REPUBLIC OF LIBERIA

Call Detail Record (CDR) ANALYSIS: REPUBLIC OF LIBERIA

Final Report

Telecommunication Development Sector

Call detail record (CDR) analysis: Republic of Liberia

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).

ISBN

978-92-61-20281-1 (Paper version)

978-92-61-20291-0 (electronic version)

978-92-61-20301-6 (EPUB version)

978-92-61-20311-5 (MOBY version)

Please consider the environment before printing this report.

Table of Contents

List of tables, figures and boxes

Tables

Figures

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 Liberia is a prime example, being heavily impacted by the Ebola outbreak, along with neighbouring countries (the Republic of Guinea and Sierra Leone), where 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 not only on a mobile communication network but on 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 Liberia, 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.

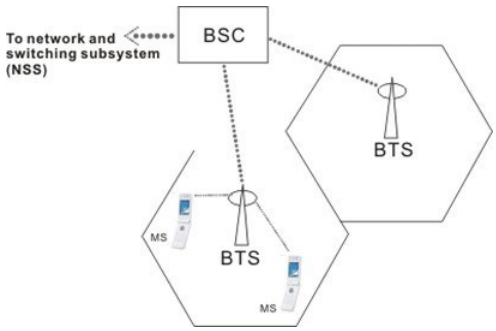
In the last outbreak of Ebola in West Africa over 11 000 people died. National-health authorities in West Africa have struggled to contain Ebola, especially how to estimate the affected population in the outbreak prone area and how to understand outbreak patterns spreading from one area to another. Currently, 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 to 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 Liberia

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

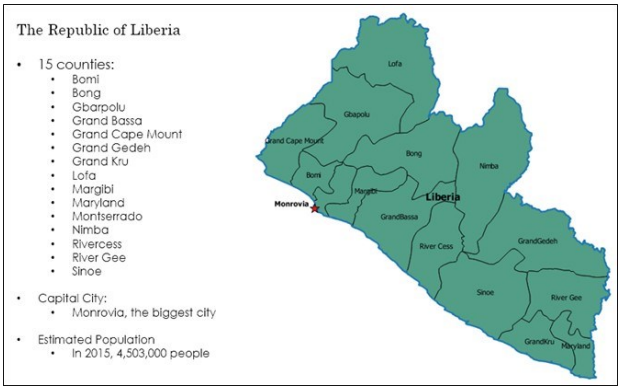
- 2G networks, first introduced in 1992, the second-generation of cellular telephone technology, was the first to use digital encryption of conversations, and the first to offer data services and SMS text messaging.
- 3G networks succeeded 2G, offering faster data transfer rates and were the first to 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 Liberia

The Republic of Liberia is composed of 15 counties that are subdivided into a total of 90 districts and further subdivided into clans. Nimba is the largest of the counties in size and Montserrado County is the smallest but the most populous. The local government takes the form of superintendents appointed by the President. The capital city is Monrovia, which is also the largest city. Liberia has an estimated population of approximately 4.3 million people[3] and approximately 73 per cent is covered by mobile networks. There are four major mobile phone operators in Liberia: Lonestar Cell, Atlantic, Cellcom and Novafone. They offer GSM services, which offer both a prepaid and post-paid service. The Liberia Telecommunications Authority (LTA) regulates the Liberia telecommunication sector.

1 GSM: 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 Liberia: www.itu.int/net4/ITU-D/idi/2015/#idi2015countrycard-tab&LBR

Figure 3.2: Administrative map of Liberia



Source: ITU

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's phones were connected to at the time of the call, the identities of sources (points of origin), the identities of destinations (endpoints), 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 2-3 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 Institute[4] 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.

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

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.

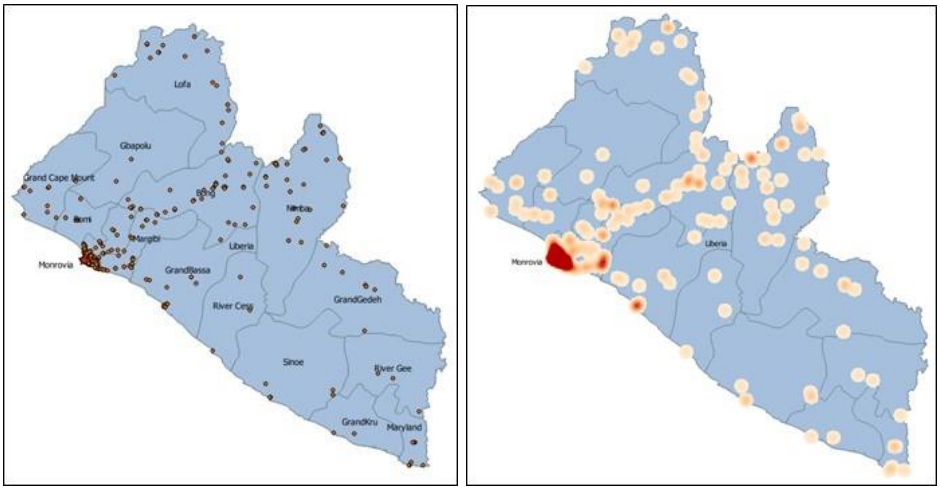
  Such information could help operators to adapt their service to specific groups of people as well as specific locations, maintain customer satisfaction with incentives, and also target customer support, for example for those who adopt specific services and features.
- **Build customer profiles**: Operators 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.
- **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 datasets and basic statistics

For the CDR analysis project, two months of CDR data - from June to July 2015 - were collected in Liberia to demonstrate how dynamic population movements could be estimated. Data was prepared by the mobile network operators in Liberia: Lonestar, Cellcom and Novafone. This meant that the data included the majority of the mobile Liberia 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. Density map on the right side show a high concentration of cell towers in Monrovia city.

Figure 5.1: Cell Tower Location at Liberia



Source: ITU

Figure 5.2 illustrates the basic information obtained in CDR data from the mobile network operators, Lonestar, Celkom and Novafone. 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 this case, data pre-processing was necessary. Lonestar provided both incoming and outgoing data; only outgoing data were used in this processing. In addition, due to the absence of IMEI code in Cellcom data, the IMSI code 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	Liberia	Lonestar	68.2 GB	June,July 2015	431,934,109	O (1,708,830)	O (1,218,484)	-	O	O	-	In and Out
2		Cellcom	84.9 GB	June,July 2015	204,715,956	O (1,576,947)	O (1,116,498)	O (1,094,203)	O	O	-	-
3		Novafone	1.7 GB	June,July 2015	8,220,071	O (126,091)	O (95499)	O (461,658)	O	O	-	-

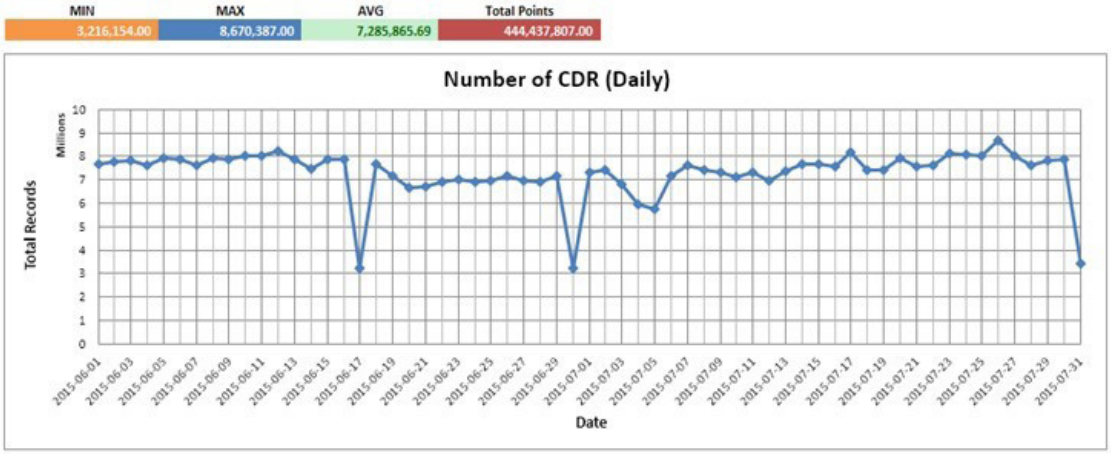
NO	Country	MNO	Data							
			Data Size	Data Period	Total record	IMEI (caller)	IMSI (caller)	Caller No	Time stamp	cell ID
1	Liberia	Lonestar	included in voice	June,July 2015	83,287,934	O (734,051)	O (671,162)	-	O	O
2		Cellcom	11.6 GB	June,July 2015	74,030,098	-	O (309,272)	O (303,502)	O	O
3		Novafone	There is an issue on data. No use. (not enough information)							

Source: ITU

CDR basic statistics

CDR data was collected from the mobile network operators and the total activity generated an average of over 7.3 million items a day. Figure 5.3 illustrates the daily call activity, which reached an average of 7 285 865.96 records.

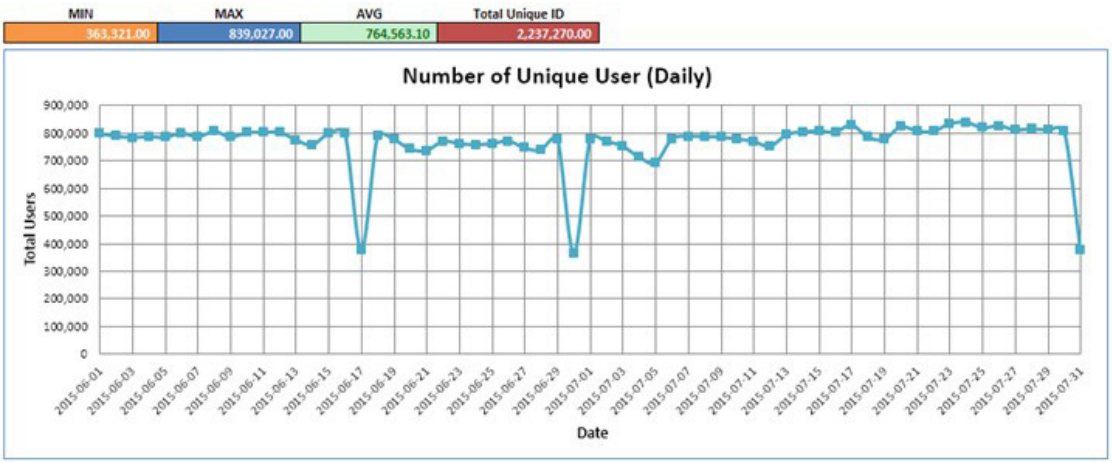
Figure 5.3: Daily call activity



Source: ITU

Analysing user behaviour during events such as during the aftermath of a disaster could use CDR data. 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 reached approximately 764 563 users and the CDR data per user averaged 9.23 per day.

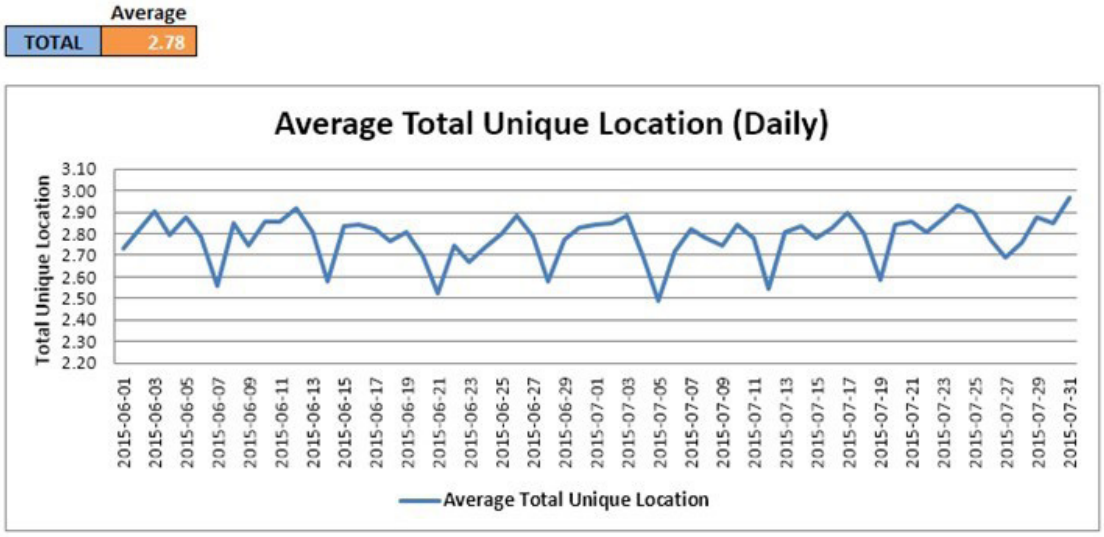
Figure 5.4: Daily unique user



Source: ITU

Figure 5.5 shows the average total unique location of people reached 2.78, which means that people visited less than three locations a day.

