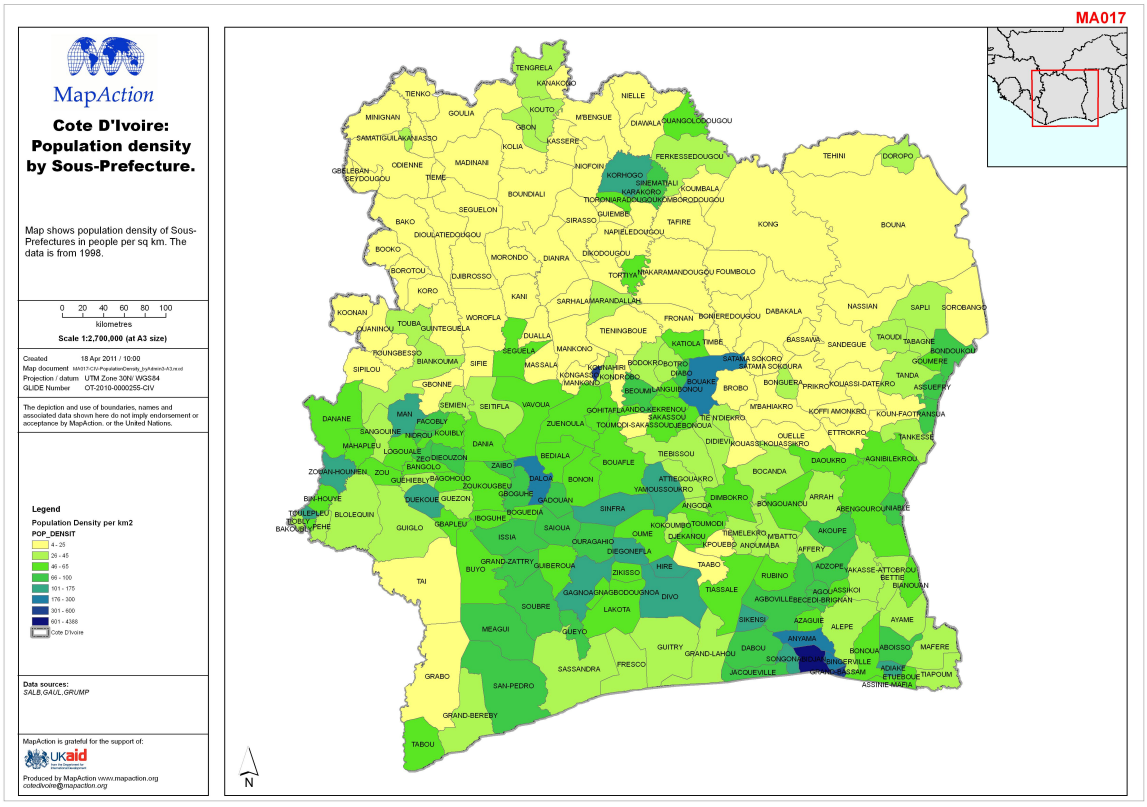

Figure 2.1. Côte d'Ivoire population distribution taken from Mapaction.

Economy is mainly dependent on agriculture, as Côte d'Ivoire is one of the major exporters of cocoa and palm oils. Côte d'Ivoire is one of the 20 poorest countries in

[4] https://www.cia.gov/library/publications/the-world-factbook/geos/iv.html

the world according to the International Fund for Agricultural Development (IFAD). Half of the population is living in the rural areas, and nearly 60% of the population is living below poverty threshold. Especially north eastern and western parts of the country are suffering from extreme poverty [5] [6] .

Côte d'Ivoire had been suffering from unstable political conditions during the data collection period. Although being ruled by Laurent Gbagbo since 2000, Alassane Ouattara won the 2010 elections This result was opposed by supporters of both parties, and carried the whole nation into a chaos. According to United Nation Human Rights Department, more than 600,000 Ivorian were displaced in the country and around 200,000 Ivorian migrated to neighboring countries in order to be in secure living conditions.

Lethal incidents significantly decreased after Gbagbo's arrest in 11 April 2011. Election have been held once more under the security of United Nations in 11 December 2011, which is within data collecting period.

The data set contains mobile call data from five million users, collected between December 2011 and April 2012. It is provided by Orange Telecommunication company. D4D is composed of four subsets of data, summarized in Table 2.1.

Table 2.1. D4D subsets.

	Subset1	**Subset2**	**Subset3**	**Subset4**
Num.User	-	50,000	500,000	5,000
Duration	5 Months	5 Months	5 Months	5 Months
Type	AntAnt.Call	UserAntenna	UserSubpref.	User-User

- SET 1 - Aggregate communication between cell towers (Antenna - Antenna)
  It consists of aggregated call logs for each hour, with the initiator antenna and the destination antenna.

[5] http://www.irinnews.org/Report/81804/COTE-D-IVOIRE-Poverty-getting-worse-study
[6] http://hdr.undp.org/external/mpi/Cote-d-Ivoire-OPHI-CountryBrief-2011.pdf

Data map is as follows :

Date - Hour - Initiating Antenna - Destination Antenna - Number of Calls - Duration

Interactive visualizations of the data set are accessible at the Geofast website [7] . Geofast is a web-based tool for the interactive exploration of mobile phone data. Below is an example from the data set:

$2012-04-28\ 23:00:00\ 1236\ 786\ 2\ 96$

$2012-04-28\ 23:00:00\ 1236\ 804\ 1\ 539$

The first entry tells us that on 28-April-2012 from 23:00 to midnight, two calls were initiated from antenna 1236 to antenna 786 with a total duration of 96 seconds.

- SET 2 - Individual Trajectories: High Spatial Resolution Data

  This data set is composed of 50,000 randomly chosen user's mobile phone usage with data resolution at antenna location level.

  Data map is as follows : *User ID - Date - Time - Antenna ID*

  Example of data in POS SAMPLE 0.TSV :

  $43690\ 2011-12-10\ 10:51:00980$

  $36462\ 2011-12-10\ 16:12:00607$

  The first entry tells us that User ID 43690, made a call on 10 December 2012 at 10:51 from antenna ID 980. Location information of sub-prefecture and antennae is given in Figure 2.2 in Set 2.

- SET 3 - Mobility traces, coarse resolution data set

  This set is composed of the same data schema as Set_2 for the whole period, but with less precision on location. The location information is given at the sub-prefecture level IDs instead of Antenna IDs. The country is divided into 255 sub-prefectures, as shown in Figure 2.2. Each sub-prefecture contains several antennae. Sub-prefecture level IDs means that for each call, only the sub-prefecture information is available. This data set is composed of CDR data from 500,000 users that are randomly chosen over five million users.

- SET 4 - Communication sub-graphs

[7] http://www.geofast.net

Figure 2.2. Côte d'Ivoire sub-prefectures with antennae locations taken from [1].

This set includes 5,000 random users considering all communications between a user and his/her contacts, at up to two degrees of separation from the user. The communications inside each sub-graph are aggregated by two-week time windows over five months.

The data map is as follows : *User ID - User ID*

Example of data in GRAPHS 0.TSV :

1052 20002

20002 20022

The first entry stands for user ID 1052 calls user ID 20002, the number of calls and time are abstracted.

The user IDs are anonymized, and the subsets are randomly sampled and re-numerated, therefore user based correlation is not possible between different subsets. Set_2 and Set_3 are similar according to the data schema, Set_2 is a small subset of Set_3. However Set_3 is more challenging to manipulate since each RAW file size is about 2

Giga Bytes in size. In Set_2, each file is around 150 Mega Bytes.

In the data set collection period, approximately 100 hours of data are missing, as declared by the Challenge Commission. In our analysis we found out that the missing part corresponds to three subsequent days {27,28,29 January 2012}, and we have removed the whole week from the data set, in order not to disrupt the weekly trend analysis. In addition to that, only antennae, which are present in all subsets, are evaluated in our study.

Similar to most developing countries, the population is dense in the capital and in the main big cities. Abidjan is the biggest city in Côte d'Ivoire including 25% of the total population. In this regard, the calls are more likely to take place in Abidjan region with around 80% of calls initiated here. This dense activity in that location suppresses other regions' call tendencies when we classify cities according to call behaviors and mobility patterns. In our analysis, we evaluate Abidjan separately. In Figure 4.11, node 60, represents the sub-prefecture Abidjan. It is very different in term of connectivity and the call volume from the rest of the country.

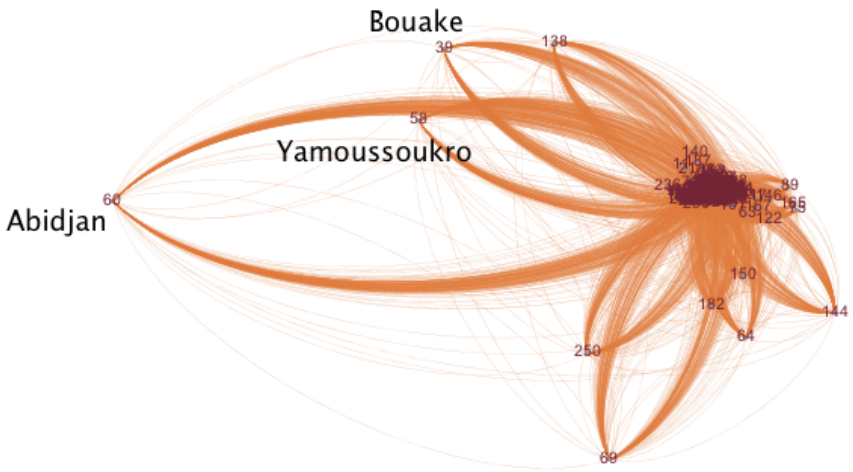

Figure 2.3. Similar cities with respect to connectivity.

2.4. Related Work

Today our phones are capable of sensing context in detail, and they can collect different types of data. This variety of data enables studies for understanding human nature, ranging from social sciences to formal sciences. We summarize different usage of mobile phone data in the literature. Then the social impact of traumatic and effective incidents and their reflections on country scale communication patterns are analyzed. In last section, we give detail about inspired study of Ihler *et al.* for understand events from one data source [5].

2.4.1. Application Areas Using Mobile Phone Data

Mobile phone data analysis make more accurate understanding of human behavior patterns possible, including patterns initiated from human behaviors such as traffic or infectious disease dissemination patterns. The findings from such analysis can be utilized for developing solutions for problems in urban planning or just for developing an intelligent assistant interacting according to subject's pattern [24, 25].

Zheng *et al.* define behavior patterns as temporal transitions between typical states of a person [26, 27]. These states may represent significant locations of the user, such as *Home, Work, etc.* or transportation states such as *via car, by foot, etc.*. Some of these locations may have general semantics, like restaurant, entertainment or shopping center. Montoliu *et al.* aim to discover the significant location of an individual from a multimodal smartphone data [28,29]. In their approaches minimum and maximum time between two *stay points* and distance between two stay points are set to predetermined thresholds in order to discover *stay locations*. These locations are clustered with two techniques: one is density based and the other is grid-base clustering. Location points are discovered with 63% success ratio, but only 4% is coming from GPS sensors, the rest are from labeled WI-FI access points. Their methodology is based on tagged WI-FI locations, this methodology can be improved from that perspective, as in real life we do not have tagged WI-FI locations.

Eagle *et al.* introduced a method called eigenbehaviors [20]. In this approach, the studied behavior patterns are daily activity logs, composed into vectors of uniform size. Each such vector is n-dimensional, where the dimensions represent the activity at a particular time of the day. Once all vectors for all users are aligned, the approach computes, via principal component analysis, the main axes of variation in the data. These are *Eigenvectors* of the data co-variance matrix, and since they represent dominant behavior patterns, are called eigenbehaviors. When examined, the researchers show that there are some eigenbehaviors correspond to typical activity on a week day, and other eigenbehaviors represent weekends.

Recommendation services are another application area. Zheng *et al.* referenced subject's activity and location history [30]. Their methodology is based on *Matrix Factorisation* [31]. The data set contains GPS data of 119 subjects, which lasts for 2.5 years, collected in Beijing. In order to compute the matrix factorization, missing entries have to be replaced with the possible values. They try to eliminate this by using collaborative filtering (CF) [32]. Like Montoliu *et al.* they first define stay points and stay regions [28]. Data are trained with 30% and 50% of the whole data set respectively. The more accurate results are obtained when the amount of training data is higher. Obtaining training data is restricted in mobile data sets, that is why they proposed a methodology, which is based on ranking. Their overall solution can be competitive, but not applicable to large data sets. Matrix factorization computation may not be feasible in real life data due to computation time.

Different than these studies, bioinformatic methodologies are implemented in Choujaa *et al.*'s research to understand human behavior patterns [33]. In both bioinformatics and human behavior analysis domains, sequences are crucial. The sequence of DNA in bioinformatics, corresponds to the sequence of human activity patterns in human routine classification domain. Global pairwise alignment [34], Unweighted Pair Group Method Using Arithmetic Averages (UPGMA) also called Average-linkage Clustering, [35] profile Hidden Markov Model (HMM) [36] are three techniques that have been applied in sequence analysis problems. Their results show that using bioinformatic techniques to overcome time shifts improves accuracy of routine classifications.

In some studies the focus is on bluetooth and application usage data, instead of GPS data. Azam *et al.* utilize bluetooth data; the proximity data provide information about the behavior pattern [37]. The users, who are predictable in their daily patterns, are stated as having low entropy. Similar to [33], they analyzed the proximity of data sequences. They used N-Gram and Correlative Matrix method. They achieve to predict for one subject's repeated pattern, and state that they will apply for larger group in their future work.

Geo statistics is also a new domain for mobile phone data analysis, Patil et. al work on virus dissemination extends the Bayesian modeling in cartography, as introducing candidate maps term, for sampling each time from the calculated distribution from the prior information, obtained from mobility data taken from CDRs [12, 13, 38].

A major problem is finding a data set with ground truth information focusing on particular problem. For that reason, Phithakkitnukoon *et al.* developed the *activity-aware map*, which is collected in Massachusetts (United States of America) with one million users with the data provided by Airsage company [22, 39]. Their work is based on data from 30 July 2009 to 12 September 2009. First they categorized the map of Massachusetts with Point of Interest (POI) data, as shown in Table 2.2, which they gathered from Yahoo API [8] , with $500m^2 \times 500m^2$ grid resolution with most probable activity from Table 2.2. They labeled each grid by one of the four categories; *Eating Region, Shopping Region, Entertainment Region, Recreational Region*, respectively. Each of these classes encapsulates information related to activities associated with these type of regions as shown in Table 2.2. After Phithakkitnukoon *et al.* created their own ground truth data, they first focused on how daily activity patterns between people who work in the same area categories are similar, and secondly how distance impacts that work in urban shopping area compared to a distant shopping area [39]. Their findings are interesting in such that people who work in similar area bring similarities into their daily activity patterns as well.

Apart from the ground truth labels, the other main problem in large mobile phone

[8] Yahoo, http://pysearch.sourceforge.net.

Table 2.2. Considered activities and keywords used for locating points of interests, taken from [39].

Activity	**Keywords Used**
Eating	Restaurant, Bakery, Coffee Shop
Shopping	Mall, Store, Market
Entertainment	Theater, Bowling, Night Club
Recreational	Park, Gym, Fitness

data sets is that they generally consist of multimodal data types in large scales, with noise and missing data. Computation of massive data to infer information, require stochastic methods rather than heuristic methods.

Table 2.3. Related mobile phone data studies.

Reference	**Dataset**	**Question**	**Techniques**	**Performance Results**
Zheng *et al.* [26, 27]	Reality Mining subset for 37 users	Understand behaviour pattern	LDA, MCMC, Gibbs Sampling	70% for Area Under Curve(AUC) of activity inference
Farrahi and Gatica-Perez [10]	RM, Nokia	Understand Behavior Pattern	LDA, MCMC, Gibbs Sampling	
Montoliu *et al.* [28]	Nokia, Controled Group of 8	Understand Significant Locations	Density and Grid based clustering	63% discover significant location points
Phithakkitnukoon *et al.* [39]	Airsage [22]	Semantic Classification of Locations / Corrolation of human behavior	Hamming Distance	Common work area profile brings similarities in daily activity patterns
Wirz *et al.* [40]	Controlled dataset of 13 people	Pedestrian Flocks	DJ Clustering	Understanding number of flocks, 78,5%, assignment of individuals to flocks 91,5%
Kang, J. and H.-S. Yong, [41]	Synthetic Data	Mining GPS Trajectories	BIRCH: Clustering with Euclidian distance	

2.4.2. Event Detection From Mobile Phone Data

Our work is based on an assumption that, if an unexpected life threatening event occurs, people tend to call for help, or inform others that they are safe. Gibson stated that, similar behavior patterns are observed under such circumstances [42]. He classified human reactions during traumatic events correlating with the phases of an incident, and proposed, *Realisation, Acknowledgment* and *Adaptation* steps. He shows that human psychology tends to share and inform about an incident right after it is realized. The event's traumatic effect may create an urge to share information. In our work, we observe that similar patterns are visible in sports events, which involve hope, excitement and competition. These motivations may bring such information diffusion with higher amplitude for longer period, observable in call count data per antenna in the affected area.

Mobile phones are extensions of modern life in developed countries. However one can question whether this is true for African countries as well. Mobile phone usage is very common, even in very poor countries. The reason is explained in a study by Duncombe and Boateng, where they state that fixed-line infrastructure for communication is either insufficient or under development in such countries [43]. Even availability to access simple banking transactions over mobile phone make Sub-Saharan people tend to invest on mobile phones. Côte d'Ivoire has a coverage of 4.8% for fixed lines, with respect to a ratio of 36.6 of people possessing mobile phones in the country [44]. Shapiro and Weidmann analyzed rebel groups and mobile phone coverage in Iraq and observed that there is a decrease in the lethal incidents within the regions under mobile phone coverage [45]. On the contrary, Pierskalla and Hollenbach found a close positive relation between phone coverage in Africa and possibility to have a violent incident. Fast information transmission from a leader to the members of an organization accelerates conflicts through out different ethnic and civil groups [46]. In our work, correlation between mobile phone coverage and the probability of conflict was not found. The densest population and also the highest coverage is in Abidjan but security related events tend to happen in mid-west of Côte d'Ivoire, where many Northern and Southern conflict groups encounters existed, as in shown in Figure 2.4.

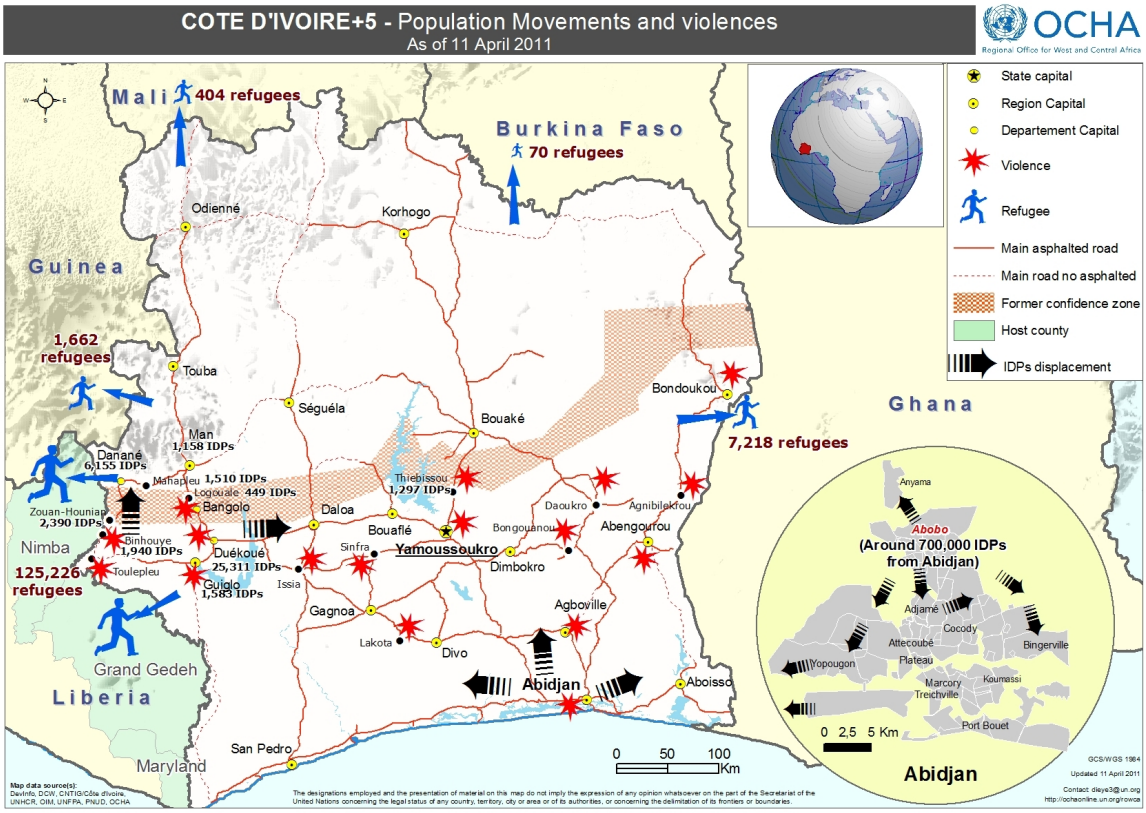

Figure 2.4. Côte d'Ivoire population movements and violence, taken from [2].

We have used a statistical Poisson distribution for modeling count data of call data per antennae and Negative Binomial distribution for rare events in given time [47], [48]. This method is widely used in different domains, for example in modeling rare incidents in psychiatric hospital [49], for assessing social influence of death penalties for deterring homicides [50], or for web traffic analysis [51, 52]. Another term for event detection is change point analysis in time series [53–55].

Our work is similar to Ihler *et al.*'s time varying Poisson process model [5]. In their work, the number of cars passing by the Dodger stadium, and number of people entering/exiting from the university campus building were analyzed. These data sources are used to predict events occurring during the time of data collection. The most common way of anomaly detection in time series is to define a baseline. Ihler *et al.* compare a Markov modulated Poisson process and a baseline model, and show that a Poisson process model can achieve almost 100% accuracy in anomaly detection. In our case, the baseline approach is not applicable, since the data come from multiple antennae and each antenna has different characteristics, depending on the time and location. In addition to their, in that method high values may suppress normal behavior in the time series model.

We detect anomalous events from the deviation of daily call volume patterns. Determining *normal* behavior is an issue, as the call data special pattern depending on time and day of the week [56]. The most prominent approach is to define a threshold value, by calculating the mean of the estimated density of the analyzed value [57].

One of the significant contributions of our work is to evaluate all antennae individually with a Poisson process and depict their effects on each other on a map when an event occurs. This keeps the spatial and temporal property of the data.

