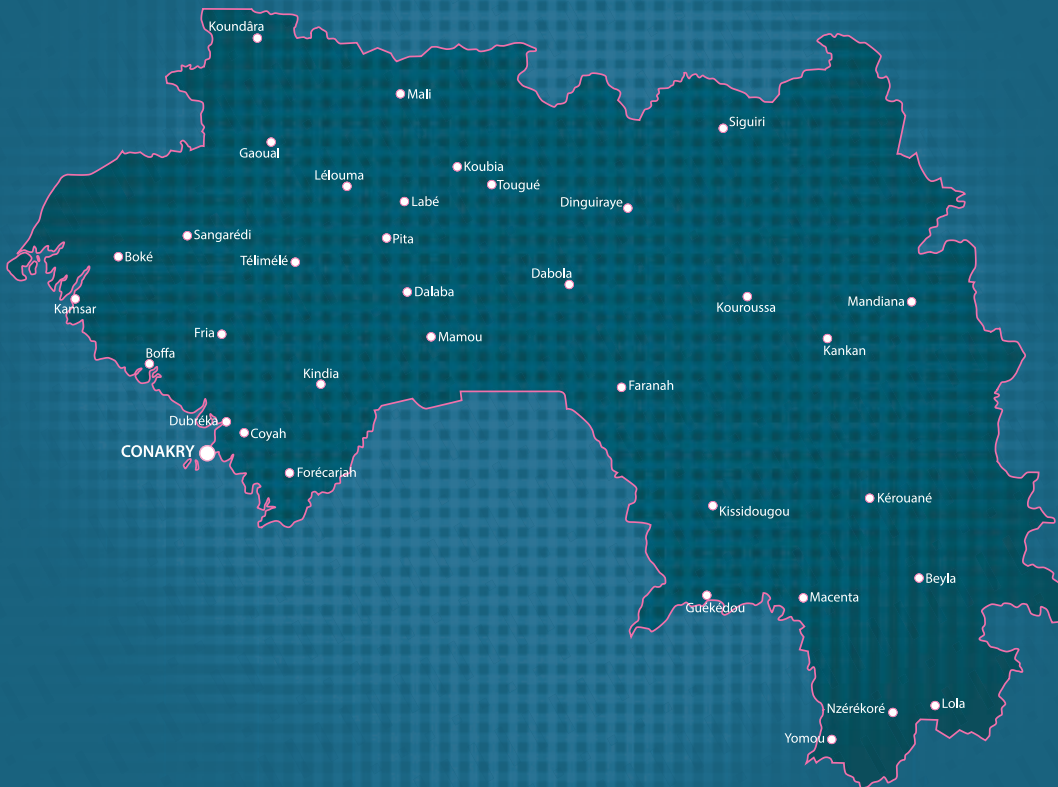


Call Detail Record (CDR) ANALYSIS : REPUBLIC OF GUINEA

Final Report



Call detail record (CDR) analysis: Republic of Guinea

This International Telecommunication Union (ITU) report was prepared by ITU expert Professor Ryosuke Shibasaki of the Center for Spatial Information Science, University of Tokyo, under the supervision of the Telecommunication Development Bureau (BDT) Least Developed Countries, Small Island Developing States and Emergency Telecommunications Division (LSE) in the Projects and Knowledge Management Department (PKM).

The designations employed and presentation of material in this publication, including maps, do not imply the expression of any opinion whatsoever on the part of ITU concerning the legal status of any country, territory, city or area, or concerning the delimitations of its frontiers or boundaries.

ISBN:

978-92-61-22281-9 (Paper version)

978-92-61-22291-8 (electronic version)

978-92-61-22301-4 (EPUB version)

978-92-61-22311-3 (MOBI version)



Please consider the environment before printing this report.

© ITU 2017

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

1	Summary	1
2	Background	1
3	Mobile communications in Guinea	2
3.1	Mobile phone technology	2
3.2	Situation of mobile subscribers in Guinea	3
4	Call detail records (CDR)	4
4.1	Extracting human mobility and spatio-temporal distribution from CDR data	4
4.2	Public and private sector impact	5
5	CDR data sets and basic statistics	6
6	CDR data: Limitations and difficulties	8
7	Methodology	9
7.1	Overview of CDR data analysis (including stakeholders)	9
7.2	Overview of CDR data analysis work flow	10
7.3	CDR data specification	10
7.4	CDR processing procedure	11
7.5	Interpolation format specification	14
7.6	Hadoop system for data processing	17
7.7	Dynamic population estimation	18
7.8	Mobile phone and SIM card cloning	20
8	Analysis results	21
8.1	Mobile phone user population by Prefecture	21
8.2	People flow visualization using interpolation result	22
8.3	Movement of people from one town (over 48 hours)	24
8.4	Tracing people who passed through a hazard area	27
8.5	Attempt to detect cloning phones and SIM cards	28
8.6	Transboundary analysis	30
8.6.1	CDR data set with transboundary IMEI	31
8.6.2	Daily activity statistics	31
8.6.3	Inbound transboundary population movement	33
8.6.4	Outbound transboundary population movement	34
9	Discussions	36
	Appendix 1: Overall concept of automated CDR data analysis	38
	Appendix 2: Building points of interest (POIs) and road networks	39

List of Tables, Figures and Boxes

Tables

Table 1: Sample CSV file data	16
-------------------------------	----

Figures

Figure 3.1: Commutation basis of GSM system	3
Figure 3.2: Administrative map of Guinea	4
Figure 5.1: Cell tower location at Guinea	6
Figure 5.2: CDR data set	6
Figure 5.3: Daily call activity	7
Figure 5.4: Daily unique user	7
Figure 5.5: Average total unique location (daily)	8
Figure 7.1: Overall concept of CDR data analysis	9
Figure 7.2: Overview of CDR analysis work flow	10
Figure 7.3: CDR processing procedure	11
Figure 7.4: CDR process of stay point extraction	12
Figure 7.5: Distribution of POIs and process of stay point reallocation	12
Figure 7.6: An example of a shortest path search result	13
Figure 7.7: Conceptual framework of the process of shortest path search	13
Figure 7.8: Distribution of POIs and process of stay point reallocation	14
Figure 7.9: Trajectory data and stay point	14
Figure 7.10: Time sequential trip points	15
Figure 7.11: Trajectory check (internal use)	17
Figure 7.12: Sample Hadoop Cluster	18
Figure 7.13: Home location identification with population magnification	19
Figure 7.14: On a daily based data aggregation for dynamic population estimation	20
Figure 8.1: Mobile phone user population by Prefecture	21
Figure 8.2: Most visited locations by the people living in Conakry	22
Figure 8.3: Second most visited locations by the people living in Conakry	22
Figure 8.4a: People flow at 6 a.m. and 9 a.m.	23
Figure 8.4b: People flow at 12 p.m. and 3 p.m.	24
Figure 8.5: Flow from a specific town (over 48 hours)	25
Figure 8.6a: Distribution of people who are/are not in a hazard area	27
Figure 8.6b: Trace people who passed through a hazard area (after 10 hours)	27
Figure 8.7a: Where did they go in next 48 hours? (After 12 hours)	28
Figure 8.7b: Where did they go in next 48 hours? (After 47.5 hours)	28
Figure 8.8: Cloning detection (cloning IMEI)	29
Figure 8.9: Cloning detection (cloning IMSI)	30
Figure 8.10: CDR data set with transboundary IMEI of Guinea	31
Figure 8.11: Daily call activity of Guinea based mobile users visiting Sierra Leone	31
Figure 8.12: Daily call activity of Guinea mobile users visiting Liberia	32
Figure 8.13: Daily unique users of Guinea mobile users visiting Sierra Leone	32
Figure 8.14: Daily unique users of Guinea mobile users visiting Liberia	33
Figure 8.15: Footprint of inbound mobile users to Guinea	33
Figure 8.16: Number of mobile users moving into Guinea	34
Figure 8.17: Footprint of visitors to Sierra Leone	34
Figure 8.18: Footprint of visitors to Liberia	35
Figure 8.19: Origin and destination of transboundary population to Sierra Leone	35
Figure 8.20: Origin and destination of transboundary population to Liberia	36
Figure A.1.1: Overall concept of CDR data analysis	38
Figure A.2.1: Building POI in West Africa	39

Figure A.2.2: Road network data in West Africa	39
Figure A.2.3: Assessment of road network in West Africa	40
Figure A.2.4: Extracting road network in West Africa	40

1 Summary

The mobile phone is one of the most ubiquitous technologies of the modern era, being used as a tool not only for communication but also for information provision in sectors such as government, finance, education, agriculture, and health. It is especially important in the field of disaster and emergency management, providing crucial estimates of the number and movement of people prior to or during such events.

The situation in the Republic of Guinea is a prime example, being heavily impacted by the Ebola outbreak, along with neighbouring countries, Liberia and Sierra Leone, where an understanding of human mobility was critical to an effective intervention policy to tackle the disease. Because Ebola is an epidemic disease, by understanding human mobility, authorities can build models that are essential when making decisions, efficient policies, and interventions to contain and tackle the disease effectively.

Call detail record (CDR) data provides information about activities on a mobile communication network and aggregated human mobility that can enable swift action against the disease. This report demonstrates how analysed CDR data can contribute to addressing specific issues related to Ebola epidemics by estimating dynamic trajectories and spatio-temporal distribution of people.

For this report, a human movement analysis at two different scales has been conducted; city-to-city movement in Guinea, and transboundary movement across three countries, Guinea, Sierra Leone, and Liberia. Results of the city-to-city scale analysis showed a strong correlation between the distribution of mobile phone users and actual populations at the district level and demonstrated the possibility of quantitatively estimating spatio-temporal distribution and movement of people.

Although the scope of this report includes only the population covered by CDR data, the results of the analysis will give government and local authorities a better and quantitative understanding of population flow patterns over time (short- and long-term) and at specific events. This enables the extraction of people's city-to-city migration data for analysis from city-to-city during an outbreak of disease that could be vital to understanding how and where the disease spreads and to manage its eventual control.

In addition, analysis of available data related to transboundary movement indicates that there are more frequent movements in neighbouring villages along the border than movement to and from central areas in neighbouring countries. This indicates that tracing movement to and from central areas is not sufficient to tackle an outbreak, however it should be noted that the data analysed was recorded during the containment stage of the Ebola outbreak.

This report explains how spatio-temporal user population movement and spatio-temporal distribution of mobile phone users can be used to analyse this and other societal issues. This report reviews the use of mobile technology and advanced methods of information gathering and dissemination. This report also demonstrates how data collected by mobile phone network operators can cost-effectively provide accurate and detailed maps of population distribution over a selected area. It also describes how mobile phone technology is generally used; it discusses GSM (Global System for Mobile Communication) and future technologies; and it presents various applications for use in sectors such as agriculture, health, finance, governance, and disaster preparedness, response and recovery.

2 Background

Spatio-temporal data of the population can be considered as a key input for policy intervention to control the Ebola epidemic. This is because Ebola is a communicable disease and the way it spreads is significantly affected by human mobility. The globalization of economies has increased the volume

and rapidity of human mobility, and it has lowered the cost of transportation and made large scale human movement easier. Consequently, call detail record (CDR) data (mobile phone data) has been attracting the attention of policy makers and researchers in various fields because of the capacity of capturing population movement patterns and trajectories.

The CDR data generally consists of randomized identity tags, time stamps of mobile communications, and approximate locations of communications, which are represented by the geographic coordinates of cell phone antennas. Although the antenna locations do not allow us to pinpoint exact locations of mobile phone users, proxy location from cell phone devices can be a powerful source to describe the general pattern of population movement.

Traditionally, city monitoring and analysis rely on a fixed location and a considerable amount of statistical data and this does not permit the identification of multi-temporal events in wide areas. However, in addition to the quantitative aspect of human mobility (e.g. such as population volume and its speed), understanding qualitative aspects of the people in CDR data (e.g. gender, age, and occupation) is expected to expand the use of CDR data to various applications. In this case, the usage logs of mobile devices will be treated as a medium for data collection.

Call detail records represent mobile communication network activity data, all of which could be used as aggregate data to estimate the target area population density.

For large scale monitoring, CDR data from mobile network base stations provides excellent spatial patterns that reflect urban life, its temporal dynamics, and it could potentially become a new way to extract or identify less evident problems. Analysis of CDR data captures human movement, however, most analysis uses CDR data of a single country or part of a region of a country because mobile network operators cannot easily trace transboundary movements.

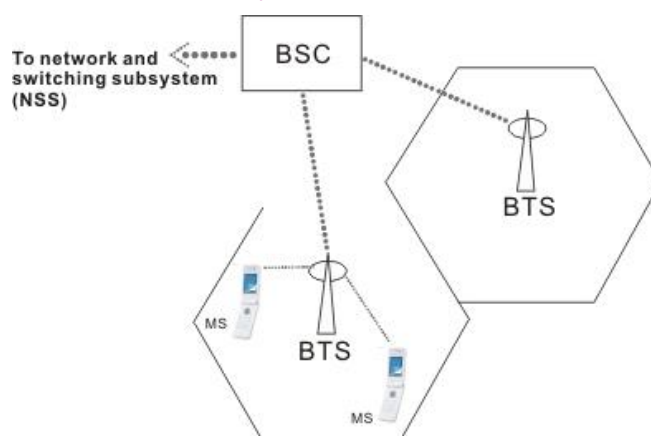
In the last outbreak of Ebola in West Africa over 11 000 people died. National-health authorities in West Africa have struggled to contain Ebola, especially how to estimate the affected population in the outbreak prone area and how to understand outbreak patterns spreading from one area to another. Currently, population movement simulations are based on statistical data that cannot provide reliable and dynamic population flows from an area where an outbreak has occurred, nor can they predict where outbreaks were likely happen next. This type of simulation can be better served by using call-data-records (CDR) generated from mobile phone data. The CDR data contains time and location information for voice, messaging, and data communication of each handset – collected for billing purposes – and it can also be used to estimate time-based and up-to-date dynamics of population movements, leading to better support and preparation and more lives saved.

3 Mobile communications in Guinea

3.1 Mobile phone technology

In real terms, the first mobile network came into existence nearly 35 years ago and has been steadily developing into the modern high-speed services that are taken for granted today.

Figure 3.1: Commutation basis of GSM system



Source: ITU

In 1982, the Conference of European Posts and Telecommunications (CEPT) established a “Group Special Mobile (GSM)”, because existing analogue systems were unable, from a subscriber point of view, to offer an acceptable service. A GSM system is basically designed as a combination of three major subsystems: the Network Switching System (NSS), the Base Station System (BSS) and the Operation Support System (OSS).¹ In general, data is collected from the Base Station Controller (BSC), which is a part of the radio subsystem. When a user makes a call, a mobile phone connects to the closest Base Transceiver Station (BTS).²

Current and next generation of mobile phone technology

- 2G networks, first introduced in 1992, the second-generation of cellular telephone technology, was the first to use digital encryption of conversations, and the first to offer data services and SMS text messaging.
- 3G networks succeeded 2G, offering faster data transfer rates and were the first to enable video calls. This makes them especially suitable for use in modern smartphones, which require constant high-speed Internet connection for many of their applications.
- 4G is the fourth generation of mobile phone communications standards. It is a successor of the 3G and provides ultra-broadband Internet access for mobile devices. The high data transfer rates make 4G networks suitable for use with wireless modems for laptops and even home Internet access.

3.2 Situation of mobile subscribers in Guinea

The Republic of Guinea is composed of four natural regions: Maritime Guinea, Middle Guinea, Upper Guinea and Forested Guinea. Guinea is divided into eight administrative regions that are subdivided into a total of thirty-three prefectures. The capital city is Conakry, which is also the largest city and economic centre. The 2017 population for Guinea has been projected to reach over 11.7 million people³.

¹ GSM: <https://www.itu.int/osg/spu/ni/3G/casestudies/GSM-FINAL.doc>

² T. Horanont. CSIS Discussion Paper. 2012. A Study on Urban Mobility and Dynamic Population Estimation by Using Aggregate Mobile Phone Sources No. 115.

³ Institut national de la statistique: www.stat-guinee.org/

Figure 3.2: Administrative map of Guinea

The Republic of Guinea

- 8 Administrative Regions:
 - Conakry Region
 - Nzérékoré Region
 - Kankan Region
 - Kindia Region
 - Boké Region
 - Labé Region
 - Faranah Region
 - Mamou Region
- Capital City:
 - Conakry, the biggest city
- Population (2017 projection):
11,702,692 habitants



As of 2014, approximately 72.1 per cent of the population was covered by mobile network operators. There are four major mobile phone operators in Guinea: Cellcom, Intercel, MTN and Orange. They offer GSM services, which work on both a prepaid basis and post-paid service. The Regulatory Authority for Posts and Telecommunications (ARPT) regulates the Guinea telecoms sector.

4 Call detail records (CDR)

Call detail record data contains basic information about mobile phone usage, such as, which cell towers the caller and recipient phones were connected to at the time of the call, the identities of sources (points of origin), the identities of destinations (end points), the duration of each call, the amount billed for each call, the total usage time in the billing period, the total free time remaining in the billing period, and the running total charged during the billing period. In the case of pinpointing people, the operator knows cell tower locations and it is possible to use CDRs to approximate the location of both parties. The spacing of cell towers, and thus the accuracy in determining caller location, varies according to expected traffic and terrain. Cell towers are typically spaced 2 to 3 km apart in rural areas and 400 to 800 m apart in densely populated areas. This geo-spatial information is extremely useful for humanitarian and development applications.