Figure 5.5: Average total unique location (daily)



Source: ITU

6 CDR data: Limitations and difficulties

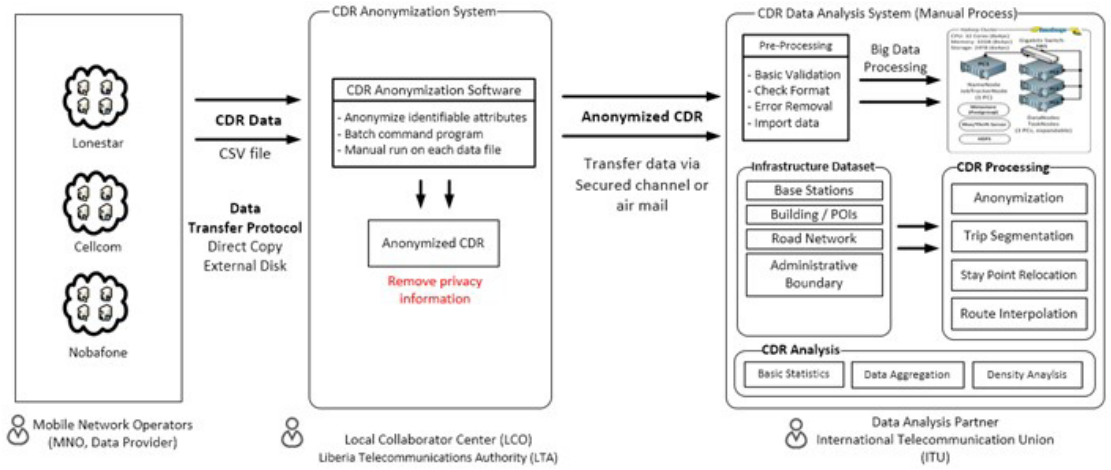
Even though applying CDR data can provide many benefits in various sectors and applications as previously described in section 4, there are some obstacles and difficulties that limit the use of CDR data.

- **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: Overview of CDR data analysis



Source: ITU

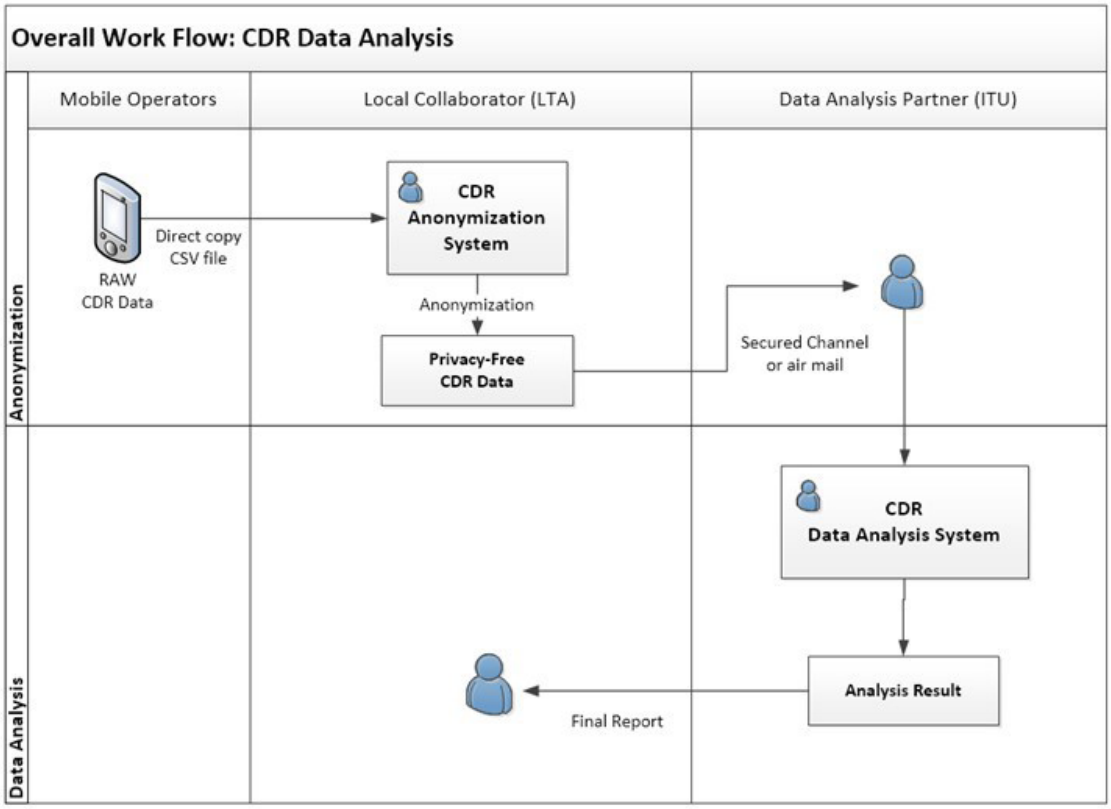
The process of CDR analysis requires a certain number of steps that could eventually be automated, including the use of anonymization software, which has been developed and shared to data providers to remove personally identifiable information from their data sets, so that the people remain anonymous. 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 Overall work flow

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

Figure 7.2: CDR data analysis work flow



Source: ITU

7.3 CDR data specification

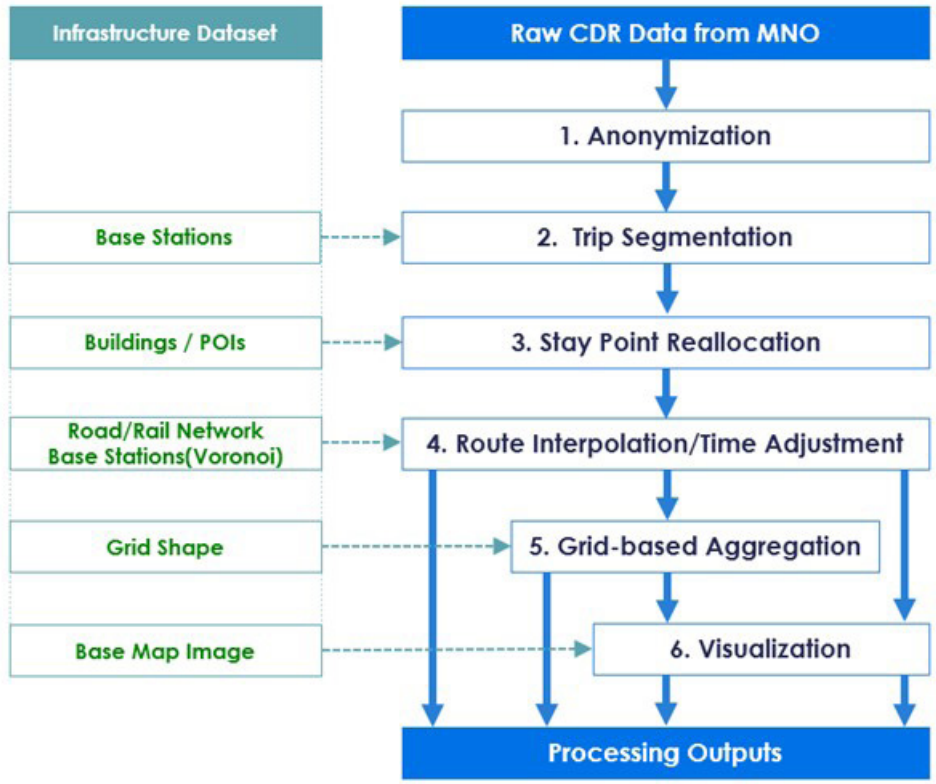
The operator is asked to provide the following standardized information:

- 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

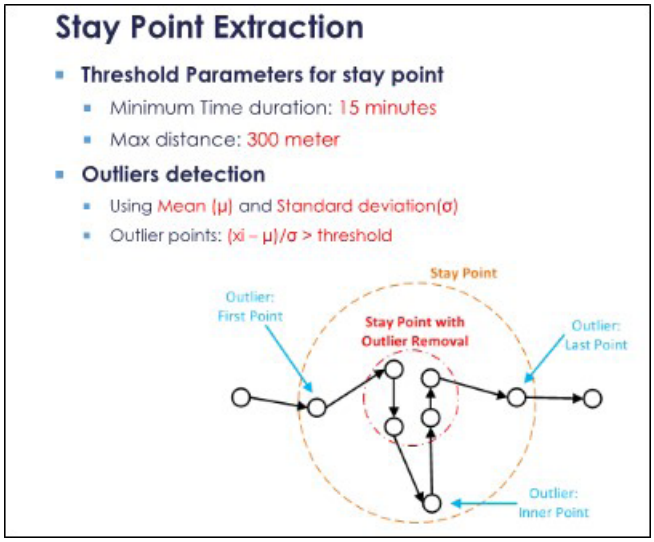


Source: ITU

1. **Anonymization (hash[5]):** 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.

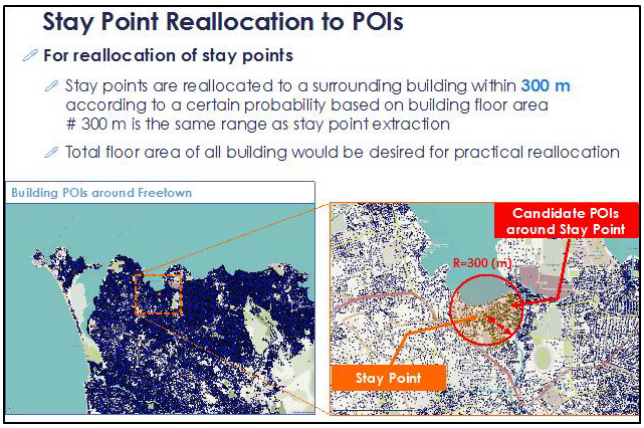
Figure 7.4: CDR process of stay point extraction



Source: ITU

3. **Stay point reallocation:** Reallocate stay points (Trip OD) to surrounding points of interest (POIs) with a certain probability and fill gap between stay/move segments. POIs are regarded as surrounding a certain cell tower if they are closer to the cell tower location than to the others (Voronoi tessellation). The 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 or likely because POIs can be considered places where people are likely to stay or visit, such as shopping areas, residential houses, villages, and to which people are reassigned rather than concentrating on cell tower locations. 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 antenna location, are reallocated.

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

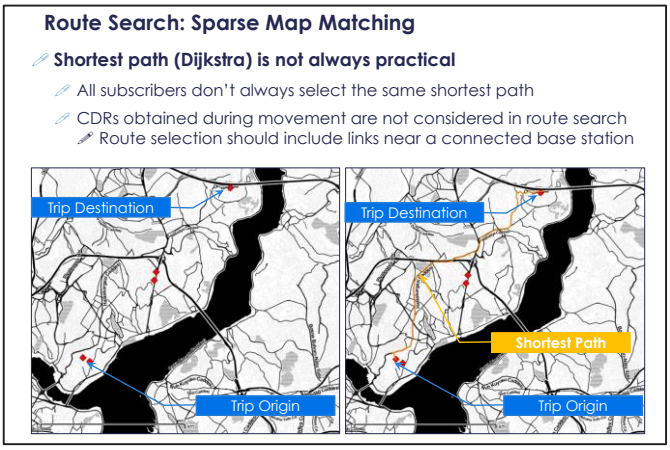


Source: ITU

4. **Route interpolation:** Routes between every pair of reallocated stay points (building POIs) are interpolated by searching the shortest path between the POIs (ODs). 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

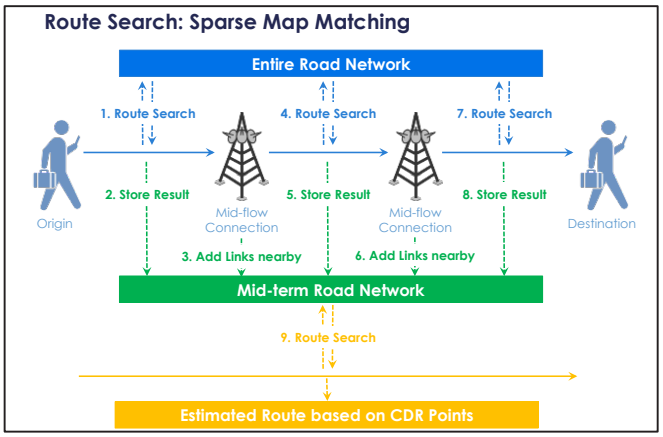
trip origin and destination. Red dots indicate the location of reallocated stay points and yellow nodes indicate a shortest path for the origin and destination. Some intermediate points, which are not stay points, are not on the shortest path because only the origin and destination of a trip is displayed. 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



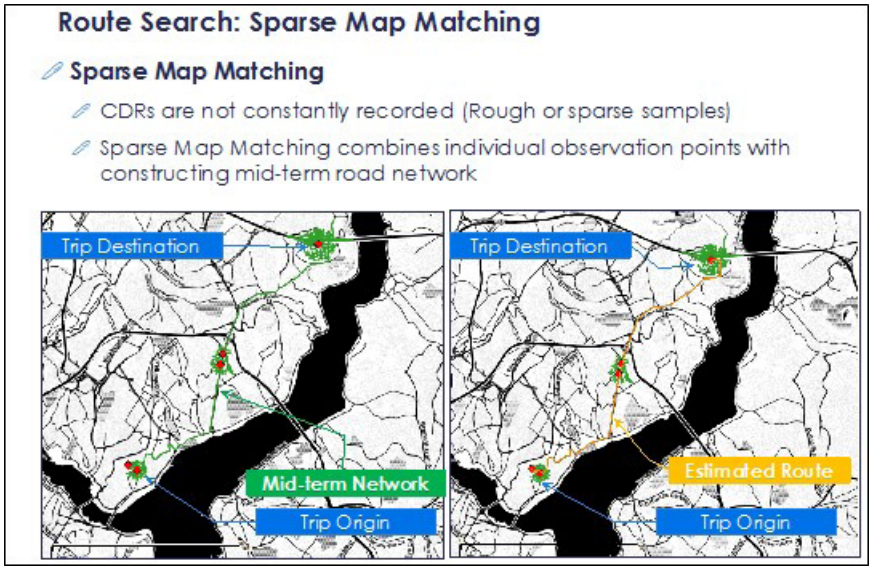
Source: ITU

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

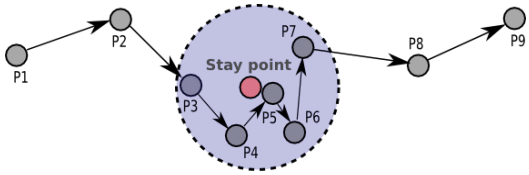
5. **Grid-based aggregation:** this refers to the aggregate measurement of population density in a certain grid (1 km) and time duration (1 hour).
6. **Visualization:** This creates an animated projection (Mobmap generated film[6]) of disaggregate/ aggregate movement of people (off-line).

