FLOWMINDER.ORG

Population and mobility estimates for the Democratic Republic of the Congo

Methodology report & description of datasets

Estimates version 2.0

Population and mobility estimates for the Democratic Republic of the Congo

Methodological report and description of datasets

Estimates version 2.0

Last updated : February 2026

List of content

1. Summary of methodology 2
2. Estimated residents per health zone 3
2.1. General information 3
2.2. Estimated residents per health zone 4
2.3. Estimated inflows per health zone 4
2.4. Estimated outflows per health zone 5
3. Estimated relocations between health zones 6
3.1. General information 6
3.2. Estimated relocations between health zones 6
4. Data sources 8
4.1. Call Detail Records (CDRs), Vodacom DRC 8
4.2. DRC microcensus 2021, Flowminder/WorldPop 8
4.3. DRC phone survey 2021, Flowminder 8
4.4. UN OCHA population estimates 2020 8
4.5. INS Statistical Yearbooks, 2017 and 2020 8
4.6. Health zone shapefiles, DHIS2 8
5. Limitations 9
5.1. “Measurement” biases 9
5.2. Representativity and biases 9

1. Summary of methodology

From the raw CDR datasets and the dataset on cell locations provided by the MNO we can identify the approximate location of an ID, based on the cell's location. In combination with the time of the event included in the CDR dataset, we assume the ID to be close to the cell location at that time.

It is important to note that in order to estimate residents we first need to estimate relocations. Estimated **residents per health zone per month** are conceptualised as the population that spent the majority of the month in that health zone - the **"de facto" population**. The calculation steps are:

- First, we detect each ID's **"home location"** per month as the health zone that contains the cell phone towers near which the ID was located the majority of the month.
- Then we define **relocations** as a change in this home location from one month to the next.
- We **weight** relocations between health zones using a SIM/ID-to-user parameter and a combination of the population coverage of CDR data in the origin health zone and the destination health zone.
- We sum up all weighted relocations to each health zone per month (Total inflows) and all weighted relocations from each health zone per month (Total outflows) to compute estimated **net relocations** for each health zone (Net flows). This corresponds to the difference in residents due to monthly mobility only (those who moved in minus those who moved out).
- We then calculate **baseline population estimates** for 2020 derived from **UN OCHA's 2020 estimates** per health zone and the **province-level estimates** from the **INS Statistical Yearbook 2020**.
- To these baseline estimates we **add net relocations** (inflows minus outflows) for each health zone between that month and the next month. Only the mobility detected in CDRs is used, in combination with weights, to estimate residents - not the count of CDR-derived home locations.
- We multiply these monthly estimates by **monthly population change rates** (also derived from the population estimates published in the INS Statistical Yearbook 2020) to account for those demographic components not covered in CDR data (births, deaths, immigration, emigration).
- We **repeat** this process each month up to the latest month available in CDR data, to estimate residents from the existing baseline population, internal mobility and other population changes that occurred since then.

2. Estimated residents per health zone

2.1. General information

2.1.1. Version

Version 2.0

2.1.2. Description

A dataset containing monthly estimates of the de-facto population of DRC health zones, since March 2020

2.1.3. Temporal units

Calendar months

2.1.4. Geographic units and coverage

The estimates cover all DRC health zones where CDR data are available. Health zone boundaries from DHIS2 2024 data are being used.

2.1.5. Redactions

Health zones for which the share of CDR-derived resident counts in the general population never surpassed 1% in the time series were dropped. All values smaller than 15 have been redacted to missing (and labelled "redacted (count <15)") for data protection reasons.

2.1.6. Baseline population estimates

The **subnational population estimates compiled by UN OCHA for 2020** were used as baseline estimates, in combination with **province-level estimates provided by the Institut National de la Statistique (INS) for 2019** and the **annual growth rate derived from the INS time series estimates** (1.033).

$$est\_pop\_2020\_03_a = est\_pop\_ocha\_2020_a * ((est\_pop\_ins\_2019\_adm1_a * ann\_growth\_ins\_adm0) / est\_pop\_ocha\_2020\_adm1_a)$$

$est\_pop\_2020\_03_{an}$	is the residents estimate for health zone *a* for March 2020
$est\_pop\_ins\_2019\_adm1_a$	is the INS residents estimate for the province (adm1) of health zone *a* for 2019
$est\_pop\_ocha\_2020\_adm1_a$	is the OCHA residents estimate for the province (adm1) of health zone *a* for 2020
ann_growth_ins_adm0	is the annual growth rate for the country total population (adm0) derived from the INS time-series estimates

2.2. Estimated residents per health zone

The estimate of de facto residents in health zone *a* for month *n* (est_pop_{an}) is calculated as the sum of the population for that health zone in the previous month *m* (est_pop_{am}) and the net relocations for that health zone between the two months ($est\_netflows_{amn}$), multiplied by a population growth factor (1 + $change\_rate_a$). The baseline month is March 2020 (*m=0*) for which the estimate of residents is based on existing population estimates.