4.1 Extracting human mobility and spatio-temporal distribution from CDR data

Sequential locations recorded in CDR data can provide a partial view of population trajectories because CDR data are individual specific, allowing individual movements to be traced. With the aggregation of individual movements in CDR data, patterns, trends, and spatio-temporal distribution of the whole population can be tracked. By overlaying such information with secondary data, relationships between human mobility and other societal factors can also be analysed.

4.2 Public and private sector impact

Social impact (country level)

This section describes how information extracted from CDR data can be utilized to address societal issues:

- **Disaster response:** Natural disasters give rise to emergency situations, when providing time-sensitive information is crucial for fast allocation of resources, which aids the response and recovery process. In 2010, a research team led by the Karolinska Institute⁴ in Sweden showed that CDRs can be used to direct emergency aid by analysing mobile phone records covering the time period when people are fleeing natural disasters.
- **Health:** Researchers have also used CDRs together with a simple disease transmission model based on infection prevalence data, and in doing so were able to map routes of disease dispersal. By analysing the regional travel patterns of millions of mobile subscribers, researchers were able to map the specific locations where disease had a higher probability of spreading.
- **Socio-economics:** CDRs can provide a proxy indicator for assessing population census, regional poverty levels, and can valuably augment national surveys in estimating changes associated with a growing economy.
- **Transportation:** Rapid urbanization in developing countries has increased pressure on infrastructure such as road networks. Roads and public transportation systems become saturated, and people lose a great deal of time traveling from home to work, which in turn has a collective economic cost. By analysing CDR data, scientists can map new routes to decongest crowded roads, which would reduce travel time.

Private sector impact (mobile phone operator)

- **Evaluating patterns of telephone use:** Many usage patterns can be extracted from CDR data. One example is the call behaviour of people in different demographic groups. This includes both spatial and temporal data such as number of calls at each time period, call average, call duration for identified user groups (such as male, female, student, worker, housewife), location (where people in the same or other groups tend to make voice or data calls), minutes of usage per user on average, local call percentage, long distance call percentage, roaming percentage, idle period local call percentage, idle period long distance call percentage, idle period roam call percentage. The requirement is that anonymized CDR data maintains links to information of the user group.

Such information could help operators to adapt their service to specific groups of people as well as specific locations, maintain customer satisfaction with incentives, and also target customer support, for example for those who adopt specific services and features.

- **Build customer profiles:** Operators can build customer profiles from the patterns and then construct a pricing structure to maximize efficiency, whether it's based on mobile user data or customer service calls, the data available in call detail records can help to improve services and promote opportunities or ways to shorten time on call.
- **Sales forecasting:** Customers make calls or use services, and operators can improve services and network efficiency by predicting service usage. Usage analysis can determine a strategy of planned obsolescence or figure out complimentary services. Forecasting also looks at the number of customers, market-share, can predict usage and estimate revenue from groups of customers.

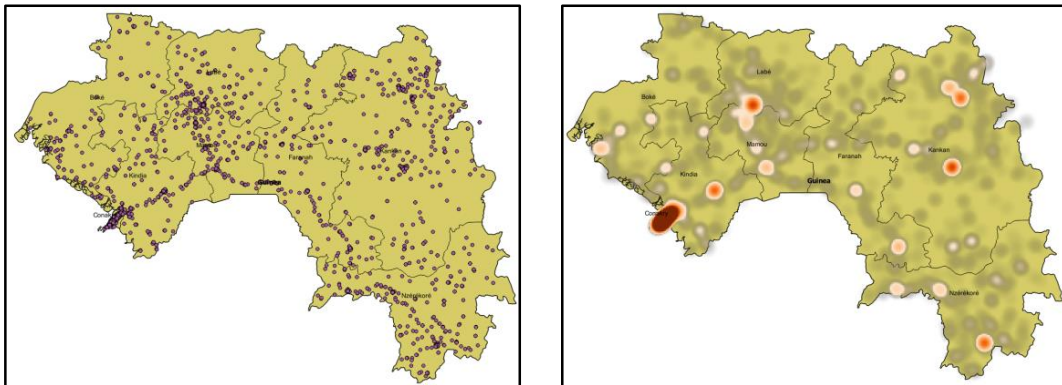
⁴ <http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001083>

- **Identifying fraud or overuse on mobile phone:** Through the use of historical information of CDRs, operators can spot unusual calling behaviour and over-use behaviour, and can help operators to limit or stop related services, and alert customers.
- **Plan for changes in resource needs:** The usage data derived from CDRs can help to plan for changes in resource needs such as to add a gateway or a media server in a certain location. It can be used to measure quality of service by capturing information on packet loss, latency, or jitter. It can also be used for network planning for example events that may result in base station capacity issues.

5 CDR data sets and basic statistics

For the Guinea project, two months of CDR data – from June to July 2015 – were collected in Guinea to demonstrate how dynamic population movements could be estimated. Data was prepared by the mobile network operators in Guinea: Cellcom, Intercel, MTN and Orange. This meant that the data included the majority of the mobile Guinea subscriber population. Subscriber privacy issues were reflected by the creation an anonymous set of information. The unique identification number was replaced using a cryptographic hash algorithm that generated new random numbers and there is no way to rebuild the original identities. Figure 5.1 depicts cell towers and their coverage area. The density map on the right side shows a high concentration of cell towers in Conakry city.

Figure 5.1: Cell tower location at Guinea



Source: ITU

Figure 5.2 illustrates the basic information obtained in CDR data from the mobile network operators. Data specifications were prepared for the operators requesting data in a specific format. Not all operators prepared the data in the specified format, and in these cases, data pre-processing was necessary. In addition, due to the absence of IMSI code in Intercell data, the caller number was used instead.

Figure 5.2: CDR data set

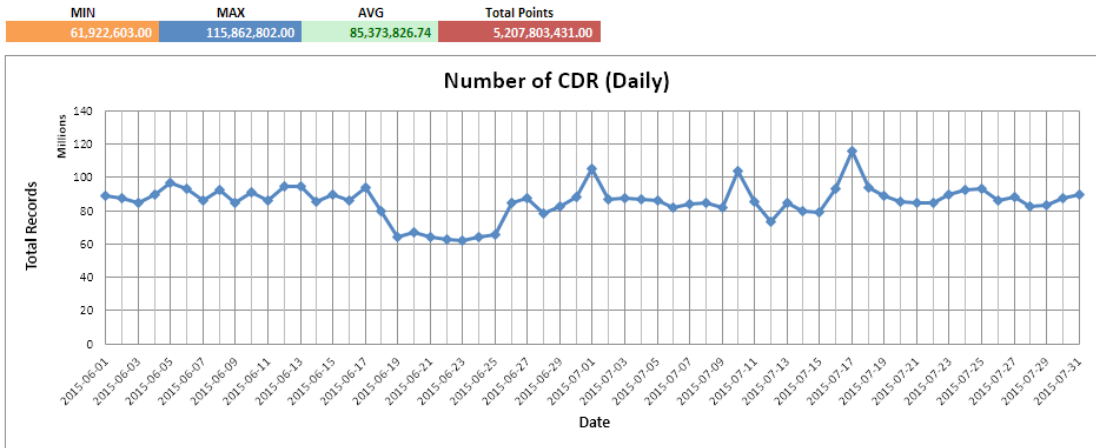
NO	Country	MNO	Voice SMS									
			Data Size	Data Period	Total record	IMEI (caller)	IMSI (caller)	Caller No	Time stamp	LAC/cell ID	Duration	Incoming Cal/Outgoing Call
1	Guinea	MTN	303 G	June, July 2015 (No June 19-25)	1,003,898,498	O (6,979,708)	O (2,699,233)	O (2,660,893)	O	O	O	-
2		Orange	1,100 G	June, July 2015	3,915,651,949	O (10,386,715)	O (5,265,388)	O (10,285,506)	O	O	O	-
3		Cellcom	53 G	June, July 2015	279,441,946	O (3,379,974)	O (1,862,923)	-	O	O	-	-
4		Intercell	2.1 G	June, July 2015	8,811,038	-	-	O (99,531)	O	O	O	-

Source: ITU

CDR basic statistics

CDR data was collected from the mobile network operators and the total activity generated an average of over 85.4 million records a day. Figure 5.3 illustrates the daily call activity.

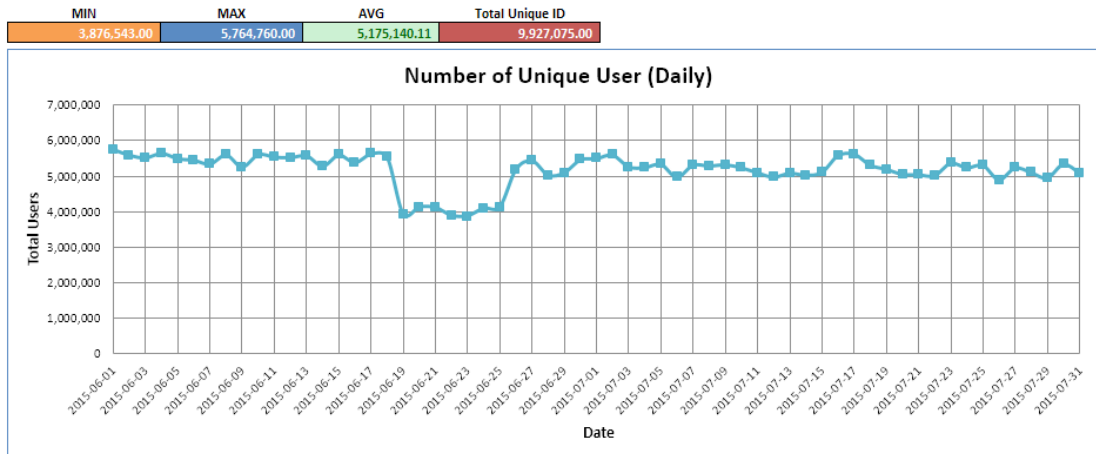
Figure 5.3: Daily call activity



Source: ITU

As seen in Figure 5.4, the number of calls is lower on some days and higher on other days, which may correspond to some important event. Daily activity averaged approximately 5 175 140 users and the CDR data per user averaged 16.47 per day.

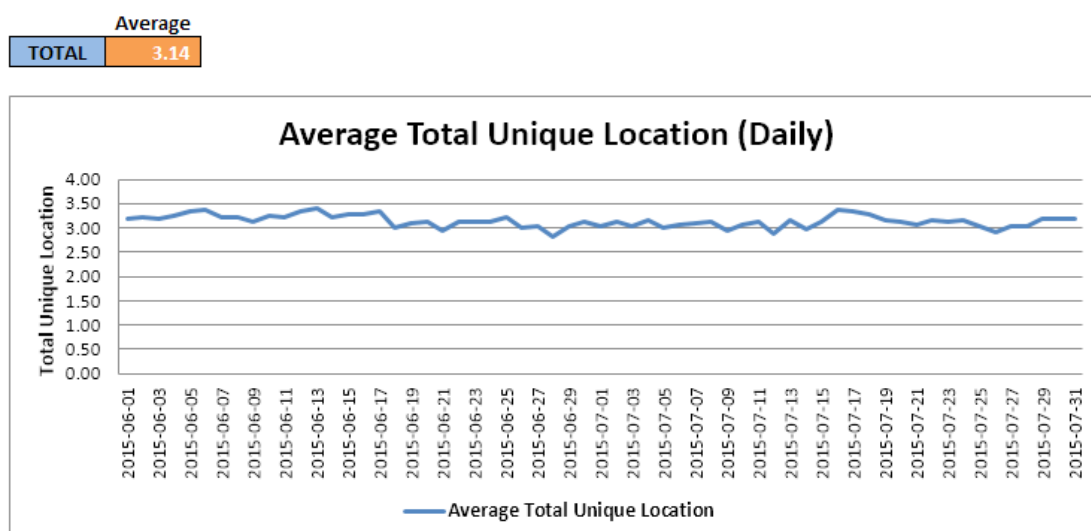
Figure 5.4: Daily unique user



Source: ITU

Figure 5.5 shows the average total unique location of people reached 3.14, which means that people visited about three locations in a day.

Figure 5.5: Average total unique location (daily)



Source: ITU

6 CDR data: Limitations and difficulties

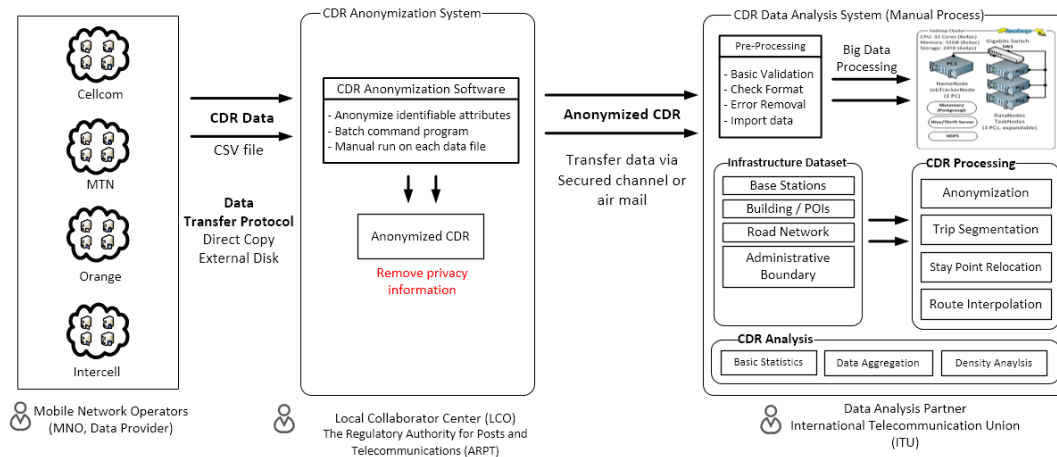
Even though applying CDR data can provide many benefits in various sectors and applications as previously described in section 4, there are some obstacles and difficulties that limit the use of CDR data.

- **Privacy concerns:** When mobile phones transmit voice they also collect and record other information, not only inherent to CDRs but also user specific and personal information, and this creates new challenges in terms of the conflict between technology innovation and rights to privacy.
- **Accuracy:** One major factor that could impact continued advancement of local analysis on these mobile devices is accuracy of their geolocation or estimated position. Almost all CDR data from telecommunication networks use the base tower location to infer the geographic location of the devices. In most cases, the accuracy from this method only varies from 50 to 300 metres in dense urban environments. To mitigate the accuracy issue, the report applies an estimation method of people stay locations and movement routes by using digital map data including POI (Point of Interest) and transportation network data.
- **Availability of data:** In most countries, the research and study of mobile phone data is limited to the availability of data from operators. Although datasets have become available in recent years and have opened the possibility for researchers to carry out large-scale urban and social impact analysis, both support from the mobile industry and data availability are still very limited.
- **Data discontinuity:** Call detail records are generated when people use their mobile phone, creating a lack of continuity in consecutive access points in CDRs, creating a core user location problem when analysing the data. The discontinuity issue is tackled by the estimation method of people stay locations and movement routes, described above. In addition, and despite industry-wide formatting standards, the variation of data format requires much data pre-processing and cleaning before analysis.

7 Methodology

7.1 Overview of CDR data analysis (including stakeholders)

Figure 7.1: Overall concept of CDR data analysis



Source: ITU

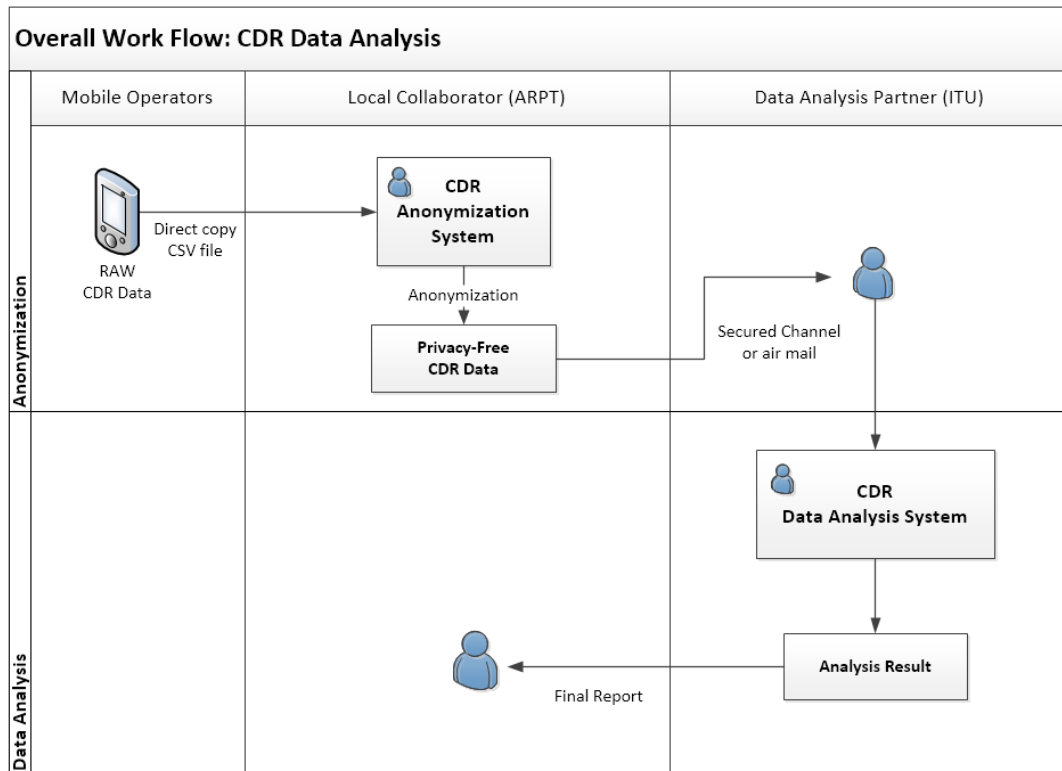
The process of CDR analysis requires a certain number of steps that could eventually be automated, including the use of anonymization software, which has been developed and shared to data providers to remove personally identifiable information from their data sets, so that the people remain anonymous. Data transfer was also carried out manually online using secure channels or offline using external disks. Details of each step and module is described below:

- **Mobile network operator (MNO):** Data is collected from operators. In this case, the mobile operator collects position data of the mobile device stored on its server. Data is then exported and transferred to local collaborator centres (LCOs) for further processing, generally provided in a compressed CSV file format. For the purposes of this project, data transferring was carried out via direct copy.
- **Local collaborator centre (LCO):** This country specific unit stores and sanitizes data (eliminates the risk of personal data disclosure). Usually, this role is carried out by the regulator or mobile operator licence provider, which is also in charge of transferring anonymized data to a designated data analysis partner.
- **Data analysis partner (DAP):** The role of the DAT is to keep and maintain all sanitized CDR data received from operators through the LCO, and is in charge of processing and analysing CDR data:
 - **CDR anonymization:** This function handles the anonymization process on CDR data. The LCO retrieves raw CDR data in csv format and manually processes it to remove all privacy related information. The CDR data can then be transferred to the data analysis partner. In this project, a macro-programme was run (a command line application) once certain parameters such as path of input, path of output, seed data and CDR format parser had been set.
 - **CDR data analysis:** This function collects all CDR data from the LCO, manually checks the data and imports it to a big data platform before it carries out deeper analysis. This analysis by DAP staff incorporates many modules including: pre-processing, big data processing, and CDR processing. CDR data processing is explained in detail in the next section.

