scientific data

OPEN

DATA DESCRIPTOR

A highly granular temporary migration dataset derived from mobile phone data in Senegal

Paul Blanchard[1,2] ✉ & Stefania Rubrichi[3]

Understanding temporary migration is crucial for addressing various socio-economic and environmental challenges in developing countries. However, traditional surveys often fail to capture such movements effectively, leading to a scarcity of reliable data, particularly in sub-Saharan Africa. This article introduces a detailed and open-access dataset that leverages mobile phone data to capture temporary migration in Senegal with unprecedented spatio-temporal detail. The dataset provides measures of migration flows and stocks across 151 locations across the country and for each half-month period from 2013 to 2015, with a specific focus on movements lasting between 20 and 180 days. The article presents a suite of methodological tools that not only includes algorithmic methods for the detection of temporary migration events in digital traces, but also addresses key challenges in aggregating individual trajectories into coherent migration statistics. These methodological advancements are not only pivotal for the intrinsic value of the dataset but also adaptable for generating systematic migration statistics from other digital trace datasets in other contexts.

Background & Summary

The movement of people across space is intricately linked with economic activity and economic development processes. Previous research examining mobility within countries have predominantly focused on the significance of permanent migration in fostering growth and structural transformation[1,2]. Such studies mostly delve into the factors influencing and hindering the reallocation of individuals from a less productive rural sector to an urban non-agricultural sector.

Yet, a growing body of research has highlighted the importance of other forms of short-term mobility in developing countries, such as temporary migration movements. These flows of internal movements have been found to be incredibly common and to largely exceed permanent moves[3–5]. They have at first been portrayed as a sign of failure of rural livelihoods[6,7] but have also been described more recently as a structural component within households' livelihood strategies[4,5,8]. Despite its proven significance, temporary migration is seldom integrated into national statistical systems in a systematic way. Short-term movements are intrinsically difficult to measure (e.g., due to attrition and recall biases) and require specialized – and oftentimes costly – surveys[9]. More importantly, the rare surveys measuring temporary migration often adopt standard definitions that do not necessarily allow to capture relatively short trips, which are nonetheless frequent[4]. Temporary migration patterns thus remain poorly documented at national scales, and especially so in sub-Saharan Africa.

This article introduces an open access dataset containing highly granular temporary migration estimates derived from mobile phone data in Senegal. The dataset includes measurements of migration flows and stocks across 151 locations spanning the entire country – these locations encompass the rural areas of 112 districts and 39 cities. We provide estimates of migration movements directly observed in the data, which are further extrapolated using correction methods to estimate total migration within the broader population. Estimates are provided for each half-month period over the 2013-2015 timeframe, considering mobility events lasting from 20 to 180 days. The unique level of granularity offered by this dataset aims to furnish researchers from various disciplines, including economists, demographers, environmental sociologists, and others, with a robust foundation of information to advance our understanding of the characteristics, causes and consequences of temporary movements. Comprehensive data on short-term movements are indeed crucially needed to inform development practitioners and policy makers on various matters and support the design of adequate policy interventions.

[1]Institut de Recherche pour le Développement, Paris, 75010, France. [2]Department of Economics, Trinity College Dublin, Dublin 2, Ireland. [3]Orange Innovation - SENSE, Châtillon, 92320, France. ✉e-mail: blanchap@tcd.ie

These interventions include, for instance, responses to the effects of environmental shocks, climate change, epidemics and conflicts on short-term population dynamics. Additionally, real migration measures derived from mobile phone data can be employed to inform and calibrate models that generate synthetic OD data, including emerging AI-based tools[10].

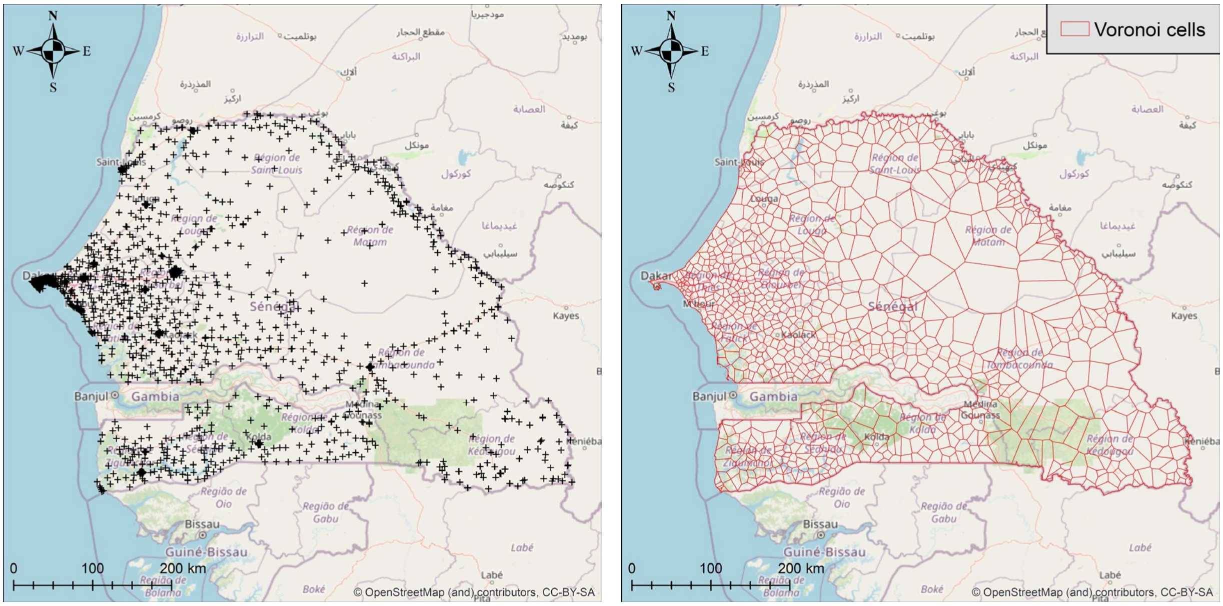
The development of the proposed dataset arises within a context where digital footprints generated by the use of digital services and devices have emerged as a promising source of big data for measuring human mobility on broader scales. More importantly, these data exhibit a proven capability to capture subtler human movements with an increased spatio-temporal granularity[11,12]. In the field of migration studies, they represent a significant opportunity to improve migration estimates and better inform policy[13–16]. In particular, some studies have leveraged mobile phone metadata to quantify seasonal and temporary migration movements in developing contexts[17,18]. However, none of the corresponding datasets of migration estimates have been made publicly accessible, and the methods usually employed to derive migration measures are subject to certain limitations. For instance, migration events are typically identified as a change in the estimated location between two consecutive time periods – e.g. calendar months – calculated as the modal location observed during those time periods[17–20]. The regularization of a user's trajectory at a harmonized but coarser temporal resolution – i.e. by calculating monthly locations – necessarily causes some measurement error on the exact start and end dates of migration events as well as on their actual duration. It also implies that relatively short migration events with a duration that is comparable to the time resolution considered are potentially missed. For example, an individual can be seen at his home location from March 1 to March 16 of some year, then temporarily migrate for 28 days from March 17 to April 14, and return home from April 15 to the end of the month. Since the majority of days in March and April are spent at the home location, a frequency-based method using monthly locations will assign the user to that location in both months and the migration event will not be detected. Moreover, those methods provide a limited characterization of the direction of migration flows. Since migration events are simply identified as a location change, it is not possible to distinguish between a departure from and a return to a primary home location. Finally, the production of time-disaggregated temporary migration measures poses a number of methodological issues which have not been clearly addressed yet. Most notably, periods of inactivity necessarily induce some degree of uncertainty in the timing and duration of temporary migration events. This in turn creates situations where, for instance, the assignment of an identified migration departure date to a particular time period (e.g. a week or a month) can be ambiguous if the user is unobserved for some period of time before the departure date.

The dataset is a product of a thorough methodological framework meticulously designed to address a number of these issues. A migration event detection algorithm is developed based on a conceptualization of human mobility on three distinct temporal scales: the micro-, meso-, and macro-scales. Micro-movements refer to short-term mobility such as daily trips, commuting, or visits to cities. Meso-movements involve temporary changes in the usual place of residence and are the primary focus of this paper. Macro-movements are long-term changes in residence, i.e. permanent migration. The granularity of mobile phone data allows to capture movements across all three scales, but tailored algorithmic methods are required to isolate movements at a specific scale. The proposed migration event detection algorithm builds on recent work by Chi *et al.*[21], who demonstrated how a clustering approach can enhance accuracy compared to traditional frequency-based methods. An important addition relies on the estimation of a primary residence location prior to detecting temporary migration events, which enables the clear characterization of migration flows' direction by distinguishing departures from and returns to a home location. A set of well-defined algorithmic rules allows to aggregate user-level migration trajectories into consistent migration statistics at a desired spatio-temporal scale. They specifically account for issues related to sampling irregularity while maximizing the retention of information contained in phone-derived trajectories. Furthermore, the validation of migration estimates incorporates systematic methods and supplementary data sources to carefully assess the representativeness of a sample of phone users for generating migration statistics. In addition to the intrinsic value of the dataset, this suite of methodological tools thus constitutes a valuable outcome, as these can be readily adapted and applied to other digital trace datasets for systematically generating migration statistics in other contexts.

Methods

Call Detail Records. We use Call Detail Records (CDR) from the main telecommunication company in Senegal (Sonatel) and spanning the period 2013 to 2015 as the primary input for the construction of temporary migration estimates. CDR are mobile phone metadata collected by telecommunication providers for billing purposes. Each data record corresponds to an instance where a user made or received a call (or a text message), and is associated with a set of attributes that typically include: the phone number of the user, the starting time and date of the call and an identifier of the phone tower that processed the call. During the study period, Sonatel was largely dominating the mobile telephony market: in 2014, 88% of mobile phone users owned a Sonatel SIM card[22].

A separate dataset provides the point coordinates of each phone tower. For the study period (2013-2015), the Sonatel network was comprised of 2,071 phone towers (Fig. 1, panel **a**). The set of phone tower coordinates is converted into a set of contiguous cells via a voronoi tesselation. Each voronoi cell coincides with the smallest area containing the point location of a device connecting to the corresponding phone tower or, equivalently, it is the approximate area covered by the phone tower. Phone towers are distributed unevenly across the country and their density typically increases with population density. Cells belonging to a single city are thus merged together for mainly two reasons. Firstly, temporary migration is conceptualized as movements across locations such as villages and cities but exclude intra-urban mobility. Secondly, equalizing the sample of cells in terms of their size helps mitigate systematic measurement errors. City polygons are defined based on the GHS Settlement Model 2015 product (GHS-SMOD)[23], which identifies 33 urban settlements in Senegal. Voronoi cells intersecting a city polygon are grouped together to form a city cell. However, some secondary urban areas are not captured by



(a) Phone towers **(b)** Final voronoi cells

Fig. 1 Distribution of phone towers (panel **a**) and final voronoi cells (panel **b**) after aggregating cells within cities.

GHS-SMOD, resulting in clusters of small cells. Clusters of phone towers that are within a distance of less than 2km from each other are thus detected and the corresponding voronoi cells are merged. This process yields an additional 6 cities. A final network of 916 cells forms a partition of the country extent (Fig. 1, panel **b**) and is comprised of 39 urban cells and 877 rural cells.

To ensure privacy, CDR were pseudonymized by the mobile phone provider via a procedure that replaces phone numbers with unique identifiers. Distinct pseudonymization procedures were applied for the year 2013 and the period 2014-2015. As a result, the unique identifier assigned to a single phone number differs between the two periods and both datasets are processed separately. The 2013 dataset has 9,386,171 unique identifiers and over 28.3 billion records, while the 2014-2015 dataset is comprised of 12,244,494 unique identifiers for over 67 billion observations. To address concerns on the presence of bots and call centers in the sample, 102,313 (resp. 98,086) identifiers that have over 100 records per day on average are removed from the 2013 (resp. 2014-2015) sample – they account for a total of over 3.5 billion (resp. 5.4 billion) records.