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,...p_n), where p = (ID, time, latitude, longitude) and n = 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}, ..., p_n}, where $\forall m < i \leq n$, Distance(p_m, p_i) ≤ dt, Distance(p_m, p_{n+1}) > dt and Time Interval(p_m, p_n) ≥ 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:

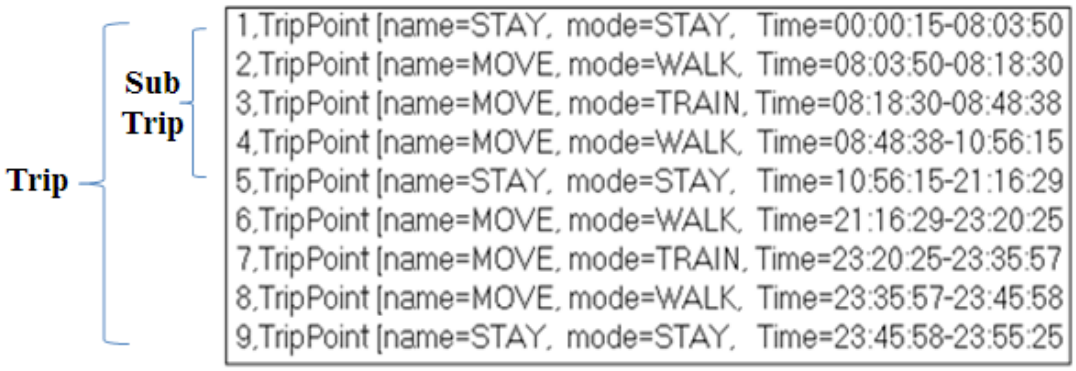
[6] http://shiba.iis.u-tokyo.ac.jp/member/ueyama/mm/

$$h = \langle s_0 \overset{\Delta t_1}{\rightarrow} s_1 \overset{\Delta t_2}{\rightarrow} \cdots \overset{\Delta t_{n-1}}{\rightarrow} s_n \rangle$$

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



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	• Column name: UID • Unique for each device • Encrypted using Hash function, irreversible
2 Date	• Column name: DATE • Date format: yyyy-MM-dd • Example: 2015-12-31
3 Trip Sequence	• Column name: TRIP_SEQUENCE • Order of sub trip in a day, start from 1
4 Mobility Type	• Column name: MOBILITY_TYPE • Value: STAY or MOVE
5 Transportation Mode	• Column name: TRANSPORT_MODE • Indicate mode of transportation of corresponding sub trip • Value: STAY, WALK, CAR
6 Total Distance	• Column name: TOTAL_DISTANCE • Total travel distance of sub trip in meter
7 Total Time	• Column name: TOTAL_TIME • Total travel time of sub trip in second
8 Start Time	• Column name: START_TIME • Indicate start time of sub trip • Format: HH24:mm:ss • Example: 23:20:00
9 End Time	• Column name: END_TIME • Indicate end time of sub trip • Format: HH24:mm:ss • Example: 23:20:00
10 Total Points	• Column name: TOTAL_POINTS • Indicate total number of point data in sub trip
11 Point lists	• 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.

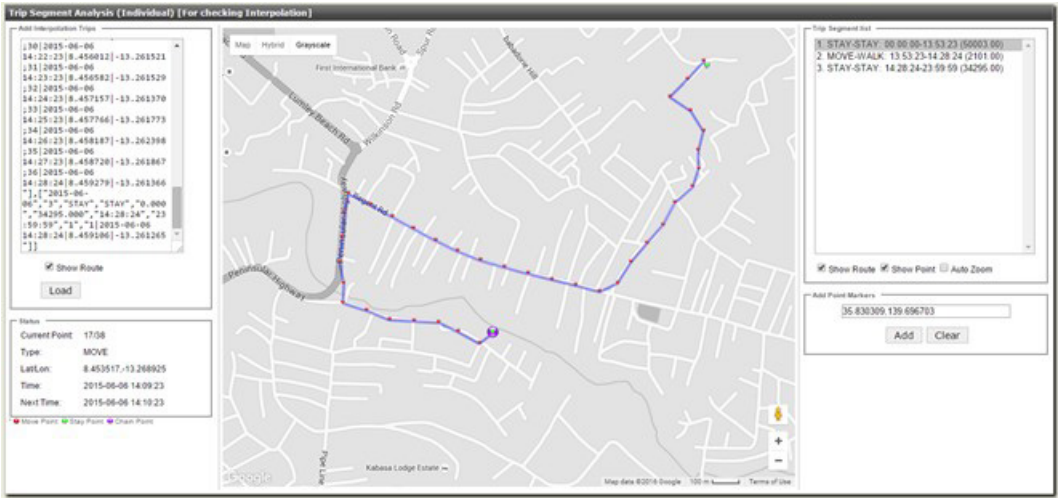
Trip segment analysis for checking interpolation result

UID, DATE (yyyy-MM-dd), TRIP_SEQUENCE_ID, MOBILITY_TYPE, TRANSPORT_MODE, TOTAL_DISTANCE, TOTAL_TIME, START_TIME, END_TIME, TOTAL_POINTS, POINT_LIST
422a837717,2015-06-06,1,STAY,STAY,0.000,57749.000,00:00:00,16:02:29,1,1|2015-06-01 00:00:00|6.373743|-10.772951

422a837717,2015-06-06,2,MOVE,WALK,3153.708,2323.000,16:02:29,16:41:12,39,1|2015-06-01 16:02:29|6.373497|-10.773267;2|2015-06-01 16:03:30|6.374243|-10.773447;3|2015-06-01 16:04:31|6.374983|-10.773652;4|2015-06-01 16:05:32|6.375711|-10.773898;5|2015-06-01 16:06:33|6.376103|-10.774265;6|2015-06-01 16:07:34|6.375691|-10.774913;7|2015-06-01 16:08:35|6.375280|-10.775561;8|2015-06-01 16:09:36|6.374868|-10.776209;9|2015-06-01 16:10:38|6.374457|-10.776858;10|2015-06-01 16:11:39|6.374046|-10.777506;11|2015-06-01 16:12:40|6.373634|-10.778154;12|2015-06-01 16:13:41|6.373223|-10.778802;13|2015-06-01 16:14:42|6.372842|-10.779469;14|2015-06-01 16:15:43|6.372488|-10.780150;15|2015-06-01 16:16:44|6.372098|-10.780811;16|2015-06-01 16:17:45|6.371690|-10.781461;17|2015-06-01 16:18:47|6.371286|-10.782115;18|2015-06-01 16:19:48|6.370883|-10.782768;19|2015-06-01 16:20:49|6.370479|-10.783421;20|2015-06-01 16:21:50|6.370076|-10.784074;21|2015-06-01 16:22:51|6.369672|-10.784727;22|2015-06-01 16:23:52|6.369238|-10.785360;23|2015-06-01 16:24:53|6.368795|-10.785988;24|2015-06-01 16:25:55|6.368358|-10.786618;25|2015-06-01 16:26:56|6.368053|-10.787320;26|2015-06-01 16:27:57|6.368022|-10.787936;27|2015-06-01 16:28:58|6.368780|-10.788057;28|2015-06-01 16:29:59|6.369542|-10.788139;29|2015-06-01 16:31:00|6.370310|-10.788158;30|2015-06-01 16:32:01|6.371075|-10.788096;31|2015-06-01 16:33:02|6.371611|-10.788376;32|2015-06-01 16:34:04|6.371883|-10.789094;33|2015-06-01 16:35:05|6.372137|-10.789817;34|2015-06-01 16:36:06|6.372195|-10.790581;35|2015-06-01 16:37:07|6.372183|-10.791348;36|2015-06-01 16:38:08|6.372054|-10.791999;37|2015-06-01 16:39:09|6.371293|-10.791979;38|2015-06-01 16:40:10|6.371313|-10.792746;39|2015-06-01 16:41:12|6.371295|-10.793513

422a837717,2015-06-06,3,STAY,STAY,0.000,34295.000,16:41:12,23:59:59,1,1|2015-06-06 16:41:12|6.371295|-10.793513

Figure 7.11: Trajectory check (internal use)



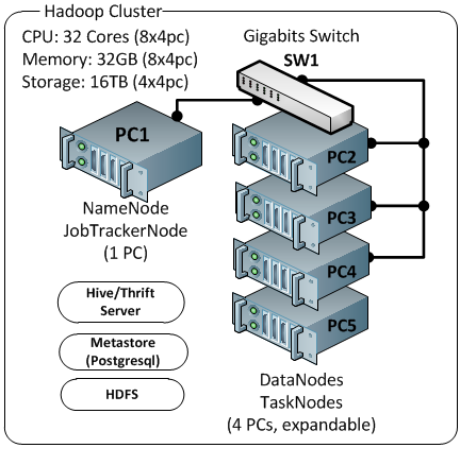
Source: ITU

7.6 Hadoop system for data processing

CDR data forms a very large dataset that ordinary computers or database systems are unable to process within an acceptable time frame, which is why the Hadoop system was introduced as 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 setup and use Hadoop for full operation mode, it required to run five components: 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 which 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.

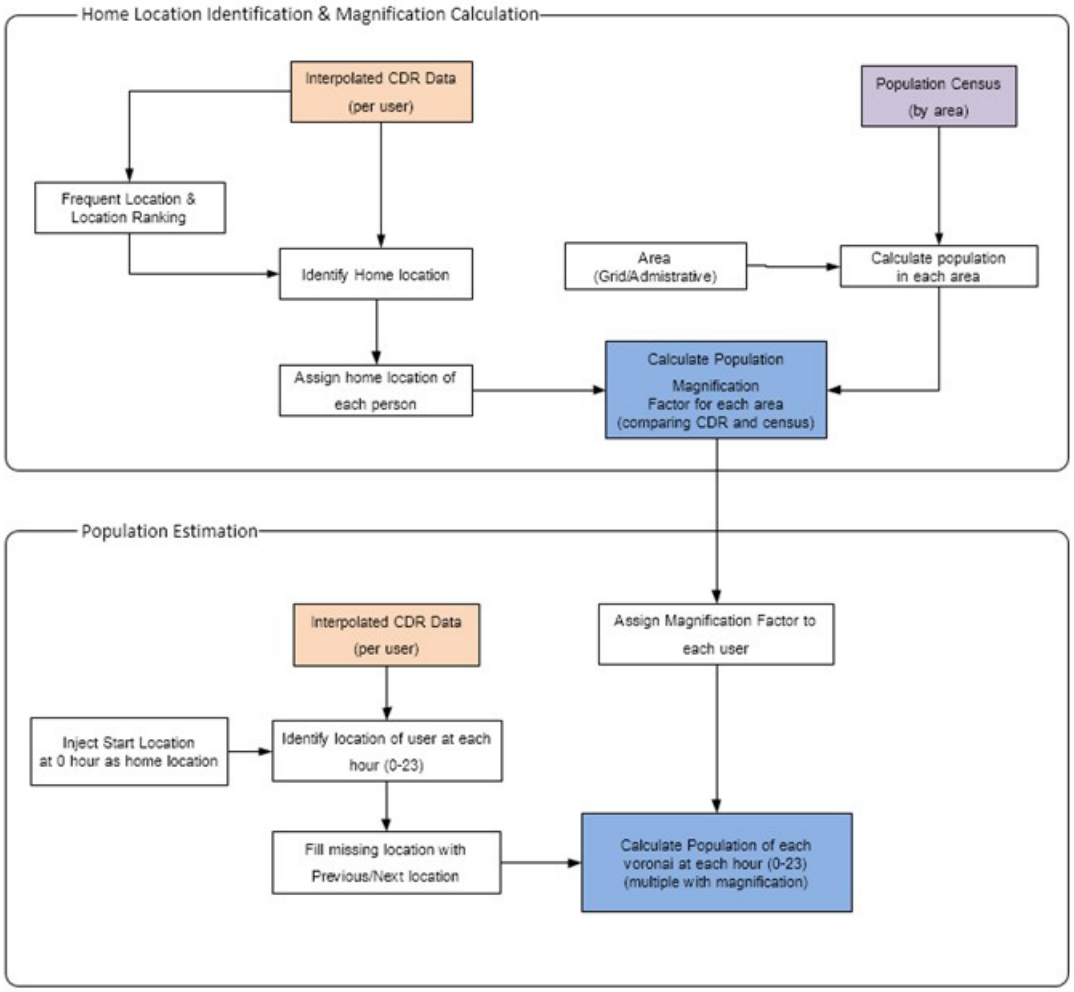
7.7 Dynamic population estimation

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 when 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
2. represent trajectory of people or where and when people are located. The magnification process is described in Figure 7.13.