7.2 Overview of CDR data analysis work flow

Figure 7.2 describes the overall work flow of the process from anonymization of raw CDR data to analysis.

Figure 7.2: Overview of CDR analysis work flow



Source: ITU

7.3 CDR data specification

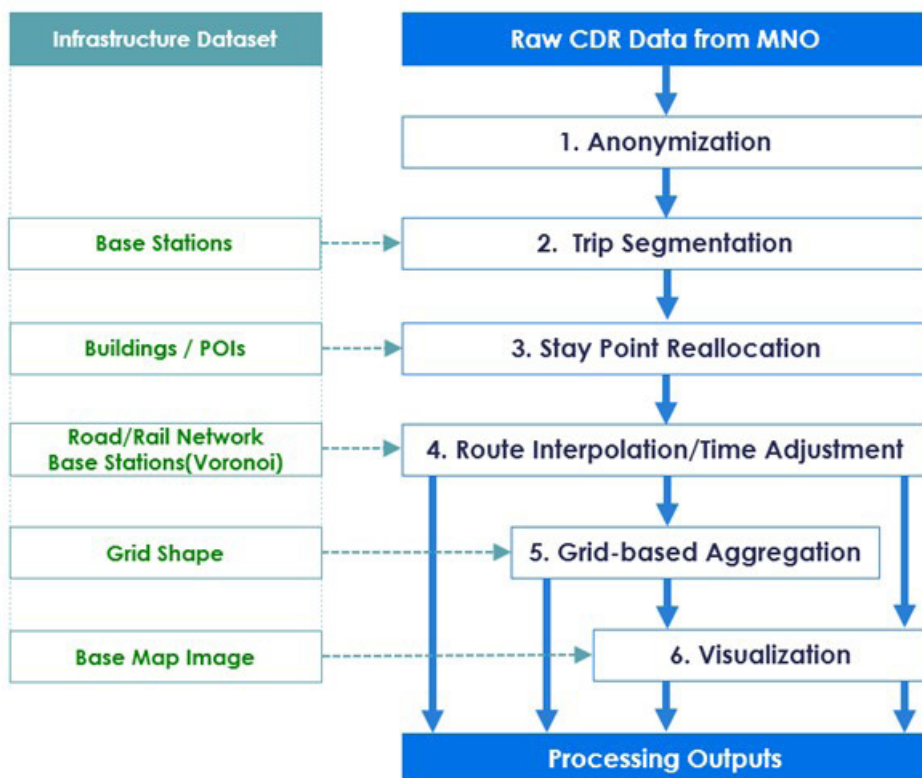
The operator is asked to provide the following standardized information:

- International mobile equipment identity (IMEI) of both calling and called party is:
 - unique for each device;
 - will be irreversibly encrypted (using hash function);
- International mobile subscriber identity (IMSI) of both calling and called party is:
 - unique for SIM (subscriber identification module) card;
 - will be encrypted using Hash function (irreversible).
- Time stamp of call-start and call-end format is YYYY-MM-DD HH24:mm:ss (e.g. 2016-07-16 12:05:22).
- Base station calling party identity and coordinates:
 - LAC, cell identity;
 - longitude, latitude.
- Base station called party identity and coordinates:

- LAC, cell identity;
- longitude, latitude.
- Mobile phone number of calling party: irreversibly encrypted using hash function.
- Activity type:
 - voice, SMS, data;
 - 2G, 3G, LTE.

7.4 CDR processing procedure

Figure 7.3: CDR processing procedure

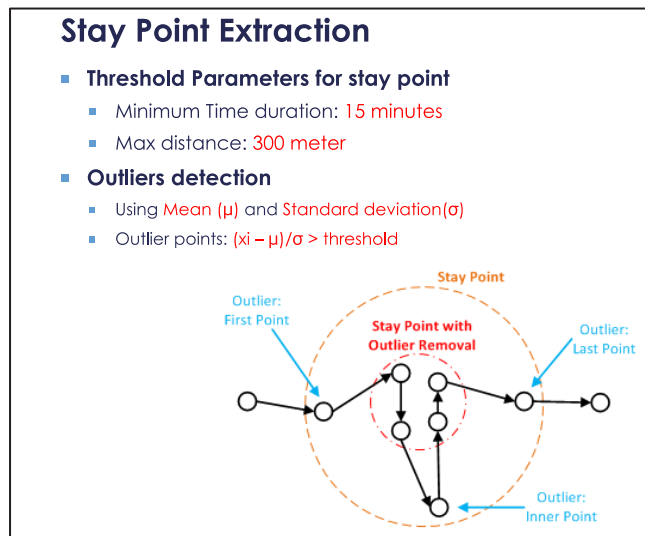


Source: ITU

- 1 Anonymization (hash⁵):** Any identifiable attributes are anonymized with a cryptographic hash function (SHA-256) to protect user privacy. Original attributes will be removed and replaced with randomly generated hashing key.
- 2 Trip segmentation:** Extract stay points from anonymized CDR data, and divide move/stay segments. Figure 7.4 explains how stay points are extracted by applying parameters and thresholds to CDR data.

⁵ Hashing works by running a program that can take text input and turn it into another unique but random value.

Figure 7.4: CDR process of stay point extraction



Source: ITU

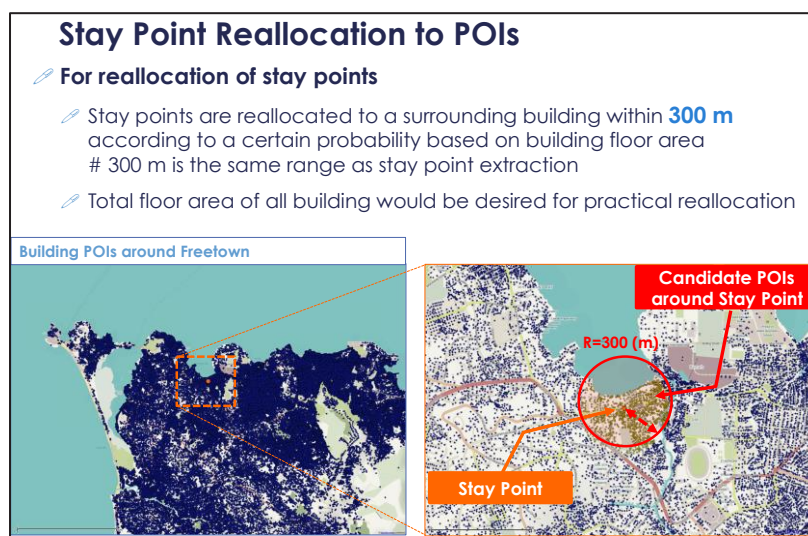
3 Stay point reallocation: This step reallocates stay points (trip origin and destination) to surrounding points of interest (POIs) – with a degree of probability – and fills gaps between stay/move segments. POIs are regarded as surrounding a certain cell tower if they are closer to that cell tower location than to any other (Voronoi tessellation).

Reallocation is necessary because CDR location data is based on cell tower location, which means that all users in the same area have the same location.

Reallocation can make the distribution of people more realistic because the area distribution of POIs to which people are reassigned – rather than cell tower locations – are where people are more likely to have been, such as shopping areas, residential houses, and villages.

A new dataset of POIs was constructed for this process by collecting data from the distribution of buildings from open access Internet data (see Appendix 2). Figure 7.5 shows how POIs are distributed in a city. Areas in blue indicate building POIs with extracted stay points, where location information originally based on cell tower location are reallocated.

Figure 7.5: Distribution of POIs and process of stay point reallocation

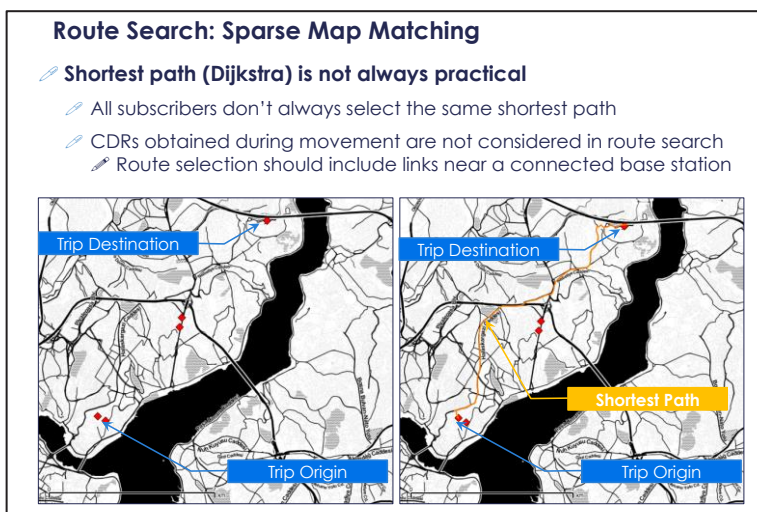


Source: ITU

4 Route interpolation: Routes between every pair of reallocated stay points (building POIs) are interpolated by searching the shortest path between the POIs (trip origin and destination). Because the time log of each record in CDR data does not coincide with the start and end time of a mobile phone user’s trip, the time of trip occurrence is adjusted to match a certain probability distribution. Road network data was created for this process using Open Street Map (OSM) data (see Appendix 2), based on the travel time for the shortest path.

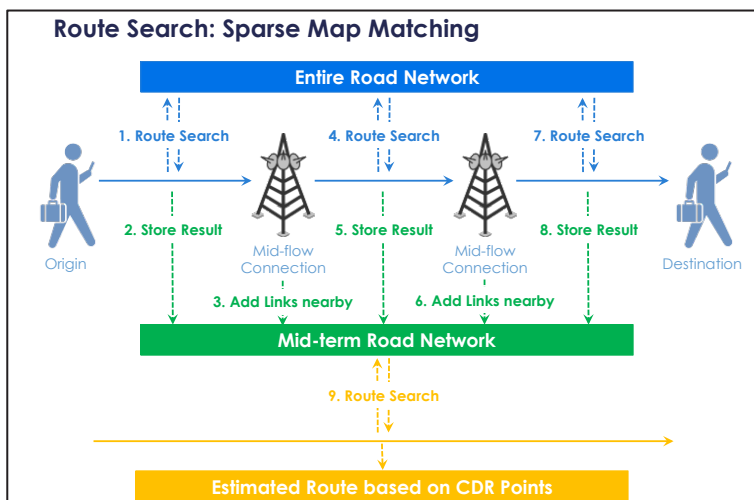
Figure 7.6 shows how a shortest path connects a trip origin and destination. Red dots indicate the location of reallocated stay points and yellow nodes indicate a shortest path for the origin and destination. Some intermediate points, which are not stay points, are not on the shortest path because only the origin and destination of a trip is displayed. An algorithm selects the shortest path that connects not only the origin and destination but also intermediate points. A framework of the process of shortest path search is illustrated in Figure 7.7 and Figure 7.8 shows a route, which also goes through intermediate points.

Figure 7.6: An example of a shortest path search result



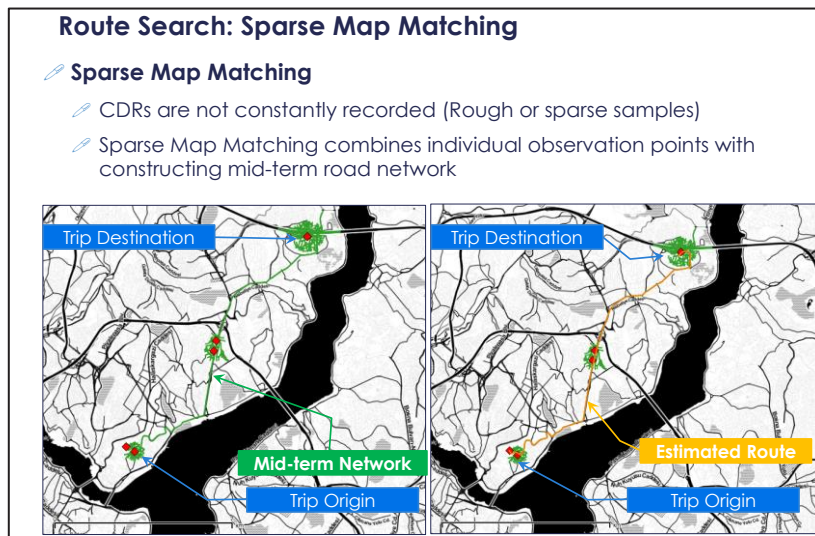
Source: ITU

Figure 7.7: Conceptual framework of the process of shortest path search



Source: ITU

Figure 7.8: Distribution of POIs and process of stay point reallocation



Source: ITU

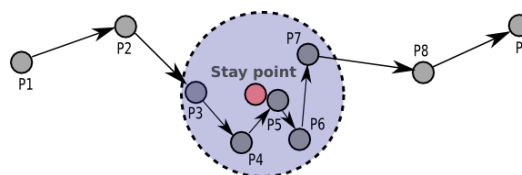
- 5 **Grid-based aggregation:** This refers to the aggregate measurement of population density in a certain grid (1 km) and time duration (1 hour).
- 6 **Visualization:** This creates an animated projection (Mobmap generated film⁶) of disaggregate/ aggregate movement of people (off-line).

7.5 Interpolation format specification

In this section, some basic definitions of trajectory data are explained. It also includes basic terminology for trajectory data mining.

- **Trajectory data:** Is a sequence of time-stamped points, $P = (p_1, p_2, \dots, p_n)$, where $p = (\text{ID}, \text{time}, \text{latitude}, \text{longitude})$ and $n = \text{a total number of points}$.

Figure 7.9: Trajectory data and stay point



Source: ITU

- **Stay point:** This is a geographical reference to a place where a user stayed over a time threshold (tt) within a distance threshold (dt). In a trajectory, stay point is characterized by a set of consecutive points $P = \{p_m, p_{m+1}, \dots, p_n\}$, where $\forall m < i \leq n, \text{Distance}(p_m, p_i) \leq dt, \text{Distance}(p_m, p_{n+1}) > dt$ and $\text{Time Interval}(p_m, p_n) \geq tt$. Therefore, $s = (x, y, t_a, t_l)$, where x, y are a centroid location of those points in a stay point.
- **Location history:** This refers to an individual's location history (h) and is represented as a sequence of stay points they have visited with corresponding transition times:

⁶ <http://shiba.iis.u-tokyo.ac.jp/member/ueyama/mm/>

$$h = \langle s_0 \xrightarrow{\Delta t_1} s_1 \xrightarrow{\Delta t_2} \dots \xrightarrow{\Delta t_{n-1}} s_n \rangle$$

Where s_i is a stay point and Δt_i is the time interval between two stay points.

- **Significant places:** For each individual, significant places are the locations where they often visits in the life activity such as home, work place and supermarket.
- **Trip point:** Is a record attached with trip properties that represent a sequence of GPS points with same activity such as stay or move as shown in Figure 7.10. The properties include time duration, distance, type, transport mode and so on.
- **Sub trip:** This represents a group of sequential trip points from one stay point to the next stay point such as stay-> walk-> train-> walk-> stay. It also depicts an activity of user such as going out from home to office.
- **Trip:** This is a set of time sequential trip points or sub trips in a day. A trip must contain at least one sub-trip. It indicates overall sequential activity in a day of user. Normally, one trip contains multiple Trip Points of STAY and MOVE. The example is shown in Figure 7.10.