2.2.1. Calculation

The estimate of residents can be expressed as a system of recursive equations:

$est\_pop_{an} = (est\_pop_{am} + est\_netflows_{amn}) * (1+change\_rate)$

$est\_pop_{a0} = est\_pop_{abase}$

Where:

est_pop_{an}	is the residents estimate for health zone *a* for the current month *n*
est_pop_{am}	is the residents estimate for health zone *a* for the previous month *m*
$est\_netflows_{amn}$	is the estimated total net relocations for health zone *a* between months *m and n*
change_rate	is the estimated average monthly population change rate in percent
est_pop_{a0}	is the residents estimate for health zone *a* for *m=0* (March 2020), the baseline population estimate (est_pop_{abase})

The net relocations estimate for health zone *a* between months *m and n* ($est\_netflows_{amn}$) is the sum of all estimated relocations to that health zone ($est\_inflows_{amn}$) minus the sum of all estimated relocations from that health zone ($est\_outflows_{amn}$):

$est\_netflows_{amn} = est\_inflows_{amn} - est\_outflows_{amn}$

See also **estimated inflows** and **estimated outflows**.

2.2.2. Upper and lower bound estimates

$est\_pop\_LB_{an}$	is the residents estimate for health zone *a* for month n, lower bound
$est\_pop\_UB_{an}$	is the residents estimate for health zone *a* for month n, upper bound

2.3. Estimated inflows per health zone

The estimated number of people who relocated (i.e. moved) to a health zone (from all other health zones) between the previous and the current month. In other words, this is the total of all inflows to a health zone.

2.3.1. Calculation

The sum of inflows to health zone *a* in month *n* is calculated as the sum of estimated relocations to health zone *a* from all other health zones *b* between months *m* and *n*:

$$est\_inflows_{amn} = \sum_{b=1}^{k} est\_flows_{bamn}$$

Where:

est_flows_{bamn} is the number of estimated relocations to health zone *a* from all health zones *b* (b≠a) between months *m and n*

For the calculation of estimated bilateral relocations (est_flow_{bamn}), see **estimated relocations**.

2.3.2. Filters and redactions

Values below 15 are redacted to missing and labelled.

2.4. Estimated outflows per health zone

The estimated number of people who relocated from a health zone (to all other health zones) between the previous and the current month. In other words, this is the total of all outflows from a health zone.

2.4.1. Calculation

The sum of outflows from health zone *a* in month *n* is calculated as the sum of estimated relocations from health zone *a* to all other health zones *b* between months *m and n*:

$$est\_outflows_{amn} = \sum_{b=1}^{k} est\_flows_{abmn}$$

Where:

est_flows_{abmn} is the number of estimated relocations from health zone a to all health zones b (b≠a), between months m and n

For the calculation of estimated bilateral relocations (est_flows_{abmn}), see **estimated relocations**.

2.4.2. Filters and redactions

Values below 15 are redacted to missing and labelled.

3. Estimated relocations between health zones

3.1. General information

3.1.1. Version

Version 2.0

3.1.2. Description

Datasets containing estimates of month-to-month relocations of population between DRC health zones.

3.1.3. Temporal units and coverage

Each month in the residents estimates covers a calendar month.

Relocation estimates, month-to-month, are operationalised as changes of estimated stay locations ("home locations") between the previous month and the reference month. For example, relocation estimates for March 2024 cover changes in the stay locations between February and March 2024.

3.1.4. Geographic units and coverage

The estimates cover **directional population flows** between DRC health zones across all 26 DRC provinces, for health zones where CDR data are available. Health zone boundaries from DHIS2 data are being used.

3.1.5. Redactions

Values smaller than 15 have been redacted (labelled "redacted (count <15)") for data protection reasons.

