FLOWMINDER.ORG

Mobility data products and analyses from mobile operator data: focus on Call Detail Records in low- and middle-income countries

Privacy-conscious data analytics in the context of COVID-19

Flowminder Foundation
08 April 2020

covid19.flowminder.org

Mobility data products and analyses from mobile operator data: focus on Call Detail Records in low- and middle-income countries

Privacy-conscious data analytics in the context of COVID-19

Last update: 09 April 2020
Version: 1.0

Glossary of terms

Call Detail Records (CDR)	The mobile phone data that are most commonly used for studying diseases in low- and middle-income countries are 'Call Detail Records' (CDRs). CDRs are generated each time a mobile phone subscriber makes or receives a call, sends or receives a SMS, or uses mobile data. Each record includes an anonymous identifier of the subscriber, a timestamp, and the cell tower that the transaction was routed through.
CDR aggregates	CDR aggregates are produced by processing the CDR data of many individual subscribers into an output that characterises the behaviour of the entire group of subscribers.
Mobility indicators	Mobility indicators are descriptive statistics of mobility generated from CDR aggregates.

About

Anonymised and aggregated data from Mobile Network Operators (MNOs) is a key data source to understand mobility patterns of populations, and improve decision making and scenario planning during the COVID-19 pandemic. This data can be analysed in near real-time and provide an overview of mobility patterns at local levels and across an entire country.

Large scale mobility changes are both a cause and an effect of the COVID-19 pandemic. On the one hand, mobility of populations affects the speed and patterns of the epidemic. On the other, government interventions (travel restrictions, curfews, closure of public places, banning of gatherings), as well as social and economic changes caused by the epidemic, alter mobility patterns and shift distributions of populations, for example, through returns of migrant workers or movements of urban populations to the countryside.

This document explains how Call Detail Records (CDR) can be leveraged for improving the understanding of population mobility in low- and middle-income countries (LMICs), in the context of the COVID-19 pandemic.

Flowminder is a non-profit foundation that specialises in the analysis of anonymised mobile phone data, satellite imagery, and household survey data for humanitarian and international development purposes. Our researchers and analysts have been working in this field for over ten years, and were the first to respond to a large-scale infectious disease outbreak using mobile operator data (Haiti cholera outbreak in 2010). Flowminder researchers were also the first to show that mobile operator data can predict the spatial spread of an infectious disease (Nature Sci. Rep. 2015).

We define in this document the key stages in the analysis of CDR data, and set a standard for the CDR aggregates to be produced, and the mobility indicators that can be derived from the aggregates, to support decision makers within the ongoing COVID-19 pandemic.

Flowminder has designed a series of aggregates and indicators to represent all dimensions of mobility and with the following considerations as a priority:

i) they should not be excessively computationally intensive to produce, even in settings with scarce compute resources,

ii) they are fully anonymous and contain no information about individual subscribers, ensuring that the privacy of subscribers is maintained at all times, and

iii) they are robust to sparse tower distribution and to infrequent phone usage (the aggregates have been developed with the assumption that most subscribers will not have a record associated with them every day, or even once every few days), both of which are common in LMICs.

With these considerations in mind, we have produced the set of CDR aggregates and indicators we present in this document (along with code, methods and guidelines), which we believe extract relevant mobility information from CDRs in LMICs, in a privacy-conscious and robust manner. Flowminder will later expand its set of indicators for other types of mobile operator data and for higher income settings.

1. Key questions that can be answered from CDR analysis

Mobile operator data can support government and public decision making during the COVID-19 pandemic. We have identified below five key areas of applications which would benefit from mobility insights extracted from CDR data:

1. Monitoring the primary effects of mobility and social distancing interventions

   We propose indicators that aim to provide a measure of change in mobility following specific government interventions and their announcement. These indicators may be used to assess whether restrictions have had the expected effect of reducing travel, dispersion, and population mixing.

   They are not a measure of the number of people who do or do not comply with mobility restrictions. Restrictions have a number of exemptions which cannot be quantified using CDR data. Exemptions include, for example, key workers (e.g. health sector, law enforcement, military, maintenance of essential service, supply chain of essential products, etc..), people returning home, people supporting their families, people in need of health care, etc… Therefore, as we cannot quantify exemptions, we cannot quantify compliance.

2. Monitoring the side effects of interventions

   Interventions designed to reduce mobility and increase social distancing may have the unintended opposite effect, which negatively impact the controlling of COVID-19 or have broader social impacts. For example, research conducted in New York[1] showed that the introduction of school closures in New York resulted in increased activity at grocery, shopping, food and outdoor places. In addition, evidence from Italy, France and Ghana[2] demonstrated that there were large scale movements of people before lockdowns were implemented.

We propose indicators that aim to provide a measure of unintended side effects following government interventions. These indicators will help planners assess whether interventions should be modified or ceased. In addition, such indicators would also help planners form a more accurate estimate of the likely impact an intervention would have, accounting for these side effects and supporting longer term planning efforts.