Akoglu and Dalvi work is similar to our work in terms of data type, and the objective [4]. However with respect of multiplication of high dimensional matrixes is not applicable to our data in term of computational cost. They construct graphs for *who-call-whom* and *who-texts-whom* and after evaluating call behaviors with eigenvalues,

detect events by the deviation from common behavior.

Our work differs from previous works with implementation and data source perspectives.

2.5. Machine Learning and Stochastic Approaches

In this section we review the probabilistic methods to understand the background of the Markov modulated Poisson process (MMPP) for event detection and Latent Dirichlet Allocation (LDA) for mobility pattern analysis. Problems that we tackle in evaluating *CDR* data generally include hidden variables, missing data in multidimensional space, and are intractable to compute. Finding an exact solution mostly is not possible. In such circumstances, the only way is to find an approximate instead of an exact result.

The first assumption in our event detection model is defining the count per antenna as a random variable, which is assumed to be populated from a Poisson distribution. In Section 2.5.1, the main properties of this distribution are summarized. It is shown that MMPP converges to the true target distribution in Section 2.5.2.2 [5, 48] .

2.5.1. Distributions

Below, the main distributions used in this thesis for derivations and model selection are shown.

2.5.1.1. Bernoulli. In Bernoulli distribution the random variable takes the value of either 1 with probability p, or 0 with probability $(1-p)$. The probability mass function is shown in Equation 2.1.

$$f(x;p) = \begin{cases} p = 1 \quad if \quad x = 1, \\ p = 1 - p \quad if \quad x = 0. \end{cases} \tag{2.1}$$

$$f(x;p) = p^x(1-p)^{1-x} \quad for \quad x \in \{0,1\}$$

2.5.1.2. Poisson. Poisson distribution is mainly used for modeling count data [58]. Poisson distribution is defined with parameter λ, which represents the mean value of the distribution, generally denoted as rate. A non-homogeneous Poisson process rate is a function of time. In our model, the rate function, $\lambda(t)$ is changing according to day of week and hour of day [59].

Let x be the random variable for count data, collected through period of time (t). Each count is independent and identically distributed (i.i.d.) from each other, with probability of π. Suppose we have $\{ x_1, x_2,..x_n\}$ observations for interval n. Probability of $P(x;n)$ can be calculated as n Bernoulli trials as shown in Equation 2.2. Probability mass function of Poisson distribution is derived by the mean value represented as λ when infinite trials take place.

$$P(x;n) = \binom{n}{x}\pi^x(1-\pi)^{n-x} \tag{2.2}$$

$$\lim_{n\to\infty}\left[\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}\right] = \frac{e^{-\lambda}\lambda^x}{x!} \tag{2.3}$$

The density, or probability mass function is shown in Equation 2.4.

$$
\begin{aligned}
Pois(x) &= \frac{e^{-\lambda}\lambda^{x}}{x!} \\
var(x) &= \mathbb{E}(x) = \lambda
\end{aligned}
\tag{2.4}
$$

2.5.1.3. Negative Binomial. The Poisson distribution is a special case of negative binomial distribution. If we recall from Equation 2.4, variance and expected value equals to λ, for Poisson distribution. In the count data model, we assume that the variance is *proportional* to the mean value, as shown in Equation 2.5. When this ratio ω is greater than one, it means over-dispersion. In another words, the variance of the distribution is greater than the expected value.

$$
var(x) = \omega\mathbb{E}(x) = \omega\lambda \tag{2.5}
$$

The negative binomial distribution shows the number of successes n in a sequence, which is generated through a Bernoulli distribution with probability p. Let $NB(n, p)$ denote the negative binomial distribution, n being the mean value and p the probability of having a rare event. The mass function can be represented as in Equation 2.6 [60,61].

$$
f(x; n, p) = \binom{n+p-1}{n} p^{x}(1-p)^{n} \tag{2.6}
$$

$$
\mathbb{E}(x) = \frac{pn}{1-p} \tag{2.7}
$$

$$
var(x) = \frac{pn}{(1-p)^{2}} \tag{2.8}
$$

Rare events with steep increase in the amount of count data, correspond to this

case [49]. Osgood applied Negative Binomial distribution for detecting crime rates [62].

2.5.1.4. Gamma Distribution. The gamma distribution is widely used as conjugate prior to exponential distribution or Poisson distribution. It is denoted with two parameter α as shape, and β as scale.

$$\begin{aligned} p(x|\alpha,\beta) = \Gamma(x;\alpha,\beta) = \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} exp(-\frac{x}{\beta}) \\ \mathbb{E}(x) = \alpha\beta \\ var(x) = \alpha\beta^2 \end{aligned} \tag{2.9}$$

2.5.1.5. Beta Distribution. Beta distribution has two parameters, which determines shape of the distribution. Beta(α, β) distribution is used model degree of belief, with some probability to the most likely probability.

$$x \in [0,1] \tag{2.10}$$

$$p(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{2.11}$$

$$\mathbb{E}(x) = \frac{\alpha}{\beta} \tag{2.12}$$

$$var(x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{2.13}$$

2.5.1.6. Dirichlet Distribution. Dirichlet distribution is multivariate generalization of beta distribution. Dir(α), which α is a vector of number of categories with the total probability equals to 1.

$$Dir(\alpha_1, \cdots, \alpha_K) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{K} \theta_j^{\alpha_j - 1} \tag{2.14}$$

$$\mathbb{E}[X_i] = \frac{\alpha_i}{\sum_k \alpha_k} \tag{2.15}$$

$$var[X_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \tag{2.16}$$

$$\alpha_0 = \sum_{i=1}^{K} = \alpha_i \tag{2.17}$$

2.5.2. Monte Carlo Methods

Monte Carlo methods arise as we need to compute in high dimensional space. This methodology is based on two theorems; *the law of large numbers* and *central limit theorem* [48]. In law of large numbers, if we can draw independent and identically distributed (i.i.d.) samples from a target distribution that we believe these samples are generated, the mean of the all samples will converge to the mean value of the target distribution, if large number of iterations take place.

2.5.2.1. Topic Models. Probabilistic topic models, also called Latent Dirichlet Allocation, are suitable to discover the hidden patterns in large scale unstructured data sets. Latent Dirichlet allocation (LDA) is an unsupervised method, which does not need any annotation. The main usage of topic models is to categorize documents. This method is not only applied on document libraries, but also in image databases, audio and music libraries and social networks.

LDA is a three-level generative hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics [11]. Generative process, enables the model to be represented as a *joint probability distribution*, which includes both hidden and observed variables. Each topic is, in turn, modeled as an infinite mixture over set of topic probabilities. In the context of text modeling, the

topic probabilities provide an explicit representation of a document. LDA is successor of probabilistic Latent Semantic Indexing (pLSI) [63]. The major difference between these two methods is; in LDA, a document composed of multinomial probability distribution over topics, however in pLSI , there is no probability distribution for documents over topics, in addition to that there is a risk of over fitting. From this perspective Blei *et al.* completes the study of Hofmann *et al.* [11, 63].

The notation of LDA is as follows,

- A *word* is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1..V\}$.
- A *document* is a sequence of N words denoted by w=$(w_1, w_2, \cdots, w_N)$, where w_N is the nth word in the sequence.
- A *corpus* is a collection of M documents denoted by M=$\{ w_1, w_2, \cdots, w_M\}$

Topic modeling is very strong, as it does not need any prior knowledge of the underlying themes. However, there are several assumptions in LDA.

- *Bag Of Words* assumption, the order of words are not important.
- Order of documents are not important.
- The number of topics are assumed to be known and fixed.

In Figure 2.5 plate notation of the LDA model can be seen. The rectangles K, N, M represent repetition in graphs. K number of topics, distributed in M documents, which is a distribution over N number of words. α, β are the hyper parameters of Dirichlet distribution. In that model, it is important to calculate the hidden variables with the observed variables. It is computed by *conditional distribution* which is also called *posterior distribution.* Blei *et al.*'s another contribution is using Dirichlet as conjugate prior of multinomial distribution for simplifying the computation of posterior distribution [11]. The probability density of a K dimensional Dirichlet distribution over a multinomial distribution $\theta = (\theta_1, \cdots, \theta_k)$. α is the parameter of the Dirichlet prior on the per-document topic distributions, β is the parameter of the Dirichlet prior

on the per-topic word distribution, θ_i is the topic distribution for document i, ϕ_k is the word distribution for topic k, z_{ij} is the topic for the jth word in document i, and w_{ij} is the specific word.

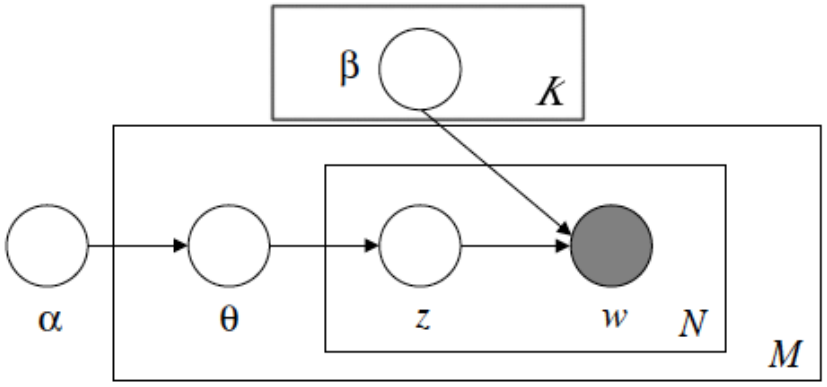

Figure 2.5. LDA graphical model, taken from [11].

(i) For $k = 1..K$:

(a) $\phi^k \propto Dirichlet(\beta)$

(ii) For each document $m \in M$

(i) $\theta_d \propto Dirichlet(\alpha)$

(ii) For each word $w_i \in m$:

(i) $z_i \propto Discrete\ (\theta_d)$

(ii) $w_i \propto Discrete\ (\phi^{(z_i)})$

Figure 2.6. Pseudo Code for LDA.

Joint probability distribution of the graphical representation is shown in Figure 2.5.

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{m,n}) \right) \quad (2.18)$$

Conditional probability distribution in Equation 2.19 shows that word counts

$w_{1:M}$, the only observed variables.

$$p(\beta_{1:K}, \theta_{1:M}, z_{1:M}|w_{1_M}) = \frac{p(\beta_{1:K}, \theta_{1:M}, z_{1:M}, w_{1:M})}{p(w_{1:M})} \tag{2.19}$$

The graphical representation of LDA as joint probability distribution as in Equation 2.18, can be converted into posterior distribution if we integrate over hidden variables. This posterior calculation should be iterated as shown in Equation 9.

2.5.2.2. Markov Chain Monte Carlo. Markov chain is sequence of random variable from a finite state space X, as shown in Equation 2.20. This sequence is called *Markov Chain* if it satisfies *Markov property* [48]. In our case, states are defined as $z_0(t)$ and $z_1(t)$ corresponds to normal, and anomalous times respectively. Here, t represents the time sequence of events.

Markov Chain is an ergodic chain, which means irreducible and aperiodic. Therefore all states can be accessible, independent from the initial state, and transition converge to a stationary state [64, 65].

X, denotes the random variables and s denotes state space.

$$p(X_{t+1} = s_j|X_0 = s_k, \cdots, X_t = s_i) = p(X_{t+1} = s_j|X_t = s_i) \tag{2.20}$$

Markov chain Monte Carlo (MCMC), is an unsupervised machine learning method. It differs from supervised learning that, it does not need any training set, instead the model converges to expected value in time. However it is assumed that random variables are generated from a distribution. The MCMC differs, as sampling from a state transition kernel [65].

Monte Carlo methods converge to approximate posterior if being iterated enough times. The disadvantage is that, the number of iterations are not known. Markov chain means transition probabilities between states only depend on the current state of the random variable as shown in Equation. 2.20.

2.5.2.3. Gibbs Sampling. Gibbs is a sampling algorithm, which samples through conditional distributions given by Markov chain [48, 66–70]. If the full joint distribution can be written in a standard distribution (Gaussian, Gamma etc.), samples can be taken directly. In a Bayesian network notation, the *Markov blanket* property, which assures node A depends on its parents, children and other parents of all its children, simplifies Gibbs sampling usage on Markov Chain Monte Carlo implementations. The popularity of Gibbs is coming from iterative usage of the conditional distribution, as shown in Algorithm 2.7.

i. Initialize $x^1 = (x_1^1, ..x_M^1)$

ii. **for** $n = 2, 3, ..$ **do**

(a) $x_1^n \sim p(x_1^n | x_2^{n-1}, x_3^{n-1}, ...x_M^{n-1})$.

(b) $x_2^n \sim p(x_2^n | x_1^{n-1}, x_3^{n-1}, ...x_M^{n-1})$.

...

(c) $x_M^n \sim p(x_m^n | x_1^{n-1}, x_3^{n-1}, ...x_{M-1}^{n-1})$.

iii. **end for**

Figure 2.7. Pseudo code for Gibbs sampling.

3. TOPIC MODEL IMPLEMENTATION TO UNDERSTAND MOVEMENT PATTERNS

This chapter is about implementing document classification techniques to CDR data to understand movement and call patterns. Movement patterns collected through sequences of call locations, which have at most antenna based precision, are evaluated as documents in the topic model. Latent Dirichlet Allocation and topic models are used as synonyms throughout the thesis. In Section 2.5.2.1, technical details of the model were given.

3.1. Implementation

In a topic model, we have latent variables. Latent variables in our model correspond to location traces that have been initiated with each users call log. The corpus is composed of documents, which correspond to phone users. Each user leaves a trace whenever he or she makes a call. This trace includes spatio-temporal information, which is the antenna location or region of the call. Figure 2.2 shows the sub-prefectures, and the antenna locations. The hidden patterns that we are searching for are the main movement patterns that subjects use, and possible immigration paths during the civil war.

Trajectories are composed of location pairs $[s_1 s_2]$, where s_1 refers to the source location, and s_2 refers to the destination location. This representation does not incorporate the time dimension. In the topic model, such a location pair corresponds to the $Word$ $[s_1 s_2]$, and multiple words of a single user make up a *Document*. In Farrahi and Gatica-Perez's works, annotated location information corresponds to {*home, work location*}, and time slots correspond to a words [9, 10, 71] .

3.1.1. Inference and Parameter Estimation

The inference problem, which means calculating posterior distribution's parameters, shown in Equation 2.19, is not easy. Equation 2.19 is intractable as Blei *et al.* stated [11]. Instead of exact inference, approximate inference methods are used. LDA presents efficient approximate inference techniques based on variational methods and Monte Carlo methods.

3.1.1.1. Variational Inference. Variational inference is used to calculate approximate true posterior. Marginalizing the log likelihood of Equation 3.1 allows us to solve it in terms of an optimization problem [72]. The disadvantage of variational inference is that, while computing maximum log likelihood, a local maxima can be reached.

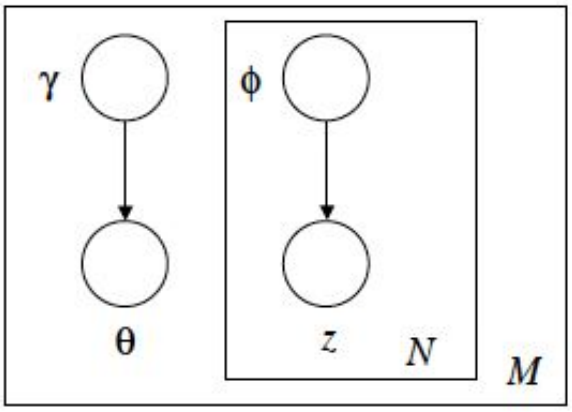

Figure 3.1. Variational inference graph representation taken from [11].

In order to compute this model, graphical model shown in Section 2.5.2.1, Figure 2.5 is simplified as in Figure 3.1. In that representation, posterior distribution can be written, as in Figure 3.1. As we recall from Section 2.5.2.1, γ stands for Dirichlet hyper parameter, and $(\phi_1, \cdots, \phi_N)$ are the multinomial parameters.

$$q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1} N q(z_n|\theta_n) \tag{3.1}$$

3.2. Experiments

We use LDA [11] in two different ways, first to understand the call pattern of the subscribers, and secondly the mass movement patterns of population.

3.2.1. Daily Call Behaviors

In order to implement LDA, we should define *documents, topics and vocabulary* in our domain.

Documents → User Mobile Data
Topics → Different Call Behaviors
Vocabulary → Daily Call Numbers

Matlab Topic Toolbox [73] is used for implementation. Data are preprocessed as vectors of words, as shown below. The number of topics or call patterns should be given to the model. We set the number of topics as 10, in our experiment.

3.2.1.1. Call Information. In the first experiment, we set the number of calls during the day as follows, { 0, 1-3, 4-7, 8-10, 10 <}. However in the first experiment, it is observed that Côte d' Ivory mobile phone users are not very talkative, and this model does not discriminate the users from each other. So a new range is set as, {0, 1-2, 3-5, 6-10, 11 <}, and instead of dividing the day into eight time slots, we divided the day into four slots, which represent, night, morning, daytime, evening, respectively. The related dictionary is shown in Table 3.1.

After running out the LDA with 10 topics with hyper parameter set $\alpha = 0.01$ and $\beta = 5$ for 300 iterations, we visualize the probability distribution of call behaviors as shown in Figure 3.2.

Table 3.1. Topic model dictionary.

Slots	**Number of Calls**	**Corresponding Word**
Slot 1-Night	0/ 1-2/ 3-5 / 6-10 / 11 <	A0/A1/A2/A3/A4
Slot 2-Morning	0/ 1-2/ 3-5 / 6-10 / 11 <	B0/B1/B2/B3/B4
Slot 3-Daytime	0/ 1-2/ 3-5 / 6-10 / 11 <	C0/C1/C2/C3/C4
Slot 4-Evening	0/ 1-2/ 3-5 / 6-10 / 11 <	D0/D1/D2/D3/D4

The vertical axe corresponds to duration of the day, such as night to morning, noon to evening. The horizontal line presents 10 topics and the last line corresponds to mobility. The matrix shows the probability distribution for each time slot and for the possible number of calls. For example, Topic 7 shows a behavior of being likely to talk during the day, and in the evening. Topic 3 corresponds to talking in the evening with high mobility. The other topics are very similar with tendency to make quite a few calls with less move. It is important to keep in mind that, we get movement information whenever the subject makes a call. If the subject does not make a call, it is not possible to infer location.

3.2.2. Movement Direction Patterns

We make a number experiments in order to understand human movement patterns. Our motivation is to analyze human movements for understanding possible migration routes. During the data collection period there was a civil war, and thousands of people were displaced inside the country and immigrated to neighboring countries. This movement and violent incidents were roughly visualized in Figure 2.4, prepared by Oxford Poverty and Human Development Initiative (OPHI). Our analysis provide a much more detailed picture of these movements.

We use LDA Topic Model, as we mentioned in Section 2.5.2.1. LDA is a model, to discover hidden structure from the unlabeled data. According to our assumptions, the corpus of the data consists of users which have commuter, immigrant and refugee

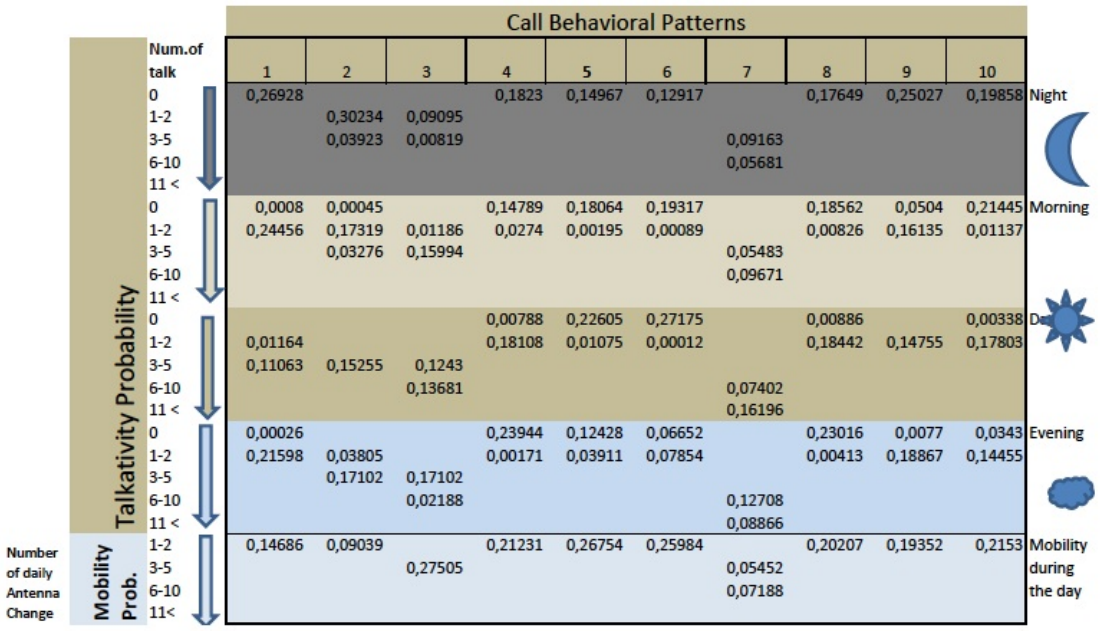

Figure 3.2. Topic distribution.

patterns.

This set of experiments have been done with the subset of D4D, Set_3, see Section 2.3.2.

Let for each user u in U, the call path be defined as follows. The sub-prefecture region is where the call takes place in, as Set_3, locations are given with sub-prefecture level precision. S is the set of sub-prefectures; $s \in \{S_1, \cdots, S_{255}\}\}$.

$< Traj >_u = \{s_{t1}, s_{t2}, \cdots, s_{tk}\}$ where $t_1 < t_2 < \cdots < t_k$ which shows the consecutive time slices. The *vocabulary*, is drawn from $\{S_M$ x $S_M\}$ and represents pairwise Origin Destination (OD) paths. If we have $\{s_1, s_2\} \in S$, it corresponds to subjects consecutive calls s_1 and s_2, respectively. $[s_1 s_2]$ is represented as a word for that subject.

$$\begin{aligned} Document &= < Traj >_u \\ &= \{[s_1 s_2], [s_2 s_3], \cdots, [s_{k-1} s_k]\} \\ Topic &= \{Commuter, Immigrant, Others\} \end{aligned}$$

Determining the number of topics is one of the weaknesses of this model. In our movement related experiments, a *document* corresponds to subjects or mobile phone users, and *words* corresponds to trajectory of the subjects.

3.2.2.1. Topic Model Experiment 1. The number of topic is set to 10. The number of iteration is set as 1000. Only one subset of the corpus is taken, for 500,000 use for two weeks period. The pre-processed file, which is also the size of the corpus, contains approximately 1 million lines.

$< Traj >_u$ contains the data whenever user change location. It is a state transition array for each user. Each sub prefecture can be denoted as nodes, and from each node.