3.2. Estimated relocations between health zones

The estimated number of persons relocating from one health zone to another health zone between the current and the previous month.

3.2.1. Calculation

Relocations from health zone *a* to health zone *b* between months *m and n* are estimated based on CDR aggregates of relocations, i.e. the number of IDs changing their home locations from health zone *a* to health zone *b* between those months.

A home location is determined as the health zone containing those cell towers which most frequently (and in at least 3 separate weeks) routed the last event of the day of an ID over a calendar month. For each ID, relocations are then detected as a change in the health zone of the home location from one month to the next.

Then CDR aggregates of relocations (cdr_flow_{abmn}) from communal section *a* to communal section *b* between months *m* and *n* are adjusted, accounting for the estimated SIM-to-user ratio ($sims_a^{-1}$)

and for the population coverage of CDR data *(median_12m(geom(pop_coverage$_{am}$, pop_coverage$_{bm}$)$^{-1}$)*.

$$est\_flows_{abmn} = cdr\_flows_{abmn} * sims_a^{-1} * median\_12m(geom(pop\_coverage_{am}, pop\_coverage_{bm})^{-1})^{att}$$

Where:

est_flows_{abmn}	is estimated relocations from health zone *a* to health zone *b* between months *m and n*
cdr_flows_{abmn}	are CDR-derived relocations from health zone *a* to health zone *b* between months *m and n*
geom()	is the geometric mean
$sims_a$	is the nr of SIMs per user in province of origin health zone a
median_12m()	is the 12-month median for months m, m-1, …, m-12
$pop\_coverage_{am}$	is the population coverage of CDR aggregates in health zone a, month m
att	is the attenuation factor for weights (power shrinkage)

The parameter $sims_a$ is calculated as a dual-frame estimate based on the 2021 microcensus data and the 2021 phone survey data and captures the average number of Vodacom SIMs per Vodacom user per origin province. The term $sims\_per\_user\_origin^{-1}$ downscales the CDR relocation aggregates to account for multiple CDR records of the same users. This parameter is constant over time.

To avoid large fluctuations in weights due to low home location counts in some months, the 12-month median of the geometric mean is used. This median is then downscaled by an exponent (*att* = 0.5), to reduce the variance of the weights.

For the **upper bound estimates**, this downscaling is implemented by using the exponent 0.7

$$est\_flows\_UB_{abmn} = cdr\_flows_{abmn} * sims_a^{-1} * median\_12m(geom(pop\_coverage_{am}, pop\_coverage_{bm})^{-1})^{0.7}$$

For the **lower bound estimates**, this downscaling is implemented by using the exponent 0.3

$$est\_flows\_LB_{abmn} = cdr\_flows_{abmn} * sims_a^{-1} * median\_12m(geom(pop\_coverage_{am}, pop\_coverage_{bm})^{-1})^{0.3}$$

NOTE: Relocations refer to directional bilateral relocations, from health zone *a* to health zone *b*. These are not usually equal to the number of relocations from *b* to *a*.

3.2.2. Filters and redactions

Values below 15 are redacted to missing and labelled.

4. Data sources

4.1. Call Detail Records (CDRs), Vodacom DRC

For the current version of the estimates, monthly stay locations and month-to-month relocations derived from Call Detail Records from Vodacom DRC are used. For billing purposes, MNOs keep a record of an ID's activities in a database. These records are generated each time an ID makes or receives a call, sends or receives an SMS, or uses mobile data. These are called Call Detail Records (CDRs). CDRs contain information about the sending and receiving ID of a call or text, and duration of a call or data volume of data session, as well as the ID of the cell routing the call.

4.2. DRC microcensus 2021, Flowminder/WorldPop

In cooperation the DRC Institut National de la Statistique (INS), the WorldPop group of the University of Southampton and the Kinshasa School for Public Health (KSPH) Flowminder commissioned and coordinated a microcensus in seven DRC provinces: Haut-Katanga, Haut-Lomami, Ituri, Kasaï, Kasaï-Oriental, Lomami and Sud-Kivu. Data were collected via face-to-face interviews between 14 March and 27 April 2021. The final dataset includes data on 85,982 households and 367,792 individuals.