The pseudonymised CDR dataset was temporarily stored within a secure platform at the operator own premises. Preprocessing and spatiotemporal aggregation was carried out in the same platform by personnel of the network operator.

The detailed CDR data are proprietary and confidential. We obtained access to these data from Orange Sonatel within the framework of a collaborative project and with the agreement of the CDP (Senegalese data protection authority). Access to the full dataset can be requested from Orange on a contractual basis.

Selection of relevant subsets of users to derive robust migration statistics.

Phone users must satisfy minimal observational constraints – high observation duration, high frequency of observation, limited periods unobserved – in order to ensure a high level of accuracy in the migration detection procedure. On the other hand, higher observational constraints come at a cost of a lower statistical power since they decrease sample size. Additionally, excluding users based on sampling characteristics may exacerbate selection biases on the cross-section since phone usage patterns can vary with individual characteristics[24] that potentially correlate with migration decisions. In this respect, previous studies have applied observational criteria aligned with their measurement objectives[17,19,20,25], but have largely disregarded the impact of those constraints on sample composition. Here, we quantify both the benefits of stricter observational constraints, i.e. reduced measurement errors in migration estimates, and their associated drawbacks, i.e. smaller sample sizes and selection biases. The corresponding quantitative analyses are detailed in the Technical Validation section.

On that basis, we define two subsets of users with distinct observational constraints, each reflecting a deliberate trade-off between the benefits and costs of applying stricter observational criteria. The primary subset, denoted as *subset A*, is created by applying relatively strict observational constraints: users must be observed for a period covering at least 330 days, on at least 80% of those days, and with periods unobserved not exceeding 15 days. The minimal length and frequency of observation are specifically designed to guarantee a baseline level of accuracy in determining users' home location, detecting temporary migration events and estimating departure and return dates. On the other hand, imposing a maximum period unobserved helps mitigate the risk of measurement biases arising from non-random attrition, i.e. periods of inactivity precisely coinciding with users being in migration.

Subset	Unique identifiers, 2013	Unique identifiers, 2014-2015	Total records
subset A	1,990,754	2,041,566	47,857,866,128
subset B	3,377,994	3,746,640	61,566,733,246

Table 1. Number of unique identifiers and total number of records in *subset A* and *subset B*.

We also define a secondary subset with lower observational requirements, denoted as *subset B*, associated with a lower level of accuracy in the migration detection procedure – but still reasonably high (see Technical Validation section) –, but a larger sample size and a lower degree of selection induced by the filtering procedure. Specifically, subset *subset B* includes users observed over a period of at least 250 days, with at least 50% of days observed, and a maximum period unobserved of 25 days.

The final dataset includes separate temporary migration estimates obtained from both *subset A* and *subset B*. *Subset A* serves as the primary sample of analysis, offering a high migration detection rate while maintaining a substantial sample size and minimal selection bias relative to the initial dataset. In contrast, migration estimates derived from *subset B* are intended to facilitate robustness checks and allow researchers to test the sensitivity of their results to variations in filtering parameters. The number of unique identifiers and total number of records for both subsets are summarized in Table 1. *Subset A* has 1,990,754 unique identifiers in 2013, 2,041,566 for the period 2014-2015, amounting to a total of 47.9 billion records. By contrast, *subset B* has 3,377,994 unique identifiers in 2013, 3,746,640 for the period 2014-2015, and a total of 61.6 billion records. For the period 2014-2015, the number of phone users in *subset A* and *subset B* corresponds to 26% and 47% of the adult population (i.e. aged 15 and older), respectively. Drawing from the sensitivity analysis of migration detection accuracy relative to observational constraints, we estimate that the algorithm detects at least 90% of migration events in *subset A*, and a reasonably high detection rate of 76% maintained in *subset B*.

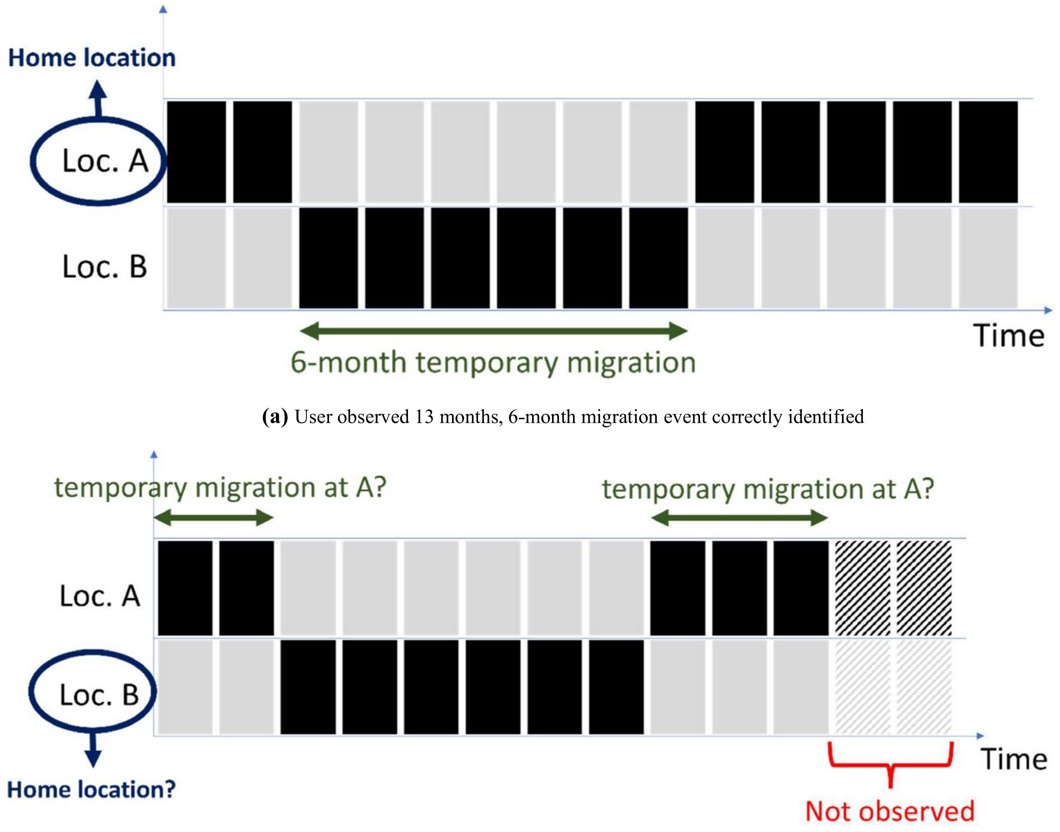
Migration event detection. The migration event detection algorithm is structured around a conceptualization of human mobility on three distinct temporal scales, considering that individuals can move within a set of predefined locations at a fixed spatial resolution (e.g., at the voronoi level). First, short-term mobility events such as daily commutes, short trips to cities or weekend getaways are characterized by a short duration, typically a few days. They correspond to movements at a *micro*-scale. Second, temporary migration events correspond to an individual moving from a primary home location to a host area for a period of time going from a couple of weeks to several months before returning to his home location. Those are movements at a *meso*-scale and form the central focus of the paper. Third, permanent migration moves imply a long-term change in the usual place of residence and are defined as movements at a *macro*-scale. The time intervals associated with these mobility events are called micro-, meso-, and macro-segments, respectively. For any given individual observed over some period of time, the sets of micro-, meso- and macro-segments constitute three layers of mobility that define the micro-, meso- and macro-location of that individual at any point in time, where those locations are defined at the same spatial resolution. Note that, in this framework, the macro-location is thus considered as the usual place of residence (i.e. the home location). Given the length of observation and the frequency with which phone users are observed, a raw CDR trajectory generally allows to capture movements at all three scales. As a result, one of the main challenges of identifying segments at a higher scale (e.g., at a meso-scale) is to develop algorithmic methods that smooth out noisy patterns created by movements at lower scales (e.g., at the micro-scale).

With these concepts in mind, a four-step methodology is developed to identify temporary migration events in individual CDR trajectories. Firstly, a hierarchical frequency-based procedure is implemented to estimate hourly, daily and monthly locations for each user over his period of observation (*Step 1*). Then, a clustering method is applied to monthly locations to detect macro-segments, which allows to define the usual place(s) of residence over the observed period (*Step 2*). A similar clustering algorithm is applied to daily locations for the detection of meso-segments (*Step 3*). Finally, temporary migration events are identified by overlaying meso- and macro-segments: they correspond to meso-segments at a location which is not the usual place of residence (*Step 4*). Several parameters are introduced throughout the detailed description of the data treatment methods below. For the reader's convenience, these parameters are summarized in Table 2 along with their definition and values used to produce the final temporary migration dataset.

To effectively differentiate between long micro-segments and short meso-segments, as well as long meso-segments and short macro-segments, empirical criteria on the duration of mobility events are essential. In this respect, the detection algorithm considers meso-segments with a duration ranging from $\tau_{meso}^{min} = 20\ days$ to $\tau_{meso}^{max} = 180\ days$. Consequently, macro-segments are naturally defined as periods of at least τ_{meso}^{max} reflecting the continuous presence of a user at a single location at the macro scale. The relatively low value for τ_{meso}^{min} allows to capture short migration events, which are more prevalent and often overlooked in survey data compared to longer-term migration spells[4]. On the other hand, the upper-bound duration τ_{meso}^{max} is constrained by users' observation duration. Specifically, and as a basic heuristic, for a period of τ_{meso}^{max} days of continuous presence of a user at some location to be identified as a temporary migration event, the user must be observed at least twice that duration. Indeed, this is the limit over which it is possible to determine that the user spent the majority of his time at a location that can effectively be identified as his primary home location, which then allows to correctly identify the period of τ_{meso}^{max} days at a distinct location as a temporary migration event. An illustrative example is provided in Fig. 2, where a hypothetical user spends consecutive months at distinct locations A and B. Considering a value for τ_{meso}^{max} equivalent to 6 months, Fig. 2 (panel (**a**)) shows that the full 13-month period of observation – i.e. more than twice τ_{meso}^{max} – effectively allows to identify location A as the home location and to

Macro- and meso-segment detection		
Parameter	**Description**	**Value**
τ_{meso}^{min}	Minimum duration imposed to classify a meso-segment as a temporary migration event	20 days
τ_{meso}^{max}	Maximum duration imposed to classify a meso-segment as a temporary migration event. It also sets the minimum duration for identifying macro-scale movements.	180 days
$\varepsilon_{gap}^{macro}$	Maximum observation gaps allowed to group consecutive observed months at the same location in the first step of the macro-segment detection.	6 months
ε_{gap}^{meso}	Maximum observation gap allowed for grouping consecutive observed days at the same location (first step of the meso-segment detection). It also sets the maximum duration permitted between consecutive groups at the same location for merging them (second step of the meso-segment detection).	7 days
ϕ	For any detected meso-segment, it is the minimum fraction of days observed at the identified meso-location required to validate the segment. This helps to limit cases where a meso-segment might capture frequent movements between multiple locations rather than a temporary migration event at a single location.	0.5
Aggregation into statistics		
Parameter	**Description**	**Value**
ε^{tol}	Tolerance parameter setting the maximum acceptable time unobserved before (resp. after) a period *t* during which a meso-segment starts (resp. ends) in order to still consider that the user effectively departed (resp. returned) during *t*.	7 days
Σ	Minimum overlap between a temporary migration segment and a period *t* to count the corresponding user as being in migration during *t*.	8 days

Table 2. Parameters for macro- and meso-detection procedures and aggregation into migration statistics.