In this process, at first, 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 ten people. The magnification factor is also used to calculate dynamic population at each time period.

Figure 7.13: Home location identification with population magnification



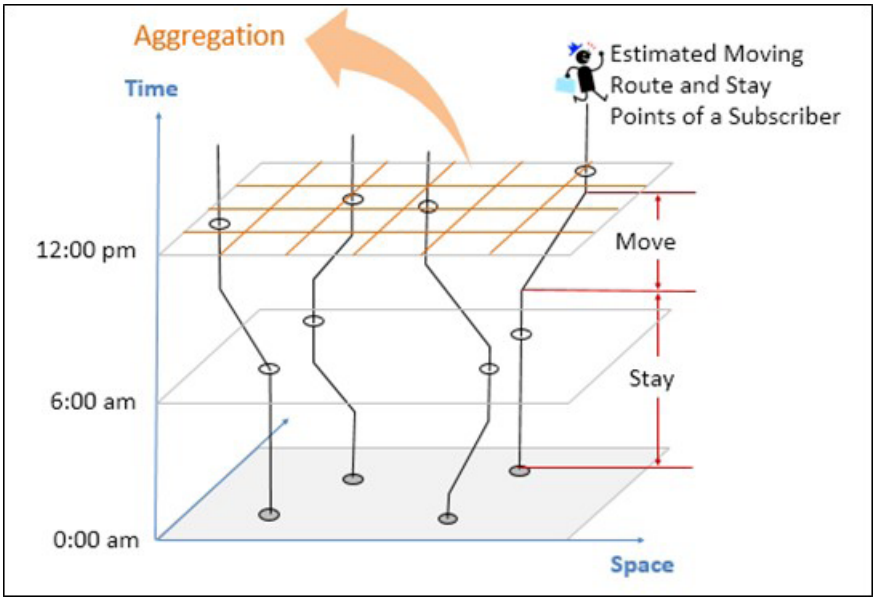
Source: ITU

Dynamic day time population (ambient population), unlike residential population, takes into account the movements of individuals through a given area (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

7.8 Mobile phone and SIM card cloning

Mobile phone cloning[7] 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 Monrovia, and five minutes later, another call is made but this time in Greenville, about 300 km away, 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,

[7] https://www.movzio.com/howto/cell-phone-cloning/

start 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.

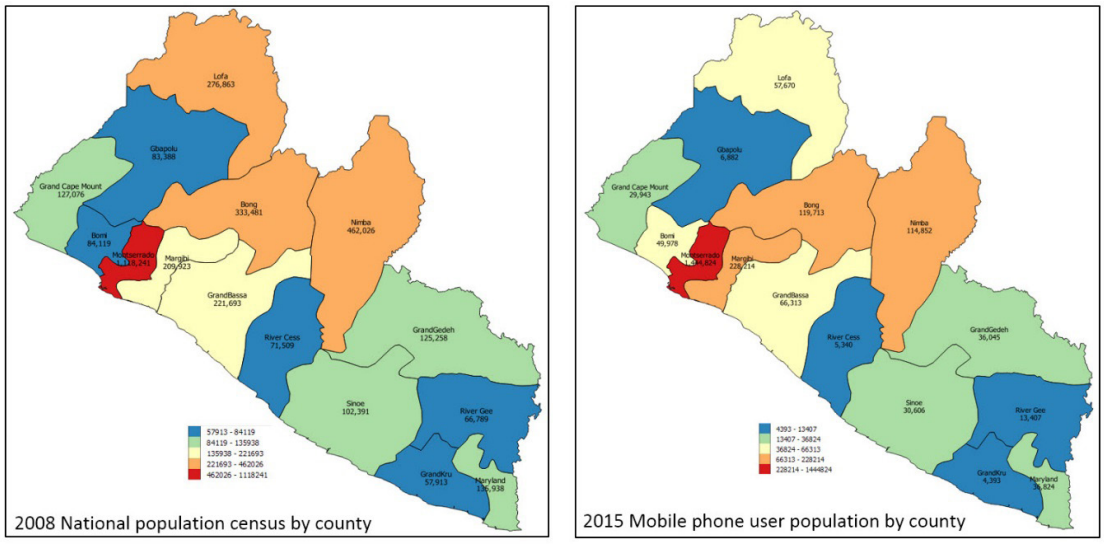
In this project, attempts were made to detect mobile cloning in GSM networks from CDRs using *big data* technology and spatial processing, and the results are expressed in the section 8.5 of this report.

8 Analysis results

8.1 Mobile phone user population by county

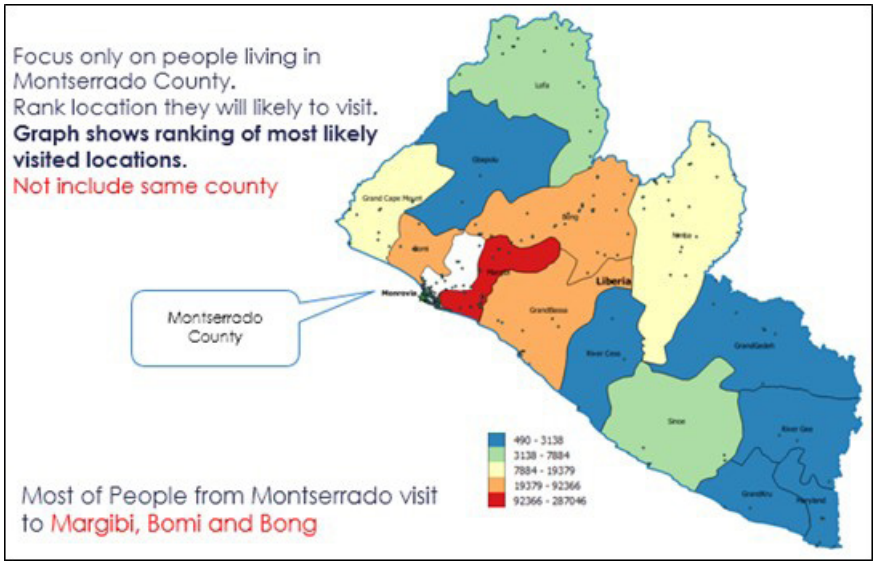
This section discusses the population distribution extracted from call detail record data. Figure 8.1 shows the number of mobile phone users computed from CDR data and population data by the national population census in the Republic of Liberia, which confirms that population of mobile phone users across districts correlates strongly with that of real populations obtained from census data. Figure 8.2 presents the most visited locations by the people living in Montserrado, revealing that Margibi, Bomi, and Bong are the most visited locations. Figure 8.3 shows Margibi is the second most visited location by the same people.

Figure 8.1: Mobile phone user population by county



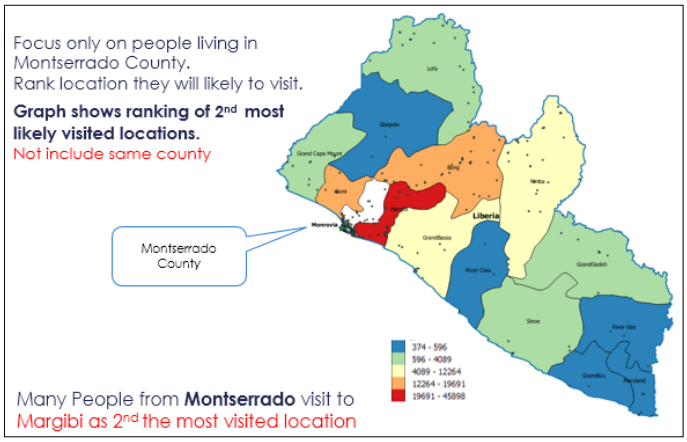
Source: ITU

Figure 8.2: Montserrado: most visited location



Source: ITU

Figure 8.3: Margibi: second most visited location

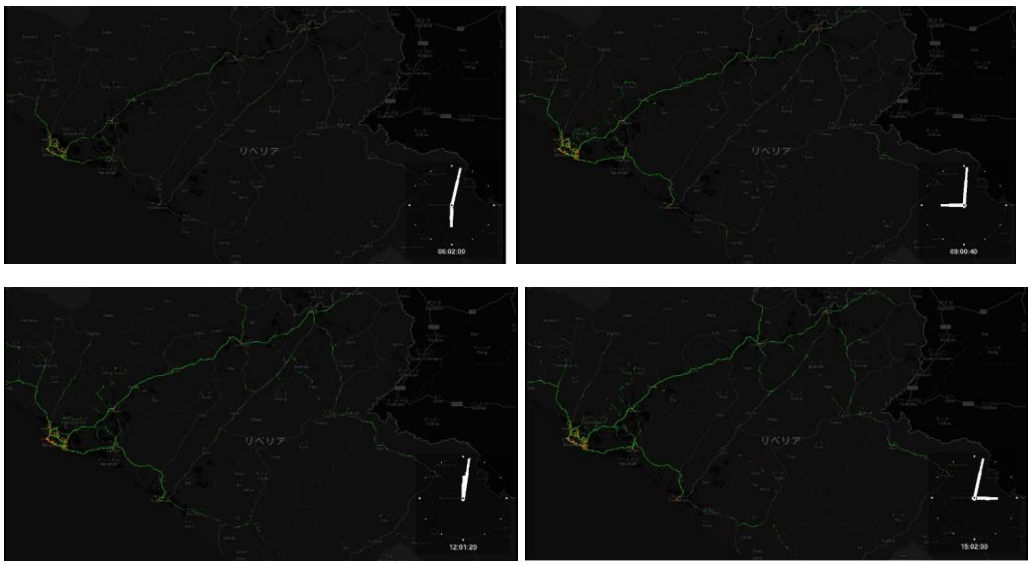


Source: ITU

8.2 People flow from interpolation result

This section provides the visualization of interpolated CDR data through the process described in Section 7. Figure 8.4 shows the people flow reconstructed from interpolation results at 6:00 am, 9 am, 12 pm, and 3 pm that shows less people in Monrovia in the early morning hours, while more people are observed 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. CDR data interpolated at one-minute intervals.