Figure 7.10: Time sequential trip points

Trip { Sub Trip }	1, TripPoint [name=STAY, mode=STAY, Time=00:00:15-08:03:50
	2, TripPoint [name=MOVE, mode=WALK, Time=08:03:50-08:18:30
	3, TripPoint [name=MOVE, mode=TRAIN, Time=08:18:30-08:48:38
	4, TripPoint [name=MOVE, mode=WALK, Time=08:48:38-10:56:15
	5, TripPoint [name=STAY, mode=STAY, Time=10:56:15-21:16:29
	6, TripPoint [name=MOVE, mode=WALK, Time=21:16:29-23:20:25
	7, TripPoint [name=MOVE, mode=TRAIN, Time=23:20:25-23:35:57
	8, TripPoint [name=MOVE, mode=WALK, Time=23:35:57-23:45:58
	9, TripPoint [name=STAY, mode=STAY, Time=23:45:58-23:55:25

Source: ITU

Table 1 gives sample data of trajectory data that includes basic terminology and required information and format used in trajectory data mining. Interpolation results are packed as trip data separated for each user and date. Output results contain 11 columns including user ID, date, trip sequence, mobility type, transportation mode, total distance, total time, start time, end time, total points and point lists. Output will be exported to a CSV file (comma-separated values), and values of each column is given.

Table 1: Sample CSV file data

Output results	CSV file column values
1 User Id	<ul style="list-style-type: none"> Column name: UID Unique for each device Encrypted using Hash function, irreversible
2 Date	<ul style="list-style-type: none"> Column name: DATE Date format: yyyy-MM-dd Example: 2015-12-31
3 Trip Sequence	<ul style="list-style-type: none"> Column name: TRIP_SEQUENCE Order of sub trip in a day, start from 1
4 Mobility Type	<ul style="list-style-type: none"> Column name: MOBILITY_TYPE Value: STAY or MOVE
5 Transportation Mode	<ul style="list-style-type: none"> Column name: TRANSPORT_MODE Indicate mode of transportation of corresponding sub trip Value: STAY, WALK, CAR
6 Total Distance	<ul style="list-style-type: none"> Column name: TOTAL_DISTANCE Total travel distance of sub trip in meter
7 Total Time	<ul style="list-style-type: none"> Column name: TOTAL_TIME Total travel time of sub trip in second
8 Start Time	<ul style="list-style-type: none"> Column name: START_TIME Indicate start time of sub trip Format: HH24:mm:ss Example: 23:20:00
9 End Time	<ul style="list-style-type: none"> Column name: END_TIME Indicate end time of sub trip Format: HH24:mm:ss Example: 23:20:00
10 Total Points	<ul style="list-style-type: none"> Column name: TOTAL_POINTS Indicate total number of point data in sub trip
11 Point lists	<ul style="list-style-type: none"> Column name: POINT_LISTS List of point data in sub trip Format: No. time latitude longitude; No. is order number start from 1. Time: yyyy-MM-dd HH24:mm:ss Latitude and longitude in decimal format. Each point is separated by “;” Example: 1 2015-06-06 13:53:23 6.373743 -10.772951

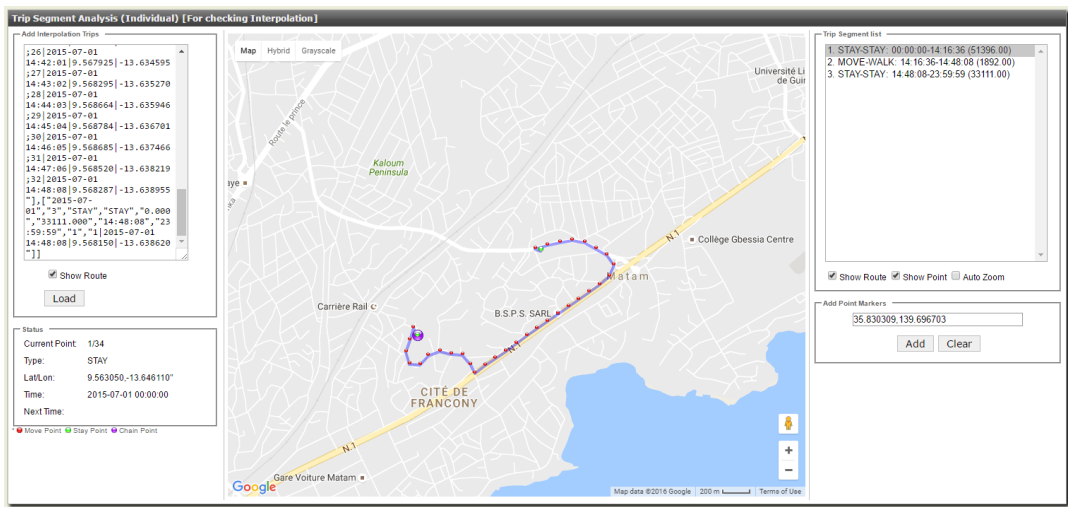
Note: In the present project, the Move segment interpolation is set to 1 minute intervals.

Trip segment analysis for checking interpolation result

```

UID, DATE (yyyy-MM-dd), TRIP_SEQUENCE_ID, MOBILITY_TYPE, TRANSPORT_MODE, TOTAL_DISTANCE,
TOTAL_TIME, START_TIME, END_TIME, TOTAL_POINTS, POINT_LIST
422a837717,2015-07-01,1,STAY,STAY,0.000,51396.000,00:00:00,14:16:36,1,1|2015-07-01
00:00:00|9.563050|-13.646110
422a837717,2015-07-01,2,MOVE,WALK,2558.719,1892.000,14:16:36,14:48:08,32,1|2015-07-01
14:16:36|9.563562|-13.646353;2|2015-07-01 14:17:37|9.562828|-13.646577;3|2015-07-01
14:18:38|9.562095|-13.646802;4|2015-07-01 14:19:39|9.561415|-13.646590;5|2015-07-01
14:20:40|9.561354|-13.645955;6|2015-07-01 14:21:41|9.561891|-13.645452;7|2015-07-01
14:22:42|9.562174|-13.644745;8|2015-07-01 14:23:43|9.561984|-13.644067;9|2015-07-01
14:24:44|9.561947|-13.643375;10|2015-07-01 14:25:45|9.561363|-13.642889;11|2015-07-01
14:26:46|9.560838|-13.642623;12|2015-07-01 14:27:47|9.561300|-13.642007;13|2015-07-01
14:28:48|9.561745|-13.641377;14|2015-07-01 14:29:49|9.562178|-13.640741;15|2015-07-01
14:30:50|9.562623|-13.640111;16|2015-07-01 14:31:51|9.563067|-13.639481;17|2015-07-01
14:32:52|9.563506|-13.638848;18|2015-07-01 14:33:53|9.563944|-13.638214;19|2015-07-01
14:34:54|9.564382|-13.637580;20|2015-07-01 14:35:55|9.564823|-13.636948;21|2015-07-01
14:36:56|9.565266|-13.636317;22|2015-07-01 14:37:57|9.565705|-13.635684;23|2015-07-01
14:38:58|9.566144|-13.635051;24|2015-07-01 14:39:59|9.566627|-13.634456;25|2015-07-01
14:41:00|9.567308|-13.634174;26|2015-07-01 14:42:01|9.567925|-13.634595;27|2015-07-01
14:43:02|9.568295|-13.635270;28|2015-07-01 14:44:03|9.568664|-13.635946;29|2015-07-01
14:45:04|9.568784|-13.636701;30|2015-07-01 14:46:05|9.568685|-13.637466;31|2015-07-01
14:47:06|9.568520|-13.638219;32|2015-07-01 14:48:08|9.568287|-13.638955
422a837717,2015-07-01,3,STAY,STAY,0.000,33111.000,14:48:08,23:59:59,1,1|2015-07-01
14:48:08|9.568150|-13.638620
    
```

Figure 7.11: Trajectory check (internal use)



Source: ITU

7.6 Hadoop system for data processing

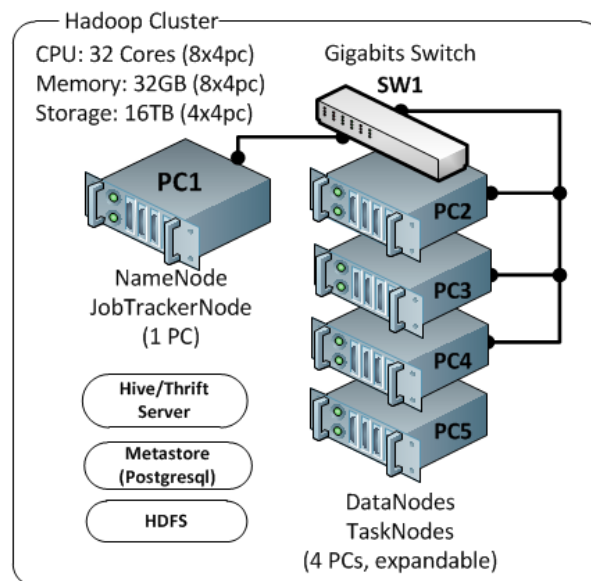
CDR data forms a very large data set that ordinary computers or database systems are unable to process within an acceptable time frame, which is why the Hadoop system was introduced as the main system for data processing for this project. Hadoop is an open source cloud computing software framework for data intensive and distributed application. There are many services and framework under Hadoop umbrella; however, in this project, Hadoop Distributed File System (HDFS) and Hive were used. To set up and use Hadoop for full operation mode, five components were required: NameNode, DataNodes, Secondary NameNode (SNN), JobTracker, and TaskTrackers.

NameNode is the bookkeeper of HDFS; it keeps track of how your files are broken down into file blocks, which nodes store those blocks, and the overall health of the distributed filesystem. DataNodes are the workhorses of the filesystem. They store and retrieve blocks when they are told to (by clients or the namenode), and they report back to the namenode periodically with lists of blocks that they are storing. Secondary NameNode (SNN) is an assistant daemon for monitoring the state of the cluster HDFS and the SNN help snapshots NameNode to help minimize the downtime and loss of data. JobTracker is the liaison between your application and Hadoop. Once you submit your code to your cluster, the JobTracker determines the execution plan by determining the files to process, assigns nodes to different tasks, and monitors all tasks as they are running. TaskTrackers are responsible for executing the individual tasks that the JobTracker assigns and manages the execution of individual tasks on each slave node.

The Hadoop Cluster experiment consisted of five computers with the same specification: Xeon 2.6 GHz, 8 GB memory, and 2x2 TB disk with CentOS 6.0 64-bit for database system, and library-based application. PostgreSQL 9.0.6 with PostGIS 1.5.3 was installed in the system. One computer runs as NameNode and the others as DataNodes and TaskNodes.

Figure 7.12 illustrates the system used for this project: the Hadoop had 32 cores, 32 GB memory and 16 TB storages and could run up to 28 tasks at the same time. The version of Hadoop was 0.20.2, the version of Hive was 0.8.0, and the version of JTS was 1.12.

Figure 7.12: Sample Hadoop Cluster



Source: ITU

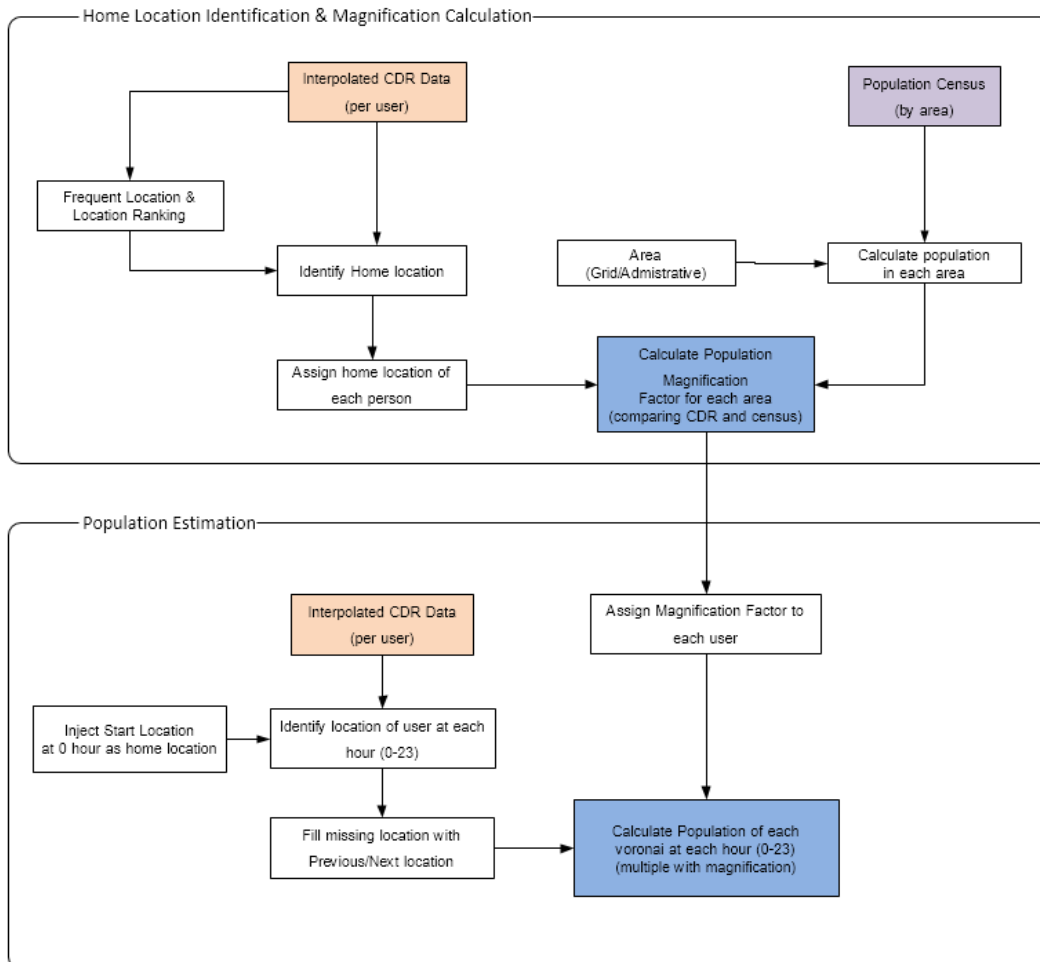
Hive is a data warehousing package built on top of Hadoop, and users should be familiar and comfortable with using SQL to carry out ad-hoc queries, summarizations, and data analysis. Web GUI (interface) and Java Database Connectivity (JDBC) are provided for interacting with Hive by issuing queries in a SQL-like language called HiveQL.

7.7 Dynamic population estimation

Knowing the population distribution rather than the distribution of mobile phone subscribers in any given area during the daytime hours or at any time slice is crucial to calculating movement trajectories and patterns, and the challenge is how to 1) magnify CDR data or the number of subscribers to the real population with population census and to 2) represent trajectory of people or where and when people are located. Magnification process is described in Figure 7.13.

In this process, CDR data is used to extract home locations of mobile phone users in terms of cell tower locations. Extracted home locations of mobile phone users are reallocated using POIs and complementary land use and house distribution data around each cell tower. Then the reallocated mobile phone users are aggregated by each census zone (tracts) of population census. The magnification factor, describing how many people a single mobile phone user represents, can be computed by comparing real population of a census zone with the aggregated number of mobile phone users. For example, at a certain census zone, one mobile phone user may represent 10 people. The magnification factor is also used to calculate dynamic population at each time period.

Figure 7.13: Home location identification with population magnification

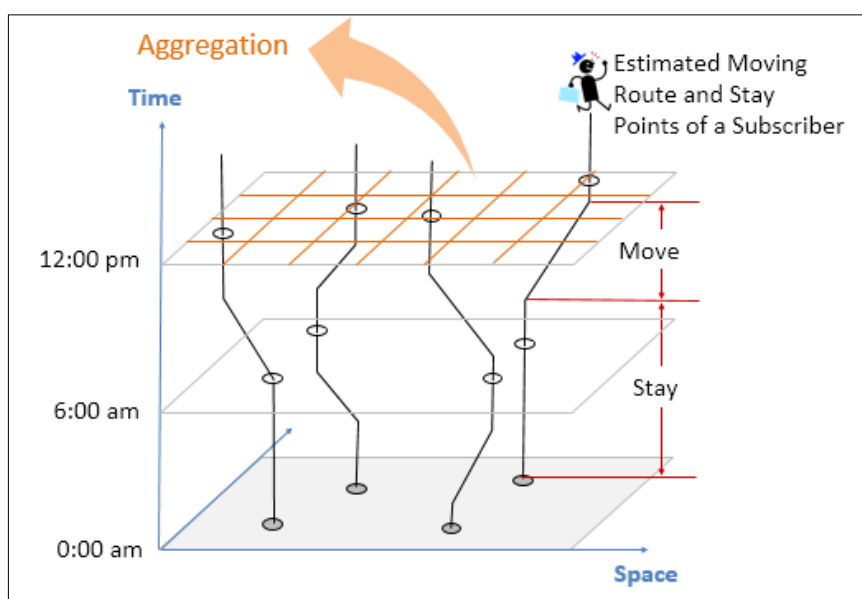


Source: ITU

Dynamic day time population (ambient population), unlike residential population, takes into account the movements of individuals through a given area as shown in Figure 7.14. For instance, no individuals live on a main road, but they do travel on it from time to time. A resident population would show no one living on the road, whereas the ambient population would indicate the presence of individuals based on factors specific to the road.

Through the CDR analysis, home locations and the other stay points, movement routes, and timing of mobile phone users are estimated to fully represent the trajectory, together with the magnification factors. Dynamic population at any time slice at any location (grid-cell) can be computed.

Figure 7.14: On a daily based data aggregation for dynamic population estimation



Source: ITU

For dynamic population estimation, CDR data of each mobile phone is processed to identify location at each time period. Voronoi values will be attached to each hour period of each mobile phone. In the case where there is no call record at some period, previous or next location will be used instead to complete all missing location data. Finally, in combination with the magnification factor of each user, the number of population for each Voronoi and time value are calculated. In the process of calculation, each mobile phone will have its own magnification factor to generate a more detailed reflection of time-based population location and movement.