Definition: Directed Acyclic Graph (DAG) index: For each u, $u \in U$, there is a $< Traj >_u$ sequence, if $(X, Y) \in < Traj >_u$ and $(Y, X) \in < Traj >_u$ it increments the DAG index. It denotes acyclic behavior of trajectories.

In addition to the words, which are composed of location pairs of the user, we put a so called Directed Acyclic Graph (DAG) index, in order to understand the subject's probability of being immigrant. High DAG index means, subject tend to come back to a significant location. For example, let each sub prefecture node denoted as s_i, if the user u_j, trajectory $< Traj >_u$ contains a transition from sub prefecture Id 69 to 61 the corresponding *word* is 69061. If the user tends to turn back there should be a word in the topic as 61069, as in Figure 3.3.

We evaluate our model, after obtaining the topics/word distribution from training data set, we calculate the topic probability for each user in the test data set. And compare the probability distribution with the training results. If we calculate the perplexity of each document, we see that it is almost predictable, very close to 1. In Figure 3.10 the log likelihood calculation per document is plotted.

People tend to return back to place they come from. In that perspective DAG index make the differentiation as we expect. As in Figure 3.4 Topic 5 and Topic 4 is not related to each other although their DAG index is very probable. Topic 2 has very small DAG index, even not visible, it has a similar characteristic in document wise Kullback-Leibler (KL) distance with the rest of the topics.

3.2.2.2. Topic Model Experiment 2. In that experiment we would like to see how the amount of data will effect the topic distribution. In first experiment, we just use the small subset of data. Now we put the test data apart, which is the last month of the data collection period, to use in testing period.

The pre-processed data composed of around 10 million lines, for 500,000 users for 4 months period. In order to understand how our model fits into data, θ_{train} is compared with θ_{test} for each subject. Number of topics is 50, and iteration number is 500. In order to visualize the training data set, we put each trajectory of each topic into Gephi [9] , and gain an intuition of the data. The meaningful trajectories can be found in that section. In order to understand the semantic of the data, as a background layer, the map of the Côte d'Ivoire is attached with a common image edit program, as you can see the map in Appendix A.1. The number of topic is a prerequisite for LDA, so some topics may have similar words. This analyze is calculated with KL divergence of document/topic distribution, which is calculated for each document, as you can see in Figure 3.7.

The Root Mean Square Error (RMSE) of Experiment 2 shows that the corpus

[9] http://gephi.github.io

Table 3.2. Experiment 2 results.

Topics	**RMSE**
3 Topics	0.5482
10 Topics	0.3733
20 Topics	0.2734
50 Topics	0.1652
100 Topics	0.1112

length do not effect the success rate of our model.

The RMSE results show that the number of topic can be set to 50. In Figure 3.6, the corresponding probability distribution of per document / topic distribution θ_d in Figure 2.5 is shown for Experiment 2.

KL divergence of the topic distributions θ_d is shown in Figure 3.7. When we analyze Topic 12, Topic 31 and Topic 44 according to Figure 3.7, which make them different from the rest of the topics, we see that in Figure 3.8 Topic 44 denotes main highway of Côte d'Ivoire, Topic 12 and 31 results do not give any specific information within our knowledge.

In Figure 3.8, are the people who tend to travel back and forth from Abidjan to Yamoussoukro, which the biggest highway in the country.

Abidjan is the biggest city in Côte d'Ivoire, most of the population is living there. In Figure 3.6, Topic 18 is the most probable topic in the corpus. Figure 3.9 shows that corpus is mainly composed of people living in Abidjan who do not tend to move much.

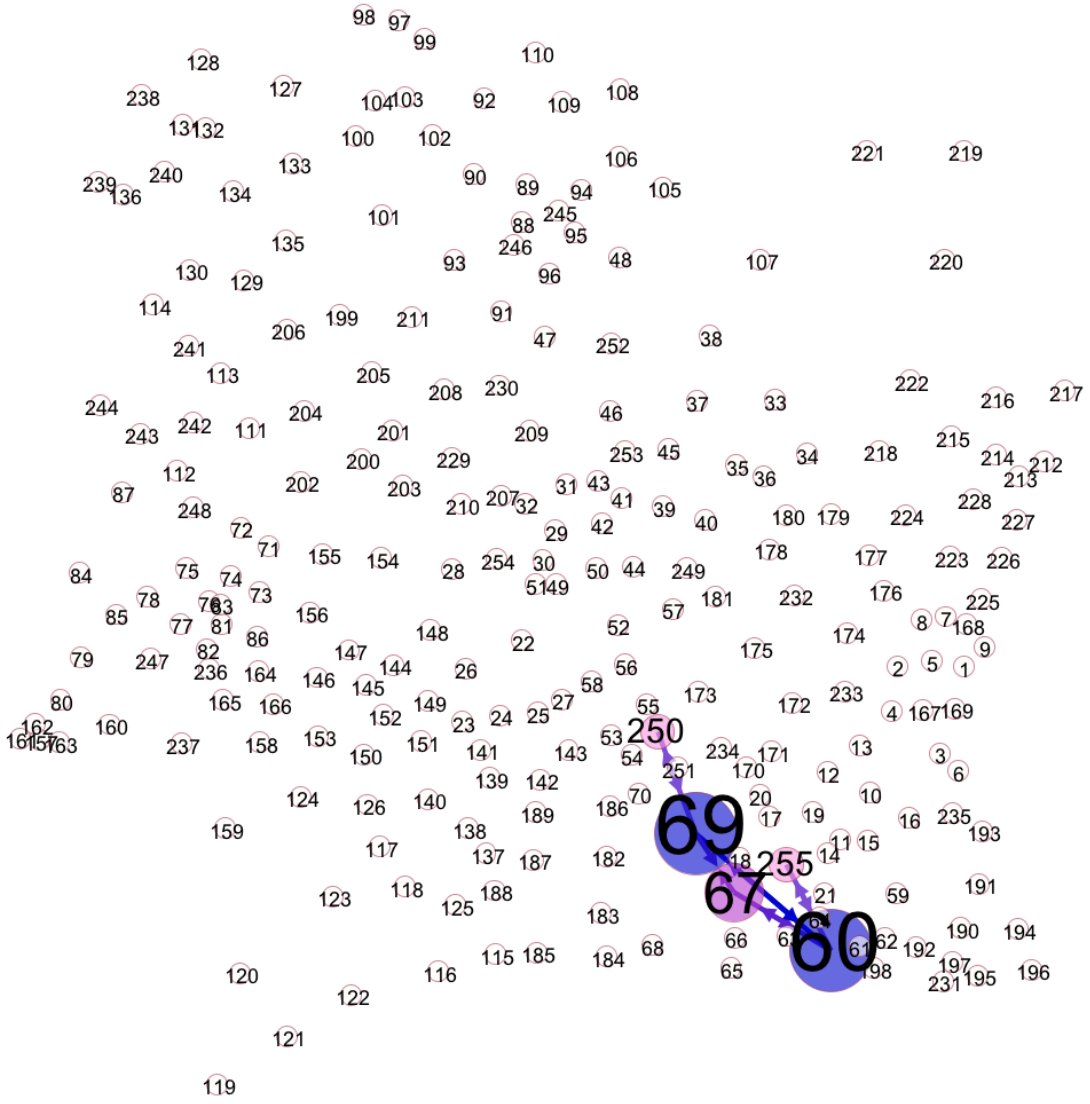

Figure 3.3. An example transition between sub-prefecture nodes.

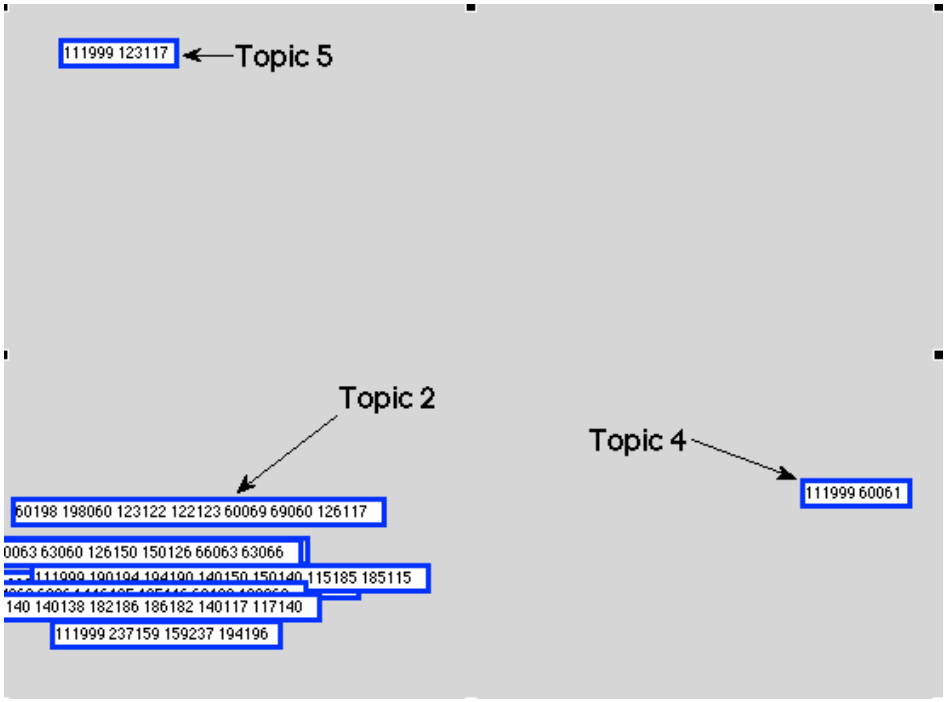

Figure 3.4. Experiment Set 1: Topic similarity for 10 topics.

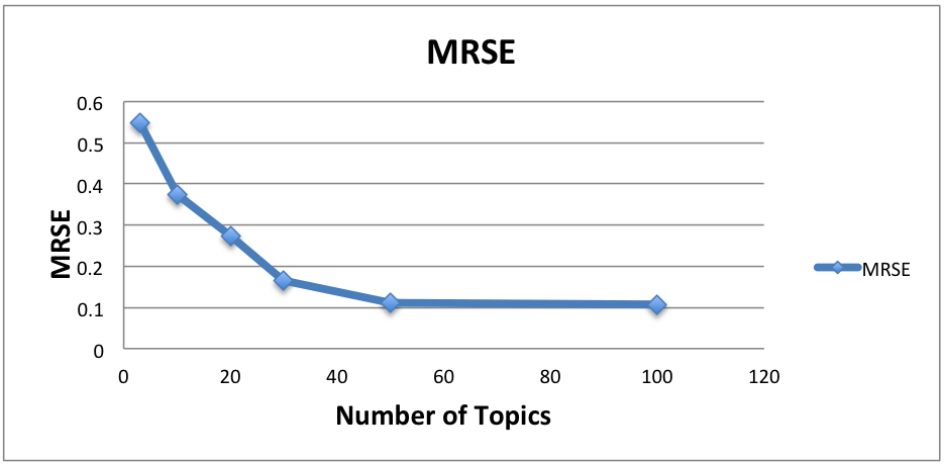

Figure 3.5. Experiment Set 2: Root Mean Square Error.

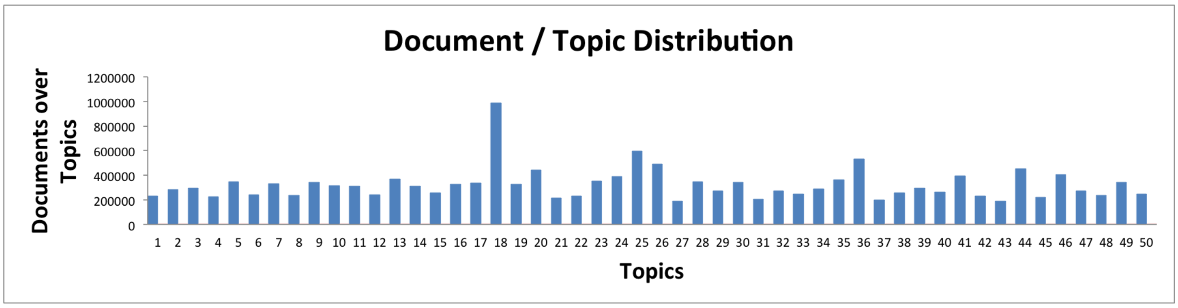

Figure 3.6. Document/topic distribution for experiment 2.

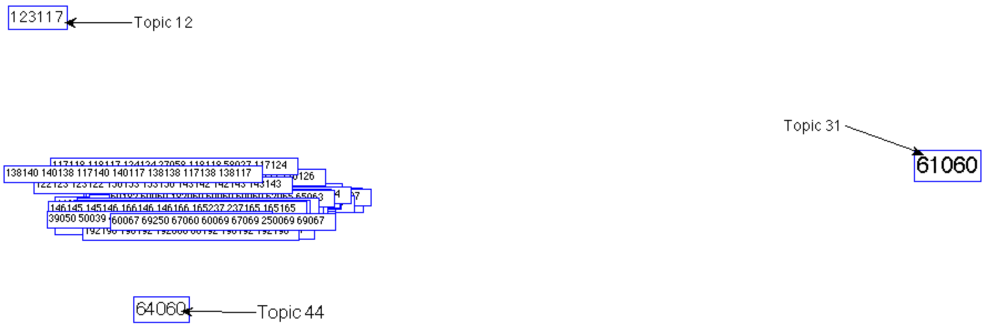

Figure 3.7. Experiment Set 2: Topic similarity.

Figure 3.8. Experiment Set 2, Topic 44 showing the main highway.

Figure 3.9. Experiment Set 2, most probable topic in the corpus, topic number 18.

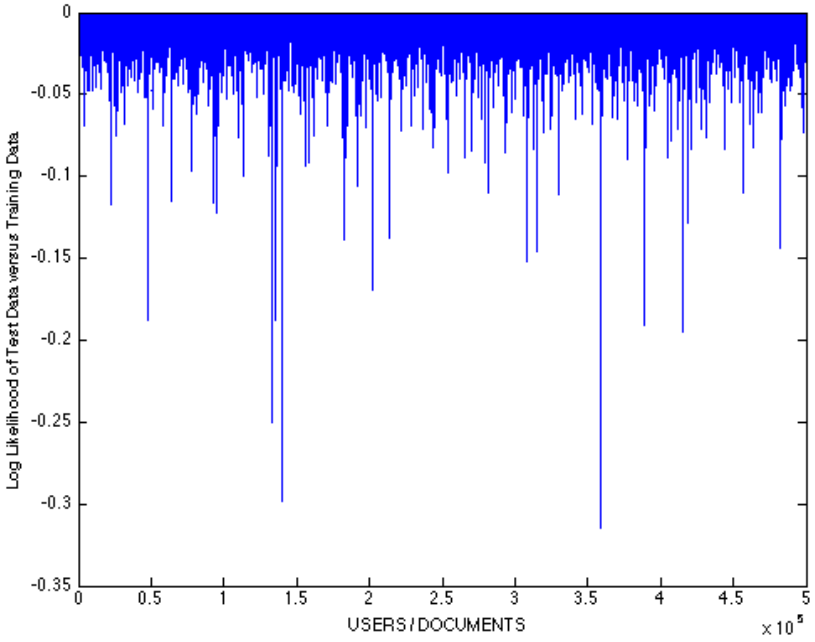

Figure 3.10. Document log likelihood of test data versus training data.

4. METHODOLOGY

In this chapter, we describe how we process big data, visualize and perform predictive analysis. In Section 4.1, we introduce our framework for pre-processing, tools and environment. In Section 4.2, explorative analysis of CDR data is given. In Section 4.3 Markov Modulated Poisson process applied to CDR data for understanding anomaly events are presented.

In Figure 4.1 the spatio-temporal representation of one antenna is shown.

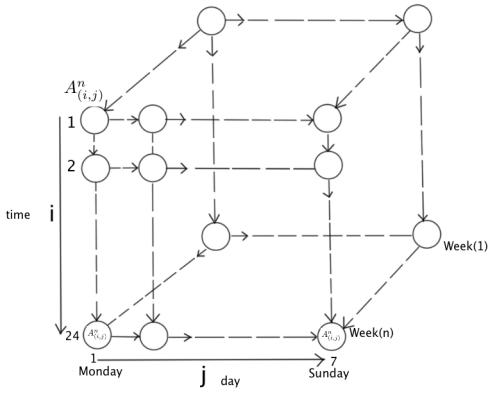

Figure 4.1. Aggregated call number per antenna in time space.

The number of calls per antenna are changing in time, such as higher during the day, and lower during the night, as shown in Figure 4.2. The *"normal"* call pattern is substantially make a peak when an anomaly happen in the region or among country.

Geo-spatial temporal statistics generally utilize different approaches compared to statistical science [74]. The most common assumption for *random variables*, *independent and identically distributed* (i.i.d.) does not hold for spatial data. Observations which are closer in space and time means, more correlated with each other in spatial space, however in our stochastic model each antenna assumed to be iid for exposing anomalies in the call patterns.Interpolation of antenna data mislead less dense antenna

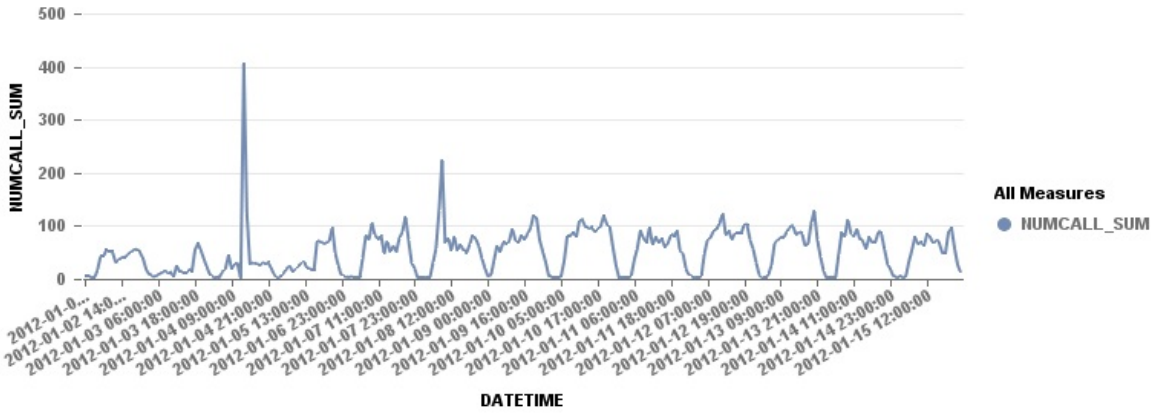

Figure 4.2. 2 Jan.-15 Jan.2012 Antenna 524 call counts.

regions, which corresponds to North East of the country. Country wide analysis, bring challenges of missing data. Here, missing data term corresponds to regions having no antenna, or not visible on the map, as very low call volume takes place.

With billions of record; processing, querying and visualizing CDR data is challenging. In order to explore data, we developed a framework as shown in Figure 4.3 to handle *Big Data* effectively. The first experiments are composed of single antenna call patterns, which are close and far from the incident location. GIS tools like ESRI,[10] Google Maps,[11] are used for GIS querying. In memory computing [75, 76] leverages querying from CDR.

Having an insight of data, is crucial to understand the patterns beneath. One type visualizing tool is not adequate most of the time. In our study we use, Gephi,[12] Tableau Software,[13] Matlab,[14] SAP Lumira[15] . In Appendix B some of the technical properties are given for Gephi. Analysis and related visualizations are found in Section 4.2.

[10]http://www.esri.com
[11]http://maps.google.com
[12]http://www.gephi.com
[13]http://www.tableausoftware.com/public/
[14]http://www.mathworks.com
[15]http://saplumira.com

4.1. Data Processing Model

For data exploration we visualize data over map. Visualizing raw data is not an efficient and not possible in most of the applications, beside that preprocessing millions of lines take days, in a regular configured personal computer, runs Matlab. We introduced a hybrid model composed of *In Memory Computing*, Matlab and Geographic Information Systems (GIS) tools to process big data to solve performance issues.

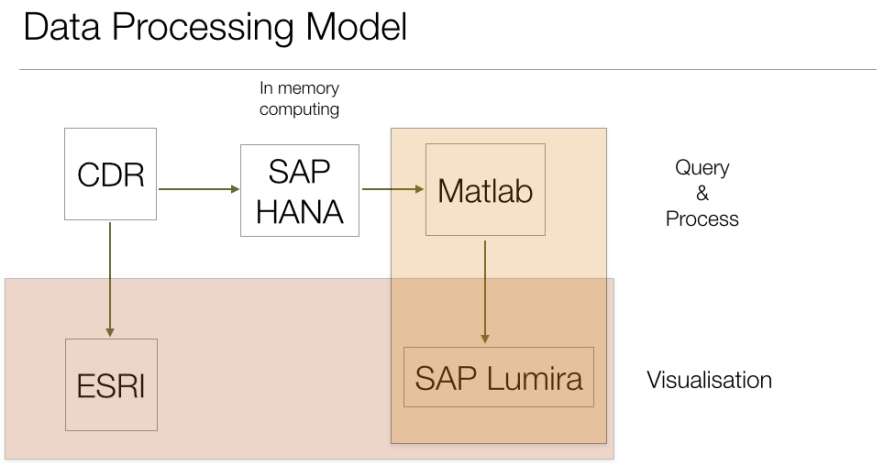

Figure 4.3. Data processing model.

4.1.1. SAP HANA - In-Memory Computing

SAP[16] is one of leader's in enterprise resource planning (ERP) sector which offers end to end solutions for various industries. In memory computing arise from the need of managing information workload that is populating from extensive amount of different data sources (i.e. Twitter[17] , Mail, Facebook[18]) [75].

[16]http://www.sap.com
[17]https://twitter.com
[18]http://www.facebook.com

In relational databases, data is stored row by row, which enables structured query languages (SQL) to fetch the related data from that query. However, recently data is generated in unstructured form, and this is problematic for traditional database, to store vast amount of sparse data sets. In-memory computing beside from it's hardware infrastructure, it brings column base storage of the data, that makes fast access to unstructured data [76].

4.1.2. Geolocation Information Systems

Locations in our world is coded by, geocoding systems, which means either latitude and longitude values or degrees for corresponding geographic location.

Location information is very valuable for analyzing behavioral patterns, Esri ArcGIS, is a GIS tool, enables spatial analysis, merge layers of information in one visualization. Information layers are stored in *Shape Files*, which store data not only as points but also as polygons which covers an area. This helps to visualize topological properties of the locations. QGIS [19] is an open source GIS tool, that covers most of the capabilities of ESRI ArcGIS.

In our incident detection model, single analysis should be performed before applying methodology over the country. Spatial querying let us find the geographic coordinates of the incident, and determining the corresponding antenna in that region.