Figure 8.4: People flow from interpolation result by time

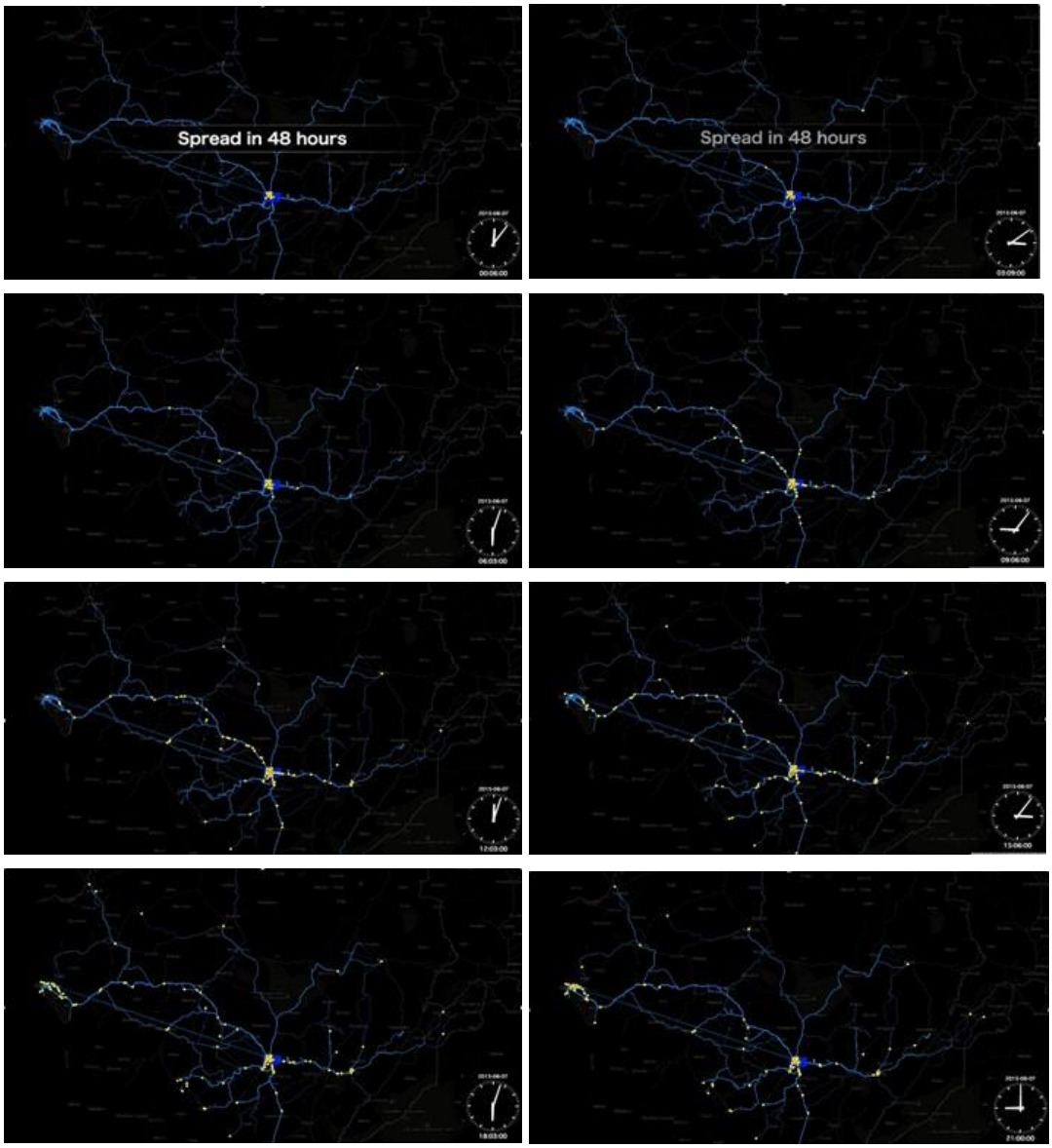


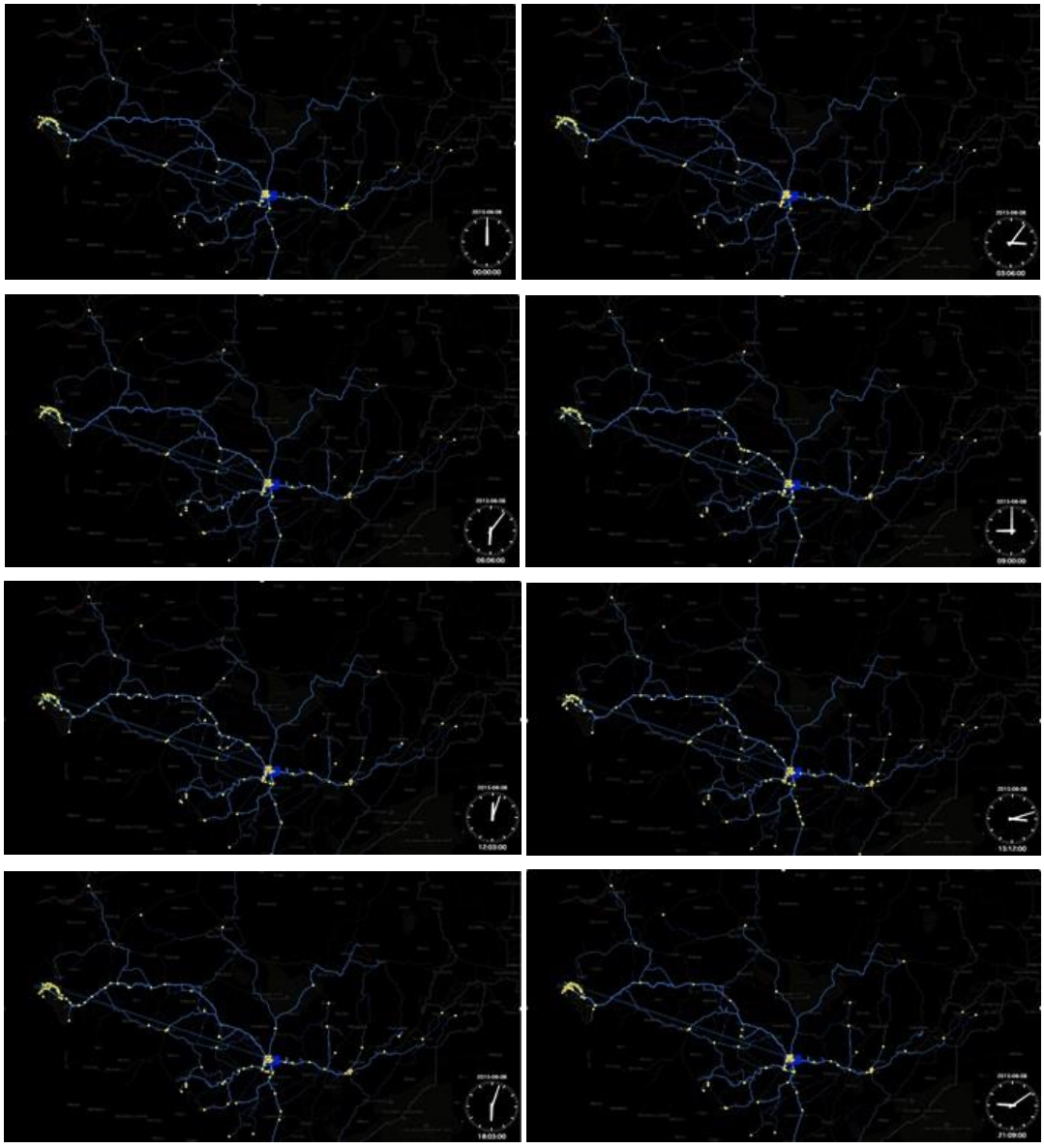
Source: ITU

8.3 Spread of people originated from a town (in 48 hours)

Figure 8.5 shows how people in a town spread at three-hourly interval over a 48 hour period. In a situation, where Ebola 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

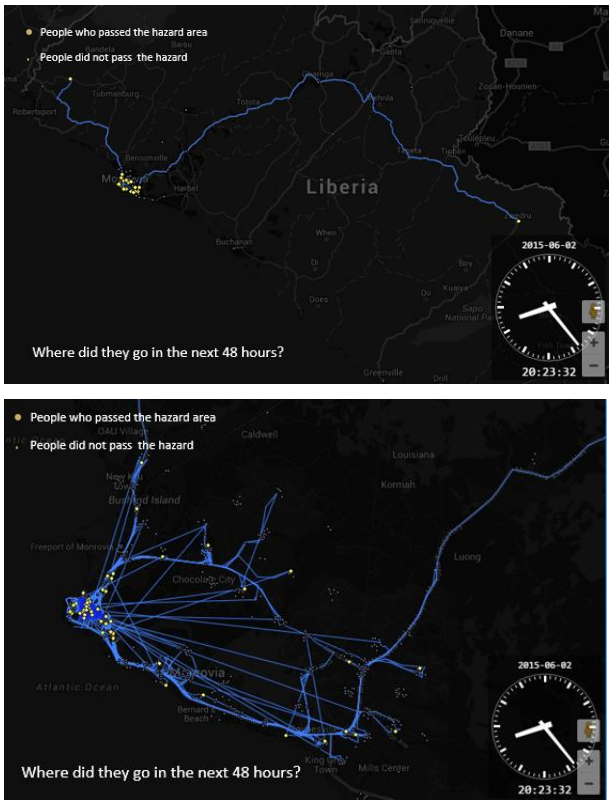
8.4 Tracing people who passed through a hazard area

Figures 8.6 and 8.7 show the distribution of people who passed/did not pass through a hazard area. By overlaying the hazard area with the human mobility estimated from CDR data, larger-yellow dots inside blue polygons indicate those who passed through, and smaller dots outside the polygons indicate those who did not pass through the hazard area.

Figure 8.6: Trace people who passed through a hazard area

Source: ITU

Figure 8.7: Where did they go in the next 48 hours?



Source: ITU

As also observed in Figure 8.7, population mobility can spread quite far from the first location even only for a short period of 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 identity (IMSI).

Figure 8.8: Cloning detection (cloning IMEI)

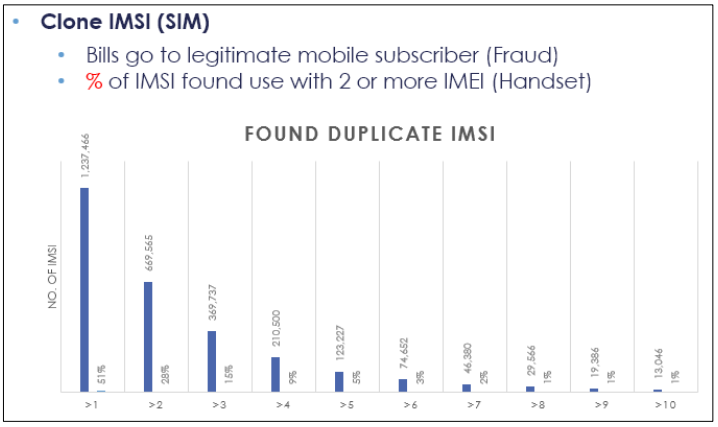
IMEI	TIME	PREV_TIME	SITE	PREV_SITE	IMSI	PREV_IMSI	DISTANCE (km)	TIME_DIFF (min)
c383	2015-06-01 11:12:01	2015-06-01 11:07:32	10181	11604	3f38796f1	2df42dfb8e88	51.90	4.48
c383	2015-06-01 11:55:07	2015-06-01 11:49:10	10145	10531	456b354b	1d4e0a862866	83.65	5.95
c383	2015-06-01 13:15:41	2015-06-01 13:07:32	10972	10314	21f2914eb	bd0cf07bdc85	75.66	8.15
c383	2015-06-01 15:45:51	2015-06-01 15:41:53	10973	10314	21f2914eb	bd0cf07bdc85	75.66	3.97
c383	2015-06-01 17:25:49	2015-06-01 17:24:00	10973	11643	21f2914eb	ccb37141d544	75.19	1.82
c383	2015-06-01 17:30:02	2015-06-01 17:25:49	31652	10973	2c67da6c9	21f2914eb506	73.99	4.22
c383	2015-06-01 17:34:28	2015-06-01 17:30:02	10972	31652	21f2914eb	2c67da6c925f3	73.99	4.43
c383	2015-06-01 20:17:17	2015-06-01 20:13:15	10972	11673	21f2914eb	db034c2b4f51	69.29	4.03
c383	2015-06-01 20:25:47	2015-06-01 20:17:17	10245	10972	e5bfd85d2	21f2914eb506	74.37	8.5
c383	2015-06-01 20:27:57	2015-06-01 20:25:47	10972	10245	21f2914eb	e5bfd85d21e4	74.37	2.17
c383	2015-06-01 20:50:52	2015-06-01 20:41:37	10086	10973	c30c7135d	21f2914eb506	75.45	9.25

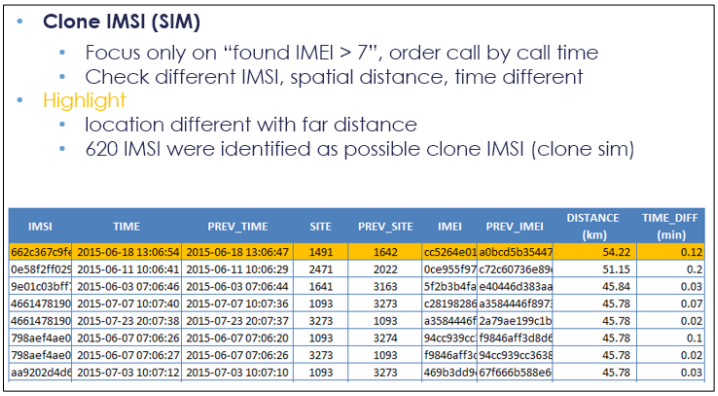
Source: ITU

As described in Figure 8.8 (upper), in the CDR data used for this project, the number of unique IMSI is almost 1.4 times the number of unique IMEI, and about 8 200 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 a minute of making a second call. Processing CDR data with IMSI and IMEI can detect cloning and this method was used for the purposes of this project.

