Making the Most of Mobility Data (CDRs):

A GUIDE FOR POLICYMAKERS

How can mobility data help governments respond during a public health crisis? Drawing from our work applying telecoms mobility data to the Covid-19 response in Sierra Leone, this guide outlines technical requirements for using call detail records for voice call and SMS, highlights when this data are most useful in crises, and provides guidance on how to work with external experts in service of local needs and policy priorities.

CIVIC DATA DESIGN LAB

MITGOV/LAB

Suggested Citation:

Ndubuisi-Obi Jr., Innocent. 2021. “Making the Most of Mobility Data (CDRs): A Guide for Policymakers”, MIT Governance Lab and Civic Data Design Lab (United States).

MIT Gov/Lab

MIT Governance Lab is a group of political scientists focusing on innovation in citizen engagement and government responsiveness. *www.mitgovlab.org.*

CDDL

The Civic Data Design Lab at MIT translates data into tools for policy change. *http://civicdatadesignlab.mit.edu.*

Acknowledgments

Thanks to our partners at Sierra Leone’s Directorate of Science, Technology & Innovation (DSTI), Africell-Sierra Leone, and Flowminder Foundation, including Aurelie Jeandron, Veronique Lefebvre, and Xavier Vollenweider. And, thanks to Sarah Williams, Lily L. Tsai, Alisa Zomer, Will Sullivan, and Maggie Biroscak for providing input into various versions of this guide. Illustrations and graphic design by Susy Tort and Gabriela Reygadas.

6 Introduction
8 Background: Mobility Study in Sierra Leone
10 Getting Started: Team, Data Agreements, and Project Goals
15 Step 1: Properly Accessing and Securing Data
21 Step 2: Checking Data Quality
26 Step 3: Facilitating Analysis and Communicating Results
32 Conclusion: Data Ethics, Privacy, and Limitations of CDRs
Data Ethics and Privacy
36 Limitations of CDR Data
37 Final Thoughts
38 Glossary of Terms

Mobility Data

INTRODUCTION

Understanding how, when, and where people move is especially important during a public health crisis. In many countries in the Global South, there is a lack of robust transit or mobility data to help ground-truth mobility patterns. Even in the cases where these data exist, they usually contain one-time snapshots rather than time series views of mobility.

That's why call detail records (CDRs) have become a popular source of mobile phone data for researchers and policymakers. **CDRs** are logs of network events (voice call or SMS) that are made on the network of a telecommunications provider. Each CDR includes the cell tower the subscriber was closest to, so the locations of cell towers can be used to approximate population mobility patterns. Importantly, CDRs are automatically generated by mobile network operators, so they do not require additional infrastructure or effort to collect like other types of mobile phone data do.[1]

Call Detail Records (CDRs): are automatically generated by the telecommunications equipment whenever a voice call or SMS is made or received by a subscriber in a telecommunications network.

FIGURE 1: **Sample CDR record**

Cell-Id	Source	Call Time	Duration	Direction	Event	Target
12103	21D7B5	2020-03-04 20:00:10	00:29:57	Incoming	Call	65H8I9

Cell id is a unique identifier representing the antenna or cell where a call was made or received. Source and Target are unique identifiers for subscribers generated by some anonymization or hashing procedure. Duration is the length of a CDR event. Direction logs if the call or SMS was initiated by the source. Event encodes whether the event is a voice call or sms.

1 In the review of the types, metrics, and applications of mobile phone data, Grantz et al. 2020 document four types of mobile phone data: CDR data, GPS location data, Bluetooth data, and Opt-in application data. For more see, Grantz, K. H., Meredith, H. R., Cummings, D. A. T., Metcalf, C. J. E., Grenfell, B. T., Giles, J. R., ... Wesolowski, A. (2020). The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nature Communications.*

Research using CDRs ranges from mapping population movement to understanding migration patterns, exploring community structure, and understanding disease spread. [2] These kinds of studies demonstrate that CDR data can be used to understand population mobility patterns, the temporal dynamics of mobility, and the relationships between socio-economic status and mobility.

This guide provides an introduction of how to use CDRs to inform public policy response during a crisis. It draws on our research team's experience working with CDRs in Sierra Leone during the early stages of the Covid-19 pandemic. We have distilled **our process into three steps** that any team interested in using CDRs should consider:

Steps

1 → **Properly accessing and securing data.** Step 1 entails getting access to the CDR dataset and storing it in a secure environment. Another important task here is ensuring that no sensitive or personally identifiable information is released — such as a name, address, or phone number.

2 → **Checking the data quality.** Step 2 focuses on ensuring that the records are valid, accurate, complete, and consistent.

3 → **Facilitating analysis and communicating results.** Step 3 is where the analytics team chooses the types of analysis they will run and configures an environment to support the processing of the data and the communication of the results.

The goal of this guide is to provide a detailed overview for non-technical policymakers who want to learn more about how CDR analysis works and what to expect throughout the process. It covers technical details of the analysis, with the aim of helping all partners collaborate effectively and understand their roles. The last section of the guide discusses data ethics, privacy, and the limitations of CDR data.

2 Zhou, C., Xu, Z. and Huang, B. (2010) 'Activity Recognition from Call Detail Record: Relation Between Mobile Behavior Pattern and Social Attribute Using Hierarchical Conditional Random Fields', in 2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing. Int'l Conference on Cyber, Physical and Social Computing (CPSCom), Hangzhou, China: IEEE, pp. 605–611. Wesolowski, A. et al. (2012) 'Quantifying the impact of human mobility on malaria', Science (New York, N.Y.), 338(6104), pp. 267–270. doi: 10.1126/science.1223467; Doyle, C. et al. (2019) 'Predicting complex user behavior from CDR based social networks', Information Sciences, 500, pp. 217–228.

BACKGROUND: MOBILITY STUDY IN SIERRA LEONE

Sierra Leone reported its first case of Covid-19 on March 30, 2020. Soon after, the Government of Sierra Leone (GoSL) implemented a three-day lockdown, followed by a 14-day, inter-district travel ban. As the pandemic developed, the government wanted to understand the effectiveness of their mobility restrictions and subsequent policy actions—did the lockdown measures actually change people's mobility patterns?

To find out, a research collaboration was formed between GoSL's Directorate of Science, Technology and Innovation (DSTI), Africell (the mobile network operator with largest market share), and researchers at the Massachusetts Institute of Technology's Civic Data Design Lab and MIT Governance Lab (MIT GOV/LAB) to answer the following questions:

- **Did the three-day lockdown and the travel ban decrease mobility in Sierra Leone?**
- **If there was a change, what were the differences across districts?**

The researchers tackled these questions using call detail records (CDRs), logs of cell tower pings that can serve as a proxy for mobility. A summary of those findings with interactive data visualizations is available online[3], while a book chapter in "Urban Informatics and Future Cities"[4] covers the research in more detail.

3 Ndubuisi-Obi Jr, I. (2021). Were Lockdowns Effective in Sierra Leone? Mobility Data Shows Compliance. Retrieved from https://mitgovlab.org/updates/were-lockdowns-effective-in-sierra-leone-mobility-data-shows-compliance/.

