УНИВЕРЗИТЕТ У НОВОМ САДУ
ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА У НОВОМ САДУ

Катарина Д. Гаврић

Истраживање великих количина података о покретним објектима

ДОКТОРСКА ДИСЕРТАЦИЈА

Нови Сад, 2017

Mining large amounts of mobile object data

Doctoral dissertation

Katarina D. Gavrić
Faculty of Technical Sciences
University of Novi Sad

Novi Sad, 2017

УНИВЕРЗИТЕТ У НОВОМ САДУ ● **ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА**
21000 НОВИ САД, Трг Доситеја Обрадовића 6

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА

Редни број, **РБР**:		
Идентификациони број, **ИБР**:		
Тип документације, **ТД**:		Монографска публикација
Тип записа, **ТЗ**:		Текстуални штампани запис
Врста рада, **ВР**:		Докторска дисертација
Аутор, **АУ**:		Катарина Д. Гаврић
Ментор, **МН**:		Др Дубравко Ћулибрк, ванредни професор
Наслов рада, **НР**:		Истраживање великих количина података о покретним објектима
Језик публикације, **ЈП**:		Енглески
Језик извода, **ЈИ**:		Српски / Енглески
Земља публиковања, **ЗП**:		Република Србија
Уже географско подручје, **УГП**:		Аутономна Покрајина Војводина
Година, **ГО**:		2017
Издавач, **ИЗ**:		Ауторски репринт
Место и адреса, **МА**:		Факултет техничких наука, Трг Доситеја Обрадовића 6, Нови Сад
Физички опис рада, **ФО**: (поглавља/страна/ цитата/табела/слика/графика/прилога)		5 поглавља/121 страна/150 цитата/27 слике/12 табела/2 прилога
Научна област, **НО**:		Индустријско инжењерство и инжењерски менаџмент
Научна дисциплина, **НД**:		Инжењерски менаџмент
Предметна одредница/Кључне речи, **ПО**:		Истраживање података, откривање знања, људско понашање, обрасци кретања, дигитална епидемиологија
УДК		Монографска документација
Чува се, **ЧУ**:		Библиотека Факултета техничких наука, Трг Доситеја Обрадовића 6, Нови Сад
Важна напомена, **ВН**:		
Извод, **ИЗ**:		Предмет и циљ истраживања докторске дисертације представља евалуација могућности коришћења све веће количине јавно доступних података о локацији и кретању људи, како би се дошло до нових сазнања, развили нови модели понашања и кретања људи који се могу применити за решавање практичних проблема као што су: анализа атрактивних туристичких локација, откривање путања кретања људи и средстава транспорта које најчешће користе, као и откривање важних параметара на основу којих се може развити стратегија за заштиту нације од инфективних болести итд. У раду је у ту сврхе спроведена практична студија на бази заштићених (агрегираних и анонимизираних) ЦДР података и метаподатака гео-референцираног мултимедијалног садржаја. Приступ је заснован на примени техника вештачке интелигенције и истраживања података.
Датум прихватања теме, **ДП**:		11.05.2017.
Датум одбране, **ДО**:		
Чланови комисије, **КО**:	Председник:	Др Александар Купусинац, ванредни професор
	Члан:	Др Дарко Стефановић, доцент
	Члан:	Др Дејан Вукобратовић, ванредни професор
	Члан:	Др Владимир Божовић, ванредни професор
	Члан, ментор:	Др Дубравко Ћулибрк, ванредни професор

UNIVERSITY OF NOVI SAD ● **FACULTY OF TECHNICAL SCIENCES**
21000 NOVI SAD, Trg Dositeja Obradovića 6

KEY WORDS DOCUMENTATION

Accession number, **ANO**:		
Identification number, **INO**:		
Document type, **DT**:		Monograph type
Type of record, **TR**:		Printed text
Contents code, **CC**:		PhD dissertation
Author, **AU**:		Katarina D. Gavrić
Mentor, **MN**:		Dubravko Ćulibrk, PhD, associate professor
Title, **TI**:		Mining large amounts of mobile object data
Language of text, **LT**:		English
Language of abstract, **LA**:		Serbian / English
Country of publication, **CP**:		Republic of Serbia
Locality of publication, **LP**:		Autonomous Province of Vojvodina
Publication year, **PY**:		2017
Publisher, **PB**:		Author's reprint
Publication place, **PP**:		Faculty of technical sciences, Trg Dositeja Obradovića 6, Novi Sad
Physical description, **PD**: (chapters/pages/ref./tables/pictures/graphs/appendixes)		5 chapters/121 pages/150 citations/27 images/12 tables/2 appendix
Scientific field, **SF**:		Industrial engineering and engineering management
Scientific discipline, **SD**:		Engineering management
Subject/Key words, **S**/**KW**:		Data mining, knowledge discovery, human behavior, mobility patterns, digital epidemiology
UC		
Holding data, **HD**:		The Library of the Faculty of Technical Sciences, Novi Sad
Note, **N**:		
Abstract, **AB**:		Within this thesis, we examined the possibilities of using an increasing amount of publicly available metadata about locations and peoples' activities in order to gain new knowledge and develop new models of behavior and movement of people. The purpose of the research conducted for this thesis was to solve practical problems, such as: analyzing attractive tourist sites, defining the most frequent routes people are taking, defining main ways of transportation, and discovering behavioral patterns in terms of defining strategies to suppress expansion of virus infections. In this thesis, a practical study was carried out on the basis of protected (aggregated and anonymous) CDR (Caller Data Records) data and metadata of geo-referenced multimedia content.
Accepted by the Scientific Board on, **ASB**:		11.05.2017.
Defended on, **DE**:		
Defended Board, **DB**:	President:	Aleksandar Kupusinac, PhD, associate professor
	Member:	Darko Stefanović, PhD, docent
	Member:	Dejan Vukobratović, PhD, associate professor
	Member:	Vladimir Božović, PhD, associate professor
	Member, Mentor:	Dubravko Ćulibrk, PhD, associate professor

“THERE IS SOME GOOD IN THIS WORLD, AND IT’S WORTH FIGHTING FOR.”
- J. R. R. Tolkien

Dušanu, Snežani i Strahinji
jer su me inspirisali i nisu dozvolili da odustanem.

Aan Steven,
voor een geweldig avontuur.

Acknowledgements

First of all, many thanks to my advisor, Prof. Dr. Dubravko Ćulibrk, for giving me the chance to step into the world of data mining research, his invaluable support and advice over the last 5 years, and for creating an exceptionally productive and positive working environment. Despite his busy schedule he was still able to help and assist me.

Thanks to the Faculty of Technical Sciences, University of Novi Sad for providing a great research environment and making this interdisciplinary work possible.

Special thanks to my colleagues and friends Danijela Gračanin, Sanja Brdar and Milan Mirković for all the fun times, creative discussions and interesting research we conducted. Also thanks to Dejan Rašić for tolerating me as a roommate. We had a great time in ITC 206.

Thanks also to my dear friends for forming such a close group and helping each other over the years, in particular Gordana Radonić, Ivana Novaković (Jerkov), and Milenko Saravolac for long discussions on science, life and all the rest that matters.

Deepest thanks to my parents for their constant love, support, and advice over all the years. You set the course for all of this and I am deeply grateful for it. Also, thanks to my brother Strahinja for tolerating and supporting me unconditionally.

Finally, thanks to Steven for an incredible year. Our relationship has become my new foundation to plan for the future and made me feel great while closing this chapter of my life.

Katarina Gavrić,
in Novi Sad/Utrecht, July 2017

Abstract

In recent years a massive and increasing amount of textual data, photos, mobile phone data and videos is generated by users on a daily basis. Such data includes records of numerous social network services (e.g., Twitter, Facebook, Instagram), communication platforms (e.g., Blogger, Viber, WhatsApp), media sharing platforms (e.g., Youtube, Flickr, Panoramio), or joint repositories (e.g., Wikipedia).

Within this thesis we examined the possibilities of using publicly available metadata attached to these records, often containing information about locations and peoples' activities, in order to gain new knowledge, develop new models of behavior and movement of people. The purpose of the research conducted for this thesis was to solve practical problems, such as: analyzing attractive tourist sites, identifying the most frequent routes people are taking, identifying main ways of transportation, and discovering behavioral patterns relevant to the definition of strategies to suppress the expansion of viral infections. Practical studies were carried out on the basis of protected (aggregated and anonymous) Caller Data Records (CDRs) data and metadata of geo–referenced multimedia content.

While investigating the continent–level dynamics and human mobility, three types of interesting results have been generated: visual results of route similarity clustering, the main directions of movements identified (10 of them) and the results of temporal analysis. The results presented indicate that user–generated videos (YouTube) that contain geo–location can be used to identify the basic patterns in human behavior, as well as analyze the temporal dynamics of their activity. Working with metadata of over 1 million geo–referenced images (obtained from Flickr), we show that the tracks of Flickr users seem to be governed by the same laws that have previously been observed in studies based on mobile phone data and a high–resolution data set of wandering albatross flights. While there is significant heterogeneity within the population, individual users exhibit significant regularity and follow trajectories whose statistics are largely indistinguishable after rescaling with the radius of gyration of a user. These results represent the first step towards an attempt of modeling and understanding human activity patterns on a world–wide scale.

In the research related to digital epidemiology we placed mobile phone data in the context of a generalized Human Immunodeficiency Virus (HIV) epidemic. Raw mobile phone data was processed in a search for patterns that could explain the spatial variation in disease prevalence. We discovered that strong ties and hubs in communication align with HIV hot spots. Strong ties in mobility revealed to us the pathways that connect regions with higher prevalence. Our intent was not to provide a final model for HIV prediction, rather to put our effort into exploring numerous features and different models that would allow epidemiologists to gain insights, make new hypotheses and enrich their studies.

Promising results in all three directions of research pursued in this thesis open a number of possibilities for future research directed towards better understanding of human behavior.

Contents

List of Figures

List of Tables

List of Abbreviations

AIDS	Acquired Immune Deficiency Syndrome
API	Application Program Interface
APP	Application
CDC	Centers for Disease Control and Prevention
CDR	Caller Data Record
D4D	Data for Development
DHS	Demographic and Health Surveys
GDK	Geographic Knowledge Discovery
GPS	Global Positioning System
HIV	Human Immunodeficiency Virus
JSON	JavaScript Object Notation
KDD	Knowledge Discovery in Databases
KML	Keyhole Markup Language
LOO	Leave-One-Out
MMS	Multimedia Messaging Service
NSID	Network Services Identification
OPTICS	Ordering Points to Identify the Clustering Structure
PDF	Probability Density Function
POI	Point of Interest
RFE	Recursive Feature Elimination
RMSE	Root Mean Square Error
SMS	Short Message Service
SVM	Support Vector Machine
SVR	Support Vector Regression
UGC	User Generated Content
UNAIDS	Joint United Nations Programme on HIV/AIDS
UN	United Nations
USA	United States of America
VGI	Volunteered Geographic Information
XML	Extensible Markup Language

Chapter 1

Introduction

1.1 Motivation

In recent years a massive and constant growth of the amount of textual data, photos, mobile phone data and videos is generated by users on a daily basis. Such data includes records of numerous social network services (e.g., Twitter, Facebook, Instagram), communication platforms (e.g., Blogger, Viber, WhatsApp), media sharing platforms (e.g., Youtube, Flickr, Panoramio), or joint repositories (e.g., Wikipedia). In addition to user–generated records, user actions, such as calls, short messages, and multimedia messages provided by mobile phone providers, are available while taking care of user anonymization.

While a certain amount of the data is publicly available and may be crawled from the Web using dedicated Application Program Interfaces (APIs), most of it is up to the service providers to decide on providing.

This data–rich environment provides new possibilities to discover and exploit knowledge about the real world. For example, records often contain geographic information (latitude and longitude) at a certain level of granularity and precision that describes where and when users were when they generated them. The past decade has witnessed a tremendous amount of scientific effort invested in the fields of information retrieval, computer vision, text mining and geographic information science, with the aim of utilizing the data to create innovative applications and to identify novel research questions. Such applications include the extraction of places and events [98], the detection of land cover [78], predicting the spread of epidemics [53], or recommending Point–of–Interest (POI) and trajectories [148], among others.

1.1.1 Applications and challenges

There are different fields of research dealing with the ubiquitously available user–generated content and creating a heterogeneous set of application–driven techniques. Some problems addressed in the different research fields are:

- *Social Sciences*: Identifying and describing city cores [65] or extracting event and place semantics [98] using Flickr tags.

- *Disaster Management and Health Control*: Identifying crime outbreaks using mobile phone data [20], tracking hurricanes [104] using Twitter data, predicting the spread of epidemics using Twitter [53, 92] and mobile phone data [22].
- *Information Retrieval*: Discovering places and events while improving data organization of image collections [98, 1], and extraction of geographical topics from tags and blogs [109].
- *Systems for Recommendation*: Recommending points of interest and routes using user positions gained from image series and travel blogs [140, 148, 147].
- *Market Research*: Forecasting opinions and predicting polls based on online image collections and search queries [46, 96].

These works are sharing a number of common challenges apart from distinct techniques, methods and models they are using to address the problems:

- Extracting significant geographic information from unstructured, sparse and complex data records.
- Handling a high level of uncertainty of selected attributes.
- Dealing with geographic heterogeneity, such as a biased and strong background population and areas with a small number of records.
- Handling a sparse distribution of received geographic information, even when large number of records is available for processing.
- Poor–quality data such as "dirty data", missing values and poor representation in data sampling.
- Privacy issues when using data coming from domains such as e–mail, instant messaging, or phone communication.

1.2 Thesis contributions and structure

The main contribution of the thesis is to generate new knowledge about people's movements, habits and behavior using publicly available data (Spatio–Temporal User–Generated Content (UGC)), or more specifically Volunteered Geographic Information (VGI); Mobile–phone generated data CDRs on their location and movement, developing new models that better describe the phenomena that could be applied to solve practical problems, such as detecting attractive locations, the frequent trajectories people are using in the observed area, their travel habits, and the transmission and spread of a virus infection in a particular geographic area.

In addition the contribution of the dissertation is also enabling a sufficiently precise prediction of patterns of behavior, habits and people's movements. By researching and performing experiments one can get an answer to the question of

which research methodologies give the best results on the geo–referenced records that people generate, and to what extent they are suitable for a specific task. Using different sets of data for different predictions led to the identification of patterns in people's behavior which would further indicate the possibility of improving the quality of their lives and improving the environment. By monitoring the influence of different factors on the movements and patterns of behavior, for example, connectivity patterns, one might notice the need to emphasize individuals and ignore other factors.

The rest of this thesis is structured as follows:

- **Chapter 2.** This chapter provides a brief review on related work: mining of mobile–phone generated data and community–contributed data, human dynamics, activity, and mobility, and mining of community–contributed data in the field of digital epidemiology.
- **Chapter 3.** This chapter presents the methodology used while conducting the research for the purpose of this thesis. We give an overview of different data sets that were used during the research work for this thesis. It includes: continent–based dynamics, estimates on HIV prevalence rate, strong ties identification, frequent trajectories, regional connectivity and graph representation, feature selection, regression models ridge regression, support vector regression and recursive feature elimination, and statistical modeling (Lévy flight and power–law distribution).
- **Chapter 4.** This chapter encompasses experiments conducted on various different scenarios using different data sets. This chapter also presents obtained results.
- **Chapter 5.** Thesis conclusion is provided in this chapter.

Chapter 2

Background and Related Work

2.1 Overview and objectives

This thesis aims to contribute to the research focused on extracting and detecting significant patterns of human mobility and dynamics, traveling habits, and digital epidemiology, using spatio–temporal UGC, or more specifically VGI, and mobile–phone generated data (CDRs). In this chapter we introduce general concepts, challenges, and applications currently in use. We also present different scenarios and attempts to investigate these concepts. We conclude this chapter with a summary and a discussion of the main issues still open in the field of knowledge extraction from mobile object data.

2.2 Research background

This dissertation focuses on discovering significant knowledge from user–generated movement–related data. The domain dealing with advanced data analysis techniques and methodology using extensive amounts of data in computer science is traditionally referred to as Data Mining and Knowledge Discovery in Databases (KDD) and is often embedded within the term Data Science [58, 121, 77]. The subdomain coping with spatio–temporal data and knowledge discovery is called Geographic Knowledge Discovery (GDK). The use of sensors that capture the information of a user's location over time is rapidly expanding and various types of sensors are becoming ubiquitous. Location sensors are integrated into a huge number of personal devices including smartphones, watches and other location–aware devices. The most popular systems used to acquire spatio–temporal data are the Global Positioning System (GPS), radiolocation, proximity sensors and others.

In the coming sections we give an overview of commonly used concepts and terms in the domain of data mining and its subdomains. For more details about mentioned terms see [58, 121, 77].

2.2.1 Knowledge discovery and data mining

Data mining refers to extracting or "mining" from large amounts of data i.e. the process of analyzing data from different perspectives and summarizing it into valuable information. As Han states [58], data mining is a misnomer since knowledge is to be mined, not the data. Still this term is commonly accepted and used.

Definitions of data mining include:

- "Data mining is the process of extracting useful models from data." [77]
- "Data mining is the process of finding interesting patterns from a huge amount of data." [58]
- "Data mining is the process of discovering patterns and potentially significant information embedded in massive datasets." [112]

The domain of data mining has a huge overlap with traditional statistics and machine learning. Compared to traditional data analysis, the unique aspects of data mining can be identified as (1) having complex input data, (2) not a strict hypothesis to test, and (3) using massive amounts of data [121]. In data mining, a *pattern* is used to describe a particular type of information. The term pattern refers to: distributions, sequences, items, text summaries, distributions, labels or parameters describing a model. The aim of data mining is to find interesting patterns among a possibly huge number of candidates. Patterns are the output of data mining and represent the mined knowledge. The input to data mining is any kind of data collection, such as relational databases, data warehouses, transactional data, textual data, multimedia data, data streams, and sensor measurements. Transforming a data collection into a valuable format such that data analysis routines can be used, is an essential part of KDD. Discovery is the process of finding interesting patterns (and hence new knowledge) without searching for them explicitly. Data mining algorithms should allow to explore data and discover new knowledge in an automated fashion with only a small amount of prior assumptions. In this sense data mining is similar to exploratory data analysis in statistics.

Han [58] describes common data mining tasks:

- *Characterization and Discrimination* relates to the summarization and description of data having a certain characterization and the comparison of different classes of data.
- *Pattern and Association Rule Mining* focuses on the determination of patterns that frequently occur in the data. This allows the discovery of association rules between attributes.
- *Classification and Regression* are the processes of finding a model that describes and distinguishes data classes or value distributions.

- *Cluster Analysis* is the process of generating class labels from a group of data and can be used to derive a taxonomy from the records. In terms of usage, it is similar to finding peaks in a probability density distribution.

Data processing in knowledge discovery

Discovering knowledge from data can be seen as an iterative process. Fig. 2.1 provides a block diagram of the data processing activities in knowledge discovery.

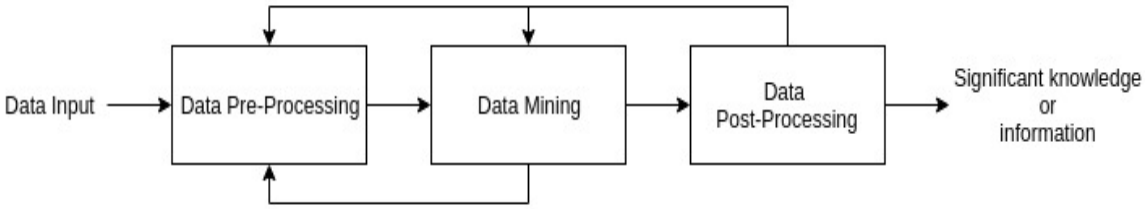

Figure 2.1: Data processing in knowledge discovery

The process includes several steps:

- *Data Pre–processing*, which includes data cleaning, integration, selection, transformation, grouping and feature (attribute) extraction from unstructured data.
- *Data mining*, which focuses on the extraction of patterns from pre–processed data.
- *Data Post–processing*, which deals with the presentation of patterns using different measures, visualizations and statistical methods.

2.2.2 Data mining of geographical data

Data mining of geographical data is a subdomain of data mining and focuses on the dimensional and spatial relationships, significant patterns, or geographical knowledge extraction. In this thesis we define time as a basic dimension of geographic space. Geographic data mining has a number of unique aspects [86] relevant to this work.

Spatial heterogeneity refers to the unequal distribution of an event or a relationship across a region (e.g., population distribution varies from region to region). *Temporal heterogeneity* refers to the variation of availability along the time. In both cases, heterogeneity can be observed on a small, medium or large scale.

Spatial dependency refers to the property of objects to be similar if they are in close spatial proximity. This indicates that closer objects are more similar to each other than distant ones. *Spatio-temporal dependency* allows us to extract significant knowledge from noisy data.

Scale is the ratio of a distance on the map to the corresponding distance on the ground.

Space and time are highly interrelated and provide a *measurement framework* for the other dimensions [86]. Therefore patterns not only occur on an attribute level,

but also in space and time. Spatio–temporal patterns are important to distinguish different type of geographic information (e.g., events, places, trends). In this work we will deal with spatio–temporal patterns of a large number of attributes (later called *features*).

Spatio–temporal data can be provided using different representations (e.g., raster, distribution, points, relationships between objects). All those types of representations can be found in user–generated data.
The geographic data mining tasks expand on the general data mining tasks and include *co–location pattern mining* and *spatial association rule mining*.

Co–location pattern mining finds subsets of spatial features frequently located at the same position [111]. The geographic aspect lies in the concept of a co–location, defined by a distance measure and a threshold in spatial or spatio–temporal space.

Spatial association rule mining finds rules between items with spatial predicates (nearby, north, contains, etc.). The input of spatial association rule mining are records with spatial and temporal features.

2.3 Community–contributed content

The growth of social platforms in last few years has enabled users to generate their own content, publish it, and interact with each other. Such services can be divided into several subcategories including, but not limited to, social networking (Facebook and LinkedIn), microblogging services (Twitter and Foursquare), photo sharing (Flickr and Instagram), video sharing (YouTube), news aggregation (Google reader and Feedburner) and social search (Google.com, Ask.com). All these platforms break the barriers between real and virtual world, facilitating new opportunities to understand individuals and extract actionable patterns from this content.

Consequently, mining of UGC has recently received considerable attention across various research areas, including data mining, spatio–temporal data analysis, machine learning, and statistics. Knowledge and meaningful patterns extracted from social media can be utilized in many application domains such as event detection [19, 104, 125], social opinion analysis [67, 118], and recommender systems [21, 87].

Such data sources (UGC) are often unstructured, including images, videos and textual messages. Commonly they contain complex relationships between different types of information e.g., locations, links, check–ins, text and others.

To indicate how many data is generated every second, here are some examples:

- According to YouTube, more than 1 billion unique users visit the site each month and consume over 6 billion hours of video–almost an hour for every person on Earth. Every second, approximately 2,314 hours worth of video is consumed. Additionally, 100 hours of new video is uploaded to YouTube each minute. [source: YouTube].

- In the 10 years since its inception, Flickr has grown its photo sharing platform to boast over 92 million users and 2 million groups. Approximately a million photos are uploaded to the site every day, which was acquired by Yahoo! in 2005. [source: TechCrunch].

- When Yahoo! acquired Tumblr in May 2014, CEO Marissa Mayer stated that the site had more than 300 million monthly visitors, 120,000 new users each day, and an incredible 900 posts per second. Tumblr currently hosts more than 105 million different blogs, but it is a constant struggle to gain more followers. This thinking is the same that professional stock shooters do–metadata is access blog content through the mobile applications. [source: Marissa Mayer, Yahoo! CEO].

- Instagram, which recently reached the 300 million monthly active user mark, continues to grow at breakneck speed. Presently, the APP sees 2.5 billion likes daily–more than 28,000 every second–and 70 million new photo uploads per day. [source: Instagram]

User–generated data sources include text documents, mobile phone communication, videos, and images. These content types contain features and spatio–temporal information (locations and time stamps) in a non–structured form (hidden in language phrases or image pixels). An important pre–processing step is to extract basic information items from the records. Before explaining pre–processing steps, we should take a look into metadata generated from Flickr and YouTube.

People use Flickr service to share and organize photos, an option also allows them to add a geographical reference. Each time a photo is virtually linked to a physical location, the Flickr system assigns a longitude and latitude and retrieves the time of capture from the Exchangeable Image File Format (EXIF) metadata embedded in the photo. The location provided by the user indicates where the photo was taken, but sometimes it denotes the photographed object. Flickr has an open API, which means that anyone can write their own program to present public Flickr data (like photos, video, tags, profiles or groups) in new and different ways. The *App Garden* is a place where developers can showcase the applications they've created and where users can find new ways to explore Flickr. The Flickr API is how users can access that data and almost all the functionality that runs Flickr is available through the API. This API is completely free to use, as a service to Flickr members as well as developers and other integrators, so they can create even more ways to interact with photos beyond Flckr. Each photo that is uploaded to Flickr can have zero or more tags. Each tag contains numerous fields, but it can also be very simple (see Fig. 2.2).

```
<tag id="1234" author="12037949754@N01" raw="woo yay">wooyay</tag>
```

Figure 2.2: Example of simple Flickr tag.

ID represents a unique identifier for the photo, *author* is the Network Services Identification (NSID) of the user who uploaded the photo, *raw* represents the version of the tag entered by the user (not modified), *tag–body* represents the "clean" version of the tag–as processed by Flickr.

All the visual content hosted on Flickr is user–contributed, and tagged by users themselves. Tagging rights are restricted to self-tagging or even permission–based tagging. All tags are freely chosen from an uncontrolled vocabulary, and thus might contain typos, inverted words etc. Each photo can have a maximum of 75 tags and this group of tags are a photo's *tagset*. The motivation that stimulates the tagging process in Flickr, as well as in other platforms, has been classified in different ways. The most popular being a macro–distinction between *categories* (users who employ shared high–level features for later browsing) and *describers* (users who accurately and precisely describe resources for later searching) [70]. One such classification is performed by Beaudoin [10]. There are 18 *post hoc* created tag categories, which include property types (e.g., adjectives, verbs), semantic classes (people, living or non–living things), places (e.g., museum, square), events/activities (e.g., wedding, Christmas), emotions etc. From all 18 types of tags identified, Beaudoin reports that the most frequent are: (1) geographical locations, (2) compounds, (3) inanimate things, (4) participants, and (5) events.

To extract a significant amount of Flickr photos, one of the publicly available search engine is *flickr.photos.search*[1]. This search engine returns a list of photos matching some criteria. Only photos visible to the calling user will be returned. To return private or semi–private photos the caller must be authenticated with "read" permissions to view the photos. Unauthenticated calls will only return public photos. One can forward various different arguments to the API based on the criteria of interests. Below is the list of possible arguments a user can forward to the service:

- *user_ID (Optional)*: The NSID of the user who's photos to search. If this parameter isn't passed then everybody's public photos will be searched.
- *tags (Optional)*: A comma–delimited list of tags.
- *text (Optional)*: A free text search. Photos who's title, description or tags contain the text will be returned.
- *accuracy (Optional)*: Recorded accuracy level of the location information. Accuracy range is from 1 to 16 where: 1–World level, 3–Country level, 11–City level, 16–Street level. If not specified, default accuracy is 16.
- *lat (Optional)*: A valid latitude in decimal format.
- *lon (Optional)*: A valid longitude in decimal format.

[1] https://www.flickr.com/services/api/flickr.photos.search.html

Another interesting search engine is *flickr.places.findByLatLon*[2]. This engine returns a place ID for a latitude, longitude and accuracy triple. It is designed to allow users to find photos for "places" and will round up to the nearest place type to which corresponding place IDs apply. For example, if you pass it a street level coordinate it will return the city that contains the point rather than the street, or building, itself. It will also truncate latitudes and longitudes to three decimal points. To use this engine the user should forward the following arguments: api_key (required), lat (required), lon (required), accuracy (optional). Fig. 2.3 shows an example of the response generated using*flickr.places.findByLatLon* in JavaScript Object Notation (JSON) format.


```
{ "photo":
      { "id": "34371182944",
        "secret": "b58cbf9598",
        "server": "4249",
        "farm": 5,
        "camera": "Canon EOS 760D",
     "exif": [
       { "tagspace": "IFD0",
         "tagspaceid": 0,
         "tag": "Make",
         "label": "Make",
         "raw": { "_content": "Canon" }}
      ],
     "places": [
       { "latitude": {"37.76513627957266"},
         "longitude": {"-122.42020770907402"},
         "accuracy|: {"16"},
         "total": {"1"},
         "place_id": {"Y12JWsKbApmnSQpbQg"},
         "woeid: {"23512048"},
         "latitude": {"37.765"},
         "longitude": {"-122.424"},
         "place_url":
{"/United+States/California/San+Francisco/Mission+Dolores"},
         "place_type": {"neighbourhood"},
         "place_type_id": {"22"},
         "timezone": {"America/Los_Angeles"},
         "name": {"Mission Dolores, San Francisco, CA, US, United
States"}
      ]
}
```


Figure 2.3: Complex response from Flickr in JSON format.

[2]https://www.flickr.com/services/api/flickr.places.findByLatLon.html

Besides Flickr, YouTube also has an open YouTube Data API allowing users to search the database based on resource type. A resource is an individual data entity with a unique identifier. There are different types of resources that user can interact with using the API, such as a video (represents a single YouTube video), a playlist (represents a single YouTube playlist, a collection of videos that can be viewed sequentially and shared with other users), a subscription (contains information about a YouTube user subscription) etc. To query lists of resources, the user should provide parameters for any API request that retrieves or returns a resource. For example, a video resource has the following parts: snippet, contentDetails, fileDetails, player, processingDetails, recordingDetails, statistics, status, suggestions, topicDetails. Fig. 2.4 gives an example of a response received using YouTube Data API in JSON format. Supported methods for each resource are LIST, INSERT, UPDATE, and DELETE.

```
{"kind": "youtube#video",
        "etag": etag,
        "id": string,
        "publishedAt": datetime,
        "title": string,
        "description": string,
        "channelTitle": string,
        "tags": [
                "categoryId": string,
                "liveBroadcastContent": string,
                "defaultLanguage": string,
                "localized": {
                "title": string,
                "description": string
                ],
        "recordingDetails": {
                "locationDescription": string,
                "location": {
                        "latitude": double,
                        "longitude": double,
                        "altitude": double
                },
        "recordingDate": datetime
                },
        "videoStreams": [{
                "widthPixels": unsigned integer,
                "heightPixels": unsigned integer,
                "frameRateFps": double,
                "aspectRatio": double,
                "codec": string
}]
```

Figure 2.4: Complex response from YouTube in JSON format.

UGC data sources often include text documents, images, and videos. These content types contain attributes and spatio–temporal information in a non–structured form, e.g., implicitly covered in language or image pixels. Often the term *unstructured* data is used to refer to records whose information is not represented by an explicit data schema or attributes [137]. The term unstructured data is slightly misleading, since the records of course have an implicit structure to convey the information [43]. An important data pre–processing step hence is to extract basic information items from the records.