4.2. Data Exploration and Mobility Analysis

In this section we explain how different features are extracted from the data different features of the data and visualize them. Since Set_1 includes aggregate communication between cell towers, it was visualized in http://www.geofast.net, we focus here on Set_2 (Mobility traces, fine resolution), and Set_3 (Mobility traces, coarse resolution).

[19] http://qgis.osgeo.org/en/site

Let U denote the set of mobile phone users, $U = \{u_1, u_2 \cdots u_n\}$, where N is the number of users. We use the same notation for all sets. These sets may have overlapping users, but since the data are anonymized this can not be predicted.

N equals 50,000 users in Set_2, 500,000 users in Set_3 denoted as U. Set_2 contains triplets of antennae usage $S2 =< u_i, a_j, t_j >$ where u_i denotes the user i, a_j denotes the antenna, and t_j denotes the time when a call takes place. Set_3 is similar, but the triplet we have $< u_i, s_j, t_j >$, with s_j that denotes the sub prefecture. Each sub prefecture contains a number of antennae.

$\forall\ a_i \in A,\ A = \{Latitude_i, Longitude_i\}$

4.2.1. Location Based Analysis

In this section we aim to discover significant locations of a user, such as home. Similar to the methodology used in [17], we focus on using threshold based analysis formulated as follows:

$$SpaceDistance(p_s, p_e) < D_{max}$$
$$TimeDifference(p_s, p_e) > T_{minutes}$$

4.2.1.1. Daily Work Locations. We have discrete spatio-temporal data. CDR data is recorded, when only subscriber makes a call. Evaluating data in hourly intervals, increasing the inference effort. Time slots are defined as shown below, to overcome this issue.

Time Slots defined for a day:

$$\begin{array}{lll} 9PM - 6AM & Slot_1 & Night \\ 6AM - 10AM & Slot_2 & Morning \\ 10AM - 5PM & Slot_3 & Day \\ 5PM - 9PM & Slot_4 & Evening \end{array}$$

In order to understand the subscribers' significant locations, we assumed that if the user makes call from the same antenna during the day, it should be annotated as the user's working place. In that work, we neglect the users, work at night and students with different patterns. To differentiate people who stay at home, these can be either housewives, or unemployed/retired people, we assume that people who commute between their homes and offices are in their office during the day, and at home at night. With the requirement of that the call locations are different during day and night, we can determine office locations. The noise is filtered with a threshold value.

Let τ is the threshold value, if we set $\tau = 0.7$ that means, at least 70% of the antenna IDs are in that time $Slot_3$.

$$\begin{aligned} &< OfficeLocation >_u = \\ &< a_j >_u = Count(< u_i, a_j, Slot_3 >_u)_\tau \end{aligned} \tag{4.1}$$

Here, we are using locations of antennae; this may be misleading us. Hence in this computation, rather than identifying exact home or work locations we can rather infer neighborhoods for living and working the same location. In that computation, for 50,000 subscribers, around 14,000 subscribers have the same daily and nightly location. It can give us, the information that they are not working, maybe house-wives or not having a structured pattern, or not using phone at home. The plot in Figure 4.5 shows the slums, it is focused on Abidjan for giving an idea about poverty zones.

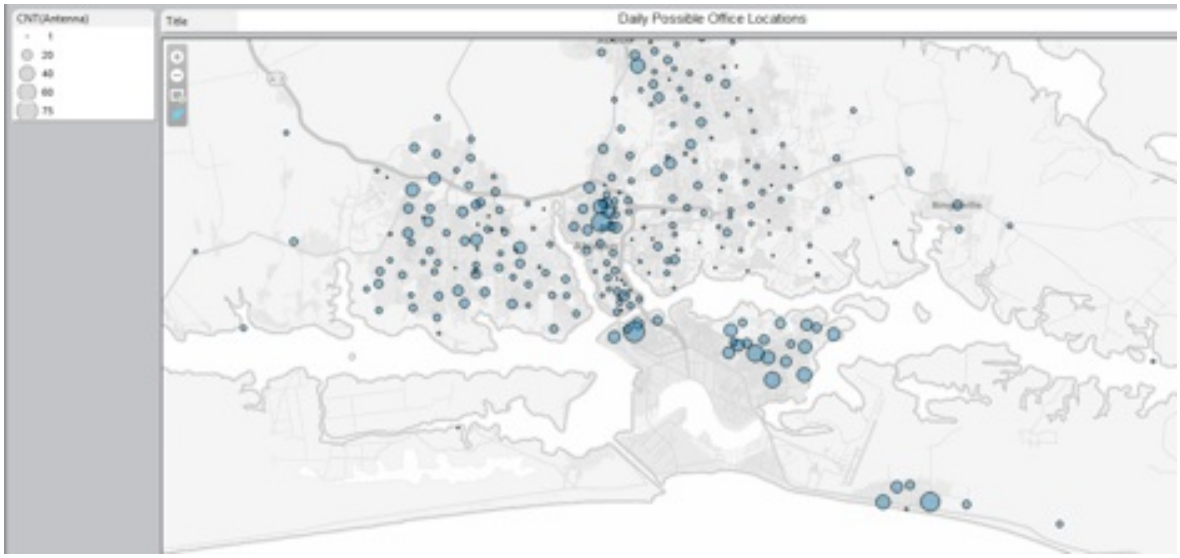

Figure 4.4. Daily work locations in Abidjan.

4.2.1.2. Home Locations. Yopougon and Northern part of Abidjan have slums. In Figure 4.6, dense home locations without schools can be evaluated as recently developing residential areas. Even in non-worker's plot in Figure 4.7, it is seen that the similar clusters of regions with low income.

If the antenna ID does not change during the night calls ; it is predicted to be the home location. (At least 70% same antenna ID, in that time slot)

Let th_1 is the threshold value, and set to $th = 0.7$

$$\begin{aligned} &< HomeLocation >_u= \\ &< a_j >_u= Count(< u_i, a_j, Slot_1 >_u)_{th_1} \end{aligned} \tag{4.2}$$

4.2.1.3. Non-Working Population. The non-working density is composed of subjects who do not have any assigned office or daily work entry in the data set.

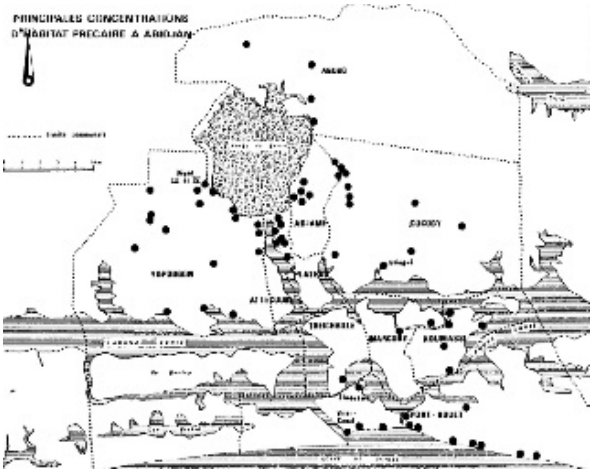

Figure 4.5. Abidjan rural residential area, taken from, (BNETD).

4.2.1.4. Mobility Diameters. In order to understand a subscriber's mobility, a pairwise longest distance was calculated for each subscriber for the given period. The Antenna locations were given as latitude, and longitude information. The longest distance was taken as the diameter, and the median of the two antennas form the circle of subscriber's mobility. In Figure 4.8, A-B-D-E-F-E-C-F-D-C-A is a sample trajectory of a user. In Figure 4.8 diameter is found, assuming the longest pair is A to D. This simple algorithm gives fast and approximate results. Blue line indicates the longest distance between pairs, and denoted as diameter, the centre is assigned the median of point A and D.

- Find the distinct nodes in the graph
- For each pair, find the distance
- Take the maximum distance of each iteration and assign as longest diameter
- Take median of pair as centre of the user

The distance between two points given in Latitude (ϕ) and Longtitude (Ω) is calculated by Haversine Formula 4.3 which is also suggested by US. Census Bureu, as the best way of calculating degree, although underestimating the ellipsoidal shape of the earth.

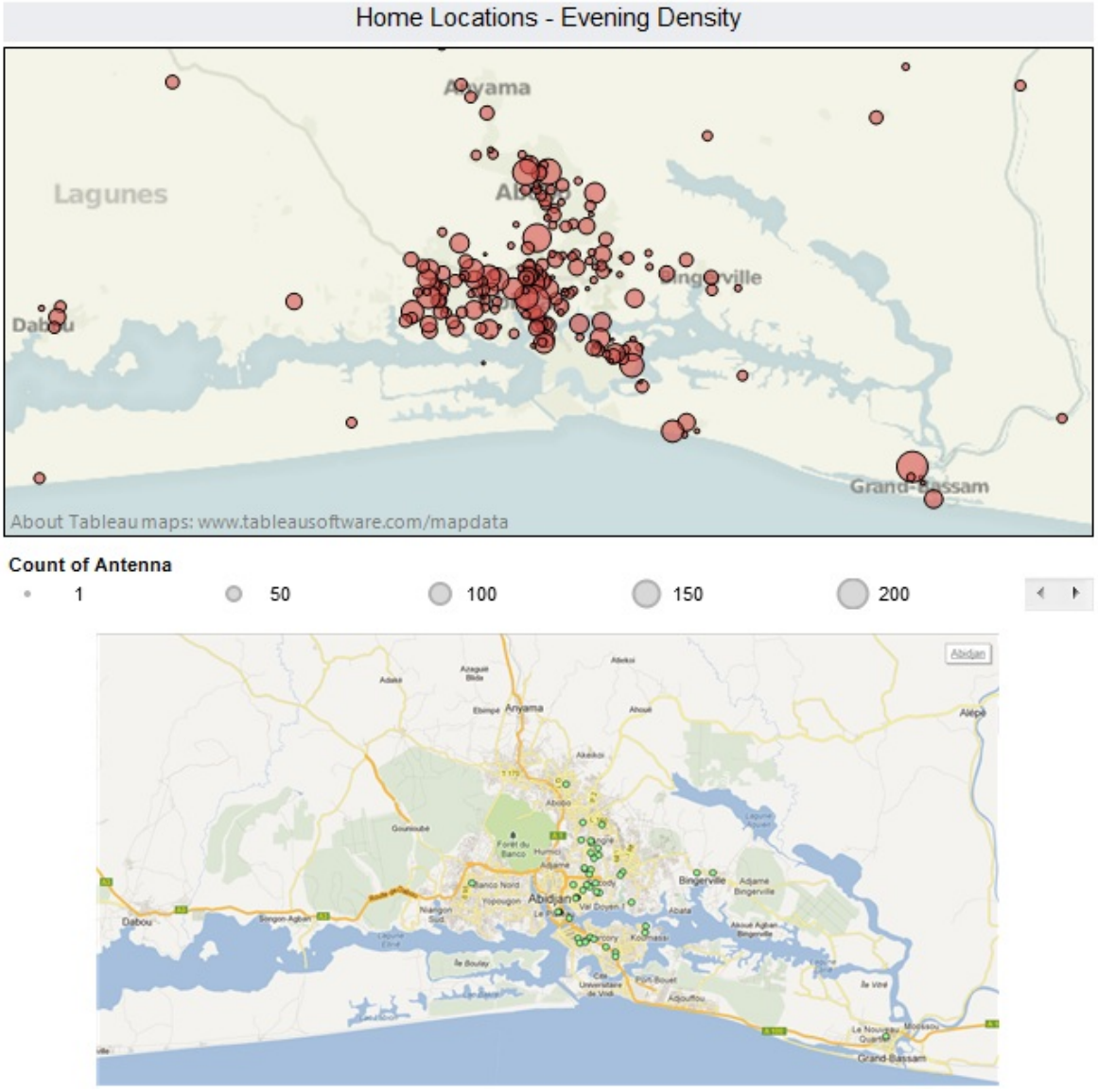

Figure 4.6. Abidjan, home locations and schools.

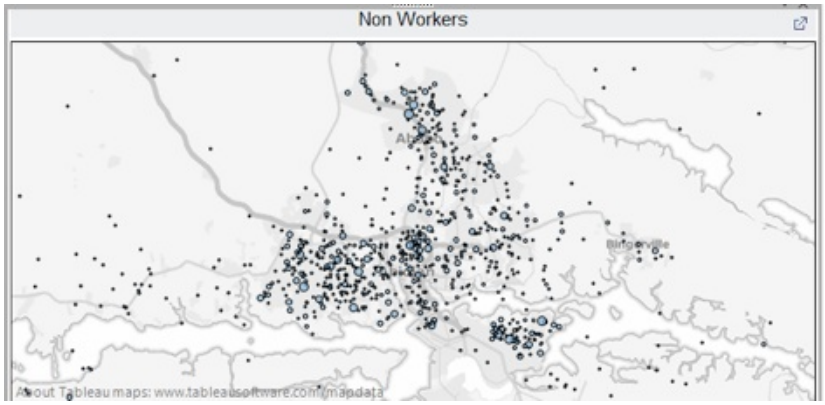

Figure 4.7. Non-Working population.

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\Omega_2 - \Omega_1}{2}\right)}\right) \tag{4.3}$$

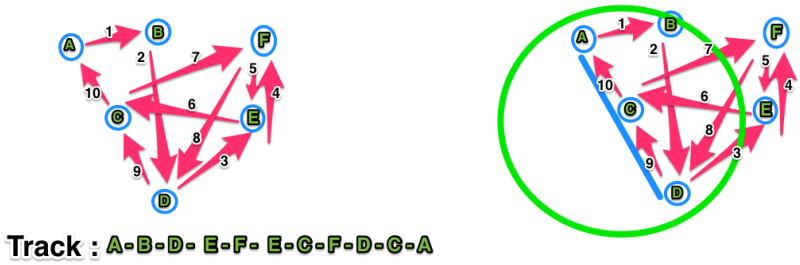

Figure 4.8. Sample path and mobility diameter of a user.

In Figure 4.9, for each user, ψ is calculated. If we denote $MobDiam(u, d)$, u represents user id, and ψ is for mobility diameter. The visualization has been plotted with Tableau Software, with low mobility and with high mobility respectively. Tableau Software is very flexible and from the link of each plot, it can be visualized over Web, for different zoom and view options.

If we analyze the plot, around the city Abidjan, people tend to move less, however the capital of the Côte d'Ivoire, Yamoussoukro, has people with higher mobility.

Abidjan is the biggest city in Côte d'Ivoire, but people visit Yamoussoukro for administrative purposes.

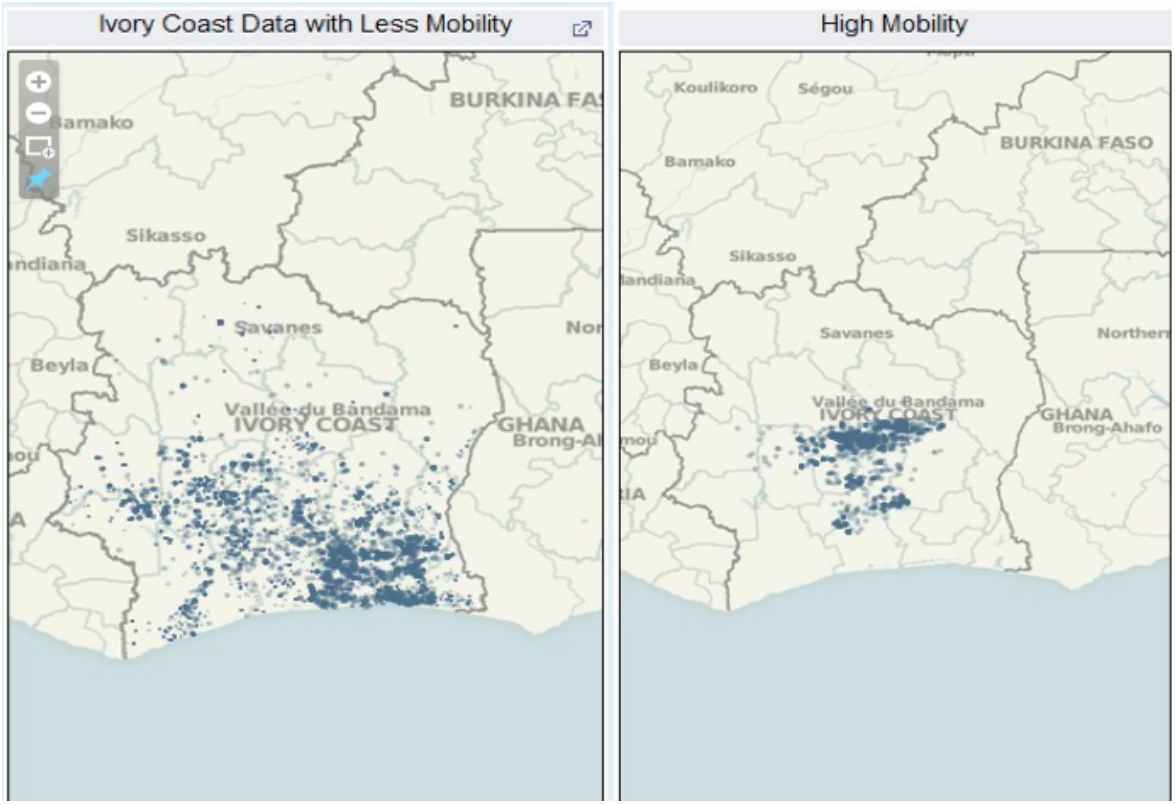

Figure 4.9. Mobility diameters.

By the help of such analysis, we figure out there is an anomaly in the region since only 30% of the subscribers can assigned to a home location. This questioning brings deeper analysis of the socio-political situation at that time period, and then we found the unstable chaotic conditions that take place.

4.2.2. Correlation Analysis

4.2.2.1. Talkativity Versus Mobility. The first visualization is on *Mobility versus Talkativity*. We define Mobility $\mu(\rho, t)$, as the number of distinct antennae used during a sampling period of ρ hours, averaged over a time frame of t hours.

Let Mobility for each user: $\mu(\rho, t)$: as the number of antennae switches during

the sampling period. $\mu(24, 240)$, shows the daily mobility average for 10 days.

The Talkativity $\tau(\rho, t)$ represents the number of outgoing calls (data set only includes call initiator's record) per sampling period of ρ hours similarly averaged over a time frame of t hours. $\tau(24, 240)$ is the daily talkativity averaged for ten days.

We use Set_2 in *Mobility Versus Talkativity* analyze, since this data set gives location sensitivity in antenna level. It is plotted as shown in Figure 4.10, they all have similar characteristics, which is tend to have less talk and less mobility, which is meaningful if we consider the daily expenditure of an average person is one Euro [20] . However this results only depend on when the user makes a call, so it may not show the exact behavior. Figure 4.10 shows a Power Law characteristics. This figure exhibits similar results with the work in [77], they observed that people devote most of their time to a few locations, although they spend their remaining time in 5 to 50 places.

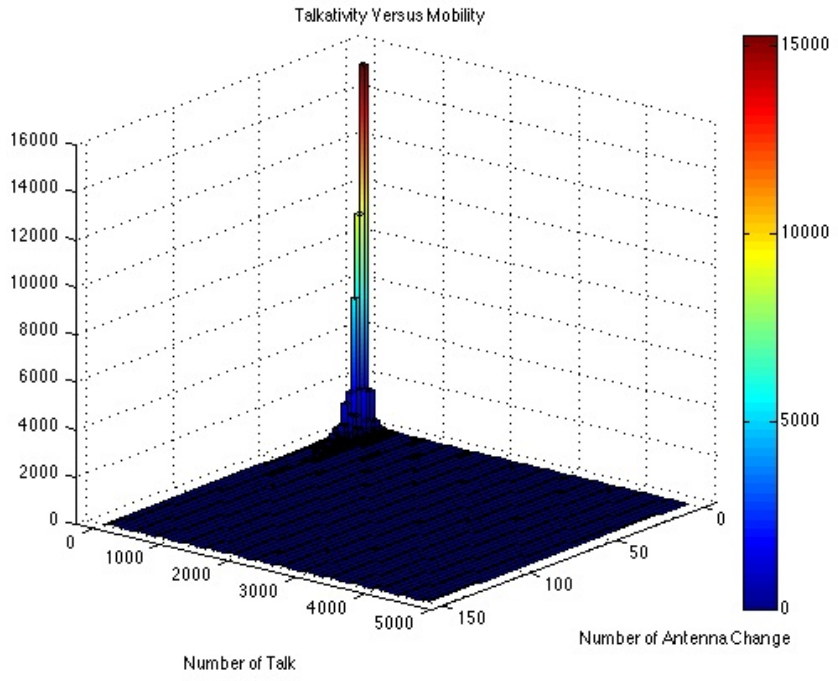

Figure 4.10. Talkativity versus mobility.

[20]http://www.ruralpovertyportal.org/country/home/tags/cote_divoire

4.2.2.2. Mobility Based Analysis. In order to understand the connectivity of each city, we analyze mobility patterns which are calculated per user in Section 4.2.1.4, and show in Figure 4.11. Bolder edge weight shows strong connectivity. Apart from their geographic closeness, people tend to visit different cities for commercial, business or social reasons. Accessibility in means of high way, or airway is also another parameter which increase connectivity. In Figure 4.11, Abidjan, the biggest city in Côte d'Ivoire is connected to Yamoussoukro as much as its neighbor cities.

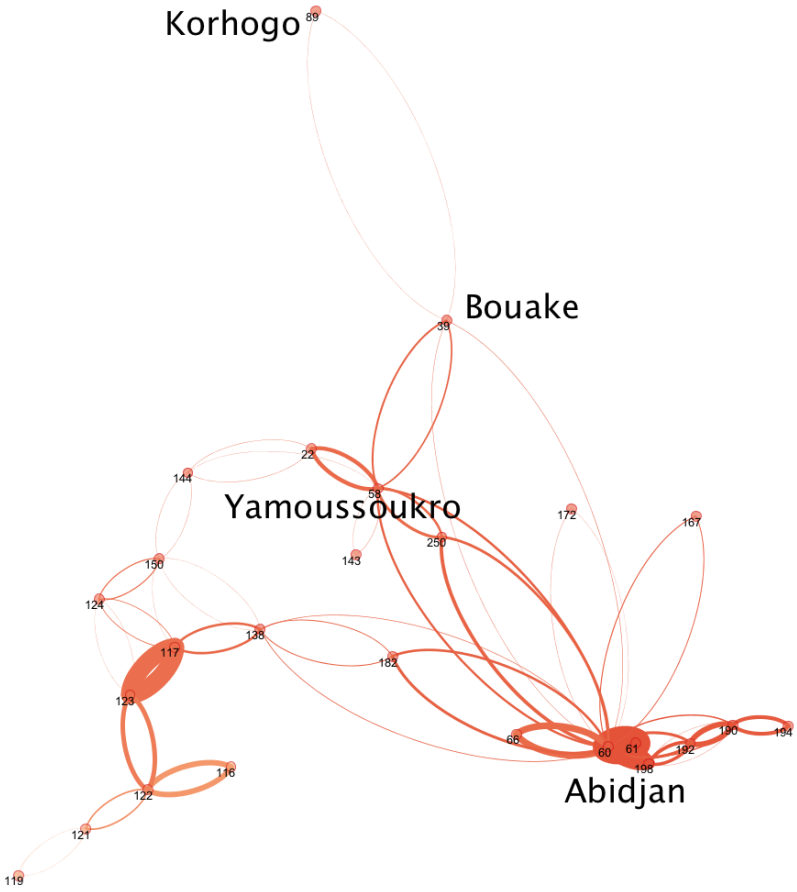

Figure 4.11. Most connected cities, edge weight represents the likelihood to travel between two cities.