4 Urban Informatics and Future Cities. (2021). Retrieved from https://www.springer.com/gp/book/9783030760588#aboutBook.

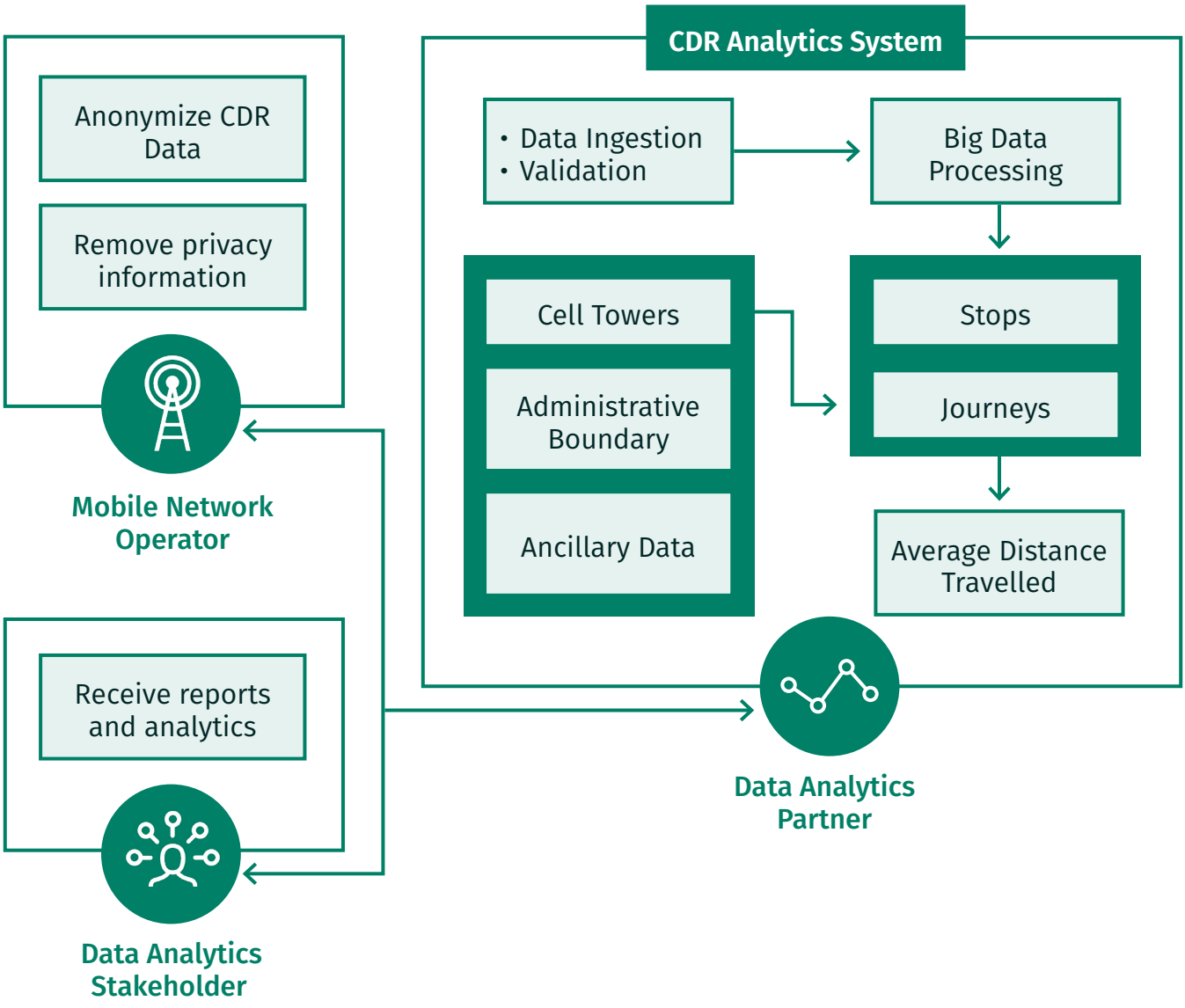
Mobility Data

Bo, Sierra Leone. Photo by Joshua Hanson via Unsplash

GETTING STARTED: TEAM, DATA AGREEMENTS, AND PROJECT GOALS

In our experience, there are particular roles that are important to the success of any CDR analytics project. These roles are the Data Analytics Stakeholder, the Mobile Network Operator, and the Data Analytics Partner (Figure 2).

FIGURE 2: **A typical CDR analytics process**



Mobility Data

The **Data Analytics Stakeholder (DAS)** is the primary client and recipient of the insights generated by the Data Analytics Partner. They initiate and define the scope of the CDR analysis — from selecting partners to determining research questions. In many cases, they are also critical in providing ancillary data that the Data Analytics Partner might need, like census data, statistical surveys, and health data.

The **Mobile Network Operator (MNO)** is the data owner, typically a telecommunications company. They are the creators of the CDR data and have the ultimate legal responsibility over its use and distribution. It is their primary role and responsibility to prepare the data for use by de-identifying and anonymizing the dataset. Once they have done this, they work closely with the Data Analytics Partner to secure appropriate access and use of the CDRs.

De-identification: is a process by which identifiers/ attributes that are known to increase the risk of identification of a person or group are removed from data.

The **Data Analytics Partner (DAP)** is responsible for preparing the analytics environment and performing the requested analytics on the CDRs. Throughout the analytics process, the DAP works in collaboration with the MNO and the DAS to ensure they have the domain knowledge and ancillary data needed to carry out CDR analysis. They communicate directly with the DAS and perform analyses that can support them in answering policy-related questions.

Data sharing agreement: is an agreement where partners explicitly acknowledge and address sensitive concerns regarding the sharing and use of their private data and establish expectations and processes that ensure the data are secure. They usually cover topics such as how and with whom data will be shared and protocols for working with the data and communicating results and findings.

← *The role of **data sharing** and use agreements*

Even with these roles, there isn't a single template for building partnerships for a CDR analytics project. In some cases these projects are developed to study purely academic research questions and in other cases they are developed to address policy questions from a governmental body. The MNO and DAP can be the same organization or a consortium of organizations. The MNO could also be a Data Analytics Stakeholder. An analytics project could include CDRs from multiple MNOs and require the engagement of multiple DAS and DAP actors.

That's why a data sharing agreement is important for formalizing and making explicit the three roles and their responsibilities. By clearly defining the roles and responsibilities of each partner, it becomes easier to understand each partner's ethical and legal obligations. Data sharing agreements can aid in answering the following questions: Who has access to the data? Who gives consent to the use of data? What can be published? What types of analysis can and should be performed?

For more details on how to set up a successful research collaboration, check out MIT GOV/LAB's engaged scholarship tools[5] and Box 1.

Sierra Leone. Photo by Annie Spratt via Unsplash

5 Zomer, A., & Lipovsek, V. (2020). How to Have Difficult Conversations. Retrieved from https://mitgovlab.org/resources/updated-guide-how-to-have-difficult-conversations/.

BOX 1.

WHAT CAN A CDR ANALYTICS PARTNERSHIP LOOK LIKE?

Example from Sierra Leone

Part of setting up a team is ensuring that each party has the information they need, without having "too many cooks" or potential privacy/data leakages.

In the example from Sierra Leone, GoSL's Directorate of Science, Technology, and Innovation was the Data Analytics Stakeholder. Africell was the Mobile Network Operator. And our team at MIT filled the role of the Data Analytics Partner.