Figure 8.9: Cloning detection (cloning IMSI)





IMSI	TIME	PREV_TIME	SITE	PREV_SITE	IMEI	PREV_IMEI	DISTANCE (km)	TIME_DIFF (min)
662c367c9fe	2015-06-18 13:06:54	2015-06-18 13:06:47	1491	1642	cc5264e01	a0bcd5b35447	54.22	0.12
0e58f2ff025	2015-06-11 10:06:41	2015-06-11 10:06:29	2471	2022	0ce955f97	c72c60736e89	51.15	0.2
9e01c03bff7	2015-06-03 07:06:46	2015-06-03 07:06:44	1641	3163	5f2b3b4fa	e40446d383aa	45.84	0.03
46614781900	2015-07-07 10:07:40	2015-07-07 10:07:36	1093	3273	c28198286	a3584446f897	45.78	0.07
46614781900	2015-07-23 20:07:38	2015-07-23 20:07:37	3273	1093	a3584446f	2a79ae199c1b	45.78	0.02
798aef4ae0	2015-06-07 07:06:26	2015-06-07 07:06:20	1093	3274	94cc939cc	f9846aff3d8d6	45.78	0.1
798aef4ae0	2015-06-07 07:06:27	2015-06-07 07:06:26	3273	1093	f9846aff3c	94cc939cc3638	45.78	0.02
aa9202d4d6	2015-07-03 10:07:12	2015-07-03 10:07:10	1093	3273	469b3dd9	67f666b588e6	45.78	0.03

Source: ITU

Figure 8.9 (upper) shows an attempt to detect SIM card cloning: there are more than 1.2 million unique sim cards that are being used in at least two handsets over the same period, and more than 46 000 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 (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, 620 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 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 Liberia, Guinea, and Sierra Leone, were merged to trace transboundary movements. After the merge, each user's home country is identified based on the frequency of called locations.

8.6.1 CDR dataset 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. Mobile network operators, Lonestar, Cellcom, and Novafone, provided CDR data (with hashed IMEI codes) that covered the movement trends of the entire mobile subscriber base.

Figure 8.10: CDR data set with transboundary IMEI of Liberia

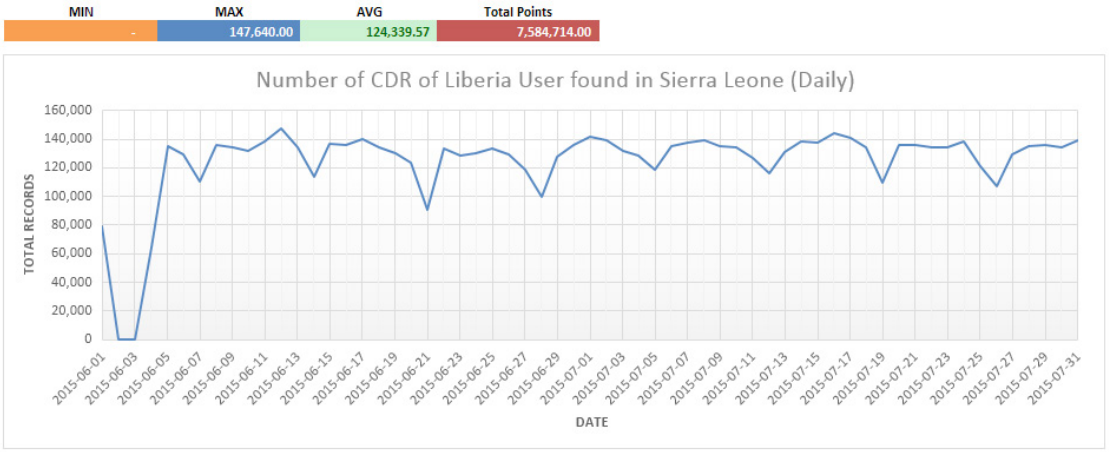
NO	Country	Operator	CDR Data Info				Transboundary (IMEI)		
			Data Size	Data Period	Total record	IMEI (caller)	Sierra Leone	Guinea	Sierra Leone & Guinea
1	Liberia	Lonestar	68.2 GB	June,July 2015	431,934,109	1,708,830	29,846	17,275	2,792
2		Cellcom	84.9 GB	June,July 2015	204,715,956	1,576,947	286	24	12
3		Novafone	1.7 GB	June,July 2015	8,220,071	126,091	1,183	1,016	258

Source: ITU

8.6.2 Daily activity statistics

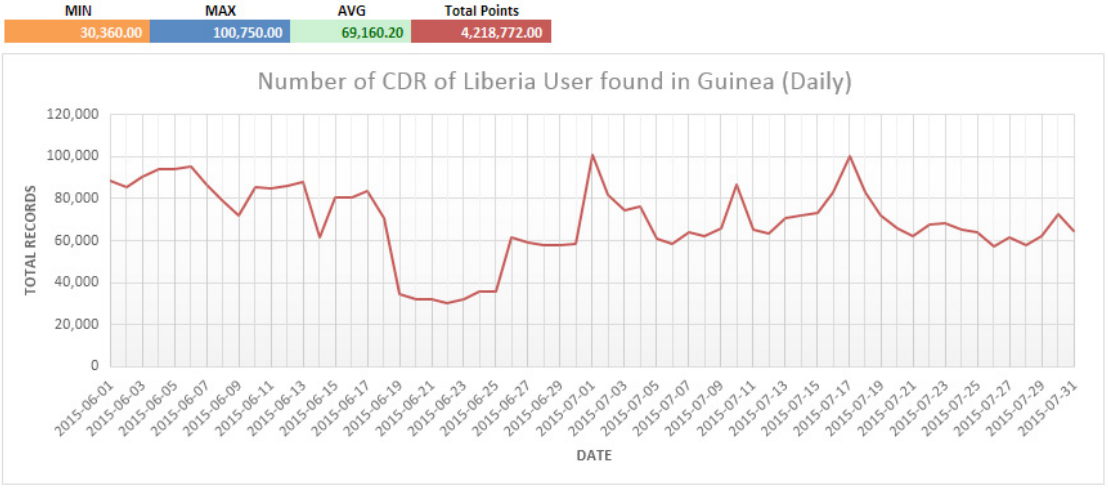
Figure 8.11 and 8.12 illustrate the daily call activity of Liberia based mobile phone subscribers being registered on neighbouring country networks (Guinea and Sierra Leone), where daily call activity reached an average of approximately 124 000 call records in Sierra Leone, compared to daily call activity of 70 000 in Guinea (Figure 8.12).

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



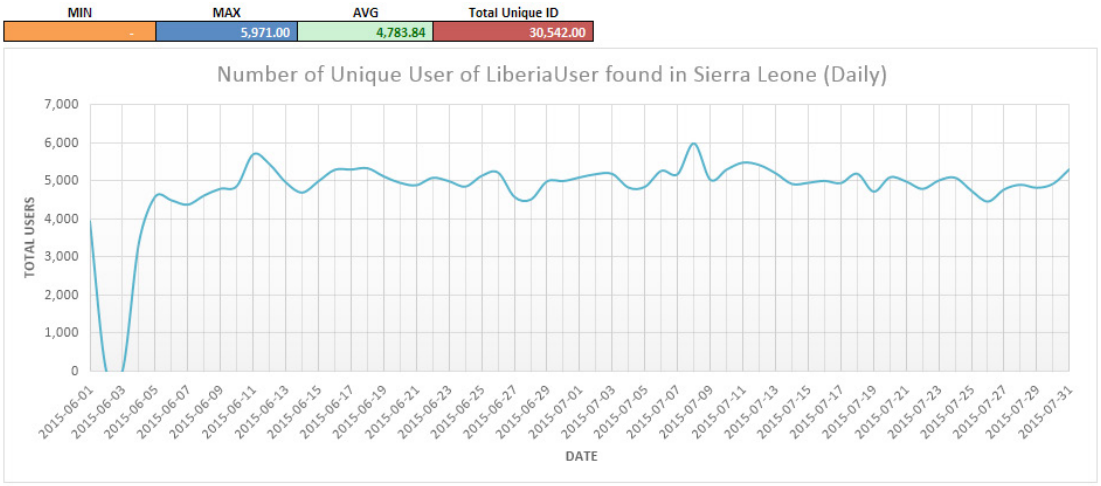
Source: ITU

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



Source: ITU

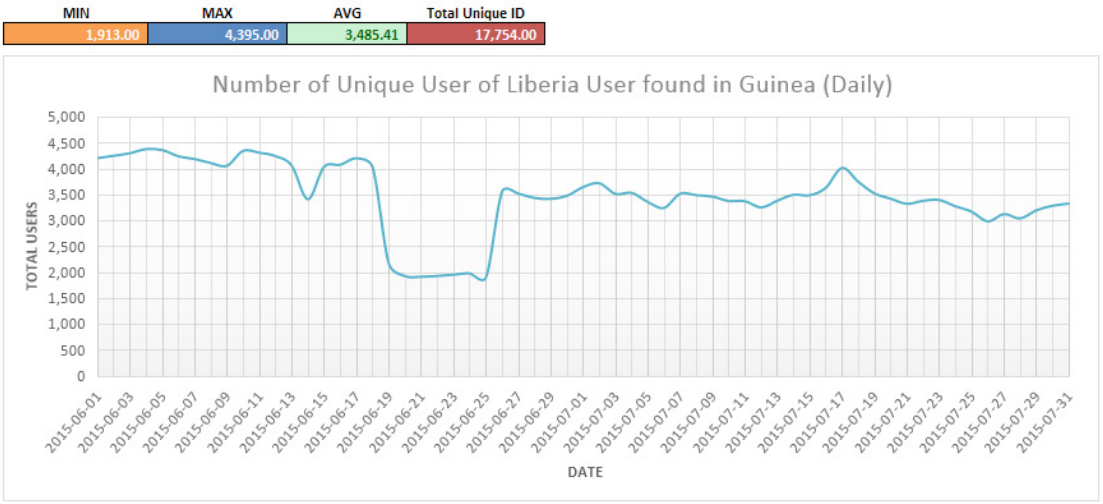
Figure 8.13: Daily unique user call activity of Liberia mobile subscribers visiting Sierra Leone



Source: ITU

Figures 8.13 and 8.14 illustrate the number of daily unique Liberia mobile users in Sierra Leone and Guinea, averaging approximately 5 000 users in Sierra Leone and 4 000 users in Guinea.

Figure 8.14: Daily unique users from Liberia in Guinea

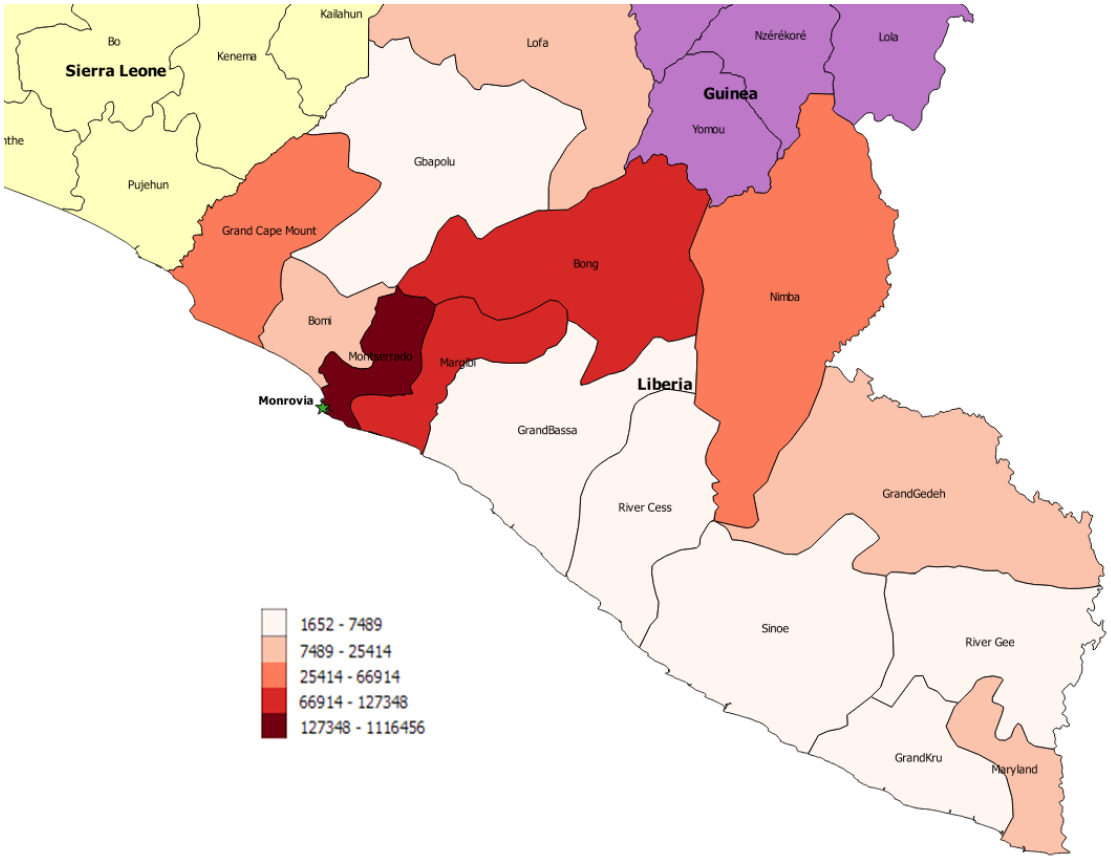


Source: ITU

8.6.3 Inbound transboundary population movement

Figure 8.15 illustrates footprints of inbound transboundary movements from Sierra Leone and Guinea to Liberia. Footprints were calculated from mobile usage activity of mobile users from Sierra Leone and Guinea, which indicate locations where users have passed through and where they have high mobile usage activities. The most visited county is Montserrado, where the capital city Monrovia of Liberia is located. The second and third rank of frequent visits are the neighbouring counties of Margibi and Bong.