Data pre–processing is a data mining technique that involves transformation of raw data into an understandable format. An important task in pre–processing unstructured data is to leverage different types of elements in the records in order to transform it from being irregular and unstructured into an explicitly structured representation [43]. Data pre–processing is an often ignored but very important step in the data mining process. Real–world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data pre–processing is a proven method of resolving such issues and is used to prepare raw data for further processing. Data pre–processing includes cleaning, integration, transformation, reduction, etc. These pre–processing operations are responsible for transforming unstructured, original–format data stored in data collections into a more explicitly structured intermediate format. Two clear ways of categorizing the totality of preparatory data structuring techniques are *according to their task* and *according to the algorithms and formal frameworks* that they use.

Data pre–processing methods are divided into following categories [40]:

- *Data Cleaning*: is applied to remove noise and correct inconsistencies in the data.
- *Data Integration*: merges data from multiple sources into a coherent data source, such as a data warehouse.
- *Data Transformation*: convert the data into appropriate forms for mining. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements.
- *Data Reduction*: reduces the data size by aggregating, eliminating redundant features, clustering etc.

The output of the process is usually a representation where the observations are described predefined features or attributes. In computer vision, features usually describe edges, textures, or shapes [36]. In text mining, features are used to describe basic information such as terms, words, or characters [43]. Features are used to represent parts of a record or an entire record, creating a feature–based (structured) representation of a record. Feature representations are then used for subsequent tasks in text mining and computer vision, such as extraction of high–level features (concepts and entities) or objects in images and videos.

2.4 Mining mobile–phone generated data

In a burst of new applications based on mobile phone data [16], we emphasize those with great practical importance such as disaster management [13], urban planning [11], deriving poverty indicators [117], transportation mode interface [130], traffic engineering [28] and even crime prediction [20].

Mobile phone communications engender the era of big data by creating huge amounts of CDRs. Cell phone service providers collect these records whenever a phone is used for text messages or calls. The records contain the time of action, identifiers of sender and receiver, and the cell tower used in communication. In this way mobile phones uncover approximate spatio–temporal localization of users and provide an immense resource for the analysis of human mobility and behavioral patterns [11, 29, 138].

Early studies [55, 131] showed that the basic laws governing human mobility can be derived using the publicly available data for a specific region. Both of these studies have similar goals of combining geo–spatial information with mathematical models in order to extract significant patterns of human motion. In [55] the authors address the challenging problem of mathematically modeling human mobility. Their study is based on two mobile phone datasets. The first was collected by tracking 100 000 anonymized mobile phone users, selected out of a sample of over 6 million users. Their position was recorded any time they initiated a call or sent an SMS over a six–month period. The second dataset captured the location of 206 users whose position was recorded every two hours, for an entire week.

Analyzing user displacements between consecutive positions, they mathematically show that their distribution is well approximated by a truncated power law [55]. The authors continue the analysis to show that this type of distribution captures a convolution of individual Lévy flight trajectories [102] and population based heterogeneity. Defining the radius of gyration (r_g) of a single user to be the typical distance travelled by the user up to time t, they show that the rescaling of the distribution of displacements with this value causes it to collapse into a single distribution, suggesting that a single relative jump size distribution characterizes all users, independent of their r_g. Finally, ranking of the locations visited by the users reveals that people devote most of their time to a few locations, while spending their remaining time in 5 to 50 places.

In [131] the authors used the similar mobile phone data to study fundamental spreading patterns that characterize a mobile virus (Bluetooth and Multimedia Messaging Service (MMS)) outbreak.

Wang *et al.* used a dataset of over 10 000 mobile phone towers trying to understand the spreading patterns of mobile phone viruses. In their study the dataset was collected by a mobile phone carrier for billing and operational purposes. It contains

the date, time and coordinates of the phone tower, and a record of each phone call and text message sent or received by approximately 6.2 million costumers. The dataset summarizes one month of activity. To ensure anonymity each user is identified by the data source with a hash key and any potential personal identifiers were removed from the dataset. Wang *et al.* used the Voronoi diagram around each mobile phone tower to obtain a reasonable estimate of the towers service area. The Voronoi diagram areas are acquired by Delaunay triangulation [45], finding that the distribution of the tower areas follow a power–law with two different scaling regions. To identify the hourly location the authors discretized time by splitting each day into 24 hour time intervals.

The methodology they used to identify the users locations with an hourly frequency used the following procedure: determine each user's travel characteristics and reconstruct the position of the user at times when there are no calls such that the fundamental characteristics of each user are preserved. To validate the proposed methodology they compared different statistics for the empirical and simulated data. With this approach they were able to show that, contrary to previous research, human trajectories show a high degree of temporal and spatial regularity and that each individual is characterized by a time independent characteristic length scale and a significant probability to return to a few highly frequented locations.

In addition to understanding the spread of human viruses, there is a growing interest in the mining of mobile phone data for true epidemiological purposes [79, 124]. The practice can advance the research in epidemiology by shedding light on relationships between disease distribution, spread and incidence on one side and migrations, everyday movements and connectivity of people on the other side. Up to now only a few studies have used mobile phone data to quantify those relationships on real disease distribution data. Wesolowski *et al.* explored the impact of the human mobility to the spread of malaria [138]. They analyzed CDR data collected by a mobile phone service provider in Kenya over a one–year period and discovered how mobility patterns contribute to the spread of the disease beyond what could be possible just by insects. The other study carried out by Martinez *et al.* investigated the effect of government alerts during an H1N1 flu outbreak in Mexico on the diameter of mobility of individuals. Bengtsson *et al.* [13] estimated population movements from a cholera outbreak area and suggested using obtained information for disease surveillance and resolving priority in relief assistance. These pioneering works announce the emerging field of digital epidemiology [17].

To the best of our knowledge, the research conducted for the purposes of this thesis is the first attempt to use mobile phone data for exploring the complex structure of HIV epidemics [22]. HIV has a devastating social, demographic, and economic effect on Africa [27, 39]. With 3.7% of the population infected [132], Ivory Coast has the highest prevalence rate in West Africa and a generalized epidemic [68, 134]. This epidemic, where the disease spreads out of risk groups and affects the general population, demands the development of national HIV prevention plans. Although the prevalence rate appears to have remained relatively stable over the past decade,

and is even decreasing, due to the prevention of mother–to–child transmission, there is still much work to be done in improving the health system for a more effective response to HIV. Deeper understanding of the epidemics can help in finding ways to suppress HIV even more. Modern technologies that deal with human mobility phenomena may help respond to that challenge. A lot of scientific effort is aimed at identifying the driving factors of HIV spread. Most frequently mentioned are poverty, social instability, violence, rapid urbanization, high mobility and modernization. The difference among these factors could explain spatial disparity in prevalence rates. In the study of Messina *et al.*, geographic patterns of HIV prevalence in the Democratic Republic of Congo were examined [85]. They showed that spatial factors: prevalence level in the range of 25 km and distance to the urban areas are strongly connected to the risk of HIV infection. The impact of migration on the spread of HIV in South Africa was studied in [33] where authors developed a mathematical model to compare the effects of migration and associated risk behavior. In the early stage of epidemics, migration impacts HIV progression by linking geographical areas of low and high risk, while in the later stage by increasing high–risk sexual behavior. However, the migration was quantified through surveys where participants were questioned about movement history and the study included only two migration destinations.

In [22] we conducted a comprehensive analysis of two data sets offered within the Data for Development (D4D) challenge [17]. Our research was guided by the following hypothesis: the risks for spreading HIV infection are associated with spatial and behavioral factors that could be detected from the available collection of data. We were particularly interested in tracking population movements and inferring the communication strength between sub–prefectures of Ivory Coast with different prevalence rates.

2.5 Mining community–contributed data

In recent years, geo–referenced community–contributed multimedia data and associated metadata, available from services such as Flickr/YouTube, have been used to help understand patterns of human mobility, behavior and habits. Such information is of importance for traffic forecasting [28], urban and trip planning [11], disease outbreak management [13], transportation mode inference [130], etc. While the data is freely available for much larger regions of the world, it is understood that the quality of such data is lower than that of data that can be obtained from mobile phone operators. This is due to the fact that there is no control over the accuracy and the sampling frequency of user–generated content. Nevertheless there is a growing amount of research work aimed at identifying human dynamics and social interaction.

Tourists face a great challenge when they search for information about places they want to visit, as there are many web sites providing different information about popular locations. Fusing all this information is a time consuming activity. As a result there have been some attempts to develop automatic tools to analyze publicly available data of tourists' behavior and use it to suggest truly attractive routes and

places. The information gathered in this way is potentially useful not only to the tourist industry but urban planners and authorities.

In [54] *et al.*, authors attempt to identify attractive locations based on data gathered from Flickr. They used a corpus of geographically referenced photos taken in the province of Florence in Italy by 4280 photographers over a period of two years. Based on the disclosure of the location of the photos, they design geo–visualisations to reveal the tourist concentration and spatio–temporal flows. In the pre–processing task, they separated visitors from the residents of the region. To accomplish that they used the presence in the area over time as the discriminating factor. They divided the time in frames of 30 days and computed the number of periods each user was active in that area. If a photographer took all her/his photos within 30 days the algorithm considers him/her as a visitor, while if the interval is greater then 30 days, the algorithm considers him/her as a resident. After this pre-processing a population of 4280 users was reduced to 3505 one-time visitors. Their initial results bring to light two aspects related to spatio-temporal density and movements of visitors: (1) Characterizing the areas of the city/region where the tourists are concentrated, and (2) Revealing spatio-temporal signatures: activity by day of the week and month of the year, and days of the year. Authors use Google Earth[3] for interactive visual synthesis of encodings generated using the combination of MySQL[4] for data storage and querying (selection and aggregation), and the software "Urban Dynamics" they developed in Java to access, process, transform, aggregate and cluster the raw data stored in the database. The Keyhole Markup Language[5] (KLM) is used to describe visual encodings and define interactions. Their approach enables the quantification of movements between tourist attractions, but also allows detection of distinct patterns of mobility among groups of tourists of different nationalities.

Similarly, Kisilevich *et al.* [69] used geo–tagged photo information collected from Flickr and Panoramio and tried to map locations attractive for tourists on vacation. Their study presents a framework in which they demonstrate a systematic approach for visualization and exploration of attractive places. They used density–based clustering and an influence weight, which allows for the estimation of the characteristics of the attractive places by retrieving photos with high influence weight. The influence weight (for a photo point p) is calculated as the sum of kernel functions between point p and all other points in a cluster whose owners are different then the owner of point p. Results were visualized using Google Earth Mashup[3].

Adrienko *et al.* [3] proposed an approach that allows interactive cluster analysis of large collection of structurally complex objects. Based on proposed methodology, the same authors continued to develop a tool (Visual Analytics Toolkit) to cope with the complexity of analyzing a large dataset of moving objects in a step wise manner

[3] www.google.com/earth/
[4] http://www.mysql.com/
[5] http://earth.google.com/klm

[4]. In a further study [3] authors tried to analyze tourists' movements within the city of Berlin based on the images collected from Panoramio. Their approach was based on building flow maps of aggregated moves between places. They considered sequences of photos taken in Berlin and divided them into sessions. Each session was treated as trajectory and trajectories clustered to form a flow map.

In this thesis we focus on detecting attractive locations and tourist dynamics using metadata from publicly available geo–referenced images collected from Flickr, tagged as recorded in Berlin, Germany [49]. The goal was to identify attractive locations in the city, assuming that an attractive place is characterized by large amounts of photos taken by many people. In addition, we attempt to identify standard routes people take when making a tour. The presented techniques use density–based clustering and route similarity and dynamics clustering. Results indicate that information obtained by analyzing Flickr photos can be used to reliably detect locations attractive to tourists' and suggest best tour routes.

2.6 Human dynamics and mobility

The introduction of location–based services gave a possibility to users to voluntarily share their activities through online social platforms, providing an exceptional amount of user–generated data on human movement and activity participation. This data contains detailed geographical information, which reflects wide–ranging knowledge about human movement behavior. Thus location–based data offers researchers a new dimension of information related to human activity categories in much more detail.

In the past few years the social science literature has shown significant attention to extracting information from community–contributed data to track and analyze human movements. Studies that use such data tend to focus on identifying human mobility patterns from which predictions can be made. Scientists are achieving extreme accomplishments by mining the user–generated location–based data which has already given great results and many novel applications such as recommendation system for physical locations (or activities) [8, 145] or potential customers or friend [103, 146]; or recommendation system for popular travel routes in a city [136].

These data have a possible impact on many other areas including travel demand modeling, epidemiology, tourismology, ubiquitous computing, urban planning, security and health monitoring. As such, an excellent opportunity exists to develop essential tools to analyze this very large–scale spatial and temporal data that allows one to understand the social and behavioral characteristics of the users of location–based services. Previous research efforts on individual travel activities and patterns over longer period of time were usually based on people's movements through traditional surveys on travel journeys [101, 82, 127].

On the other hand, there are recent studies of human mobility, activity and dynamics that have used distance–based measures to extract significant patterns using alternative datasets collected from mobile phones [55], bank note movements [24], user–generated content [61] and subway smart–card transactions [60] etc. These studies however limit the understanding of the interaction between a selection of destinations for different purposes and mobility dynamics due to the lack of information about the purposes behind these movements. In this context, user–generated location–based data has received increasing attention in the research community, as the rich information in the data connects each geographical location with a venue category, indicating the purpose of the performed activity. In late studies, Cheng *et al.* [31] examined 22 million check–ins and observed a similar mobility pattern found in previous researches [55, 24], which is a variety of short, random movements with occasional long jumps. In [32] authors investigated the relation between human mobility and social relationship using data from Gowalla[6] and Brightkite[7]. They found that social relations can explain 10% to 30% of all human movements, while periodic behavior explains 50% to 70%.

However the dimension of human activity lack in both of the researches. In [61] by also considering the temporal dimension and activity categories (i.e. purpose or goal) into the analysis, authors discover more realistic and detailed descriptions of human mobility dynamics. Including the activity goal in the analysis enabled researchers to develop advanced models for predicting mobility decisions. In their study they consider the user–generated location–based data obtained from Twitter to characterize urban human activity and mobility patterns. They first investigate the characterization and visualization of aggregate human mobility and activity patterns by creating a grid reference of a city map into square cells of 200 x 200 meters. They discover a relation between the popularity of a cell and the probability of visiting the cell. Spatial distributions of visiting different places are also determined for various activity goals by counting the number of goal–specific visits within each cell and computing the proportion of visits to each cell for each activity category. They classify different activity categories based on the type of the visited locations (Table 2.1).

Table 2.1: Activity category classification

Activity category	**Type of visited location**
home	home(private), residential building (apartment)
work	office, design studio
eating	restaurant, pizza place, coffee shop
entertainment	pub, nightclub, bar, concert hall
recreation	park, gym, playground
shopping	supermarket, store, bookstore

[6]https://en.wikipedia.org/wiki/Gowalla
[7]https://en.wikipedia.org/wiki/Brightkite

This generates activity distribution maps showing the popular places within a city and the functionality of each part of the urban area. Check–in distributions appear to be different for different activity categories suggesting a strong influence of urban context on users' destination choices. Using Kernel density estimation methods they construct time–dependent activity density maps. Using this approach authors also visualize different human activities in a city and thus capture the pulse of urban human activities. Furthermore, they investigate the characterization of the spatio–temporal aspects of individual mobility patterns. They determine a set of statistical parameters to characterize human mobility based on check–in data from Twitter. First, they observe the timing of visiting different places depending on activity category. Second, they explore the frequency of visiting a place with respect to the rank of the place in individual visits. Finally, they used Zipf's law to measure the visitation frequency of the *Lth* most visited location.

In [95], authors take the example of the Haiti earthquake [13] to provide motivation for big data driven crisis response that helps the affected people of the earthquake. Huge volumes of data related to the crisis (including Short Message Service (SMS) from onsite victims, social media data from onsite–users, journalists, and aid organizations) was subsequently collected. Sifting through the voluminous "big crisis data" (big data collected during a crisis situation [84]) to find relevant information about the affected population is a challenging task. This challenge was tackled by the *digital humanitarians* by employing techniques such as crowd–sourcing to acquire data to produce crisis maps. The online crowd–sourcing platform used to collect data for the Haiti earthquake was Ushahidi[8]. Ushahidi is a mobile–based platform for developing "crowd maps" through collecting, visualizing, and mapping citizen–supplied (or crowd–sourced) data.

Even though most of the research is conducted in a sphere of how to use community–contributed data in natural disaster management, this field also includes: migration/refugee crises, epidemic crises, natural disasters, crowd control problems, terrorist attacks, civil wars, public violence and disorder, industrial accidents, infrastructural failures etc.

Overall, recent literature reflects the success in usaging of community–contributed data in human dynamics and mobility; it is clear that smartphones, as they gain greater traction and popularity, will be the bridge that enables those data to be used to great extent.

2.7 Mining community–contributed data in digital epidemiology

Traditional disease surveillance has been a key element in any public health portfolio for many years. Disease surveillance is widely recognized as one of the most important tools to collect, assess, predict, and diminish infectious disease outbreaks.

[8]https://www.ushahidi.com/

Traditional disease surveillance is based on data collected by health institutions, and the data typically consist of information such as: demographic data, morbidity and mortality data, laboratory reports, individual case reports, field investigations, and surveys. They are generally collected by hospitals, laboratories, and other health providers and institutions. Improvements in technology and computers have affected traditional disease surveillance systems by improving the convenience of data and by increasing the speed at which data are transmitted between institutions. However, the ongoing Internet and mobile phone revolution has a qualitatively distinct effect: in addition to making epidemiologic data available faster and more broadly, new data is generated directly by the public, often on platforms not primarily designed for health purposes (like social platforms). These streams of user–generated data are almost always bypassing traditional public health channels. They are the data streams on which digital epidemiology is generally based [105, 106].

One of the first and most prominent examples of digital disease surveillance was Google Flu Trends [53]. Google Flu Trends was essentially an analytical estimate of the level of weekly influenza activity based on the search queries that Google received. The analytical estimate was derived by a model selected by generating the best fit to the Centers for Disease Control and Prevention's (CDC). Influenza–like illness (ILI) data from a number of different regions in the United States of America were used. A mean correlation between the results of the original model and CDC data was estimated to 0.9. A few years later, in the summer of 2015, Google decided to shut down the public website of Google Flu Trends. Instead they gave data access to selected academic and public health institutions. This achievement followed numerous reports [34, 90, 76] that systematically assessed Google Flu Trends' over-estimation of influenza activity, attributing it to a combination of a term "big–data hubris" and algorithm dynamics. The first refers to the assumption that the novel big–data streams are a substitute, rather than a supplement, to traditional data collection efforts. The second refers to the observation that, while the Google search algorithm receives updates on a weekly or even daily basis, the Google Flu Trends model received updates only rarely. This led to a situation where the model did not keep in sync with the changing nature of the data from which it was supposed to generate predictions.

Despite the problems of Google Flu Trends, the system was an important example of the promises on which digital epidemiology is based: to use novel data streams, often generated for purposes quite distinct from public health, to extract additional public health signals, such as those relevant for disease surveillance. While Google makes some search pattern data available through an interface called Google Trends, the raw search–query data that Google Flu Trends was based on is not publicly available. In recent years two other digital data sources have attracted the attention of digital epidemiologists: Twitter, the popular microblogging service, and Wikipedia, the world's largest open–access encyclopedia. Twitter data are openly accessible through an application programming interface, which allows any third party to stream Twitter data in real time to their own application. Twitter has been extensively used

to assess influenza activity [25, 114]. Wikipedia access logs, a public data source, have recently attracted the attention of the research community as a proxy of search engine query logs. Early analyses of Wikipedia access logs have shown great promise in providing real–time estimates of the prevalence of a number of infectious diseases [51, 83].

Real–time, mobile, precise: the ongoing revolution in the way of how people communicate opened a new course for research in the sphere of epidemiology. Digital data sources, when exploited properly, can provide geographical information about disease and health dynamics on a world-wide scale. Using digital data sources such as chat rooms, social networks, news platforms, and blogs can provide a possibility to capture and disseminate (in almost real–time) some types of infections, outbreaks and chronic diseases. Results on a global health level obtained by using these online sources is often different [26] from the picture created by traditional surveillance systems. In fact, publicly available data streams became valuable data sources for a new generation of public health surveillance, because it complements existing traditional surveillance, fill gaps in public health infrastructure, and operate across international borders and therefore is not related to only one specific region [63, 64].

Constant movements of humans create the dynamic links that connect populations on different scales and enable detection of geographic spread and sustained transmission of diseases. Up to the last few years those types of human movements and dynamics were estimated using travel surveys, road networks, or small–scale GPS studies. Mobile phone data in the form of call data records (containing information about the location of the mobile phone tower used during a call from a mobile phone) provide one of today's most promising opportunities to study human mobility [55] and its influence on disease dynamics. Developments in the sphere of smartphones improved the capability to track human activities and migrations at high spatial and temporal resolution [30], providing a much deeper understanding of social behavior [93], and enriching previous studies based on large–scale surveys [88].

Objective measurements, that can be obtained from mobile–phone generated data, allow for a more accurate description of infectious disease dynamics. Those measurements can then improve parameterization of large–scale computer simulation disease models. The introduction of these models has enabled wider research in terms of large numbers of individuals, rather than population aggregates. Mobile–phone generated data have already been used to create realistic models of human mobility [55], predict the rate of spread of drug resistance [81], assess the prospects of malaria eradication [138], monitor population movements during the Haiti cholera outbreak in near real–time [13], and used mobile phone data in the context of a generalized HIV epidemic [22]. Models based on recorded sequences and links between humans can provide information potentially useful for the design of targeted immunization strategies [107]. Large–scale mobility data can be used to map the worldwide measuring of emerging infectious diseases such as the 2009 H1N1 epidemic [7].

While search engine logs, social media posts, and Wikipedia access logs are a few examples of big–data sets that have emerged following the ongoing Internet penetration worldwide, there is also another source of data that is increasingly relevant for disease surveillance–the public itself: web–based participatory surveillance systems. In those surveys patients are asked to report symptoms and other data directly online. Those types of surveillance systems have shown great promise in the case of influenza. In Europe the*Influenzanet* project[9] has been successfully collecting data on ILI activity in a number of European countries [91]. In the United States of America (USA) *Flu Near You*[10] has emerged as a leading crowd–sourced influenza surveillance system and there are other similar projects around the world. Given the widespread use of smartphones with broadband Internet access worldwide, we can expect many more participatory public health applications in the near future, complementing traditional surveillance systems.

Importantly, because many of the data streams of digital epidemiology have not been generated for the disease surveillance niche, much broader insights can be gained from these data sources. While a lot of the earlier work on digital epidemiology has focused on user–generated descriptions of symptoms, later work has increasingly focused on the analysis of health behaviors and sentiments/opinions, particularly as they relate to infectious diseases. For example Twitter data have been mined for signals of vaccine sentiments to estimate vaccine uptake rates. During the 2009 influenza of H1N1, vaccination sentiments measured on Twitter correlated positively with prospectively reported vaccination uptake rates which indicates that these new data streams can help in the public health decision–making process. Last but not least data from most of these services are increasingly generated on mobile phones and other devices, increasing the probability that high–resolution geographic information is associated with the data, a phenomenon that will become increasingly important, given the spatial dynamics of disease spread. Mobile phone–generated data are improving the development of computer simulations with the goal of providing better scenario analysis for the policy making process and crisis management.

[9]https://www.influenzanet.eu/

[10]https://flunearyou.org/

Chapter 3

Methodology

3.1 Data and setup

This section describes all the datasets that were used for the research done within the scope of this thesis: geo–referenced community–contributed metadata (from Flickr and YouTube), continent population distribution (for Africa), an HIV data set containing information about the Acquired Immune Deficiency Syndrome (AIDS) pandemic and HIV seroprevalence, and a D4D data set containing mobile phone data records for the Ivory Coast.

Flickr data sets

In the research conducted on geo–referenced community–contributed metadata, aimed at deriving the basic laws of human mobility [48], we used a dataset of 1 million images from Flickr relevant to the San Francisco and San Diego area. The content has been automatically downloaded using a tool developed in our lab, which in turn relies on the Flickr public API and uses the cURL library (Flickcurl). We used two datasets to explore the mobility patterns of individuals. The first one (SET1) was the entire set consisting of meta data from geo–referenced images, and the second one (SET2) was a subset of the first set, containing the data uploaded by users who contributed images over a period of time longer than a week. This was done in an attempt to eliminate the contribution of tourists from set SET2, as we assumed that users with just a few images over a short period of time fall in this category. To better compare two different approaches, ours and the one conducted by Gonzalez [55], we used the same data set as the authors in the mentioned study.

Another data set, collected for a study described in [49], was used to detect attractive locations and tourist dynamics. For each image an Extensible Markup Language (XML) metadata file was obtained which consisted of a photo ID, username, date and time at which the image was taken, geo–referenced tags (latitude and longitude) and accuracy of tag (values between 1 (lowest) and 16 (highest)). The crawler was executed in the middle of March 2011 and returned around 600.000 results. They corresponded to 387.524 unique users and the observed time frame was between March,

1962 and March, 2011. All the images were tagged as recorded in Berlin, Germany. The reason behind using this particular data set lies in the fact that Berlin is one of the most visited cities in Europe, so we assumed that that the number of available photos will be large.

YouTube data sets

For the study focused on determining major routes of movement across the African continent, start/duration/end of trip and basic means of transportation [50], we used a dataset of 113.157 unique metadata records associated to videos collected from YouTube. The records collected contain:

- ID: unique string for video identification (assigned by YouTube).
- Category: one of the 15 categories (entertainment, education, music etc.), selected by the user.
- Viewed: number of times the video was viewed, at the time the data were obtained.
- Published: time stamp when the video was published.
- Duration: duration of the video in seconds.
- Latitude: latitude of the location where the video was recorded.
- Longitude: longitude of the location where the video was recorded.
- User: username of the user who published the video.

All videos were tagged as recorded in Africa. The observed time frame was between September, 2006 and April, 2011. We decided to use this data set because: (1) Africa is a large and rapidly growing market for research, with a great potential; (2) Africa is the second biggest continent in the World with a large geographic scale.

Population data

For the studies described in [22, 47] we used a data set made available on AfriPop by Linard [80], which contains full details on population distribution, summarized on the level of countries. Linard *et al.* developed a new high–resolution population distribution data set for Africa and analyzed rural accessibility to population centers. Contemporary population data was combined with detailed satellite–derived settlement extents to map population distribution across Africa at a finer spatial resolution.

HIV data

Two data sets related to the prevalence of HIV were used. The first is provided by the United States Census Bureau, which contains the information about AIDS pandemic and HIV seroprevalence (infection) in population groups in the developing countries [134]. The second includes the results of Demographic and Health Surveys (DHS) [132] which provides data about the health status of countries. We used data collected in a survey conducted between 2008–2012. This data provides estimates for ten administrative regions of the Ivory Coast. The results of the estimation are presented in Fig. 4.11 (a).

Using the first data set we estimated the HIV prevalence rate between 2008–2010. The results are presented in Table 3.1.

Table 3.1: Estimated prevalence rate between 2008–2010

Regions	**2005**	**2008–2010**
Center-East (Moyen-Comoè)	5.8	9.17
South (Lagunes, Agnèby, Sud Comoè, Sud Bandama)	5.5	8.91
Center (Lacs, N'zi Comoè, part of Vallèe du Bandama)	4.8	8.56
South-West(Bas Sassandra)	4.2	6.94
Center-West (Fromager, Haut Sassandra, Marahouè)	3.7	4.85
Center-North (part of Vallèe du Bandama)	3.6	6.29
West (Dix-Huit Montagnes, Moyen Cavally)	3.5	4.76
North-East (Zanzan)	3.3	4.56
North (Savanes)	3.2	5.46
North-West (Bafing, Denguèlè, Worodougou)	1.7	4

D4D data

Mobile phone data sets originate from the service provider "Orange" in Ivory Coast and are further processed into four different D4D sets. Two of these were used in our study: SET1 and SET3. SET1 contains antenna–to–antenna communication traffic flow of five million Orange costumers aggregated to one hour time resolutions. Each record contains data about originating and terminating antennas of calls, number of calls and overall duration. SET2 observes users in consecutive two–week periods, which do not significantly influence HIV transmission patterns. On the other hand, insight into the long term mobility (5–months–long observation period) is possible trough SET3. Spatial resolution in this set is reduced from towers to sub–prefectures (255 spatial units). Records of this set contain user id, time stamp and sub–prefecture ids. Although SET4 provides connectivity at the user level and could be very informative for HIV epidemiology, it lacks spatial information. Users' ids cannot be related to the ids in the second or third set and therefore we were not able to approximate their home locations.

3.2 Continent–level dynamics

For the study focused on continent–level dynamics [50] YouTube was chosen as a source for data collection, as it is the most popular platform for video sharing and allows users to provide the geographical location where the uploaded videos were acquired. The obtained data was pre–processed and then analyzed using several geo–visualization techniques proposed by Andrienko *et al.* [3]. The process is driven by a human analyst through an interactive visual interface. Initially, density–based clustering is used to determine hubs that most travelers go through. Subsequently, we used route similarity clustering to identify the standard paths taken among the hubs. Once the routes have been identified, spatio–temporal clustering was used to identify the major time periods when the videos were taken. This yields the distribution of travels in time at a somewhat coarse resolution of days. Finally, a cluster analysis of spatio–temporal distance was conducted on longer paths in order to obtain times of day at which each of these characteristic events occur. This enabled us to identify the precise carriers and even departure times that were used by a majority of travelers.

The first step was to pre–process the data to eliminate entries that had invalid time and spatial content, as well as duplicate entries uploaded by the same user at the same geographical location. This was achieved by creating a grid of $1000mx1000m$ cells across the entire continent and removing the duplicate entries. Based on the pre–processed data set, we identified 43.917 trajectories. The entire data processing flow can be seen on Fig. 3.1.

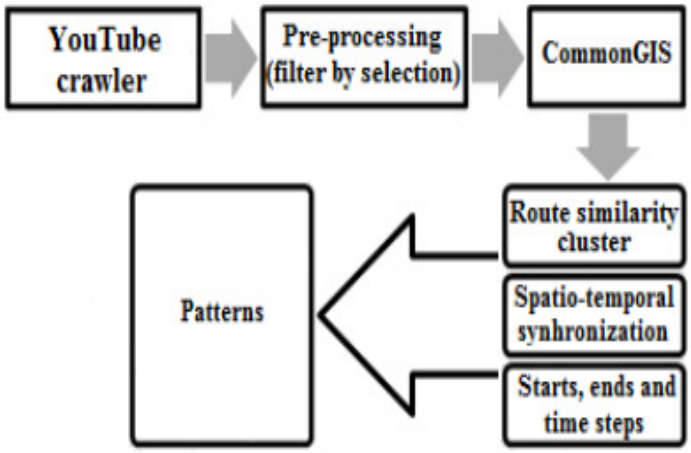

Figure 3.1: The data processing procedure

The clustering was performed using the OPTICS algorithm that allows different distance functions to be applied (see subsection **3.2.1**). A trajectory spatio–temporal sequence is a sequence of triples $T =< x_1, y_1, t_1 >, ..., < x_i, y_i, t_i >$, where $t_i (i = 1, ..., n)$ denotes a time stamp such that $\forall 1 \leq i < nt_i < t_{i+1}$ and (x_i, y_i) are points in $\mathbb{R}^2$.