MIT had already spent considerable time with DSTI and Africell to ensure that we had access to anonymized and de-identified data. So when colleagues from the Flowminder Foundation also joined the research effort, they did not need direct access to anonymized and de-identified CDR data.

Since MIT already had a secure computing environment set up for data storage and processing, we worked together with Flowminder to run their SQL queries to generate the aggregate statistics needed for their reports. There are two unique aspects of this partnership.

1/ **Multiple Data Analytics Partners.** While both teams generated reports using CDRs for DSTI, we were focused on very different statistics. Our team developed a more fine-grained measure of mobility, and the Flowminder team generated coarse-grained analytics. Our two analyses were complementary and allowed our teams to work in parallel while also collaborating to compare our results before sharing with the DAS.

2/ **Bilateral Data Use Agreements.** Both the MIT and the Flowminder teams created separate data sharing agreements with the MNO and DAS. Even though Flowminder could have gained access to CDR aggregates without needing an agreement in place, they felt it was useful to have a separate agreement formally outlining the ethical and legal expectations of their use of the data.

The partnership reduced duplication and wasted effort. We wanted to work together to ensure that DSTI, as the Data Analytics Stakeholder, received the best and most relevant analysis to aid and guide their policy decisions. In addition, our actions streamlined issues related to data quality and also ensured that Africell only needed to interact with one partner on matters related to the data.

Is CDR analysis right for you?

As you go through the guide, please consider the following questions to determine if you should use CDRs instead of other types of mobile phone data to answer mobility-related policy questions.

1. **Do you need fine-grained information on individual patterns, or are population-level estimates enough?** *For example, if you need to understand how infected individuals contributed to the spread of a virus, then CDRs will have little utility. On the other hand, if you wanted to understand if areas with more mobility had fewer cases then CDRs are a perfect fit.*

---

2. **Do you already have access to a source of mobile phone data? If not, which source will be the quickest to gain access to?** *In some cases, it might be easier to use available data than to create new agreements and partnerships to gain access to new data. If time is a constraint, go with the data that is quicker to access.*

---

3. **Are there regulations that limit the type of data you can have access to?** Are there restrictions based on data privacy? *Each nation and organization faces different regulations that govern the release of data. These regulations might not allow the use of CDRs and might favor other types of mobile phone data.*

We'll get into more detail on the technical aspects of this decision later.

Step #1

PROPERLY ACCESSING AND SECURING DATA

This section introduces you to the basic process of getting access to CDR data and the type of computing and storage environment needed to securely use the data. We review the difference between on-premise and cloud computing solutions and discuss the various factors that might make one more suitable than another.

Before any analysis begins, it is important that the Data Analytics Partner (DAP) and the Mobile Network Operator (MNO) work together to ensure that the CDR dataset is securely accessed and stored. In most cases, the MNO provides anonymized CDR data. In this scenario, a data sharing agreement will detail the standards for how the anonymized dataset should be stored, accessed, and distributed. In cases where the MNO provides the DAP with direct access to raw CDR records, it is critical that the DAP de-identify and anonymize the CDRs before any data processing.

An important aspect of providing secure access to the CDR data is choosing where data is stored. Due to their sensitive nature, CDR records must be stored in a computing environment where access and authorization policies can be easily managed. Whether in government or industry, organizations have internal and external protocols governing how data should be secured and accessed. In both cases, IT administrators for the Data Analytics Stakeholder (DAS), DAP, and MNO have the best understanding of the systems and protocols that make sense for storing, accessing, and analyzing CDR data. The IT administrators at the DAP will understand the limitations of their chosen computing environment and the internal protocols like those recommended by an Institutional Review Board (IRB). The administrator at the DAS and MNO both might have regulatory and institutional frameworks that guide the use and release of CDR data.

Step #1

On-premises: software and services that are installed and run on computers on the premises of the organization using the software, rather than a remote server operated by another organization or cloud provider. It refers to hardware and software that are within the organization's internal systems.

Cloud: computing resources, especially data storage and compute, that are owned, operated, and maintained remotely and require no direct active management by the user.

Computing resources: technological and human inputs needed to carry out a computing task.

Below we highlight two approaches to storage and computing: an **on-premises** and a **cloud** solution and their implications for the DAS.

- **On-premises.** The MNO or the DAP might be required by regulators to store data and run services on its premises, or they might just prefer to do so. An organization that uses an on-premises solution has full control over its storage and computing decisions. Below we highlight two different approaches to on-premises CDR processing.

←— For example, in a research project conducted in Pakistan, researchers worked with **computing resources** provided by the research arm of a major telecommunications provider Telenor Research. This decision was taken in accordance with Telenor's privacy policies and national laws and regulations related to data privacy. The researchers highlight that "the CDR/mobility data were processed on a backup and recovery server made available by Telenor Pakistan. Only Telenor employees have access to the detailed CDR/mobility data."[6] This research was carried out completely within the premises of Telenor, which provided researchers with the computing resources needed to access and query CDRs.

The second example of an on-premises solution is the Flowminder Foundation's Flowkit,[7] an open-source toolkit for facilitating the analysis of CDR data aimed at development and humanitarian practitioners. Flowminder has developed, validated, and operationalized the use of mobile operator data to monitor population displacement in a humanitarian emergency and to use in epidemiology and public health, urban planning, service access assessment, and migration. When using FlowKit, the data would remain with the MNO and the FlowKit system would simplify the process of running SQL queries on the CDRs for the DAP. This type of solution builds capacity in the MNO in providing analytical support for CDR analysis and assuages concerns related to data sharing and privacy. In cases where the DAS is concerned about data privacy, FlowKit is a great on-premises solution.

6 Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., ... Singer, B. H. (2015). Impact of human mobility on the emergence of dengue epidemics in Pakistan. Proceedings of the National Academy of Sciences of the United States of America.

7 Foundation, F. (n.d.). FlowKit. Retrieved from https://flowkit.xyz/.

- **Cloud-based.** Vendors such as Amazon Web Service (AWS), Microsoft Azure, and even specialized providers like Databricks provide computing resources that simplify the deployment of a big data analytics environment. Rather than having to install, manage, and maintain on-premises computing resources, these vendors provide managed environments that can be provisioned as needed. In the case of AWS, a diversity of services are available: S3 provides secure storage, EC2 provides access to virtual machines, and EMR provides a managed Hadoop cluster that can run a number of data analytics and processing tools. Box 2 provides an example of a CDR analysis carried out by researchers at the World Bank. It highlights their use of Microsoft Azure as their chosen cloud computing platform.

Choosing between on-premises and cloud-based

Irrespective of who does the analysis, a solution is considered 'on-premises' if the data storage and processing is done on computing resources owned by the DAP, MNO, or DAS. For example, the MNO might not have the computing resources to support running CDR analysis on its servers and might instead choose to allow the DAP to copy anonymized and de-identified CDR records to a computing environment owned by the DAP. In this situation, the data is now local to the DAP's servers. Or the MNO might be constrained by data laws to not share data outside of its organization boundaries, and instead choose to set up an OLAP[8] which would allow the DAP to query CDR data as needed through a **business intelligence** (BI) tool. In both cases, the computing resources (servers and storage) are fully owned and operated by either the MNO or DAP. On the other hand, cloud-based solutions are more appropriate in cases where the MNO or DAP might not have the capacity or do not want to pay the cost of running and managing a large computing resource. They may also face regulations or policies that allow the use of verified cloud vendors as long as the proper security requirements are fulfilled. As we've mentioned already, there may be resource limitations,

→ **Business intelligence:** strategies and technologies used by organizations to support a variety of data analysis needs.