4.2.3. Call Number per Antenna Analysis

In this section, we analyzed each antenna's call number counts per hour between December 2011 to 14 March 2012. We expect to see peaks in call count histograms, for either global or local events. In addition to that, we calculate hourly average call numbers per day, and difference from that average is plotted as in Figure 4.12. The blue line represents observations, the green lines represents average daily/hour calls, and black straight line shows the mean for the whole period.

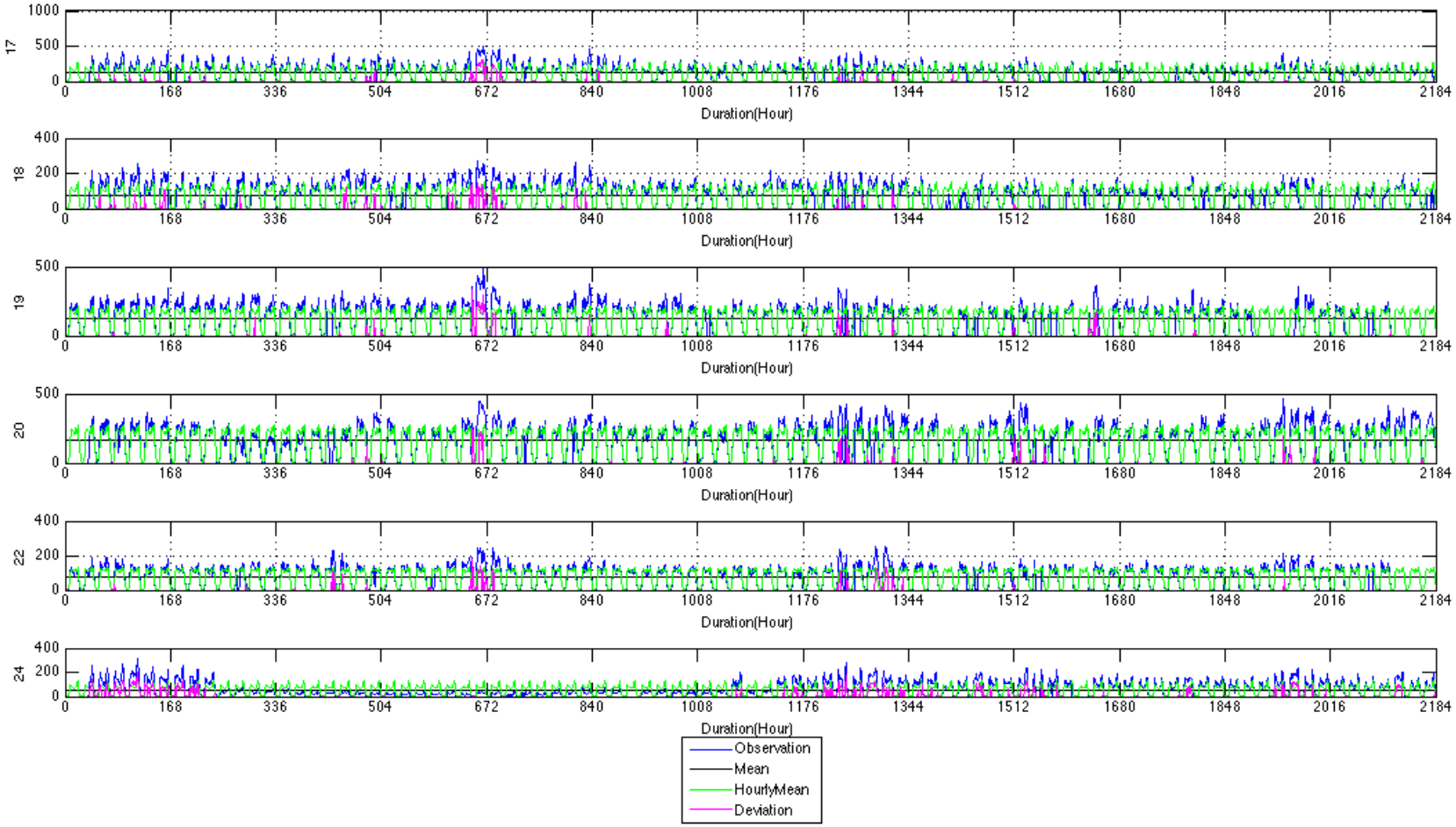

Figure 4.12. Mean and derivation from mean per antenna.

We expect to have a normal call behavior for each antenna, as in Figure 4.13.

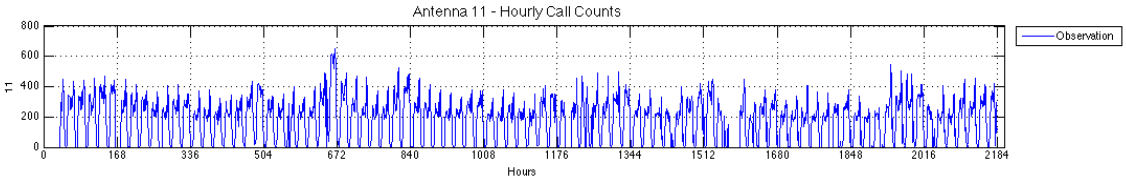

Figure 4.13. Antenna 11 call numbers per hour.

In the visualizations, it is detected that 322 antennae over 1028 show a very low call count between 13 December 2011 and 14 January 2012, like in Figure 4.14.

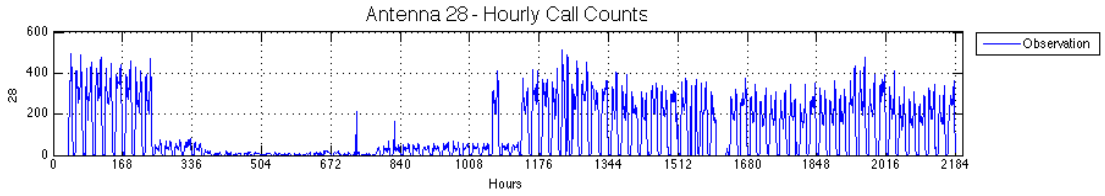

Figure 4.14. Antenna 28 call numbers per hour.

In order to understand if there is a displacement trend in that region for the given period, a cumulative call count histogram is depicted in Figure 4.15. The figure shows that this low call tendency in that region presents even in country level.

Map Figure 4.16 shows antennae with irregular patterns on map. The western part of the country presents below average call number from December 14 to January 14. The possible reasons is discussed in Chapter 6.

4.3. Markov Modulated Poisson Process Implementation

In this section Markov Modulated Poisson process implementation details are given for event detection through call counts per antenna per hour.

The model is represented as two state Markov chain, normal call behavior and abnormal event state. As detailed in Section 2.5.2.2, the transition between these states

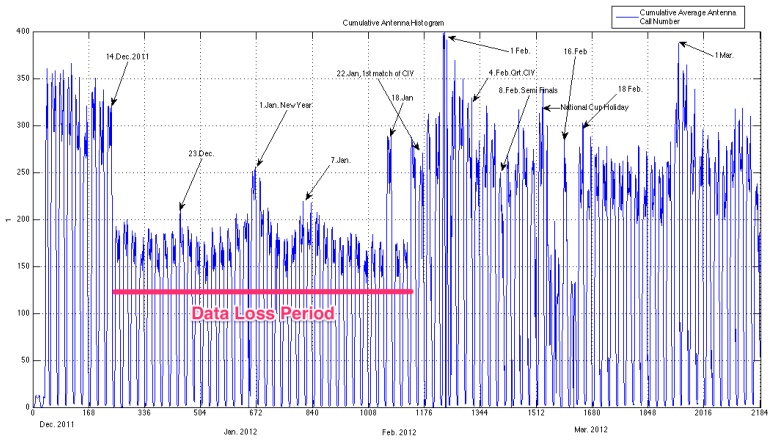

Figure 4.15. Cumulative antennae call numbers per hour with possible data loss period.

are defined with a transition matrix in Equation 4.4, which is time independent but dependent on previous transition probabilities.

$$z(t) = \begin{cases} 1, & \text{if there is an event in time t.} \\ 0, & \text{otherwise.} \end{cases} \qquad M_z = \begin{pmatrix} 1 - z_0 & z_1 \\ z_0 & 1 - z_1 \end{pmatrix} \tag{4.4}$$

As we can recall from Section 2.5.2.2, we use Monte Carlo methods when the functions that represent the data, are too complex and become intractable while we integrate out. Instead we assume the data is generated from a distribution and sampling from that distribution will converge to the expected value when we iterate enough times. Here, we assume that the number of calls are coming from a Poisson distribution Pois(N,$\lambda(t)$), with rate function $\lambda(t)$. We model change in the count data with periodicity depending on day and hour of the day and t represents the time interval in that perspective. In Section 4.3.1, this property will be used to calculate the posterior distributions.

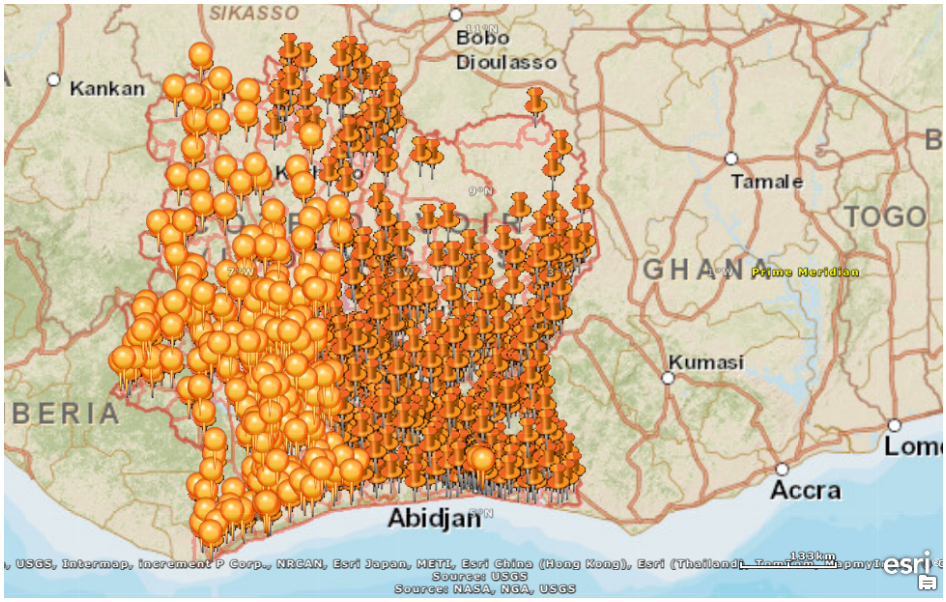

Figure 4.16. Characteristic antenna's location.

Let $N^k(t)$, represents total number of call for antenna k for time t. As shown in Equation 4.3 this call volume may be the reason of a normal call activity $N_0^k(t)$ and may include an activity coming from an anomaly $N_E^k(t)$ [5].

$$N^k(t) = N_0^k(t) + N_E^k(t) \tag{4.5}$$

The normal call pattern is dependent on which day (i) of the week (δ_i^k) and which hour(j) of the day $\eta_{j,i}^k$ and λ_0^k represents the average rate of antenna in one week. As in shown in Figure 4.18, we can formulate rate function of Poisson distribution $\lambda^k(t)$ as products of initial rate λ_0^k, $\delta_d^k(t)$ and $\eta_{d(t),h(t)}^k$. Plate notation is used to depict repeating form of the data. In our work, (t) is hour of the day (T), and model is periodic with respect to that notation. Antenna (k) based notation is neglected for simplicity in the rest of the thesis.

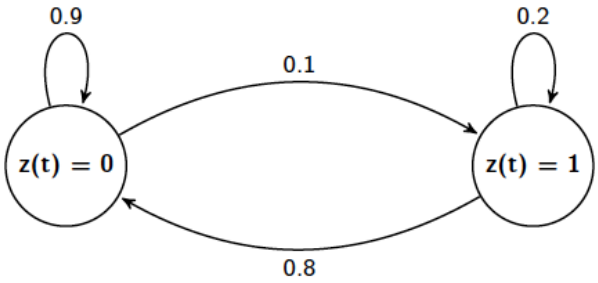

Figure 4.17. Event transition state diagram.

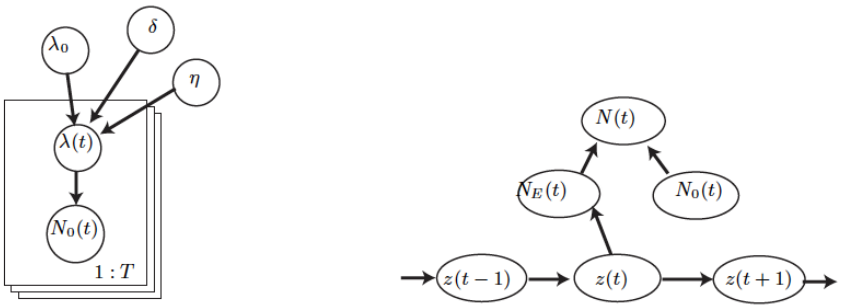

Figure 4.18. Markov chain representation of Poisson process, taken from [5].

.

Our observations denoted as $N(t)$, and hidden variables are as follows; the amount of call from normal call pattern $N_0(t)$, the amount of call initiated from an anomalous event $N_E(t)$ and the transition probabilities of events $z(t)$ as in Equation 4.4.

$$Pois(N;\lambda(t)) = e^{-\lambda(t)}(\lambda(t)^N/N!) \tag{4.6}$$

$$\lambda^k(t) = \lambda_0^k \delta_{d(t)}^k \eta_{d(t),h(t)}^k \tag{4.7}$$

Conjugate prior, ensures that the posterior distribution is coming from the same family as the prior. Gamma distribution is conjugate prior to Poisson distribution. By choosing conjugate prior to posterior distribution, our model become tractable, and can be written in close form. δ, η represents a series of value which total probability one, as shown Equation 4.8, means they can be generated from multinomial distribution.

$$\begin{aligned} \sum_{i=1}^{7} \delta_i &= 7 \\ \sum_{i=1}^{D} \nu_{j,i} &= D \forall j, \end{aligned} \tag{4.8}$$

Random variables δ, η that denotes day of the week and hour of the day, are coming from multinomial distribution so the conjugate prior is selected as Dirichlet distribution.

$$\lambda_0 \sim \Gamma(\lambda; a^L, b^L) \tag{4.9}$$

$$\frac{1}{7}[\delta_1, ..\delta_7] \sim Dir(\alpha_1^d ... \alpha_7^d) \tag{4.10}$$

$$\frac{1}{D}[\eta_{j,i}, ...\eta_{j,D}] \sim Dir(\alpha_1^h ... \alpha_D^h) \tag{4.11}$$

In Figure 4.18, $Z_{(t-1)}$, $Z_{(t)}$ and $Z_{(t+1)}$ shows the time series property of the event transition. $N_0(t)$ and $N_E(t)$ are hidden variables, and $N(t)$ is our observations in that respect.

$$z_0 \sim \beta(z; a_0^Z, 0^Z) \tag{4.12}$$

$$z_1 \sim \beta(z; a_1^Z, b_1^Z) \tag{4.13}$$

$$N_E(t) = \begin{cases} 0, & z(t) = 0 \\ P(N; \lambda(t)), & z(t) = 1. \end{cases} \tag{4.14}$$

$$\lambda(t) \sim \Gamma(\lambda; a^E, b^E) \tag{4.15}$$

Event probabilities can be sampled through densities of normal event $P(N; \gamma)$ times probability have an anomalous event $\gamma(t)$ as in Equation 4.16.

$$\int P(N; \gamma)\Gamma(\gamma; a^E, b^E) = Nbin(N; a^E, \frac{b^E}{1 + b^E}) \tag{4.16}$$

Hyper parameters of Gamma Distribution Γ, a^E and b^E, set the distribution's either sharpness or smoothness between transition states Z_0 and Z_1. In traumatic events, this transition generates a sharp peak, as shown in Figure 4.19. Estimation of these two parameters are shown in Section 4.3.1.

4.3.1. Inference and Parameter Estimation

If we know, { $N_0(t)$, $N_E(t)$, $z(t)$ } it would be straight forward to estimate the parameters of the distribution, by computing Maximum a Posteriori, as all other variables

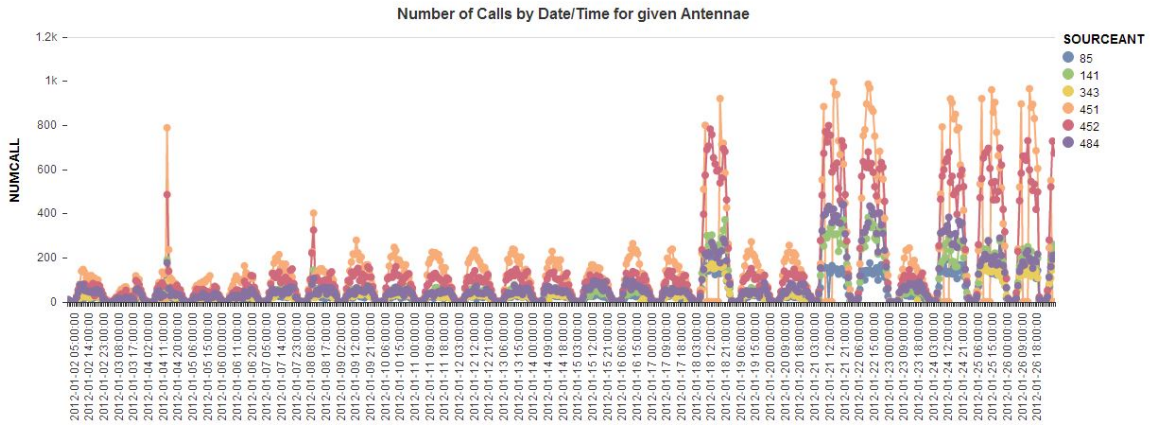

Figure 4.19. January call counts for sub-prefecture 138 Gagnoa.

are conditionally independent. However as in Figure 4.18, only $N(t)$ is observed.

In our model, we need to estimate the parameters for Gamma distribution $\Gamma\left(\gamma; a^E, b^E\right)$ for modeling probability of anomaly event $p(z(t)^k|N(t)^k)$. The rest of the distribution's parameters are estimated through mean value of the observations. However since we do not know when an event takes place, either we compute the parameters or estimate within the given information. In that respect, we take one antenna, which is known that has an entry in UN Security report, and calculate the deviation from the mean value of hour of the day, and day of the week from the observation. After smoothing this observation for single antenna, we take this as $N_E(t)$ for the specific antenna.

We estimate the Gamma distribution parameters from the data in two forms, first moment base estimation [67, 78], and second maximum likelihood estimation of the posterior distribution [79].

4.3.1.1. Moment Base. Moment base computation is as follows; M_1 and M_2 stands for first and second order moments of gamma distribution, respectively.

$$M_1 = \mathbb{E}[N_E(t)] = a^E b^E \tag{4.17a}$$

$$M_2 = \mathbb{E}[N_E(t)^2] = a^E (b^E)^2 + (a^E)^2 (b^E)^2 \tag{4.17b}$$

and from that definition;

$$a^E = \frac{M_1^2}{M_2 - M_1} \tag{4.18a}$$

$$b^E = \frac{M_2 - M_1^2}{M_1} \tag{4.18b}$$

The expectation value of $\widehat{M_1}$ and $\widehat{M_2}$ can be calculated through Equation 4.19 and expectation values can be replaced in Equation 4.18 to calculate estimated values of a^E and b^E.

$$\widehat{M_1} = \frac{1}{N} \sum_{n=1}^{N} N_E(t) \tag{4.19a}$$

$$\widehat{M_2} = \frac{1}{N} \sum_{n=1}^{N} N_E(t)^2 \tag{4.19b}$$

4.3.1.2. Maximum Likelihood. As second alternative maximum log likelihood approximation of event distribution, which is denoted as $N_E(t)$ for one antenna.

If we open the Gamma distribution's probability mass function as in Section 2.5.1.4.

Let we have T observations of event, all denoted as N_E.

$$N_E(t) \sim \gamma(t) \sim \Gamma(N_E; a^E, b^E) = \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} exp(-\frac{x}{\beta}) \tag{4.20}$$

$$N_E(1:T) \sim log(p(N_E; a^E, b^E)) \tag{4.21}$$

$$= \sum_{i=1}^{T} logp(N_E(i)|a^E, b^E) \tag{4.22}$$

$$= \sum_{i=1}^{T}((a^E - 1)log(N_E(i)) - b^E N_E(i) \tag{4.23}$$

$$- log(\Gamma(a^E) + a^E log b^E)) \tag{4.24}$$

If we calculate the derivatives of Equation 4.21 with respect to a^E and b^E, and find the solution which make these computation equals to 0. Derivation with b^E is straightforward as you see in Equation 4.25.

$$\frac{\partial}{\partial b^E} = \sum_{i=1}^{T}\{-N_E(i)\} + Talog(b^E) \tag{4.25}$$

$$b^E = \frac{Ta^E}{\sum_{i=1}^{T} N_E(i)} \tag{4.26}$$

However $\frac{\partial}{\partial a^E}$ can not be written in closed form, since log$\Gamma(a^E)$ derivation is intractable [79].

4.3.1.3. Approximation from Negative Binomial Distribution. Greenwood and Yule derived how Poisson function behave like negative binomial distribution where the variance of the distribution is greater than the mean of the Poisson distribution [80]. If we

recall properties of negative binomial distribution in Section 2.5.1.3, scale and mean value can be calculated as follows.