The idea is that two trajectories (P, Q) are repeatedly scanned in search for the closest pair of positions with (D)–distance threshold. In the course of scanning two derivative distances are computed: the mean distance between the corresponding positions and a penalty distance. Skipping a position increases the penalty distance. Finding corresponding positions decreases the penalty distance. The final result is the sum of the two derivative distances. The size of the clusters obtained, based on this distance measure, represents how frequently the route was traveled [69].

Initial trajectory clustering was performed using a distance function dubbed "route similarity". The algorithm for calculating this distance is presented below.

```
-------------------------------------------------
dist = 0; pen = 0  //distance and penalty
n = 0              //number of corresponding points
i = 1; j = 1       //indices of points in P and Q
   WHILE i <= P.length AND j <= Q.length
      d = point_distance (Pi, Qj)
      WHILE i+1 <= P.length AND
         point_distance (Pi+1, Qj) < d
      pen = pen + point_distance (Pi, Pi+1)
      i = i+1; d = point_distance (Pi, Qj)
   END WHILE
   WHILE j+1 <= Q.length AND
         point_distance (Pi, Qj+1) < d
      pen = pen + point_distance (Qj, Qj+1)
      j = j+1; d = point_distance (Pi, Qj)
   END WHILE
   dist = dist + d; n = n + 1
   IF dist /n > D THEN RETURN D*2 END IF
   pen = pen - (D - d)
   i = i + 1; j = j + 1
END WHILE
dist = dist / n
WHILE i <= P.length
   pen = pen + point_distance (Pi-1, Pi)
END WHILE
WHILE j <= Q.length
   pen = pen + point_distance (Qj-1, Qj)
END WHILE

RETURN dist + pen
-------------------------------------------------
```

To perform pattern analysis with respect to the time of travelers' movements, the data set was restricted to users with 30 or more trajectory segments. This resulted in 14.167 unique users. We used the spatio–temporal distance function which computes the distance in space and time. It asks the user for an additional parameter: the

temporal distance threshold $maxT$, which is assumed to be equivalent to the spatial distance threshold $maxD$. The function finds the spatial distance d between the times of their occurrence. Then it proportionally transforms t into an equivalent spatial distance d' and combines d and d' in a single distance according to the equation for Euclidean distance shown by Eq. 3.1 and 3.2.

$$\phi(\delta, \delta') = \sqrt{(\delta_1 - \delta'_1) + ... + (\delta_n - \delta'_n)^2} \tag{3.1}$$

$$\phi(\delta, \delta') = \sqrt{\sum_{i=1}^{n} (\delta_i - \delta'_i)^2} \tag{3.2}$$

3.2.1 OPTICS algorithm

The OPTICS algorithm was proposed in order to overcome the difficulty in using one set of global parameters in clustering analysis [5]. OPTICS generalizes density–based clustering by creating an ordering of the points that allows the extraction of clusters with arbitrary values for ϵ. OPTICS does not explicitly produce a data set clustering, it outputs the cluster ordering. It also does not require the user to provide specific density thresholds.

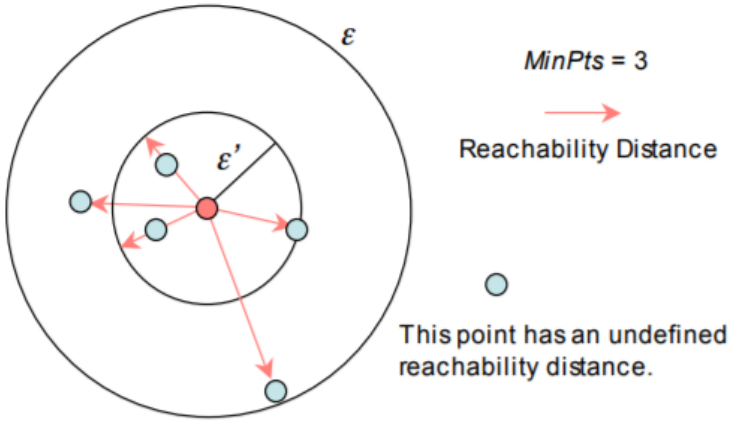

Figure 3.2: OPTICS illustration

The generating–distance ϵ is the largest distance considered for clusters. Clusters can be extracted for all ϵ_i such that $0 \leq \epsilon_i \leq \epsilon$. The core–distance is the smallest distance ϵ' between p and an object in its ϵ–neighborhood such that p would be a core object. The reachability–distance of p is the smallest distance such that p is density–reachable from a core object o. Fig. 3.2 illustrates the OPTICS algorithm and following is the code of the algorithm.

```
-----------------------------------------------
OPTICS(Objects, e, MinPts, OrderFile):
  for each unprocessed obj in objects:
    neighbors = Objects.getNeighbors(obj, e)
    obj.setCoreDistance(neighbors, e, MinPts)
    OrderFile.write(obj)
    if obj.coreDistance != NULL:
      orderSeeds.update(neighbors, obj)
      for obj in orderSeeds:
        neighbors = Objects.getNeighbors(obj, e)
        obj.setCoreDistance(neighbors, e, MinPts)
        OrderFile.write(obj)
        if obj.coreDistance != NULL:
          orderSeeds.update(neighbors, obj)
-----------------------------------------------
```

```
-----------------------------------------------
OrderSeeds: update(neighbors, centerObj):
  d = centerObj.coreDistance
  for each unprocessed obj in neighbors:
    newRdist = max(d, dist(obj, centerObj))
    if obj.reachability == NULL:
      obj.reachability = newRdist
      insert(obj, newRdist)
    elif newRdist < obj.reachability:
      obj.reachability = newRdist
      decrease(obj, newRdist)
-----------------------------------------------
```

3.3 Estimates on HIV prevalence at region level

National estimates on HIV hide the heterogeneity that exists within the country. To unveil subnational prevalence rates, a recently proposed package, *prevR* [75], relies on the estimation function and DHS measurements to generate a surface of HIV prevalence. This package performs a methodological approach for spatial estimation of regional trends of prevalence using data from surveys using a stratified two–stage sample design (as demographic and health surveys respectively). Estimates are based on a Gaussian kernel density function with adaptive bandwidth. In order to link aggregated behavioral patterns to HIV prevalence rates, high precision was required while doing the calculation. The estimate on HIV prevalence in a point in space (x, y) is determined by Eq. 3.3.

$$prev(x,y) = \sum_{i}^{n} \frac{1}{{h_i}^2} K\left(\frac{d_i}{h_i}\right) \tag{3.3}$$

where n is the number of samples, d_i the geometrical distance between sample i and point (x, y), K is the kernel function and h_i the bandwidth used for sample i. Additionally an indicator of the quality of the estimates was assigned to each region based on the survey sampling size [74]. Some estimates are very uncertain and should be interpreted with caution. See Table 4.3 in subsection **4.4.1** for estimated values and quality indicators.

3.4 Identification of strong social ties

Estimating the strength of social ties among people and classifying them as strong or weak helps in understanding socio–geographical relations [94]. In our study we aim at quantifying the tie strengths among administrative regions and identifying the strong ones. The ties between regions i and j are expressed through the communication or mobility flow ω_{ij} directed from i to j and quantified as the number of calls or mobilities originating from i and terminating at j divided by the population of the originating region. To categorize those connectivity ties as strong or weak we adopted the approach proposed in [110] where Eq. 3.4 is used as disparity filter for detecting the significant links. For code representation, see Appendix B, Module1: SET1 connectivity matrix, pp. 107.

$$\alpha_{ij} = 1 - (k-1)\int_0^{p_{ij}} (1-x)^{k-2}\mathrm{d}x < \alpha. \tag{3.4}$$

Where i and j are the indices of the administrative regions, α_{ij} is the significance of the tie between region i and region j, k is the degree of a node under consideration and $p_{ij} = \omega_{ij}/s_i$ corresponds to weights normalized by the node (region) strength $s_i = \sum_j \omega_{ij}$. The degree (k) of each node in our communication and mobility graphs is 49, inner loops are not taken into account. Ties with $\alpha_{ij} < \alpha$ are classified as strong ties, statistically significant at the level α. The underlying null hypothesis used in the significance inference assumes that the normalized weights of edges linking node i with its neighbors are produced by a random assignment from uniform distribution. In our experiment the statistical significance α is set at 0.05. The filter works locally at the level of nodes, but globally allows to preserve relevant fluctuations at different scales. After the filtering procedure, for the purpose of visualization, the directed graph was transformed into undirected by summing ω_{ij} and ω_{ji}.

3.5 Frequent trajectories

SET2 was analyzed using several geo–visualization techniques, with an emphasis on trajectory aggregation and clustering. The initial idea was to determine the hubs with the highest level of connectivity and to identify the major routes taken among the hubs. Subsequently, we used route similarity clustering to identify standard paths taken among the hubs.

The clustering was performed using the Ordering Points to Identify Clustering Structure (OPTICS) algorithm that allows different distance functions to be applied (See subsection **3.2.1**).

The size of the clusters obtained, based on this distance measure, represents how frequently the route was used. We analyzed each cluster set separately and then combined them in order to get a representation for the entire set. From Fig. 3.3 it is obvious that the main hubs are situated in the center of each region and that communication is higher in the southern part of the country. This confirms the hypothesis that the closer people are to the main routes and hubs, the higher the chance to get infected with a virus and to transfer it.

Figure 3.3: Trajectory aggregation model based on OPTICS "route similarity" clusterization

3.6 Regional connectivity and graph representation

We analyzed the spatial distribution and regional connectivity of HIV prevalence rates in Ivory Coast. Although valuable results have been achieved, poor data availability limited the spatial resolution of our study. Therefore we focus on regions as spatial units in order to be able to relate knowledge extracted from D4D sets to a spatial prevalence distribution.

A graph–based analysis was carried out for SET1 and SET3. We inferred the pairwise connectivity of regions by measuring the flow of communications and migrations between them (see Appendix B, Module 2: SET1 strong ties inference, pp. 109). We hypothesized that regions with higher HIV prevalence rates are more connected than those with lower.

To infer the inter–region communication graph we teased out the information from SET1. The first step was to assign each antenna to its region. Then, the communication flow is further aggregated at the region level by aggregating all antenna level flows between different regions. Nodes in the graph represent regions and the edges measure the strength of human interaction expressed through the number of calls. Graph loops–edges that start and end in the same region were excluded from this part of the study.

The next step was the normalization of edge weights. We took unevenly populated regions into account by dividing the edge weights w_{ij} (sum of all calls during the 6–month period between regions i and j and vice versa) with a product of population numbers N_i and N_j (population estimates from 2010 [133] were used). The product of $N_i \cdot N_j$ is an approximation of all possible communication links between people in the two regions. Finally, we filtered all of the obtained pairwise weights to create a 3NN graph (kNN–k Nearest Neighbor) [44]. By adding only the three strongest links for each region, we can inspect the major directions and hubs of communication flow in Ivory Coast. The graph is presented at Fig. 3.4.

Nodes are geographically ordered and their colors indicate the HIV infection rate: from red that denotes the regions severely affected by HIV to yellow for moderately affected regions. Added edges are presented with different widths and color intensities to highlight differences in their strength. We notice that the graph structure corresponds well to the spatial distribution of HIV. Southeastern and southern parts of the country that have higher prevalence rates, turn out to be the more connected part of the graph in terms of incidence edges, hubs and edge widths. The major hubs, Lagunes with 8 incidence edges, Bas Sassandra with 7 and Lacs with 6, are located in the area of the highest risk. The northwestern part of the country is notably sparser in the graph with no more than 3 thin incident edges. An interesting property revealed by this graph is that the gravitational law previously observed in inter–city commutations [71] is only partly supported. Although the proximity of regions is correlated with the strength of the link between them, we found some exceptions. One of the exceptions are the links from Zanzan and Denguélé to the distant Lagunes region, rather than to some closer region. The reason probably lies in the fact that Abidjan is the economic capital of Ivory Coast and that it holds 20 percent of the

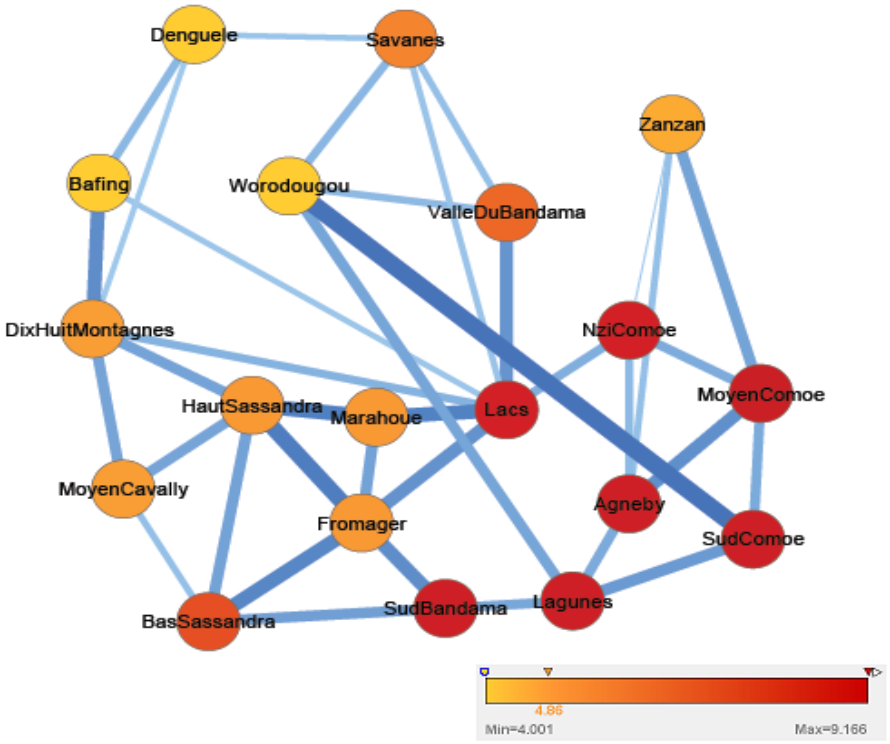

Figure 3.4: 3NN communication graph: Nodes represent Ivory Coast regions, arranged in geographical order and colored according to HIV prevalence rates. Links are inferred from inter–region communication flow during six months. Their color and width is proportional to the normalized flow between regions

overall population of the country. The inference of the migration graph was done in a way similar to that used for the communication graph. The data source used is SET3.

First, we estimated for each user his home location in order to be able to assign him a home region. Then, we followed all his movements through time and detected transitions into other regions. Upon detection of a regional transition the pairwise matrix of region to region migrations is updated and the value increased at position (home, detected host region). All users were processed in this manner and the final result is a pairwise matrix of overall migrations between regions during the 6–month period. Before creating the 3NN graph, the values in the matrix were normalized. The applied normalization was different than for SET1. From the estimated home locations, we can obtain the number of residents that were tracked in SET3 for each region. Edge weights w_{ij} (sum of all migrations between regions i and j and vice versa) with the sum of obtained resident numbers for region i and j: N_i and N_j. After this normalization step, the 3NN graph was built. Its structure is presented in Fig. 3.5.

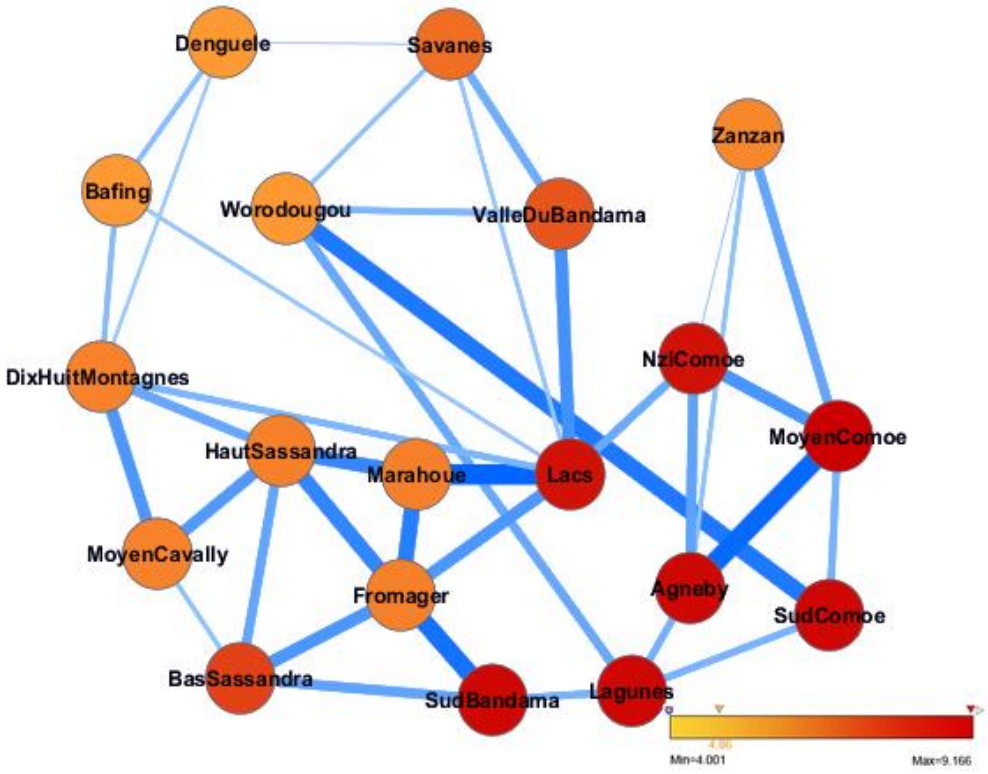

Figure 3.5: 3NN migration graph: The nodes represent Ivory Coast regions, arranged in geographical order and colored according to HIV prevalence rates. Links are inferred from inter–region migration flows during six months. Their color and width are proportional to normalized flow between regions.

Normalized migration flows also support the spatial HIV distribution, as the graph is denser in the high risk area. The only link that significantly departs from the distribution and our expectations is the link between Worodougou and the Sud Comoè region. A possible reason is that only 8,653 residents of Worodougou were covered by SET3 and this number should be near 12,000 to be in accordance with the real population distribution. Nevertheless, even with a correction that link would still exist. Better to look at that link as a new insight. The strong connection may indicate the next hot spot of HIV epidemics and we can utilize that for prioritizing areas for intervention. Outliers can also uncover strange occurrences in the field. In this particular case the outlier could be related to the mining of diamonds from Toubabouko field located in Worodougou and their export despite a United Nations (UN) ban.

3.7 Feature selection

The initial part of the study was more descriptive and focused on linking regions' connectivity inferred from D4D data with HIV spatial distribution. Numerous features were extracted in order to quantify behavioral and mobility patterns for each region. We then built regression models and evaluated their performance when it comes to predicting regional prevalence rates. D4D SET1, 2 and 3 were analyzed separately. From SET1 we extracted features related to intra–region communications. For each region we created average profiles of communication flow in one–hour time resolutions for weekdays and weekends. Profiles contain the average number of calls and their average duration and are normalized by the number of people in the region. Additionally, we created aggregated features for night hours (22h–05h) during weekdays, weekends and whole days. In total we derived 104 features from SET1. Very often the limited knowledge of an individual's trajectories can be significant for human mobility monitoring because individuals can be traced during a certain period of time. SET2 contained high resolution trajectories of randomly sampled individuals over two–week periods. With feature extraction we gained the intervals when people are more active (working and non working days, working and non working hours, weekends, nights etc.), as well as the home and visited regions per each user. We assumed that the time when people are more active increase the chance of infection, as well of virus transmission.
SET3 is first analyzed at the user level and then based on home location estimates, individual patterns of daily movements were aggregated into region–level features. We calculated various aspects of mobility such as gyration, radius, diameter and approximate sum of all distances that users travel [55, 35], counted the number of distinct sub–prefectures that users visited within the home region and out of it during the 6–month period and under specific time constraints: only during night hours or weekends. By tracking moving trajectories of users we determine in and out migrations for each region at different time scales: staying in host region more than 3, 5 or 10 days. We also measured how long, on average non–residents stayed in each region. Overall number of features from SET3 is 23.

3.8 Regression models

3.8.1 Elastic net predictive model

In statistics the Elastic Net is a regularized regression method that linearly combines the l^1 and l^2 norm penalties of the lasso [123] and ridge methods [42] (the Elastic Net adds a quadratic part to the penalty $||\beta^2||$). The l^1 norm penalty generates a sparse model and l^2 norm penalty: removes the limitation on the number of selected variables, encourages grouping effect, and stabilizes the l^1 regularization path. For equation, see Eq. 3.5.

$$argmin_{\beta} \sum_{i} (y_i - \beta' x_i)^2 + \lambda_1 \sum_{k=1}^{K} |\beta_k| + \lambda_2 \sum_{k=1}^{K} \beta_k^2 \tag{3.5}$$

The Elastic Net predictive model performs automatic variable selection and continuous shrinkage. It produces a sparse model with prediction accuracy, while encouraging grouping. The empirical results and simulations demonstrated good performance of the elastic net and its superiority over the lasso predictive model. The Elastic Net is particularly useful for problems where the number of features is higher than the number of samples ($p >> n$). The prediction procedure can be divided into three steps: approximation of the unpenalized log–likelihood using iteratively re–weighted least squares; application of soft–thresholding to take care of lasso contribution to the penalty and application of proportional shrinkage for the ridge penalty [150].

In late 2014 it was demonstrated that the Elastic Net can be reduced to the linear support vector machine [149]. This reduction is called the kernel Elastic Net and it enables the estimation of $p(y|x)$ in Support Vector Machine (SVM). Eq. 3.6 and 3.7. shows linear SVM and kernel Elastic Net reduction (including *loss function*).

$$\begin{aligned} \hat{\alpha} = arg \min_{\alpha} \frac{1}{n} \sum_{i=1}^{n} \phi(y_i \sum_{i=1}^{n} \alpha_i k(x_i, x)) + \lambda_2 \alpha^T K \alpha \\ \phi(y, f) = (1 - yf)_+ \end{aligned} \tag{3.6}$$

$$\begin{aligned} \hat{\alpha} = arg \min_{\alpha} \frac{1}{n} \sum_{i=1}^{n} \phi(y_i \sum_{i=1}^{n} \alpha_i k(x_i, x)) + \lambda_2 \alpha^T K \alpha + \lambda_1 \sum_{i=1}^{n} |\alpha_i| \\ \phi(y, f) = log(1 + exp(-yf)) \end{aligned} \tag{3.7}$$

The estimation of the $p(y|x)$ form is for the loss function of kernel Elastic Net. The implication of this reduction is that SVM solvers can also be used for Elastic Net problems.

A similar reduction has previously also been demonstrated for the Lasso in 2014. The authors show that for every instance of the Elastic Net, an artificial binary classification problem can be constructed such that the hyper–plane solution of a linear SVM is identical to the solution β (after re–scaling). The reduction immediately enables the use of highly optimized SVM solvers for Elastic Net problems. It also enables the use of GPU acceleration, which is often already used for large–scale SVM solvers. Fig. 3.6 shows the geometric illustration of Elastic Net, ridge regression, and Lasso.

3.8.2 Ridge regression

Ridge regression is a variant of ordinary multiple linear regressions whose goal is to circumvent the problem of instability arising, amongst other, from collinearity of the predictor variables. There are three main variants of ridge regression: ordinary

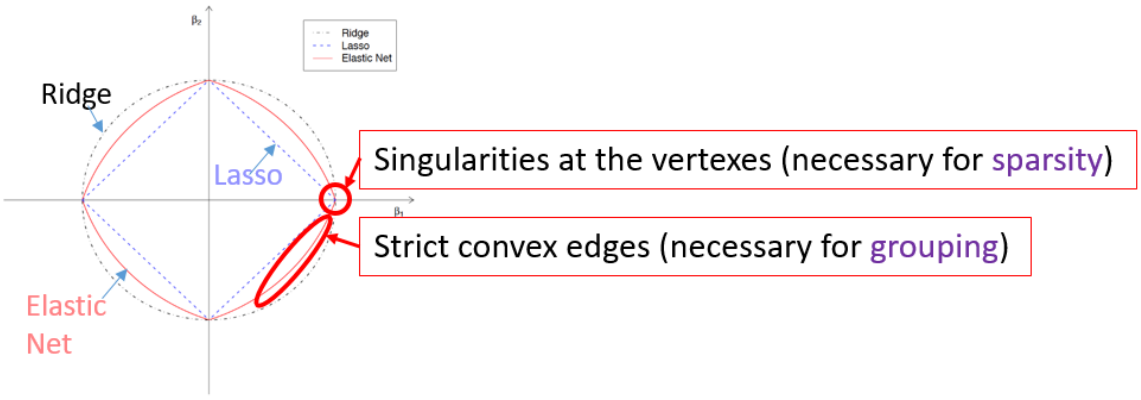

Figure 3.6: Geometric illustration of Elastic Net, ridge regression and LASSO

ridge regression, generalized ridge regression, and directed ridge regression. Given a response vector $y \in R^n$ and a predictor matrix $X \in R^{nxp}$, the ridge regression coefficients are defined by Eq. 3.8.

$$\hat{\beta}^{ridge} = \underset{\beta \in R^p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_1 - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \underset{\beta \in R^p}{\operatorname{argmin}} \underbrace{||y - X\beta||_2^2}_{\text{Loss}} + \underbrace{\lambda ||\beta_2^2||}_{\text{Penalty}} \tag{3.8}$$

Ridge regression works with the original variables and tries to minimize penalized sum of squares. Like ordinary least squares, ridge regression includes all predictor variables, though typically with smaller coefficients, depending upon the value of the complexity (tuning) parameter λ, which controls the strength of the penalty term. Note that:

- When $\lambda = 0$, it is a linear regression estimate.
- When $\lambda = \infty$, it is $\hat{\beta}^{ridge} = 0$.
- When $0 < \lambda < \infty$ in between, fitting a linear model of y on X, and shrinking the coefficient.

The selection of the ridge parameter λ plays an important role, it multiplies the ridge penalty and thus controls the strength of shrinkage of coefficients toward zero [42]. Value of λ is estimated through leave–one–out validation.

Ridge regression performs particularly well when there is a subset of true coefficients that are small (close to zero) or even zero. It doesn't do as well when all of the true coefficients are moderately large (means more shrinkage). However, in this case it can still outperform linear regression over a pretty narrow range of (small) λ values.

3.8.3 Support vector regression

Support vector machines are a set of supervised learning methods used for classification and regression analysis, motivated by results of statistical learning theory [126]. Originally developed for pattern recognition, they represent the decision boundary since the weights w_i of the decision function $D(x)$ are a function only of a small subset of the training examples called the "support vectors".

A type of SVM for regression analysis is called Support Vector Regression (SVR). SVR searches for optimal repression function, but allows a tolerance margin – ε. This way a tube around regression functions is created ignoring errors in predictions on training data. While measuring the loss incurred for the observed pattern, there is a large area where zero loss is accrued: whenever a pattern is on the correct side of the decision, and does not touch the margin, it does not contribute any loss to the objective function. Correspondingly, it does not carry any information about the position of the decision surface, the latter is computed by minimizing that very objective function. A loss function for regression estimation must also have an insensitive zone, hence ϵ–insensitive loss is used.

In order for this to apply to SVR, [126] devised the ε–insensitive loss function $|y - f(x)|_\varepsilon = max\{0, |y - f(x)| - \varepsilon\}$, which does not penalize errors below some $\varepsilon > 0$, chosen a priori. SVR algorithm, seeks to estimate functions (see Eq. 3.9, 3.10 and 3.11).

$$f(x) = (w \cdot x) + b, w, x \in R^N, b \in R \tag{3.9}$$

Based on data

$$(x_1, y_1), ..., , (x_t, y_t) \in R^N xR \tag{3.10}$$

by minimizing the regularized risk function

$$\frac{||w||^2}{2} + C \cdot R^\varepsilon_{emp} \tag{3.11}$$

where C is a constant determining the trade–off between minimizing training errors and minimizing the model complexity term $||w||^2$, and $R^\varepsilon_{emp} := \frac{1}{l}\sum_{i=1}^{l} |y_i - f(x_i)|_\varepsilon$

The parameter ε can be useful if the desired accuracy of the approximation can be specified beforehand.

The method also includes regularization parameter in a form of cost parameter–C, that penalize the training errors outside the tube. In our experiments we used linear kernel, default $\varepsilon = 0.1$, and value of C is estimated though leave–one–out validation.

3.8.4 Recursive feature elimination

Recursive Feature Elimination (RFE) is a greedy method for selecting defined a number of features. It starts from an initial set of features build in a model (in our case SVM or Ridge), it assigns weights to each feature based on an estimate from the predictive model, eliminates the lowest ranked feature and then recursively repeats

this procedure on the remaining set of features until it reaches the desired number of features. The criteria $DJ(i)$ or $(w_i)^2$ estimates the effect of removing one feature at a time on the objective function. This causes the method to be very sub–optimal when it comes to eliminating various features at a time, which is needed to obtain the small subset of features. This problem can be overcome by using the following iterative process which is called RFE:

- Train the classifier (optimize the weights w_i with respect to J (cost function)).
- Compute the ranking criteria for all the features ($DJ(i)$ or $(w_i)^2$).
- Remove the feature with the smallest ranking criteria.

For calculation purposes it is more efficient to remove several features at a time, at the expense of possible classification performance degradation. In such a case the method produces a feature subset ranking, as opposed to a feature ranking. Feature subsets are nested $F_1|F_2| \dots F_m$. If features are removed one at a time there is also a corresponding feature ranking. However, the features that are top ranked (eliminated last) are not necessarily the ones that are individually most relevant. Only taken together the features F_m are optimal. The output is a top ranked feature subset obtained through this recursive procedure [57]. What is also important to mention here is that RFE has no effect on correlation methods since the ranking criteria is calculated with information about a single feature.

3.9 Statistical modeling

A statistical model represents a set of assumptions concerning the generation of some sample data (of a larger population) and is a class of mathematical model. The assumptions embodied by a statistical model describe a set of probability distributions, some of which are assumed to adequately approximate the distribution from which a particular data set is sampled. The probability distributions inherent in statistical models are what distinguishes statistical models from other, non-statistical, mathematical models. Distributions of interest for this thesis are heavy tailed and power–law.

3.9.1 Lévy flight

A Lévy flight is a random walk in which the step–lengths have a probability distribution that is heavy–tailed. When defined as a walk in a space of greater than one dimension, the steps made are in isotropic random directions. It describes a class of random walks whose step lengths follow a power–law tailed distribution [113]. Lévy flights are Markovian random processes whose underlying "length of the jump" distribution displays the long–tailed form $\lambda(x) = |x| - 1 - \alpha$, with $0 < \alpha < 2$. Probability density in a homogeneous environment is defined through the function $P(k,t) = FP(x,t) = exp(D|k|\alpha t)$. As such, Lévy flights are a natural generalization

of Gaussian diffusion processes ($\alpha = 2$).