8 "OLAP is an acronym for Online Analytical Processing. OLAP performs multidimensional analysis of business data and provides the capability for complex calculations, trend analysis, and sophisticated data modeling. It is the foundation for many kinds of business applications for Business Performance Management, Planning, Budgeting, Forecasting, Financial Reporting, Analysis, Simulation Models, Knowledge Discovery, and Data Warehouse Reporting. OLAP enables end-users to perform ad hoc analysis of data in multiple dimensions, thereby providing the insight and understanding they need for better decision making." Retrieved from https://olap.com/olap-definition.

Step #1

regulatory constraints, and internal preferences that result in the preference of one option over another. As the DAP or MNO ponders this decision, the DAS should ask the following questions:

- **Do you have the technology to support the lifecycle of the analysis?** What are the types of technologies you are planning to use? What are the costs associated with these technologies? Do these technologies meet the security requirements of your organization? Is this technology open-source or is it a proprietary technology?

- **Do you have the support needed to engage in the analysis?** Do the technologies you intend to use provide support and technical assistance? Can the solution be maintained over time? Will it require other organizational resources to maintain?

- **Do you have the funding and resources to cover the costs of data storage and hosting?** Is an on-premises solution more cost effective over the lifetime of the CDR analytics project? Is there funding available to cover the maintenance of the chosen technologies? What is the organizational burden of managing and maintaining the technologies needed to support the CDR project?

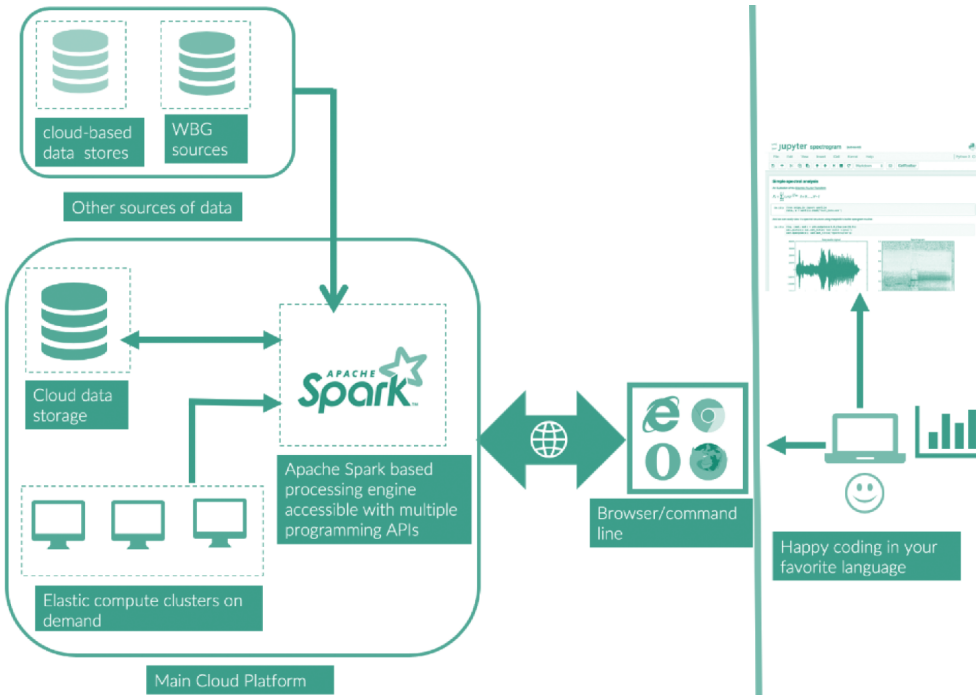
- **Do you have the appropriate skills and training needed to carry out the analysis?** Are the skills needed to run and maintain the technologies for the analysis available in your organizations? Is there a plan for training members of the MNO or DAP?

BOX 2.

Processing CDRs in Microsoft Azure at the World Bank

As part of the World Bank's Integrated and Resilient Urban Mobility Project (IRUMP) in Freetown, Sierra Leone, researchers from the World Bank and UC Berkeley set out to use CDRs to understand urban mobility in Sierra Leone. Their study aimed to understand the impacts of the rainy season on mobility, most popular locations and areas in Freetown, and the relationship between long trips and poverty. Researchers chose to use a cloud-based provider, Microsoft Azure, to provide secure storage and computing resources. The following quote below is extracted from their report and discusses their computational setup for CDR analysis.

“For the four months, we generated over 250 gigabytes (GB) of data. In order to store and process this massive data set, we use the Hadoop ecosystem of tools. The core components of Hadoop include the Hadoop Distributed File System (HDFS) for distributed data storage on a cluster of computers. We use HDFS for data storage. Next, we need a processing engine to interact with the data. For this, we use an open-source software tool called Apache Spark. Some of the advantages of using Apache Spark include the following: it is open- source, it has many programming language APIs, including Python and R, it is very popular and therefore has a big community for technical support. There are many options for deploying Apache Spark with different cloud computing providers. In most cloud platforms, the data are stored within the cloud infrastructure’s distributed system, while computing is also provided by the same cloud provider. In figure 5 we show a schematic of how the different components in this setup interact. In this setup, the Microsoft Azure environment is deployed within the World Bank secure system, which ensures that the data are secure.”



Data storage and processing with Apache Spark within a cloud-based, computing enviroment.

Text and image extracted from: Matekenya, D., Espinet Alegre, X., Arroyo Arroyo, F., & Gonzalez, M. (2021). Using Mobile Data to Understand Urban Mobility Patterns in Freetown, Sierra Leone. Retrieved from https://openknowledge.worldbank.org/bitstream/handle/10986/35033/Using-Mobile-Data-to-Understand-Urban-Mobility-Patterns-in-Freetown-Sierra-Leone.pdf.

STEP #1 RECAP:

Call details records are sensitive data that require proper and secure access and storage. The security of the data is the responsibility of all project partners. Standards stipulating how the data can be used and disclosed can be written into a Data Use Agreement. The MNO and the DAP can choose between an on-premises and a cloud-based solution for data access and storage. This decision must align with the ethical obligations and legal requirements of all project partners.

The Mobile Network Operator should...

- provide anonymized and de-identified CDRs to the DAP
- understand institutional constraints around the use and disclosure of sensitive data

The Data Analytics Partner should...

- create a secure environment for access and storing CDRs
- understand institutional constraints around the use and disclosure of sensitive data

The Data Analytics Stakeholder should...