Home location
Loc. A
Loc. B
Time
6-month temporary migration
temporary migration at A?
temporary migration at A?
Loc. A
Loc. B
Home location?
Time
Not observed

(a) User observed 13 months, 6-month migration event correctly identified

(b) User observed 11 months, failure to detect a 6-month migration event

Fig. 2 Illustration of the importance of users' observation duration for the choice of τ_{meso}^{max}. Black bars represent the location of a hypothetical user over consecutive months. For simplicity, all months are assumed to have 30 days and τ_{meso}^{max} is assumed to be set to 180 days. Panel (**a**) depicts the full 13-month observation period, correctly identifying location A as the home location (the user spends 7 months out of 13 at location A) and detecting the 6-month period at location B as a temporary migration event. Panel (**b**) assumes the last two months are unobserved, resulting in an 11-month dataset where location B appears to dominate, leading to its misidentification as the home location and a failure to identify the 6-month period at location B as a temporary migration event. This graphic illustration of a CDR-derived trajectory is inspired from Fig. 1 in Chi *et al.*[21].

detect a 6-month temporary migration event at location B. Figure 2 (panel (**b**)) considers a case where we observe the user for the first 11 months only, with the last two months being unobserved. Over this period of observation, the user spends the majority of months at location B that then defines his home location, so that the 6-month period at location B cannot be identified as a temporary migration event. Given that we consider users with a minimum length of observation of approximately a year, we set $\tau_{meso}^{max} = 180\ days$. However, parameters τ_{meso}^{min} and τ_{meso}^{max} could be flexibly adjusted in future applications based on specific research needs and data constraints.

Step 1: hourly, daily, monthly locations. First, some useful notations and definitions are in order. The studied area is partitioned into contiguous, non-overlapping spatial units that define the full set of potential locations where users can be observed, denoted by $\mathscr{L} = (\ell_k)_{k\in[1;L]}$, with L the total number of locations. In the present case, $\mathscr{L}$ is the set of voronoi cells introduced above so that $L = 916$. The raw CDR trajectory of a user i is denoted by $(x_{t_1}^i, x_{t_2}^i, \ldots, x_{t_{T_i}}^i)$, where each $x_t^i \in \mathscr{L}$ represents i's observed location at timestamp t. T_i is i's total number of CDR.

Consistent with conventional methodologies outlined in previous studies[17–20,26], a hierarchical frequency-based method is implemented to determine hourly, daily, and monthly locations. For a user i, the hourly location $x_{h,d}^i$ for an hour h of day d is defined as the most frequently visited location during that one-hour time interval, denoted h_d:

$$x_{h,d}^i = mode\{\, x_t^i \mid t \in (t_1, \ldots, t_{T_i}),\ \ t \in h_d\} \tag{1}$$

Hourly locations are then used to estimate a location for each day d, i.e. a daily location x_d^i. The daily location x_d^i is defined as the most frequent hourly location during nighttime (6pm-8am) when available, and the most frequent hourly location during daytime otherwise. As is customary in the literature, night hours are preferred in order to mitigate the influence of daytime location shifts (e.g. commuting) and maximize the likelihood that the inferred location effectively coincides with the location where the corresponding user spends the night[17,19,27]. That said, we also use daytime locations when a nighttime location is not observed in order to limit the loss of information induced by this filtering procedure. More formally, the set of night hours for day d is denoted by $\mathscr{N}_d = \{(h, d)|(h, d) \in \{(18, d), \ldots, (23, d)\} \cup \{(0, d+1), \ldots, (7, d+1)\}\}$ and the set of daytime hours is $\overline{\mathscr{N}_d} = \{(h, d) \mid (h, d) \in \{(8, d), \ldots, (17, d)\}\}$. With these notations, nighttime and daytime locations for day d can be defined as:

$$\begin{cases} x_{d,nighttime}^i = mode\ \{x_{h,d}^i \mid (h, d) \in \mathscr{N}_d\} \\ x_{d,daytime}^i = mode\ \{x_{h,d}^i \mid (h, d) \in \overline{\mathscr{N}_d}\} \end{cases} \tag{2}$$

As a result, the estimated daily location x_d^i can be written as:

$$x_d^i = \begin{cases} x_{d,nighttime}^i, & \text{if } \{x_{h,d}^i \mid (h, d) \in \mathscr{N}_d\} \neq \varnothing \\ x_{d,daytime}^i, & \text{otherwise} \end{cases} \tag{3}$$

Of course, x_d^i is undefined for days when neither nighttime nor daytime observations are available. For any user i, let $\mathscr{D}_i = \{d_1^i, \ldots, d_{D_i}^i\}$ represent the set of D_i days for which a daily location is determined.

Finally, monthly locations are calculated as the modal daily location over a month, with a minimum of 10 days observed imposed in order to guarantee some degree of confidence in the estimated monthly location.

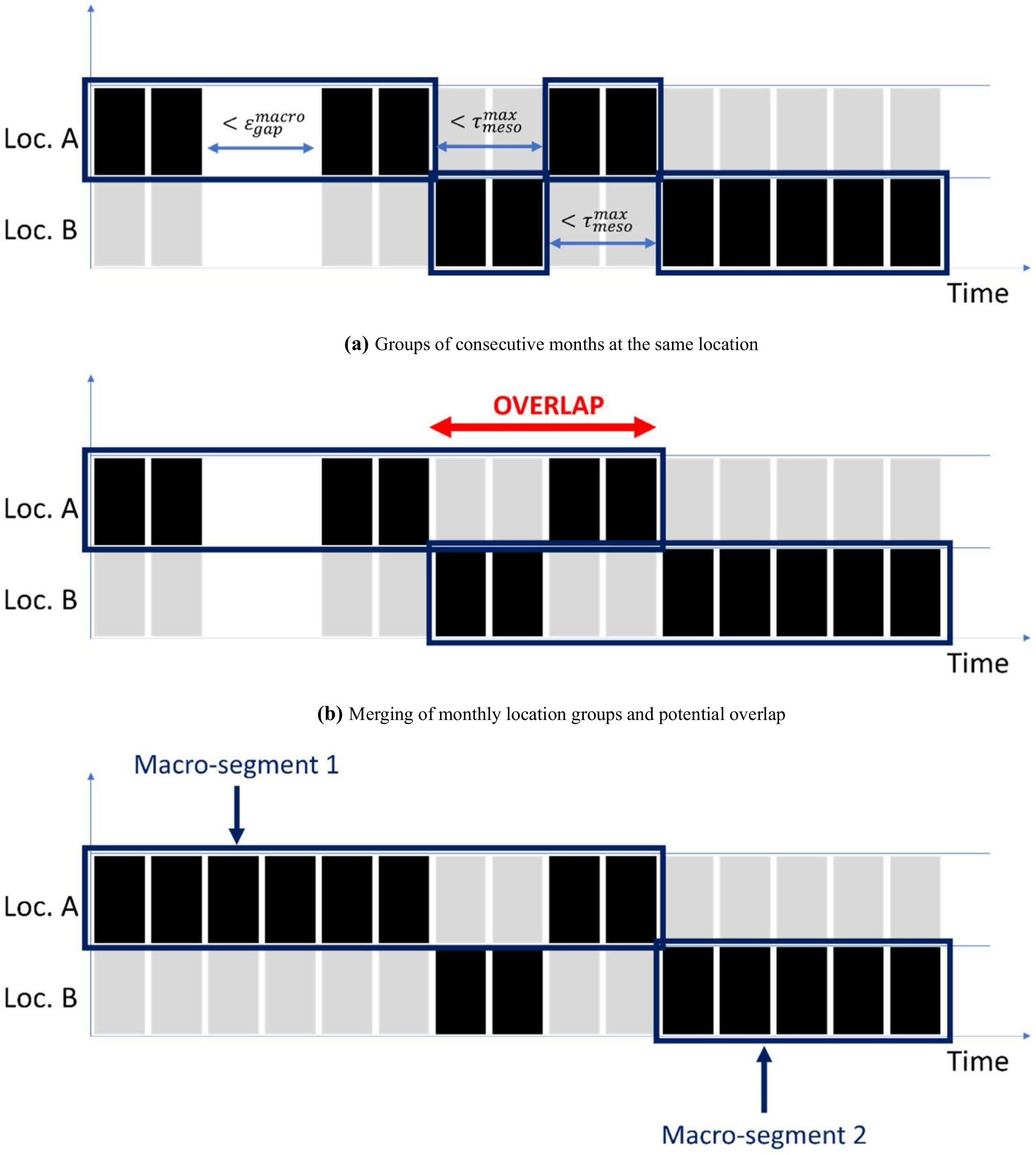
Step 2: Macro-segment detection. Step 2 focuses on the identification of macro-segments, defined as periods of at least τ_{meso}^{max} during which a user remains consistently present at a single location, while permitting short-term movements (i.e. micro-segments) and temporary migration (i.e. meso-segments) at other locations. The macro-segment detection algorithm uses a clustering procedure on monthly locations, as the frequency-based approach outlined above serves as a simple method to smooth out micro-segments from a raw CDR trajectory. Then, the clustering technique follows the main principles outlined in Chi *et al.*'s[21] methodology and proceeds in four steps:

(i) Preliminary unique home location estimation: A default unique home location $\overline{home_i}$ is estimated for each user i. It corresponds to the most frequently observed daily location over i's period of observation:

$$\overline{home_i} = mode\ \{x_d^i \mid d \in \mathscr{D}_i\} \tag{4}$$

The following three steps are designed to identify potential macro-movements, defined as consecutive macro-segments at distinct locations. In most cases, where this process does not yield multiple macro-segments, the user is considered to have a single macro-segment at location $\overline{home_i}$ spanning the entire observation period, defining his unique home location for that period.

(ii) Detect contiguous monthly locations: Consecutive months at the same location are grouped together, allowing for observation gaps of at most $\varepsilon_{gap}^{macro}$ months. $\varepsilon_{gap}^{macro}$ is set such that no movement at the macro-scale (i.e. a permanent migration) could occur during unobserved periods. $\varepsilon_{gap}^{macro}$ is thus set to 6 months



(a) Groups of consecutive months at the same location

(b) Merging of monthly location groups and potential overlap

(c) Overlap resolved and final macro-segments

Fig. 3 Illustration of the macro-segment detection procedure (step 2). Black bars represent monthly locations for a hypothetical user across two locations A and B. Panel (**a**) shows the grouping of consecutive months at the same location. Panel (**b**) illustrates the merging of monthly location groups when they are less than τ_{meso}^{max} months apart, and how this process can generate overlap between the resulting groups. Panel (**c**) shows the final macro-segments detected after those overlaps are resolved. This graphic illustration of a CDR-derived trajectory is inspired from Fig. 1 in Chi *et al.*[21].

so that it coincides with the minimum duration defining a macro-segment ($\tau_{meso}^{max} = 180$ days). Note that, in practice, observation gaps are much shorter given the constraint imposed on the maximum period of non-observation. This process is illustrated in Fig. 3 (panel (**a**)) where a hypothetical user is observed across two locations, A and B. Groups of consecutive months at the same location are illustrated with dark frames.

(iii) Merge monthly location groups: Groups of months at a single location are then merged when they are separated by one or more groups accounting for a total duration strictly less than τ_{meso}^{max}. This process essentially groups home stays that may be interspersed with temporary migration spells. As explained in Chi *et al.*[21], this clustering approach occasionally generates overlapping groups of monthly locations. This is illustrated in Fig. 3 (panel (**b**)).

(iv) Resolve overlap: Next, the overlap between merged groups that may result from the previous step is resolved. First, merged groups with a duration strictly lower than τ_{meso}^{max} are removed: as per the definition adopted, they cannot be macro-segments. For two consecutive overlapping groups, overlapping months are assigned to the longest group, as showed in the illustrative example provided in Fig. 3 (panel (**b**)). Start and end dates of merged groups are updated accordingly and merged groups which now have a duration strictly lower than τ_{meso}^{max} are removed. To address rare cases of multiple overlaps, this procedure is iterated until no overlapping groups are left. For each user, the final merged groups form his set of detected macro-segments. Given relatively low rates of permanent migration and limitations due to the length of observation relative to τ_{meso}^{max}, the vast majority of users end up with only one macro-segment detected. Those users are assigned the default unique home location $\overline{home_i}$ determined in the preliminary step, which defines a unique macro-segment for the entire period of observation.