In the course of the random walk governed by $\lambda(x)$ with $0 < \alpha < 2$, longer jump lengths are characteristic, leading to unique trajectories with fractal dimension α. Because of that, processes with an underlying Lévy stable jump length distribution are called Lévy flights [113]. A difference between the trajectory of a Gaussian and a Lévy flight trajectory is shown in Fig. 3.7, for the same number of jumps (approx. 7000).

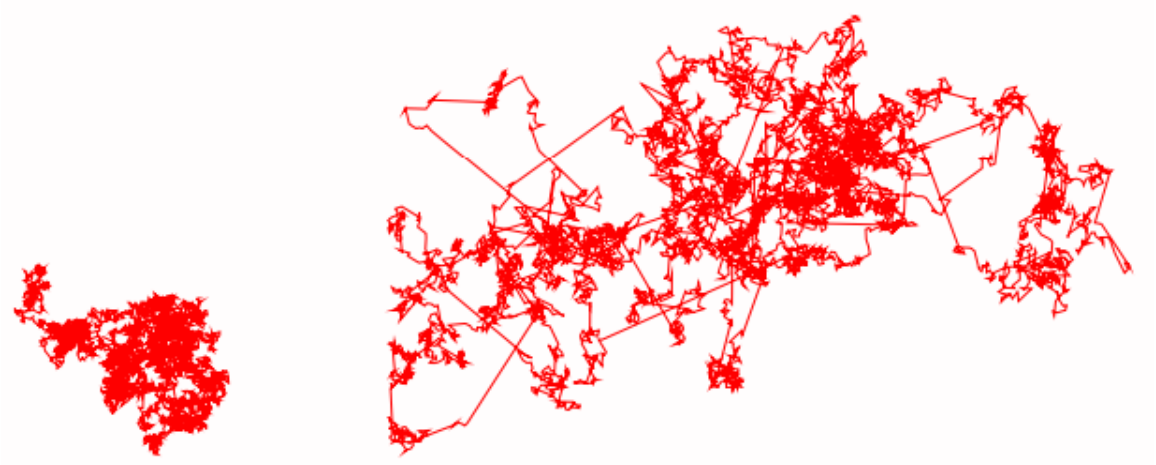

Figure 3.7: Difference between Gaussian (left) and Lévy (right) trajectories

A Lévy flight distribution is commonly observed in nature. It has been proven to be the best strategy that can be adopted in random searches [129]. Lévy flights have been observed in the patterns of movement of different species ranging from albatrosses [128] to marine predators [66, 115], monkeys [97], and mussels [37]. Lévy flights were also used to model groundwater flow [14].

In a Lévy flight the probability $P(d)$ that the walker performs a jump of length d is given by Eq. 3.12.

$$P(d)\ d^{-\gamma} \tag{3.12}$$

This function is scale free, it exhibits the same patterns regardless of the range over which they are observed. Lévy flights underlie many aspects of human dynamics and behavior, and have the potential to become a merge concept for problems traditionally addressed by different disciplines, such as social and cognitive science and urban planning. Lévy flights represent a widely used tool in the description of anomalous stochastic processes. Their statistical limit distribution emerges from independent identically distributed random variables, by virtue of the central limit theorem. Despite the quite straightforward definition, Lévy flights are less well understood than one might assume. This is due to their strongly non–local character in space. These long–range correlations spanning the entire available geometry.

3.9.2 Power–law distribution

Discovering behavior that follows a power–law distribution in both natural and human–made systems is challenging.

The power–law, or scaling law, states that a change in one quantity creates the same proportional relative change in another. The easiest way to explain this is with the relation between the side of a square and its area. A square with sides of $10cm$ has an area of $100cm^2$. If you change the length of the sides to $100cm$ the area will change to $10,000cm^2$. A ten–fold increase in one quantity results in a hundred–fold increase in another. A power–law distribution has the form of $Y = kX_a$, where X and Y are variables of interest, a is the law's exponent and k is a constant.

The fastest way to find out if a distribution is within a power–law is by plotting two quantities against each other with logarithmic axes. If this shows a linear relation, it is a power–law distribution. One of the most common known power laws is the Pareto principle, also know as the 80/20 rule. This states that 80% of the work is a result of 20% of the effort. Power–law relations are interesting for many (theoretical) reasons. Though it is easy to mistake a distribution for a power law distribution. Primarely log–normal distributions are often mistaken for power–law distributions, mostly because of too little data. It will be approximately linear for large values, but will drop off quickly for small values. This is illustrated in a log–log plot that is curving downwards.

There are many different mechanisms for producing power–laws and different ones are applicable to different cases. One of the studies that was conducted shows the cumulative distribution of the number of calls received on a single day by 51 million users of AT&T (telephone service in United States of America) [2]. Another study was investigating cumulative distribution of the intensity of 119 wars from 1816 to 1980 [116]. War intensity is defined by taking the number of battle deaths among all participant sides in a war, dividing by the total combined populations of the countries and multiplying it with 10 thousand. The comprehensive study of geographic information in this context has been conducted on Wikipedia articles, where [59] shows that editors of Wikipedia follow a power–law in number of contributions.

Mathematically speaking, a power–law distribution has a probability $p(x)dx$ of taking a value in the interval from x to $x + dx$, where

$$p(x) = Cx^{-\alpha} \tag{3.13}$$

and $\alpha > 0$.

Power–law distributions occur in many situations and bring significant information for our understanding of natural and human–made phenomena. Unfortunately, the detection and characterization of power–laws is complicated by the large fluctuations that occur in the tail of the distribution. Commonly used methods for analyzing power–law data, such as least–squares fitting, can produce substantially inaccurate estimates of parameters for power–law distributions.

Chapter 4

Results

This chapter provide an insight into the conducted research and their results. In the first part we describe continent–level dynamics, limitations and challenges we faced. Secondly, we discuss human mobility patterns and tourist dynamics. Finally, we give a detailed overview of the research results obtained in the field of digital epidemiology.

4.1 Continent–level dynamics

While investigating the continent–level dynamics and human mobility [50], three types of interesting results have been generated: graphic results of route similarity clustering, the main directions of movement identified (10 of them) and the results of the temporal analysis. The dominant patterns identified have been further analyzed by a human expert with respect to other data available online, allowing us to identify the carriers and connections taken by the majority of travelers (see Table 4.1).

Table 4.1: Basic data statistic for used data set

Number of videos	143 532
Time frame	2006 (September)–2011(April)
Mean number of views	6697
Average path length	4500
Unique users	113 157

Fig. 4.1 shows the results of initial trajectory clustering. The main hubs form three triangles, with initial points in Spain, Morocco, Egypt, and Israel and end points in Tanzania and South Africa. A potential reason for this kind of division lies in the fact that these are the main tourist destinations on the continent, but also because these cities represent the main and busiest airport centers in Africa[1]. An examination of the start and end points of the trajectories provides the main direction of travel. Fig. 4.2 shows the ten main routes (with the primary direction) of movement across the continent. As can be seen from both Fig. 4.1 and 4.2, the amount of movement detected in the central part of the continent is relatively small. The clustering of

[1] www.myairnigeria.com

the trajectories by time, coupled with the extensive search of other online resources, allowed for the analysis of all main routes from the point of time, transportation mode and carriers (air, bus and ship companies).

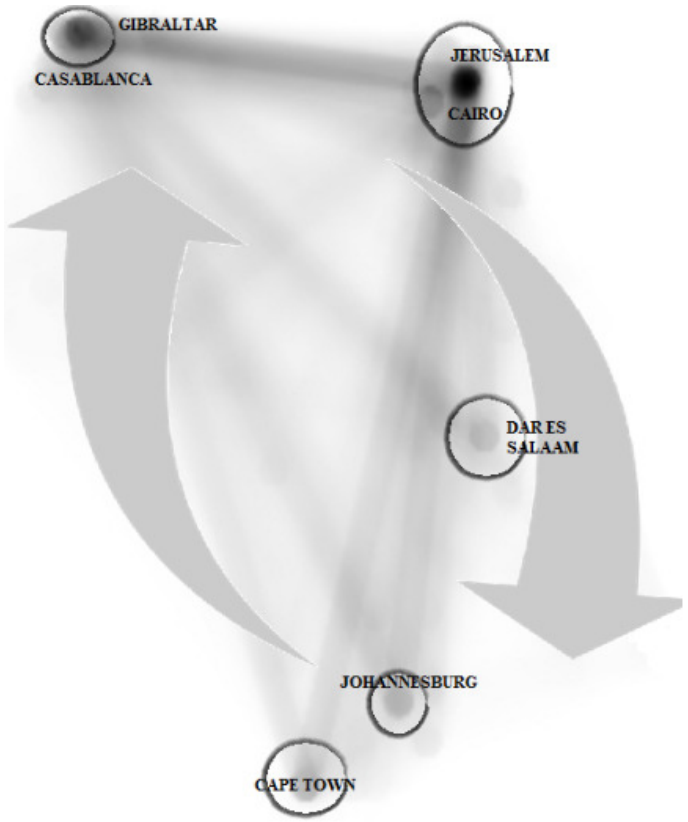

Figure 4.1: Results obtained by route similarity clustering

We used Expedia as a source of data about transportation companies and their timetables, as well as web sites of local airlines (Egypt air, Kenya Airways, Ethiopian Airways etc.). Based on the length of the trip (approximately 15 hours) and the distance traveled (approximately 5000 km), we assumed that the main mode of transportation mode is by airplane, except between Cairo and Hurghada (where the distance is approximately 500 km). We assumed that the main mode of transportation is by bus, which corresponds well with the obtained departure times. The results of the data analysis showed that the most common time of departure for trips between those two cities is early in the morning (around 7 am) and in the evening (around 9 pm), which coincides with the departure of local buses. Table 4.2 shows results of temporal clustering for observed results.

Table 4.2: Results of temporal clustering

Routes	Departure Time	Comments
Rabat–Cairo	8.00 am	
Antananarivo–Cape Town	10.00 am	
Brazzaville–Conakry	12.00 am	
Cotonou–Johannesburg	17.40 pm	corresponds to Air Nigeria flight[1]

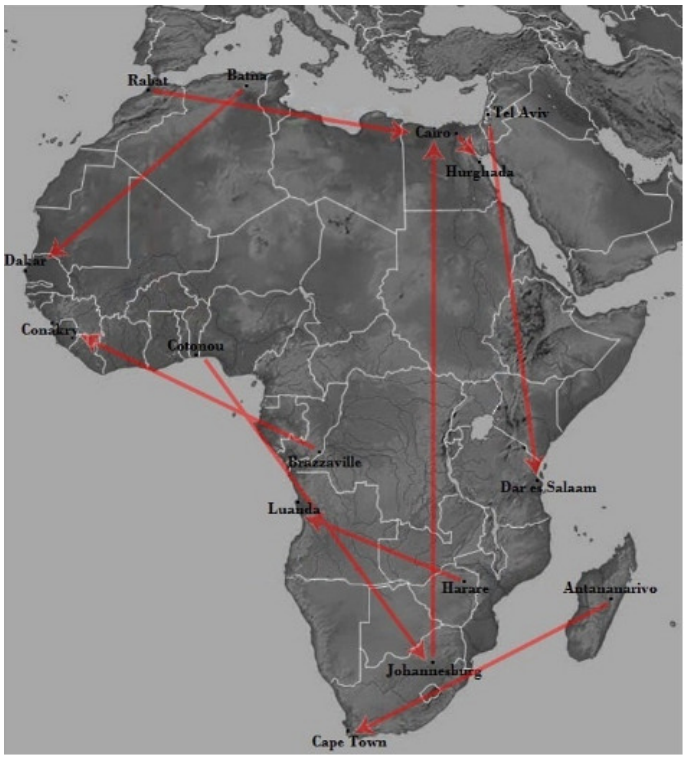

Figure 4.2: Main movement directions on the continent

For the route between Cape Town and Antananarivo it is noticeable that the number of departures in the summer period (May–August) is larger than in rest of the year due to the tidal sea, but also that there is a stop in Mombasa (Kenya), which seems to correspond to the possibility to use boat transport during the summer (May–September). The boat stops in Mombasa, from where people mostly use airplanes to Cape Town (based on travel times obtained from clustering). The flights favored by the YouTube users are those with no stops or plane changes. Also those that stop in African countries are preferred over those stopping in Europe or Turkey.

Regarding carrier companies, one can conclude that the most widely used local companies are Ethiopian Airways and South African Airways, while Air France is the dominant carrier among international air companies. This was inferred by correlating the observed times of departure, arrival and duration of a trip and available flights of each of the major routes.

4.1.1 Limitations and challenges

There were several limitations of the conducted study that should be considered:

- Africa is an under–developed continent in terms of information and communication systems, which is why the data set obtained by the YouTube crawler is relatively small.
- Earth curvature and the extent of the territory considered makes it hard to determine relevant route similarity clustering parameters.
- The results obtained by temporal clustering had to take into account the different time zones in Africa.

The first issue can possibly be addressed by conducting research using additional publicly available data, such as that from Flickr or Panoramio. Other limitations were managed by exploring a wide range of different cluster parameters in order to get the optimal solution.

4.2 Human mobility patterns

To explore the statistical properties of Flickr users' mobility patterns, we first took a look at the displacements between a user's successive positions. We found that the distribution of displacements can be described well using the truncated power–law presented by Eq. 4.1.

$$P(\Delta r) = (\Delta r + \Delta r_o)^{-\beta} exp(\frac{-\Delta r}{k}) \tag{4.1}$$

with exponent value $\beta = 1.65 \pm 0.15$ (for SET1) and $\beta = 1.70 \pm 0.18$ (for SET2) (mean $\pm$ standard deviation), $\Delta r_o = 1km$ and cut–off value $k = 50km$ (Fig. 4.3). Note that the observed scaling exponent is between $\beta = 1.75 \pm 0.15$ observed in [55] for a mobile phone dataset and $\beta = 1.59$ observed in [41] for a high–resolution data set of wandering albatross flights.
This suggests that all three distributions capture the same fundamental mechanism driving human mobility patterns. Values for Δr_o and cut–off value k ($\Delta r_o = 1.5km$ and $k = 80km$) are also close to what was obtained in [55]. The difference in the value of Δr_o may be due to the fact that the data used in this study is actually more precise in terms of users' position, as the mobile phone data had to be approximated to the center of the network cell. A plot of the Probability Density Function (PDF) of the displacements is shown on Fig. 4.4. As the figure indicates, SET2 fits the power–law better, but the general trend is present in both datasets.

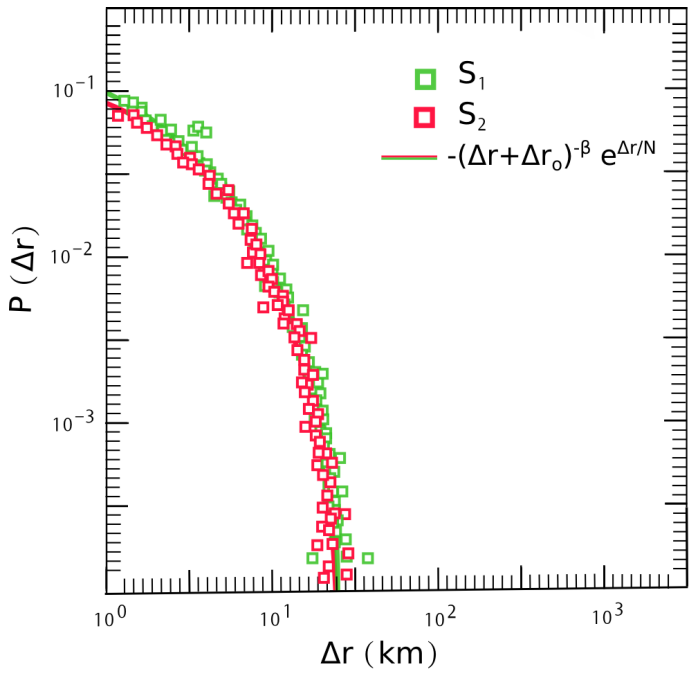

Figure 4.3: Probability density function ($P\Delta r_o$) of travel distances for the entire dataset

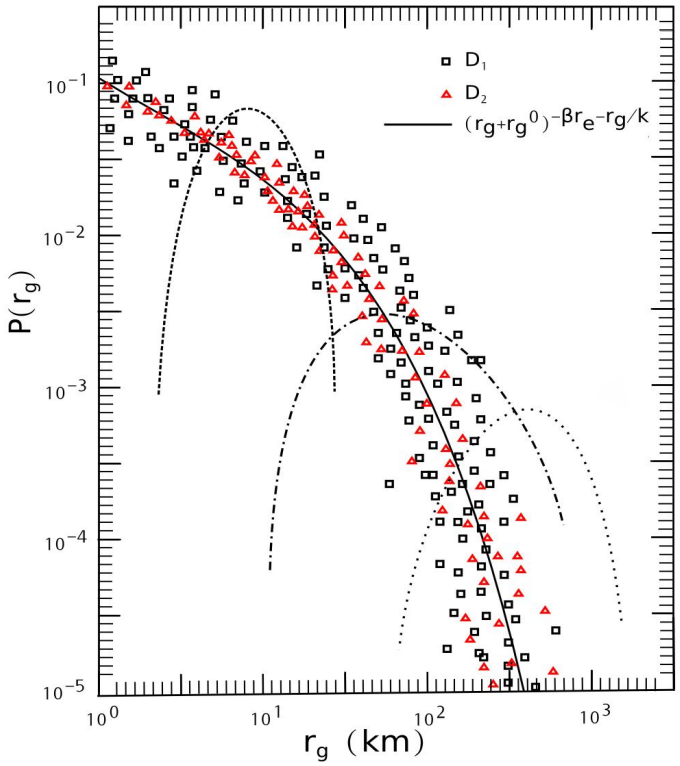

Figure 4.4: The distribution of $P(r_g)$, where $r_g(T)$ was measured after $T = 12$ months of observation. The dotted, dashed and dot–dashed curves show $P(r_g)$ obtained from the standard null models (Random walk, Lévy flight and truncated Lévy flight)

Next we attempted to see if individual users exhibit the same regularities of motion observed in [55]. To do so we first determine the radius of gyration for all Flickr users in SET1 and SET2, Eq. 4.2.

$$r_g = \sqrt{\frac{1}{n}(\sum_{i=1}^{n}(x_i - x_{cm})^2 + (y - y_{cm})^2)} \tag{4.2}$$

where x_{cm} and y_{cm} represent the center of mass position calculated by Eq. 4.3 and Eq. 4.4 respectfully:

$$x_{cm} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{4.3}$$

$$y_{cm} = \sum_{i=1}^{n} \frac{y_i}{n} \tag{4.4}$$

where (x_i, y_i) are the x and y coordinates of the center of the cluster of positions visited by a single user and n is the number of positions. We find that the distribution of the radius of gyration $P(r_g)$, shown on Fig. 4.4, can also be approximated with a truncated power–law by Eq. 4.5.

$$P(r_g) = (r_g + r_g^0)^{-\beta r} exp(\frac{-r_g}{k}) \tag{4.5}$$

With $r_g^0 = 8km$, $\beta_r = 1.75 \pm 0.25$ and $k = 50km$.

Lévy flight is characterized by a high level of heterogeneity, giving the possibility that Eq. 4.5 could emerge from an ensemble of identical agents, each following a Lévy flight. Therefore, we compare $P(r_g)$ with the distributions of r_g generated by an ensemble of agents following a random walk, Lévy flight and truncated Lévy flight [100, 9]. Even though an ensemble of Lévy flight agents display a significant degree of heterogeneity in r_g, it is not sufficient to explain the truncated power–law distribution $P(r_g)$ exhibited by Flickr phone users. A similar effect has been observed in [55] for mobile phone users.

Taken together, Fig. 4.3 and 4.4 suggest that the difference in the range of typical mobility patterns of individuals (r_g) has a strong impact on the truncated Lévy flight behavior described by Eq. 4.1.

If individual trajectories are described by a Lévy flight or a truncated Lévy flight, the r_g should increase with time as $r_g(t) \sim t^{3(2+\beta)}$ [62], or for a random walk, $r_g(t) \sim t^{\frac{1}{2}}$ [55]. The longer we observe a user, the higher the chances are that she/he will travel to areas not visited before. This has been proven for mobile phone users in [55]. We expect the Flick user to behave in a similar fashion.

To check this assumption we measured the time dependence of the radius of gyration for users whose radius would be considered small ($r_g(T) \leq 5km$), medium ($10 < r_g(T) \leq 15km$) or large ($T > 30km$). The result is shown in Fig. 4.5. As is the case with mobile phone users, the time dependence is better approximated by a logarithmic increase, than what we would expect for a Lévy flight or random walk models.

Finally, following the procedure done in [55], we selected users with similar asymptotic $r_g(T)$ after $T = 12$ months and examined the jump size distribution $P(\Delta r \mid r_g)$ for each group. The authors used this approach to observe that the users with a small r_g usually travel over small distances, while those with larger r_g have a tendency to make longer trips. This cannot be corroborated with our data. However, once when the distribution is rescaled by r_g, the variance is reduced and the data collapsed into a single curve, suggesting that a single jump distribution characterizes all users independent of their r_g.

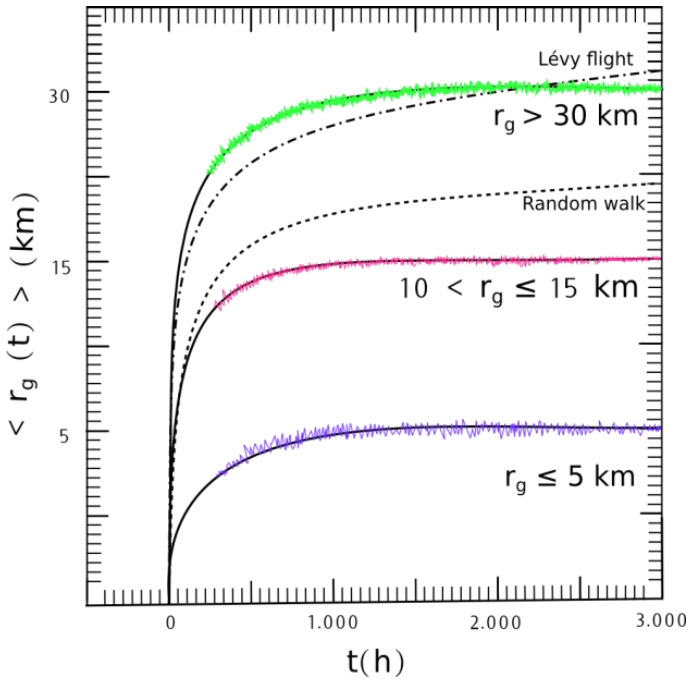

Figure 4.5: Radius of gyration versus time, separated into three groups according to their final $r_g(T)$, $T = 12$ months

The results presented in this section were substantiated using the Kolmogorov-Smirnov test (K–S test) for the goodness of fit of empirical data to the fitted distribution [102].

4.3 Tourist dynamics

In this section we present the results of an approach to detect locations attractive to tourists and discover tourist dynamics based on a data set obtained from metadata

attached to publicly available geo–referenced images on Flickr [49]. The goal was to identify attractive locations in a city (Berlin), assuming that an attractive place is characterized by large amounts of images taken by many people. In addition, we attempted to identify standard routes people take when making a tour. To achieve this we proposed an approach based on density–based clustering and route similarity and dynamics clustering.

Fig. 4.6 shows a plot of the spatial positions at which the images in the data set were taken. Each dot represents a single image in the data set. The result is a strikingly accurate map: city center boundaries are easily recognized and it is easy to see the hot–spots of photo activity corresponding to locations of major tourist attractions.

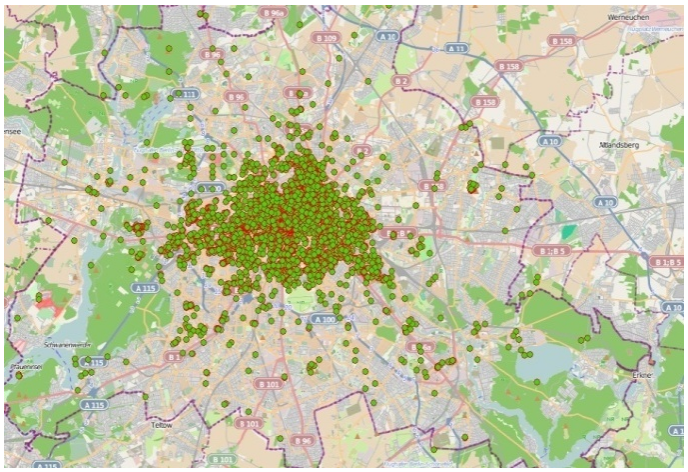

Figure 4.6: Images taken in the Berlin city center

The results of cluster analysis using the OPTICS algorithm can be seen in Fig. 4.7. The figure shows that the data is tightly–grouped near the center of the city, since the most dots (green) cover that area, where Brandenburg Gate, Reichstag, Unter den Linden, Berlin Cathedral etc. are situated. There is also big cluster in the area of the Olympiastadion, the seat of Hertha BSC (soccer club), which is logical since they hosted the FIFA World Cup in 2006 and a lot of matches are still played on this field.

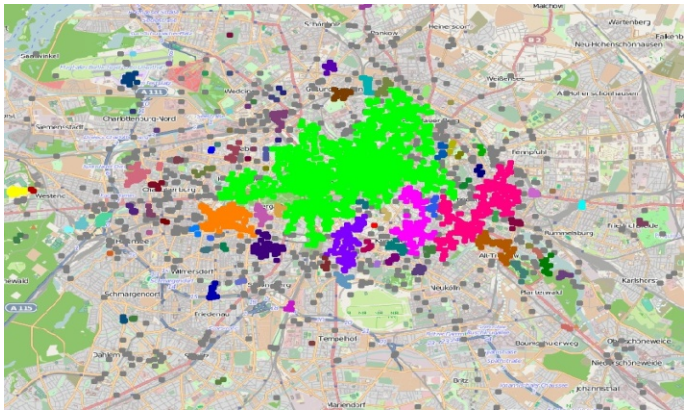

Figure 4.7: Results of OPTICS clustering for the entire city

Fig. 4.8 presents the map obtained though the density–based clustering of the largest cluster, which shows popular locations in the city center. This approach identified 28 of the most attractive locations in Berlin, which include: Victory Column, Checkpoint Charlie, Reichstag, Brandenburg Gate, Charlottenburg Palace, Kaiser Wilhelm Church, Central Station, German Federal Parliament, Topography of Terror, Postdam Square, Holocaust Memorial, Remains of Berlin wall etc. In addition to these locations, there are several locations supported by less data, but still significant: Jewish Museum, DDR Museum, Old Library, Bellevue Palace and Karl–Marx Allee.

Figure 4.8: Results of OPTICS clustering for the city center

To enhance visualization, the results were exported to KML format, used by Google Earth[2], providing a map with placements corresponding to the results of clustering, so it would be easier to track locations but also to see pictures of them obtained through Google Earth. Fig. 4.9 and 4.10 show the main directions of tourists' routes through the city center, based on the two distance functions considered. The clusters are presented in a summarized form as flow maps, so the routes can easily be seen.

[2]www.google.com/earth

Fig. 4.9 shows that a central point is Brandenburg Gate from where a path is leading towards Reichstag, Victory Column, Postdam Square and through the Unter den Linden (a historical part of Berlin) to Alexanderplatz.

Figure 4.9: The flow of tourists through Berlin based on route similarity clustering

Fig. 4.10 presents the map based on route similarity and dynamics, and it yielded similar results as the previous approach. It can again be seen that the main concentration of tourists is in the central area, which presents almost a straight line that connects Reichstag, Brandenburg gate and Postdam Square.

Figure 4.10: The flow of tourists through Berlin based on route similarity and dynamics clustering

It should be noted that the obtained results are similar to those of Andrienko *et al.*, although they used other techniques and images collected from Panoramio [4]. Their results also indicate that a large amount of images is taken in the city center, where the main attractions are, but also that the main routes are following the Unter den Linden street. This suggests that there is a consistent underlying pattern of tourist mobility that can be discovered using different mining methods and data sources.

4.4 Digital epidemiology

4.4.1 HIV spatial distribution

The Demographic and Health Surveys (DHS) program collects and disseminates accurate, nationally representative data on fertility, family planning, maternal and child health, gender, HIV/AIDS, malaria, and nutrition. To determine the health status of a population, DHS periodically organizes surveys to gather relevant data for the observed country. In our study the results of the DHS are collected in Ivory Coast for the period between 2008 and 2012 [132]. Based on measurements the DHS provides estimates on HIV at sub–national level, but with low spatial resolution determined by 10 administrative regions (Fig. 4.11 (a)). Estimates of the HIV prevalence range from 2.2% to 5.1% and reveal the spatial variability of the distribution of HIV–infected across the country.

Due to initiatives to further unearth spatial heterogeneity of HIV [135], new methods emerged aiming to provide HIV estimates at a finer resolution. One approach, that employs kernel estimation on spatial DHS measurements with additional adjustment to UNAIDS data, made available estimates [74] for 50 regions of Ivory Coast (see Table 4.3). This table provides HIV estimates at region level from national population–based surveys. Using the DHS dataset we also estimated HIV prevalence rate by administrative region (see Table 4.4). After redistributing disease frequencies across 50 regions, HIV prevalence map (Fig. 4.11 (b)) uncovers higher spatial variability (from 0.6% to 5.7%) of disease distribution. We notice hot spots of epidemics–regions severely hit by HIV. The map also enables us to explore links between the connectivity and mobility patterns derived from D4D data and HIV prevalence at a better spatial resolution. Although quality of HIV estimates (imposed by the DHS sampling measurement) at region level varies from high and moderate to uncertain, which is the highest spatial resolution currently available for studying the HIV epidemic in Ivory Coast. We used open source QGIS software[122] to create maps from (a) DHS data [132] and (b) UNAIDS estimates [73].

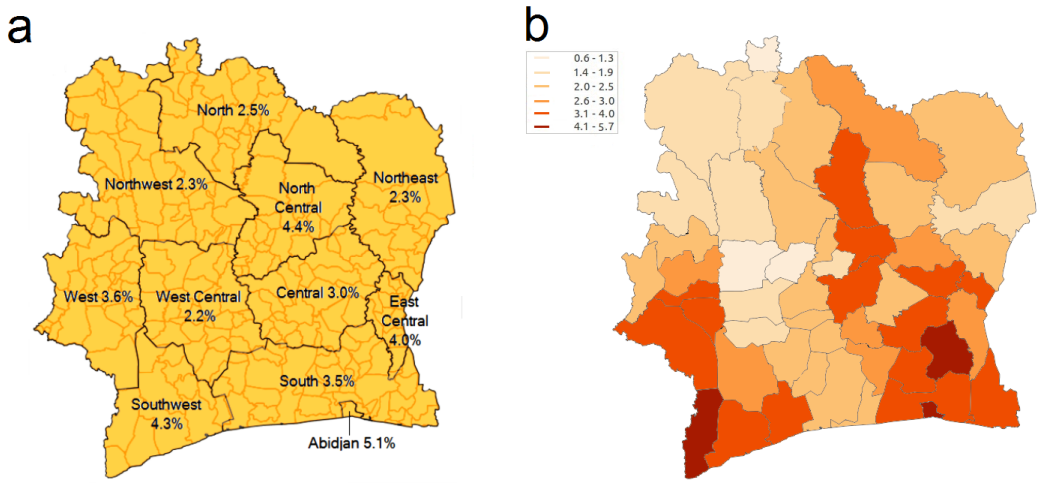

Figure 4.11: (a) HIV prevalence rate by administrative regions (b) HIV prevalence rate by regions for 15–49 year–olds; estimated values range between 0.6% and 5.7%.