If we assume that for one antenna the peak values shows that there is an event.

$$N_E(t) = N(t) - N_0(t) \tag{4.27}$$

If we take $\hat{N_0}(t)$ as mean value of the distribution, with respect to the hour of the day and day of the week means. Therefore we can approximate negative binomial distribution's parameters, as in Equation 4.28.

$$\mathbb{E}(x) = \frac{pn}{1-p} \tag{4.28}$$

5. EXPERIMENTS AND RESULTS

In this chapter we applied our model which is described in detail Section 4.3, to D4D data to detect anomalous events. We used Set_1, as it includes antenna base call counts with hour precision. As we explained in Section 4.1 we store data in SAP Hana, offered by Amazon as a service to overcome querying performance of *Big Data*, analyze with Matlab and visualize with various tools. Our aim is to report the results in terms of accuracy, for this purpose United Nations Security Council, United Nations Refugee Agency (UNHCR), International Fund for Agricultural Development (IFAD), United Nations Office for the Coordination of Humanitarian Affairs (OCHA) databases, are scanned to obtain correlative information with D4D data. In Appendix C the security related incidents are listed and in Appendix D country base main events are merged with UN Security report data.

5.1. Experimental Setup

Before evaluating the performance of our methodology of using Markov modulated Poisson process for event detection, in order to gain insight about the events data, we visualized the temporal changes in terms of number of total calls per antenna. For this purpose, we plot the hourly event probability distributions and created a video from these plots that visualize the events happening in an hourly basis. Table C.1 is used as a baseline, and we plot the incidents for the given time and region and compare the probable activity map, side by side.

Our experimental setup is as follows and the details of each step will be presented in Section 5.1.1 to 5.1.4;

- Annotation : Manual querying of Côte d'Ivoire map for antenna locations.
- Training : Markov modulated Poisson process parameter tuning.
- Visualization (Qualitative Evaluation) : Histogram and map visualization .
- Matching (Quantitative Evaluation) : Evaluating results with Appendix C.1.

From 12 December 2011 to 30 January 2012, the possible data loss or corruption in the data set, let us decide to leave this period out for 322 antennae, as mentioned in Section 4.2.3.

5.1.1. Annotation

As the first step, we determine a date that an incident took place from Appendix C.1 from United Nations Security report and Appendix D are merged nation wide events obtained from several data sources. According to our ground truth, a couple of violent incidents took place in Western Côte d'Ivoire, and total of 16 deaths have been declared. More specifically, in Peite Guiglo a small village close to Guiglo, on 4 January 2014 an incident happened. We annotate Peite Guiglo and antennae around that region as shown in Figure 5.1.

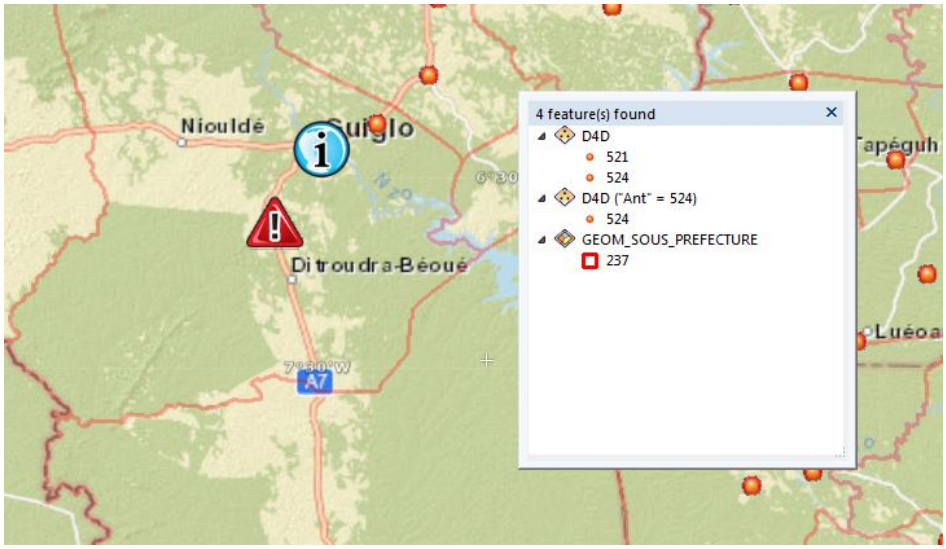

Figure 5.1. Petie Guiglo antenna locations.

After determining the antenna numbers, as {524,501,503}, the raw data of each antenna is visulized as shown in Figure 5.2. The peak for the event on 4th of January and another peak on 7th of January is visible, possibly there was an event but this was not reported in the sources we utilized.

Figure 5.2. Petie Guiglo antenna histogram for January 2-16.

5.1.2. Training

Markov modulated Poisson process is an unsupervised learning method, however the model's performance for event detection is highly dependent on priors. We define transition kernel and gamma distribution's parameters which generates $\gamma(t)$ as prior. Gamma distribution's parameter for event is defined as a^E and b^E. We first implement a similar works' priors as input for a single antenna. As stated in Section 4.3.1, we estimate gamma distribution parameters with two different methodology. First moment base approximation, and second negative binomial approximation.

Instead of evaluating whole antennae set, we run the algorithm and parameter sets for a single antenna. The most probable event distribution of a single antenna is given as you can see in Figure 5.3. The reason for selecting Antenna 113 is, the location. The western part of the country has a data inconsistency, for approximately 6 weeks. Therefore we analyze an antenna in western region, with an entry in the ground truth table of United Nations [2].

For unit test we run our Markov modulated Poisson process (MMPP) on antenna

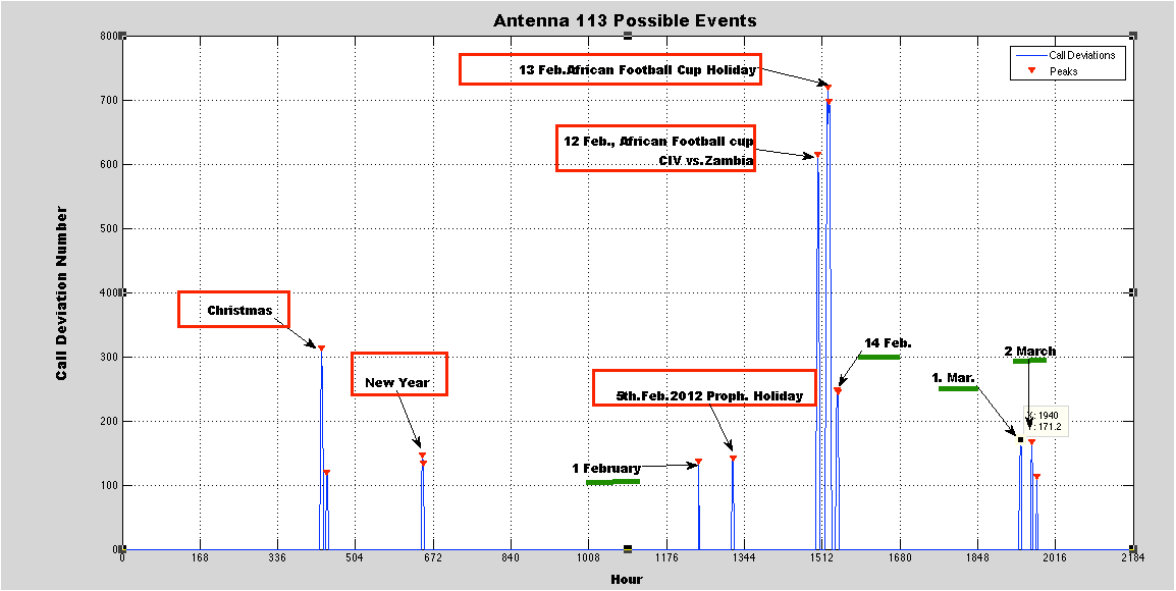

Figure 5.3. Antenna 113, calls possibly coming from an event.

113, with the parameters ($a^E = 5$ and $b^E = 0.333$) used in [5]. The reason for testing with the same set of parameters was that, we aimed to see how MMPP model performs in a different setting; in a single sensor was used to predict the events taking place in a building, whereas here we are looking at the total number of calls served by an antenna. The result of the posterior distribution from December 2011 to March 2012 is shown in Figure 5.4. The blue line represents the observation of hourly number of calls. Red line denotes the generated data, coming from the model. The results, in Figure 5.4 show that almost everyday there is a possibility of having an event. Although we do not have any information regarding these days, it is much more probable that we need to tune the parameters. Maximum likelihood equals to -82347.7 for this parameter set.

As the next step, we explored whether different values for the parameter set would exhibit better results. We implement moment based parameter learning algorithm for gamma probability distribution (Section 4.3.1). The calculation results are as follows; $a^E = 397$ and $b^E = 0.02$. As in Figure5.5, the posterior of the event represents very low probability profile which is almost very unlikely to have an event. However, if the probability distributions is magnified, this parameter set behaves better than first set, as events predicted in right time. In both experiments model is iterated 100 times with

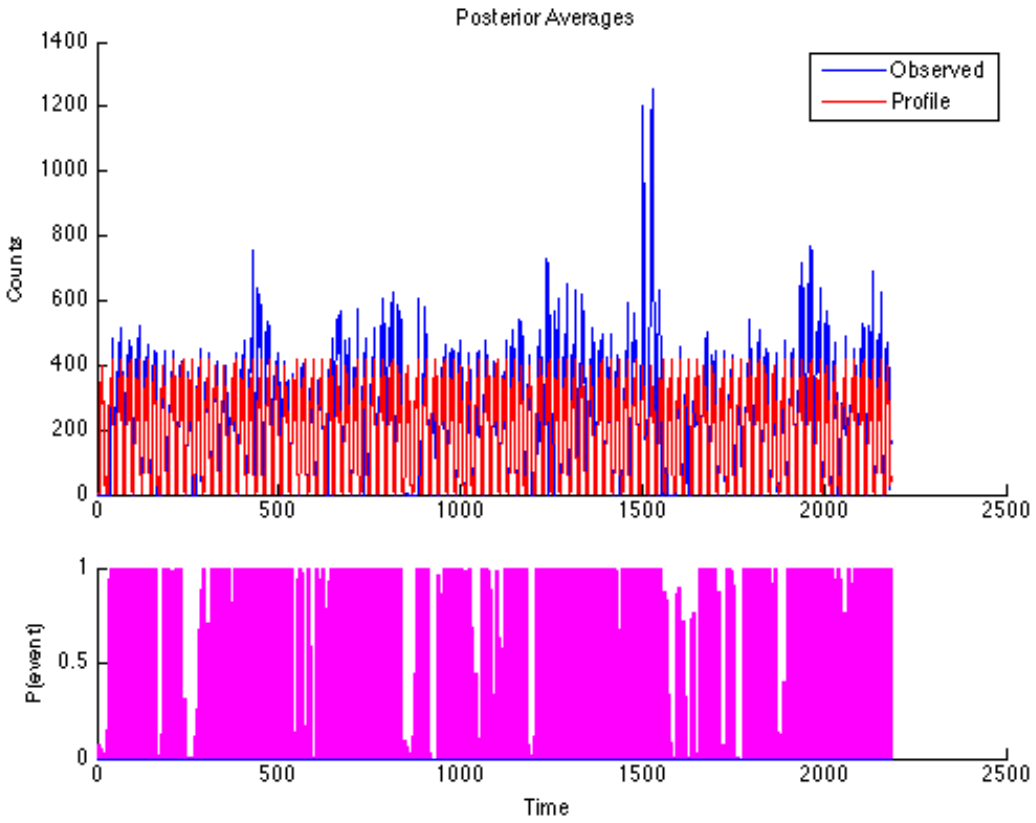

Figure 5.4. Experiment-1 Arrah town Antenna 113.

20 burn in period. So we test the iteration amount 200 and burn in period 50, but this change does not effect the results as ML changed from -102307.2 to -102307.1. The reason would be the duration of the experiment. As stated, the number of samples is important for accurate results [78].

The next experiment set is to approximate possible event distribution for antenna 113 to Negative binomial distribution. The mean value of the event intervals are, calculated as shown in Section 2.5.1.3, $a^E = \frac{\text{Time Intervals between peaks}}{\text{Number of Peaks}} = 154$. And b^E denotes the probability to have an event. In our case, it means ;

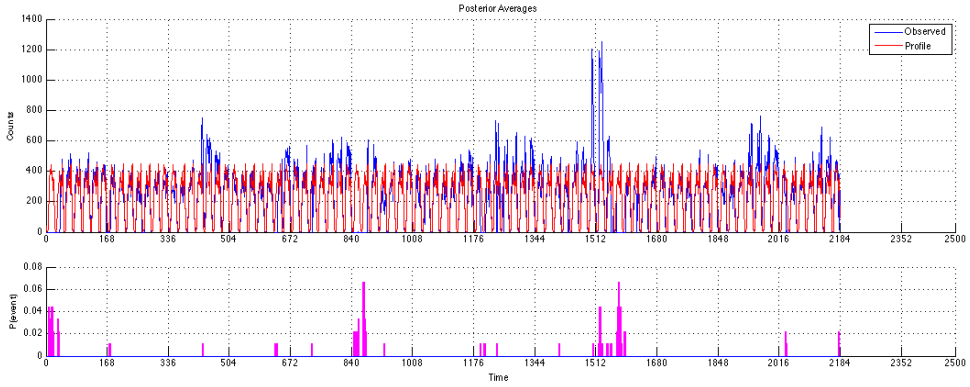

Figure 5.5. Moment based parameter estimation result [78].

$$b^E = \frac{\text{Number of Events}}{\text{Duration of the test}} = 13/91 = 0.14$$

The results of MMPP for antenna 113 is shown in Figure 5.6. Our method correctly detect almost all anomalous events annotated in Figure 5.3. In Table 5.1, the experiments results are given for model parameter estimation.

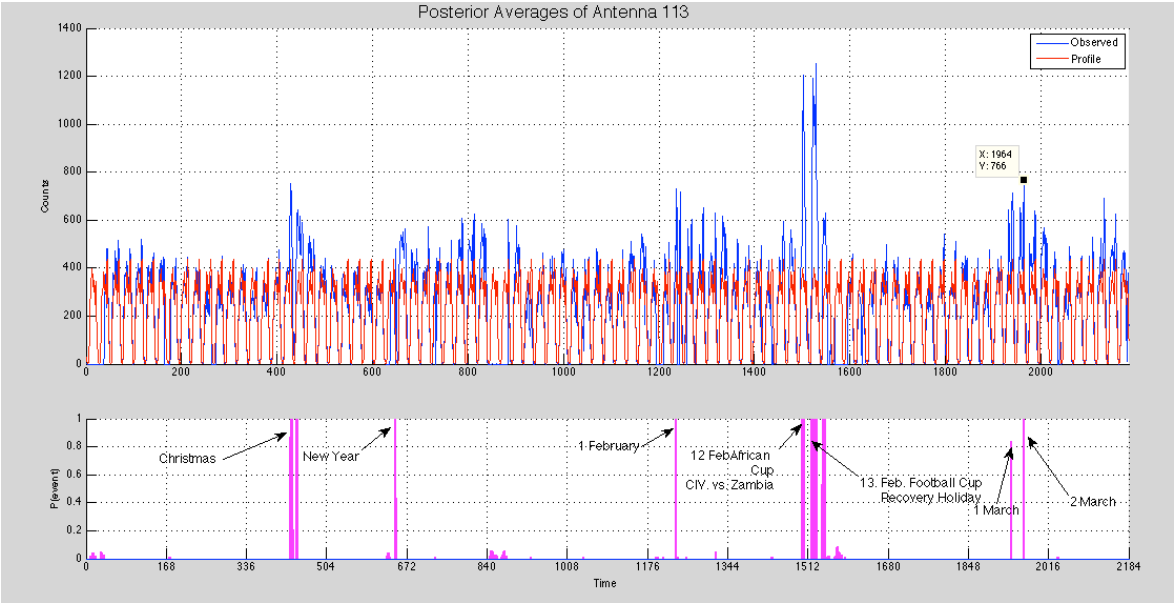

Figure 5.6. Antenna 113, negative binomial fitting.

Table 5.1. Experiment setup matrix for single antenna.

Exp.No	**Method**	a^E	b^E	**Iter./Burn.**	**Trans. Kernel**	**MLE**	**Comments**
1	Ihler [5]	5	0.333	100/20	z01=0.01, z00=0.99, z10=0.25, z11=0.75	-82347.7	Overestimated
2	Moment base	397	0.02	100/20	z01=0.01, z00=0.99, z10=0.25, z11=0.75	-102307.2	Very low probabiliy profile
3	Moment base	397	0.02	200/50	z01=0.01, z00=0.99, z10=0.25, z11=0.75	-102307.2	Iteration does not effect
4	NB Par.	154	0.14	100/20	z01=0.01, z00=0.99, z10=0.25, z11=0.75	-90692.6	Better results
5	NB Par.	154	0.1	100/20	z01=0.01, z00=0.99, z10=0.75, z11=0.25	-92422.9	Tuning NB fits better, trans. kernel not so effective

We assume in our model, the peaks from the mean values from day of week and hour of day, is taken as events as shown in Figure 5.3. We compare the events annotated in Figure 5.3 with the computed probabilities in Figure 5.6. We annotated totally eight events, five events are existing in the event database in Figure 5.3, two events are showing a periodic pattern in the beginning of each month, and one day 14th of February which is not existing in the database. Since we do not know all events, 14th February is also taken as event for evaluation. The model observed, five event out of six annotated event, and observed both begin of the month events and 14th of February unknown event.

The accuracy and precision of the experiments are given as follows:

Table 5.2. Results evaluation matrix.

		Ground Truth	
		Positive	Negative
Findings	Positive	7	-
	Negative	1	-

5.1.3. Visualization

In the unit test, we analyzed the data in micro level just for one antenna for a given period. We have 1024 antennae in our data set which cover the whole country. The visualization of posterior of each antenna over the map in time series, would enable us to have more insight for the country. A sample visualization is shown in Figure 5.7.

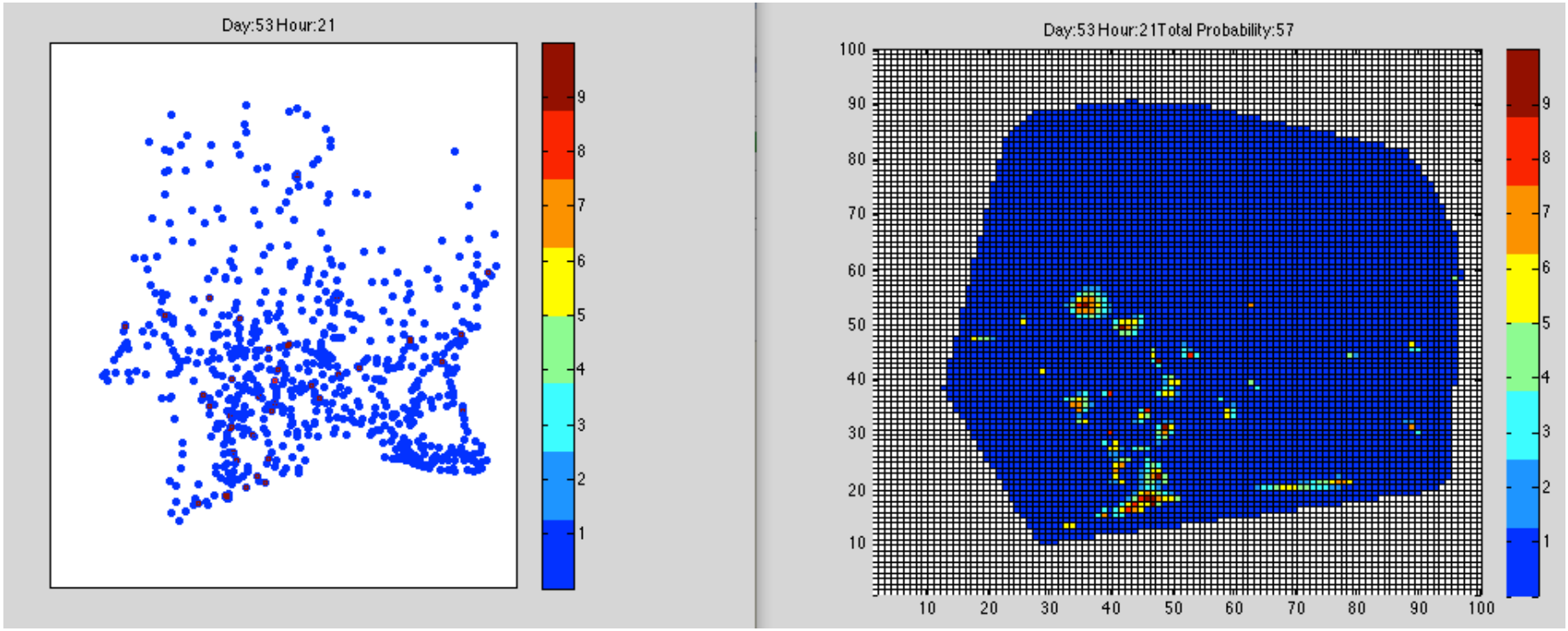

Figure 5.7. Snapshot from country event probability map.

5.1.4. Classification

As the final step, we report the performance of the method in terms of correctly classified events, with respect to Figures D.1, D.2 and Appendix C.1.

5.1.5. Experiments

In this section, the data for the whole period from 2nd of December 2012 to 11st of March coming from all antennae are analyzed using the MMPP method with the parameters explained in Section 5.1.2. The best result set obtained by negative binomial fitting, is applied to all antenna's call count data.

Figure 5.8. Bouake Katiola road.

We first show single antenna analyze, then analyze all antennae in that region, and at last aggregated call volume for given sub-prefecture, and evaluate the results. In that experiment we analyze the event happened in 3 February 2012, in Bouaké-Katiola road, as shown in Figure D.2 and Appendix C. This sub-prefecture is called Katiola, and it is located in the east part of the country. There is four antennae in that region $\{583, 74, 616, 964\}$ which are obtained by Esri[21] . Antenna 74 is not present during the

[21]http://www.esri.com

whole data collection period, therefore it is not included in the analyze.

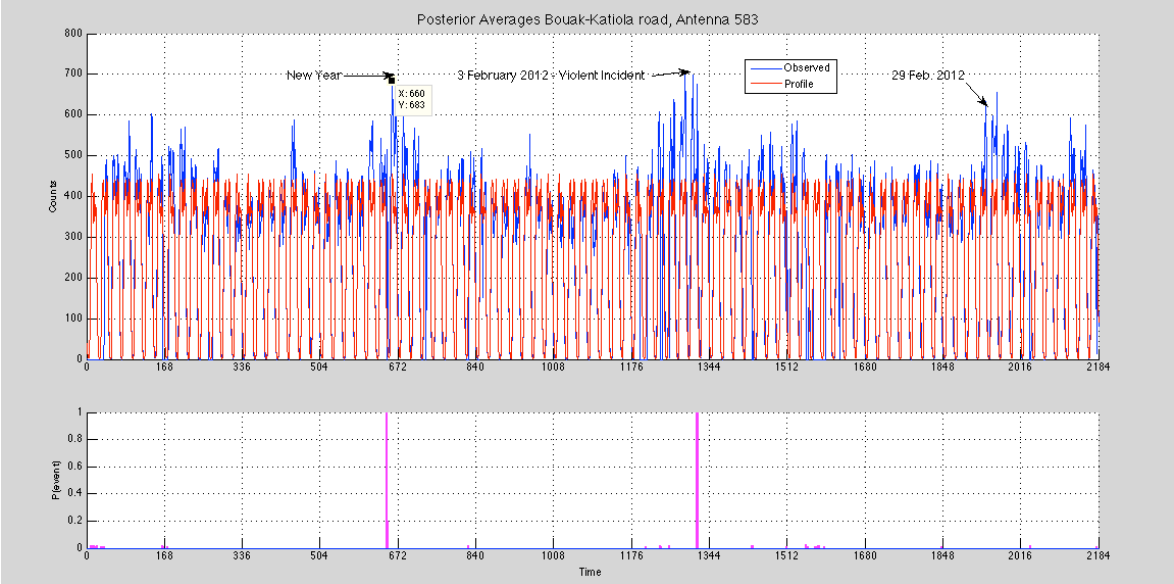

Figure 5.9. Antenna 583, Bouake Katiola road.