3. Identifying routine mobility patterns (to plan interventions and assess risk levels)

We propose indicators which extract current patterns of mobility, helping to identify hotspots (places with receiving large crowds and with high population mixing), most travelled routes and secluded regions. This information would help decision-makers to plan interventions and restrictions, and to target areas to send information messages to. In addition, the indicators will enable planners to conduct scenario testing for different types of measures (restrictions on travel, closures of public places) and for the staged relaxation of measures.

4. Monitoring changes in density of population (dynamic population mapping)

There can be large scale movements of populations as a result of the COVID-19 epidemic, with, for example, countries experiencing high levels of movements from urban to rural areas. Such changes can have unintended consequences on public services, food supplies and other critical infrastructure.

We propose indicators that monitor the changes in population densities from CDRs which, combined with existing population estimates, will help provide more accurate estimates of population during the epidemic. This will provide a useful indicator both during the outbreak and after restrictions have been lifted, as planners will be able to understand how long it takes for populations to return to the pre-crisis state.

5. Mobility data as an input to predictive models and analyses with ancillary data

The indicators we propose can be used in further analyses, predictive modelling and research. The indicators reflect all dimensions of mobility and can be used to support decision-making and investigations across a wide range of domains. This may include epidemiological modelling, resource planning, provisioning of services, and longer term research into preparedness for epidemics or the effect of mobility restrictions on the environment.

Table 1 provides an overview of these five key issues affecting decision-making that can be informed by CDR analysis. For each main issue, the table lists topics to be assessed, insights required for its assessment, and which of our proposed indicators may be used to produce these insights. These

[1] http://curveflattening.media.mit.edu/Social_Distancing_New_York_City.pdf
[2] https://web.flowminder.org/publications/mobility-analysis-to-support-the-government-of-ghana-in-responding-to-the-covid-19-outbreak

insights can take the forms of graphs, maps, statistics or reports, which can be used by governments, journalists and other scientists at a national or global scale.

Within the table, there may be multiple indicators which are listed for a single insight or topic. In these cases, indicators may need to be combined with each other for double checking, or are needed to answer the multiple questions in the 'insight column' of the same row. More details will be provided on this analysis and interpretation in further guidance.

Table 1: Key areas that can be informed from for mobility data, and insights provided by the mobility indicators we propose.

Topic under assessment	Examples of insights provided	Indicator(s)
1. Monitoring the primary effects of mobility and social distancing interventions		
Limitation on travel distance	How much has long-distance travel decreased? Do regions receive fewer incoming visitors?	▪ Inter-regional travel: distribution of distances travelled ▪ Inter-regional travel: dispersion ▪ Population mixing
Promotion or regulation about work from home	How much have urban 'commuting'-type trips been reduced? Has population mixing been reduced in urban centres?	▪ Hourly presence and variance, per area ▪ Crowdedness ▪ Intra-regional travel distribution ▪ Number of regions visited, per home location ▪ Population mixing
Lockdown	How much have urban trips and population mixing decreased? Do more people appear to remain at home more than before?	▪ Intra-regional travel distribution ▪ Number of regions visited, per home location ▪ Number of subscribers visiting only their home region ▪ Population mixing: visitors vs residents ▪ Crowdedness
Curfew	How much has travel reduced during night time?	▪ Daytime / nighttime presence ratio ▪ Intra- and inter-regional travel ▪ Crowdedness
Closure of public places	By how much does the closure of specific establishments (religion, entertainment, shops, education) reduce travel and population mixing?	▪ Hourly presence and variance, per area ▪ Crowdedness ▪ Intra-regional travel ▪ Number of regions visited, per home location ▪ Population mixing
Ban on public gatherings and events	How much have crowd sizes and crowd frequencies decreased?	▪ Crowdedness: crowd size and frequency