Table 4.3: HIV prevalence estimates by administrative regions

Regions	**2008**	**2010–2012**
Centre–East (Moyen–Comoé)	5.8	9.17
South (Lagunes, Agnéby, Sud Comoé, Sud Bandama)	5.5	8.91
Centre (Lacs, N'zi Comoé, part of Vallée du Bandama)	4.8	8.56
South–West (Bas Sassandra)	4.2	6.94
Centre–West (Fromager, Haut Sassandra, Marahoué)	3.7	4.85
Centre–North (part of Vallée du Bandama)	3.6	6.29
West (Dix–Huit Montagnes, Moyen Cavally)	3.5	4.76
North–East (Zanzan)	3.3	4.56
North (Savanes)	3.2	5.46
North–West (Bafing, Denguélé, Worodougou)	1.7	4

Table 4.4: HIV prevalence rate by administrative regions

Department ID	Department Name	HIV prevalence	Quality
1	Abengourou	2.70	high
2	Abidjan	3.60	high
3	Aboisso	3.20	uncertain
4	Adzopé	4.70	uncertain
5	Agboville	3.10	uncertain
6	Agnibilékrou	3.20	moderately high
7	Bangolo	3.30	uncertain
8	Béoumi	2.50	uncertain
9	Biankouma	2.00	uncertain
10	Bondoukou	1.70	moderately high
11	Bongouanou	3.30	uncertain
12	Bouaflé	2.50	uncertain
13	Bouaké	3.10	high
14	Bouna	2.00	uncertain
15	Boundiali	1.40	uncertain
16	Dabakala	2.50	uncertain
17	Daloa	1.90	uncertain
18	Danané	2.10	uncertain
19	Daoukro	3.80	uncertain
20	Dimbokro	2.30	uncertain
21	Divo	2.00	uncertain
22	Duékoué	3.80	uncertain
23	Ferkessédougou	2.90	uncertain
24	Gagnoa	2.00	uncertain
25	Grand–Lahou	2.20	uncertain
26	Guiglo	3.20	moderately high
27	Issia	1.90	uncertain
28	Katiola	4.00	uncertain
29	Korhogo	2.30	high
30	Lakota	2.00	uncertain
31	Man	2.80	moderately high
32	Mankono	2.10	moderately high
33	Mbahiakro	3.00	uncertain
34	Odienné	1.70	moderately high
35	Oumé	2.50	uncertain
36	Sakassou	1.70	uncertain
37	San–Pédro	3.30	moderately high
38	Sassandra	3.50	uncertain
39	Séguéla	1.80	uncertain
40	Sinfra	2.50	uncertain
41	Soubré	2.80	moderately high
42	Tabou	5.70	uncertain
43	Tanda	2.10	moderately high
44	Tengréla	0.60	uncertain
45	Tiassalé	2.60	uncertain
46	Touba	1.60	moderately high
47	Toumodi	2.60	uncertain
48	Vavoua	1.30	uncertain
49	Yamoussoukro	3.10	moderately high
50	Zuénoula	1.10	uncertain

4.4.2 Spatial distribution and mobility

Three types of experiments were performed, using one of the regression models for each set. For SET1 and SET3 we used an Elastic Net predictive model and for SET2 we used a ridge regression. Before learning the regression model we normalized the feature–space by dividing each feature with its mean value. Parameters of models were estimated with leave–one–out (LOO) cross–validation. The results of the selected models are expressed through a correlation coefficient and a Root Mean Square Error (RMSE) in Table 4.5.

Table 4.5: Correlation coefficient and RMSE for models

SET	Correlation Coefficient	RMSE
1	0.96	0.59
2	0.55	1.67
3	0.71	1.46

The obtained models have different predictive powers. The model that learned on SET1 performed the best. Most likely due to the high number of very detailed features that we extracted. The other reason could be the fact that SET1 encompasses communications of 5 million people during 6 months. This was enough to detect regional patterns. Extracting the features from SET2 and SET3 is harder since complex dynamics of human movements are involved. Furthermore, the data sets covered fewer people and also shorter periods of time in the case of SET2.

In Table 4.6 we report on a few interesting features which we observed were stable in all validated methods–they did not change the sign and they have a high coefficient.

Table 4.6: Features coefficient weights for 3 data sets

SETS	Description	Weight	
1	Duration–w0h5	Avg. duration of calls during weekdays (05–06h)	0.88
1	NbVoice–w1h2	Avg. number of calls during weekend (02–03h)	0.84
1	Duration–w1h22	Avg. duration of calls during weekend (22–23h)	-1.00
2	WeD	Time spent in each region during weekend days	0.68
2	NHo_We	Time spent in each region during weekend night hours (00–05h)	0.34
2	WoH	Time spent in each region during working hours (08–18h)	0.19
3	RadiusNight	Maximum radius from home location during the night hours	-1.00
3	Gyration	Standard deviation from the average location of user	-0.77
3	InMigration	Counted movement of non–resident users from their home regions to observed region	0.46

Features that stand out from SET1 are those related to the communication activities during night hours. The late night calls and their longer duration are positively associated with epidemic rate and they can be seen as indicators of risk behavior. On the other side, duration of the early night calls is negatively associated. After the application of ridge regression prediction on features extracted from SET2, the results showed that each of them have a different informative weight. The most informative features were time spent in each region during weekend days, and time spent in each region during working hours. Values across the regions indicate that higher activity

in the sense of migration are related to the regions with higher risk. The values of features across all regions showed also that:

- Migration of people is higher during the working hours, as well as one hour before and after work which can be explained with performing daily business duties and travel to and from the work place.
- Migration of people is higher during the weekend due to the lack of specific contents in their own environment (malls, cinemas, sport contents etc.). Therefore people are forced to travel to larger centers nearby in order to fulfill some of their secondary obligations.
- Migration of people is higher during the weekend night hours (00–05h) which is the most important indicator. In this period people usually go to larger centers looking for fun and entertainment, which significantly increases the risk of infection and transmission of infection.

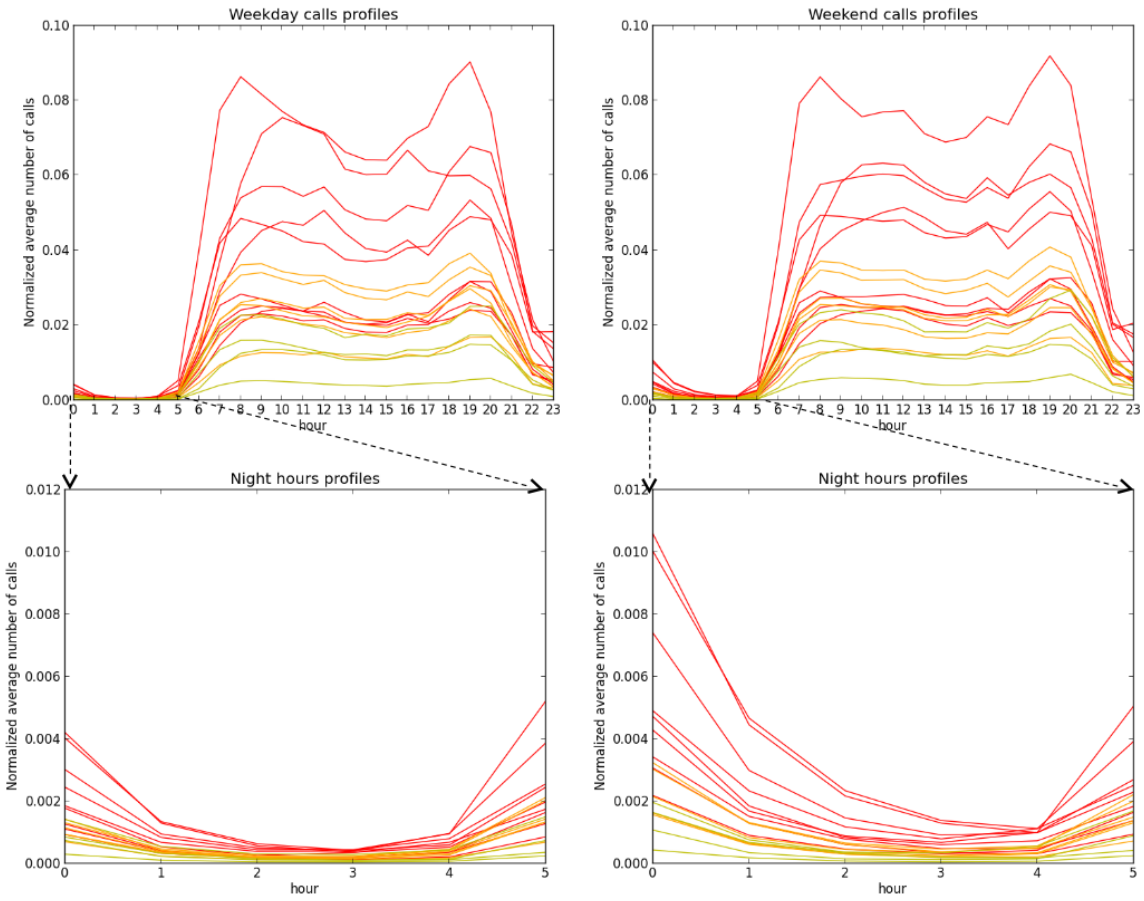

Figure 4.12: Normalized average number of calls for SET1 (left up and bottom) and SET3 (right up and bottom)

In SET3 the coefficient for regional average gyration and radius during night hours have a negative sign. This is not surprising: studies have already shown that in denser urban areas there are higher expectations of shorter movements [89]. The third feature listed in the table indicates that the more migration in the region, the higher the epidemic rate. This is something that we would intuitively expect and here the trained model learned it from our data.

4.4.3 Communication and mobility patterns

Social interactions and mobility mediate the spread of infectious diseases [124, 99, 12]. When examined in a spatio–temporal context, they can uncover how disease propagates and finally explain variability in the prevalence distribution. To better understand spatial epidemiology of HIV in Ivory Coast, we analyzed collective connections at the level of regions. We estimated pairwise connections among sub–prefectures by measuring communication and mobility flows. To accomplish that we explored "antenna–to–antenna data" (D4D SET1) and "long term individual trajectories" (D4D SET3) [17]. SET1 provided us with insight into the communication flow between each pair of antennas on an hourly basis. The strength of communication flow is expressed through the number of calls. We assigned each antenna to the corresponding region and then aggregated the number of calls at the region level over a 5–month observation period. SET3 shed light on people's mobility through geographic locations of users performing actions on the phones (calls and sending messages). Since records in SET3 contain user ids, locations at the sub–prefecture resolution and timestamps of phone activity we can estimate users' home locations. Based on the most frequent location we assign each user its home region. Then we counted the changes from their home to other locations through the entire 5–month period and aggregated users' movements at the region level.

In the obtained pairwise communication and mobility matrices, we identified strong ties for each region that represent links to other regions with strength higher than average (see section **3** subsection **3.4**). Before searching for strong ties, we normalized matrices with corresponding population sizes. SET1 encompasses 5 million users. We distributed them into regions according to population frequencies from Afripop data [80] and used the obtained populations per region to normalize communication flows. To normalize migration flows we estimated users' home locations to calculate population size of the originating region. Overall flow between two regions was then quantified as a sum of normalized flows in both directions. This enables us to identify truly strong links that are not biased by the population sizes.

The strong ties discovered in communication flows are presented in Fig. 4.13 (a). The visualization further emphasizes the strongest links and communication hubs. The hubs correspond to HIV hot spots and we notice that larger hubs have higher prevalence rates. Additionally, we visualized nighttime communication, constrained by the 01–05h time interval, and obtained a connectivity graph with a similar structure (Fig. 4.13 (b)). The links correspond to the relative (divided by the maximum value in the set of detected strong ties) rather than absolute flow. On both graphs

we notice how regions in the northern part of the country have weaker links and this may explain why they have smaller prevalence. We used open source QGIS software [122] to create maps.

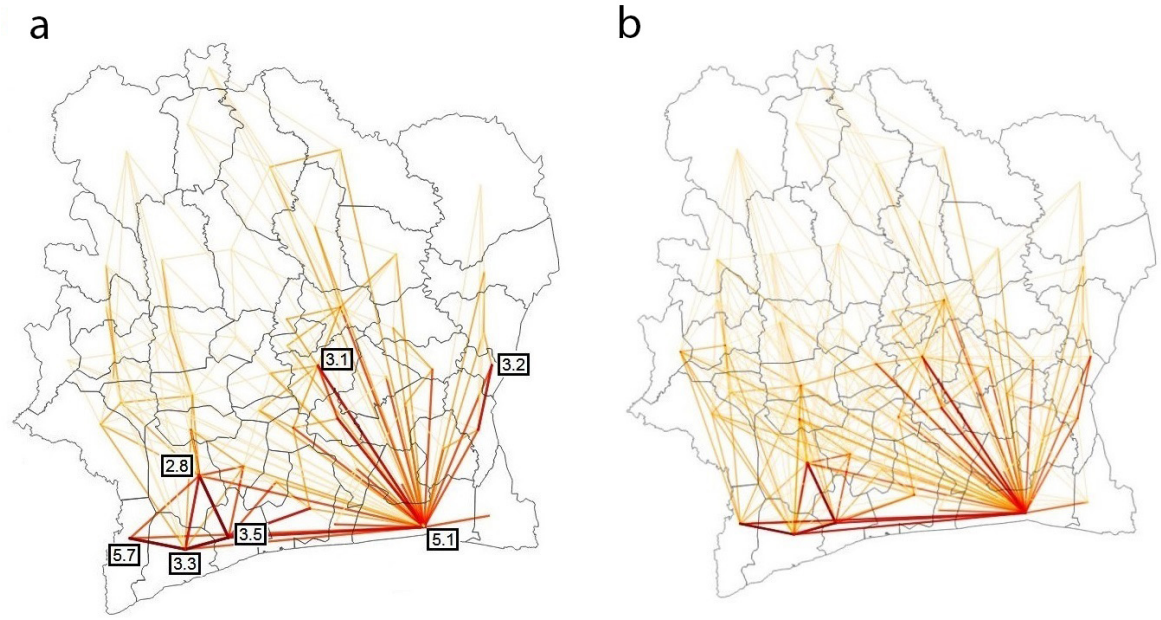

Figure 4.13: Strong connectivity ties for (a) overall communication (b) nighttime communication. The hubs are labeled with the corresponding HIV prevalence rate shown in Fig. 4.11 (b). Link width and color, ranging from yellow to red, are proportional to the strength of communication flow.

The strong ties discovered in mobility flows (Fig. 4.14 (a)) have an obvious localized character. They connect geographically close regions, but on a global scale, we can observe strong pathways. One connects two large hubs-largest city Abidjan (5.1% prevalence rate) and the capital city Yamoussoukro (3.1% prevalence rate). From the center of the country we notice strong pathways to the Western region (3.6% prevalence rate, Fig. 4.11 (a)) and North–central region (4.0% prevalence rate, Fig. 4.11 (a)). East–Central region, with a prevalence rate of 4.0%, is strongly connected to Abidjan. The map of mobility flows revealed to us the pathways that connect regions with higher prevalence. In addition to observing the mobility of users, we explored longer–term mobility. We measured how long users stay at their destinations and in migration analysis counted only those where users stayed for more than 3 days. The strong ties discovered in the longer–term mobility flows are presented in Fig. 4.14 (b). The obtained connectivity graph reveals how longer–term migrations link sub–prefectures at larger distances. Interestingly, Abidjan emerged as a huge hub for those migrations. In this light we can denote this city with the largest prevalence rate and high connectivity as a driver for the epidemic in Ivory Coast. As such, Abidjan needs careful long–term monitoring of mobility flows, in order to prioritize interven-

tions and control further spread of HIV. Again, we used open source QGIS software [122] to create maps.

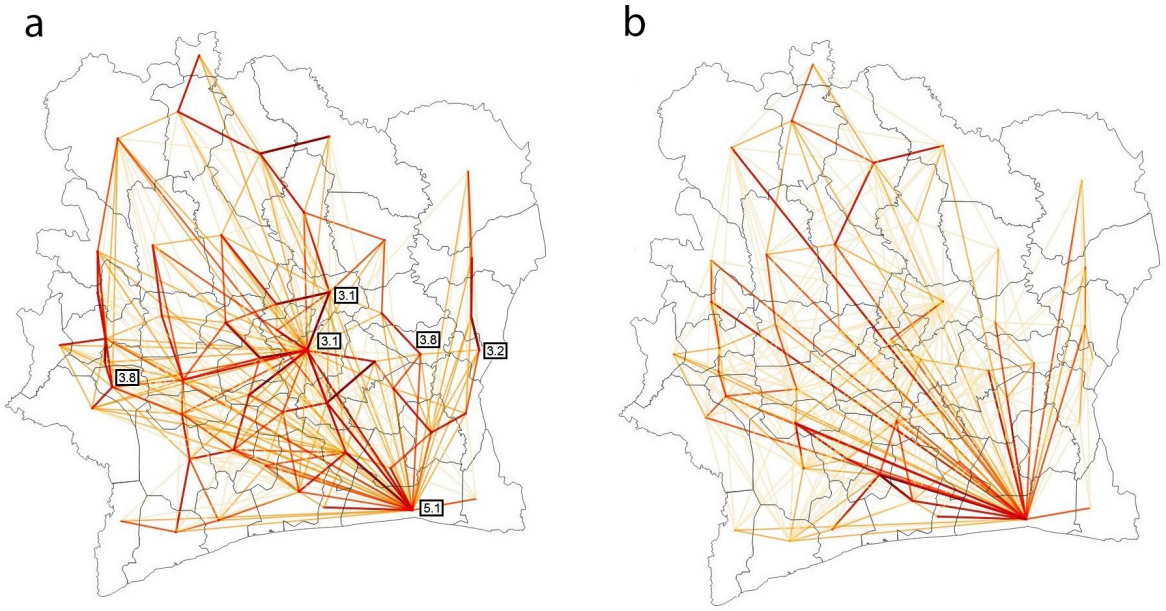

Figure 4.14: Strong mobility ties discovered through summarizing (a) all mobilities (b) mobilities with 3 days or longer spent at the destination. The hubs are labeled with the corresponding HIV prevalence rates shown in Fig. 4.11 (b). The link width and color, ranging from yellow to red, are proportional to the strength of mobility flow.

4.4.4 Features for learning and analysis

For each region of the Ivory Coast, numerous features were extracted from raw data [17] with the aim to quantify the behavioral and mobility patterns potentially relevant to the measured HIV prevalence rate. Overall, we extracted 224 different features and grouped them into 4 categories: connectivity, spatial, migration and activity (mobile phone use).

The connectivity features were obtained from SET1. The communication flow is expressed in SET1 through the number of calls and their duration. Using information on originating and terminating antenna, we summed for each region its inner, originating, terminating and overall communication. The overall communication was further analyzed through the type of day and the time of day constraints. We considered two types of days: weekdays and weekends. For the time of day we used 1–hour time slots (00–01h, 01–02h, ..., 23–24h) and 8–hour slots (00–08h, 08–16h, 16–24h). Features related to the number of calls represent a cumulative sum for the five months observation period. We further normalized them by corresponding regions' population sizes, estimated from Afripop data [80] and rescaled to fit 5 million of users. Features related to the duration of calls represent their average

values. Overall 120 connectivity features in different time slots and type of days were extracted. Half of them correspond to the number of calls and the other half to the average duration.

Spatial, migration and activity features were obtained from SET3. To craft spatial features we explored the positions and the distribution of locations visited by users. We measured the radius of gyration, area and perimeter of convex hull and diameter [55, 139, 35]. For measuring those features we used all locations. We considered different subsets of locations: recorded in night hours, during weekdays and weekends, weekdays' and weekends' nights. Additionally, we calculated the total distance traveled by each user. In total we created 25 spatial features that represent the 95th percentile of values across users matched to regions based on their home location. We first considered averaged instead of the 95th percentile of values for users in the corresponding regions, but for predictive models better results are achieved when spatial features capture only the top five percent of users; i.e. the patterns of users that cover larger space through their mobility have higher predictive power on the epidemic of HIV. For migration features we tracked changes in locations. Every time a user changed location we added one migration link from his home to the observed location.

We summarized all movements into a pairwise migration matrix by iterating this procedure for all users. Besides quantifying all movements, we also identified those where users were away from home for more than defined number of days (1, 2, ..., 10) to explore longer–term migrations. Features are further divided according to the direction of "in–or–out" migration. Their total number is 22. Activity features were extracted in a way similar to what was done with connectivity features, but in SET3 we cannot distinguish the direction of communication ("in–or–out"), nor do we have the duration of communication. We refer to those features simply as activity, since they can count only when and where users were active. As with connectivity features, we considered two types of days (weekdays and weekends) and with respect to the time of day we observed 1–hour time slots, 8–hour time slots and whole day intervals. The total number of activity features extracted is 57.

All features capture the cumulative effect of human connectivity or mobility observed over a 5–month period. We focused on this long term perspective in extracting features to better understand the spatial distribution of HIV prevalence.

4.4.5 Predictive models

HIV prevalence rates across the regions of Ivory Coast range from 0.6% to 5.7%. Each of the 50 regions is represented with extracted feature vectors and the corresponding prevalence rate. Using these features we built regression models and evaluated their performance in predicting a region's prevalence rates. Before learning the models, we normalized features by dividing each feature with its mean value. We experimented with two regression methods: Ridge [42], and SVR [56] on four groups of features. Both regression methods were wrapped up with an RFE [57] method to select a smaller subset of features. Additionally, we considered one ensemble with

an approach–stacked regression [23] through which we fused 4 heterogeneous feature sets. Predicting disease levels needs careful evaluation [18] to ensure a situation where models on randomly generated data work comparatively as well as those created on possibly meaningful data. Therefore, we created random datasets by randomly permuting values for each feature. To estimate the predictive capacity of the model, we measured the prediction errors and correlations between the predicted and actual values.

The experiments were divided into two parts: the first stage focused on the 15 regions with high and moderate estimates of HIV prevalence, while in the second we used the data for all 50 regions. In Tables 4.7 and 4.8, we report on correlation coefficients (ρ) and relative root mean square errors ($RRMSE$) produced by models during LOO cross–validation for the two experimental setups (15 and 50 regions). LOO evaluation enabled us to select the best model among those we built. On the subsample of 15 regions, the models built with SVR and RFE performed best. In the best models RFE reduced the initial set of features to a subset of 60, 6, 3, 4 for connectivity, spatial, migration and activity features respectively.

Table 4.7: Evaluation of predictive models on high and moderate HIV estimates–correlation coefficient (RRMSE): ρ ($RRMSE$)

	Predictive models			
Features	**Ridge**	**Ridge+RFE**	**SVR**	**SVR+RFE**
(SET1)	0.624 (0.331)	0.626 (0.331)	0.661 (0.306)	**0.669 (0.301)**
(SET3)	0.639 (0.434)	0.703 (0.376)	0.544 (0.351)	**0.753 (0.294)**
(SET3)	0.585 (0.369)	0.585 (0.369)	0.678 (0.307)	**0.691 (0.288)**
(SET3)	0.618 (0.339)	0.645 (0.325)	0.633 (0.316)	**0.664 (0.302)**
Ensemble	0.610 (0.327)	0.601(0.327)	0.659 (0.305)	**0.710 (0.287)**
Best Random	-0.231(0.511)	-0.066 (0.480)	-0.065 (0.479)	0.070 (0.441)

Table 4.8: Evaluation of predictive models on all HIV estimates–correlation coefficient (RRMSE): ρ ($RRMSE$)

	Predictive models			
Features	**Ridge**	**Ridge+RFE**	**SVR**	**SVR+RFE**
(SET1)	0.467 (0.556)	0.481 (0.546)	0.501 (0.516)	**0.508 (0.514))**
(SET3)	0.363 (0.540)	**0.431 (0.523)**	0.310 (0.552)	0.336 (0.545)
(SET3)	0.269 (0.630)	0.315 (0.613)	0.291 (0.637)	**0.375 (0.599)**
(SET3)	0.511 (0.542)	0.542 (0.535)	0.522 (0.537)	**0.627 (0.509)**
Ensemble	0.500 (0.527)	**0.543** (0.519)	0.535 (0.515)	0.518 **(0.514)**
Best Random	0.020 (0.760)	0.202 (0.657)	0.139 (0.630)	0.038 (0.607)

The selected features are highlighted in Tables 4.9 and 4.10 and includes all the features and their descriptions. The SVR models surpassed Ridge and reducing the size of the feature set with RFE improved performance of both, but the

SVR method benefited more from the RFE procedure than Ridge. The highest correlation coefficient (0.753) between the predicted and actual values is achieved with the SVR on a reduced set of 6 most relevant spatial features. The lowest error of 0.287 is reached by combining regression models trained on different sets of features. Through the linear combination of the four models, the ensemble approach predicts HIV prevalence values that are well correlated with the actual ($\rho = 0.710$) values. All models built on the real features outperformed their random counterparts.

The second part of the experiments evaluated the proposed methods and extracted features on the full set of 50 regions, including those with uncertain estimates on HIV. Table 4.8 reports on obtained results. As expected the performance declined. Predictions are moderately correlated with actual values. The best result $\rho = 0.627$, $RRMSE = 0.509$ is achieved with the SVR model on a reduced subset of activity features. The ensemble approach that combines four SVR+RFE models results in $\rho = 0.518$ and $RRMSE = 0.514$. The models created on randomly permuted features predict HIV with higher errors and without correlation with actual values and underperform those built on real features.

Table 4.9: Features descriptions for SET1

Feature Name	**Description**
Connectivity–Inner*	Sum of inner calls divided by population
Connectivity–Orig*	Sum of calls originating from region divided by population
Connectivity–Term*	Sum of calls terminating in region divided by population
Connectivity–All*	Sum of all calls in region divided by population
Connectivity–Weekday*	Sum of all calls in region during weekdays divided by population
Connectivity–Weekend	Sum of all calls in region during weekends divided by population
Connectivity–Weekday-00-08h	Sum of all calls during weekdays (00-08h) divided by population
Connectivity–Weekday-08-16h*	for (08-16h)
Connectivity–Weekday-16-24h*	for (16-24h)
Connectivity–Weekend-00-08h*	Sum of all calls in region during weekends (00-08h) divided by population
Connectivity–Weekend-08-16h	for (08-16h)
Connectivity–Weekend-16-24h	for (16-24h)
Connectivity–Weekday-00h	Sum of all calls in region during weekdays (00-01h) divided by population
Connectivity-Weekday-01h*	for (01-02h)
Connectivity-Weekday-02h*	for (02-03h)
Connectivity-Weekday-03h*	for (03-04h)

Feature Name	Description
Connectivity-Weekday-04h	for (04-05h)
Connectivity-Weekday-05h	for (05-06h)
Connectivity-Weekday-06h*	for (06-07h)
Connectivity-Weekday-07h	for (07-08h)
Connectivity-Weekday-08h*	for (08-09h)
Connectivity-Weekday-09h*	for (09-10h)
Connectivity-Weekday-10h*	for (10-11h)
Connectivity-Weekday-11h*	for (11-12h)
Connectivity-Weekday-12h*	for (12-13h)
Connectivity-Weekday-13h*	for (13-14h)
Connectivity-Weekday-14h*	for (14-15h)
Connectivity-Weekday-15h*	for (15-16h)
Connectivity-Weekday-16h*	for (16-17h)
Connectivity-Weekday-17h*	for (17-18h)
Connectivity-Weekday-18h	for (18-19h)
Connectivity-Weekday-19h	for (19-20h)
Connectivity-Weekday-20h	for (20-21h)
Connectivity-Weekday-21h	for (21-22h)
Connectivity-Weekday-22h	for (22-23h)
Connectivity-Weekday-23h*	for (23-00h)
Connectivity-Weekend-00h*	Sum of all calls in region during weekends (00-01h) divided by population
Connectivity-Weekend-01h*	for (01-02h)
Connectivity-Weekend-02h*	for (02-03h)
Connectivity-Weekend-03h*	for (03-04h)
Connectivity-Weekend-04h	for (04-05h)
Connectivity-Weekend-05h	for (05-06h)
Connectivity-Weekend-06h*	for (06-07h)
Connectivity-Weekend-07h*	for (07-08h)
Connectivity-Weekend-08h	for (08-09h)
Connectivity-Weekend-09h*	for (09-10h)
Connectivity-Weekend-10h*	for (10-11h)
Connectivity-Weekend-11h*	for (11-12h)
Connectivity-Weekend-12h	for (12-13h)
Connectivity-Weekend-13h	for(13-14h)
Connectivity-Weekend-14h	for (14-15h)
Connectivity-Weekend-15h	for (15-16h)
Connectivity-Weekend-16h	for (16-17h)
Connectivity-Weekend-17h*	for (17-18h)
Connectivity-Weekend-18h	for (18-19h)
Connectivity-Weekend-19h	for (19-20h)
Connectivity-Weekend-20h	for (20-21h)
Connectivity-Weekend-21h	for (21-22h)

Feature Name	Description
Connectivity-Weekend-22h	for (22-23h)
Connectivity-Weekend-23h	for (23-00h)
Average-Inner-Call-Duration	Sum of durations of inner calls in region divided by number of calls
Average-Orig-Call-Duration	Sum of durations of originating calls in region divided by number of calls
Average-Term-Call-Duration	Sum of durations of originating calls in region divided by number of calls
Average-Call-Duration	Sum of durations of all calls in region divided by number of calls
Average-Call-Duration-Weekday	Sum of durations of all calls in region during weekdays divided by number of calls
Average-Call-Duration-Weekend*	Sum of durations of all calls in region during weekends divided by number of calls
Average-Call-Duration-Weekday-00-08h*	Sum of durations of all calls in region during weekdays (00-08h) divided by number of calls
Average-Call-Duration-Weekday-08-16h	for (08-16h)
Average-Call-Duration-Weekday-16-24h	for (16-24h)
Average-Call-Duration-Weekend-00-08h*	Sum of durations of all calls in region during weekends (00-08h) divided by number of calls
Average-Call-Duration-Weekend-08-016h*	for (08-16h)
Average-Call-Duration-Weekend-16-24h	for (16-24h)
Average-Call-Duration-Weekday-00h	Sum of durations of all calls in region during weekdays (00-01h) divided by number of calls
Average-Call-Duration-Weekday-01h*	for (01-02h)
Average-Call-Duration-Weekday-02h*	for (02-03h)
Average-Call-Duration-Weekday-03h*	for (03-04h)
Average-Call-Duration-Weekday-04h	for (04-05h)
Average-Call-Duration-Weekday-05h*	for (05-06h)
Average-Call-Duration-Weekday-06h*	for (06-07h)
Average-Call-Duration-Weekday-07h*	for (07-08h)
Average-Call-Duration-Weekday-08h	for (08-09h)