7.8 Mobile phone and SIM card cloning

Mobile phone cloning⁷ is a technique used to copy private data (identity theft) from one mobile phone onto another phone, which then becomes the exact replica of the original phone. Consequently, while calls can be made from and received by both phones, only the legal subscriber is charged for the bill as the mobile network operator does not have a way to differentiate between the legitimate phone and the 'cloned' phone. So when a subscriber is surprised by an enormous bill, there is a chance that the phone has been cloned. Many mobile phones, be it GSM or CDMA, run the risk of being cloned.

There are several kinds of user specific identity information depending on mobile network system. For a GSM network, every handset has a unique IMEI number and it needs a SIM card inserted to be able to operate. SIM cards can be cloned and put in other handsets, and because inexpensive handsets can often have the same IMEI (to reduce production costs), cloning of SIM cards is generating a serious fraud problem. For CDMA networks, every handset has its unique ESN number (electronic serial number) and MIN (mobile identification number) that are burned into the chip in the handset, and cloning of this type of handset involves replacing these identity numbers.

Cloning can be detected from the duplication of identity where the mobile phone use is transferred from one place to another place at impossible speeds (velocity trap). For example, if a call is first made in Conakry, and five minutes later, another call is made but this time in Mali, about 300 km away, this indicates that there are two phones with the same identity on the network. Mobile network operator reactions to mobile cloning once it is detected is often to shut them all off so that the real customer contacts operator customer services to complain. Call detail records such as identity number, start

⁷ <https://www.movzio.com/howto/cell-phone-cloning/>

call-time, and cell tower location are significant keys to identify cloning behaviour. However, such analysis requires the processing of a lot of data, which requires high performance systems.

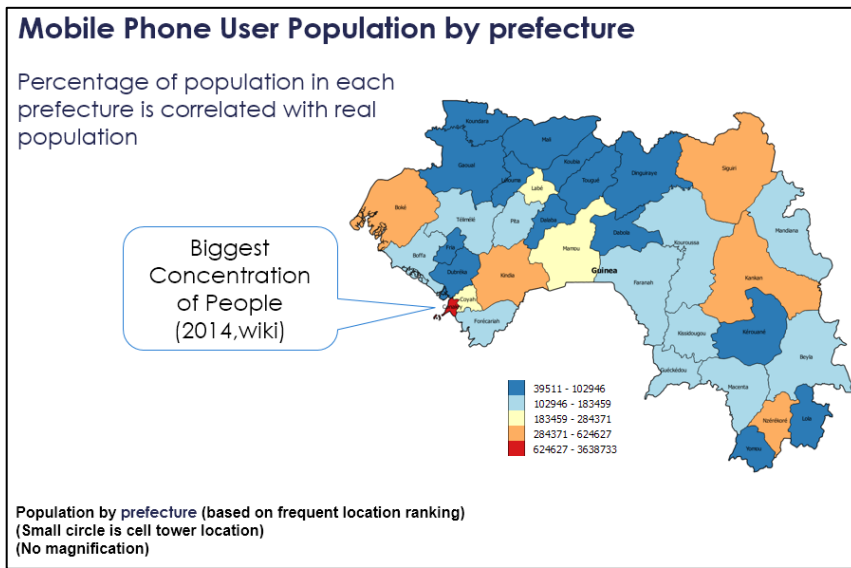
In this project, attempts were made to detect mobile cloning in GSM networks from CDRs using big data technology and spatial processing, and the results are expressed in the section 8.5 of this report.

8 Analysis results

8.1 Mobile phone user population by Prefecture

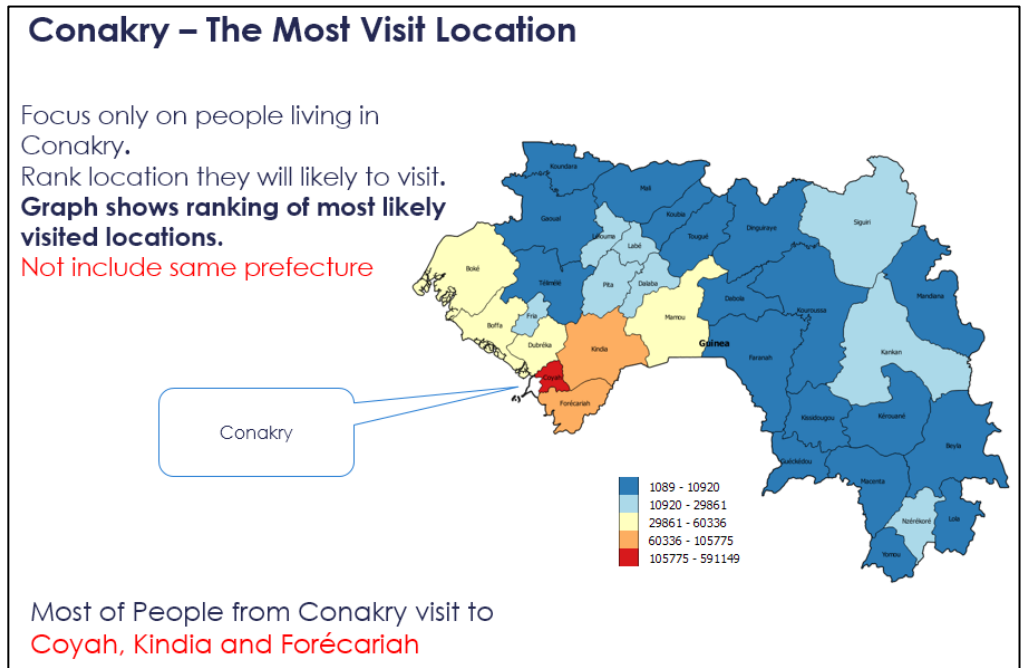
This section discusses the population distribution extracted from CDR data. Figure 8.1 shows the number of mobile phone users estimated from CDR data and population data by the national population census in the Republic of Guinea, which confirms that the proportion of mobile phone users across prefectures correlates strongly with that of real populations obtained from census data. Figure 8.2 presents the most visited locations by the people living in Conakry, revealing that Coyah, Kindia, and Forécariah are the most visited locations. Figure 8.3 shows Coyah is the second most visited location by the same people.

Figure 8.1: Mobile phone user population by Prefecture



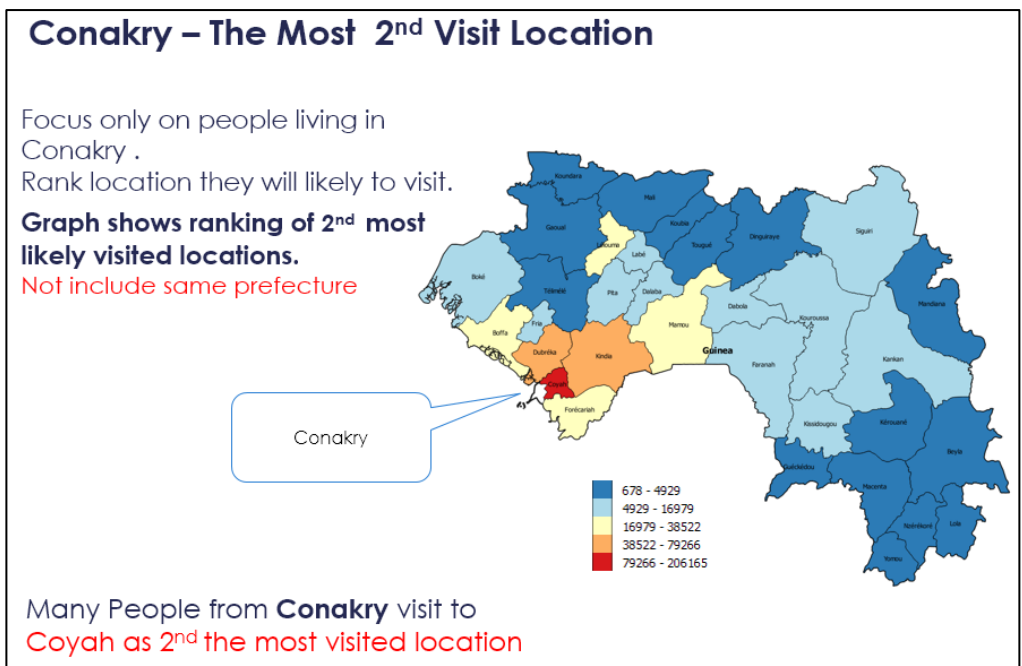
Source: ITU

Figure 8.2: Most visited locations by the people living in Conakry



Source: ITU

Figure 8.3: Second most visited locations by the people living in Conakry



Source: ITU

8.2 People flow visualization using interpolation result

This section provides the visualization of interpolated CDR data through the process described in Section 7. Figures 8.4a and 8.4b show the people flow reconstructed from interpolation results at 6:00 a.m., 9 a.m., 12 p.m., and 3 p.m. In Monrovia, there are less people early in the morning and more people in the afternoon. Interpolation results contain 11 columns including user id, date, trip

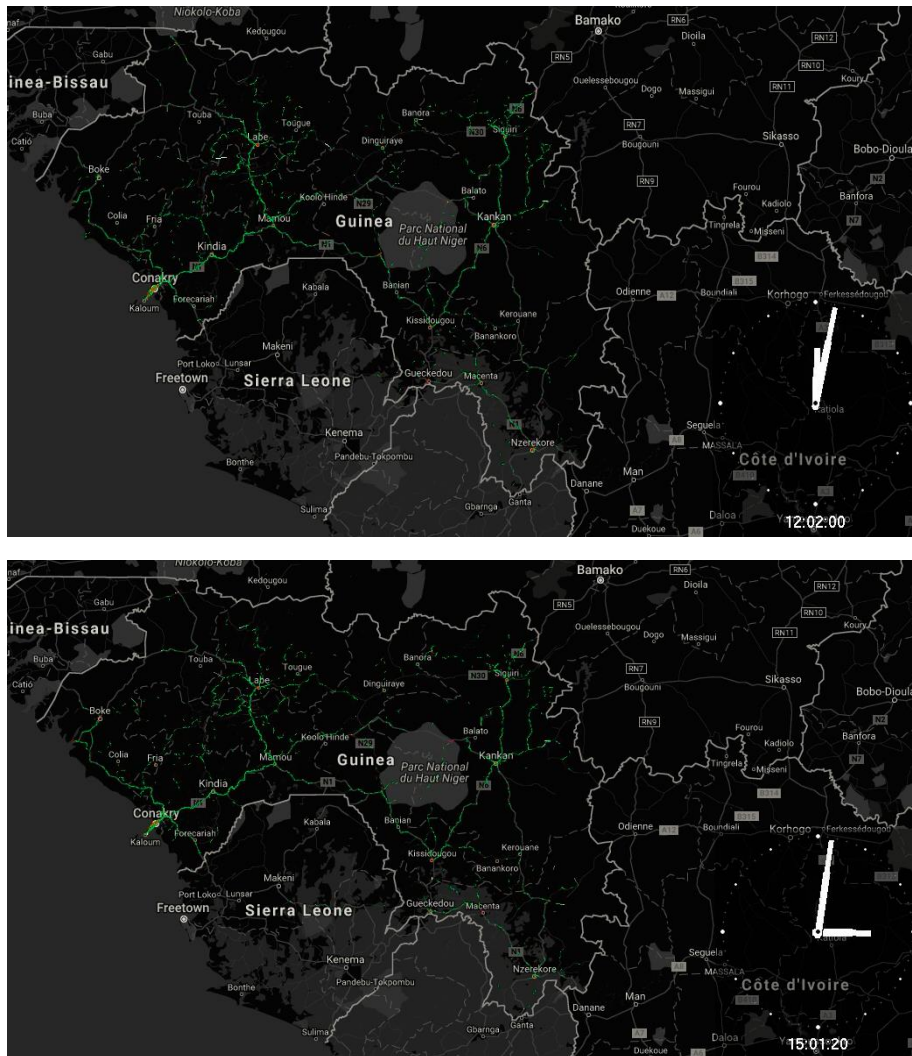
sequence, mobility type, transportation mode, total distance, total time, start time, end time, total points and point lists (see section 7.5.) CDR data is interpolated at one-minute intervals.

Figure 8.4a: People flow at 6 a.m. and 9 a.m.



Source: ITU

Figure 8.4b: People flow at 12 p.m. and 3 p.m.

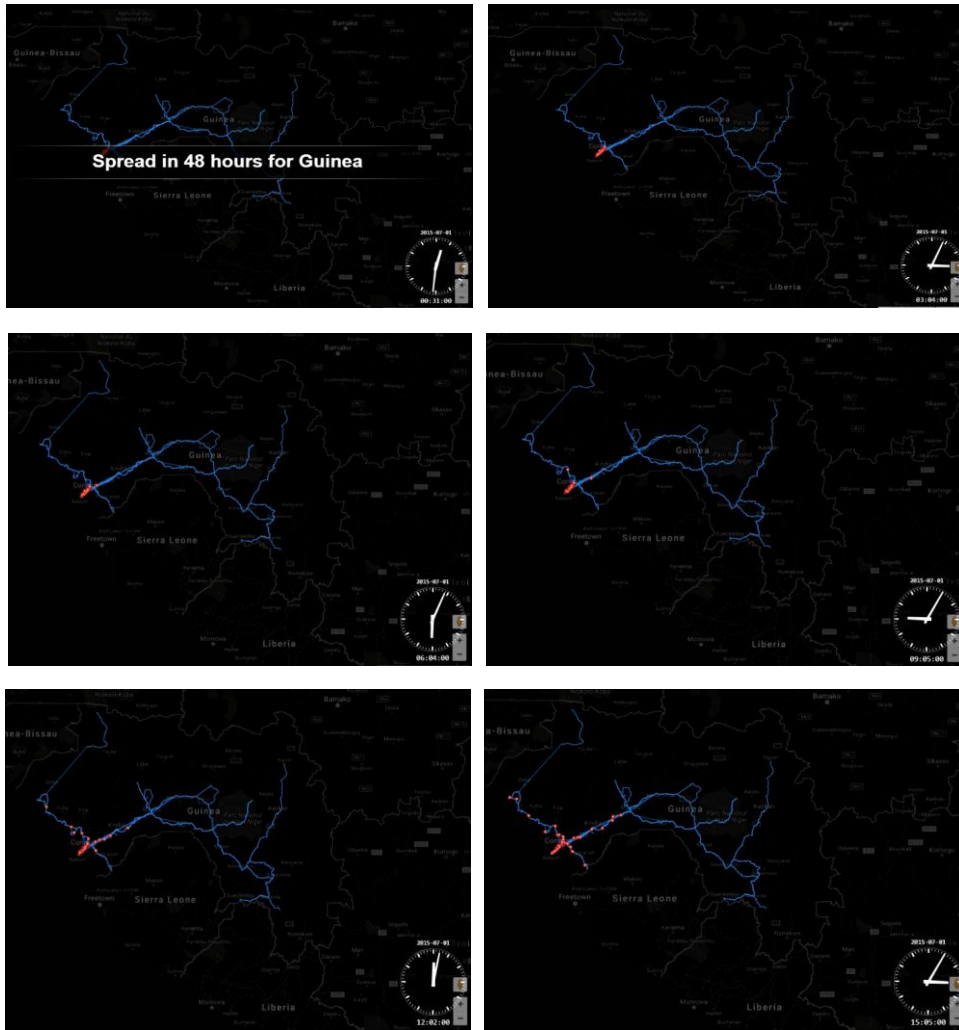


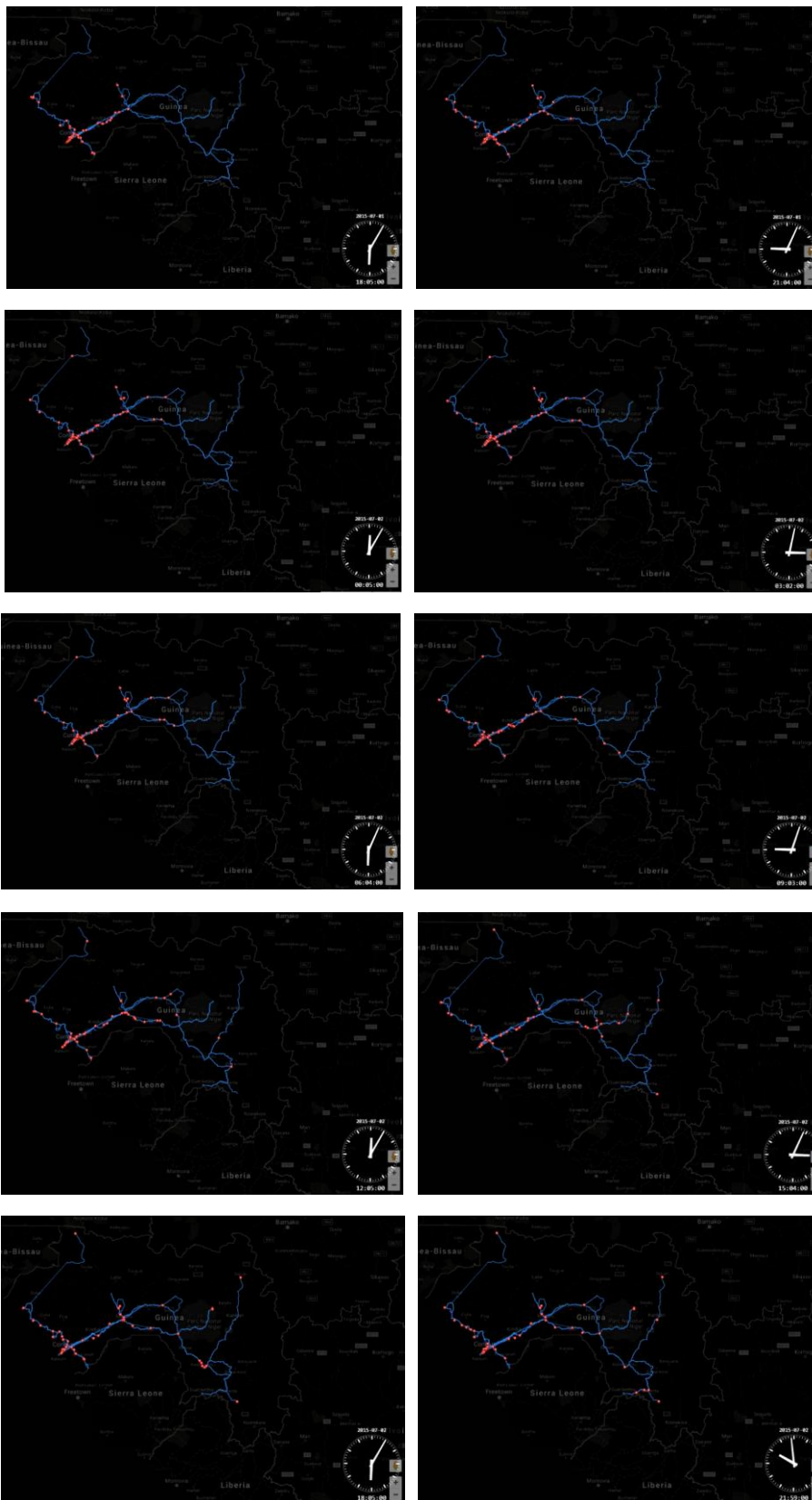
Source: ITU

8.3 Movement of people from one town (over 48 hours)

Figures 8.5 shows how people in a town spread at three-hourly intervals over a 48 hour period. In a situation, where incidents are observed in the town, it is clear how quickly the communicable disease could spread due to the mobility of people.