Step 3: Meso-segment detection. A comparable approach is used to detect meso-segments, with an illustrative example provided in Fig. 4 to aid the reader's understanding. The procedure can be decomposed in three steps:

(i) Detect contiguous daily locations: Consecutive days at a single location are grouped together, allowing for observation gaps of at most ε_{gap}^{meso}. While small values of ε_{gap}^{meso} may fail to smooth out short-mobility events, larger values are associated with significant overlap between groups of days detected. We rely on Chi *et al.*[21] to determine a reasonable value for ε_{gap}^{meso} and we set it to the optimal value of 7 days they infer from a cross-validation exercise. In the illustrative example shown in Fig. 4 (panel (**a**)), this process results in three groups of observations at location A and two at location B. Note that an observation gap can be observed within the first group. Since this gap is shorter than ε_{gap}^{meso}, the consecutive days at location A before and after the gap are combined into a single group of daily locations.

(ii) Merge daily location groups: Those groups of consecutive days at the same location are merged when they correspond to the same location and are less than ε_{gap}^{meso} days apart. For each user, this results in a set of intermediary meso-segments. In the example provided in Fig. 4 (panel (**a**)), the first two groups at location A are less than ε_{gap}^{meso} and are merged into a single intermediary meso-segment (Fig. 4, panel (**b**)). The same applies to the first two groups at location B. However, the second and third groups at location A are separated by a duration exceeding ε_{gap}^{meso}, so they remain distinct segments.
Similar to Chi *et al.*[21], we filter out meso-segments with a proportion of days at the identified location lower than some parameter ϕ, that we set to 0.5. This helps to limit cases where a meso-segment might capture frequent movements between multiple locations rather than a temporary migration event at a single location.

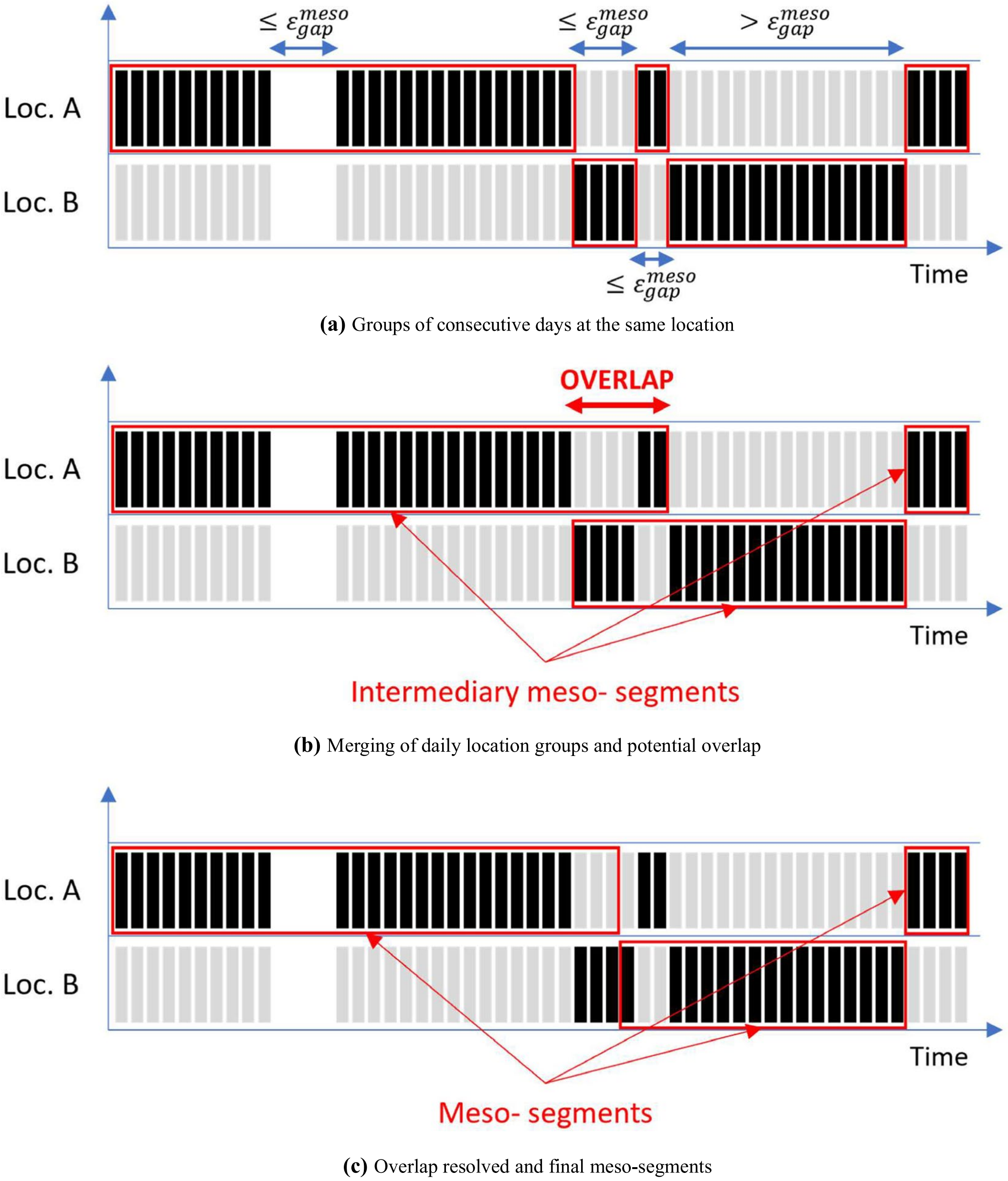
(iii) Resolve overlap: As in the macro-segment detection procedure, merging groups of days at a single location can lead to some overlap between intermediary meso-segments (see Fig. 4, panel (**b**)). The overlap between pairs of consecutive segments is resolved by taking the middle of the overlap as the end date of the first segment and the following day as the start date of the second one (Fig. 4, panel (**c**)). This process is iterated until no overlap is left.

Each meso-segment is associated with four main attributes: a meso-location, the macro-location associated with the period covered by the segment, a minimum duration and a maximum duration. The meso-location is a direct output of the meso-segment detection procedure. The macro-location associated with the meso-segment is straightforward for users with a unique home location, which constitutes the majority of cases. For other users with multiple macro-segments across the period of observation, if a meso-segment is entirely covered by a macro-segment, it is assigned the corresponding macro-location. If a meso-segment overlaps between two macro-segments, it is assigned the macro-location of the macro-segment with the largest overlap. Finally, due to potential observation gaps before and after a meso-segment, accurately determining its exact duration is not always feasible. For that reason, for any segment S_i of a user i, we define both a lower-bound duration $minDuration(S_i)$ – referred to as the "observed duration" – and an upper-bound duration $maxDuration(S_i)$ – referred to as the "maximum duration". As illustrated in Fig. 5, $minDuration(S_i)$ is calculated as the time elapsed between the segment's identified start and end dates, while $maxDuration(S_i)$ represents the time between the last observed day before the segment and the first observed day after S_i.

Step 4: Identification of migration events. Temporary migration events are identified as meso-segments with a duration of at least τ_{meso}^{min}, occurring at a destination that differs from the macro-location – which defines the home location at the time of the mobility event. For instance, in the illustration provided in Fig. 6, a hypothetical user has a unique home location (location A), and three meso-segments are detected (highlighted in red frames). They all have an observed duration of at least 20 days. Among them, only the second meso-segment exhibits a meso-location (location B) distinct from the macro-location (location A) and is therefore identified as a temporary migration event.

From user-level migration history to migration statistics.

Weighting scheme. In conventional surveys, statistics on a target population are derived from a sample of individuals. The extrapolation from the sample to the population level is permitted by a meticulously defined sampling process. Individuals are selected from a sampling frame, which represents the target population, using a well-defined sampling design. However, mobile phone data simply provide a selected subset of the population, which composition is not governed by a similar sampling procedure. Because phone ownership and usage patterns vary among different demographic groups[20,24,28], directly inferring population-level statistics from a sample of mobile phone data is inherently subject to sampling biases. The size of these biases ultimately depend on the magnitude of migration behavior


$\leq \varepsilon_{gap}^{meso}$
$\leq \varepsilon_{gap}^{meso}$
$> \varepsilon_{gap}^{meso}$
Loc. A
Loc. B
$\leq \varepsilon_{gap}^{meso}$
Time
OVERLAP
Loc. A
Loc. B
Time
Intermediary meso- segments
Loc. A
Loc. B
Time
Meso- segments

(a) Groups of consecutive days at the same location

(b) Merging of daily location groups and potential overlap

(c) Overlap resolved and final meso-segments

Fig. 4 Illustration of the meso-segment detection procedure (step 3). Black bars represent daily locations for a hypothetical user who is seen across two locations, A and B. Panel (**a**) shows the identification of groups of consecutive days at a single location. Panel (**b**) illustrates how these groups are merged when they are at most ε_{gap}^{meso} days apart. This process generates an overlap between the first two intermediary meso-segments. Panel (**c**) represents the final meso-segments detected after the overlap is resolved. This CDR trajectory representation is inspired from Fig. 1 in Chi *et al.*[21].

differentials between phone users and non-users, combined with the prominence of non-users within the target population. Moreover, since a statistical bias represents the difference between a sample-based statistic and the true value in a target population, its magnitude is contingent on how this target population is actually defined.

With this in mind, the selection issue in the production of phone-based migration statistics can be addressed in mainly two ways. First, the target population can be simply restricted to a minimal subset that the data effectively represent. Second, some degree of extrapolation to a larger target population can be achieved by using observable characteristics of users to implement correction methods. Both approaches are considered in two distinct sets of migration estimates.

The first one is comprised of statistics directly derived from a given subset (i.e., *subset A* or *subset B*). They are referred to as the *unweighted* estimates. Operating under the minimal assumption that the migration outcomes of users in the subset are comparable to those of the overall population of phone users, the sample of users is considered as representative of that population. Evidence supporting this assumption is provided in the Technical Validation section. As a result, the target population associated with the *unweighted* estimates is

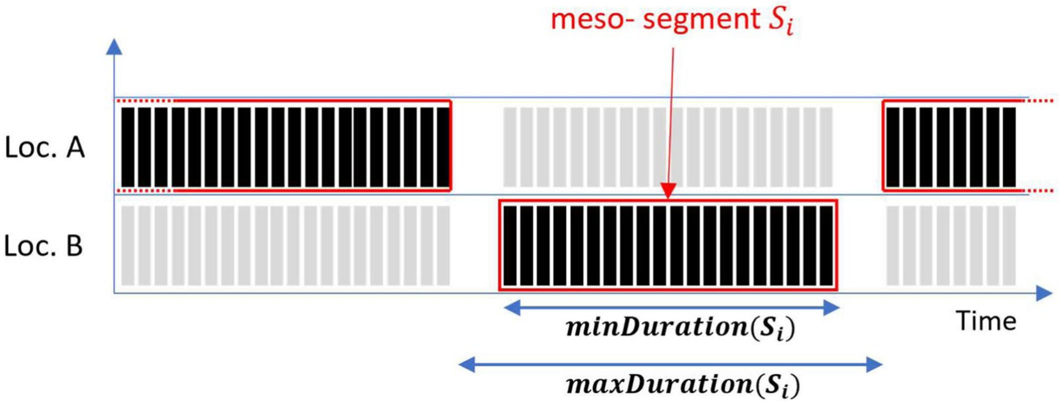


Fig. 5 Observed duration ($minDuration(S_i)$) and maximum duration ($maxDuration(S_i)$) of a meso-segment S_i.