2. Monitoring the side effects of interventions		
Home relocations	How many people have relocated (changed their -presumed- region of residence) as a result of interventions, announcements or news? How many people are leaving urban centres for rural areas?	▪ Number of subscribers who changed their home location ▪ Number of subscribers who relocated between any 2 regions (origin-destination flow)
Crowds formation	Are some places visited by a larger number of people following interventions? (e.g. shops after school closure)	▪ Hourly presence, per area ▪ Crowdedness ▪ Intra-regional travel ▪ Population mixing
Duration of disruption	How long before a 'new normal' settles in and daily repeated patterns re-appear?	▪ Hourly and daily variance of subscriber presence, per area
Travel range	Do people tend to travel longer distances?	▪ Inter-regional travel: distribution of distances travelled
Geographic dispersion	Do the interventions, announcements, news, result in more dispersion?	▪ Inter-regional travel: dispersion ▪ Number of regions visited, per home location ▪ Population mixing
Effect of border closure	Are many citizens coming back to the country?	[TBC: This may not be measurable with CDRs. It depends on several factors (information on roaming, duration of SIM validity, frequency of phone usage, ..), and we are currently investigating possible indicators of new SIMs and new arrivals to points of entry]
3. Identifying routine mobility patterns (to plan interventions and assess risk levels)		
Selecting and prioritising locations for implementing mobility restrictions	Which locations receive large crowds and/or have a large amount of population mixing (hotspots)? What are the main travelled routes and routes linking hotspots? Are there areas in the country already partially shielded from the rest (secluded)?	Hotspot locations ▪ Crowdedness - size and frequency ▪ Population mixing - mixing factor Regional connectivity structure: ▪ isolated clusters of regions ▪ main travelled routes ▪ main travelled routes through hotspots
Targeting locations for communications (where to send information, e.g. via SMS, voice or billboards)	Which locations receive large crowds and/or have a large amount of population mixing (hotspots)? Which locations do people who travel a lot come from?	Hotspot locations ▪ Crowdedness - size and frequency ▪ Population mixing - mixing factor ▪ Inter-regional travel - number of regions visited, per home location

Identifying places where people may be at higher risk of spreading or contracting the virus (based on mobility profile only - not infectious cases)	Which locations receive large crowds and/or have a large amount of population mixing (hotspots)? What are the main travelled routes and routes linking hotsports)?	Hotspot locations ▪ Crowdedness - size and frequency ▪ Population mixing - mixing factor Regional connectivity structure: ▪ main travelled routes through hotspots
Estimating the number of people in each region who may be at higher risk of spreading or contracting the virus?	How many people in each area visit a hotspot / multiple hotspots? How many people in each area visit high-mixing areas?	▪ Number of residents of hotspots ▪ Average number of hotspots visited per subscriber, obtained from: ▪ Crowdedness ▪ Population mixing ▪ Inter-regional travel - number of regions visited, per home location
4. Monitoring changes in density of the population (Dynamic population mapping)		
Changes to the population size of each region during the crisis (either driven by interventions or news or events related to the epidemic)	How many people move out of their home and relocate to a new region?	▪ Changes to resident subscriber population of each region (weekly)
Duration of relocations	Does the population distribution return to its pre-crisis state and how long does it take?	▪ Average time for a region to get back to its pre-crisis subscriber population size
5. Mobility data as an input to predictive models and analyses with ancillary data		
Prediction of the spatial spread of SARS-COV-2 (assuming good case data)	Which areas of the country will be most affected? In which order may this happen? When and where to relax interventions and which ones?	▪ Inter-regional travel ▪ Population mixing ▪ Changes to resident subscriber population of each region
Resource planning and provision of services (assuming data on resources and location of services)	How to optimise resource provision (e.g. health care needs) given population movements?	▪ Population mixing ▪ Changes to resident subscriber population of each region
Longer term research in e.g. preparedness to epidemics, economics, ecology, social sciences	Example of questions: How long did it take for mobility patterns to settle to a new normal after interventions? How disruptive has the crisis been and for how long (resilience)? How are changes in economic variables (production, seasonal agricultural work, poverty) and environment (air pollution, noise pollution, electricity consumption) related to changes in mobility? How changes in air pollution affect the transmission and severity of the disease?	▪ Hourly and daily variance, per area ▪ Changes to resident subscriber population of each region ▪ Intra and inter-regional travel ▪ Crowdedness

Caveats

While CDR data has an important role to play in the response to COVID-19, it is important that considerations are made to the suitability of its use. This includes a range of technical and social reasons as described below.

There are rightfully concerns about the potential for misuse of mobility indicators derived from CDR data. It is important to note that indicators will not be provided in real-time but with a delay of at least a few days, to serve for analysis. This delay would prevent its use to identify ongoing gatherings and to be used as a policing tool for example. In addition, indications of changes in mobility are not indications of level of compliance to restrictions, as to measure this would require knowledge of exemptions and a finer spatial resolution in most cases.

Secondly, as the indicators are designed to be exhaustive and capture all dimensions of mobility to inform a range of possible questions, the exact usefulness of each and every indicator will depend heavily on the country context. This not only includes the needs of the country, but factors influencing the availability and representativity of the data (coverage, density of towers, frequency of phone usage).

Finally, CDR data from a mobile network operator only contains information about the people who use a SIM card from that operator. This subset of people is unlikely to be perfectly representative of the entire population of a country because not everyone uses a mobile phone, and not all mobile phone users are subscribed to a single mobile network operator. For example, young children and elderly people in many countries do not use a mobile phone, and in low-income countries, individuals in the lowest socioeconomic strata may not own a mobile phone. Additionally, the geographic, demographic, and socioeconomic distribution of subscribers to each mobile network operator is often not representative of the full population.