- understand institutional constraints around the use and disclosure of sensitive data
- ensure that partners have the resources, both financial and in-kind, to carry out CDR analysis

Step #2

CHECKING DATA QUALITY

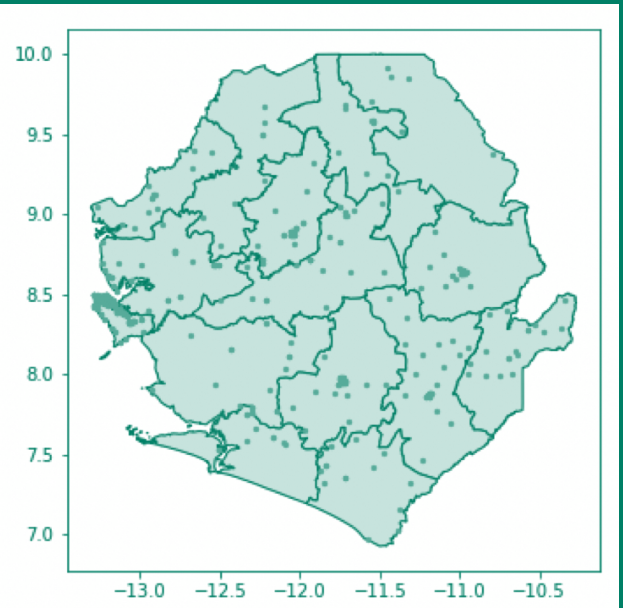
This section discusses the process of ensuring data quality after access has been granted and storage sorted. We discuss some data sparsity issues with CDRs, dimensions of data quality, and data quality checks, and highlight important considerations for validating CDR data.

After acquiring or gaining access to the data, the Data Analytics Partner (DAP) must carry out checks to ensure that the data is of high quality. CDR analysis oftentimes requires combining not only CDR records, but also tower metadata, geospatial files, and other public datasets like census data and land use information. Each of these sets of data introduce their own classes of errors. For example, CDRs are subject to both spatial and temporal data sparsity problems (See Box 3). If these issues are not addressed, the validity of conclusions reached from any analysis may be questioned.

In order to increase confidence in the results of analyses, it is important for researchers to design checks on the data to explore quality issues along four dimensions: validity, accuracy, completeness, and consistency. Although the DAS is not core to this step, it is important that the DAS understands the various dimensions of data quality and how they can support in ensuring the data used for CDR analysis is of the highest quality.

BOX 3.

Spatial and temporal sparsity of CDRs in Sierra Leone



Spatial Distribution of Africell Towers in Sierra Leone

From the image to the left, it is clear that the towers (green dots) are not equally distributed in Sierra Leone. CDRs have spatial and temporal data sparsity issues. Unlike fine-grained GPS data, CDR data is spatially sparse because it maps a subscriber's movement to the coordinates of the closest cell tower. As a result, there are significant limits to what we can infer about a subscriber's displacement pattern. The temporal sparsity of CDR records is due to the fact that CDR records are only created when a subscriber makes or receives a call or text message. The fewer calls or SMS messages a subscriber receives, the less accurate their CDRs are in tracing their mobility patterns. These data sparsity issues produce noisy signals that introduce errors when using CDRs as mobility traces.

Dimensions of Data Quality

- **Validity** refers to the degree to which the data conforms to the expected structure as determined by schemas or metadata.
- **Accuracy** refers to whether or not the data correctly refers to a ground-truth observation[9] and whether it can be considered to come from a verifiable source.
- **Completeness** refers to whether the data contains all the expected fields.
- **Consistency** refers to whether the data records represent the same information across different collections and over a given period of time.

9 For example, this might entail that a CDR record refers to an existing tower and that the GPS coordinates in the metadata are correct.

Examples of quality checks from Sierra Leone

Validate that the cell tower data are the most current. CDR records uniquely identify the tower where that call or SMS was recorded. However, metadata containing information like the GPS coordinates of the tower are not included in the CDR record. This information and other tower metadata (name, location, coordinate, service type (2G/3G)) are usually stored separately. Sometimes they are publicly available, but in most cases the most authoritative tower metadata is held by the MNO. Relying on publicly available information could result in the use of outdated or incomplete tower metadata. If the tower metadata is incomplete, then there is a high probability that the DAP might incorrectly associate CDR records to the wrong tower and GPS coordinate. Therefore, when CDR data is provided to the DAP, the DAS should ensure that the MNO also provides the most authoritative tower and antenna metadata. This type of check will ensure validity, accuracy, completeness, and consistency of tower and antenna data.

Confirm that shapefiles and other ancillary data are valid and accurate. When running CDR data analysis, the DAS is usually interested in a specific geographic scale (country-level, regions, districts, states, etc). In order to analyze CDR records at the correct geographic scale, it is important that the DAS provide the DAP access to the most current shapefiles that can support analysis at the geographic scale of interest, and the DAS should validate the selected shapefile. Given the presence of many open-data initiatives, without the support of the DAS, the DAP might use erroneous geospatial data found online. In addition, CDR analysis requires using other datasets like census information and public health information. The DAS is also critical in availing this and other relevant data to the DAP so that the DAPcan be sure the datasets are correct and valid.

Shapefile:
is a digital vector storage format for storing geospatial data for geographic information systems.

In the case of our work with GoSL, we worked with DSTI to gain access to the most authoritative geospatial data for Sierra Leone, in addition to the census data and daily reporting on Covid-19 infections, deaths, and recoveries. This partnership was key, as there were no publicly available geospatial data that reflected the Government's most recent redistricting efforts. In addition, we were able to access spatial data at different geographic scales, which allowed us the freedom to choose the spatial resolution relevant for our analysis. The availability of census data also allowed our team to map socioeconomic information to the relevant administrative regions. This type of check ensures the validity, accuracy, completeness, and consistency of geospatial and other ancillary data.

Step #2

Review summary statistics to understand statistical properties of data. The statistical properties of CDR records are a function of a variety of parameters, including population, poverty levels, and tower placement. It is important that the DAP runs high-level summary statistics to understand the statistical properties of their CDR records and detect the presence of errors, missing data, or any other irregularities. A thorough understanding of the CDR records and other ancillary data will inform approaches to identifying and filtering outliers and will inform decisions on the relevance and applicability of different types of analyses and visualizations. By engaging in this process, the DAS can build an understanding of the limitations of the data and set clear expectations of what types of insights can be provided by CDR analysis. This type of check will ensure validity, completeness, and consistency of CDR data and can be extremely useful in identifying missing data and data corruption issues.

Contextualize data and results with in-country experts. Oftentimes, usually due to lack of domain knowledge of the telecommunications industry, the DAP might overlook important context that would change their understanding of the phenomena represented in their statistical analyses. For example, when working with CDRs, our team noticed a downward trend in network events (total calls and SMS messages) in Sierra Leone far before any Covid-19 mobility restrictions were put in place. We were not sure what was driving this behavior and were concerned that this unknown effect would confound the effect of the mitigation measures. Through conversation with our partners in DSTI and Africell, we soon discovered that on March 7 the national telecommunications regulatory body, NATCOM, implemented a price floor of Le590. From conversations with mobile network operators in Sierra Leone, they believed that the price floor reduced voice and SMS significantly—50% in the case of one mobile network operator. This exogenous event complicated the task of building a relevant baseline for our analysis. As a result, our team decided that we needed to use a measure that was not sensitive to this price change.

The DAS stakeholder should engage often with the DAP, particularly during the inception and early stages of any analysis, to ensure that the DAP has an understanding of the context that will support them in interpreting the data correctly. This type of check will ensure validity and accuracy of CDR data analyses.