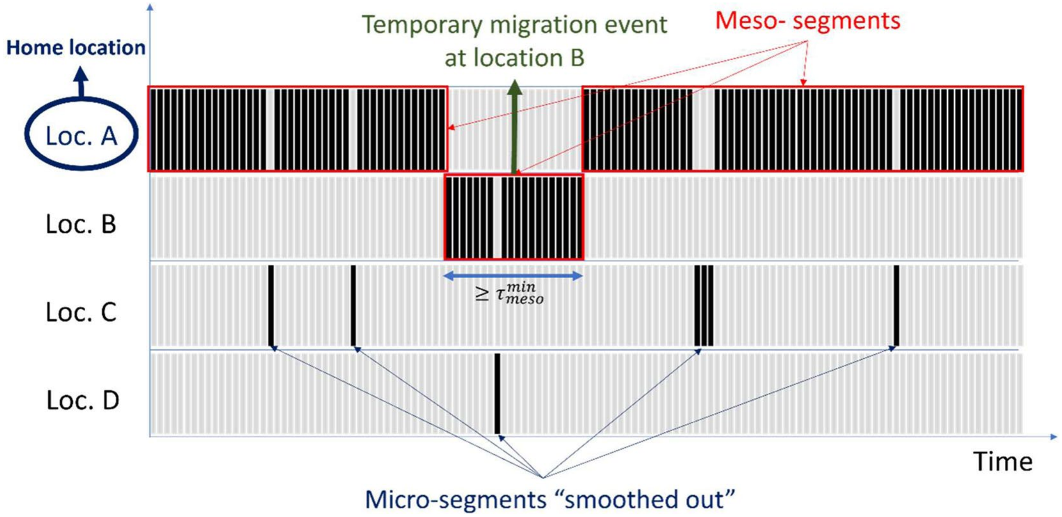


Fig. 6 Illustration of the identification of temporary migration events (step 4). Black bars represent daily locations for a hypothetical user who is seen across four locations: A, B, C and D. A well-behaved trajectory is assumed: the user has a unique home location (location A) and the meso-segment detection does not generate overlapping groups of daily locations. Red frames describe meso-segments detected with the clustering procedure on daily locations step 3. The second meso-segment is the only meso-segment detected at a non-home location (B) and with a duration of at least τ_{meso}^{min}. It is thus classified as a temporary migration segment. This CDR trajectory representation is inspired from Fig. 1 in Chi *et al.*[21].

confined to the subset of mobile phone users, which constitutes a sizable portion of the adult population (see Fig. 7). According to the 2014 *Listening to Senegal* survey[22], mobile phone users comprised 72% of the population over 18, thus constituting at least 37% of the entire population.

A second set of migration estimates, called the *weighted* estimates, is produced with a correction method that allows to consider a broader target population, extending to the entire adult population (i.e. individuals aged 15 and above). This choice is motivated by the fact that mobile phone ownership among individuals below 15 is indeed considered as negligible. This segment of the population is entirely absent from the mobile phone dataset, and their movements are unlikely to be captured. In fact, estimates of mobile phone ownership by age derived from the 2017 Demographic and Health Survey (DHS)[29] effectively reveal a notable decrease among individuals aged between 20 and 15, from over 75% to 23%. Moreover, it is assumed that local differences in migration outcomes between users in the sample and individuals in the target population are small, which we refer to as the *local representativeness* assumption. Specifically, we define 39 urban strata – coinciding with the 39 identified cities in Senegal – and 185 rural strata, which correspond to the rural areas of third-level administrative units further segmented into areas with low population density (i.e., below the rural median) and high population density (i.e., above the rural median). A map showing these 224 strata is provided in Fig. 8. Differences in migration outcomes between users and the target population are then presumed limited within each individual stratum. In the Technical Validation section, we provide evidence supporting the notion that differences at a local level between phone users and the target population are effectively limited. Then, a correction method is implemented to neutralize imbalances in a key characteristic that is easily observable in CDR data: their home

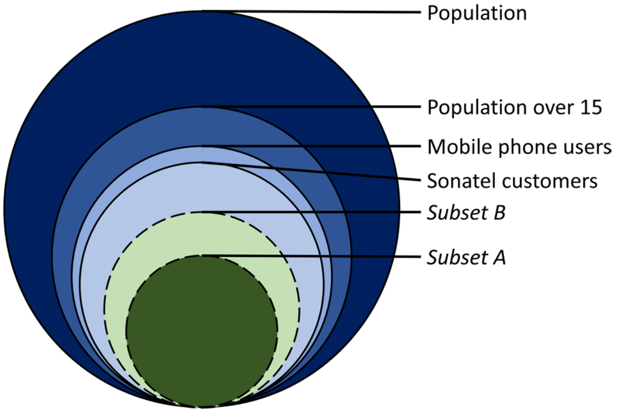


Fig. 7 Scaled Venn diagram illustrating the relative sizes of various subsets within the Senegalese population, including *Subset A* and *Subset B*. The diagram reflects the following hierarchy: $Subset\,A \subset Subset\,B \subset Sonatel\,Customers \subset Mobile\,Phone\,Users \subset Population\,Aged\,15\,and\,Above \subset Total\,Population$. The area of each disk is proportional to the size of the corresponding subset. *Subset A* and *Subset B* are represented in shades of green, while the broader population subsets are depicted in shades of blue.

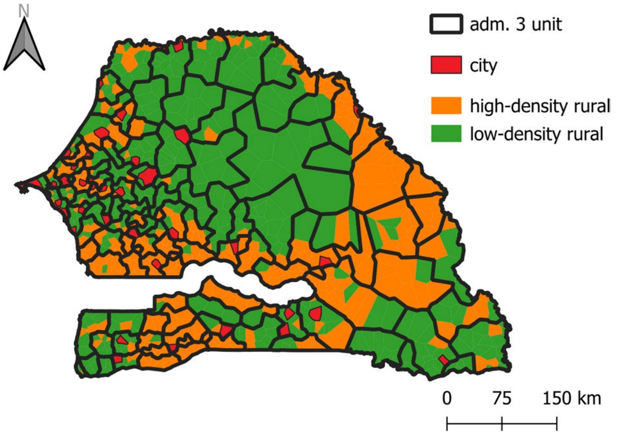


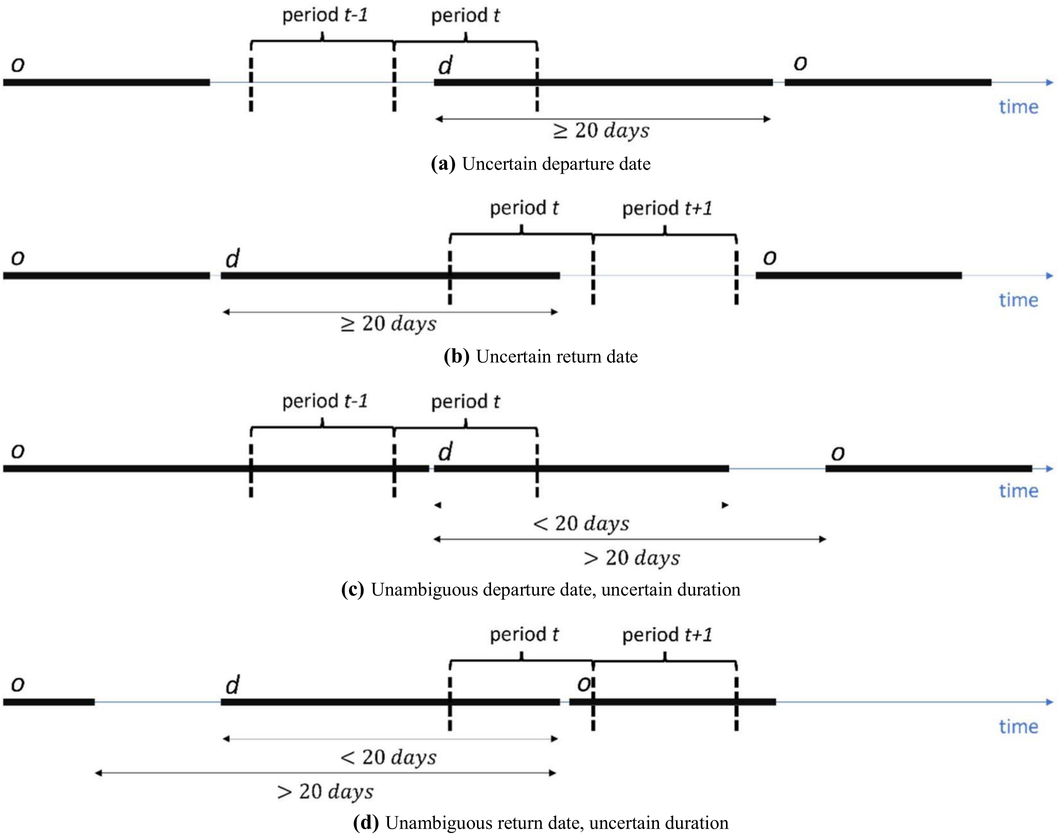
Fig. 8 Strata used to design the weighting scheme employed in the *weighted* migration estimates. Each of the 39 individual cities constitutes an urban stratum (in red). Black lines delineate groups of voronoi cells belonging to a single third-level administrative unit, where voronoi cells are assigned an administrative unit based on a maximum population criterion. The median population density across rural voronoi cells (1550 inh./km^2) is used to define low- and high-density rural areas, represented in green and orange respectively. Each (administrative unit,rural density category) couple constitutes a rural stratum. Note that voronoi-level population estimates are obtained by overlaying voronoi polygons with the 2017 100m-resolution gridded population product from the WorldPop Research Group[33].

location. Within each stratum, users are assigned a weight equal to the ratio of the stratum-level target population over the total number of observed users identified as residing in that stratum. Moreover, we allow for weights to vary over time to accommodate the fluctuating number of users actually observed across time units. For any location ℓ and time period t, the value of the weight $w_{\ell t}$ is then:

$$w_{\ell t} = \frac{pop_{\ell t}}{N_{\ell t}^{users}} \tag{5}$$

where $pop_{\ell t}$ is the size of the target population in location ℓ at time t and $N_{\ell t}^{users}$ is the total number of users residing in ℓ who are effectively observed during time period t.

Consequently, for any given time period, the sum of weights across users is equal to the total population aged 15 and above. In short, the weighting scheme corrects for disparities in the population-to-users ratio across strata, which are primarily caused by variations in mobile phone ownership and usage. For instance, urban areas are generally over-represented in the sample: in *subset A*, 80% of users live in cities whereas those only account for 54% of the population aged 15 and above. As a result, under the *local representativeness* assumption, the *weighted* migration estimates are unbiased estimators for the true migration outcomes of the population aged 15 and above.



(a) Uncertain departure date

(b) Uncertain return date

(c) Unambiguous departure date, uncertain duration

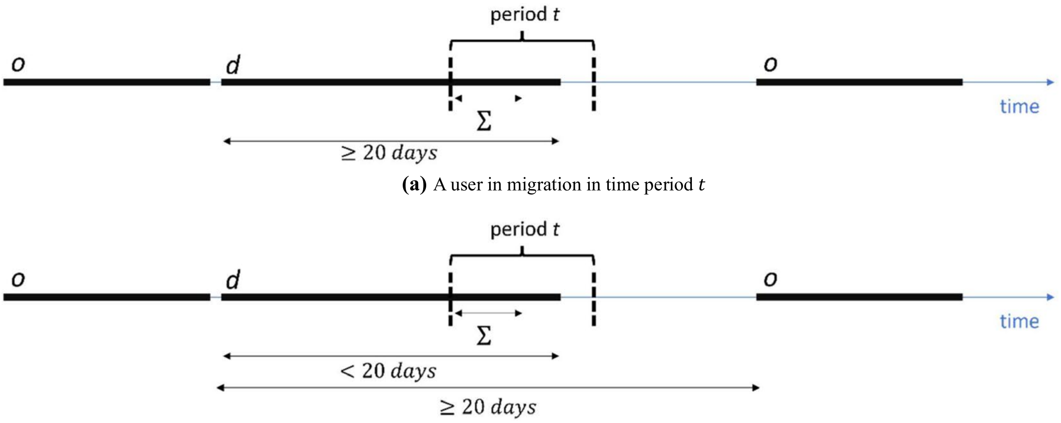
(d) Unambiguous return date, uncertain duration

Fig. 9 Uncertainty in the calculation of migration flows by time unit.