2. From raw data to insights: the CDR analysis process

Several key stages are required to convert raw CDR data into actionable insights. These are categorised as follows:

1. Raw CDR data: CDR data are owned by MNOs and are generated each time a mobile phone subscriber makes or receives a call/text or uses mobile data. Each record includes an anonymous identifier of the subscriber, a timestamp, and the cell tower that the transaction was routed through. CDRs can be useful mobility resources as they contain a record of each subscriber's approximate location (the location of the cell tower) each time they use their phone.

2. Production of CDR aggregates: CDR aggregates are obtained from the processing of the raw CDR data corresponding to groups of individual subscribers. These aggregates can be produced by MNOs and regulators using Flowminder's code available on GitHub (github.com/Flowminder/COVID-19). Examples of aggregates are the counts of subscribers actively using their phone in a given region and time interval, or counts of subscribers travelling from one region to another.

3. Production of mobility indicators from CDR aggregates: Data analysts can extract information from CDR aggregates to produce descriptive statistics of mobility. Flowminder

analysts propose here a range of indicators that capture mobility characteristics that are relevant to the COVID-19 context.

4. Turning mobility indicators into actionable insights for decision making: The interpretation of mobility indicators, by incorporating local knowledge and additional data sources (such as the locations of schools or shopping districts), can provide decision makers actors with a wide range of insights, examples of which are provided in Table 1..

A summary of the analysis chain is given in Figure 1.

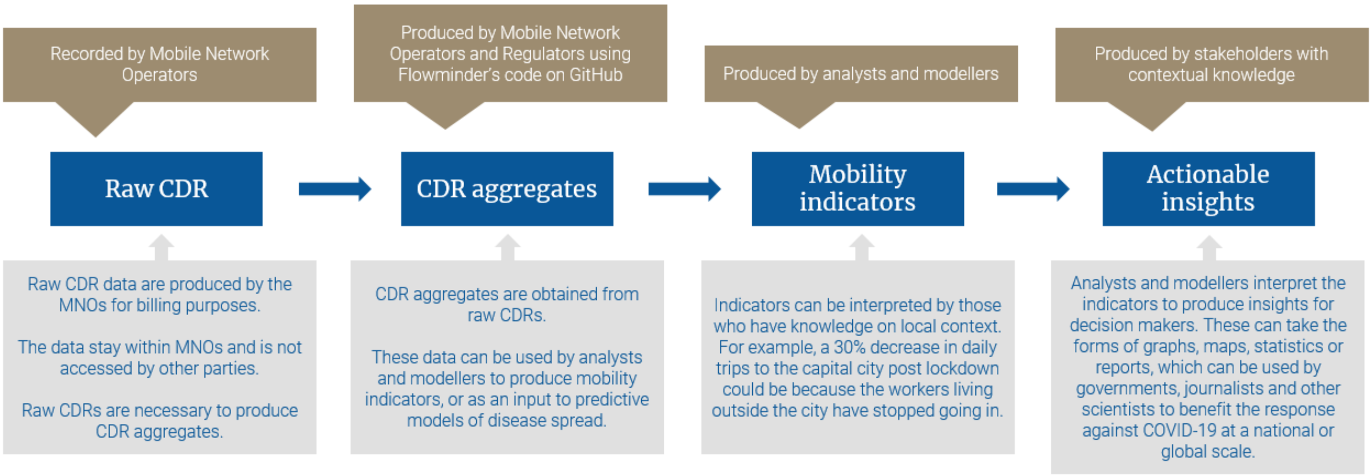

Figure 1. Analysis chain.

3. CDR aggregates: 'Building blocks'

This section outlines the CDR aggregates which can be used as the basis of a range of analyses for supporting low-and middle-income countries.

Table 2 provides a summary of the categories and CDR aggregates that can be produced from CDRs to build mobility indicators. Code used to derive these aggregates can be accessed from GitHub (https://github.com/Flowminder/COVID-19), and descriptions are also listed on our website (https://covid19.flowminder.org/resources-for-mnos/list-of-cdr-aggregates) as well as the description of methods to produce them:
https://covid19.flowminder.org/resources-for-analysts/methods-cdr-aggregates

3.1. The basics of CDR aggregates

MNOs have datasets that comprise records that are generated each time a subscriber makes or receives a call, sends or receives a SMS, or uses mobile data on their phone. The combination of events (calls, SMS, data sessions) included in the dataset depends on each MNO. We are only able to tell if a subscriber was at a particular location at a certain time if they used their phone at that location and time, for an event that is included in the dataset. In this case, we say that we 'recorded' the subscriber as being at that location at that time.

Aggregates have been selected to represent all dimensions of mobility and the following three criteria: 1) they are fast and easy to compute for MNOs with limited resources; 2) they are fully anonymous and contain no information about individual subscribers; and 3) the aggregates are robust to infrequent phone usage. The aggregates have been developed with the assumption that most subscribers will not have a record associated with them every day, or even once every few days. This is especially common in low- and middle-income countries.