STEP #2 RECAP:

CDRs are subject to data sparsity problems that can affect the validity of results based on them. All project partners should ensure that the appropriate checks are put into place to inspect the quality of the data. There are four dimensions of data quality that are relevant when checking CDRs: validity, accuracy, completeness, and consistency. The goal of the checks is to ensure confidence in the data and the results. Oftentimes very little emphasis is placed on this step in research planning.

The Mobile Network Operator should...

- provide the most current and accurate tower data
- participate in conversations that aid in contextualizing CDR data and any results from analysis

The Data Analytics Partner should...

- engage MNO and the DAS to ensure they have access to the most current and relevant data
- generate summary statistics before beginning any other analytics
- work with in-country experts to contextualize events, such as religious holidays, regulatory changes, or strikes that could influence the data

The Data Analytics Stakeholder should...

- engage in early conversations with the Data Analytics Partner and the MNO to contextualize the CDR records and other ancillary data
- provide the DAP access to relevant data like census records, health surveys, and other statistical surveys
- participate in conversations that aid in contextualizing CDR data and any results from analysis

Step #3

FACILITATING ANALYSIS AND COMMUNICATING RESULTS

This section explains how the type of analysis can change depending on the needs of the DAS, gives examples of the different types of questions CDR analysis can answer, and discusses the importance of effectively communicating results.

Once the data has been accessed, stored, and checked for quality, we can now focus on carrying out analyses. Nowadays, there are many tools and frameworks that the Data Analytics Partner (DAP) can use to carry out big data analysis. The choice of tools and frameworks is dependent on the capacity and capability of the DAP. Irrespective of the analytical tool and framework, from our point of view, the DAP can meet two types of information needs of the Data Analytics Stakeholder (DAS): situational awareness and long-term planning.

Information for situational awareness prioritizes supporting policymakers and relevant stakeholders with measures and metrics that provide an overview of an evolving situation. It is descriptive in nature and, while not used to tie cause to effect, can be particularly useful in rapid-response situations.

Information for long-term planning goes deeper and attempts to understand cause and effect. It is usually in the form of regression analysis and other rigorous studies and experiments. This type of information can be used to pose “what-if” questions and support policymakers in answering questions pertaining to the effective distribution of resources.

Mobility information generated from CDR data can be useful in supporting both situational awareness and planning. Due to the time-intensive nature of rapid-response situations, the DAS will find

analyses that generate information for situational awareness more useful. Below we highlight some examples of analyses from our work and that of other researchers that provide information for situational awareness. These are as follow:

1. **Understanding mobility and compliance**
2. **Understanding interaction between socioeconomic groups**
3. **Understanding and classifying public health risks**

Understanding mobility and compliance

Some of the most basic analysis can help identify and visualize important patterns in mobility. We can combine CDR data to create what are called **Origin-destination (OD) matrices**, which count the number of trips from a given origin to a given destination[10]. Based on how we choose to run our analysis, our origins and destinations could be towers or administrative levels. For example, in Sierra Leone, we used districts instead of towers for our OD matrix. The OD matrix records the numbers of trips from one district to the next in a given period of time.

Origin-destination matrix:
a matrix in which each cell represents the number of trips from origin (row) to destination (column).

FIGURE 3: **Example of an Origin-destination matrix:**

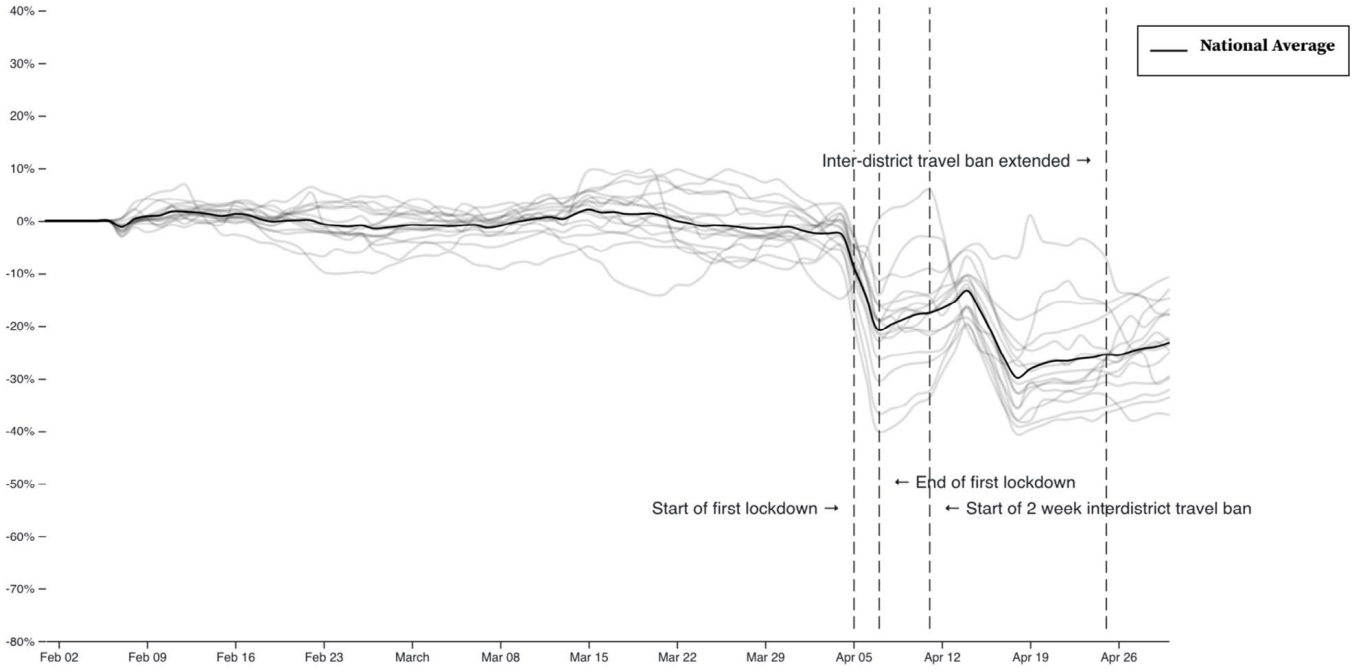
	A	B	C	D	E
A	1	20	89	59	3
B	4	0	23	96	33
C	3	98	72	134	10
D	44	7	0	121	47
E	67	32	90	77	81

10 Mamei, M., Bicocchi, N., Lippi, M., Mariani, S., & Zambonelli, F. (2019). Evaluating origin–destination matrices obtained from CDR data. Sensors (Switzerland).

Step #3

The OD matrix tells us, for example, the number of trips from one tower to the next, and our geospatial data tells us the distance between those towers. We used these two pieces of information to see if the average distance people traveled for inter-district trips decreased after the inter-district travel ban was put into place. As you can see in Figure 4, there is a significant drop in average distance traveled after both the three-day lockdown and inter-district travel ban were enacted. This type of work provides descriptive evidence to policymakers of the impact of their policies. More so, it also provides a base for analyzing the differential impact of mobility restrictions on mobility as a function of geography, work patterns, incomes, and other factors.