Note that the weighting scheme is designed as if the sample had been randomly drawn at the stratum-level, with the fraction of individuals selected from the target population varying across strata. However, in practice, other forces drive the underlying selection mechanism, and a limitation of the rectification method is its failure to account for these biases. For instance, the sample of users often disproportionately represents men, even within strata. Future research could enhance the proposed weighting scheme by incorporating socio-demographic information – either made available by the data provider or inferred from usage patterns[30,31] – in order to address these local sampling biases. Despite the acknowledged and documented issue of selection in CDR samples, the literature has generally overlooked concrete correction methods to address it for constructing representative mobility measures. Hence, we argue that the proposed weighting scheme and its underlying logic represent a significant improvement for producing near-representative migration statistics from a non-random sample of digital traces.

Regularizing unbalanced user-level trajectories. The migration detection model furnishes the location history of each individual user in the form of successive meso-segments. Migration statistics are derived by aggregating these heterogeneous trajectories at a specific spatio-temporal resolution. For any given time unit t and pair of locations o and d, it is possible to calculate migrations flows during t, i.e. the number of migration departures from o to d and returns from d back to o, as well as the migration stock, which corresponds to the number of users residing in o being in migration at destination d during t. These calculations involve various methodological challenges, which we address and resolve below.

Assigning a migration departure preceded (resp. return followed) by an observation gap to a specific time unit: First, we focus on the calculation of migration flows for any origin-destination pair and time unit. A user i residing in o is considered to have departed for migration to destination d at time t if he has a migration meso-segment at d that started during t. Similarly, user i returned from d to his home location o at time t if a migration segment at d ended during t. However, observational gaps imply some degree of uncertainty regarding the actual start and end dates of meso-segments, thereby complicating the computation of migration flows in practice. Illustrative examples are shown in Fig. 9 considering a minimum duration of $\tau_{meso}^{min} = 20\ days$ to define temporary migration events. In Fig. 9 (panel (**a**)), user i residing in o has a migration segment at destination d with an observed departure date within period t. However, i is unobserved in period $t-1$, rendering it uncertain as to the specific period when the migration departure actually occurred (i.e. t, $t-1$, or $t-2$). Similarly, in Fig. 9 (panel (**b**)), a migration segment ends within period t but the observational gap that follows raises the possibility for user i to have returned home in period $t+1$ or $t+2$. These ambiguities are partly resolved by introducing a tolerance parameter ϵ^{tol}. In situations analogous to that of Fig. 9 (panel (**a**)), ϵ^{tol} sets the



(a) A user in migration in time period t

(b) Ambiguous migration status due to uncertainty in segment duration

Fig. 10 Determining the migration status of a user in period t.

maximum acceptable time unobserved before the start of period t to still consider that user i departed for migration at d during period t. Likewise, for migration returns, ϵ^{tol} sets the maximum acceptable time unobserved after the period when the user is seen returning home in order to consider that the user effectively returned during that time period. The migration statistics disaggregated by half-month provided in the dataset are produced with ϵ^{tol} equal to 7 days.

Addressing uncertainty in segment duration for migration flow calculation: Then, uncertainty on the start and end dates of meso-segments naturally leads to some level of uncertainty on the actual duration of meso-segments, which gives raise to a second category of ambiguous cases. For example, in Fig. 9 (panel (**c**)), the start date of the segment at destination d unequivocally falls within period t, but the observed duration is lower than 20 days and the segment is not classified as a migration segment. Yet, the observational gap following the segment indicates that its actual duration may possibly be greater than 20 days, in which case user i should be regarded as having departed for migration at time t. Figure 9 (panel (**d**)) shows a similar situation where the return date is unambiguously assigned to period t but the uncertain duration induced by the observational gap preceding the segment complicates its classification as a migration segment. The migration estimates provided in the dataset are simply based on meso-segments with an observed duration (*minDuration*) greater than τ_{meso}^{min}, which are referred to as *high-confidence* estimates. "Lower-confidence" estimates of migration departures and returns were also produced considering meso-segments with an observed duration lower than τ_{meso}^{min} but a maximum duration (*maxDuration*) greater than τ_{meso}^{min}, similar to scenarios depicted in Fig. 9 (panel (**c**)–(**d**)). Many such configurations giving rise to ambiguous cases are possible, and an exhaustive set of algorithmic rules were implemented to address each one of them. In practice, due to the rather strict observational constraints imposed (as outlined in the Filtering procedure section), uncertainty on the actual duration of meso-segments remains minimal. As a result, these lower-confidence estimates show negligible difference with the primary high-confidence estimations, and the dataset therefore exclusively contains high-confidence estimates. Nonetheless, a comprehensive description of the algorithm's treatment of various configurations, along with illustrative diagrams, is furnished in the Supplementary Material. This empowers researchers to apply the methodology to alternative digital trace datasets not necessarily exhibiting comparably high sampling frequencies, while also aiding in the understanding of the code.

Determining a user's migration status (i.e. in migration or not) for any given time unit: We determine aggregation rules allowing to calculate migration stocks for each origin-destination pair and time unit. The migration stock from o to d during a time period t is calculated by aggregating the migration status of users, i.e. whether a user is in migration or not at time t. A user i is defined as being in migration at time t if that user exhibits a migration segment that overlaps time period t on at least Σ days (see Fig. 10, panel (**a**)). Migration stock estimates provided in the dataset are generated with a value of Σ equal to 8 days. With this value, we simply impose that the overlap represents at least half the time unit since half-months have a duration of at most 16 days.

Addressing uncertainty in segment duration for migration stock calculation: Determining the migration status of user i for a time period t can also be subject to some ambiguities, arising primarily from uncertainty in the duration of meso-segments. For instance, in Fig. 10 (panel (**b**)), user i may or may not be in migration at destination d in period t, depending on his actual location during the following observation gap. The possibility exists that the segment has an actual duration greater than 20 days, in which case i should be considered as being in migration at time t. High-confidence estimates of migration stocks are derived exclusively from meso-segments with an observed duration greater than τ_{meso}^{min}. Lower-confidence estimates were also calculated considering meso-segments with an observed duration below τ_{meso}^{min} but a maximum duration greater than τ_{meso}^{min}. Again, these two sets of estimates show little difference in practice and only high-confidence estimates are included in the dataset, but a comprehensive description of algorithmic rules along with illustrative diagrams are provided in the Supplementary Material.

Calculating the total number of users observed for each time unit (used to calculate migration rates): Finally, estimating time-disaggregated migration rates requires some measures of the actual number of users observed at any given time period *t*, serving as the denominator for such rates. Similar to the computation of migration flows and stock, the presence of observational gaps and attrition introduces variations in the number of users observed over time in each location. In essence, a user *i* is classified as "observed at time *t*" for a specific migration measure (i.e. departures, returns, or stock) if their trajectory allows to unambiguously determine his migration status at time *t* for that migration measure – e.g. *i* departed for migration during *t* or did not depart for migration during *t*. For example, in the scenario depicted in Fig. 9 (panel (**a**)) and assuming that the tolerance parameter is exceeded, the user would be deemed unobserved for the calculation of the departure rates at time period *t*. Again, all possible configurations in the trajectory of users are analyzed and all cases where users are considered as unobserved for measures of migration departures, returns and stock, are identified. Details on each of those cases along with illustrative diagrams are provided in the Supplementary Material, facilitating the understanding of the rules implemented as well as the corresponding code. It is important to highlight that the conditions defining the observational status of a user for a time period *t* depend on the migration measure considered, as well as the minimum migration duration threshold τ_{meso}^{min}, the tolerance parameter ϵ^{tol} and the parameter Σ. Additionally, these numbers are also employed in the calculation of weights used to produce the *weighted* estimates.

Data Records

The dataset is available for download at a public *figshare* repository[32]. The complete dataset is broken down into multiple files as described above, with each file associated with a type of estimates (*weighted* or *unweighted*), a specific subset (A or B), and a minimum migration event duration threshold.

Migration estimates are provided at the (origin*destination*time)-level. The origin and destination locations considered are comprised of 39 cities and 112 rural areas of third-level administrative units (i.e., districts), defining the spatial resolution of the dataset. Note that all 39 cities are considered as individual spatial units. Therefore, the dataset contains city-level migration estimates rather than estimations at the level of urban areas for each district. Time units coincide with "half-months", defined as the periods going from the 1st to the 15th, and from the 16th to the end of each month. Each year is thus comprised of 24 half-months and the dataset covers the period 2013–2015.

The full dataset is organized in 12 datasets. Each dataset provides either *weighted* or *unweighted* estimates and is derived from either *subset A* or *subset B*. In addition, for each of these four combinations, separate datasets provide estimates considering only migration events with a duration of at least 20, 30, or 60 days. A standard file name `type_X_DDdays.csv.gz` is used to uniquely identify each dataset, with `type` being either `weighted` or `unweighted`, `X` ∈ {`A`,`B`} denotes the subset from which migration estimates are derived, and `DD` is either `20`, `30`, or `60`. For instance, the file `weighted_A_20days.csv.gz` contains weighted estimates derived from *subset A* considering temporary migration events of at least 20 days.

Each dataset contains time series of migration departures, migration returns, and migration stock, both in absolute terms and as a fraction of the total number of users observed at origin. These metrics are provided for each origin-destination pair over the period from 2013 to 2015, with time units defined as half-month intervals. Note that half-month periods that coincide with the start and end of each CDR dataset (i.e., the 2013 and 2014-2015 datasets) are excluded from the final estimates due to increased uncertainty at the boundaries of the observation periods. For example, considering migration events of at least 20 days, all users seen at a non-home location from January 1, 2013 to at least January 8, 2013 and returning to their home location before January 20, 2013 are not classified as migrants during the first half of January. However, if they had departed to the non-home location before December 31, 2012, they should indeed be identified as migrants. The extent of these high-uncertainty periods at the edges of CDR datasets depends on the minimum duration used to define temporary migration events. Therefore, for migration estimates based on events of at least 20 days, we only exclude estimates for the first and last half-months of 2013, the first half-month of 2014 and the last half-month of 2015. For events of at least 30 days, estimates for the first and last two half-months of 2013, the first two half-months of 2014 and the last two half-months of 2015 are excluded. For events of at least 60 days, we exclude estimates for the first and last four half-months of 2013, the first four half-months of 2014, and the last two half-months of 2015. All variable names along with detailed descriptions are provided in Table 3.

Weighted estimates incorporate population dynamics during the 2013–2015 period, meaning that target population values at the voronoi level (i.e., individuals aged 15 and above) vary over time. The calculation of target population estimates for each voronoi cell and half-month period involves four steps. First, a baseline population distribution is established by overlaying voronoi polygons with the 100m-resolution gridded population product from the WorldPop Research Group[33]. Second, population values for the midpoint of the study period (i.e. the first half of June 2014) are obtained by applying this distribution to the 2014 total population. Third, department-level estimates, broken down by rural and urban zones, of the fraction of the population aged 15 and above from the 2013 census[34] are applied to calculate voronoi-level estimates of the target population for the first half of June 2014. Fourth, using annual population growth projections by region and zone for 2013–2015[34], half-monthly population growth rates are determined and applied to estimate the target population for each voronoi cell and half-month period.