CDR aggregates do not expose any information about individual subscribers and cannot be used to re-identify an individual. In line with international standards, aggregates are only produced for groups of at least 15 subscribers. This means that, for example, when the number of active subscribers in a location is less than 15 subscribers, that count is not included in the aggregate. This means that the data can be shared with third parties. These parties include epidemiologists who may study the aggregates, in combination with case data, to assess whether mobility changes are having an effect on the spread of the disease, and to predict the evolution of the spread.

3.2. Spatial and temporal resolutions of the CDR aggregates

How CDR aggregates are broken down in time (temporal resolution) and space (spatial resolution) is key to which types of questions they can inform. Temporal resolution ranges from weeks to 15 minute intervals, and as described in the following section, the selection of this time interval is dependent on the indicator required. Spatial resolution is often based around administrative boundary datasets, from sub-county level (level 4) up to state level (level 1), although clusters of towers can be used for fine scale analyses (especially relevant in urban areas)[3]. Computing CDR aggregates at different temporal and spatial resolution is essential to obtain the full range of mobility indicators.

Whilst it might seem preferable to produce CDR aggregates at the finest spatial and temporal resolution, this may not always be beneficial. As the resolution is increased, a larger proportion of counts will fall beneath the 15 subscriber threshold required for data privacy (and statistical significance), resulting in more data being removed from the outputs. We recommend selecting the coarsest spatial and temporal resolution that will meet the requirements of the use case. This will ensure that statistically significant and privacy-preserving outputs are available for the maximum number of spatial regions and time periods.

Additionally, aggregates need to be produced at different resolutions (e.g. both for days and for weeks, and both for admin 4 and 3 level) which are combined to produce indicators of population mixing and intra-regional travel (cf next section).

Note: the maximum spatial resolution that can reasonably be achieved is typically dependent on the density of cell towers within the studied region. In regions where there are many cell towers (e.g. highly urbanised regions), it is possible to divide up the region into multiple sub-regions by clustering towers, each cluster containing several cell towers. Each sub-region is therefore likely to have a statistically significant number of data records associated with it. However, in regions with a low number of cell towers (e.g. rural regions), it is usually not possible to divide up the region and obtain

[3] We also provides guidelines to MNOs to aggregate their towers into clusters or by administrative regions: https://covid19.flowminder.org/resources-for-analysts/methods-cdr-aggregates/geospatial-details

sufficient data in each sub-region.

3.3. Defining home locations

Several of the aggregates are based on the concept of defining a home location for each subscriber. We define this to be a 'reference location' based on where the subscriber most frequently used their phone for the last time each day over a four-week period, and updated every week. Once calculated, it can be used to produce mobility indicators and understand mobility patterns with respect to each subscriber's reference location (e.g. whether they are in their home region or visiting another region). We define as 'resident of a cluster of towers or region' the subscribers who have been assigned this cluster or region as home location.

3.4. Applications in non-LMIC contexts

Although not the focus of this work, the listed aggregates are relevant for high-income settings in addition to low- and middle-income settings. However, in high-income settings, there can be additional mobile operator data available (including signalling data) which can enable more advanced aggregates to be developed. For example, in scenarios where high-frequency signalling data are available, we recommend extracting modal daytime locations for each subscriber, and the duration of visits and time away from home.

Table 2: CDR Aggregates

CDR aggregate	Temp res	Spatial res.	Indicators used in	prio
1. Count of active subscribers - the number of unique subscribers recorded in each region during the specified period				
Count_subscribers (hour, local)	hour	Cluster or admin4	Subscriber presence Population mixing Intra-regional travel Hotspots	2
Count_subscribers (day, local)	day	Cluster or admin4		
Count_subscribers (week, local)	week	Cluster or admin 4		
Count_subscribers (day, regional)	day	Admin 3, 2,1		
Count_subscribers (week, regional)	week	Admin 3, 2,1		
Count_subscribers (15min, urban cluster)	15 min	Urban clusters	Crowdedness Hotspots	
2. Count of residents (home locations) - The number of subscribers assigned to each region as their home location during the specified period.				
Count_residents (week, local)	week	Cluster or admin 4	Home location	1
3. Count of travellers (origin-destination matrix) - The number of subscribers that travel between any two locations				