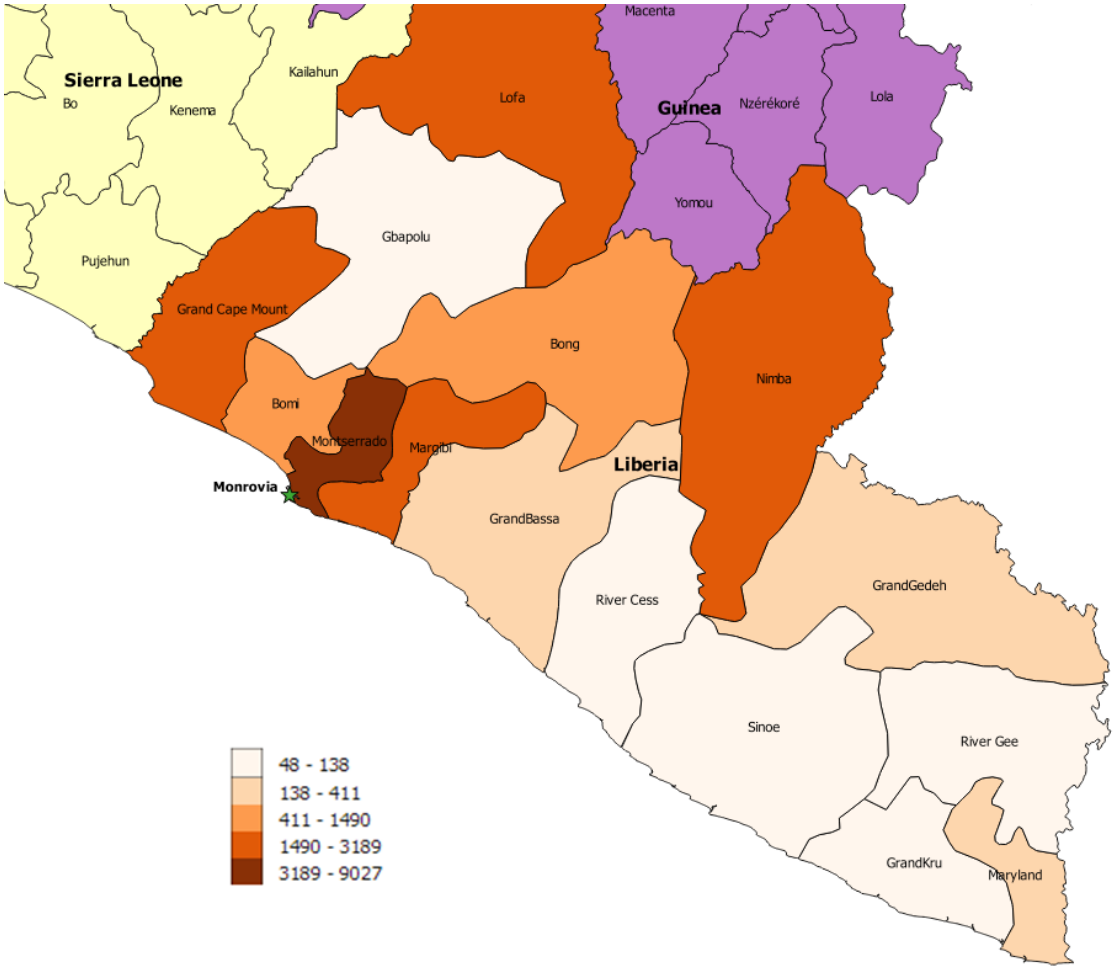
Figure 8.15: Footprint of mobile users from Sierra Leone and Guinea visiting Liberia



Source: ITU

Figure 8.16 illustrates the transboundary population of mobile users from Sierra Leone and Guinea who visited Liberia. Population was calculated from the total number of unique mobile users who visited Liberia. It shows that the most visited counties by inbound transboundary population to Liberia are Montserrado, Grand Cape Mount, and Margibi. Additionally, Grand Cape Mount seems to be the main boundary county for crossing from Sierra Leone into Liberia. For Guinea, Nimba seems to be the main boundary county for those crossing from Guinea to Liberia.

Figure 8.16: Number of mobile users from Sierra Leone and Guinea visiting Liberia

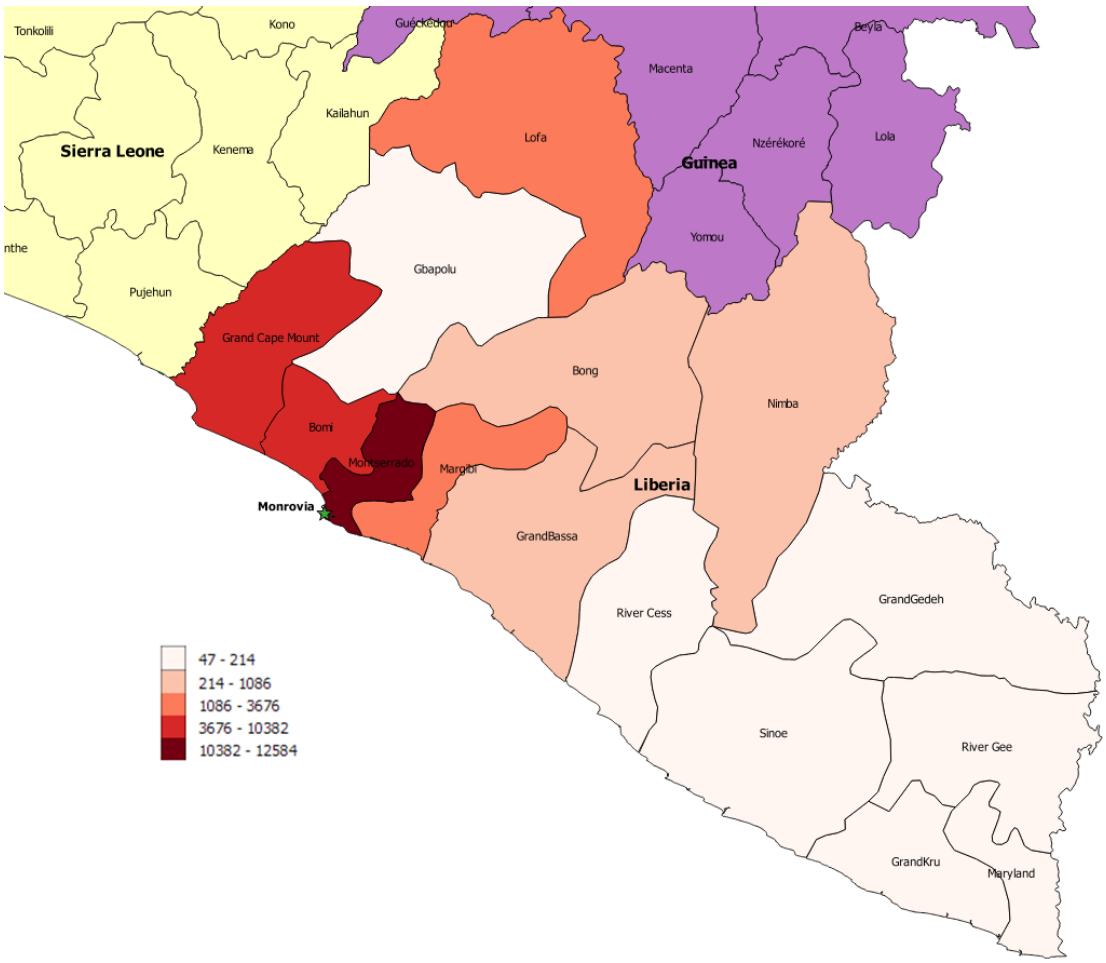


Source: ITU

8.6.4 Outbound transboundary population movement

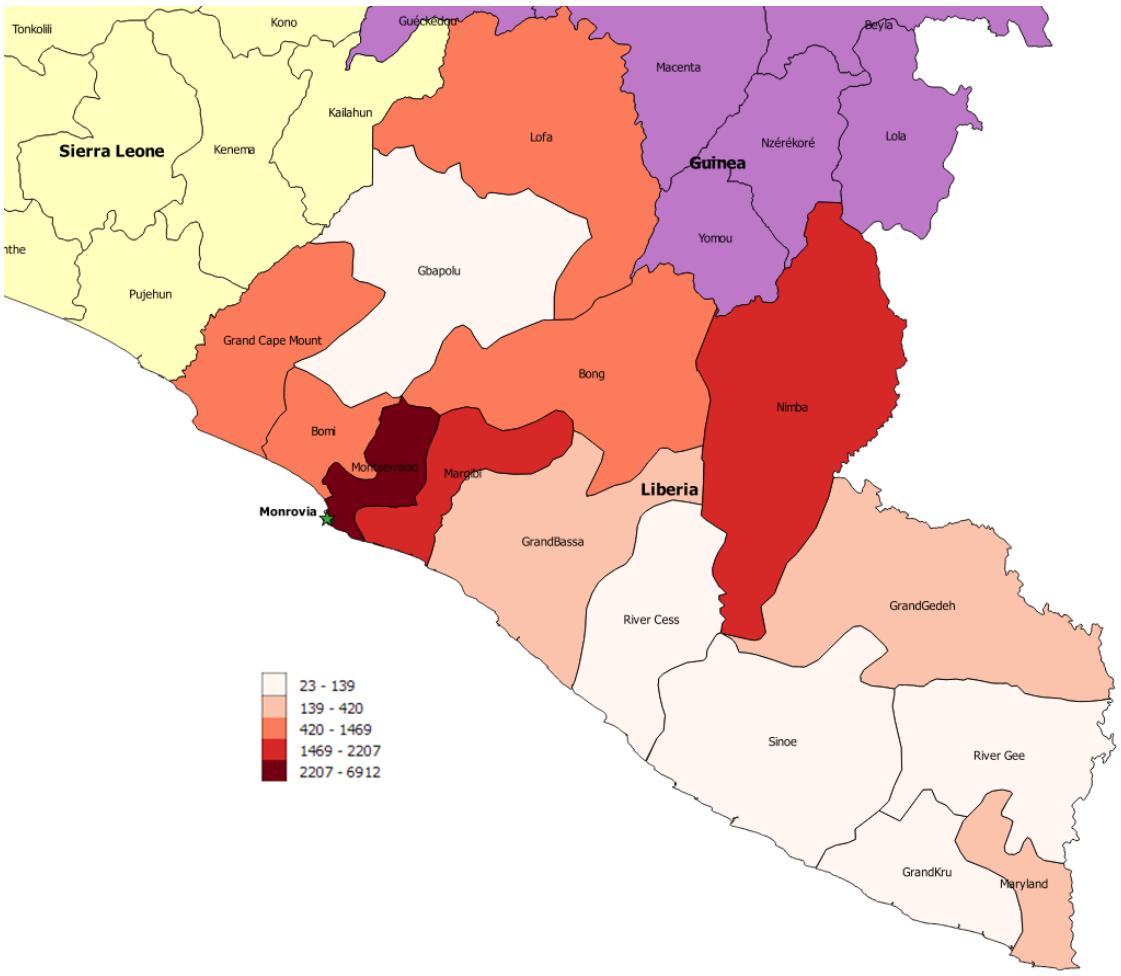
Figure 8.17 illustrates the transboundary movement of mobile users visiting Sierra Leone from Liberia. It shows that the three most visited counties are Montserrado, Grand Cape Mount, and Bomi.