Figure 8.5: Flow from a specific town (over 48 hours)





Source: ITU

8.4 Tracing people who passed through a hazard area

Figures 8.6a, 8.6b, 8.7a and 8.7b show the distribution of people who passed/did not pass through a hazard area. A hazard area is colored in blue. In Figure 8.6a larger-yellow dots inside blue polygons indicate those who are in the hazard area the hazard area, and smaller dots outside the polygons indicate those are not in the hazard area at the given time. Figures 8.6b, 8.7a, and 8.7b show how those who are in the hazard area (at the time of Figure 8.6a) spread in the following hours.

Figure 8.6a: Distribution of people who are/are not in a hazard area



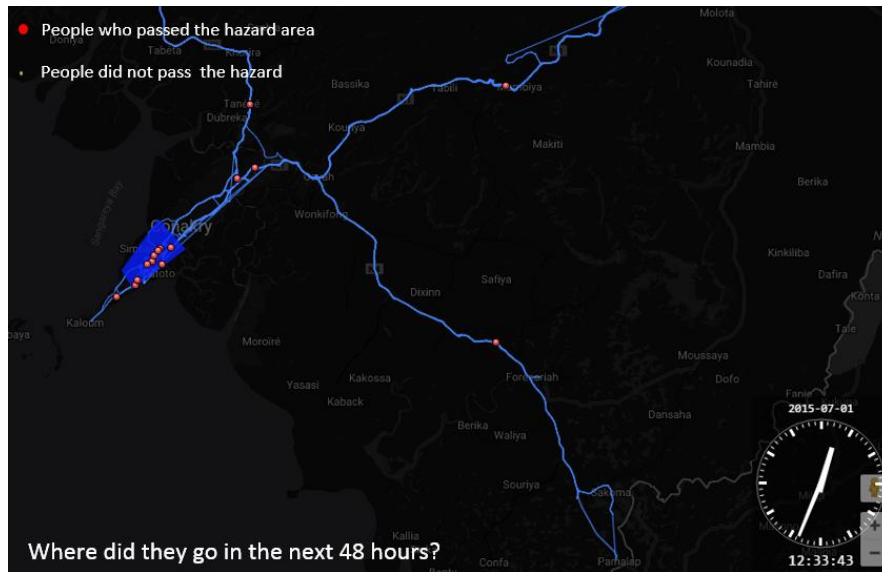
Source: ITU

Figure 8.6b: Trace people who passed through a hazard area (after 10 hours)



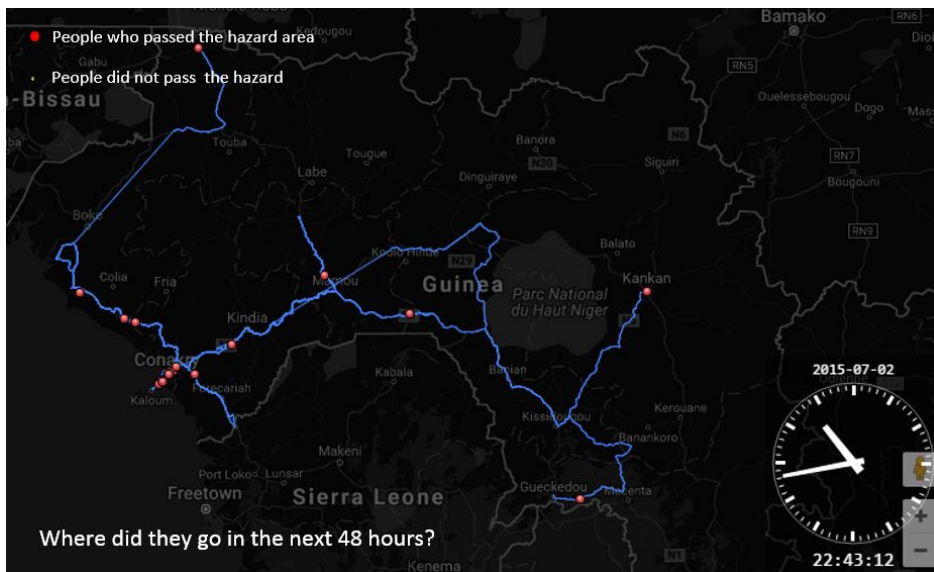
Source: ITU

Figure 8.7a: Where did they go in next 48 hours? (After 12 hours)



Source: ITU

Figure 8.7b: Where did they go in next 48 hours? (After 47.5 hours)



Source: ITU

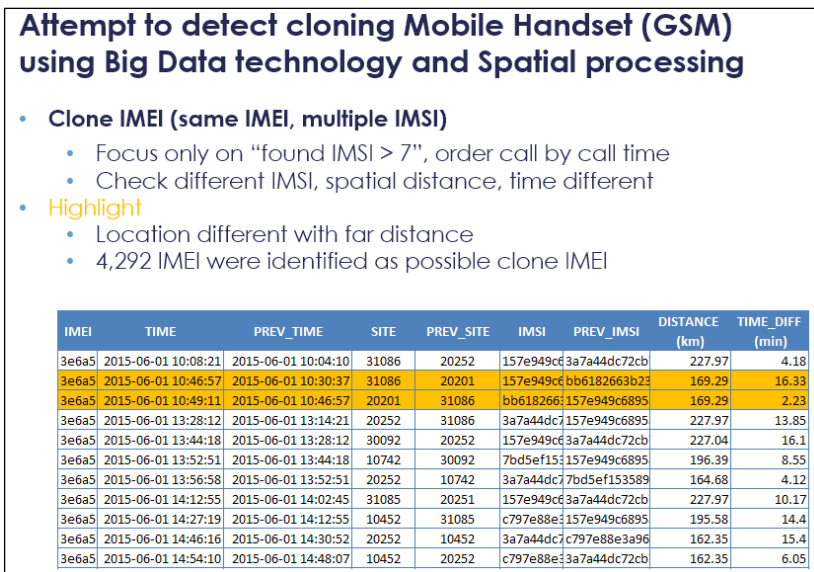
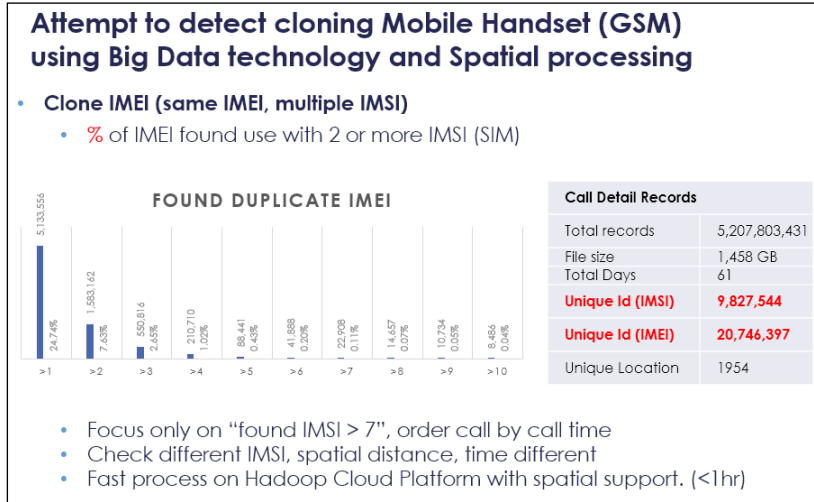
As also observed in Figures 8.7a and 8.7b, population mobility can spread far from the first location in 48 hours, which potentially threatens a very large area in the case of an Ebola epidemic.

8.5 Attempt to detect cloning phones and SIM cards

As mentioned in section 7.8 above, CDR data can be used to detect cloning. Figure 8.8 shows the proportion of IMEIs (a single mobile phone record in CDR data) that are originating call records from multiple SIMs or international mobile subscriber identities (IMSI). As described in Figure 8.8 (upper), in the CDR data used for this project, the number of unique IMEI IDs is almost twice the number of unique IMSI IDs and about 22 980 unique mobile handset numbers were used with more than seven SIM cards (IMSI) indicating probable cases of mobile cloning. For this project, IMEI used with more than seven IMSI were filtered out.

Figure 8.8 (lower) shows two call records from identical IMEI with different IMSIs. Their locations are too distant to have made one call within less than a minute of a second call. Processing CDR data with IMSI and IMEI can consequently help to detect cloning and this method was used for the purposes of this project.

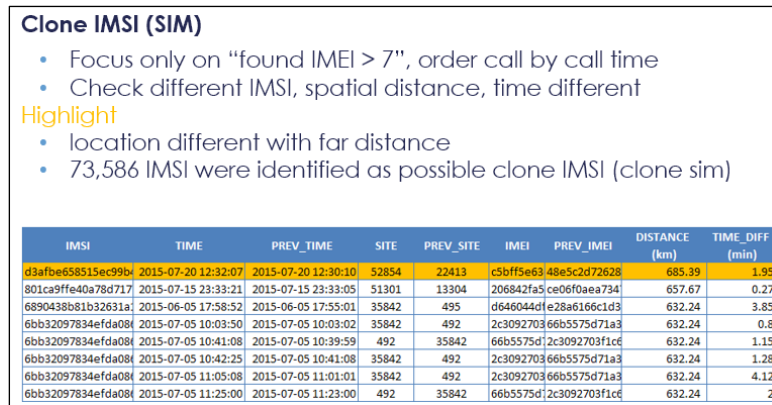
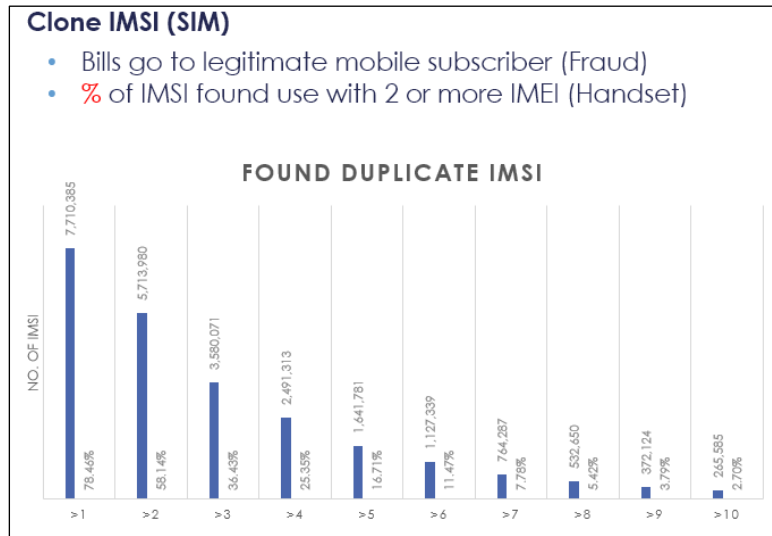
Figure 8.8: Cloning detection (cloning IMEI)



Source: ITU

Figure 8.9 (upper) shows an attempt to detect SIM card cloning: there are more than 7.7 million unique sim cards that are being used in at least two handsets over the same period, and more than 764 287 SIM cards were being used in seven handsets, which can probably be attributed to SIM card cloning. For this project, the focus and filters were on IMSI found on more than seven IMEI.

Figure 8.9: Cloning detection (cloning IMSI)



Source: ITU

Figure 8.9 (lower) shows the record (highlight) from identical IMSI with different IMEIs. Their locations are too distant for two calls to have been made with the same SIM card. From this series of data, 73 586 IMSIs were identified as possible clones (SIM card cloning).

8.6 Transboundary analysis

This section discusses the results of transboundary movement analysis. It demonstrates how CDR data can be used to understand population movements in border areas of neighbouring countries.

There are two types of transboundary population movement:

- **inbound** transboundary movement is when a person returns to the home country from a neighbouring country; and
- **outbound** transboundary movement is when a person leaves the home country to visit a neighbouring country.

Transboundary analysis uses the IMEI as a key, and CDR data from both Liberia and Sierra Leone were merged with data from Guinea to trace transboundary movements. After the merge, each user home country is identified based on the frequency of called locations.

8.6.1 CDR data set with transboundary IMEI

Figure 8.10 illustrates the transboundary information obtained from CDR data including IMEI, which indicated the source of mobile phones being used in neighbouring countries. Three mobile network operators (MNOs), MTN, Orange, and Cellcom, provided CDR data (with hashed IMEI code). However, only MTN IMEI and Cellcom IMEI could be matched. This means that analysis results in the following sections do not include the transboundary movements of Orange subscribers. Based on the data size, the number of Orange subscribers is assumed to be the largest among the given MNOs. So, there is a possibility that some other movement trends, represented by Orange subscribers, were missed.

Figure 8.10: CDR data set with transboundary IMEI of Guinea

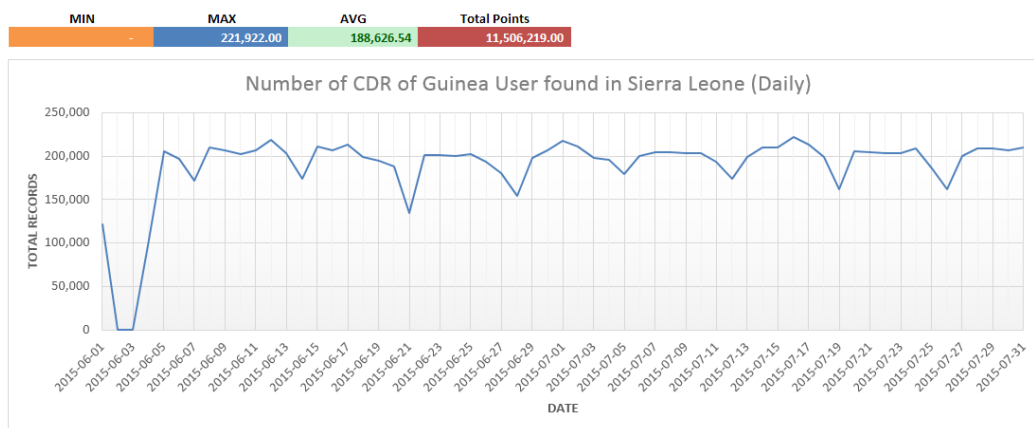
NO	Country	Operator	CDR Data Info				Transboundary (IMEI)		
			Data Size	Data Period	Total record	IMEI	Liberia	Sierra Leone	Liberia & Sierra Leone
1	Guinea	MTN	303 G	June, July 2015 (No June 19-25)	1,003,898,498	6,979,708	10,521	28,578	1,949
2		Orange	1,100 G	June, July 2015	3,915,651,949	10,386,715	0	0	0
3		Cellcom	53 G	June, July 2015	279,441,946	3,379,974	10,877	23,293	1,811
4		Intercell	2.1 G	June, July 2015	8,811,038	-	-	-	-

Source: ITU

8.6.2 Daily activity statistics

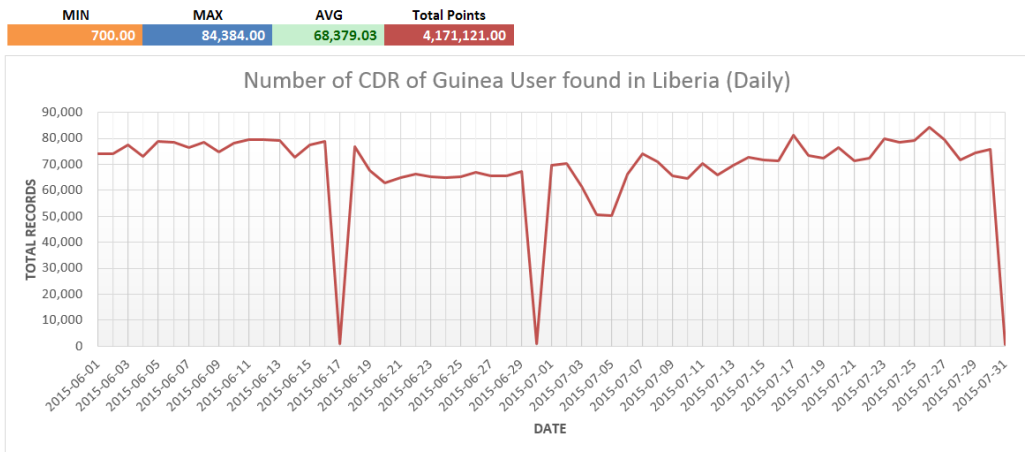
Figure 8.11 and 8.12 illustrate the daily call activity of Guinea based mobile phone subscribers being registered on neighbouring country networks (Sierra Leone and Liberia). Daily call activity averaged approximately 188 626 for Sierra Leone and 68 379 for Liberia, indicating that Guinea subscribers are more than twice as likely to go to Sierra Leone than to Liberia.