within the time period (there are 2 types: all locations (all pairs of locations visited in a trip: a subscriber travelling from A to B then to C, is counted between A->B, B->C but also A->C), consecutive locations only (A->B, B->C)				
od_matrix_directed_all_pairs (hour, local)	hour	Cluster or admin 4	Inter-regional travel: travel distance, dispersion, mixing factor, Regional connectivity	
od_matrix_directed_all_pairs (hour, local)	day	Cluster or admin 4		
od_matrix_directed_all_pairs (day, regional)	day	Admin 3, 2,1		3
od_matrix_directed_consecutive_pairs (hour, local)	hour	Cluster or admin 4	Inter-regional travel: flows	
od_matrix_directed_consecutive_pairs (hour, local)	day	Cluster or admin 4		
od_matrix_directed_consecutive_pairs (day, regional)	day	Admin 3, 2,1		
4. Count of travellers (connections triangular matrix) - : alternative to OD matrix with no direction of movement. The number of subscribers that travel between any two locations within the time period, irrespective of the direction of travel				
od_matrix_undirected_all_pairs (hour, local)	hour	Cluster or admin 4	Inter-regional travel: travel distance, dispersion	
od_matrix_undirected_all_pairs (hour, local)	day	Cluster or admin 4		
od_matrix_undirected_all_pairs(day, regional)	day	Admin 3, 2,1		
5. Count of visits at home and away (home-away matrix) The number of times a subscriber is recorded in their home region, and the number of times they are seen at a location that is not their home, within the time period, given for each pair of home region and visited region				
Count_visits_home_away (hour, local)	hour	Cluster or admin 4	Inter-regional travel from home (per subscriber aggregate equivalent)	
Count_visits_home_away (day, regional)	day	Admin 3		
6. Count of home relocations (home origin-destination matrix) The number of subscribers that have changed home region, given for each pair of regions as previous home and new home				
Count_home_relocations (week, regional)	week	Admin 3	Home location	
7. Count of subscribers only seen in 1 region - The number of subscribers that are only recorded within a single region				
Count_subscribers_single_region (day, regional)	day	Admin 3	Inter-regional travel	
Count_subscribers_single_region (week, regional)	week	Admin 3		

8. Count of subscribers only seen in home region - The number of subscribers that are only recorded within their home region during the specified time period				
Count_subscribers_home_region (day, regional)	day	Admin 3	Inter-regional travel	
Count_subscribers_home_region (week, regional)	week	Admin 3		
9. Count of events - The total number of data records for the specified time period.				
Count_events (hour, local)	hour (or 15 min)	Cluster or admin 4	Sample size / data quality indicators	4
10. Count of active residents - The number of subscribers that are recorded to have been within their home region, within the specified time period				
Count_active_residents (hour, local)	Hour	Cluster or admin 4	Sample size / data quality indicators	5
Count_active_residents (day, regional)	Day	Admin 3		
TBC: temporary 'stay' location (assigning each subscriber each day a 'stay' location which may be different from the home location, e.g. stays from 2 to 3 days. This could help categorize mobility types.				
			Inter-regional travel	

3.5 [TBC] Further description of CDR aggregates

Home-away matrix:

		home		
		A	B	...
visits	A	Nb subscribers who 'live' in A and visited A in time interval	Nb subscribers who 'live' in B and visited A in time interval	..
	B	Nb subscribers who live in A and visited B in time interval	..	..
	...	..	..	..

Home origin-destination matrix:

	New home		
	A	B	...

Previous home	A	Nb subscribers who did not change home location and 'live' in A	Nb subscribers who used to 'live' in A and now 'live' in B	..
	B	Nb subscribers who used to 'live' in B and now 'live' in A	..	..
	...	..	..	..

4. Indicators built from CDR aggregates

Using a combination of the CDR aggregates described above, we can generate a range of mobility indicators which are useful for planning and evaluating interventions related to mobility in the context of the COVID-19 response. Table 1 provides examples of insights that can be generated from the indicators, and all indicators and the CDR aggregates they are based on are listed in Table 3 below.

4.1. Categories of indicators

Indicators we suggest can be regrouped in the following categories, based on the main CDR aggregates they are derived from:

- Subscriber presence: how visited places are over time and regularity of pattern
- Home locations: changes to the population distribution
- Crowdedness: how busy and how often busy places are
- Population mixing: how many different people visit the same place over time
- Intra-regional travel: quantity of movement within a region
- Inter-regional travel: quantity of movement across regions for all pairs of regions and quantity of movement from home
- Locations of interest: hotspots, main travelled routes and secluded areas
- Sample size / data quality indicators

4.2. Methods to produce the indicators

Analysis, including data cleaning, aggregation of data across time periods, or rescaling to account for biases within the data, is required to produce these indicators. Full details will be provided on our website.

Most indicators are formed using the following standard operations:

- Computing a baseline (for comparison or for computing a % change)
- Computing a % change compared to baseline (often but not always)
- Rescaling the data to account for changes in phone usage (active subscribers and number of events)

To do this all CDR aggregates need to be computed for a period of at least four weeks before interventions were introduced, or before the initial COVID-19 cases were announced in the country. This is to obtain indicators about routine baseline mobility, against which the post-intervention (or post-news) indicators can be compared. CDR aggregates should then be produced continuously throughout the epidemic, and output with a frequency that meets the needs of the data consumers.

