Correcting measurement biases in the detection of long and short stay locations in sparse Call Detail Records (CDRs)

Galina Veres*, Jono Gray, James Harrison, Véronique Lefebvre
Flowminder Foundation
*corresponding author email: galina.veres@flowminder.org

Introduction. With mobile phone technologies continuing to spread around the world, Call Detail Records (CDRs) represent an attractive additional data source for inferring internal migration in addition to traditional data sources such as surveys and administrative records, especially in low and middle-income countries (LMICs), where traditional statistics can be either unavailable or difficult to collect. Stay locations are the basis of most mobility statistics derived from location data and CDRs. Detecting stay location is crucial to migration statistics (Where do people live? When and to where do they change their residence?), to disaster statistics (Where have people lost their homes? Where are they displaced to? Did they then return home and when?), and to a multitude of other applications from informing disease spread to tourism statistics. Though a number of methods were proposed in the literature to detect stay locations from CDR data, some challenges remain especially in LMICs - where CDRs tend to be sparser - due to irregularity and low frequency of phone use, network instability, so called 'ping-pong' effect as artefact of mobile communications, and resulting conflation between changes in phone usage and changes in mobility. We present our solution to detect stay locations (long and short stays) and relocations from CDRs, which addresses the above issues related to particularly sparse data in LMICs and can be run on mobile operator infrastructure (constrained in memory and compute power) to ensure data privacy.

Problem formulation. In this paper, we address two problems of using CDR data for migration and disaster statistics: 1) robust detection of stay locations and relocations in individual CDR traces and 2) developing aggregated mobility indicators corrected for biases stemming from changes in phone usage. Stay location is often assumed to be a location of the last call of the day for an overnight stay location detection, or the modal location of the last call of the day for longer periods (e.g. detecting 'home location'). However, such methods of relocation detection between two stay locations lead to ~ 79% of false discoveries at the daily level, from experiments we conducted on CDRs traces using a subset of 781 Digicel subscribers in Haiti for whom we manually labelled stay locations (at administrative level 3) and relocations. This indicates the need for better performing methods to detect relocations and stay locations from individual CDR traces. Another source of error arises from summing the detected stay locations for each region each day or each month to estimate the number of 'residents' of each region (or subscribers who 'stay' in each region). However, numbers of residents and their temporal variations computed as 'stay location counts' contain both variations in phone usage and variations in internal mobility. We quantified the proportion of temporal variation in 'stay location counts' due either to phone usage or to mobility, using CDRs from Digicel Haiti, and found that only 23% of the monthly 'stay location counts' variations (on average across regions) in a 24-month study period are attributable to mobility in this case. This indicates the need to derive resident numbers directly from observed mobility (relocations).

Method development and validation. We propose a fast and elegant solution to fix such measurement biases, while ensuring it is operationally feasible: in near real time (updating every day), and on infrastructure constrained in compute power and memory. To ensure data privacy constraints we impose for all computations on individual data to be done on a server located at the mobile operator premises, behind their firewall.

We improved on the common methodology of capturing the modal location of the last call of the day by using a system of two moving windows: a short window to estimate a daily overnight location and a longer window (length depending on the type of stay to be detected) within which we check for a dominant location. For the short window, we tested several methods and concluded the last call of a day method is a trade-off between performance and required execution time for operational purposes. We compared the modal last call of a day, the modal distinct day and the anchor methods for detecting stay locations within 7-day windows at the fine spatial level of the group of cell towers. The last call of a day and anchor methods performed similarly for stay location detection, based on our labelled subset. Recall was better by 5% for the last call of a day, precision and false positive rate was better for anchor method by 19% and 24% respectively for relocation detection taking location of relocation into consideration. However, execution time increased by 2.5 times for the anchor method in comparison to the last call of a day, which was retained as the method to use as 'first pass' on the data in the short window. We use a 7 day rolling window (short window) and assign a daily stay location as a modal location of the last call of a day as a 'first pass' over individual trajectory. In a 'second pass' over the time series, we use a longer window (e.g. 28 days) to search for a dominant location within the locations returned by the modal last call of day location method. This is particularly relevant when searching for home location, when the subscriber could be absent from their home. If there is a dominant location, then a subscriber is assigned this location as their residence location, otherwise the subscriber is 'unlocatable', i.e. avoiding to assign a residence or stay location when a subscriber has been mainly on the move. This effectively reduces the number of false relocations compared to the simpler common method, and creates a sample of subscribers who are stable enough and active enough so that their stay location can be detected. This two windows method also permits an approximation of stay duration that is robust to missing data and noise. Then a relocation is simply detected as a change in stay location for each subscriber. We tested the proposed method for detecting residence location and relocation on a manually labelled subset of subscribers and found that false discoveries were reduced by 33% for daily relocation detection, from ~79% for a simple modal last call of a day location method to 53% (our novel method). We are working towards further reduction of false discoveries in our work of suppressing the ping-pong (or re-routing) artefact of CDR data by weighting the number of relocations between regions.

Secondly, while this approach creates a more robust detection of stay location at the individual level, the number of residents (or stays) in a location cannot simply be computed by summing the number of stay locations detected, as this may mainly be driven by the number subscribers becoming (un)locatable. To ensure that we only retain variations in the number of residents derived from mobility (changes in stay locations), we propose to estimate the number of residents directly from the number of relocations (e.g. change in stay locations between two months). The difference in the number of residents between two months is equal to the total number of subscribers relocating into the location minus the number who relocated out of the location. The number of estimated residents is then the cumulative sum of these differences added to a baseline or starting time as described by

$$\widehat{residents}(i,t) = \widehat{residents}\,(i,0) + \sum_{j=1}^{t} (reloc\_to(i, i\,-\,1, j) - reloc\_from(i, j\,-\,1, j)),$$

where $i\,(i\,=\,1,..., N)$ is a location i from N locations in the area of interest, $\widehat{residents}(i,0)$ is an estimated number of residents in the baseline period or start point. Using this method, large fluctuations in subscriptions (often observed in urban areas) are suppressed. The method ensures that the variations of our resident indicator are now derived from observed mobility, at least for short time periods (changes in the subscriber base will still create a drift over years, which needs to be addressed by further weighting factors and auxiliary survey data - a problem we are also working on).

Operational use. We have been able to use our stay detection method and resident indicator design on an operational level, computing resident estimates and internal migration monthly in 3 countries (Haiti, Ghana and the DRC), and adjusting them for representation biases using survey data. Such data can be used for service provision planning and for refining other statistics taking population mobility into account such as disease incidence and prevalence. We also use the method with shorter time windows to detect disaster-driven displacements and returns and inform disaster management.