A shapefile delineating the boundaries of all spatial units used in the dataset (i.e., origin and destination locations) is provided in the *figshare* repository. This allows users to map temporary migration estimates and combine the dataset with other spatial data, such as climate or land use information. The spatial units are constructed from the voronoi cells showed in Fig. 1 (panel (**b**)). Cells classified as urban form individual spatial units. On the other hand, rural cells are assigned to a unique district based on a maximum population criterion and cells that belong to a unique district are grouped. This process results in 112 rural locations and 39 urban locations (i.e. cities) for a total of 151 distinct locations. The shapefile is provided as a GeoPackage file (`spatial_units_SciData.gpkg`) with a single layer. The attribute table has three columns: i) *id* is a unique identifier for each

Variable name	Description
N_depart	Number of migration departures
N_return	Number of migration returns
N_migrants	Number of individuals in migration, i.e. the stock of migrants
N_users_observed_depart	Total number of users residing in origin *o* observed at time *t*, for departure counts
N_users_observed_return	Total number of users residing in origin *o* observed at time *t*, for return counts
N_users_observed_stock	Total number of users residing in origin *o* observed at time *t*, for migration stock
rate_depart	Rate of departures calculated as $\frac{N\_depart}{N\_users\_observed\_depart}$
rate_return	Rate of migration returns calculated as $\frac{N\_return}{N\_users\_observed\_return}$
rate_migrants	Migration stock rate calculated as $\frac{N\_migrants}{N\_users\_observed\_stock}$

Table 3. Description of variables provided in the dataset. Each row of the dataset provides measures of temporary migration from an origin *o*, to a destination *d*, for a half-month denoted by *t*. In datasets providing *weighted* estimates, variable names include the suffix "*_adj*" to indicate that variables were adjusted with the weighting scheme outlined in the Methods section (e.g. *N_depart_adj* instead of *N_depart*). Note that, by construction, the "adjusted" number of users observed provided in weighted estimates datasets coincides with the target population for the corresponding origin location.

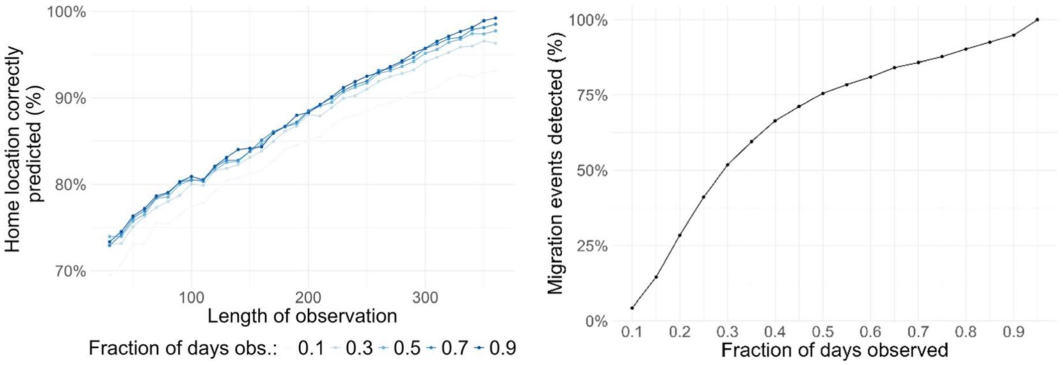
spatial unit, ii) *name* gives the name of the corresponding spatial unit, which is either the city name or a name of the form *NAME-rural* where *NAME* is the name of the corresponding district, and iii) *zone_category* indicates whether the spatial unit is classified as rural or urban.

Technical Validation

Migration event detection accuracy. Observational requirements for the measure of human mobility necessarily depend on the type of movements one aims to capture. Broadly speaking, measuring long-term changes in the place of residence requires extended periods of observation (e.g. several years) with modest sampling frequencies, while capturing commuting movements asks for high sampling frequencies (e.g. multiple observations per day) over potentially shorter observation periods. Minimal sampling characteristics for the measure of temporary migration movements is qualitatively somewhere in between. The proposed migration detection algorithm essentially requires that users are seen often enough during a sufficiently long period of time in order to be able to (i) confidently identify a home location and (ii) detect the temporary changes in the usual location observed – that we have called "meso-movements". We investigate this issue in quantitative terms by conducting a sensitivity analysis of the proposed migration detection algorithm with respect to users' observational characteristics. More specifically, we evaluate the impact of the length of time a user is observed and the fraction of days with observations (i.e. the frequency of observation) on the level of accuracy associated with both the prediction of home locations and the detection of temporary migration events. This contributes to the validation of the choice of filtering parameters considered to define *subset A* and *subset B*.

To achieve this, we consider a benchmark subset of users selected from the entire 2013 dataset, who meet stringent observational constraints: they are observed for at least 360 days and on at least 95% of days (195,070 such users satisfy those constraints). To ensure computational feasibility, a random subset of 10,000 users is selected among those with a unique home location identified and at least one migration event of at least 20 days detected. The strict observational constraints imposed to select this subset ensure that the migration detection outputs accurately represent (i) the users' true home locations and (ii) their actual temporary migration movements. Subsequently, we impose increasingly stringent observational constraints on these users and evaluate how the accuracy of the migration detection procedure declines as a result. Specifically, we consider values for the duration of observation varying between 30 and 360 days, and fractions of days observed ranging from 0.1 to 0.95. For each pair of values, we select a subset of observations for each user that meets the corresponding constraints. For instance, consider a user observed for 360 days on 95% of days and observational constraints set to 360 days and 0.8 respectively, the subset is formed by randomly removing 15% of observations from the initial trajectory. We re-apply the detection algorithm to these subsets of observations and compare the outputs with those obtained with the full trajectories to assess the model's accuracy under the specified observational constraints. This process is repeated for various pairs of values within the range specified above, which allows to appreciate the overall sensitivity of the detection model to observational characteristics.

First, we evaluate the impact of the length of observation Δ and frequency of observation Ω (henceforth also referred to as the "density" of the trajectory) on the accuracy of home location predictions. For each set of parameters (Δ, Ω), the model accuracy is simply defined as the fraction of users with a correctly predicted home location. Figure 11 (panel (**a**)) shows estimates of the model accuracy for lengths of observation ranging between 30 and 360 days and for different values of Ω. It is clear that the density of trajectories Ω has little incidence on the accuracy of home location predictions. For instance, even with only 10% of days observed, the level of accuracy continues to exceed 90% for lengths of observation of at least 290 days. More generally, for any given length of observation, the level of accuracy only varies by a few percentage points with values of Ω ranging from 0.1 to 0.9. On the other hand, accuracy seems to increase linearly with the length of observation. For $\Omega = 0.9$, it increases from 73% to 99% when the length of observation imposed increases from 30 to 360 days.



(a) Home estimation accuracy **(b)** Migration detection accuracy

Fig. 11 Model accuracy for home location predictions as a function of the length and frequency of observation (panel (**a**)) and accuracy of the migration detection algorithm as a function of the fraction of days observed (panel (**b**)).

Second, we focus on the impact of the frequency of observation Ω on the accuracy of the migration detection model, holding Δ fixed. Note that looking at the impact of Δ on the ability of the algorithm to detect migration events is not particularly relevant. Shorter lengths of observation simply imply missing migration events occurring during the period unobserved. Here, the model accuracy for any given value of Ω is defined as the fraction of real migration segments (i.e. those detected in the benchmark subset) that are effectively identified in subsets of CDR of density Ω. Removing observations from a full trajectory can lead to migration events being still detected although with slightly different start and end dates. Therefore, a real migration segment is considered as identified in a subset of density Ω if a migration segment overlapping at least half the real migration segment is detected in this subset of CDR. Results for Δ set to 360 days and Ω varying from 0.1 to 0.95 are provided in Fig. 11 (panel (**b**)). Unsurprisingly, the frequency of observation has a significant impact on the accuracy of the migration detection model. From 95% of migration events detected with a density of 0.9, it decreases to as low as 4% when the fraction of days observed is equal to 0.1. The convex shape of the relationship indicates that the level of accuracy starts to deteriorates sharply when Ω falls below approximately 0.5, and drops below 50% for values of Ω that are less than 0.3. On the other hand, densities greater than 0.8 allow to sustain a high level of accuracy beyond 90%.

The accuracy of both home location estimation and migration event detection is then assessed for the observational constraints associated with *subset A* and *subset B* respectively. Results are summarized in Table 4. Home detection accuracy is high for both subsets: 92% for *subset A* and 98% for *subset B*. We estimate that at least 90% of migration events are effectively detected by the algorithm in *subset A*. As expected, this figure is lower for *subset B* but still relatively high at 76%.

Validity of unweighted estimates: Are Sonatel users representative of the population of phone users?

Under the assumption that users in *subset A* and *subset B* have temporary migration outcomes that are comparable to those of the overall population of mobile phone users, *unweighted* estimates are considered as representative of that population.

Given the absence of CDR data for the entire population of mobile phone users, conducting a direct comparison of migration outcomes between our subsets of users and the broader population of mobile phone users is unfeasible. Likewise, without personal information available in CDR data, we cannot compare the characteristics of users in *subset A* and *subset B* with those of the population phone users. However, leveraging ICT Access Surveys[35] data, we perform statistical tests to at least evaluate whether Sonatel users differ from the population of phone users across various observable characteristics, including gender, age, education, zone of residence and wealth. Results are showed in Table 5. Overall, the findings suggest that Sonatel users are generally comparable to the broader population of phone users, particularly in terms of characteristics potentially associated with temporary migration determinants (e.g. assets ownership, wealth, gender). One notable difference is that Sonatel users tend to be slightly more urban than users of other operators (57.4% against 50.5%). Nevertheless, the existence of potential differences between Sonatel users and users from other operators remains a minor concern in the context of Senegal. Indeed, as illustrated in Fig. 7, the market was largely dominated by Sonatel during the study period: over 88% of mobile phone users have a Sonatel SIM card and 77% report Sonatel as their main provider[22]. However, this issue may assume greater significance in settings where the telephony market is more fragmented.

Validity of weighted estimates: local representativeness assumption and selection on home locations.

The *local representativeness* assumption crucially underpins the production of the *weighted* estimates, allowing to extend the target population to the population aged 15 and above. This assumption posits that differences in temporary migration outcomes between users in the sample and the target population remain limited. Although this cannot be directly verified in practice due to the absence of data on temporary movements,

	subset A	*subset B*
Home detection accuracy	98%	92%
Migration event detection accuracy	90%	76%

Table 4. Estimated accuracy of home location estimation and temporary migration event detection in *subset A* and *subset B*.

	Sonatel users	**All users**	**Diff.**
	(1)	**(2)**	**(1)–(2)**
Male dummy	0.559	0.554	0.005
Age	37.237	36.975	0.262
Years of education	6.785	6.101	0.685***
Urban dummy	0.574	0.505	0.07***
Has electricity	0.910	0.897	0.013
Has piped water	0.869	0.847	0.022
Has a fridge	0.447	0.415	0.033**
Has a radio	0.710	0.726	− 0.016
Has a TV	0.761	0.743	0.018
Richest quintile dummy	0.170	0.170	0
Poorest quintile dummy	0.178	0.196	− 0.019

Table 5. Comparison of Sonatel users with the overall population of phone users. Statistics were derived from the individual-level Access Survey dataset for Senegal conducted by Research ICT Africa[35].

we adopt a second-best approach and conduct two distinct exercises that help evaluate the validity of the *local representativeness* assumption. First, we use secondary survey data to compare mobile phone users with the adult population at large at a local level along a number of observable characteristics. Second, we compare the specific phone users in the selected subsets with the overall population with respect to the only characteristic that is readily observable with CDR data, i.e. the home location.