Feature Name	Description
Average-Call-Duration-Weekday-09h	for (09-10h)
Average-Call-Duration-Weekday-10h	for (10-11h)
Average-Call-Duration-Weekday-11h	for (11-12h)
Average-Call-Duration-Weekday-12h*	for (12-13h)
Average-Call-Duration-Weekday-13h*	for (13-14h)
Average-Call-Duration-Weekday-14h	for (14-15h)
Average-Call-Duration-Weekday-15h	for (15-16h)
Average-Call-Duration-Weekday-16h	for (16-17h)
Average-Call-Duration-Weekday-17h	for (17-18h)
Average-Call-Duration-Weekday-18h	for (18-19h)
Average-Call-Duration-Weekday-19h	for (19-20h)
Average-Call-Duration-Weekday-20h	for (20-21h)
Average-Call-Duration-Weekday-21h*	for (21-22h)
Average-Call-Duration-Weekday-22h	for (22-23h)
Average-Call-Duration-Weekday-23h*	for (23-00h)
Average-Call-Duration-Weekend-00h	Sum of durations of all calls in region during weekends (00-01h) divided by number of calls
Average-Call-Duration-Weekend-01h*	for (01-02h)
Average-Call-Duration-Weekend-02h*	for (02-03h)
Average-Call-Duration-Weekend-03h*	for (03-04h)
Average-Call-Duration-Weekend-04h	for (04-05h)
Average-Call-Duration-Weekend-05h*	for (05-06h)
Average-Call-Duration-Weekend-06h*	for (06-07h)

Feature Name	Description
Average-Call-Duration-Weekend-07h*	for (07-08h)
Average-Call-Duration-Weekend-08h*	for (08-09h)
Average-Call-Duration-Weekend-09h	for (09-10h)
Average-Call-Duration-Weekend-10h	for (10-11h)
Average-Call-Duration-Weekend-11h	for (11-12h)
Average-Call-Duration-Weekend-12h*	for (12-13h)
Average-Call-Duration-Weekend-13h*	for (13-14h)
Average-Call-Duration-Weekend-14h*	for (14-15h)
Average-Call-Duration-Weekend-15h*	for (15-16h)
Average-Call-Duration-Weekend-16h*	for (16-17h)
Average-Call-Duration-Weekend-17h	for (17-18h)
Average-Call-Duration-Weekend-18h	for (18-19h)
Average-Call-Duration-Weekend-19h	for (19-20h)
Average-Call-Duration-Weekend-20h	for (20-21h)
Average-Call-Duration-Weekend-21h	for (21-22h)
Average-Call-Duration-Weekend-22h	for (22-23h)
Average-Call-Duration-Weekend-23h*	for (23-00h)

Table 4.10: Features descriptions for SET3

Feature Name	**Description**
Gyration-All	95 percentile of the distribution of user's radius of gyration for all visited locations
Gyration-Night*	95 percentile of the distribution of user's radius of gyration for locations visited during nigth hours 22h-05h
Gyration-Weekday	95 percentile of the distribution of user's radius of gyration for locations visited during weekdays
Gyration-Weekday-Night*	95 percentile of the distribution of user's radius of gyration for locations visited during weekdays nigth hours 22h-05h
Gyration-Weekend	95 percentile of the distribution of user's radius of gyration for locations visited during weekend hours
Gyration-Weekend-Night*	95 percentile of the distribution of user's radius of gyration for locations visited during weekend nigth hours 22h-05h
Diameter-All	95 percentile of the distribution of user's diameter of convex hull for all visited locations
Diameter-Night	95 percentile of the distribution of user's diameter of convex hull for locations visited during nigth hours 22h-05h
Diameter-Weekday	95 percentile of the distribution of user's diameter of convex hull for locations visited during weekdays
Diameter-Weekday-Night	95 percentile of the distribution of user's diameter of convex hull for locations visited during weekdays nigth hours 22h-05h
Diameter-Weekend	95 percentile of the distribution of user's diameter of convex hull for locations visited during weekend hours
Diameter-Weekend-Night	95 percentile of the distribution of user's diameter of convex hull for locations visited during weekend night hours 22h-05h
Area-All	95 percentile of the distribution of user's area of convex hull for all visited locations
Area-Night	95 percentile of the distribution of user's area of convex hull for locations visited during night hours 22h-05h
Area-Weekday*	Locations visited during weekday hours
Area-Weekday-Night*	Locations visited during weekday night hours

Feature Name	Description
Area-Weekend*	Locations visited during weekend hours
Area-Weekend-Night	Locations visited during weekend night hours
Perimeter-All	95 percentile of the distribution of user's perimeter of convex hull for all visited locations
Perimeter-Night	95 percentile of the distribution of user's perimeter of convex hull during night hours
Perimeter-Weekday	95 percentile of the distribution of user's perimeter of convex hull during weekday hours
Perimeter-Weekday-Night	95 percentile of the distribution of user's perimeter of convex hull during weekday night hours
Perimeter-Weekend	95 percentile of the distribution of user's perimeter of convex hull during weekend hours
Perimeter-Weekend-Night	95 percentile of the distribution of user's perimeter of convex hull during weekend night hours
Distance-All	95 percentile of the distribution of user's sum of distances derived from sequences of visited locations
Out-Migration-Overall*	Sum of all users' mobilities out of home region divided by population of home region
Out-Migration-1day	Sum of users' mobilities out of home region that last more than 1 day divided by population of home region
Out-Migration-2days	for 2 days
Out-Migration-3days	for 3 days
Out-Migration-4days	for 4 days
Out-Migration-5days	for 5 days
Out-Migration-6days	for 6 days
Out-Migration-7days	for 7 days
Out-Migration-8days	for 8 days
Out-Migration-9days	for 9 days
Out-Migration-10days*	for 10 days
In-Migration-Overall*	Sum of all users' mobilities into observed region from other regions divided by corresponding populations
In-Migration-1day	Sum of all users' mobilities into observed region from other regions that last more than 1 day divided by corresponding populations
In-Migration-2day	for 2 days

Feature Name	Description
In-Migration-3day	for 3 days
In-Migration-4day	for 4 days
In-Migration-5day	for 5 days
In-Migration-6day	for 6 days
In-Migration-7day	for 7 days
In-Migration-8day	for 8 days
In-Migration-9days	for 9 days
In-Migration-10days	for 10 days
Activity-All	Sum of all users' activities (calls/SMS) in region divided by population
Activity-Weekday	Sum of all users' activities (calls/SMS) in region during weekdays divided by population
Activity-Weekend	Sum of all users' activities (calls/SMS) in region during weekends divided by population
Activity-Weekday-00-08h	Sum of all users' activities (calls/SMS) in region during (00-08h) divided by population
Activity-Weekday-08-16h	for (08-16h)
Activity-Weekday-16-24h	for (16-24h)
Activity-Weekend-00-08h	for (00-08h)
Activity-Weekend-08-016h	Sum of all users' activities (calls/SMS) in region during weekends (08-16h) divided by population
Activity-Weekend-16-24h	for (16-24h)
Activity-Weekday-00h*	Sum of all users' activities (calls/SMS) in region during (00-01h) divided by population
Activity-Weekday-01h	for (01-02h)
Activity-Weekday-02h	for (02-03h)
Activity-Weekday-03h	for (03-04h)
Activity-Weekday-04h	for (04-05h)
Activity-Weekday-05h	for (05-06h)
Activity-Weekday-06h	for (06-07h)
Activity-Weekday-07h	for (07-08h)
Activity-Weekday-08h	for (08-09h)
Activity-Weekday-09h	for (09-10h)
Activity-Weekday-10h	for (10-11h)
Activity-Weekday-11h	for (11-12h)
Activity-Weekday-12h	for (12-13h)
Activity-Weekday-13h	for (13-14h)
Activity-Weekday-14h	for (14-15h)
Activity-Weekday-15h	for (15-16h)
Activity-Weekday-16h	for (16-17h)
Activity-Weekday-17h	for (17-18h)
Activity-Weekday-18h	for (18-19h)

Feature Name	Description
Activity-Weekday-19h	for (19-20h)
Activity-Weekday-20h	for (20-21h)
Activity-Weekday-21h	for (21-22h)
Activity-Weekday-22h	for (22-23h)
Activity-Weekday-23h	for (23-24h)
Activity-Weekend-00h*	Sum of all users' activities (calls/SMS) in region during weekends (00-01h) divided by population
Activity-Weekend-01h*	for (01-02h)
Activity-Weekend-02h*	for (02-03h)
Activity-Weekend-03h	for (03-04h)
Activity-Weekend-04h	for (04-05h)
Activity-Weekend-05h	for (05-06h)
Activity-Weekend-06h	for (06-07h)
Activity-Weekend-07h	for (07-08h)
Activity-Weekend-08h	for (08-09h)
Activity-Weekend-09h	for (09-10h)
Activity-Weekend-10h	for (10-11h)
Activity-Weekend-11h	for (11-12h)
Activity-Weekend-12h	for (12-13h)
Activity-Weekend-13h	for (13-14h)
Activity-Weekend-14h	for (14-15h)
Activity-Weekend-15h	for (15-16h)
Activity-Weekend-16h	for (16-17h)
Activity-Weekend-17h	for (17-18h)
Activity-Weekend-18h	for (18-19h)
Activity-Weekend-19h	for (19-20h)
Activity-Weekend-20h	for (20-21h)
Activity-Weekend-21h	for (21-22h)
Activity-Weekend-22h	for (22-23h)
Activity-Weekend-23h	for (23-24h)

Fig. 4.15 reveals results of the features contribution analysis for features ranked from 4^{th} to 6^{th} place in the recursive feature elimination procedure.

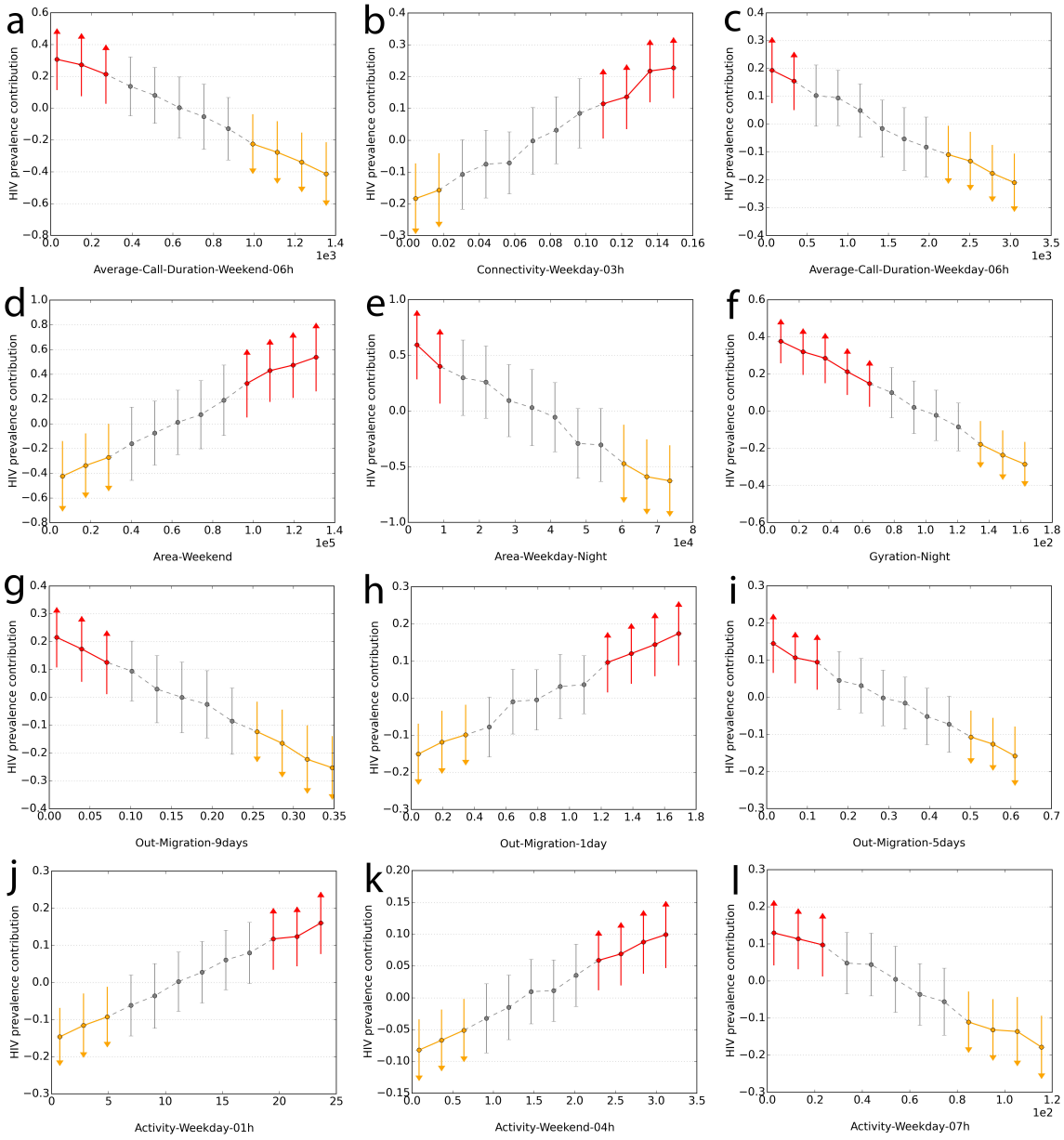

Figure 4.15: Feature contribution graphs for 12 features, ranked from 4^{th} to 6^{th} place for 4 types of features. Points correspond to the mean contribution and error bars correspond to the standard deviation. Red color indicates strong association to higher HIV, and orange to lower HIV prevalence

4.4.6 Feature contribution

Once the regression model is trained, we can use it to estimate the risk of disease in defined spatial units. Furthermore, we can examine what the model learned from the data. Model explanation techniques [119, 120] can unveil black–box predictive models by estimating the contribution of each feature on the whole range of its input values. For example, we can examine how changes in activity features affect the value of the

HIV prevalence rate obtained by the trained model. The outcome is a graph with the contribution as a function of feature values. This model's explanation procedure provides us with the opportunity to identify specific features that impact prevalence rate significantly and to quantify their contribution. The features identified can later be continuously measured and leveraged for changes and creating early warning signs for a possible increase in the HIV prevalence rate.

In the feature contribution analysis we used the best model (SVR+RFE) for each set of features, since the ensemble method is just an additive combination of models built on different sets. In the analysis we used models built on a subsample of regions (15 with a high or moderate HIV estimation) and selected the top 3 features by running the RFE procedure until 3 features remained. Top features have the highest impact on predicting the HIV prevalence. For the selected features ($f_{t,i}$, where t denotes the set of features and i is the index of the feature in that set) we performed a contribution analysis. We calculated the contribution for each feature within the range from its minimum to maximum value in m equally distributed points.

The contribution analysis includes a randomization process to create two instances as input to the regression model. The first instance is a vector where each feature value is sampled at random from the feature set t. The second instance differs in the i^{th} feature, which is not random but takes a particular value from the set of previously defined m values that is currently under contribution analysis. The contribution of the feature is the difference between the outputs of the regression model produced by using the first and the second instances as input. Due to the randomization process, this procedure is repeated for a defined number of iterations. By averaging the results of all iterations we obtained a final value for the contribution. We also report on the standard deviation for a mean contribution that provides information on its stability and quantifies complex interactions among features. We created graphs (Fig. 4.16) for 12 features, top 3 for each of four data sets, sampled in $m = 12$ points with contributions calculated through 100 iterations. Additionally, 12 graphs that correspond to features ranked from 4^{th} to 6^{th} place for each data set are provided in Fig. 4.15. All graphs contain points of a mean contribution and error bars in length of the standard deviation. Red color indicates points with feature values that are associated with increased HIV and orange color indicates feature values that are associated with decreased HIV prevalence rates. The gray part of the graph denotes the range where the standard distribution crosses zero, meaning that the contribution is neither strongly positive nor negative.

The contributions of the three connectivity features are presented in Fig. 4.16 (a), (b) and (c). The top three features represent communication flow expressed as the number of calls per resident of a region during weekends in time slots 01–02h, 02–03h and 03–04h, over a 5–month period. We can notice that the top connection features are related to weekend nighttime communication and all have a positive slope. A similar graph (Fig. 4.15) is obtained for the 5^{th} ranked feature, related to weekday 03–04h communication. According to the model, the regions with higher

nighttime communication have a higher prevalence rate. In further analysis of the contribution graph in Fig. 4.16 (a), values higher than 0.2 can be seen as risky behavior and thus critical for HIV. For example, for the region where this feature has the maximum value, the expectation on HIV prevalence is by 0.3 ($\pm$ 0.15) higher than average. Graphs for features ranked at 4^{th} and 6^{th} place (Fig. 4.15 (a) and (c)) refer to average call duration during early morning (06–07h) and contribute to HIV prevalence in a different manner. Those graphs have a negative slope indicating that, for regions were people have longer talks in early morning, we can expect lower HIV prevalence. We can observe this as a social signature [108], and may hypothesize that longer talks early in the morning could be an indicator of emotionally close relationships and lower–risk behavior.

In the contribution analysis of spatial features, area and gyration stand out with higher impact. Area is measured over weekdays and gyration over weekday and weekend nights. The model suggests that regions where people that, through their overall movement, tend to cover a larger area, have a higher HIV prevalence (Fig. 4.16 (d)). This is confirmed by the 4^{th} ranked feature that measures the area covered over weekends (Fig. 4.15). On the contrary, gyration, a measure of standard deviation from the mean location, negatively impacts HIV (Fig. 4.16 (e),(f)) and also Fig. 4.15). But it is no surprise that small gyration indicates higher HIV, since it was already shown that in denser urban areas there is a higher expectation of shorter movements [89], and those urban areas are often more affected by HIV. Interestingly, when the area covered is tracked only during the night the contribution graph has a negative slope, as is the case with gyration (see Fig. 4.15).

The contributions of overall in and out migration features are presented in Fig. 4.16 (g), (h). Both graphs indicate that larger migration flows are associated with a higher HIV prevalence. We notice the strong impact of incoming migrations for the region where this feature has the maximum value, the expectation of HIV prevalence is by 1.0 ($\pm$ 0.5) higher than average. Among the top three features is the one that quantifies the number of out migrations per resident of the region with the time clause of staying more than 10 days. Its contribution graph at Fig. 4.16 (i), shows a negative impact. Graphs for features ranked from 4^{th} to 6^{th} place (Fig. 4.15) further reveal that outbound migrations, lasting more than one day, have a positive slope, and those lasting more than 5 or 9 days have a negative slope. The contribution analysis of migration features uncovers interesting phenomena. The overall amount of migrations is linked to a higher HIV prevalence and this positive slope remains for migrations up to a few days, but beyond that the slope becomes negative. The slope changes at a time clause of 4 days for outbound migrations and 3 days for inbound migrations. The model learned to suggest that risk comes from shorter retention at the host regions and higher dynamics in migrations, while the longer retentions are associated with lower HIV prevalence.

The contribution of activity features, expressed through the number of calls and SMS per residents of a region, are presented in Fig. 4.16 (j), (k), (l). As with the

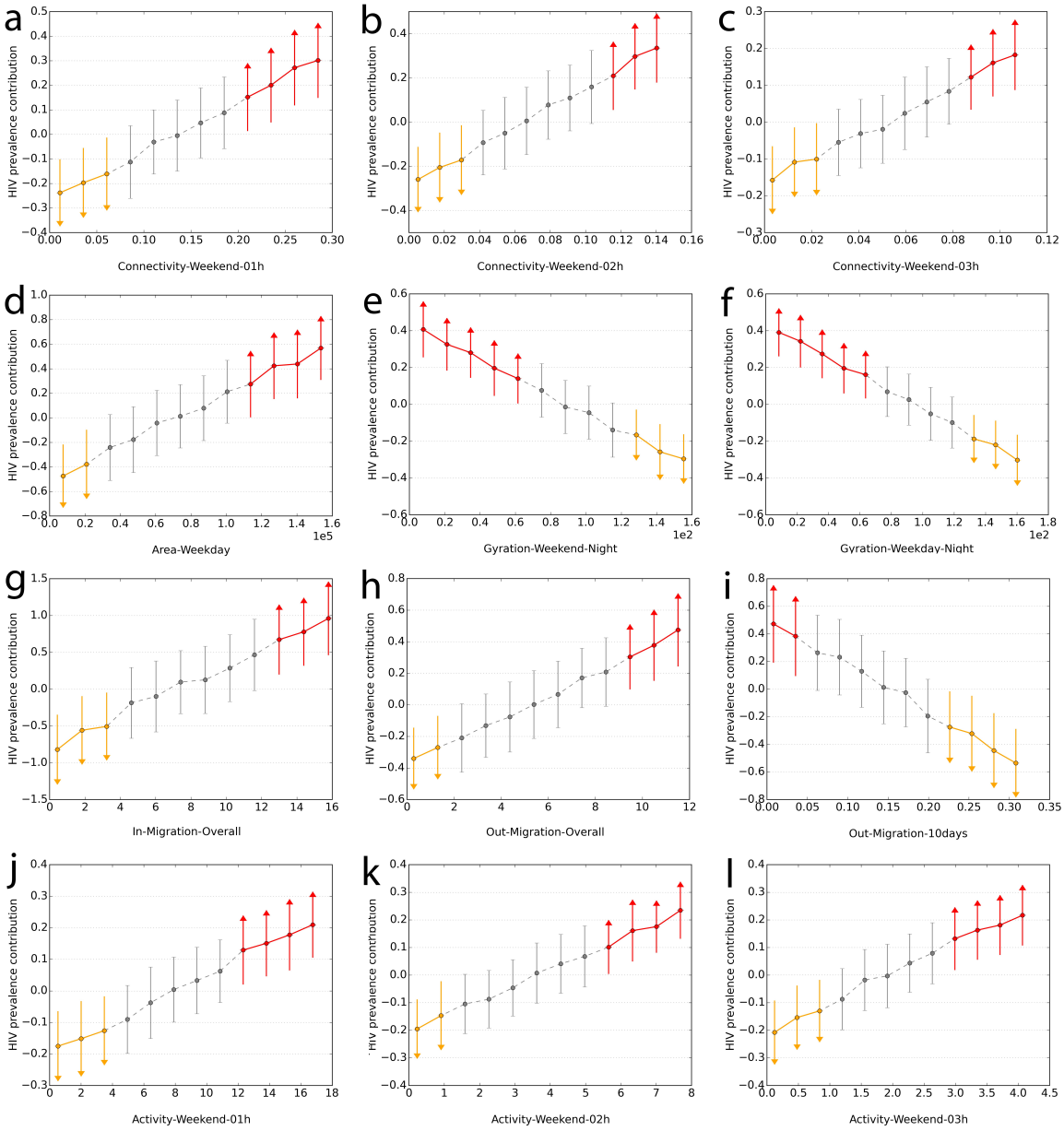

Figure 4.16: Feature contribution graphs for 12 features, the top 3 for 4 types of features. Points correspond to the mean contribution and error bars correspond to the standard deviation. Red color indicates strong association to higher HIV prevalence and orange to lower HIV prevalence

connectivity features, night hours are strongly linked to HIV. Higher activity at this time implies higher prevalence rates. This is also confirmed by the contribution graphs of the 4^{th} and 5^{th} ranked features that encompasses the activity on weekdays between 01–02h and weekends between 04–05h respectfully. On the contrary, the feature ranked 6^{th}, which refers to early morning activity (07–08h), has a negative

slope.

The contribution analysis presented uncovers what the trained models learned from data. All features work in synergy to provide the output–prediction on HIV prevalence. Nevertheless this method helps us to identify the subset of stronger factors. Resulting graphs may serve to create new hypotheses in epidemiology concerning disease distribution and spread, and later for quantifying risk of increase in the prevalence of HIV.

Chapter 5

Conclusions and Future Work

Our work is the first comprehensive study of a common framework for utilizing user–generated data as a general source of information for understanding geographic phenomena. This raises a lot of questions about the models proposed and design decisions, as well as about the particular challenges addressed in this work. The conclusions presented in the rest of this section are organized by the main direction of research, in order to address the specifics of each study.

Using mobile phone data that can unveil patterns of human interactions and mobility is gaining increased attention in epidemiology. In the study in the area of digital epidemiology we place the mobile phone data in the context of a generalized HIV epidemic. Raw data were processed in search for patterns that could explain spatial variation in disease prevalence. We discovered that strong ties and hubs in communication align with HIV hot spots. Strong ties in mobility revealed the pathways that connect regions with higher prevalence. Abidjan, the city most affected by HIV, emerged as the center of migration.

The other part of our study was focused on extracting features related to connectivity and mobility of users at the level of spatial units (regions) that could be predictive for HIV. Several regression methods were used to address that task and results obtained on a subset of high estimates of HIV are promising and can lead to the generation of new hypotheses. The initial set of 224 features was reduced with a recursive feature elimination procedure that allowed us to identify those with the largest impact. It turned out that night connectivity and activity, spatial area covered by users and overall migrations are strongly linked to HIV. Models built on spatial features (gyration, area, perimeter of convex hull, diameter and distance) have a highly predictive power ($\rho = 0.753,\ RRMSE = 0.294$).

A few real–world applications emerged from the obtained results. The first is in the possibility to use machine learning algorithms coupled with feature engineering for predicting disease prevalence. HIV surveys are expensive and difficult to carry out. Rather than increasing the sampling size of a survey at the level of a whole country, only representative parts could be sampled in such a way as to provide high estimates of HIV for selected spatial units. Then a predictive model could

be trained, evaluated and finally applied to the rest of the country to generalize on unobserved spatial units. This would enable faster access to HIV estimates. A second application arises from a recent study [52] on geo–spatial modeling of roll–out plans for anti–retroviral drugs distribution. The study reveals that a utilitarian plan in the allocation of limited resources could prevent more infections than an egalitarian plan, i.e. geographic targeting is better than geographic equity. Underlying optimizations rank spatial units based on the efficiency of interventions and HIV incidence rates. Features and predictions inferred from mobile phone data can be also included into a strategy for geographic resource allocation and thus help with epidemic control. Integrating mobility and connectivity patterns into the decision making process for resource allocation would be beneficial especially in adjusting the plan over time as new resources become available and changes in HIV incidences, mobility and connectivity patterns occur. A third application targets epidemiologists.

The purpose of our study was to draw attention of HIV epidemiologists to a rich source of information such as mobile phone data that are constantly collected by service phone providers. Nowadays their data are becoming available for other African countries and the rest of the world [39]. Our results shed light on HIV epidemiology in Ivory Coast, but our intent was not to provide a final model for HIV prediction. Rather we spent our time on exploring numerous features and different models that will allow epidemiologists to gain insights, make new hypotheses and enrich their studies. A final model should be selected together with epidemiologists by taking into account complexity, relevant features and background knowledge. Further evaluations can be made on other African countries with a generalized epidemiology.

The limitations of our study arise from the spatial and temporal scale of available data. On one side, HIV data are limited by the measurement strategy of DHS, UNAIDS or other relevant entities. The quality and spatial resolution of such data are determined by sampling the design–frequency and distribution of measurements. Variability in HIV prevalence across Ivory Coast is certainly higher than modeled on region–level, but we lacked more precise measurements to better account for this. The time resolution is even more scarce. HIV measurement campaigns are organized only once every few years (for Ivory Coast 2001, 2005, 2008, 2012). Our findings linked aggregated behavioral patterns to HIV prevalence rates, but discovered correlations do not imply causation. To explore causation we would need more estimates on changes of HIV prevalence over time. This could soon be overcome by a new device that easily connects to a smartphone [72]. The device performs an ELISA test and discovers disease markers from a tiny drop of blood in just 15 minutes.

This approach has a high acceptance rate among the population and will enable large–scale screening. On the other side, mobile phone's data spatial resolution is restricted by a carrier's antennas distribution and time resolution is conditioned by users' phone activity (calls or SMS). The major constraint for using mobile phone data are the privacy concerns [38]. Besides mandatory user anonymization,

mobile phone data are further spatially obfuscated and/or aggregated over time, or a part of information is removed. For example, antennas are aggregated at the level of larger geographical units, time is aggregated at hourly intervals, and communication graphs at the level of users are detached from any spatial information.

Human mobility and digital epidemiology is an area of active research and still leaves space for a lot of interesting work in the future. In terms of digital epidemiology, future work should include a detailed analysis of the spatio–temporal dynamics of human motion in the context of primary and subsidiary habitats [6], where the first denotes frequently visited locations during typical daily activities and the second captures additional travel. In D4D data sources mobile phone datasets are aggregated to one–hour time slots, but with the spatial resolution of 1250 antennas preserved or spatially restricted by 250 sub–prefectures, without the time aggregation. Even with data aggregation, mobile phone data is a richer source of information than the HIV estimates that are available across 50 regions. Only regarding individual communication graphs (D4D SET4), where spatial information is completely removed, we lose any chance to link it with the HIV distribution. Those communication graphs, if geographically determined, would be an immense source of information for uncovering the connectivity at a more detailed scale. If such data becomes available in a privacy–acceptable form, further progress in the domain of modeling the spread of communicable diseases [15] will be enabled. In summary, our study showed how raw real–world data can be used for significant knowledge extraction. We believe our work, which is a first attempt to link mobile phone data and HIV epidemiology, lays a foundation for further research in explaining heterogeneity of HIV and building predictive tools aimed at advancing public–health campaigns and decision making for HIV interventions. Together with other "big data" approaches to HIV epidemiology [142] that rely on Twitter data [144] and social networks [141, 143] our work fits well into the wider initiative of digital epidemiology [105].

While detecting human mobility patterns and working with metadata of over 1 million Flickr images [49] we show that the tracks of Flickr users seem to be governed by the same laws that have previously been observed in studies based on mobile phone data and a high–resolution dataset of wandering albatross flights. While there is heterogeneity within the population, individual users exhibit significant regularity and follow trajectories whose statistics are largely indistinguishable after rescaling with the radius of gyration of a user. These results represent the first step towards an attempt of modeling and understanding human activity patterns on a world–wide scale. Our results indicate that the quality of the data available through the Flickr online data sharing and management system is comparable to mobile phone data and to high–resolution dataset of wandering albatross flights that has been used in similar studies before. Flickr data is readily available and covers most of the world, which the former sources cannot match.

In terms of continent–level mobility [50], we presented a case study where data obtained from the metadata attached to publicly available videos was used to identify

major routes of movement between countries, to analyze when people start their trip and the basic means of (public) transportation. The results indicate the main mode of transportation and companies in Africa correspond to the information provided by other sources. Concerning the amount of publicly available videos, there is a large number in the period between June and July 2010, which indicates their association to the FIFA World Championship that took place in South Africa in 2010. We assume that the number of videos would be smaller if they were not the hosts of this competition. Results presented indicate that YouTube videos can be used to identify some basic patterns in human behavior, as well as analyze the temporal dynamics of their activity.

