Flowminder standards in producing mobility and population estimates from call details records in low- and middle-income countries

Véronique Lefebvre*[a], James Harrison[a], Jonathan Gray[a], Roland Hosner[a], Galina Veres[a], Chris Brooks[a], Robert Eyre[a], Thomas Smallwood[a], Romain Goldenberg[a], Zachary Strain-Fajth, Joachim Jellinek, John Roberts, Xavier Vollenweider, Sophie Delaporte, Cathy Riley, Daniel Power, Linus Bengtsson

Flowminder Foundation

* Corresponding author. email: veronique.lefebvre@flowminder.org

[a] Equal significant contribution

Processing and analysing CDRs in LMICs: Extracting the population mobility information contained in Call Detail Records (CDRs) is of critical importance in data poor contexts such as in low- and middle-income countries (LMICs), where it can support humanitarian and human development efforts. Such contexts however present additional challenges compared to high-income countries (HICs) for mobility analysis from mobile operator data: often only CDRs are available and they are sparser over time and space, mobile networks are more unstable, particular in crises, which are often more frequent, and the geographic coordinates of cells are sometimes missing and erroneous. Further, the proportion of the general population using mobile phones is significantly lower in LMICs (e.g. 47% of households on average in 7 provinces of the DRC, down to 35% in the more rural provinces) and therefore differences in the mobility of phone users and non users have a larger impact on the representativity and applicability of CDR-derived statistics. At Flowminder we have specialised in addressing such challenges and we present here an overview of our live systems, from ingestion and automated quality assurance (QA) checks of pseudonymised CDR data and cell data, to the extraction of mobility information from CDRs and bias correction using survey data, resulting in the semi-automated production of a set of standard indicators, ready to be disseminated to decision makers in LMICs through dashboards, standard reports or as data sheets.

Flowminder system: At Flowminder we made the choice to conduct all CDR data processing within the firewall of the Mobile Network Operator (MNO). While this comes with constraints on compute power and memory, it is essential to protect subscribers' data privacy. We also ensure we do not have access to subscriber phone numbers and instruct MNOs on how to pseudonymise the records. We built the 'FlowKit software' to handle all data processing, from ingesting the pseudonymised CDRs to outputting mobility and population estimates. FlowKit is an open-source CDR data processing toolkit, consisting of a PostgreSQL database along with tools for automating data ingestion and QA, implementations of our methods for extracting mobility information from CDRs, scaling, combining and formatting the mobility estimates for end usage. We describe below the different steps of the pipeline as per our overview diagram in **Figure 1**.

Pseudonymised CDR ingestion and checks: Pseudonymised CDR data are received from the MNOs daily, and are automatically ingested into the FlowKit database, ready for processing. QA checks are run on the CDR data during ingestion to ensure data quality, and exclude incomplete or unreliable data from analysis. Cell locations are received from the MNO at regular intervals (ideally monthly), and QA checks are performed to identify mislocated cells. We also set up a monitoring system that tracks the status of our at-MNO servers, FlowKit installations and automated QA checks, and sends alerts so that any detected problem can be addressed quickly. This is important particularly for crisis preparedness so that CDRs are available and specific processing can begin promptly if a crisis occurs or is forecast.

Mobility extraction and bias correction: We use FlowKit to process the pseudonymised CDR data within the MNO's firewall to extract mobility information. Recent method development led to a more robust detection of subscriber's home (or 'stay') location from sparse CDRs, and estimates of relocations. Further we collect field and phone survey data to correct for representation biases in the number of relocations between each subregion summed over subscribers. We then derive monthly population estimates (residents) from the adjusted relocations and from baseline population estimates, which scales the relocations and attenuates the effect of phone usage changes and subscriber churn on changes of resident numbers. Similar work to disambiguate between phone usage and mobility for population presence and travel and to scale these indicators is ongoing.

Reporting and dissemination: We have standardised the way we provide analytical reports and data (via datasheets or dashboards) to inform a range of applications, from disaster management, to official migration statistics and immunisation

planning, and are working towards further automating the production of these end products for each application. Currently we produce mobility and population estimates for Haiti, Ghana and the DRC.