To interpret and visualise the indicators the boundary datasets used to produce the spatial aggregation needs to be provided along with the CDR aggregates. We also provides guidelines to MNOs to aggregate their towers into clusters or by administrative regions: https://covid19.flowminder.org/resources-for-analysts/methods-cdr-aggregates/geospatial-details

Table 3: Indicators

Indicator	Unit	CDR aggregates needed
Subscriber presence - Indicators based on time series of *Count_subscribers* aggregates		
Subscriber presence - hourly/daily presence (weekday), per region Number of subscribers recorded in each region each hour/day, during weekdays: this describes how busy each region is during weekdays, and during each hour of weekdays	% change relative to baseline	*Count_subscribers (hour, local)* *Count_subscribers (day, local)* *Count_active_residents* *Count_events*
Subscriber presence - hourly/daily presence (weekend), per region Number of subscribers recorded in each region each hour/day, during weekends: this describes how busy each region is during weekends, and during each hour of weekends	% change relative to baseline	*Count_subscribers (hour, local)* *Count_subscribers (day, local)* *Count_active_residents* *Count_events*
Subscriber presence - changes to variance of hourly/daily presence, per region The variance in hourly/daily patterns, across different days/weeks: this provides information about the extent to which previously predictable, regular patterns have been disrupted. **Disruption duration:** A further indicator may be the time it takes for presence patterns to become regular again.	% change relative to baseline	*Count_subscribers (hour, local)* *Count_subscribers (day, local)* *Count_events*
Subscriber presence - changes to day/night subscriber presence ratio, per region Ratio of the number of subscribers recorded in each region during 'daytime' hours, and the number during 'nighttime' hours: this provides information about the number of residents (who are likely to be present during the nighttime) versus the number of non-residents (who are likely to be present during the daytime) that visit the region. It can also provide information about which regions are popular night-	%change relative to baseline	*Count_subscribers (hour, local)* *Count_subscribers (day, local)* *Count_active_residents* *Count_events*

life spots (a high proportion of nighttime visitors).		
Home locations - Indicators based on time series of Count_residents and *Count_home_relocations* aggregates		
Home locations - changes to the number of resident subscribers per region Number of subscribers that are assigned to a region as their home region	% change relative to baseline	Count_residents
Home locations - changes to the number of subscribers who move home, per pairs of regions Number of subscribers whose home location has changed, between any 2 regions or for specific regions	% change relative to baseline	*Count_home_relocations* *Count_events*
Home locations - duration of disruption, per region Average time for a region to get back to its pre-crisis subscriber population size (if at all)	weeks	*Count_home_relocations* *Count_events* *Count_active_residents*
Crowdedness measure - Indicators based on time series of *Count_subscribers (15min, urban cluster)*		
Crowdedness measure - crowd size per cluster Number of subscribers recorded at each cluster within a 15 minute interval (urban areas only): this measures the size of a 'crowd'	% of residents	*Count_subscribers (15min, urban cluster)*
Crowdedness measure - crowd frequency per cluster Number of 'crowds' (defined to be more than X subscribers recorded at the same cluster within a 15 minute interval) recorded each day: this measures the frequency with which crowds congregate	N per day	*Count_subscribers (15min, urban cluster)*
Population mixing - Indicators based on time series of Count_subscribers (day, local) and Count_subscribers (week, local) and Count_residents (local)		
Population mixing - visitors vs residents, for each region Number of unique subscribers recorded in the region each day/week, normalised by the number of subscribers that have that region assigned to them as their 'home region': this is an indicator of how many non-resident 'outside' visitors come to the region, relative to the size of the resident population	% of residents and % change relative to baseline	*Count_subscribers (day, local)* Count_residents *Count_active_residents*
Population mixing - 'mixing factor', for each region Average number of unique visitors each day, averaged over a week, divided by the number of unique subscribers to the region in a week: this indicates the degree to which the same people visit a region each day, versus different people visiting	% of unique week visitors And % change	*Mean Count_subscribers (day, local) divided by Count_subscribers (week, local)*