In Figure 5.9, we detect the event in 3rd of February and New Year. However, we could not detect this event in antenna 616 and 964. The reason for this is possibly the event did not happen close to those antennae. In the meanwhile, aggregating data in that region to detect events, results are more than expected as shown in Figure 5.12.

We find out that, local security incidents, only effect closest antenna. In our example, the probable location of the incident, which is defined in UN Security reports is close to Katiola, rather than Bouake.

We leave antenna 74 from single MMPP analyse, as it does not exist in the whole data collection period. But put the existing data of all four antenna, and aggregate in Figure 5.12. As you can see from the experiment, the aggregated data does not successful to catch the events precisely. It is tend to give false-positive events instead.

As a last experiment we evaluate the aggregated call data for all antennae, as shown in Figure 5.13. The annotation shows the most important dates and also the visible peaks which are out of our event database. Then we plug this call volume data,

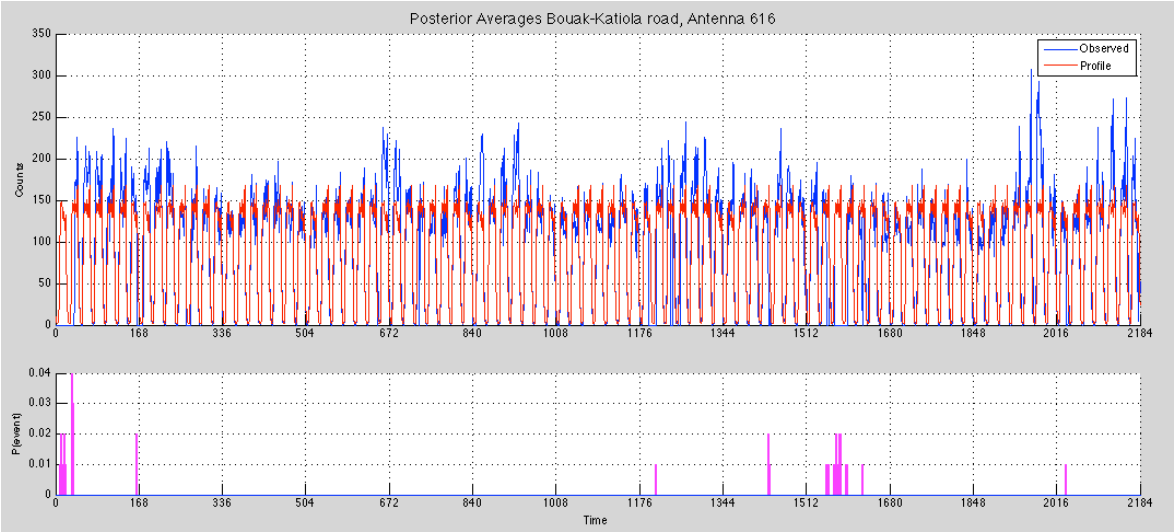

Figure 5.10. Antenna 616, Bouake Katiola Road.

into MMPP model, visualized in Figure 5.14. We expect to see the global events with that experiment, because local effects are suppressed by the higher call volume regions. However, if an event happens in Abidjan, it dominate the country profile as well.

Our data set is composed of both events exists in our ground truth, and events that we do not know the existence. In our experiments, we observe that, there are some time windows in some regions that it is very highly an anomaly event take place. However we can not prove this is an event, with our current knowledge. That is why we evaluate our model based on our ground truth and we do not evaluate *False-Positive* and *True-Negative* events.

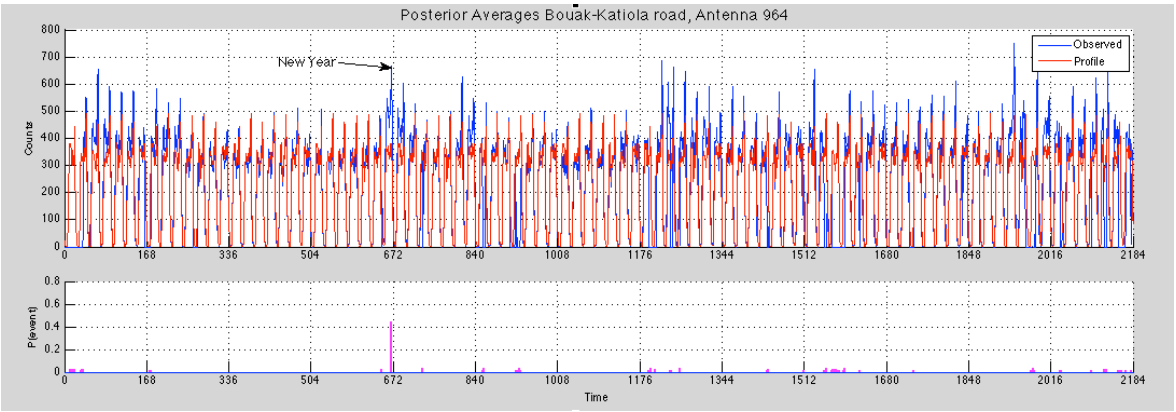

Figure 5.11. Antenna 964, Bouke Katiola road.

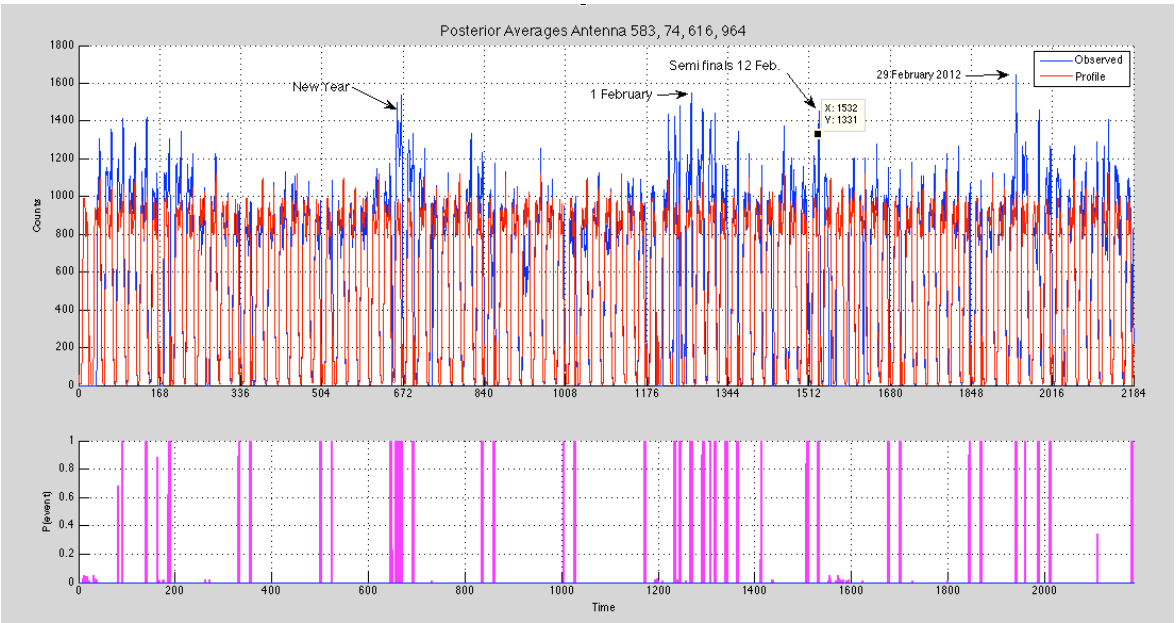

Figure 5.12. Aggregated results for Bouake Katiola road.

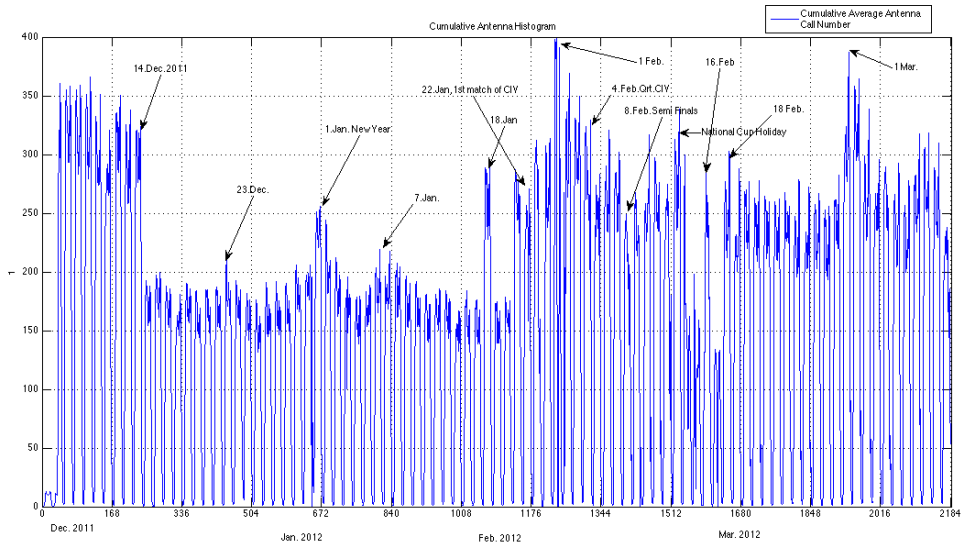

Figure 5.13. Aggregated call count per antenna for CIV.

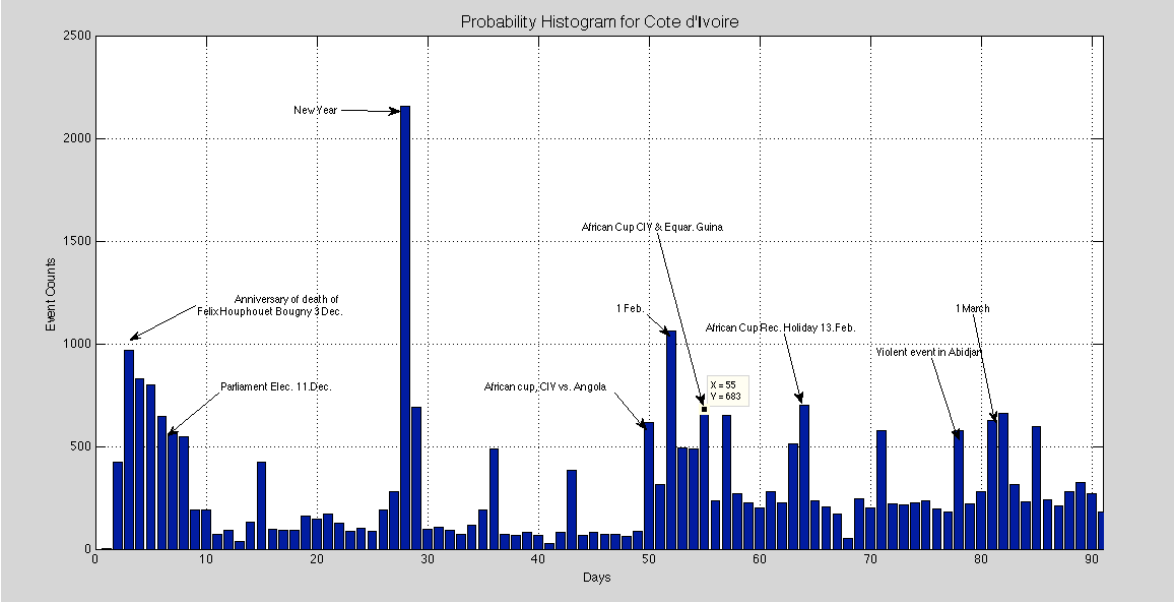

Figure 5.14. Aggregated event probabilities for CIV.

6. CONCLUSIONS AND FUTURE WORK

In this thesis, we explore how we can analyze human behaviors by using a mobile phone data set, named D4D, collected by a telecom operator company in Côte d'Ivoire in the beginning of 2012. At first, we focus on the analysis of significant locations of users, (i.e. home and office). However, since the vast amount of people do not exhibit a regular home-office-home pattern; we rather focus on event detection from mobile phone data. Additionally, since there was a civil war going on in the country during the data collection period, we explore the identification of events through the spatio temporal differences in call volumes measured in the country.

We utilized the Markov modulated Poisson process method to analyze the data. The reason for selecting a probabilistic method is that, being able to model anomalous event distribution in an analytic way. During the analysis, we expand the time information from its' linear dimension to a multidimensional matrix. Information such as day of the week and hour of the day for consecutive weeks are analyzed. The different patterns for weekday, weekend and the hourly differences are clustered. We visualize the results of this analysis for 1238 antennae over the country. In order to quantify the success of our model, incidents reported in the United Nations security reports are used as the ground truth information. In our evaluations, the effect of Markov modulated Poisson process parameters and the effect of the duration of the experiment are analyzed and the results are listed in terms of True-Positive ratios. In addition to detecting events, we evaluated the performance of our model to predict social events, such as African Football Cup that took place during the data collection period.

As a future work, we plan to analyze D4D Senegal challenge [22] with the same methodology and compare with D4D Côte d'Ivoire results. In addition to that, instead of batch processing of call data, another methodology can be examined for stream processing, to assess instant action to anomalous events.

[22] http://www.d4d.orange.com/en/presentation/the-objectives-of-the-challenge

APPENDIX A: Côte d'Ivoire Map

Figure A.1. Road map of Côte d'Ivoire, taken from Ezilon.com.

APPENDIX B: Data Visualization Tool: Gephi

Gephi Software is an open platform for Graph Visualization. It is Java based and works on Windows, Mac OS, and Linux. It is open licence. It is very powerful, to handle large networks up to 50,000 nodes and 1,000,000 edges. It is flexible to put filters on Graph properties such as, In Degree, Out Degree, managed in overview cockpit Figure B.3. In Figure B.1 a simple histogram of user call location is seen. The first step to work with Gephi is, to prepare node and edge file in CSV(Comma Seperated Values) format. The first row must includes *Source*, *Target* column names, in transferred file. There are couple of layouts in Gephi, in this thesis *Geo Layout* is used, the graph nodes are defined as latitude and longitude information, shown in Figure B.2.

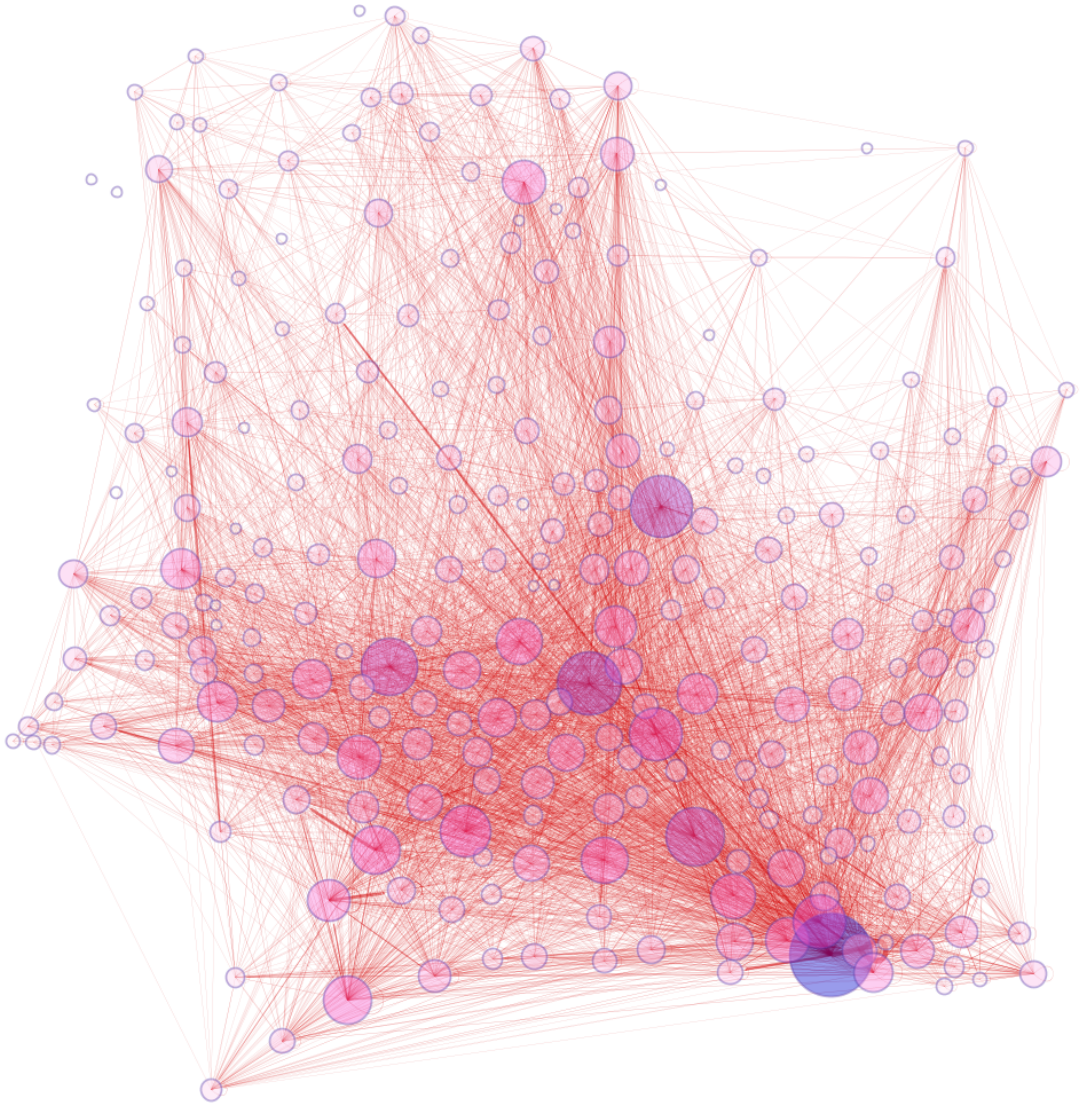

Figure B.1. Trajectory histogram visualized by Gephi.

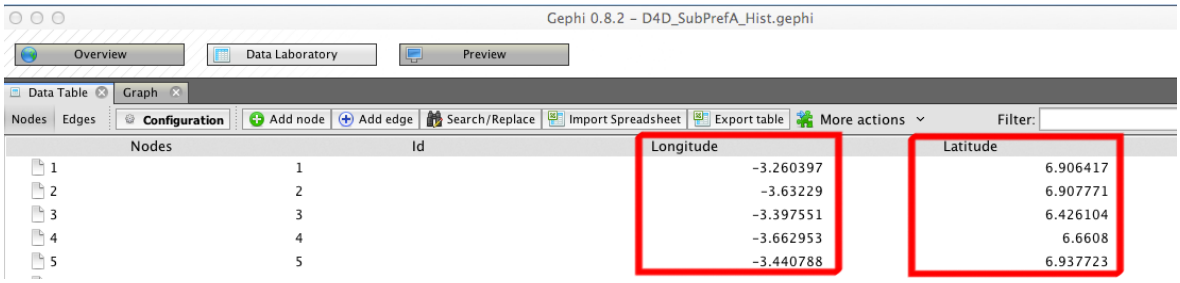

Figure B.2. Gephi data laboratory shows the graph nodes with latitude and longitude data.

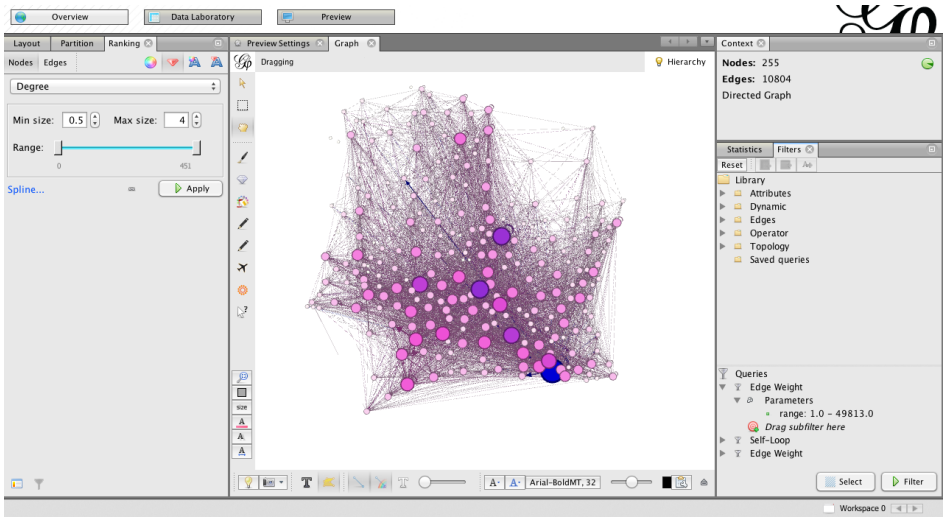

Figure B.3. Gephi preview.

APPENDIX C: United Nation Security Council Data

Below, United Nations Security Council report for Côte d'Ivoire is listed during the data collection period.

Table C.1: United Nations Security Council Reports 1st Quarter 2012.

Date	**Locations**	**Subpref.No**	**SubPref. Name**	**Antenna**	**Comment**
4.Jan.2012	Peite Guiglio	237	Guiglou	524, 521, 503	Total 16 Death
5.Jan. 2012	Dobia	150	Issia	555, 556, 753, 784, 319, 156, 447, 1202	Total 16 Death
6.Jan.2012	Toa Zeo near Duékoué	165	Duékoué	426, 884, 534, 533, 165	Total 16 Death
15.Jan.2012	Gagnoa	138	Gagnoa	141, 85, 451, 138, 452, 484, 343	Total 16 Death
7 Jan.2012	Daloa	144	Daloa	331, 645, 873, 1063, 1032, 536, 299, 997, 849, 414	4 Wounded Confrontation between farmers and cattle feeders
9 Jan.2012	Ayame	191	Aboisso	134, 66, 400, 740, 1145	Property destruction

Table C.1United Nations Security Council Reports 1st Quarter 2012. – (cont.)