Flowminder standards: While specific to CDRs and to data issues encountered in LMICs, we propose that our infrastructure characteristics, QA checks, automation framework, methods for mobility information extraction and for correcting representation biases, as well as our standard set of mobility indicators may be of use to anyone attempting to produce mobility statistics from CDRs or related data types in any country.

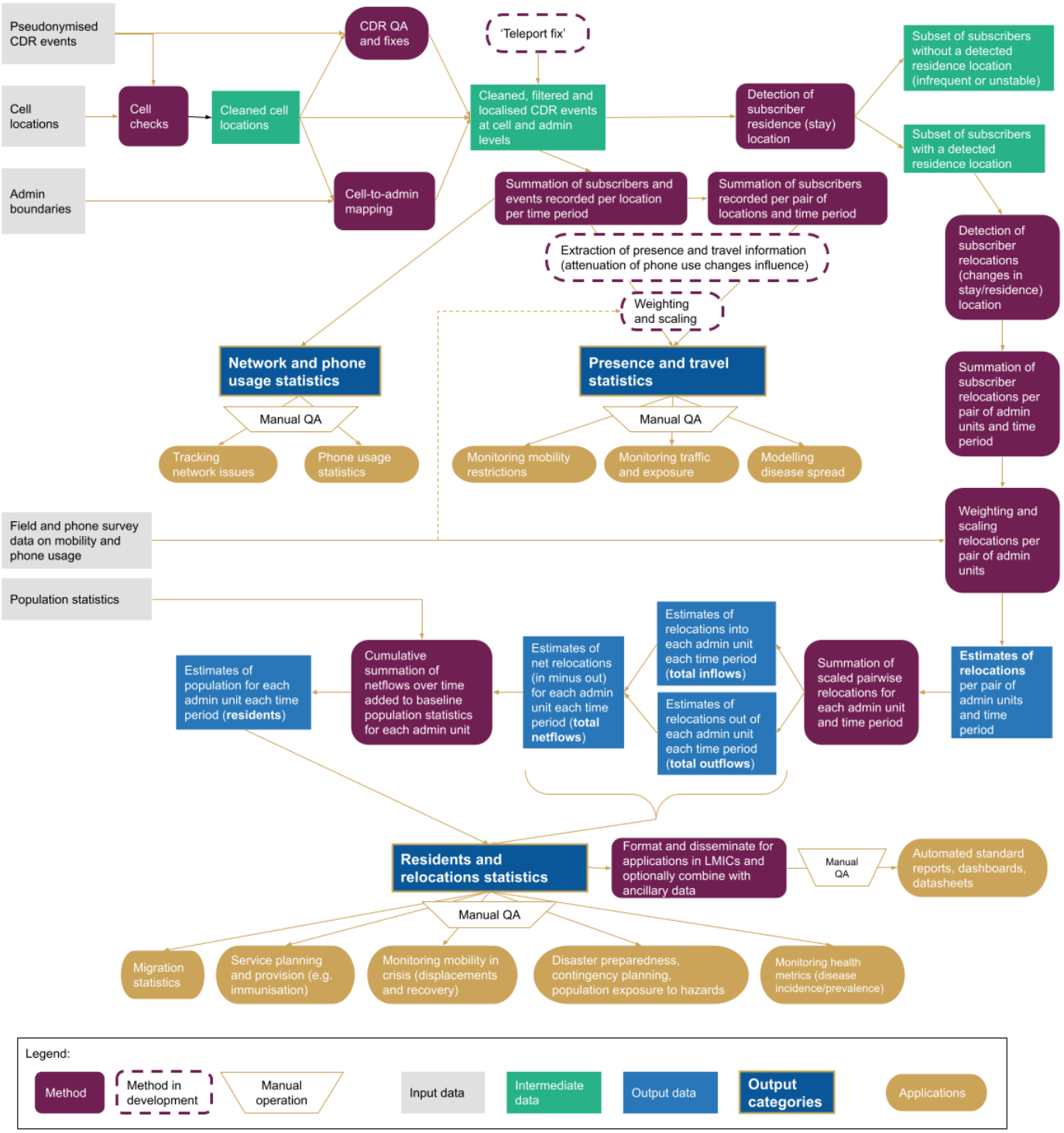

Figure 1: Flowminder pipeline for processing and analysing call detail records to inform humanitarian and development applications in low- and middle-income countries.