FLOWMINDER.ORG

Using survey data to correct for representation biases in mobility indicators derived from mobile operator data to produce high-frequency estimates of population and internal migration

Roland Hosner (roland.hosner@flowminder.org), Zachary Strain-Fajth, Véronique Lefebvre
Flowminder Foundation

Background

While mobile operator data constitute an important source of evidence on population mobility, particularly in data-poor settings such as Low- and Middle-Income Countries (LMICs), they remain partial in terms of population coverage, and prone to representation biases which are difficult to measure, control, and correct for in the absence of independent auxiliary data.

The use of mobile operator data (including Call Detail Records, CDRs) to estimate changes of sub-regional population counts over time and population mobility often relies on a series of assumptions, including that movements observable for mobile phone users are similar to the movements of the general population. But mobile phone users have been shown to be different from the general population in many characteristics (gender, age, socio-economic status (SES) including education, degree of urbanisation of place of residence), and such characteristics also lead to differences in their mobility compared to that of the non-phone-users. As a result, CDRs cannot be used as a source of population statistics unless representation biases are corrected for.

Although methodological development and research on the application of CDR data has progressed in recent years, few solutions have been offered so far for the adjustment of such representation biases in CDR-based indicators. The inherent biases of these data require correction through joint modelling with traditional data sources such as surveys. Unadjusted indicators may severely misrepresent population movements and deducted population distributions, particularly in regions where phone use is low and subscriber numbers are small.

We see from the available survey data that a range of important parameters differs between relocations (per combination of origin and destination locations): the average number of SIM cards used, the number of travellers per SIM card, the share of phone users and the MNO market shares do differ between relocation flows. This also means that we observe mobility differences between phone users and non-users.

These problems of selection bias and representativity in general are common to all big data analyses, where often the erroneous assumption is made that the quantity of data renders representation bias negligeable. We propose here a method that corrects for such biases in CDR-derived indicators. What is more, the method could be generalised to other types of estimates and big data datasets.

Data

Flowminder has ongoing access to CDR data in Haiti, Ghana and the DRC, and has commissioned or joined primary survey data collection in these countries. Our method relies on CDR aggregates, survey data and existing population estimates from National Statistical Offices, the United Nations, WorldPop and other sources.

For the DRC, we use CDR aggregates from Vodacom with survey data from a microcensus and a telephone survey in 2021 to estimate relocations and residents by health zone for all months since February 2020.

For Haiti, we use CDR aggregates from Digicel with survey data from a general population survey in 2022 to estimate relocations and residents by communal section for all months since January 2020.

For Ghana, we will combine CDR aggregates from Vodafone with census data from 2021, survey data from a telephone survey and a general population survey in 2022 to estimate relocations and residents by district for all months since December 2019.

Method

As a first step, we assess all available national, regional and sub-regional population estimates from National Statistical Offices and other sources. Sub-regional population estimates for the baseline month - the first month for which CDR aggregates are available - are then derived from existing estimates or projections. Importantly, we are not using CDR-derived counts of home locations for scaling to estimate residents, because our analyses show these counts strongly depend on phone coverage and phone use, and are not suitable for direct scaling.

In the second step, we apply adjustment and scaling factors only to CDR-derived relocations, i.e. detected changes of home locations or stay locations over time. These CDR-derived inflows and outflows of (frequent and locatable) phone users are adjusted and scaled based on survey-derived parameters. The approach is to scale the bilateral relocations between sub-regions (flows) from one month to the next, aggregate total scaled inflows and outflows per sub-region, and add scaled net flows in a cumulative manner over time to baseline estimates to arrive at time-series estimates of residents. As a result, the difference between total inflows to and outflows from a subregion corresponds to the change in estimated residents from one month to the next.

As a final step, we apply factors to adjust for overall population growth or decline, derived from estimates provided by National Statistics Offices or the United Nations.

Discussion

Our method corrects for differences in mobility between phone users and non-users and further biases observed in the survey data. However, there are few data sources to validate such estimates. Additionally, to enable the ongoing production of population and mobility statistics from mobile operator data, longitudinal survey data are needed to capture potential changes in these mobility differentials of the phone using and non-phone using populations.

In summary, we have been able to develop and apply a bias-correcting methodology to produce estimates of internal migration and sub-regional population change in two LMIC countries (Haiti and the DRC) and have started work on a third country (Ghana). This data can be used for a wide range of use cases, from the health sector to humanitarian work, disaster preparedness, and to official statistics, including reporting on development indicators. For example, these estimates can highlight sub-regions with monotonic population increases (above natural population growth), or sub-regions with fluctuating populations such as those affected by large population displacements in the DRC and Haiti.