FIGURE 4: **Seven-day moving average of the change in average distance traveled**



Understanding interactions between socioeconomic groups

A unique aspect of CDRs is that the patterns of calls or SMS can give us a picture of the social networks that underlie these network events. Researchers at MIT's Senseable City Lab used this insight to study the distribution of socioeconomic groups in Singapore and its relationship to inequality.[11] This work highlights how large scale mobile phone data can reveal the dynamics of spatial and social segregation in cities. Recently, during the Covid-19 pandemic, others have applied this same intuition to understand the differential impact of Covid-19 on different socioeconomic groups — on migration and job opportunities. When studying reduction in mobility due to lockdowns and other confinement measures, researchers found that in Africa and Latin America, poverty influences the intensity of the reduction. In other words, for a variety of reasons, it is harder for poor individuals to reduce their mobility.[12] This type of information and analyses can provide important insight that aids governments in understanding the distributional impact of their policy decisions — whether in relation to mitigation of virus spread or distribution of resources.

Understanding and classifying public health risks

Researchers have also used CDRs and other types of mobility data to study disease spread. For example, researchers have used CDR data to understand how human mobility and connectivity patterns affect the spread of HIV. Researchers have used mobile phone data to quantify malaria importation rates, identify high-risk travelers, and assess transmission.[13]

---

11 Xu, Y., Belyi, A., Santi, P., & Ratti, C. (2019). Quantifying segregation in an integrated urban physical-social space. Journal of the Royal Society Interface. https://doi.org/10.1098/rsif.2019.0536.

12 Bargain, O. and Aminjonov, U. (2020) 'Poverty and COVID-19 in Developing Countries', Groupe de Recherche en Economie Théorique et Appliquée (GREThA). Retrieved from https://ideas.repec.org/p/grt/bdxewp/2020-08.html.

13 Le Menach, A., Tatem, A. J., Cohen, J. M., Hay, S. I., Randell, H., Patil, A. P., & Smith, D. L. (2011). Travel risk, malaria importation and malaria transmission in Zanzibar. Scientific Reports.

Step #3

CDRs have also been used to study the role of mass gatherings in cholera outbreaks in Senegal[14] and the impact of human mobility on dengue outbreaks in Pakistan.[15] In addition, researchers building on work from network sciences have created an alternative network measure for analyzing disease diffusion in a commuting network.[16]

Communicating results

Beyond generating and producing these analyses, it is also important that the DAP provides useful, relevant, and interpretable visualizations of the main results and insights. The Data Analytics Partner must be sure that all analysis is easily communicated to the Data Analytics Stakeholder(s). There must also be intentional and continuous effort put towards ensuring that the capacity and capability to replicate the analysis exists with the Data Analytics Stakeholder(s). There are a number of available open-source and commercially available visualization tools and resources to aid in this task. In addition to communicating the results, it is important that the DAP share the limitations of their analysis and discuss the types of policy decisions that can (and cannot) be supported by their analyses.

14 Finger, F., Genolet, T., Mari, L., De Magny, G. C., Manga, N. M., Rinaldo, A., & Bertuzzo, E. (2016). Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks. Proceedings of the National Academy of Sciences of the United States of America.

15 Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., ... Singer, B. H. (2015). Impact of human mobility on the emergence of dengue epidemics in Pakistan. Proceedings of the National Academy of Sciences of the United States of America.

16 Huang, C. Y., Chin, W. C. B., Wen, T. H., Fu, Y. H., & Tsai, Y. S. (2019). EpiRank: Modeling Bidirectional Disease Spread in Asymmetric Commuting Networks. Scientific Reports.

STEP #3 RECAP:

The DAP can meet two types of information needs of the DAS: ***situational awareness*** and ***long-term planning***. Information for situational awareness prioritizes supporting policymakers and relevant stakeholders with different measures and metrics that can provide an overview of an evolving situation. Information for long-term planning goes deeper and attempts to understand cause and effect. Due to the time-intensive nature of rapid response situations, the DAS will find analyses that generate information for situation awareness more useful.

The Data Analytics Partner should...

- focus on generating analysis that can support the DAS in improving their situational awareness
- ensure that all analysis is easily communicated and meaningful to the DAS

The Data Analytics Stakeholder should...

- require meaningful data visualization and communication from the DAP
- ensure that there is capacity and capability within the DAS to replicate the results shared by the DAP

CONCLUSION:
DATA ETHICS, PRIVACY, AND LIMITATIONS OF CDRS

Data Ethics and Privacy

This section discusses data ethics and privacy challenges posed by using CDRs. We propose three questions to initiate conversation on data privacy and align incentives across stakeholders.

As Linnet Taylor, a data ethics and governance researcher who leads the Global Data Justice project at the Tilburg Institute for Law, Technology, and Society puts it, "human mobility is becoming legible in new, more detailed ways, and [...] this carries with it the dual risk of rendering certain groups invisible and of misinterpreting what is visible. Thus, this emerging ability to track movement in real time offers both the possibility of improved responses to conflict and forced migration, but also unprecedented power to surveil and control unwanted population movement".[17] These concerns, particularly in the context of developing countries, raise the following issues: Is anonymization[18] and de-identification enough to protect the privacy of subscribers? Who gives consent to the use of CDR data? What are potential harms caused in the release of data?

17 Taylor, L. (2016). No place to hide? The ethics and analytics of tracking mobility using mobile phone data. Environment and Planning D: Society and Space.

18 Anonymization: a process by which data is altered in such a way that a person or group can no longer be identified (is made anonymous).

When anonymization and de-identification isn't enough

In this guide and many like it, there is a focus on using anonymization and de-identification to protect the privacy of individuals in a dataset. As Box 5 discusses, this is not enough. There are privacy leakages that can occur when two or more unrelated datasets are published. As data privacy standards change over time, current approaches may prove to be inadequate in the future.

This realization is further complicated by the fact that there are many dimensions to privacy, and these dimensions are highly context-specific. For the MNO, DAP, and DAS, there are oftentimes organizational, professional, institutional, and regulatory standards guiding the use of datasets with and without sensitive information. Even though privacy leakages from multiple data sets are still an issue, there is considerable effort being made today to develop technical solutions that mitigate the privacy risks of using mobile phone data.

BOX 4.
The mosaic effect

The core idea of the mosaic effect is that datasets that individually do not leak private information can create significant security concerns when combined with other datasets containing similar or complementary information. The combination reveals new information previously thought to be undisclosed. The Data Responsibility Team at the Centre of Humanitarian Data write in their blog that "while such information could be used to gain insight, it could also be used by bad actors to do harm. In a humanitarian context, this could happen through the combination of key variables, such as age and gender, from different surveys to reveal the identity and location of people from an ehtnic minority, for instance."

The impacts of the mosaic effect can be described via four types of re-identification attacks: a database cross match, a specific individual match, an arbitrary individual match, and a group match.

Source: Team, D. R. (2020). Exploring The Mosaic Effect On HDX Datasets. Retrieved from https://centre.humdata.org/exploring-the-mosaic-effect-on-hdx-datasets/.

CONCLUSION

What is consent in CDR data analysis?