each day which would result in higher 'mixing'.	relative to baseline	
Intra-regional travel - Indicator based on time series of Count_subscribers (day, local) and Count_subscribers (day, regional)		
Intra-regional travel - changes to travel within each region Average number of sub-regions visited in each region per visitors to the region: this measures the amount of movement within a region	% change relative to baseline	*Mean Count_subscribers (day, local) divided by Count_subscribers (day, regional)*
Inter-regional travel - Indicators based on time series of od_matrix_directed_all_pairs, od_matrix_directed_consecutive_pairs and Count_visits_home_away		
Inter-regional travel - changes to incoming flow to each region Number of subscribers entering a region from another region	% change relative to baseline	*od_matrix_directed_consecutive_pairs Count_events*
Inter-regional travel - changes to outgoing flow from each region Number of subscribers leaving a region and travelling to another region	% change relative to baseline	*od_matrix_directed_consecutive_pairs Count_events*
Inter-regional travel - changes to net flow of each region The difference between incoming and outgoing flows: this measures how subscribers are redistributing amongst regions	% change relative to baseline	*od_matrix_directed_consecutive_pairs Count_events*
Inter-regional travel - changes to the distribution of distances travelled The median distance (and quartiles) travelled by all subscribers that travel to the region from another region: this indicates whether a lot of people come to the region from nearby regions, or from regions that are farther away. Can be aggregated across regions.	% change relative to baseline And km	*od_matrix_directed_all_pairs*
Inter-regional travel - changes to the dispersion distribution For each destination region, this is the average percentage of subscribers that originated from each region: this indicates whether a lot of people originate from the same region, or whether there are many origin regions with a few people coming from each one. (variant: standard deviation of coordinates of origin regions)	% change relative to baseline And km	*od_matrix_directed_all_pairs*
Inter-regional travel - changes to the mixing factor Average 'mixing factor' of origin regions: indicates whether visitors are coming from a 'high-risk' (or 'high-contribution')	% change relative to baseline	*od_matrix_directed_all_pairs*

region		
Inter-regional travel - changes to the number of regions visited, per home location Average number of regions visited by the resident subscribers of each region (Variant: **Exposure (or contribution): changes to the number of hotspots visited, per home location** Average number of hotspots visited by the resident subscribers of each region)	% change relative to baseline	*Count_visits_home_away*
Inter-regional travel - changes to the line distance travelled, per home location Average distance to visited regions by the resident subscribers of each region	% change relative to baseline And km	*Count_visits_home_away*
Inter-regional travel - changes to the dispersion of visited regions per home location Standard deviation of coordinates of visited regions by the resident subscribers of each region	% change relative to baseline And km	*Count_visits_home_away*
Inter-regional travel - changes to the number of subscribers visiting a single region, per region More specific indicator on the number of subscribers not crossing admin boundaries	% change relative to baseline	*Count_subscribers_single_region*
Inter-regional travel - changes to the number of subscribers visiting only their home region, per home region More specific indicator on the number of subscribers not recorded outside of their home region	% change relative to baseline And km	*Count_subscribers_home_region*
Locations of interest - Locations selected from the analysis of population mixing indicators and the analysis of the graph formed by the inter-regional travel aggregate *od_matrix_directed_all_pairs*		
Regional connectivity structure: isolated clusters of regions Are there areas of the country isolated for the rest? (which may be easier to protect)	Names, coordinates	*od_matrix_directed_all_pairs*
Regional connectivity structure: main travel routes - list of locations most connected to each other forming a path through the country (along which the virus may spread)	Names, coordinates	*od_matrix_directed_all_pairs*
Regional connectivity structure: main travel routes through hotspots - path along which most population mixing is occuring (along which the virus may spread)	Names, coordinates	*od_matrix_directed_all_pairs*
Hotspots: locations of hotspots - Locations with high population mixing values and/or crowdedness	Coordinates	Population mixing and crowdedness indicators

Sample size/data quality indicators - These are not indicators of mobility but serve to control for changes in the CDRs that may lead to a misinterpretation of mobility indicators, and to assess uncertainty from the variations of the sample size (number of events and subscribers used to compute the indicators)		
Sample size/data quality indicators - number of events The total number of calls/SMS/data sessions in each region each hour/day: this is to provide an indicator of the sample (data) size that we are working with in order to know if the analysis for particular regions and times should be discounted due to insufficient data. It also enables to check whether an increase in 'subscriber counts' may actually just be due to the normal number of subscribers using their phone more, and to easily determine if there is an issue with e.g. missing data one day.	N per region per time interval	*Count_events*
Sample size/data quality indicators - 'residents' activity level The proportion of residents in a region who use their phone each hour/day, divided by the total number of residents (where 'residency' is defined via a home region): this indicates how active the subscribers in each region are	% of residents	Count_residents *Count_active_residents*
Sample size/data quality indicators - new subscribers Percentage of all subscribers who are 'new' (appeared in the dataset for the first time within the last week): this provides information about how much data we have for subscribers (to inform statistical validity of analyses).	% of subscribers	Count_residents

Further descriptions of indicators, methods to compute them and analysis guidelines to use them will be added to our website:
https://covid19.flowminder.org/resources-for-analysts

Worked Example

Mobility analysis to support the Government of Ghana in responding to the COVID-19 outbreak

5. TBC: Guidance on analysis

Flowminder will provide guidelines on the following:
- GIS analysis of indicators with ancillary data (admin boundaries, location of services, urban/rural segmentation maps, population density maps
- Controlling for possible bias and anomalies in time series analyses
- Guidelines on interpretation and common biases in CDR data
- Visualisation of flows
- Code wrappers to enable use of aggregates and indicators in other applications (e.g. predictive models)