Date	**Locations**	**Subpref.No**	**SubPref. Name**	**Antenna**	**Comment**
21.Jan.2012	Abidjan	60	Abidjan	124, 307, 746, 892, 321, 919, 1125, 1111, 245, 747, 738, 143, 742, 344, 307, 746, 735, 737, 741, 743	1 dead
4.Jan.2012	Béoumi near Bouaké	29	Beoumi	1119, 186	Confrontation between farmers and cattle feeders
20.Jan.2012	Touba	242	Touba	1101, 1102	Several injuries 200 displacement
22Jan.2012	Konsou, Gnabra and Zedekan	29	Beoumi	1119	Confrontation between farmers and cattle feeders
3Feb.2012	Zibabo Yablo village near Duékoué	165	Duékoué	426, 884, 534, 533, 165	3 deaths

Table C.1United Nations Security Council Reports 1st Quarter 2012. – (cont.)

Date	**Locations**	**Subpref.No**	**SubPref. Name**	**Antenna**	**Comment**
3Feb2012	Bouaké-Katiola road	253	Katiola	583, 74, 616, 964	Total 7 dead in that 3 incidents in 3-19 feb and 8 march many injuries

APPENDIX D: Merged Events

Figure D.1 and D.2 show the merged event list for Côte d'Ivoire (CIV). The first column show the line number, red color indicates, data loss in western part of the country, and leave out from analyze for that region. In order to keep weekly periodicity, before and after the duration, denoted with dark green color days are also neglected. Beside from United Nations Security reports, are shown in orange, social events such as, New Year, Football cup and religious holidays are also listed.

#	Date		Effected Area	Event Name 1	Event Name 2
1	Dec-11	5			
2		6			
3		7	CIV	Anniversary of death of Felix Houphouet Bougny	
4		8			
5		9			
6		10			
7		11	CIV	New Parliament Selection	
8		12			
9		13			
10		14			
11		15			
12		16		Yale violent inc.	
13		17		Yale violent inc.	
14		18			
15		19			
16		20			
17		21			
18		22			
19		23			
20		24			
21		25	CIV	Christmas holiday	
22		26			
23		27			
24		28			
25		29			
26		30			
27		31			
28	Jan-12	1	CIV	New Years Day	
29		2			
30		3			
31		4		Peite Guiglio	Béoumi near Bouaké
32		5		Dobia	
33		6		Toa Zeo near Duékoué	Daloa
34		7	Abidjan	Hillary Clinton Visit	Kofi Annan Visit
35		8	CIV	Baptisim of Lord Jesus fest	Kofi Annan Visit
36		9		Ayamé	
37		10			
38		11			
39		12			
40		13			
41		14	CIV	Arbeen Iman Huseyin fest	
42		15		Gagnoa	
43		16			
44		17			
45		18			
46		19			
47		20		Touba	
48		21	Abidjan	Meeting	
49		22	CIV	CUP CIV & Sudan 17-18	Konsou, Gnabra and Zedekan
50		30	CIV	CUP CIV & Angola 19-20	
51		31			

Figure D.1. Merged event list 1.

52	Feb-12	1			
53		2			
54		3		Zibabo Yablo village near Buékoué,	Bouaké-Katiola road,
55		4	CIV	Mavlid an Nabi(Sunni)	CUP CIV & Equar. Guina20-21
56		5	CIV	Day after prophet holiday	
57		6			
58		7			
59		8		CUP CIV & Mali 20-21	
60		9	CIV	Mavlid an Nabi(Shia)	Zibabo Yablo nearDuékoué,
61		10			
62		11		Arrah in eastern Côte d'Ivoire	
63		12		CUP CIV & Zambia 20:30-21:30	
64		13	CIV	African cup recovery holiday	Arrah in eastern Côte d'Ivoire
65		14			
66		15			
67		16		Tuého village near Man	
68		17			
69		18	Abidjan	Abidjan meeting	
70		19		Bouaké-Katiola road,	Westernpart of the country
71		20			
72		21		Zriglo in the border area with Liberia	
73		22	CIV	Ash Wednesday festival	
74		23			
75		24			
76		25			
77		26		Bonon and Facobly election was attacked	
78		27	Abidjan	Abidjan	
79		28			
80		29		Séguéla,	
81	Mar-12	1			
82		2			
83		3		Daloa	
84		4			
85		5		Agboville near Abidjan	
86		6			
87		7			
88		8		Bouaké-Katiola road	
89		9			
90		10			
91		11			

Figure D.2. Merged event list 2.

APPENDIX E: Levinson Age Chart

Under *"Levinson"* column the findings of Levinson is shown, under *"Expected Behavior"* column, reflection to mobile phone usage is expressed.

Table E.1: Levinson theory on life span of human being.

Gen.	**Age**	**Levinson**	**Expected Behavior**
Y	17-22	Step-out, Explore	GPS data is very active, spend more time outside the house, SMS and Call traffic is high, YouTube, chat is frequently used, Social networking, (Facebook, instagram, foursquare etc. Is used)
Y	22-28	Start a family; Pursue a dream	More structural GPS data, home-work-etc. Stay significant of time in one place during the day, evening spend time on socializing. Diverse call numbers, social networking apps, mail usage starts
Y	28-33	See flaws; re-evaluate, changes occur in life structure, either a moderate change or, more often, a severe and stressful crisis	Job locations change, call traffic is high, SMS is getting less, high tech app usage is high, social network usage high (while the phone in silent mode –predict as s/he is in a meeting)

Table E.1Levinson theory on life span of human being. – (cont.)

Gen.	**Age**	**Levinson**	**Expected Behavior**
X	33-40	Concentrate on family, community; strive to achieve dream, establish a niche in society, progress on a timetable, in both family and career accomplishments; are expected to think and behave like a parent so they are facing more demanding roles and expectations.	Spending more time at home, Call in/out from Unique contacts, less SMS, using Calendar frequently, regular daily patterns, social network usage moderate (while the phone in silent mode –predict as s/he is in a meeting)
X	40-45	Questioning; maybe a crisis	Started small drifts in regular patterns in daily life, few SMS, calendar usage
Baby Boomer	45-50	Create a new life structure, maybe w/ new job; explore recreation, grad school, etc.	Spend more time outside, divorce/affair (?), SMS usage could be high, job location change, home location conflicts/change? Spending time for recreation.
Baby Boomers	50-55	Re-evaluate; crisis possible, especially if none during the midlife transition	Moderate amount of significant locations, few SMS, few APP usage, Diverse incoming and outgoing calls
Baby Boomers	55-60	Satisfying era (similar to earlier Settle down stage) if man has adjusted to role changes.	Moderate amount of significant locations, few SMS, few APP usage, Diverse incoming and outgoing calls

Table E.1Levinson theory on life span of human being. – (cont.)

Gen.	**Age**	**Levinson**	**Expected Behavior**
Baby Boomers	60-65	Prepare for retirement and coming physical decline; major turning point	No Job location, no APP, less than 4 significant locations during the day, no SMS, Incoming calls > Outgoing calls.

REFERENCES

1. Orange Telecommunication, *Data For Development*, 2013, `http://www.d4d.orange.com/home`, [Accessed July 2014].

2. United Nations, *United Nations Security Council*, 2012, `http://www.un.org/en/peacekeeping /missions/unoci/reports.shtml`, [Accessed July 2014].

3. Farrahi, K. and D. Gatica-Perez, "A Probabilistic Approach to Mining Mobile Phone Data Sequences", *Personal and Ubiquitous Computing*, Vol. 18, No. 1, pp. 223–238, 2014.

4. Akoglu, L. and B. Dalvi, "Structure, Tie Persistence and Event Detection in Large Phone and SMS Networks", *Proceedings of the 8th Workshop on Mining and Learning with Graphs*, pp. 10–17, 2010.

5. Ihler, A., J. Hutchins and P. Smyth, "Adaptive Event Detection with Time-varying Poisson Processes", *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 207–216, 2006.

6. Paraskevopoulos, P., T. Dinh, Z. Dashdorj, T. Palpanas and L. Serafini, "Identification and Characterization of Human Behavior Patterns from Mobile Phone Data", *International Conference the Analysis of Mobile Phone Datasets (NetMob 2013), Special Session on the Data for Development (D4D) Challenge*, 2013.

7. Levinson, D. J., *The Seasons of a Man's Life*, Random House LLC, New York, 1978.

8. Kiukkonen, N., J. Blom, O. Dousse, D. Gatica-Perez and J. Laurila, "Towards Rich Mobile Phone Datasets: Lausanne Data Collection Campaign", *Proceedings of 7th International Conference on Pervasive Services (ICPS), Berlin*, 2010.

9. Farrahi, K. and D. Gatica-Perez, "What did you do today? Discovering Daily Routines from Large-Scale Mobile Data", *Proceedings of the 16th ACM International Conference on Multimedia*, pp. 849–852, New York, NY, USA, 2008.

10. Farrahi, K. and D. Gatica-Perez, "Probabilistic Mining of Socio-Geographic Routines from Mobile Phone Data", *Journal of Selected Topics in Signal Processing*, Vol. 4, No. 4, pp. 746–755, 2010.

11. Blei, D. M., A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation", *The Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

12. Patil, A. P., P. W. Gething, F. B. Piel and S. I. Hay, "Bayesian Geostatistics in Health Cartography: The Perspective of Malaria", *Trends in Parasitology*, Vol. 27, No. 6, pp. 246–253, 2011.

13. Gething, P. W., A. P. Patil and S. I. Hay, "Quantifying Aggregated Uncertainty in Plasmodium falciparum Malaria Prevalence and Populations at Risk via Efficient Space-time Geostatistical Joint Simulation", *PLoS Computational Biology*, Vol. 6, No. 4, p. e1000724, 2010.

14. Balcan, D., V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco and A. Vespignani, "Multiscale Mobility Networks and the Spatial Spreading of Infectious Diseases", *Proceedings of the National Academy of Sciences*, Vol. 106, No. 51, pp. 21484–21489, 2009.

15. Belik, V., T. Geisel and D. Brockmann, "Natural Human Mobility Patterns and Spatial Spread of Infectious Diseases", *Physical Review X*, Vol. 1, No. 1, p. 011001, 2011.

16. Frias-Martinez, E., G. Williamson and V. Frias-Martinez, "An Agent-based Model of Epidemic Spread Using Human Mobility and Social Network Information", *IEEE Third International Conference on Privacy, Security, Risk and Trust (passat) and 2011 IEEE Third International Conference on Social Computing (socialcom)*, pp.

57–64, 2011.

17. Wesolowski, A., N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow and C. O. Buckee, "Quantifying the Impact of Human Mobility on Malaria", *Science*, Vol. 338, No. 6104, pp. 267–270, 2012.

18. Pawling, A., P. Yan, J. Candia, T. Schoenharl and G. Madey, "Anomaly Detection in Streaming Sensor Data", *arXiv preprint arXiv:0810.5157*, 2008.

19. Eagle, N. and A. Pentland, "Reality Mining: Sensing Complex Social Systems", *Personal and Ubiquitous Computing*, Vol. 10, No. 4, pp. 255–268, 2006.

20. Eagle, N. and A. S. Pentland, "Eigenbehaviors: Identifying Structure in Routine", *Behavioral Ecology and Sociobiology*, Vol. 63, No. 7, pp. 1057–1066, 2009.

21. Aharony, N., W. Pan, C. Ip, I. Khayal and A. Pentland, "Social fMRI: Investigating and Shaping Social Mechanisms in the Real World", *Pervasive and Mobile Computing*, Vol. 7, No. 6, pp. 643–659, 2011.

22. AIRSAGE, *Power of Where and When*, 2014, `http://www.airsage.com`, [Accessed July 2014].

23. Altshuler, Y., N. Aharony, M. Fire, Y. Elovici and A. S. Pentland, "Incremental Learning with Accuracy Prediction of Social and Individual Properties from Mobile-Phone Data", *IEEE International Conference on Privacy, Security, Risk and Trust (PASSAT), and International Confernece on Social Computing (SocialCom)*, pp. 969–974, 2012.

24. Vieira, M. R., V. Frias-Martinez, N. Oliver and E. Frias-Martinez, "Characterizing Dense Urban Areas from Mobile Phone-Call Data: Discovery and Social Dynamics", *IEEE Second International Conference on Social Computing (SocialCom)*, pp. 241–248, 2010.

25. Aron, J., "How Innovative is Apple's New Voice Assistant, Siri?", *New Scientist*,

Vol. 212, No. 2836, p. 24, 2011.

26. Zheng, J. and L. M. Ni, "An Unsupervised Framework for Sensing Individual and Cluster Behavior Patterns from Human Mobile Data", *Proceedings of the ACM Conference on Ubiquitous Computing*, pp. 153–162, 2012.

27. Zheng, J., S. Liu and L. M. Ni, "Effective Routine Behavior Pattern Discovery from Sparse Mobile Phone Data via Collaborative Filtering", *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, Vol. 18, p. 22, 2013.

28. Montoliu, R., J. Blom and D. Gatica-Perez, "Discovering Places of Interest in Everyday Life from Smartphone Data", *Multimedia Tools and Applications*, Vol. 62, No. 1, pp. 179–207, 2013.

29. Zheng, Y., L. Zhang, X. Xie and W.-Y. Ma, "Mining Interesting Locations and Travel Sequences from GPS Trajectories", *Proceedings of the 18th International Conference on World Wide Web*, pp. 791–800, New York, NY, USA, 2009.

30. Zheng, V. W., Y. Zheng, X. Xie and Q. Yang, "Towards Mobile Intelligence: Learning from GPS History Data for Collaborative Recommendation", *Artificial Intelligence*, Vol. 184, pp. 17–37, 2012.

31. Lee, D. D. and H. S. Seung, "Learning the Parts of Objects by Non-negative Matrix Factorization", *Nature*, Vol. 401, No. 6755, pp. 788–791, 1999.

32. Breese, J. S., D. Heckerman and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", *Proceedings of the 14th conference on Uncertainty in Artificial Intelligence*, pp. 43–52, 1998.

33. Choujaa, D. and N. Dulay, "Routine Classification Through Sequence Alignment", *Proceedings of the 17th ACM International Conference on Multimedia*, pp. 737–740, 2009.

34. Durbin, R., *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK, 1998.

35. Sokal, R. R. and C. D. Michener, *A Statistical Method for Evaluating Systematic Relationships*, University of Kansas, Kansas, 1958.

36. Eddy, S. R., "Profile Hidden Markov Models", *Bioinformatics*, Vol. 14, No. 9, pp. 755–763, 1998.

37. Azam, M. A., J. Loo, A. Lasebae, S. K. A. Khan and W. Ejaz, "Behavioural Analysis of Low Entropy Mobile People Using Contextual Information", *IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pp. 590–595, 2012.

38. Wesolowski, A., C. O. Buckee, D. K. Pindolia, N. Eagle, D. L. Smith, A. J. Garcia and A. J. Tatem, "The Use of Census Migration Data to Approximate Human Movement Patterns Across Temporal Scales", *PloS One*, Vol. 8, No. 1, p. e52971, 2013.

39. Phithakkitnukoon, S., T. Horanont, G. Di Lorenzo, R. Shibasaki and C. Ratti, "Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data", *Human Behavior Understanding*, pp. 14–25, Springer, 2010.

40. Wirz, M., P. Schläpfer, M. B. Kjærgaard, D. Roggen, S. Feese and G. Tröster, "Towards an Online Detection of Pedestrian Flocks in Urban Canyons by Smoothed Spatio-Temporal Clustering of GPS Trajectories", *Proceedings of the 3rd SIGSPATIAL International Workshop on Location-Based Social Networks*, pp. 17–24, 2011.

41. Kang, J. and H.-S. Yong, "Mining Spatio-Temporal Patterns in Trajectory Data.", *JIPS*, Vol. 6, No. 4, pp. 521–536, 2010.

42. Gibson, M., *Order from Chaos: Responding to Traumatic Events*, The Policy Press, Bristol, 2006.

43. Duncombe, R. and R. Boateng, "Mobile Phones and Financial Services in Developing Countries: A review of Concepts, Methods, Issues, Evidence and Future Research Directions", *Third World Quarterly*, Vol. 30, No. 7, pp. 1237–1258, 2009.

44. Chabossou, A., C. Stork, M. Stork and P. Zahonogo, "Mobile Telephony Access and Usage in Africa", *African Journal of Information and Communication*, Vol. 1, No. 9, pp. 17–41, 2008.

45. Shapiro, J. N. and N. B. Weidmann, *Talking about Killing: Cell phones, Collective Action, and Insurgent Violence in Iraq*, Tech. rep., DTIC Document, 2011.

46. Pierskalla, J. H. and F. M. Hollenbach, "Technology and Collective Action: The Effect of Cell Phone Coverage on Political Violence in Africa", *American Political Science Review*, Vol. 107, No. 2, pp. 207–224, 2013.

47. Pitman, J., *Probability*, Springer-Verlag, New York, 1993.

48. Liu, J. S., *Monte Carlo Strategies in Scientific Computing*, springer, U.S.A., 2008.

49. Gardner, W., E. P. Mulvey and E. C. Shaw, "Regression Analyses of Counts and Rates: Poisson, Overdispersed Poisson, and Negative Binomial Models.", *Psychological Bulletin*, Vol. 118, No. 3, p. 392, 1995.

50. Grogger, J., "The Deterrent Effect of Capital Punishment: An Analysis of Daily Homicide Counts", *Journal of the American Statistical Association*, Vol. 85, No. 410, pp. 295–303, 1990.

51. Scott, S. L. and P. Smyth, "The Markov Modulated Poisson Process and Markov Poisson Cascade with Applications to Web Traffic Data", *Bayesian Statistics*, Vol. 7, pp. 671–680, 2003.

52. Yoshihara, T., S. Kasahara and Y. Takahashi, "Practical Time-Scale Fitting of Self-Similar Traffic with Markov-Modulated Poisson Process", *Telecommunication Systems*, Vol. 17, No. 1-2, pp. 185–211, 2001.

53. Venter, J. and S. Steel, "Finding Multiple Abrupt Change Points", *Computational Statistics & Data Analysis*, Vol. 22, No. 5, pp. 481–504, 1996.

54. Hawkins, D. M., "Fitting Multiple Change-point Models to Data", *Computational Statistics & Data Analysis*, Vol. 37, No. 3, pp. 323–341, 2001.

55. Chib, S., "Estimation and Comparison of Multiple Change-point Models", *Journal of Econometrics*, Vol. 86, No. 2, pp. 221–241, 1998.

56. Akoglu, L. and C. Faloutsos, "Event Detection in Time Series of Mobile Communication Graphs", *Army Science Conference*, 2010.

57. Faulkner, M., M. Olson, R. Chandy, J. Krause, K. M. Chandy and A. Krause, "The Next Big One: Detecting Earthquakes and Other Rare Events from Community-based Sensors", *10th International Conference on Information Processing in Sensor Networks (IPSN)*, pp. 13–24, 2011.

58. Cameron, A. C. and P. K. Trivedi, *Regression Analysis of Count Data*, 53, Cambridge University Press, Cambridge, UK, 2013.

59. Jewell, N. P. and A. Hubbard, *Analysis of Longitudinal Studies in Epidemiology*, Crc, Boca Raton, 2006.

60. Gamerman, D. and H. F. Lopes, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, CRC Press, Boca Raton, 2006.

61. Ross, S., *A First Course in Probability 8th Edition*, Pearson, U.S.A., 2009.

62. Osgood, D. W., "Poisson-based Regression Analysis of Aggregate Crime Rates", *Journal of Quantitative Criminology*, Vol. 16, No. 1, pp. 21–43, 2000.

63. Hofmann, T., "Probabilistic Latent Semantic Indexing", *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57, 1999.

64. Barber, D., *Bayesian Reasoning and Machine Learning*, Cambridge University Press, Cambridge, 2012.

65. Cemgil, A. T., *A Tutorial Introduction to Monte Carlo Methods, Markov Chain Monte Carlo and Particle Filtering*, 2012, `http://www2.cmpe.boun.edu.tr/courses/cmpe58N/spring2012/mc_chapter.pdf`, [Accessed July 2014].

66. Geman, S. and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6, No. 6, pp. 721–741, 1984.

67. Barber, D., A. T. Cemgil and S. Chiappa, *Bayesian Time Series Models*, Cambridge University Press, Cambridge, UK, 2011.

68. Heckerman, D., *A Tutorial on Learning with Bayesian Networks*, Springer, Amsterdam, 2008.

69. MacKay, D. J., "Introduction to Monte Carlo Methods", *Learning in Graphical Models*, pp. 175–204, Springer, Erice, Italy, 1998.

70. Andrieu, C., N. De Freitas, A. Doucet and M. I. Jordan, "An Introduction to MCMC for Machine Learning", *Machine Learning*, Vol. 50, No. 1-2, pp. 5–43, 2003.

71. Farrahi, K. and D. Gatica-Perez, "Learning and Predicting Multimodal Daily Life Patterns from Cell Phones", *Proceedings of the International Conference on Multimodal Interfaces*, pp. 277–280, 2009.

72. Jordan, M. I., Z. Ghahramani, T. S. Jaakkola and L. K. Saul, "An Introduction to Variational Methods for Graphical Models", *Machine Learning*, Vol. 37, No. 2, pp. 183–233, 1999.

73. Steyvers, M. and T. Griffiths, "Probabilistic topic models", *Handbook of Latent*

Semantic Analysis, Vol. 427, No. 7, pp. 424–440, 2007.

74. Cressie, N. and C. K. Wikle, *Statistics for Spatio-temporal Data*, John Wiley & Sons, Hoboken, 2011.

75. Farber, F., N. May, W. Lehner, P. Große, I. Müller, H. Rauhe and J. Dees, "The SAP HANA Database–An Architecture Overview", *IEEE Data Eng. Bull.*, Vol. 35, No. 1, pp. 28–33, 2012.

76. Plattner, H. and A. Zeier, *In-memory Data Management: An Inflection Point for Enterprise Applications*, Springer, Heidelberg, 2011.

77. Gonzalez, M. C., C. A. Hidalgo and A.-L. Barabasi, "Understanding Individual Human Mobility Patterns", *Nature*, Vol. 453, No. 7196, pp. 779–782, 2008.

78. Subakan, C., *Probabilistic Time Series Classification*, Master's Thesis, Bogazici University, 2013.

79. Choi, S. and R. Wette, "Maximum Likelihood Estimation of the Parameters of the Gamma Distribution and their Bias", *Technometrics*, Vol. 11, No. 4, pp. 683–690, 1969.

80. Greenwood, M. and G. U. Yule, "An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents", *Journal of the Royal Statistical Society*, pp. 255–279, 1920.