It is important to think of consent as both an ethical obligation and legal requirement. As an ethical obligation, it requires researchers to disclose the objectives and risks of their research and receive the uncoerced permission from the subjects of research. It ensures that research carried out by a researcher is done in a manner that upholds and maintains the rights of the subjects of their research. As a legal requirement, it is informed by legal standards and regulatory norms in the jurisdiction where the research is approved and/or carried out. For a deeper discussion on consent, we recommend the Handbook of the Modern Development Specialist.[19]

When it comes to big data, particular CDR data, the concept of consent is murky. CDR data is owned by the MNO whose equipment is used to generate the data, and cell network users have typically consented to the companies' Terms of Service and Privacy Policies. However, the data represents information on citizens of a country, which should require data standards at a national level that can regulate appropriate use and disclosure of that data. However, in many low and middle income countries, data protection laws are either weak or non-existent.[20] The limited regulation or enforcement of data privacy laws in developing countries creates gray areas in the ethical obligations and legal requirements in the use of big datasets like CDRs. For example, does approval of CDR analysis by a DAS that is a government agency qualify as consent? Are there cases where approval from the DAS would not qualify as consent? Does the MNO need consent from the DAS to engage in a partnership that discloses CDRs to the DAP? In some cases, these questions can be answered in the form of a Data Use Agreement. The point this guide hopes to stress is that consent is not a one-off issue. Responsible use of CDR data requires continuous evaluation of potential harms and the reevaluation of data agreements in light of any risk to the privacy and security of individuals.

19 Antin, K., Byrne, R., Geber, T., van Geffen, S., Hoffman, J., Jayaram, M., ... Wilson, C. (2016). The Hand-book of the Modern Development Specialist: Being, a Complete, Illustrated Guide to Responsible Data Usage, Manners, and General Deportment. Retrieved from https://responsibledata.io.

20 Piper, D. (n.d.). Data Protection Laws of the World. Retrieved from https://www.dlapiperdataprotection.com//#handbook/world-map-section.

Below we provide some questions that can guide the evaluation of harms and inform the appropriate mitigations:

1. **Is the data properly de-identified and anonymized?** *It is important that before sharing with the DAP, all personally identifiable or sensitive information is removed from the dataset.*

2. **Is the data securely stored and accessed?** *It is important to ensure that anonymized CDR data is kept in a secure environment accessible only to those with the proper authorization. Once secured, the data should not be copied outside of the chosen storage and analytics environment except with proper access and authorization.*

3. **Are all partners committed to the highest standards of ethical use?** *All analytics partners and stakeholders must ensure that no deliberate attempt is made to deanonymize or re-identify individuals in the dataset. This requires the DAP to inform the data provider, the MNO, of any exposure or leakage of sensitive information. Finally, research outputs generated by any partner or stakeholder should not pose the risk of reidentification if combined with other datasets.*

Freetown, Sierra Leone. Photo by Random Institute via Unsplash

Limitations of CDR Data

As we've demonstrated above, CDRs can be used to generate new information to inform policy decisions. However, they are not without their limitations. Below we briefly discuss some of the pressing concerns and limitations of CDR data.

- **Ground-truthing is oftentimes not available.** In many cases, CDR data is the only source of robust mobility information. This means that there is no other valid method of ensuring that CDR data represents to a significant degree the 'on-the-ground' mobility behavior of individuals. Therefore, there is always a level of uncertainty in the results provided through CDR data until some method is available to assess its validity.

- **Low tower density limits the informativeness of analyses.** The data sparsity issues also engender another significant statistical limitation to CDRs. When we process CDRs, we build trip and journey patterns. The patterns are a sequence of time-ordered towers. For example, {A, B, D, F, D, C} would represent a trip on a given day that started at tower A and ended at tower C. If I live in an area with a high density of towers, the types of journeys and trips I can create will have more variety and a high degree of uniqueness. However, if the density of towers in my area is lower, my trips will look less varied and unique. If the above example of a trip took place in a district with only two towers, it might look like this: {A, A, A, A, A, B}. In this case, a significant number of users will have exactly the same trace. In short, districts with more towers provide more information, while districts with fewer towers provide less information.

- **CDRs and phone use.** CDR data alone cannot target a specific individual, but can provide significant information on general travel patterns. Susan Erickson writes that "one cell phone does not equal an individualized, unique self."[21] In her work, Erickson documents how commonly cell phones are loaned, traded, and passed around family and friends. She highlights that for many Sierra Leoneans, having more than one cell phone with multiple SIM cards is common. In short, it is simple to make assumptions about what 'realities' CDR data represents. However, these assumptions become problematic when they lead us to misinterpret the CDRs we

21 Erikson, S. L. (2018). Cell Phones ≠ Self and Other Problems with Big Data Detection and Containment during Epidemics. Medical Anthropology Quarterly.

receive and subsequently the results of the analysis we carry out on them. It also provides a clear limitation to the use of CDR data: that anthropological methods might be more appropriate to understand individual behavior and spread of diseases.

Final Thoughts

CDR data is a great source of information of population mobility patterns, particularly in situations where no other mobility data is available. However, if they are to be relied upon for policy decisions, it is important that policy makers understand their limitations. Some of these limitations are statistical, some methodological, and some structural. We believe CDR data can be used as an important input in many policy decisions; however, we do not believe it can be used as a predictive tool in scenarios where the aforementioned limitations are still present.

Please let us know if you use the guide.
Feedback welcome (mitgovlab@mit.edu).

Freetown, Sierra Leone. Photo by Random Institute via Unsplash.

GLOSSARY OF TERMS

1. **Anonymization:** a process by which data is altered in such a way that a person or group can no longer be identified (is made anonymous).
2. **Business intelligence:** strategies and technologies used by organizations to support a variety of data analysis needs.
3. **Call Detail Records (CDRs):** records automatically generated by the telecommunications equipment whenever a voice call or SMS is made or received by a subscriber in a telecommunications network.
4. **Cloud:** computing resources, especially data storage and compute, that are owned, operated, and maintained remotely and require no direct active management by the user.
5. **Computing resources:** technological and human inputs needed to carry out a computing task.
6. **Data Analytics Stakeholder (DAS)** is the primary client and recipient of the insights generated by the Data Analytics Partner; they initiate and define the scope of the CDR analysis - from selecting partners to determining research questions.
7. **Data Analytics Partner (DAP)** has the responsibility of preparing the analytics environment and performing the requested analytics on the CDRs.
8. **De-identification:** a process by which identifiers/attributes that are known to increase the risk of identification of a person or group are removed from data.
9. **Data sharing agreement:** is an agreement where partners explicitly acknowledge and address sensitive concerns regarding the sharing and use of their private data and establish expectations and processes that ensure the data are secure. They usually cover topics such as how and with whom data will be shared and protocols for working with the data and communicating results and findings.
10. **Mobile Network Operator (MNO)** is the data owner. They are the creators of the CDR data and have the ultimate legal responsibility over its use and distribution.
11. **On-premises:** software and services that are installed and run on computers on the premises of the organization using software, rather than a remote server operated by another organization or cloud provider. It refers to hardware and software that are within the organization's internal systems.
12. **Shapefile:** is a digital vector storage format for storing geospatial data for geographic information systems.
13. **Origin-destination matrix:** a matrix in which each cell represents the number of trips from origin (row) to destination (column).

CIVIC DESIGN DATA LAB

MIT GOV/LAB

www.mitgovlab.org mitgovlab@mit.edu @mitgovlab