Future research in the field of continent–level mobility could focus on the analysis of data obtained from Flickr or some other online sharing storage and to compare results with the one presented in [50]. Although it is possible that the Flickr dataset could be bigger, giving a chance to investigate a central part of the African continent. It should be noted that the identified patterns are restricted to people who uploaded videos to YouTube and they most likely do not represent the mobility of average people in Africa, since it is an under–developed continent in terms of information and communication systems.

Appendix A

Produženi apstrakt na srpskom jeziku

Živimo u vremenu u kome primena savremenih informacionih tehnologija dovodi do generisanja ogromne količine podataka o lokaciji ljudi i pokretnih objekata. Ovi podaci predstavljaju resurs koji se može iskoristiti kako bi se postigao mnogo viši stepen razumevanja navika, kretanja i obrazaca ponašanja ljudi u cilju razvijanja modela od značaja za različite naučne oblasti i unapređenje tehničkih rešenja u granama poput urbanizma, marketinga, epidemiologije itd. Primarni izvor podataka o lokaciji i kretanju ljudi danas predstavljaju mobilni telefoni, bilo da se radi o zapisima o aktivnostima korisnika mobilnih telefona (engl. *Caller Data Records (CDRs)*) koje prikupljaju telefonske kompanije za potrebe naplate usluga, ili o direktnim zapisima o lokaciji koje generišu pametni telefoni korišćenjem globalnog pozicionog sistema (engl. *Global Positioning System (GPS)*).

A.1 Predmet i ciljevi istraživanja

Iako imaju veliki potencijal, usled brige o privatnosti i sigurnosti, neobrađeni ("sirovi") podaci o lokaciji i kretanju ljudi iz ovih izvora su retko dostupni naučnicima i inženjerima, osim u vrlo agregiranoj i anonimiziranoj formi. S druge strane, korisnici na dnevnom nivou svojevoljno generišu velike količine multimedijalnog sadržaja (fotografija i video zapisa), čiji se metapodaci često mogu iskoristiti za analizu kretanja korisnika, budući da sadrže geografske informacije (geografsku širinu i dužinu) sa različitim nivoima preciznosti i tačnosti kojima se opisuje gde i kada se korisnik nalazio u trenutku nastanka multimedijalnog sadržaja.

Protekla decenija svedoči velikoj količini istraživanja u oblasti pronalaženja informacija (engl. *Information Retrieval*), istraživanja tekstualnih sadržaja (engl. *Text Mining*), računarske vizije (engl. *Computer Vision*), i geo–informacionih nauka sa ciljem eksploatacije velike količine danas dostupnih podataka za stvaranje inovativnih aplikacija i rešavanje istraživačkih problema. Takve aplikacije uključuju izdvajanje informacija o mestima i događajima [98], predviđanje širenja epidemije [53], predviđanje tačaka interesovanja (engl. *Point of Interest (POI)*) i trajektorija kojima se ljudi kreću [148].

Predmet istraživanja sprovedenih za potrebe ove disertacije je mogućnost primene različitih modela za istraživanje podataka - klasterizacije, grupisanja podataka, identifikovanja značajnih obeležja (engl. *features*), pronalaženja obrazaca ponašanja (engl. *patterns*), definisanja njihovih međusobnih relacija, a sve u cilju boljeg razumevanja obrazaca i ponašanja, navika i kretanja ljudi. Istraživanja sprovedena za potrebe ove teze su motivisana pre svega različitim studijama koje za cilj imaju kombinovanje geo–prostornih informacija sa matematičkim modelima radi izdvajanja značajnih obrazaca ljudskog kretanja.

Cilj istraživanja se sastojao u evaluaciji mogućnosti korišćenja sve veće količine javno dostupnih podataka o lokaciji i kretanju ljudi kako bi se došlo do novih saznanja, razvili modeli ponašanja i kretanja ljudi i primenili za rešavanje praktičnih problema, poput analize atraktivnih turističkih lokacija i zaštite populacije od virusa side. Pristup je zasnovan na primeni tehnika veštačke inteligencije i istraživanja podataka. U tezi je u ove svrhe sprovedena praktična studija na bazi zaštićnih (agregiranih i anonimiziranih) CDR podataka o upotrebi mobilnih telefona i metapodataka geo–refenciranog multimedijalnog sadržaja. Osnovni ciljevi disertacije su obuhvatali razvijanje inovativnih modela kretanja ljudi za različite namene:

1. Otkrivanje lokacija koje korisnici posećuju (detekcija atraktivnih lokacija primenom tehnika klasterizacije i vizualizacije podataka).
2. Otkrivanje karakterističnih trajektorija korisnika (putanja kretanja u posmatranom području) i analizi njihovih navika vezanih za putovanja.
3. Estimacija budućih vrednosti rizika širenja infektivnih bolesti.

Iako u današnje vreme postoji značajan broj studija koje se bave ovim ili sličnim istraživačkim pitanjima, još uvek ne postoji adekvatan model za razumevanje visoko informativne dimenzije geografskog prostora na osnovu podataka koje korisnici generišu. Pojavom i brzim rastom "netradicionalnih" podataka, otvorene su mogućnosti za pronalaženje zanimljivih, korisnih obrazaca u kretanju, navikama i ponašanju korisnika, što je svakako predstavljao izazov za istraživanje ovih i sličnih pitanja u okviru ove disertacije.

A.2 Korišćeni alati

Za prikupljanje podataka o javno dostupnom, korisnički–generisanom multimedijalnom sadržaju je razvijen poseban alat koji na automatizovan način prikuplja metapodatke (podaci o podacima) nekoliko desetina hiljada video zapisa odnosno fotografija kreiranih na određenom geografskom području koji su dostupni na internet servisima YouTube[1] i Flickr[2].

[1] https://www.youtube.com/
[2] https://www.flickr.com/

Za analizu ovako prikupljenih podataka su korišćene metode za istraživanje podataka koje su dostupne u softerskim paketima WEKA[3] (Waikato Environment for Knowledge Analysis), Orange[4] odnosno CommonGIS[5]. WEKA softver je razvijen na Univerzitetu Waikato na Novom Zelandu i predstavlja kolekciju metoda, algoritama i programa za mašinsko učenje i istraživanje podataka i može se upotrebljavati besplatno u okviru GNU licence. CommonGIS je interaktivni geo–informacioni sistem baziran na Java programskom jeziku koji obezbeđuje standardnu GIS funkcionalnost, a može biti upotrebljen i kao alat za vizualnu i eksploratornu analizu geografski referenciranih podataka. Takođe, korišćen je i niz alata u programskom jeziku Python, razvijenih za potrebe ove teze.

A.3 Korišćeni skupovi podataka

U istraživanju sprovedenom u cilju otkrivanja osnovnih zakona ljudske mobilnosti [48], kao i identifikovanju atraktivnih lokacija i dinamike kretanja turista [49], korišćeni su javno dostupni metapodaci korisnički–generisanog multimedijalnog sadržaja i to:

- Metapodaci 1 miliona fotografija prikupljenih sa platforme Flickr[2] za teritoriju San Franciska i San Dijega (S1). Iz skupa je za potrebe dalje analize izdvojen podskup S2, koji su činili metapodaci samo onih fotografija čiji je vlasnik postavljao iste u vremenskom periodu dužem od nedelju dana. S2 je izdvojen u cilju eliminacije uticaja turista uz pretpostavku da su vlasnici fotografija koji ih imaju samo nekoliko sa određene lokacije turisti.

- Metapodaci 600 000 fotografija prikupljenih sa iste platforme za teritoriju grada Berlina. Svaki XML[6] sa metapodacima se sastojao od jedinstvenog identifikacionog broja fotografije, korisničkog imena, datuma i vremena kada je fotografija objavljena, geografske dužine i širine lokacije gde je fotografija napravljena, kao i stepen preciznosti oznake (engl. *tag*) kojim je opisana fotografija (1–izuzetno niska preciznost, 16–izuzetno visoka). Nakon filtriranja podataka kako bi se otklonili neispravni podaci (ponavljanje istog korisnika na istoj lokaciji više puta, finalni skup se sastojao od 227 794 jedinstvenih zapisa.

- U oblasti transporta sa ciljem otkrivanja glavnih putanja kojima se ljudi najčešće kreću, vremena početka, trajanja, i završetka putovanja, kao i glavnih transportnih sredstava u upotrebi [50] na većim geografskim razmerama (kontinent), korišćeni su metapodaci 113 157 jedinstvena geo–referencirana video zapisa sa platforme YouTube. Video zapisi koje su korisnici postavljali su bili vezani za afrički kontinent i obuhvatali su period od septembra 2006. do aprila 2011. godine.

[3] http://www.cs.waikato.ac.nz/ml/weka/
[4] https://orange.biolab.si/
[5] http://geoanalytics.net/V-Analytics/
[6] https://en.wikipedia.org/wiki/XML

Na kraju, najveća studija koja je sprovedena u svrhu ove teze, sa ciljem boljeg razumevanja prostorne distribucije i širenja epidemije infektivnih bolesti [22], korišćena su 3 skupa podataka:

- Podaci o stanovništvu (populaciji) na afričkom kontinentu, preuzeti sa stranice Afripop [80]. Ovi podaci sadrže informacije o polu, starosti, statusu, i ostalim demografskim karakteristikama populacije.
- Podaci o zdravstvenom stanju nacije (engl. *Demographic and Health Surveys (DHS)*), sa glavnim fokusom na rasprostranjenost virusa side. Podaci su prikupljeni za teritoriju Obale Slonovače, u periodu od 2008. do 2012. godine.
- Zapisi mobilnih telefona, preuzeti od mobilnog operatera Oranž [17], prikupljeni za teritoriju Obale Slonovače. Ovi podaci su sadržali informacije o lokacijama izvornih i odredišnih antena mobilnog provajdera, broju poziva, ukupnom trajanju poziva, i ostale informacije koje mobilni operateri skupljaju u cilju naprlaćivanja svojih usluga korisnicima.

A.4 Metodologija

Za analizu ovako prikupljenih podataka pored navedenih, korišćene su metode za istraživanje podataka: selekcija obeležja i algoritami za numeričku predikciju, klasterizacija, tehnike vizualizacije velikih količina geo–referenciranih podataka (toplotne mape, OPTICS (*Ordering Points to Identify Clustering Structure*) klasterizacija [5]), statistički modeli i funkcije gustine verovatnoće (engl. *Probability Density Function (PDF)*).
Podaci zapisa mobilnih telefona su analizirani primenom sledećih statističkih metoda: regresija grebena (engl. *Ridge Regression*), regresija vektora (engl. *Support Vector Regression (SVR)*), te rekurzivna eliminacija atributa (engl. *Recursive Feature Elimination (RFE)*).

A.4.1 Dinamika kretanja korisnika na kontinentalnom nivou

Za istraživanje na nivou većih geografskih razmera (npr. kontinent), korišćena je platforma za deljenje video zapisa YouTube, koja omogućava korisnicima da postavljaju materijale koji sadrže jasno definisanu lokaciju na kojoj je materijal snimljen. Sakupljeni metapodaci su obrađeni a zatim i analizirani korišćenjem nekoliko tehnika za geo–vizualizaciju. U početku je korišćeno klasterovanje na osnovu gustine (engl. *Density–Based Clustering*) kako bi se definisala žarišta koja korisnici najčešće posećuju. Zatim su poređene putanje kojima se korisnici najčešće kreću (engl. *Route Similarity Clustering*) kako bi se ustanovile standarne rute između žarišta. Na ovako dobijenim rezultatima je zatim primenjena metoda prostorno–vremenskog klasterovanja (engl. *Spatio–temporal Clustering*) kako bi se definisale vremenske odrednice u kojima su videi najčešće snimani. Na kraju, korišćene su najduže dobijene putanje kako bi se izvršila "fina" analiza i dobile informacije o transportim

sredstvima i delovima dana kada korisnici najčešće putuju (engl. *"Starts, ends and time steps" Clustering*).

Prvi korak je bio da se obrade podaci kako bi se eliminisali zapisi koji su imali invalidnu vremensku i prostornu odrednicu i kako bi se uklonili duplikati koji su postavljeni od istog korisnika u isto vreme i na istoj lokaciji. Kako bi se to postiglo posmatrani geografski prostor je podeljen u mrežu ćelija dimenzije 1000x1000cm. Na osnovu ove obrade je definisano 43 917 putanja (trajektorija). Ovaj proces je prikazan na slici 3.1 u poglavlju **4.2.** Klasterovanje je izvršeno korišćenjem algoritma OPTICS koji omogućava primenu različitih funkcija distance. Prostorno–vremenska sekvenca putanje se može formulisati kao uređena n–torka $T =< x_1, y_1, t_1 >, ..., < x_i, y_i, t_i >$, gde $t_i (i = 1, ..., n)$ označava vremensku odrednicu takvu da su $\forall 1 \leq i < nt_i < t_{i+1}$ i (x_i, y_i) tačke u $\mathbb{R}^2$. Za klasterovanje putanja korišćena je funkcija distance "route similarity", algoritam čija je ideja da se dve najbliže putanje (P, Q) pretražuju kako bi se pronašao najbliži par lokacija (D), odnosno, kako bi se ustanovio prag udaljenosti. U toku pretraživanja, rastojanje između dve putanje se računa kao srednja vrednost razdaljine odgovarajuće pozicije i posmatrane razdaljine. Preskakanjem određene pozicije, povećava se razdaljina, dok se pronalaženjem odgovarajuće pozicije smanjuje razdaljina. Veličina klastera dobijena na osnovu merenja distance predstavlja najfrekventniju rutu kojom korisnici putuju [69].

Za izvođenje analize u odnosu na vreme kretanja korisnika, skup podataka je ograničen na korisnike sa 30 i više putanja što je rezultiralo broju od 14 167 jedinstvenih korisnika. Primenjeni algoritam klasterovanja nad ovim podacima uzima kao dodatni parametar i prag vremenskog rastojanja $maxT$. Algoritam funkcije pronalazi prostornu udaljenost d između dve tačke u odnosnu na vreme kada je taj zapis kreiran. Zatim se dobijeno vreme t proporcijonalno transformiše u odgovarajuću prostornu udaljenost d'. Kombinacijom d i d' se dobija jedinstvena udaljenost na osnovu formule za Euklidsku udaljenost koja se računa na sledeći način:

$$\phi(\delta, \delta') = \sqrt{(\delta_1 - \delta'_1) + ... + (\delta_n - \delta'_n)^2} \tag{A.1}$$

$$\phi(\delta, \delta') = \sqrt{\sum_{i=1}^{n} (\delta_i - \delta'_i)^2} \tag{A.2}$$

Algoritam funkcije za izračunavanje sličnosti putanja *"route similarity"* je prikazan na slici **??** u poglavlju **4.2.**

A.4.2 Procena rasprostranjenosti infekcije virusa na nivou manjih geografskih jedinica

U sledećoj fazi istraživanja, bavili smo se otkrivanjem prostorne epidemiologije virusa side korišćenjem zapisa mobilnih telefona. Procenom rasprostranjenosti infekcije virusa na nacionalnom nivou smanjenjuje se heterogenost koja postoji unutar zemlje.

Kako bi se izračunala rasprostranjenost na nivou manjih geografskih jedinica, estimacija je vršena korišćenjem Gausove funkcije gustine sa prilagodljivom propusnom moći. Procena rasprostranjenosti virusa u prostornoj tački (x, y) je računata po sledećoj formuli:

$$prev(x, y) = \sum_{i}^{n} \frac{1}{{h_i}^2} K\left(\frac{d_i}{h_i}\right) \tag{A.3}$$

gde n predstavlja broj uzoraka, d geometrijsku distancu između posmatranog uzorka i i tačke x, y, K je kernel funkcija, a h propusna moć (engl. *bandwidth*) koja je korišćena nad uzorkom i. Pored toga, svakom posmatranom regionu je dodeljen pokazatelj procene kvaliteta na osnovu veličine uzorka [74]. Za detaljne informacije o procenjenim vrednostima i indikatorima kvaliteta, pogledati tabelu 4.3 u poglavlju **3.5.**

A.4.3 Identifikacija "jakih" veza između tačaka u prostoru

Estimacija "jakih" odnosno "slabih" veza između tačaka u prostoru pomaže istraživačima da bolje razumeju socio–geografske veze među korisnicima [94]. Tokom istraživanja prostorne epidemiologije virusa, težili smo tome da kvantifikujemo jačinu veza između tačaka određenog regiona kako bi ustanovili one koji imaju najveći značaj. Veze između tačaka i i j koje se nalaze u posmatranom regionu predstavljene su kao komunikacija ili mobilnost protoka w_{ij} usmerena od i do j u cilju kvantifikacije broja poziva ili kretanja od jedne do druge pozicije, podeljena sa brojem populacije posmatranog regiona. Kako bi se izvršila kategorizacija veza na slabe i jake, korišćen je pristup [110] gde su primenom filtera dispariteta definisani značajni linkovi među tačkama u prostoru prema sledećoj formuli:

$$\alpha_{ij} = 1 - (k-1) \int_{0}^{p_{ij}} (1-x)^{k-2} \mathrm{d}x < \alpha. \tag{A.4}$$

gde su i i j indeksi posmatranih regiona, α_{ij} je stepen značajnosti posmatrane veze od i do j, k je stepen čvora koji se razmatra, a $p_{ij} = \omega_{ij}/s_i$ je odgovarajuća težina normalizovana jačinom čvora $s_i = \sum_j \omega_{ij}$. Veze gde je $\alpha_{ij} < \alpha$ su karakterisane kao jake veze, statistički značajne na nivou α. Tokom našeg eksperimenta, statistička značajnost α je postavljena na 0.05. Filter dispariteta se određuje na nivou čvora čime se omogućava da se sačuva fluktuacija na globalnom nivou na različitim razmerama. Nakon filtriranja, u svrhu vizualizacije, direktni graf je transformisan u indirektni sumiranjem ω_{ij} i ω_{ij}.

A.4.4 Identifikacija učestalih putanja kretanja

Korišćenjem različitih tehnika geo–vizualizacije (sa naglaskom na agregaciju i klasterizaciju), vršena je analiza zarad identifikacije učestalih putanja kretanja ljudi. Prvobitna ideja je bila odrediti čvorišta sa najvećim stepenom međusobne povezanosti,

kao i identifikovati putanje kojima su čvorišta međusobno povezana korišćenjem algoritma za klasterovanja OPTICS [5]. Ovaj algoritam omogućava primenu različitih funkcija distance u cilju dobijanja standardnih putanja (iz skupa svih putanja) među čvorištima. Osnovna ideja ovog algoritma je da se dve putanje (trajektorije) P i K skeniraju više puta uzastopno u potrazi za najbližim parom pozicija (D predstavlja prag razdaljine). Prilikom skeniranja trajektorija računa se i srednja vrednost udaljenosti između odgovarajućih pozicija, kao i kazneno rastojanje (engl. *penalty distance*). Pronalaženjem odgovarajućih pozicija se smanjuje vrednost kaznenog rastojanja. Veličina klastera koja je dobijena na ovaj način predstavlja vrednost učestalosti korišćenja određene putanje. Svaki podskup trajektorija smo analizirali zasebno, a zatim smo ih spojili u jedan skup. Iz slike 3.3 je očigledno da se glavna čvorišta svake regije nalaze u južnom delu države, te da je verovatnoća da ljudi koji su bliži glavnim čvorištima imaju veće šanse da budu prenosioci virusa.

A.4.5 Grafovi i regionalna povezanost

Za prostornu distribuciju i regionalnu povezanost, koristili smo stopu prevalencije HIV–a za teritoriju Obale Slonovače. Fokusirali smo se na regije kao prostorne jedinice kako bismo na što bolji način izdvojili znanje iz dostupnog skupa podataka [17]. Koristili smo metodu grafa kako bi izračunali povezanost među regionima merenjem komunikacije i migracije između njih. Ovim pristupom smo želeli da istražimo da li su regioni sa višom stopom prevalencije HIV–a bolje povezani od onih sa manjom. Sam postupak određivanja se sastojao iz nekoliko koraka: (1) dodeliti svaku antenu iz skupa pripadajućem regionu (na osnovu geografske dužine i širine); (2) agregirati komunikaciju na nivou regiona prikupljanjem svih antena koje su korišćene za komunikaciju među regionima. Granične vrednosti (veze koje započinju i završavaju se u istom regionu) su bile isključene iz studije. Sledeći korak je bio da se normalizuju težinske vrednosti ivice (engl. *edge weights*). Pošto su regioni neravnomerno naseljeni, mi smo podelili težinske vrednosti ivica w_{ij} (zbir svih poziva između regiona i i j za period od šest meseci) sa proizvodom broja stanovništva N_i i N_j (za 2011. godinu). Na kraju smo prebacili sve dobijene parove težinskih vrednosti kako bismo napravili 3NN graf (engl. *kNN–k Nearest Neighbor*) [44].

Dodavanjem tri najsnažnije veze za svaki od regiona mogli smo da proverimo glavne pravce i čvorišta komunikacije, što smo predstavili na slici 3.4. Čvorovi predstavljaju geografsku lokaciju antene, a njihove boje pokazuju stopu infekcije HIV–om: od žute koja označava regione sa umereno visokom stopom infekcije, do crvene koja označava one sa visokom.

Što se tiče grafa migracije, njega smo napravili na sličan način kao i graf komunikacije. Prvo smo za svakog korisnika odredili koji region predstavlja njegov matični region, zatim smo pratili njegovo kretanje kroz vreme kako bi ispratili ka kojim se regionima najčešće kreće. Nakon otkrivanja tranzicija među regionima, napravili smo matricu parova od regiona do regiona. Krajnji rezultat su činile sve migracije između

regiona za period od šest meseci. Nakon izračunavanja težinskih ivica, izvršena je normalizacija i rezultat je prikaz na slici 3.5.

A.5 Rezultati

Pomenute metode evaluirane su na različitim skupovima podataka i rezultati su upoređeni sa realnim podacima iz tog područja. U poglavlju **4** su detaljno opisani eksperimenti i predstavljeni dobijeni rezultati. U potpoglavlju **4.1** su prikazani rezultati analize sprovedene nad 113 157 jedinstvenih YouTube video zapisa. Bili smo u stanju da identifikujemo glavne putne pravce, vidove transporta, pa čak i avionske letove i avio kompanije koje su ljudi najčešće koristili za putovanje širom Afrike. Takođe, otkrili smo da postoji veliki broj snimaka za period između juna i jula 2010. godine, što ukazuje na FIFA Svetsko prvenstvo koje je održano u Južnoafričkoj republici iste godine. Rezultati ove studije su pokazali da YouTube video snimke možemo koristiti da identifikujemo neke osnovne obrasce u ljudskom ponašanju, kao i za analizu vremenske dinamike i njihove aktivnosti. Rezultati analize sličnosti putanja metodom klasterovanja prikazani su na slici 4.1. Na slici 4.2 su prikazani glavni pravci kretanja korisnika na kontinentu, dok tabela 4.2 prikazuje rezultate vremenske analize.

U delu **4.2** su prikazani rezultati studije u kojoj smo istražili mogućnost korišćenja Flickr metapodataka kao alternativu zapisima mobilnih telefona kada je u pitanju analiziranje ljudske mobilnosti. Da bi to uradili, analizirali smo podatke mobilnih telefona na osnovu trajektorija 6404 Flickr korisnika, izvedene iz metapodataka 1 miliona slika koje se odnose na oblast San Franciska/San Dijega. Naš cilj je bio da pokažemo da se zakonitosti koje možemo izdvojiti posmatranjem podataka mobilnih telefona mogu pronaći i u Flickr podacima, kao i da javno dostupni podaci imaju potencijal da omoguće istraživačima da sprovode analize na većim prostornim razmerama (kontinent/širom sveta). Dobijeni rezultati su pokazali da se Flickr metapodaci mogu koristiti za ove ili slične studije jer predstavljaju odličnu alternativu za zapise mobilnih telefona.

U sledećoj studiji čiji su rezultati prikazani u potpoglavlju **4.3**, izvršena je analiza na osnovu metoda geografske vizualizacije, a podaci koji su korišćeni su metapodaci pridruženi javno dostupnim geo–referenciranim fotografijama prikupljenim sa stranice Flickr, za područje glavnog grada Nemačke–Berlina. Cilj je bio da se identifikuju atraktivne lokacije u gradu, pod pretpostavkom da se atraktivno mesto karakteriše velikom količinom fotografija koje su ljudi postavili. Pored toga, identifikovane su standardne rute kojima se ljudi kreću tokom istraživanja grada. Prikazane tehnike koriste gustinu na bazi grupisanja, sličnosti između putanja i dinamike kretanja. Rezultati ukazuju na to da se informacije dobijene analizom metapodataka Flickr fotografija mogu koristiti za pouzdano otkrivanje lokacija koje su privlačne za turiste, kao i za predlaganje najboljih putanja koje bi trebali koristiti u cilju što efikasnijeg istraživanja grada. Kada je u pitanju turistička dinamika, izvedeno je nekoliko

zaključaka: većina putovanja su kratka lokalna putovanja koje ne vode u centar grada već prate krug puteva oko centra jer stanovnici grada izbegavaju puteve na kojima ima velik broj turista. Zbog toga se najveće turističke atrakcije nalaze baš u centru. Na slici 4.7 su vizuelno prikazani rezultati klasterovanja korišćenjem OPTICS algoritma na nivou čitavog grada, dok slika 4.8 prikazuje primenu istog algoritma za područje centra grada. Na slici 4.9 i 4.10 su prikazani rezultati tokova kretanja ljudi korišćenjem metode zasnovane na klasterovanju putanja kojima se ljudi kreću na osnovu njihove sličnosti.

U poslednjem delu poglavlja **4**, prikazani su rezultati studije koja je pokazala da se neobrađeni ("sirovi") podaci o stvarnom svetu mogu koristiti za izdvajanje znanja značajnog za različite namene. Primarni izvor podataka o lokaciji i kretanju ljudi danas predstavljaju mobilni telefoni, bilo da se radi o zapisima o aktivnostima korisnika mobilnih telefona (CDR) koje prikupljaju telefonske kompanije, ili o direktnim zapisima o lokaciji koju generišu pametni telefoni korišćenjem globalnog pozicionog sistema (GPS). U ovoj studiji, obradili smo problem prostorne distribucije HIV/side na osnovu analize ljudske aktivnosti i mobilnosti u prostoru korišćenjem izdvojenih obeležja, odnosno, razumevanja prostorne varijacije epidemije u globalu. Istražili smo skupove podatka prikupljene od strane pružalaca usluga mobilne telefonije i povezali ih sa prostornom prevalencijom stope HIV virusa koja je procenjena iz javno dostupnih analiza. U tu svrhu, 224 obeležja su izdvojena na osnovu mobilnosti, konektivnosti i kretanja korisnika. Korišćenjem regresionog modela smo identifikovali ključne elemente koji utiču na stopu HIV infekcije a vizualizacijom čestih putanja, međuregionalnih migracija i komunikacija među korisnicima smo nastojali da objasnimo prostornu strukturu epidemije. U slučaju nekoliko modela predviđanja dobijen je visok stepen povezanosti HIV prevalencije sa standarnim vrednostima (>0.7). Kroz analizu doprinosa, identifikovali smo ključne elemente koji utiču na stopu infekcije. Naši rezultati ukazuju na to da su komunikacija tokom noći, prostorna distribucija korisnika i ukupne migracije stanovništva snažno povezane sa HIV–om. Vizualizacijom komunikacije i tokova mobilnosti, nastojali smo da objasnimo prostornu strukturu epidemije. Zaključak je da se jake veze i čvorišta u komunikaciji i mobilnosti mogu uskladiti sa HIV žaristima. Međutim, rezultati ostavljaju mnogo prostora za poboljšanja, naročito u polju definisanja funkcije vektora koja najviše utiče na prenos bolesti. Detaljna studija o načinu odabira takve funkcije je potreban i moguć pravac kretanja za dalja istraživanja. Takođe, grafikoni pojedinačne komunikacije koji su dobijeni na osnovu geografske distribucije bi bili ogroman izvor informacija za otkrivanje veza na nivou većeg stepena detaljnosti. Navedene mogućnosti imaju značajan potencijal u omogućavanju daljeg napretka u domenu modeliranja zaraznih bolesti [15].

A.6 Zaključak

Područje istraživanja doktorske disertacije je bilo usmereno ka istraživanju velikih količina podataka i njihove primene u pronalaženju zanimljivih, korisnih obrazaca o kretanju, ponašanju i navikama ljudi. Pristup je zasnovan na primeni tehnika veštačke inteligencije i istraživanja podataka. Metode su uspešno implementirane i detaljno analizirane na bazi zaštićenih (agregiranih i anonimiziranih) CDR podataka i metapodataka geo–referenciranog multimedijalnog sadržaja. Veći deo disertacije je posvećen primeni pomenutih podataka u oblasti digitalne epidemiologije. Evaluacija je izvršena na podacima iz domena turizmologije, transporta, demografije i kontrole širenja infektivnih bolesti.

Glavni doprinos disertacije je generisanje novih znanja o kretanju, navikama i ponašanju ljudi korišćenjem javno dostupnih podataka o njihovoj lokaciji i kretanju, razvijanje novih modela kojima bi se bolje opisale pomenute pojave i koji se mogu primeniti za rešavanje praktičnih problema, kao što su detekcija atraktivnih lokacija, putanja kretanja ljudi u posmatranom području, njihove navike putovanja, kao i prenošenje virusa HIV na određenom geografskom području. Pored toga, doprinos disertacije je i da se omogući dovoljno precizno predviđanje obrazaca ponašanja, navika i kretanja ljudi. Istraživanjem i izvršavanjem eksperimenata dobijen je odgovor na pitanje koje istraživačke metode predikcije daju najbolje rezultate na geo–referenciranim zapisima koje ljudi generišu, kao i u kojoj meri su pogodni za konkretan zadatak. Primenom različitih skupova podataka za različita predviđanja uočeni su šabloni u ponašanju ljudi što svakako daje značajnu mogućnost unapređenja kvaliteta života ljudi i životne sredine uopšte. Praćenjem uticaja različitih faktora na kretanje i različite obrasce ponašanja, kao što su na primer, navike u komunikaciji ljudi (engl. *Connectivity Patterns*), uočene su potrebe za naglašavanjem pojedinih i zanemarivanje drugih faktora.

Dobijeni rezultati imaju i širu primenu jer se predloženi postupci analize zasnivaju na matematičkim i statističkim metodama koje se lako mogu primeniti na novim problemima iz oblasti u kojima su vršena istraživanja, ali i na problemima iz drugih domena kao što su upravljanje vanrednim situacijama, urbanističko planiranje, upravljanje širenjem stope siromaštva itd.