Figure 8.11: Daily call activity of Guinea based mobile users visiting Sierra Leone



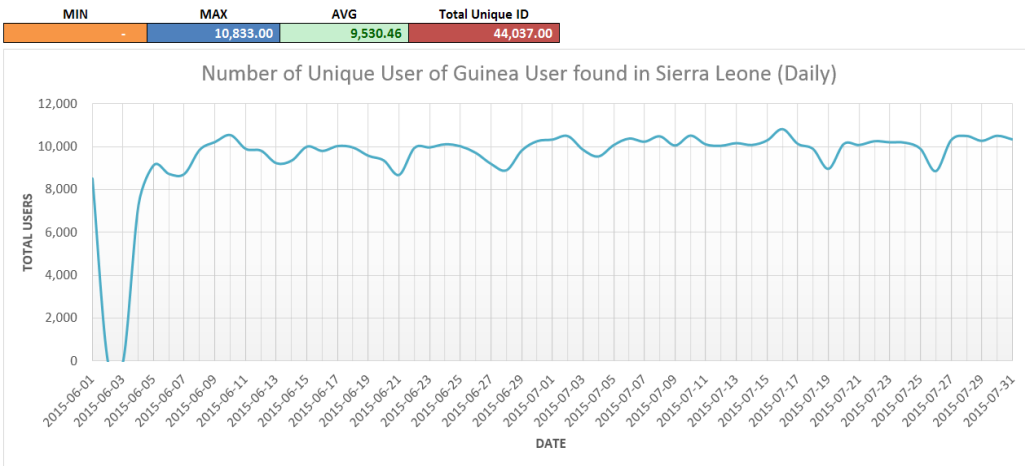
Source: ITU

Figure 8.12: Daily call activity of Guinea mobile users visiting Liberia



Source: ITU

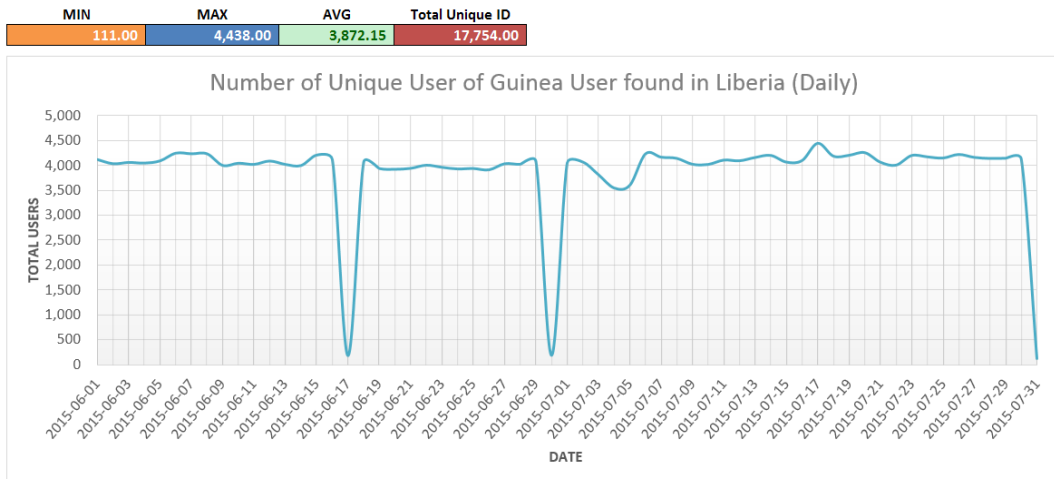
Figure 8.13: Daily unique users of Guinea mobile users visiting Sierra Leone



Source: ITU

Figures 8.13 and 8.14 illustrate the number of daily unique Guinea mobile users visiting Sierra Leone and Liberia, averaging approximately 9 530 users visiting Sierra Leone and 3 872 users visiting Liberia.

Figure 8.14: Daily unique users of Guinea mobile users visiting Liberia

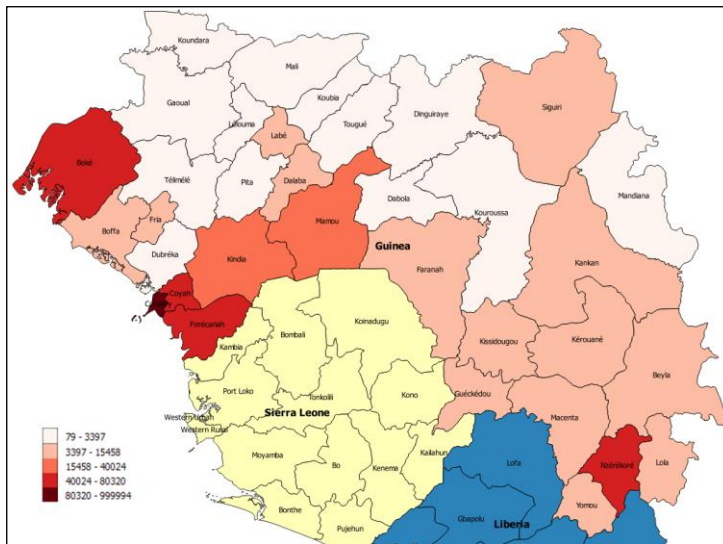


Source: ITU

8.6.3 Inbound transboundary population movement

Figure 8.15 illustrates the footprint of inbound transboundary movements from Sierra Leone and Liberia to Guinea. Footprints were calculated from mobile usage activity of users, which indicated the places they passed through and where high mobile usage activities were recorded. The most visited prefecture is Conakry where the capital city of Guinea is located. Second and third rank of frequent visits are Coyah and Boké.

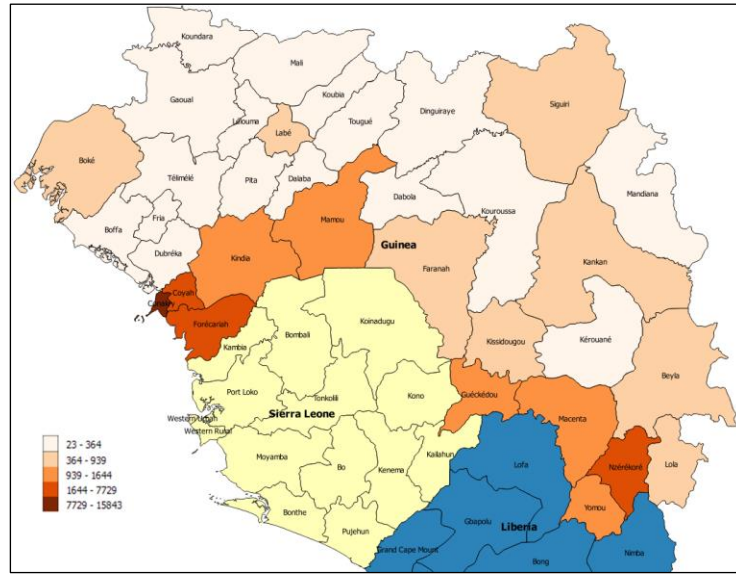
Figure 8.15: Footprint of inbound mobile users to Guinea



Source: ITU

Figure 8.16 illustrates the transboundary movement of mobile users from Sierra Leone and Liberia to Guinea. This population was calculated from the total number of unique mobile users who come to Guinea. It shows that the most visited prefectures by inbound transboundary population to Guinea are Conakry, Forécariah and Coyah. Additionally, Forécariah seems to be the main boundary prefecture for crossing from Sierra Leone to Guinea. Nzérékoré seems to be the main boundary district from Liberia to Guinea.

Figure 8.16: Number of mobile users moving into Guinea

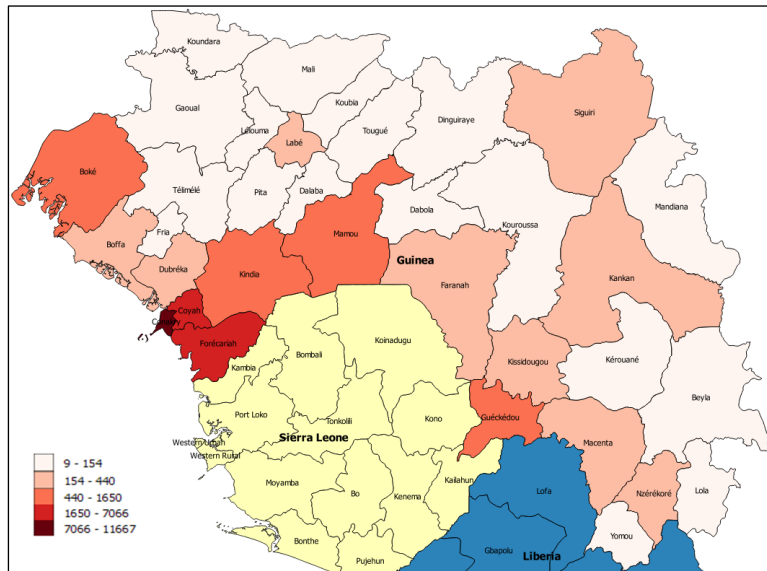


Source: ITU

8.6.4 Outbound transboundary population movement

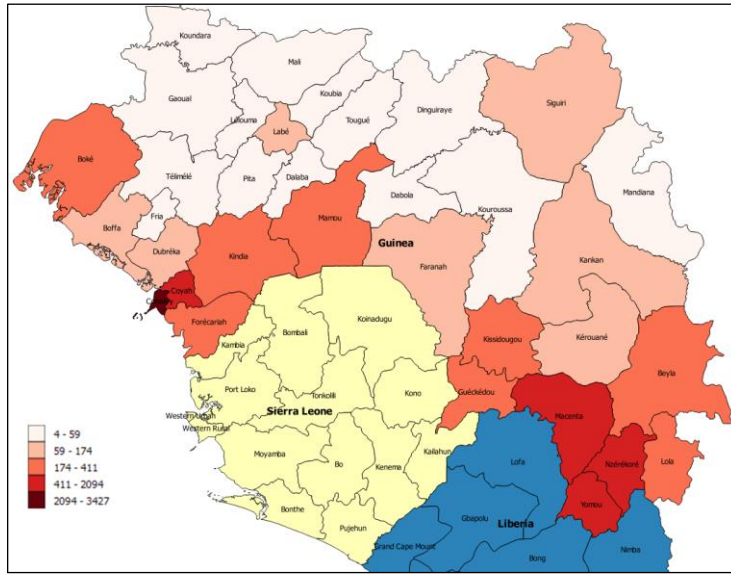
Figure 8.17 illustrates the transboundary movement of mobile users visiting Sierra Leone from Guinea. It shows that the three most visited prefectures are Conakry, Forécariah and Coyah.

Figure 8.17: Footprint of visitors to Sierra Leone



Source: ITU

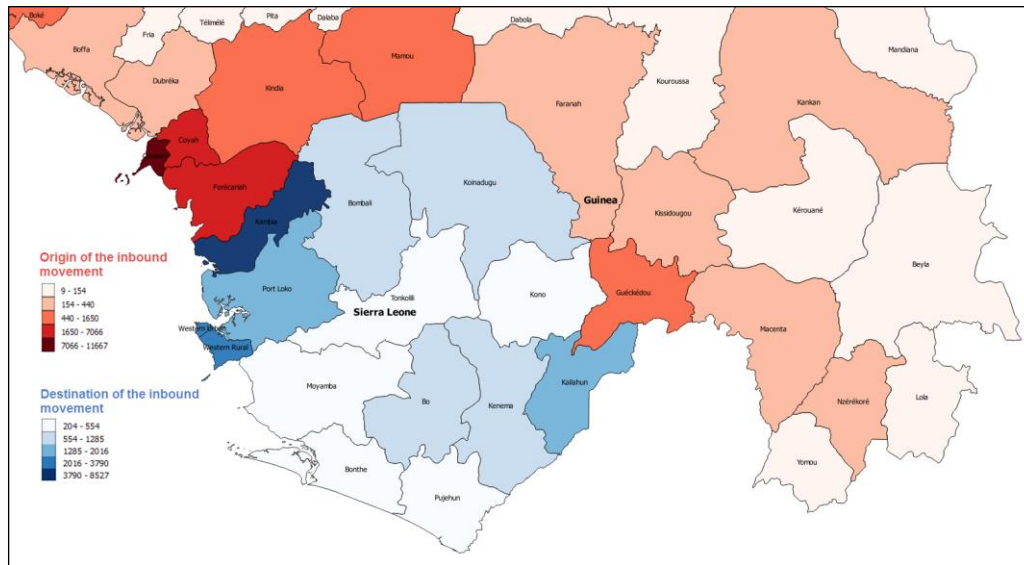
Figure 8.18: Footprint of visitors to Liberia



Source: ITU

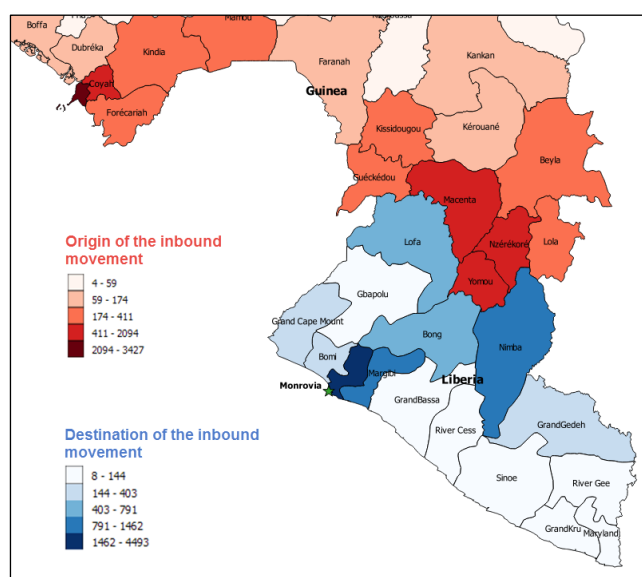
Figure 8.18 illustrates transboundary movement of mobile users visiting Liberia from Guinea. It shows that the three most visited prefectures with high transboundary population are Conakry, Nzérékoré, and Yomou. Figures 8.19 and 8.20 illustrate the cities that transboundary population visit Sierra Leone and Liberia.

Figure 8.19: Origin and destination of transboundary population to Sierra Leone



Source: ITU

Figure 8.20: Origin and destination of transboundary population to Liberia



Source: ITU

9 Discussions

The lack of timely and accurate information about movements of people during a natural disaster, emergency or outbreak of disease, can seriously limit or damage the effectiveness of humanitarian preparedness, response and recovery. However, the ubiquitous nature of mobile phones has revealed new opportunities for accessing such information, and CDRs provide valuable insights into the behaviour of people during disaster or emergency situations. For example, variations in the number of active phones connected to each cell tower reveal activity patterns in specific locations during and after an event, and can be used to determine population displacement and social connectivity between and among cities. CDR analysis suggests a strong potential to improve early warning and emergency management mechanisms, and even to reveal potential disease outbreak patterns.

This report describes how CDR data could contribute specifically to monitor and control epidemics, such as Ebola, by estimating the human trajectories and spatio-temporal distribution of populations from CDR data at the local (rural and urban) level as well as border areas.

Results of the analysis have shown that there exists a strong correlation between the distribution of mobile phone users extracted from CDR data and actual populations extracted from national census data, and high transboundary mobility from rural and urban areas in countries neighbouring Guinea was also observed.

The use of CDR data has shown enormous potential as a tool to help authorities during emergencies, disasters, and epidemic outbreaks. Ebola is an epidemic disease, and understanding the movement of potential disease carriers not just at the city-to-city level but on a national scale is critical for effective policy intervention.

In this project, CDR data was analysed and then aggregated to generate statistical information across a given area. The geographic locations of mobile phone devices are determined by the location of the mobile network base transceiver stations to which people connect when making or receiving a call. This is very important in terms of actual population calculations and estimates.

The analysis of the daily positions of millions of mobile devices collected over two months was used to estimate the home location of the mobile user population in Guinea, which also explained the phenomenon of increasing and decreasing population numbers during the day at specific locations due to a unique method of correlating mobile phone activity with the population.

However, this raises the question of how to preserve user privacy. Ensuring privacy is a significant limitation to using mobile phone data, and is important not only to maintain public acceptance of data usage, but also because there is a potential for data to be used to compromise human rights, especially those of the vulnerable. In most cases, access to the data imposes data anonymization to preserve user privacy. For this reason, software has been developed that can be provided to operators or local authorities who, with minimal training, can perform the data anonymization task. Despite this, the use of anonymized data shows very promising evidence of its effectiveness in emergency situations.