Figure 8.17: Footprint of visitors to Sierra Leone



Source: ITU

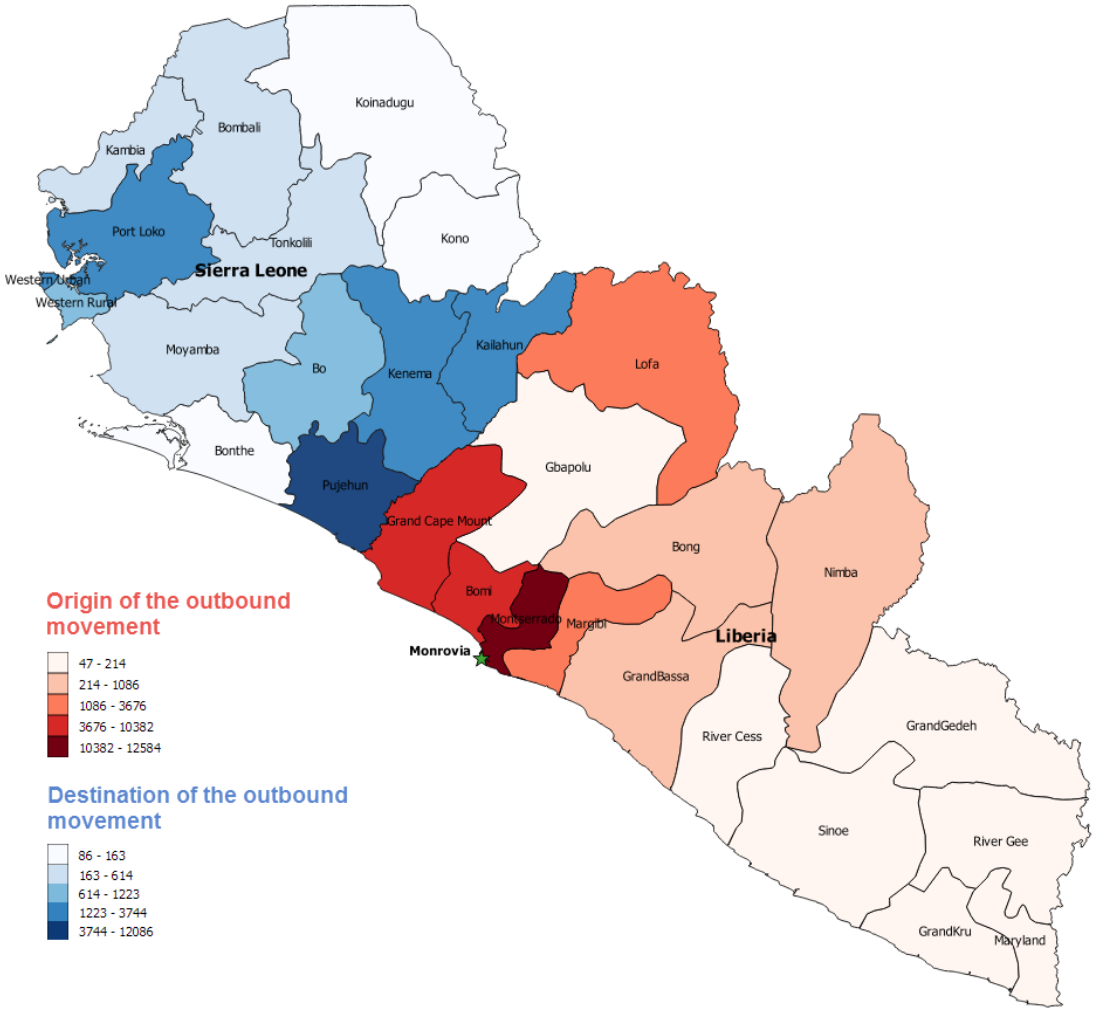
Figure 8.18: Footprint of visitors to Guinea



Source: ITU

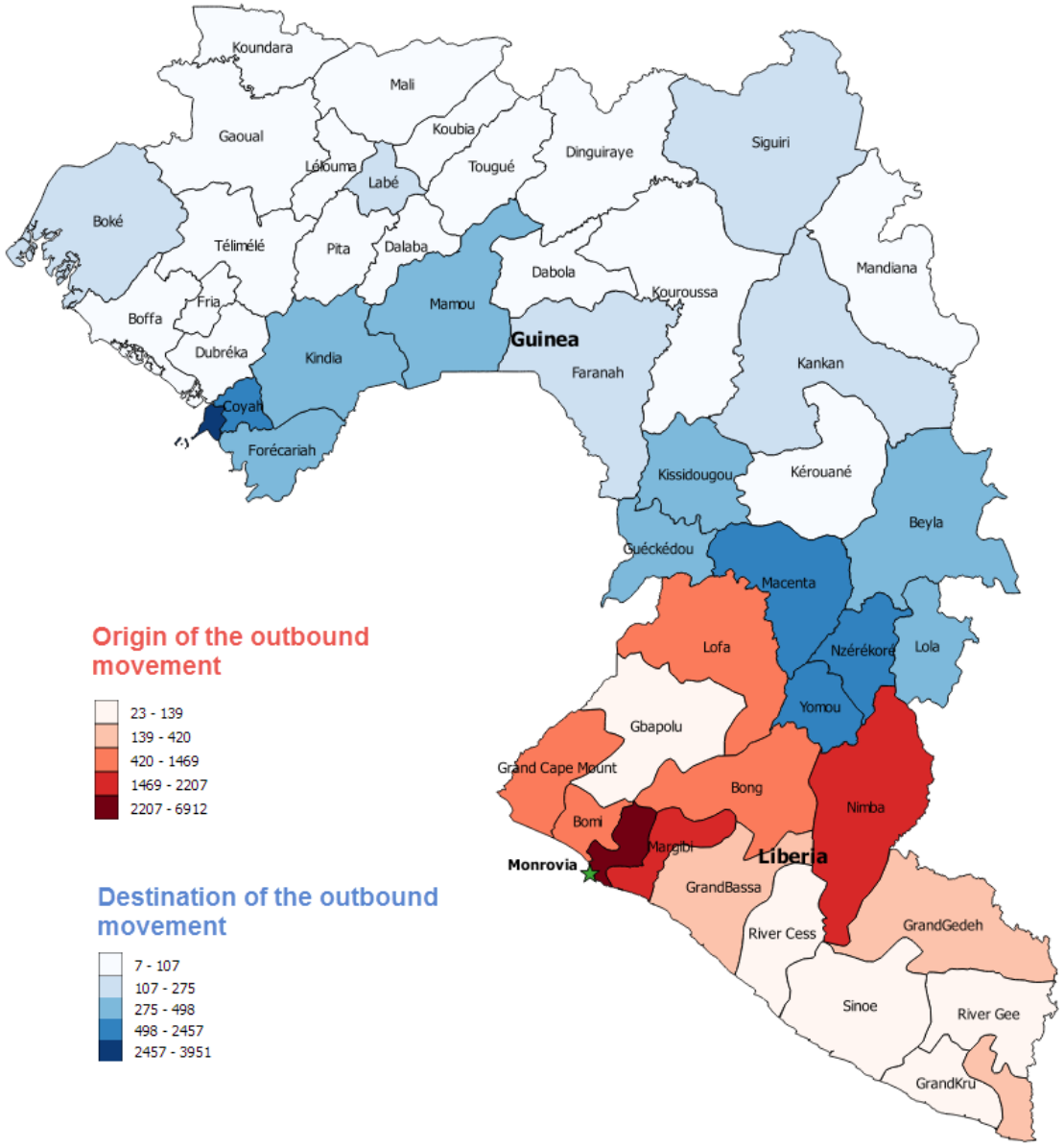
Figure 8.18 illustrates the transboundary population of mobile users visiting Guinea from Liberia, where the three most visited counties (with high transboundary movements) are Montserrado, Margibi, and Nimba. Figures 8.19 and 8.20 illustrate target areas of transboundary population visiting Sierra Leone and Guinea.

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



Source: ITU

Figure 8.20: Origin and destination of transboundary population to Guinea



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 Liberia 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 Liberia, 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.

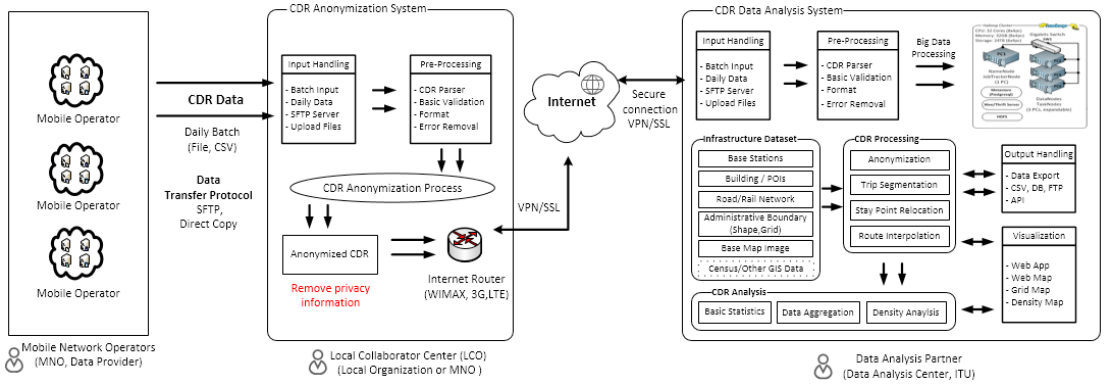
From the operator perspective, any data release involves both effort and risk, including contractual constraints over release of customer data, or at least a perceived loss of consumer trust if records are transferred outside company control. Nevertheless, operators may well release data either before 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 A1.1: Overall concept of automated CDR data analysis



Source: ITU

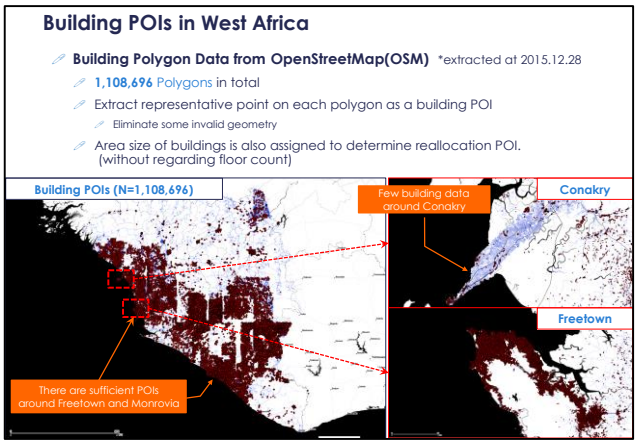
For this project, the entire process still required manual operation such as input handling, pre-processing and data analysis. The automated CDR data analysis system allows the whole system to work automatically with less operator involvement. This system is recommended in case of expansion to large-scale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

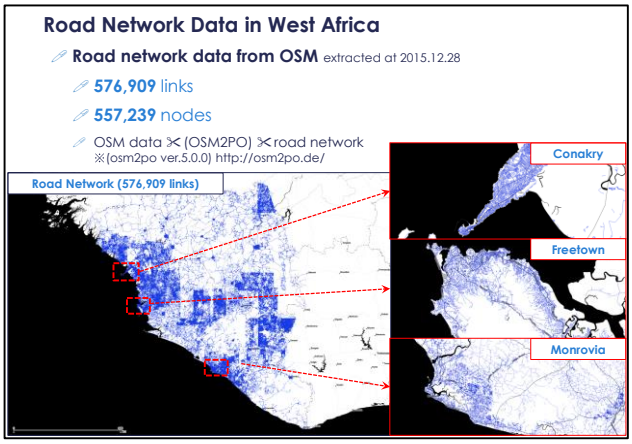
Figure A2.1: Building POI in West Africa



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.

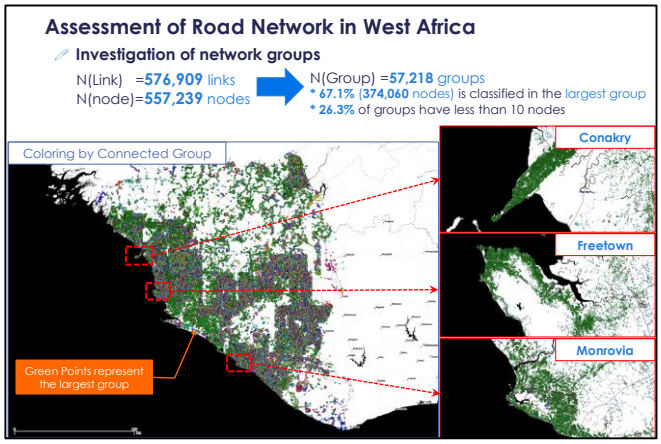
Figure A2.2: Road network data in West Africa



Source: ITU

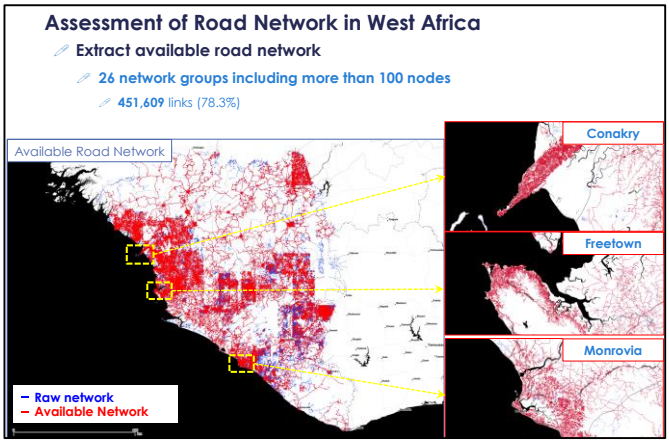
[8] https://www.openstreetmap.org/#map=5/51.500/-0.100

Figure A2.3: Assessment of road network in West Africa



Source: ITU

Figure A2.4: 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 Environmnent 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, 3e 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, 4e 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"
11º andar, Ala Sul
Ed. Luis Eduardo Magalhães (Anatel)
70070-940 Brasilia, 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: ituteguciglapa@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: itumoskow@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: 978-92-61-20291-0

Printed in Switzerland
Geneva, 2017