Appendix B

Used Code

```
#-------------------------------------------------------------------------------
# Name:       Module1: SET1 connectivity matrix
# Purpose:    Regional connectivity and graph representation
# Language:   Python
#
# Author:     Sanja Brdar, Katarina Gavric
#
# Created:    01/10/2014
# Copyright:  (c) Sanja and Katarina 2014
# Licence:
#-------------------------------------------------------------------------------

from Orange.data.sql import *
import numpy as np
import pickle
import datetime
import os

queries = ['SELECT distinct(month(time_date)) FROM `set1r`;']

def main():
    r = SQLReader()
    r.connect('mysql://root@localhost/d4d')
    w = SQLWriter('mysql://root@localhost/d4d')
    print 'connected'

    departments = [i for i in range(1, 51)]
    os.chdir('../data/CONNECTIVITY')

    connectivity = np.zeros((len(departments),len(departments)))
    # Overall connectivity
    for i in range(len(departments)):
        for j in range(len(departments)):
```

```
            try:
                sql_query = 'SELECT sum(nb_voice_calls) FROM set1d WHERE
                    originating_department = ' + str(departments[i]) + ' and
                    terminating_department = ' + str(departments[j]) + ';'
                print sql_query
                r.execute(sql_query)
                data = r.data()
                print i, j, data[0][0]
                connectivity[i,j] = data[0][0]
            except:
                print i, j, '0'
                connectivity[i,j] = 0
    f = file('set1_sum_nb_voice_connectivity.pkl', 'w')
    pickle.dump(connectivity, f)
    f.close()
    # Night connectivity
    for i in range(len(departments)):
        for j in range(len(departments)):
            sql_query = 'SELECT sum(nb_voice_calls) FROM set1d WHERE
                originating_department = ' + str(departments[i]) + ' AND
                terminating_department = ' + str(departments[j]) + ' AND
                HOUR(time_date) BETWEEN 0 AND 5;'
            try:
                r.execute(sql_query)
                data = r.data()
                print i, j, data[0][0]
                connectivity[i,j] = data[0][0]
            except:
                print i, j, '0'
                connectivity[i,j] = 0

    f = file('set1_sum_nb_voice_connectivity_night05.pkl', 'w') #wb write
        binary
    pickle.dump(connectivity, f)
    f.close()

if __name__ == '__main__':
    main()
```

```
#-------------------------------------------------------------------------------
# Name:       Module2: SET1 strong ties inference
# Purpose:    Strong ties identification within SET1
# Language:   Python
#
# Author:     Sanja Brdar, Katarina Gavric
#
# Created:    02/10/2014
# Copyright:  (c) Sanja and Katarina 2014
# Licence:
#-------------------------------------------------------------------------------

import os
import pickle
from scipy.stats.stats import pearsonr
from operator import itemgetter
from Orange.data.sql import *
from scipy.integrate import quad

def extract_mutual(graph):
    keys = graph.keys()
    remove_keys = [(key[0], key[1]) for key in keys if (key[1], key[0]) not
        in keys]
    for key in remove_keys:
        del graph[key]
    return graph

def remove_loops(graph):
    keys = graph.keys()
    remove_keys = [(key[0], key[1]) for key in keys if key[0] == key[1]]
    remove_keys.sort()
    print remove_keys
    print len(remove_keys)
    for key in remove_keys:
        del graph[key]
    return graph

def sum_pairwise(graph):
    keys = graph.keys()
    l1, l2 = [],[]
    for key in keys:
        if key[0] < key[1]:
            l1.append(graph[(key[0], key[1])])
            l2.append(graph[(key[1], key[0])])
##            print str(graph[(key[0], key[1])]) + '\t' + str(graph[(key[1],
    key[0])])
```

```
            print str(graph[(key[0], key[1])]/(graph[(key[0], key[1])]+
                graph[(key[1], key[0])])) + '\t' + str(graph[(key[1],
                key[0])]/(graph[(key[0], key[1])]+ graph[(key[1], key[0])]))
            graph[(key[0], key[1])] += graph[(key[1], key[0])]
    print 'Directed weights correlation', pearsonr(l1, l2)
    remove_keys = [(key[0], key[1]) for key in keys if key[0] > key[1]]
    for key in remove_keys:
        del graph[key]
    return graph

# Data preparation for QGIS
def print_dict(graph):
    r = SQLReader()
    r.connect('mysql://root@localhost/d4d')
    print 'department_id1 department_id2 lat1 long1  lat2   long2
        sum_mobility'
    keys = graph.keys()
    for key in keys:
        r.execute("SELECT lat, lon from hiv_departments_georef where
            department_id = " + str(key[0]))
        data0 = r.data()
        r.execute("SELECT lat, lon from hiv_departments_georef where
            department_id = " + str(key[1]))
        data1 = r.data()
        print str(key[0]) + '\t' + str(key[1]) + '\t' + str(data0[0][0]) +
            '\t' + str(data0[0][1]) + '\t' + str(data1[0][0]) + '\t' +
            str(data1[0][1]) + '\t' + str(graph[key])

def integrand(x):
    return (1-x)**(N-3)

def main():
    os.chdir('../../data/CONNECTIVITY/')
    # Import data
    c = file('set1_sum_nb_voice_connectivity.pkl', 'r')
    connectivity = pickle.load(c)
    c.close()

    N = connectivity.shape[0]

    # Sum weights of directed graph
    for i in range(N-1):
        for j in range(i+1, N):
            temp = connectivity[i,j] + connectivity[j,i]
            connectivity[i,j] = temp
            connectivity[j,i] = temp
```

```
# Strong and weak ties
strong_ties = {}
thr = 0.05
for i in range(N):
    node_strength = connectivity[i,:].sum() - connectivity[i,i]
    node_strength = connectivity[i,:].sum()
    for j in range(N):
        weight = connectivity[i,j]/node_strength
        alpha = 1- (N-2) * quad(integrand, 0, weight)[0]
        alpha = 1- (N-1) * quad(integrand, 0, weight)[0]
        if alpha < thr:
            print quad(integrand, 0, weight)[0], quad(integrand, 0,
                weight)[1]
            strong_ties[i+1, j+1] = connectivity[i,j]
            strong_ties[j+1, i+1] = connectivity[j,i]

strong_ties = remove_loops(strong_ties)
strong_ties = extract_mutual(strong_ties)
print strong_ties
print len(strong_ties)
strong_ties = sum_pairwise(strong_ties)
max_value = max(strong_ties.values())
print max_value
for key in strong_ties.keys():
    strong_ties[key] = strong_ties[key]/max_value
print strong_ties
strong_ties_sorted = sorted(strong_ties.items(), key=itemgetter(1),
    reverse=True)
print strong_ties_sorted
print 'Number of significant edges', len(strong_ties)
print_dict(strong_ties)
```

Bibliography

[1] S. Ahern, M. Naaman, R. Nair, and J. Yang. World explorer: Visualizing aggregate data from unstructured text in geo–referenced collections. *In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 1–10, 2007.

[2] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. *In Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*, pages 171–180, 2000.

[3] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, and D. Pedreschi. Visual analytics toolkit for cluster–based classification of mobility data. *Springer*, pages 432–435, 2009.

[4] G. Andrienko, N. Andrienko, and S. Wrobbel. Visual analytics tool for analysis of movement data. *SIGKDD Explorations Newsletter.*, 9(2):38–46, 2007.

[5] M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering points to identify the clustering structure. *In Proceedings of the International Conference on Management of Data*, pages 49–60, 1999.

[6] J. Bagrow and Y.-R. Lin. Mesoscopic structure and social aspects of human mobility. *PloS one*, 7(5), 2012.

[7] D. Balcan, H. Hu, B. Goncalves, P. Bajardi, and C. Polleto. Seasonal transmission potential and activity peaks of the new influenza a(H1N1): A Monte Carlo likelihood analysis based on human mobility. *BMC Medicine*, 7(45), 2009.

[8] J. Bao, Y. Zheng, and M. F. Mokbel. Location–based and preference–aware recommendation using sparse geo–social networking data. *In Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pages 199–208, 2012.

[9] A.-L. Barabasi. The origin of burst and heavy tails in human dynamics. *Nature*, 435:207–211, 2005.

[10] J. Beaudoing. Flickr image tagging: Patterns made visible. *Bulleting of the Association for Information Science and Technology*, 34:26–29, 2007.

[11] R. Becker, R. Caceres, K. Hanson, J.-M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):18–26, 2011.

[12] V. Belik, T. Geisel, and D. Brockmann. Natural human mobility patterns and spatial spread of infectious diseases. *Physical Review X, arXiv preprint: 1103.6224*, 1(1), 2011.

[13] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, and J. Von Schreeb. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post–earthquake geospatial study in Haiti. *PLoS medicine*, 8(8):1–9, 2011.

[14] R. Benson, M. Schumer, M. Meerschaert, and S. Wheatcraft. Fractional dispersion, Lévy motion, and the MADE tests. *Dispersion in Heterogeneous Geological Formations*, pages 211–240, 2001.

[15] L. Bian. Spatial approaches to modeling dispersion of communicable diseases: A review. *Transactions in GIS*, 17(1):1–17, 2013.

[16] V. Blondel, A. Decuyper, and G. Krings. A survey of results on mobile phone datasets analysis. *ArXiv preprint: 1502.03406*, 2015.

[17] V. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. Data for development: The D4D challenge on mobile phone data. *ArXiv preprint: 1210.0137*, 2012.

[18] T. Bodnar and M. Salathé. Validating models for disease detection using Twitter. *In Proceedings of the 22nd international conference on World Wide Web companion*, pages 699–702, 2013.

[19] A. Boettcher and L. Dongman. Eventradar: A real–time local event detection scheme using Twitter stream. *In IEEE International Conference on Green Computing and Communications*, pages 358–367, 2012.

[20] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland. Once upon a crime: Towards crime prediction from demographics and mobile data. *In Proceedings of the 16th International Conference on Multimodal Interaction*, pages 427–434, 2014.

[21] I. Boutsis, S. Karanikolaou, and V. Kalogeraki. Personalized event recommendations using social networks. *In the 16th IEEE International Conference on Mobile Data Management*, pages 43–48, 2015.

[22] S. Brdar, K. Gavrić, D. Ćulibrk, and V. Crnojević. Unveiling spatial epidemiology of HIV with mobile phone data. *Scientific Reports*, 2016.

[23] L. Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

[24] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439:287–288, 2006.

[25] D. Broniatowski, M. Paul, and M. Dredze. National and local influenza surveillance through Twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS One*, 8(12):e83672, 2013.

[26] J. Brownstein, C. Freifeld, B. Reis, and M. K. Surveillance sans frontiéres: Internet–based emerging infectious disease intelligence and the HealthMap project. *PLoS Med*, 5(7):e151, 2008.

[27] A. Buvé, K. Bishikwabo-Nsarhaza, and G. Mutangadura. The spread and effect of HIV–1 infection in sub–Saharan Africa. *The Lancet*, 359(9322):2011–2017, 2002.

[28] N. Caceres, L. Romero, F. Benitez, and J. Del Castillo. Traffic flow estimation models using cellular phone data. *Intelligent Transportation Systems, IEEE Transactions on*, 13(3):1430–1441, 2012.

[29] J. Candia, M. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[30] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, and J.-F. Pinton. Dynamics of person–to–person interactions from distributed RFID sensor networks. *PLoS ONE*, 5(7):e11596, 2010.

[31] Z. Cheng, J. Caverlee, K. Lee, and D. Sui. Exploring millions of footprints in location sharing services. *In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[32] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. *In Proceedings of 11th Conference on Knowledge Discovery and Data Mining*, pages 243–276, 2011.

[33] M. Coffee, M. Lurie, and G. Garnett. Modeling the impact of migration on the HIV epidemic in South Africa. *Aids*, 21(3):343–350, 2007.

[34] S. Cook, C. Conrad, A. Fowlkes, and M. Mohebbi. Assessing Google Flu Trends performance in the United States during the 2009 influenza virus a (H1N1) pandemic. *PloS One*, 6(8):e23610, 2011.

[35] B. Csáji, A. Browet, V. Traag, J.-C. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, and V. Blondel. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 392(6):1459–1473, 2013.

[36] E. R. Davies. Machine vision: Theory, algorithms, practicalities. *Elsevier*, 2012.

[37] M. de Jager, F. Weissing, P. Herman, B. Nolet, and J. van de Koppel. Lévy walks evolve through interaction between movement and environmental complexity. *Science*, 332(6037), 2011.

[38] Y.-A. de Montjoye, C. Hidalgo, M. Verleysen, and V. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1376), 2013.

[39] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. Blondel. D4D–Senegal: The second mobile phone data for development challenge. *arXiv preprint arXiv:1407.4885*, 2014.

[40] R. Dharmarajan and R. Vijayasanth. An overview on data preprocessing methods in data mining. *International Journal for Scientific Research and Development*, 3, 2015.

[41] A. Edwards, R. Phillips, and N. Watkins. Revisiting Lévy flight search patterns of wandering albatrosses, bumblebees and deer. *Nature*, 449:1044–1049, 2007.

[42] M. El-Dereny and N. Rashwan. Solving multicollinearity problem using ridge regression models. *International Journal of Contemporary Mathematical Sciences*, 6(12):585–600, 2011.

[43] R. Feldman and J. Sanger. The text mining handbook. *Cambridge University Press.*, 2007.

[44] P. Fränti, R. Mariescu-Istodor, and C. Zhong. XNN graph. *In Proceedings of Structural, Syntactic, and Statistical Pattern Recognition*, pages 207–217, 2016.

[45] T. Fu, X. Yin, and Y. Zhang. Voronoi algorithm model and the realization of its program. *Computer Simulation*, 23:89–91, 2006.

[46] A. Gallagher. The wisdom of social multimedia: Using Flickr for prediction and forecast. *In Proceedings of the International Conference on Multimedia*, pages 1235–1244, 2010.

[47] K. Gavrić, S. Brdar, D. Ćulibrk, and V. Crnojević. Linking the human mobility and connectivity patterns with spatial HIV distribution. *3rd International Conference on the Analysis of Mobile Phone Datasets–NetMob*, 2013.

[48] K. Gavrić, D. Ćulibrk, and V. Crnojević. Deriving basic law of human mobility using community–contributed multimedia data. *International Conference on Pattern Recognition Applications and Methods*, pages 543–546, 2013.

[49] K. Gavrić, D. Ćulibrk, P. Lugonja, M. Mirković, and V. Crnojević. Detecting attractive locations and tourists' dynamics using geo–referenced images. *10th International Conference on Telecommunication in Modern Satellite Cable and Broadcasting Services*, pages 208–2011, 2011.

[50] K. Gavrić, D. Ćulibrk, M. Mirković, and V. Crnojević. Using YouTube data to analyze human continent–level mobility. *International Conference on Computational Aspects of Social Networks*, pages 207–210, 2011.

[51] N. Generous, G. Fairchild, A. Deshpande, S. Valle, and R. Priedhorsky. Global disease monitoring and forecasting with Wikipedia. *PLoS Computational Biology*, 10(11):e1003892, 2014.

[52] D. Gerberry, G. Bradley, J. Wagner, G. Garcia-Lerma, W. Heneine, and S. Blowera. Using geospatial modeling to optimize the rollout of antiretroviral–based pre–exposure HIV interventions in sub–Saharan Africa. *Nature Communications*, 5(5454), 2014.

[53] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[54] F. Girardin, D. Fioreb, C. Rattib, and J. Blata. Leveraging explicitly disclosed location information to understand tourist dynamics: A case study. *Journal of Location Based Services*, 2(1):41–56, 2008.

[55] M. Gonzalez, C. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[56] S. Gunn. Support vector machines for classification and regression. *ISIS technical report*, 14, 1998.

[57] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning, Springer*, 46(1-3):389–422, 2002.

[58] J. Han, J. Kamber, and J. Pei. *Data mining: Concepts and techniques*. Morgan Kaufmann, 2012.

[59] D. Hardy. Discovering behavioral patterns in collective authorship of place–based information. *In Proceedings of 9th International Conference of the Association of Internet Researchers*, 2008.

[60] S. Hasan, C. Schneider, S. Ukkusuri, and M. Gonzalez. Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, 15:304–318, 2012.

[61] S. Hasan, X. Zhan, and S. Ukkusuri. Understanding urban human activity and mobility patterns using large–scale location–based data from online social media. *In Proceedings of the 2nd International Workshop on Urban Computing*, pages 6:1–6:8, 2013.

[62] S. Havlin and D. Ben-Avraham. Diffusion in disordered media. *Physics*, 51:187–292, 2002.

[63] D. Heymann and G. Rodier. Hot spots in a wired world: WHO surveillance of emerging and re–emerging infectious diseases. *Lancet Infect Dis*, 1:345–353, 2001.

[64] D. Heymann and G. Rodier. Global surveillance, national surveillance, and SARS. *Emerging Infect Dis*, 10:173–175, 2004.

[65] L. Hollenstein and R. Purves. Exploring place through user–generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science*, 1(1):21–48, 2010.

[66] N. Humphries, N. Queiroz, J. Dyer, N. Pade, and M. Musyl. Environmental context explains Lévy and Brownian movement patterns of marine predators. *Nature*, 465(7301), 2010.

[67] X. Jin, A. Gallagher, L. Cao, J. Luo, and J. Han. The wisdom of social multimedia: Using Flickr for prediction and forecast. *In Proceedings of the International Conference on Multimedia*, pages 1235–1244, 2010.

[68] E. Kalipeni and L. Zulu. HIV and AIDS in Africa: A geographic analysis at multiple spatial scales. *GeoJournal*, 77(4):505–523, 2012.

[69] S. Kisilevich, F. Mansmann, P. Bak, D. Keim, and A. Tchaikin. Where would you go on your next vacation? A framework for visual exploration of attractive places. *2nd International Conference on Advanced Geographic Information Systems, Applications, and Services*, pages 21–26, 2010.

[70] C. Körner, D. Benz, A. Hotho, M. Strohmaier, and G. Stumme. Stop thinking, start tagging: Tag semantics arise from collaborative verbosity. *In Proceedings of the 19th International Conference on World Wide Web*, 2010.

[71] G. Krings, F. Calabrese, C. Ratti, and V. Blondel. Urban gravity: A model for inter–city telecommunication flows. *Journal of statistical mechanics: theory and experiment*, 2009(7), 2009.

[72] T. Laksanasopin, T. Guo, S. Nayak, and A. Sridhara. A smartphone dongle for diagnosis of infectious diseases at the point of care. *Science translational medicine*, 7(273), 2015.

[73] J. Larmarange. UNAIDS: Developing subnational estimates of HIV prevalence and the number of people living with HIV from survey data. http://www.unaids.org/, December 2014. Date of access:12/12/2014.

[74] J. Larmarange and V. Bendaud. HIV estimates at second subnational level from national population–based surveys. *AIDS*, 28, 2014.

[75] J. Larmarange, R. Vallo, S. Yaro, P. Msellati, and N. Méda. Methods for mapping regional trends of HIV prevalence from demographic and health surveys (DHS). *CyberGeo: European Journal of Geography*, 558, 2011.

[76] D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of Google Flu: Traps in big data analysis. *Science*, 343, 2014.

[77] J. Leskovec, A. Rajaraman, and J. Ullman. Mining of massive datasets. *Stanford Technical Report*, 2014.

[78] D. Leung and S. Newsam. Proximate sensing: Inferring what–is–where from geo–referenced photo collections. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2955–2962, 2010.

[79] A. Lima, M. De Domenico, V. Pejović, and M. Musolesi. Disease containment strategies based on mobility and information dissemination. *Scientific reports*, 5, 2015.

[80] C. Linard, M. Gilbert, R. Snow, A. Noor, and A. Tatem. Population distribution, settlement patterns and accessibility across Africa in 2010. *PloS one*, 7(2):e31743, 2012.

[81] C. Lynch and R. C. The transit phase of migration: Circulation of malaria and its multidrug–resistant forms in Africa. *PLoS Med*, 8(5):e1001040, 2011.

[82] K. Maat, B. Van Wee, and D. Stead. Land use and travel behavior: Expected effects from the perspective of utility theory and activity–based theories. *Environment and Planning B: Urban Analytics and City Science*, 32(1):33–46, 2005.

[83] D. McIver and J. Brownstein. Wikipedia usage estimates prevalence of influenza–like illness in the United States in near real–time. *PLoS Computational Biology*, 10(4):e1003581s, 2014.

[84] P. Meier. Digital humanitarians: How big data is changing the face of humanitarian response. *Crc Press*, 2015.

[85] J. Messina, M. Emch, J. Muwonga, K. Mwandagalirwa, S. Edidi, N. Mama, A. Okenge, and S. Meshnick. Spatial and socio–behavioral patterns of HIV prevalence in the Democratic Republic of Congo. *Social Science & Medicine*, 71(8):1428–1435, 2010.

[86] H. Miller and J. Han. Geographic data mining and knowledge discovery: An overview. *In Geographic Data Mining and Knowledge Discovery*, pages 1–26, 2009.

[87] E. Minkov, B. Charrow, J. Ledlie, S. Teller, and T. Jaakkola. Collaborative future event recommendation. *In Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 819–828, 2010.

[88] J. Mossong, N. Hens, M. Jit, P. Beutels, and K. Auranen. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med*, 5(3):e74, 2008.

[89] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: Universal patterns in human urban mobility. *PloS one*, 7(5):e37027, 2012.

[90] D. Olson, K. Konty, M. Paladini, C. Viboud, and L. Simonsen. Reassessing Google Flu trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. *PLoS Computational Biology*, 9(10):e1003256, 2013.

[91] D. Paolotti, A. Carnahan, and C. V. Web–based participatory surveillance of infectious diseases: The Influenzanet participatory surveillance experience. *Clinical Microbiology and Infection*, 20:17–21, 2014.

[92] M. Paul and M. Dredze. You are what you tweet: Analyzing Twitter for public health. *In Proceedings of Fifth International AAAI Conference on Weblogs and Social Media*, pages 265–272, 2011.

[93] A. Pentland and S. Pentland. Honest signals: How they shape our world. *Cambridge (MA): MIT Press.*, 2008.

[94] S. Phithakkitnukoon, Z. Smoreda, and P. Olivier. Socio–geography of human mobility: A study using longitudinal mobile phone data. *PloS one 7*, 7(6):e39253, 2012.

[95] J. Qadir, A. Ali, R. Rasool, and A. Zwitter. Crisis analytics: Big data–driven crisis response. *Journal of International Humanitarian Action*, 1(12), 2016.

[96] K. Radinsky, J. Teeven, A. Bocharov, and E. Horvitz. Modeling and predicting behavioral dynamics on the web. *In Proceedings of the 21st International Conference on World Wide Web*, pages 599–608, 2012.

[97] G. Ramos-Fernández, J.-L. Mateos, O. Miramontes, G. Cocho, H. Larralde, and A.-O. B. Lévy walk patterns in the foraging movements of spider monkeys (Ateles geoffroyi). *Behavioral Ecology and Sociobiology*, 55(3), 2004.

[98] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from Flickr tags. *In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 103–110, 2007.

[99] J. Read, K. Eames, and W. Edmunds. Dynamic social networks and the implications for the spread of infectious disease. *Journal of The Royal Society Interface*, 5(26):1001–1007, 2008.

[100] A. Redner. A guide to first–passage process. *Cambridge University Press*, 2005.

[101] E. Reid and R. Cervero. Travel and the built environment: A synthesis. *Journal of the Transportation Research Board*, 1780:87–113, 2001.

[102] D. Righton and J. Pitchford. Minimizing errors in identifying Lévy flight behavior of organisms. *Journal of Animal Ecology*, 76:222–229, 2007.

[103] D. Saez-Trumper, D. Quercia, and J. Crowcroft. Ads and the city: Considering geographic distance goes a long way. *In Proceedings of the Sixth ACM Conference on Recommender Systems*, pages 187–194, 2012.

[104] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: Real–time event detection by social sensors. *In Proceedings of the 19th International Conference on World Wide Web*, pages 851–860, 2010.

[105] M. Salathé, L. Bengtsson, T. Bodnar, D. Brewer, J. Brownstein, C. Buckee, E. Campbell, C. Cattuto, S. Khandelwal, P. Mabry, and A. Vespignani. Digital epidemiology. *PLoS computational biology*, 8(7):e1002616, 2012.

[106] M. Salathé, C. Freifeld, S. Mekaru, A. Tomasulo, and J. Brownstein. Influenza a (H7N9) and the importance of digital epidemiology. *The New England Journal of Medicine.*, 369:401–404, 2013.

[107] M. Salathé, M. Kazandjieva, L. J.-W., P. Levis, M. Feldman, and J. Jones. A high–resolution human contact network for infectious disease transmission. *In Proceedings of the National Academy of Sciences of the United States of America*, 107(51), 2010.

[108] J. Saramäki, E. Leicht, E. López, S. Roberts, F. Reed-Tsochas, and R. Dunbar. Persistence of social signatures in human communication. *In Proceedings of the National Academy of Sciences*, 111(3):942–947, 2014.

[109] P. Serdyukov, V. Murdock, and R. Zwol. Placing Flickr photos on a map. *In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 484–491, 2009.

[110] A. Serrano, M. Boguná, and A. Vespignani. Extracting the multiscale backbone of complex weighted networks. *In Proceedings of the National Academy of Sciences*, 106(16):6483–6488, 2009.

[111] S. Shekar and Y. Huang. Discovering spatial co–location patterns: A summary of results. *In Proceedings of the 7th International Symposium on Advances in Spatial and Temporal Databases*, pages 236–256, 2001.

[112] S. Shekhar and S. Chawla. *Spatial databases: A tour*. Pearson Education, 2003.

[113] M. Shlesinger, G. Zaslavsky, and J. Klafter. Strange kinetics. *Nature*, 363(6424), 1993.

[114] A. Signorini, A. Segre, and P. Polgreen. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza a H1N1 pandemic. *PLoS One*, 6(5):e19467, 2011.

[115] D. Sims, E. Southall, N. Humphries, G. Hays, C. Bradshaw, and J. Pitchford. Scaling laws of marine predator search behaviour. *Nature*, 451(7182), 2008.

[116] M. Small and D. Singer. Resort to arms: International and civil wars, 1816–1980. *Sage Publications*, 1982.

[117] C. Smith-Clarke, A. Mashhadi, and L. Capra. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. *In Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 511–520, 2014.

[118] P. Sobkowicz, M. Kaschesky, and G. Bouchard. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarter*, 29:470–479, 2012.

[119] E. Štrumbelj and I. Kononenko. A general method for visualizing and explaining black–box regression models. *Adaptive and Natural Computing Algorithms*, pages 21–30, 2011.

[120] E. Štrumbelj and I. Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014.

[121] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to data mining*. Pearson, 2006.

[122] Q. D. Team. *QGIS geographic information system*. Open Source Geospatial Foundation, 2009.

[123] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society.*, 58:267–288, 1996.

[124] M. Tizzoni, P. Bajardi, A. Decuyper, G. King, C. Schneider, V. Blondel, Z. Smoreda, M. González, and V. Colizza. On the use of human mobility proxies for modeling epidemics. *PLoS computational biology*, 10(7):e1003716, 2014.

[125] G. Valkanas and D. Gunopulos. How the live web feels about events. *In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, pages 639–648, 2013.

[126] V. Vapnik. The nature of statistical learning theory. *Springer Verlag*, 1995.

[127] B. Vilhelmson. Daily mobility and the use of time for different activities. *GeoJournal*, 48(3):177–185, 1999.

[128] G. Viswanathan, V. Afanasyev, S. Buldyrev, E. Murphy, P. Prince, and H. Stanley. Lévy flight search patterns of wandering albatrosses. *Nature*, 381(6581), 1996.

[129] G. Viswanathan, S. Buldyrev, S. Havlin, M. Da Luz, E. Raposo, and H. Stanley. Optimizing the success of random searches. *Nature*, 401(6756), 1999.

[130] H. Wang, F. Calabrese, G. Di Lorenzo, and C. Ratti. Transportation mode inference from anonymized and aggregated mobile phone call detail records. *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 318–323, 2010.

[131] P. Wang, M. Gonzalez, C. Hidalgo, and A.-L. Barabasi. Understanding the spreading patterns of mobile phone users. *Science*, 324:1071–1076, 2009.

[132] Cote d'Ivoire dhs, 2008–12–HIV fact sheet (French), publication (ID HF46). https://www.dhsprogram.com, September 2013. Date of access:14/10/2013.

[133] UN OCHA: Humanitarian response. https://www.humanitarianresponse.info/, 2010.

[134] Joint United Nations programme on HIV/AIDS–UNAIDS (HIV and AIDS estimates. http://www.unaids.org/, September 2013. Date of access:14/10/2013.

[135] Identifying populations at greatest risk of infection, geographic hotspots and key populations. UNAIDS reference group on estimates, modeling and projections. http://www.epidem.org/resources/, September 2013.

[136] L.-Y. Wei, Y. Zheng, and W.-C. Peng. Constructing popular routes from uncertain trajectories. *In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 195–203, 2012.

[137] S. Weiss, N. Indurkhya, T. Zhang, and F. Damerau. Text mining–predictive methods for analyzing unstructured information. *Springer*, 2005.

[138] A. Wesolowski, N. Eagle, A. Tatem, D. Smith, A. Noor, R. Snow, and C. Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.

[139] N. Williams, T. Thomas, M. Dunbar, N. Eagle, and A. Dobra. Measures of human mobility using mobile phone records enhanced with GIS data. *Plos One*, 10(7):e0133630, 2014.

[140] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting geographical influence for collaborative point–of–interest recommendation. *In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 325–334, 2011.

[141] S. Young. Recommended guidelines on using social networking technologies for HIV prevention research. *AIDS and Behavior*, 16(7):1743–1745, 2012.

[142] S. D. Young. A "big data" approach to HIV epidemiology and prevention. *Preventive medicine*, 70:17–18, 2015.

[143] S. D. Young, W. G. Cumberland, S.-J. Lee, D. Jaganath, G. Szekeres, and T. Coates. Social networking technologies as an emerging tool for hiv prevention: A cluster randomized trial. *Annals of internal medicine*, 159(5):318–324, 2013.

[144] S. D. Young, C. Rivers, and B. Lewis. Methods of using real–time social media technologies for detection and remote monitoring of hiv outcomes. *Preventive medicine*, 63:112–115, 2014.

[145] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with GPS history data. *In Proceedings of the 19th International Conference on World Wide Web*, pages 1029–1038, 2010.

[146] Y. Zheng. Location–based social networks: Users. *In Computing with Spatial Trajectories*, pages 243–276, 2011.

[147] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W. Y. Ma. Recommending friends and locations based on individual location history. *ACM Transactions on the Web*, 5:1–44, 2011.

[148] Y. Zheng, L. Zhang, X. Xie, and M. Wei-Ying. Mining interesting locations and travel sequences from GPS trajectories. *In Proceedings of the 18th International Conference on World Wide Web*, pages 791–800, 2009.

[149] Q. Zhou, W. Chen, S. Song, G. J., K. Weinberger, and Y. Chen. A reduction of the Elastic Net to support vector machines with an application to GPU computing. *Association for the Advancement of Artificial Intelligence.*, 2014.

[150] H. Zou and T. Hastie. Regression shrinkage and selection via the Elastic Net, with application to microarrays. *Journal of the Royal Statistical Society*, (67):301–320, 2006.