In this pilot project, populations under study have been restricted to mobile phone users. To present the population distributions of real populations, including those who are not present in CDR data, further examination to set scaling factors, for instance, is needed. In addition, this project focused on quantitative aspects of mobile phone users, such as the number of people or groups of people and their trajectories. However, without qualitative aspects, such as gender and age of users, it is difficult to analyse how extracted movements of the population can relate to other socio-economic factors. The inclusion of personal attributes could greatly expand the potential of CDR data to address societal issues including control of epidemic diseases.

While anonymized CDR data does not include any personal information, it is possible to estimate basic demographic attributes. A key to estimating demographic attributes of CDR data is the use of supplementary data that relates calling behaviour and demographic attributes. The data can be collected from mobile phone users through a field survey, for instance, and can be used as training data and validation data for the estimation. Given the significance of having attribute information for utilizing CDR data for societal issues, it is vital to collect supplementary data from mobile phone users.

Human trajectories and spatio-temporal population distribution extracted from CDR data can be used to analyse the impact of human mobility on the spread of communicable diseases, and to estimate the number of evacuees under disaster or emergency conditions at a given time and location.

Hourly population distribution maps are useful when extracting statistical information, such as numbers of people at the disaggregated level rather than at the administrative unit. In addition, it can cover difficult to reach populations, without using an advanced national registration system, because mobile phone penetration is so wide-spread.

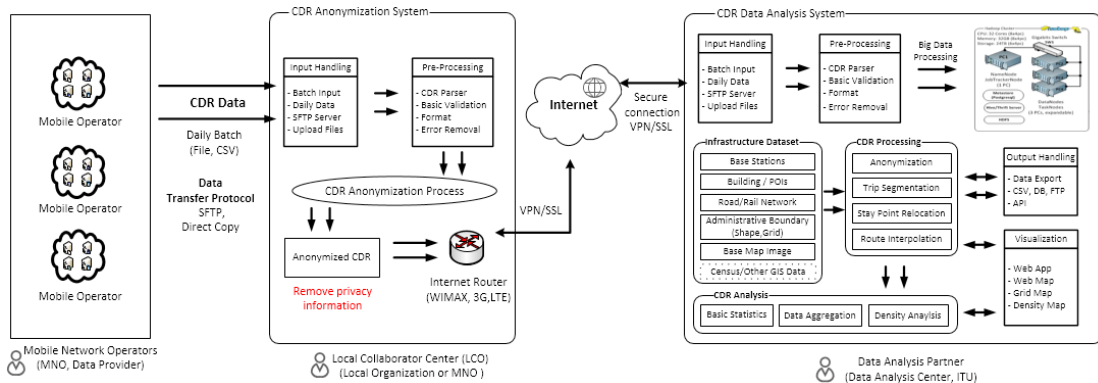
From the operator perspective, any data release involves both effort and risk, including contractual constraints over release of customer data, or at least a perceived loss of consumer trust if records are transferred outside company control. Nevertheless, operators may well release data either before or during an emergency driven by corporate social responsibility considerations. In addition, because global issues are borderless, well organized coordination of local governments and operators in multiple countries is important for effective data utilization.

To drive the use of CDR data for disaster/outbreak risk management, the increasing role of mobile technology in communicating information during emergencies and disasters (data collection and data dissemination) needs to be recognised.

Finally, it is recommended that new policies be drawn up that permit the use of CDR data for security and emergency response so that the CDR data can be recorded, retrieved, analysed and used to plan for and mitigate the effects of future outbreaks.

Appendix 1: Overall concept of automated CDR data analysis

Figure A.1.1: Overall concept of CDR data analysis



Source: ITU

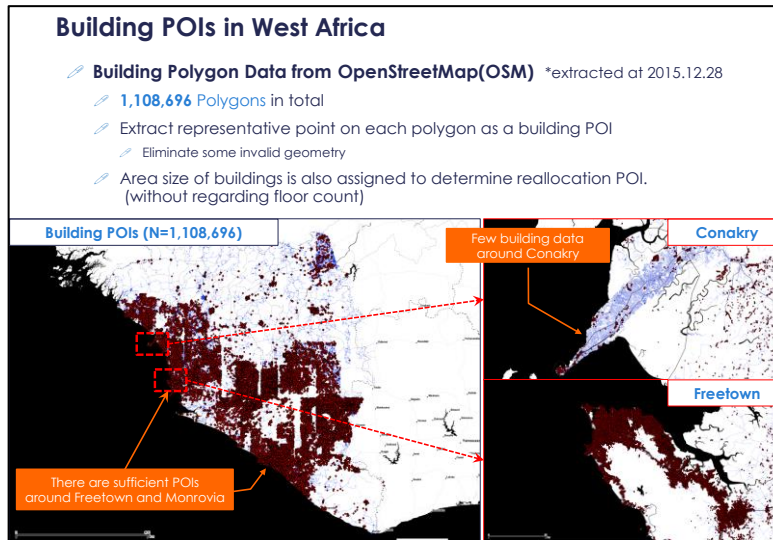
For this project, the entire process still required manual operation such as input handling, pre-processing and data analysis. The automated CDR data analysis system allows the whole system to work automatically with less operator involvement. This system is recommended in case of expansion to large-scale deployment.

- Data provider (mobile operator):** The data is collected from the mobile network operator (data provider). In this case, the operator collects mobile device location data and stores it, before exporting or transferring it to a local collaborator centre (LCO) for further processing. The data mainly comes in a CSV file format with compression (zip). Any data transfers between the data provider and LCO must be done via secure channel (SFTP, VPN) or by direct copy.
- Local collaborator centre (LCO):** The LCO stores the anonymized CDR data (usually carried out by the regulator or mobile operator licence provider) however, the data provider can also act as LCO. In such cases, the data will be anonymized by the LCO (e.g. operator) in addition to transferring data to a cloud storage system (data storage partner).
- Data analysis partner (DAP):** The DAP maintains all sanitized CDR data (possibly multiple operators). It has to ensure that all data and information are secure. This unit will provide cloud storage system to allow other data providers to upload their data to the cloud storage system for further analysis. User management and authorization to upload or download will be controlled by this unit.
- CDR anonymization system:** This system handles the anonymization process on CDR data. The system retrieves raw CDR data in csv format and processes it to remove all privacy related information that will then be uploaded to the cloud storage system. At this phase, anonymization is a command line application used to supply parameters such as path of input, path of output, seed data and CDR format parser. Since the CDR data format may be different among data providers, specific CDR format parsers may need to be developed. Basic validation, error checking and removal are also included in the program.
- CDR data analysis system:** This system handles deep analysis and maintains all CDR data from all data providers. It incorporates many modules such as: Input handling, Pre-processing, Big data processing unit and CDR Processing. Population estimation is a product of this system, however, the delivery of this system has not been included at this phase.
- Data transferring (LCO->DAP):** Secure connections (VPN/SSL) should be used to encrypt all transferring data. Connections will be established and fully controlled by the local collaborator. Internet connections can be through a 3G mobile network, WiMAX, and fixed line connection. This connection can also be used for remote management and provisioning of the system. LCO may also upload data via secure FTP (SFTP).

Appendix 2: Building points of interest (POIs) and road networks

Figure A.2.1 illustrates the POIs data constructed for this project. It consists of 1 108 696 polygons, each of which represents a building. Area size of buildings is used to calculate the area of POI buildings, to which stay point locations are reassigned.

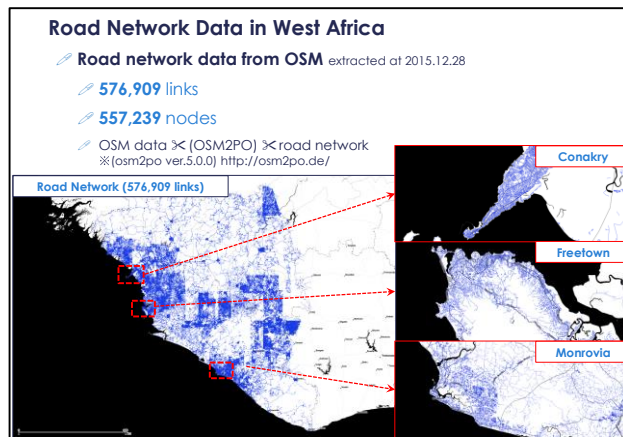
Figure A.2.1: Building POI in West Africa



Source: ITU

The road network data constructed for this project is based on OSM (OpenStreetMap)⁸ data as of 28 December, 2015, and includes 576 909 links and 557 239 nodes.

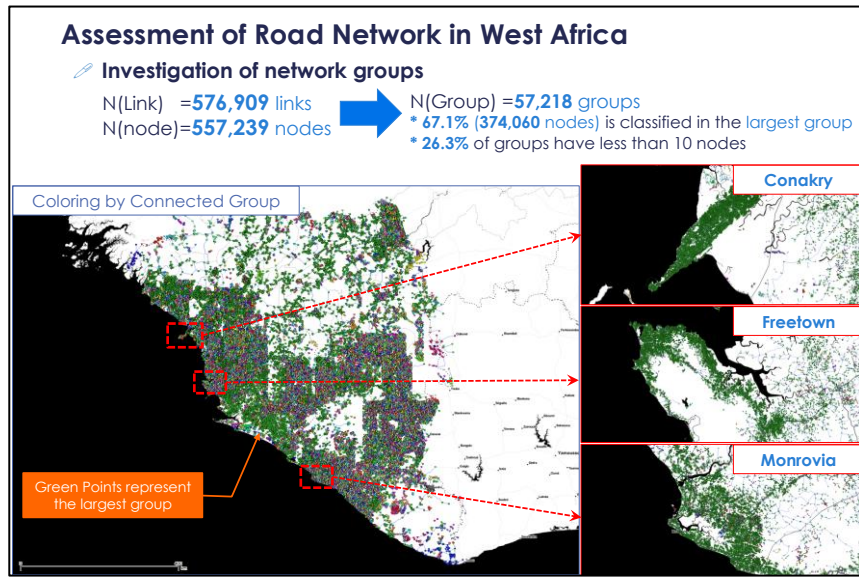
Figure A.2.2: Road network data in West Africa



Source: ITU

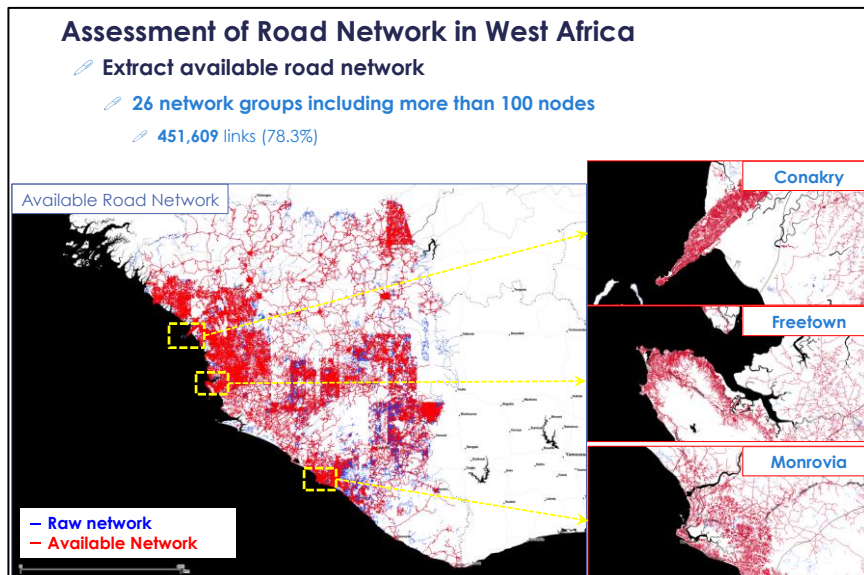
⁸ <https://www.openstreetmap.org/#map=5/51.500/-0.100>

Figure A.2.3: Assessment of road network in West Africa



Source: ITU

Figure A.2.4: Extracting road network in West Africa



Source: ITU

International Telecommunication Union (ITU)
Telecommunication Development Bureau (BDT)
Office of the Director
Place des Nations
CH-1211 Geneva 20 – Switzerland
Email: bdtdirector@itu.int
Tel.: +41 22 730 5035/5435
Fax: +41 22 730 5484

Deputy to the Director and
Director, Administration and
Operations Coordination
Department (DDR)
Email: bdtdeputydir@itu.int
Tel.: +41 22 730 5784
Fax: +41 22 730 5484

Infrastructure Enabling
Environment and
e-Applications Department (IEE)
Email: bdtiee@itu.int
Tel.: +41 22 730 5421
Fax: +41 22 730 5484

Innovation and Partnership
Department (IP)
Email: bdtip@itu.int
Tel.: +41 22 730 5900
Fax: +41 22 730 5484

Project Support and Knowledge
Management Department (PKM)
Email: bdtpkm@itu.int
Tel.: +41 22 730 5447
Fax: +41 22 730 5484

Africa

Ethiopia
International Telecommunication
Union (ITU)
Regional Office
P.O. Box 60 005
Gambia Rd., Leghar ETC Building
3rd floor
Addis Ababa – Ethiopia

Email: itu-addis@itu.int
Tel.: +251 11 551 4977
Tel.: +251 11 551 4855
Tel.: +251 11 551 8328
Fax: +251 11 551 7299

Cameroon
Union internationale des
télécommunications (UIT)
Bureau de zone
Immeuble CAMPOST, 3^e étage
Boulevard du 20 mai
Boîte postale 11017
Yaoundé – Cameroon

Email: itu-yaounde@itu.int
Tel.: +237 22 22 9292
Tel.: +237 22 22 9291
Fax: +237 22 22 9297

Senegal
Union internationale des
télécommunications (UIT)
Bureau de zone
19, Rue Parchappe x Amadou
Assane Ndoye
Immeuble Fayçal, 4^e étage
B.P. 50202 Dakar RP
Dakar – Senegal

Email: itu-dakar@itu.int
Tel.: +221 33 849 7720
Fax: +221 33 822 8013

Zimbabwe
International Telecommunication
Union (ITU)
Area Office
TelOne Centre for Learning
Corner Samora Machel and
Hampton Road
P.O. Box BE 792 Belvedere
Harare – Zimbabwe

Email: itu-harare@itu.int
Tel.: +263 4 77 5939
Tel.: +263 4 77 5941
Fax: +263 4 77 1257

Americas

Brazil
União Internacional de
Telecomunicações (UIT)
Regional Office
SAUS Quadra 06, Bloco "E"
11^o andar, Ala Sul
Ed. Luis Eduardo Magalhães (Anatel)
70070-940 Brasília, DF – Brazil

Email: itubrasilia@itu.int
Tel.: +55 61 2312 2730-1
Tel.: +55 61 2312 2733-5
Fax: +55 61 2312 2738

Barbados
International Telecommunication
Union (ITU)
Area Office
United Nations House
Marine Gardens
Hastings, Christ Church
P.O. Box 1047
Bridgetown – Barbados

Email: itubridgetown@itu.int
Tel.: +1 246 431 0343/4
Fax: +1 246 437 7403

Chile
Unión Internacional de
Telecomunicaciones (UIT)
Oficina de Representación de Área
Merced 753, Piso 4
Casilla 50484, Plaza de Armas
Santiago de Chile – Chile

Email: itusantiago@itu.int
Tel.: +56 2 632 6134/6147
Fax: +56 2 632 6154

Honduras
Unión Internacional de
Telecomunicaciones (UIT)
Oficina de Representación de Área
Colonia Palmira, Avenida Brasil
Ed. COMTELCA/UIT, 4.º piso
P.O. Box 976
Tegucigalpa – Honduras

Email: ituftegucigalpa@itu.int
Tel.: +504 22 201 074
Fax: +504 22 201 075

Arab States

Egypt
International Telecommunication
Union (ITU)
Regional Office
Smart Village, Building B 147, 3rd floor
Km 28 Cairo – Alexandria Desert Road
Giza Governorate
Cairo – Egypt

Email: itucairo@itu.int
Tel.: +202 3537 1777
Fax: +202 3537 1888

Asia and the Pacific

Thailand
International Telecommunication
Union (ITU)
Regional Office
Thailand Post Training Center, 5th
floor,
111 Chaengwattana Road, Laksi
Bangkok 10210 – Thailand

Mailing address
P.O. Box 178, Laksi Post Office
Laksi, Bangkok 10210 – Thailand

Email: itubangkok@itu.int
Tel.: +66 2 575 0055
Fax: +66 2 575 3507

Indonesia
International Telecommunication
Union (ITU)
Area Office
Sapta Pesona Building, 13th floor
Jl. Merdan Merdeka Barat No. 17
Jakarta 10001 – Indonesia

Mailing address:
c/o UNDP – P.O. Box 2338
Jakarta 10001 – Indonesia

Email: itujakarta@itu.int
Tel.: +62 21 381 3572
Tel.: +62 21 380 2322
Tel.: +62 21 380 2324
Fax: +62 21 389 05521

CIS countries

Russian Federation
International Telecommunication
Union (ITU)
Area Office
4, Building 1
Sergiy Radonezhsky Str.
Moscow 105120
Russian Federation

Mailing address:
P.O. Box 25 – Moscow 105120
Russian Federation

Email: itumoscow@itu.int
Tel.: +7 495 926 6070
Fax: +7 495 926 6073

Europe

Switzerland
International Telecommunication
Union (ITU)
Telecommunication Development
Bureau (BDT)
Europe Unit (EUR)
Place des Nations
CH-1211 Geneva 20 – Switzerland
Switzerland
Email: eurregion@itu.int
Tel.: +41 22 730 5111



International Telecommunication Union
Telecommunication Development Bureau
Place des Nations
CH-1211 Geneva 20
Switzerland
www.itu.int

ISBN-13: 978-92-61-22291-8



9 789261 222918

Printed in Switzerland
Geneva, 2017