4.3. DRC phone survey 2021, Flowminder

Flowminder commissioned and coordinated a telephone survey among phone users in the DRC in order to gain empirical insights into phone use and mobility in the DRC. The objective of the phone survey was also to validate Flowminder's methods of production of mobility estimates based on Call Detail Records (CDRs), particularly of key concepts (home location, mobility, migration), and to learn about the socio-demographic structure of phone users in the DRC. Data collection ran from 15 October to 8 November 2021. CATI interviews with 7,523 respondents were completed.

4.4. UN OCHA population estimates 2020

The 2020 population estimates by health zone published by UN OCHA on the HDX platform.

4.5. INS Statistical Yearbooks, 2017 and 2020

The DRC's *Institut National de la Statistique (INS)* published statistical yearbooks in 2017 and 2020 which contained population projections per province for the years 2012 to 2019. The population estimates for 2019 and the average annual growth rate (+3.3%) derived from the time series of projections are used to calibrate the province totals.

4.6. Health zone shapefiles, DHIS2

To align with the administrative units used in the District Health Information System 2.0 (DHIS2) software, the aggregates and estimates are produced based on a mapping of network cells to DHIS2 health zone shapefiles. The relevant shapefile, dated 2024, covers 519 distinct DRC health zones.

5. Limitations

5.1. "Measurement" biases

The accuracy, precision and validity of CDR-derived statistics depends on both "measurement" and representation errors or biases.

In general, the **density of cell towers** in an area affects the **precision** of the **location** estimates. The density of cell towers is usually higher in urban areas and lower in rural areas.

The extent to which CDR-derived statistics are a reliable proxy for population mobility depends on the correspondence between **activities observable** from available CDR data and users' actual activities. Travel episodes which take place unpunctuated by communications (e.g. when SIM cards are not in use or are outside of covered areas) are non-retrievable.

Conversely, an increased **frequency of phone usage** either at the individual or at the aggregate level may correspond to increased measured mobility, as movements previously occurring in-between communications are then captured. Also, CDRs recorded for each ID are not distributed uniformly over time and depend on the frequency with which users initiate and/or receive calls, send text messages and/or use data sessions for internet connectivity on their phones. We only 'see' a subscriber when they use their phone. If they don't use their phone on a particular day, we can't confidently say where they are on that day. Subscribers with low use frequency tend to be dropped from analysis - as missing episodes from individual trajectories cannot be inferred, or as they contribute a negligible amount of information towards group aggregates. Lower frequency use could, however, be associated with different mobility patterns and typically lower mobility.

Another source of bias is the extent to which the CDRs corresponding to an individual ID actually correspond to a **single individual**. Individuals may use multiple SIM cards concurrently, or frequently change the SIM card they use, and a SIM card may be shared among multiple individuals. Ownership may also be transferred to a different user or groups of users.

Measurement biases complicate the extraction of information from CDRs (on mobility and on locations of meaningful places such as the home location) and complicate inference on the basis of behavioural traces in CDRs.

5.2. Representativity and biases

Biases in terms of representation can be categorised into three main categories. The main **coverage and selection errors** of CDR-derived statistics are linked to

- the **share of mobile phone users** in a particular area's population
- the **market share of particular MNOs** in the area, and
- the **subset of sufficiently active users** in the area that is used for analyses

Not everyone has access to a mobile phone. In Lower and Middle Income Countries, mobile phone ownership and (to a lesser extent) access has been shown to be disproportionately skewed

towards individuals who are male, educated, urban-resident, wealthy and working-age, whose mobility may differ from that of the general population.

Where CDR data are only available from a single MNO, this can add an undercoverage error, as it is not uncommon for the characteristics of a particular operators' subscriber base to be skewed towards relatively wealthier or relatively poorer individuals.

That is, CDRs capture only a **non-random subset of the population**, which may also change over time due to changes in phone and SIM prices, extensions in signal coverage, marketing and pricing strategies of MNOs, and demographic and societal changes. This may result in increased phone use for previous low-use groups, and lead to changes in population representation. These undercoverage errors and their temporal variations can distort the interpretation and use of observations derived from CDRs.

MNOs market shares and coverage areas change over time, which will have an impact on the analysis of CDR data. Furthermore, the error rate in MNO's data management and due to network issues is generally high, rendering analysis of their data more difficult due to data gaps in time and space as well as inconsistencies. These issues in the raw data are monitored and corrected by Flowminder before the analysis stage.

Contact us

For queries or information about the DRC estimates, the methods presented in this document or on mobile data analytics, please contact us at **info@flowminder.org**.