In the first validation exercise, we leverage data from the 2017 Senegal Demographic and Health Survey (DHS)[29], which focuses on individuals aged 15 and above. The survey provides information on individuals' mobile phone ownership along with various characteristics such as wealth, occupation, financial inclusion, and education level. To assess the *local representativeness* assumption, we employ statistical tests (t-tests) to examine differences between phone users and the overall population across these dimensions. These tests are conducted separately for each zone (rural/urban) within the 14 regions of Senegal to ensure consistency with the representativeness level of the DHS. Results are illustrated for the region of Kaolack in Table 6 while results for the 13 other regions are left in the Supplementary Material. The most notable observation is that very few coefficients exhibit statistical significance, a pattern that we consistently observe across all regions and zones. In particular, phone users are generally found to be slightly wealthier compared to the overall population, but those differences are never significant, even at a 10% level. Similarly, across specific regions and zones, phone users generally tend to have better access to amenities like drinking water, sanitation, and electricity, higher levels of education, lower unemployment rates, and greater participation in the agricultural sector. Again, these disparities are minor and statistically insignificant. These results tend to confirm that, at a local level, phone users and the broader adult population are statistically indistinguishable along numerous key dimensions. Assuming some degree of correlation between those characteristics and mobility choices, this supports the notion that, at a local level, temporary migration outcomes of phone users do not significantly differ from those of the adult population as a whole. Two exceptions are worth highlighting. Unsurprisingly, phone users are older than the overall population aged 15 and above given lower ownership rates in the 15-20 age category. Most notably, a clear gender divide in mobile phone ownership exists and phone users are disproportionately more male, especially in rural areas. In rural Kaolack, 56% of phone users are male against 48% in the population aged 15 and above and this difference is statistically significant at a 5% level. Comparable differences are found in 11 of the 14 rural regions of Senegal. Therefore, a plausible constraint on the validity of the *local representativeness* assumption is that mobile phone data under-represent women and younger individuals.

Secondly, we directly compare the phone users of our sample with the adult population based on a characteristic that is easily observable from CDR data: their residence location. Figure 12 (panel (**a**)) shows the number of users by voronoi cell against the population aged 15 and above, revealing a positive but imperfect correlation for both *subset A* and *subset B*. In Fig. 12 (panel (**b**)), we calculate the distribution of phone users in *subset A* and *subset B* across three categories of locations: Dakar, other urban locations, and rural locations. Comparing this with the distribution of the adult population, we see that a disproportionately high fraction of users are in Dakar, while a lower fraction are in rural areas compared to the overall adult population.

To go further, we explicitly investigate the relationship between the distribution of users and population density. Voronoi cells are ordered by population density and grouped into ten bins each accounting for 10% of the population aged 15 and above. The degree of selection of home locations with respect to population density is then assessed by calculating the distribution of users' home locations across these density bins. Note that in

	Urban			Rural		
	Phone users	All	Diff.	Phone users	All	Diff.
	(1)	(2)	(1)–(2)	(3)	(4)	(3)–(4)
wealth group: richest	0.095	0.085	0.01	0.096	0.087	0.01
wealth group: richer	0.200	0.184	0.016	0.263	0.223	0.04
wealth group: middle	0.189	0.192	– 0.003	0.290	0.291	– 0.001
wealth group: poorer	0.255	0.262	– 0.007	0.280	0.295	– 0.015
wealth group: poorest	0.261	0.277	– 0.016	0.071	0.104	– 0.033
Years of education	5.439	5.272	0.167	3.074	2.812	0.262
Age	30.371	28.446	1.925**	31.408	29.337	2.071*
Male	0.468	0.451	0.017	0.564	0.476	0.088**
Married	0.525	0.461	0.064*	0.693	0.621	0.072
Has a bank account	0.226	0.179	0.047	0.150	0.090	0.06
occupation: not working	0.302	0.343	– 0.042	0.200	0.213	– 0.013
occupation: agriculture	0.037	0.036	0.001	0.530	0.548	– 0.019
occupation: sales	0.253	0.237	0.016	0.087	0.071	0.016
occupation: household/domestic	0.033	0.044	– 0.01	0.013	0.012	0.001
occupation: unskilled	0.138	0.148	– 0.01	0.099	0.097	0.002
Household size	10.462	10.474	– 0.012	14.009	13.704	0.305
Water access	0.697	0.688	0.009	0.518	0.489	0.03
Access to sanitation	0.845	0.825	0.02	0.364	0.323	0.042
Electricity	0.844	0.836	0.008	0.320	0.282	0.037

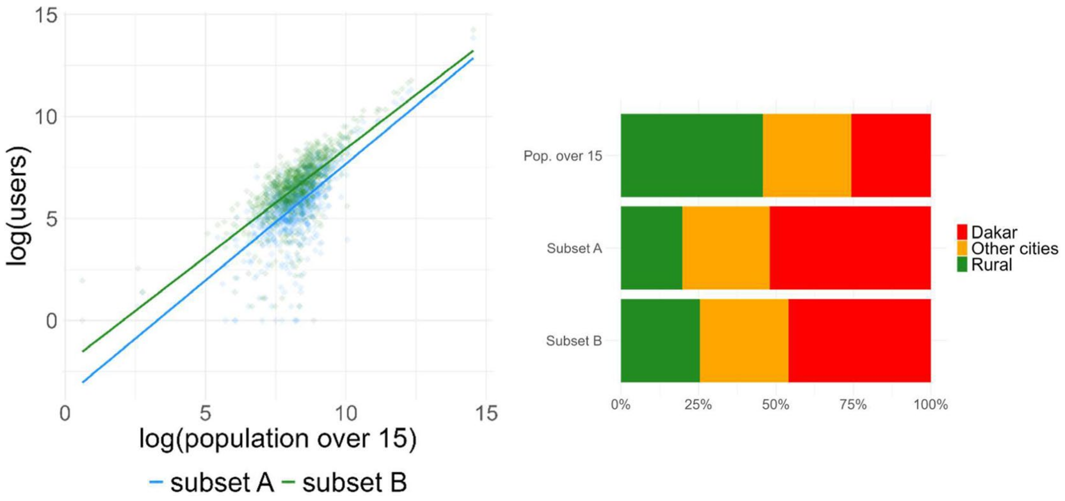
Table 6. Differences in characteristics between phone users and the population, Kaolack. Statistics were derived from the Senegal 2017 DHS men and women individual datasets. Results for all other regions are left in the Supplementary Material.

the absence of selection, the fraction of users found in each bin should match the share of the population it hosts (i.e. 10%). Results are showed in Fig. 13 (panel (**a**)) and a clear pattern of selection emerges where the fraction of users increases with population density, for both *subset A* and *subset B*. In short, although phone users are broadly similar to the overall population aged 15 and above locally (*local representativeness*), our samples tend to over-represent individuals residing in denser areas. This analysis highlights the significance of this selection pattern and underscores the relevance of the weighting scheme described in the Methods section, which precisely addresses these imbalances in the mobile phone data sample composition. An alternative way to see this is provided in Fig. 13 (panel (**b**)) that represents the population-to-users ratio against population density, calculated at the level of strata used as weighting units and defined in the Methods section. The graph reveals a negative correlation indicating that denser areas are associated with higher numbers of users relative to the local population (i.e. lower population-to-users ratios). The proposed weighting scheme allows to neutralize this systematic bias by making the ratio of the adult population over the (weighted) number of users constant and equal to 1 across strata.

Validation of filtering parameters: their impact on sample size and selection on home location.

The subsets used to compute temporary migration statistics (i.e. *subset A* and *subset B*) are constructed via a filtering procedure detailed in the Methods section. This involves imposing minimal constraints on users' frequency and length of observation, as well as the maximum time non-observed, primarily to ensure accuracy in migration detection outcomes. However, these higher observational constraints also result in smaller sample sizes and may exacerbate selection biases on the cross-section. In the first sub-section of the Technical Validation, we perform a quantitative analysis to examine the relationship between migration detection model accuracy and observational constraints. This analysis supports the notion that the filtering parameters used to construct *subset A* and *subset B* allow to maintain high levels of accuracy. In this sub-section, we validate the choice of filtering parameters by quantifying the costs associated with higher observational constraints, specifically in terms of reduced sample size and increased selection bias with respect to the distribution of home locations across space.

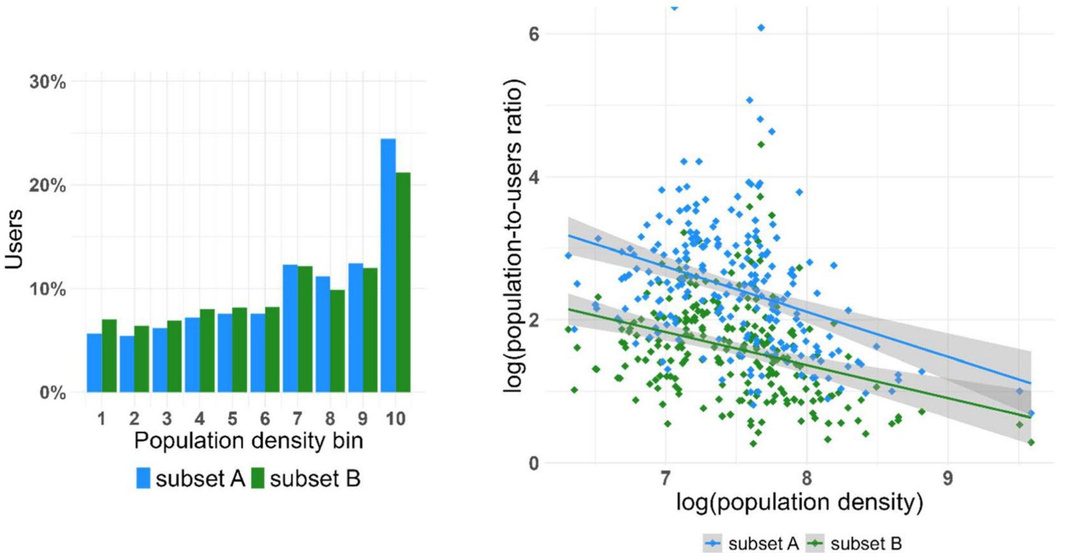
We first examine the impact of our three main observational constraints (frequency of observation, length of observation, maximum time non-observed) on sample size. To facilitate the visualization of the results, we evaluate the joint impact of any pair of constraints while holding the third fixed. Figure 14 (panel (**a**)) shows a three-dimensional surface representing the number of users remaining in a subset as a function of the minimal frequency (Ω) and length of observation (Δ) imposed. While increasing the constraint on Δ has a negative but limited impact on sample size, the frequency of observation has a much larger impact on sample size. Then, Fig. 14 (panel (**b**)) represents the sample size as a function of the minimal length of observation and the maximum observational gap allowed. Consistent with Fig. 14 (panel (**a**)), the constraint on the length of observation has a relatively minor impact on sample size. However, a clear non-linear impact of the maximum observational gap allowed is observed, with sample size decreasing sharply for values below 10–15 days. Finally, Fig. 14 (panel (**c**)) provides results consistent with those of Fig. 14 (panel (**a**)–(**b**)), where the minimum fraction of days



(a) Users versus adult population

(b) Distribution of users across three categories

Fig. 12 Comparison of the distribution of users across locations with the distribution of the population aged 15 and above. Panel (**a**) shows the logged number of users by voronoi cell in the 2013 dataset against the population aged 15 and above. The voronoi-level population aged 15 and above is estimated by combining census-based department-level estimations of the fraction of the population aged 15 and above[36] with voronoi-level total population estimates obtained by overlaying voronoi polygons with the 2017 100m-resolution gridded population product from the WorldPop Research Group[33]. Panel (**b**) shows the distribution of users in *subset A* and *subset B* in the 2013 dataset across three categories of voronoi cells: Dakar, other cells classified as urban, and cells classified as rural. The figure also provides the distribution of the target population aged 15 and above across these categories for comparison.



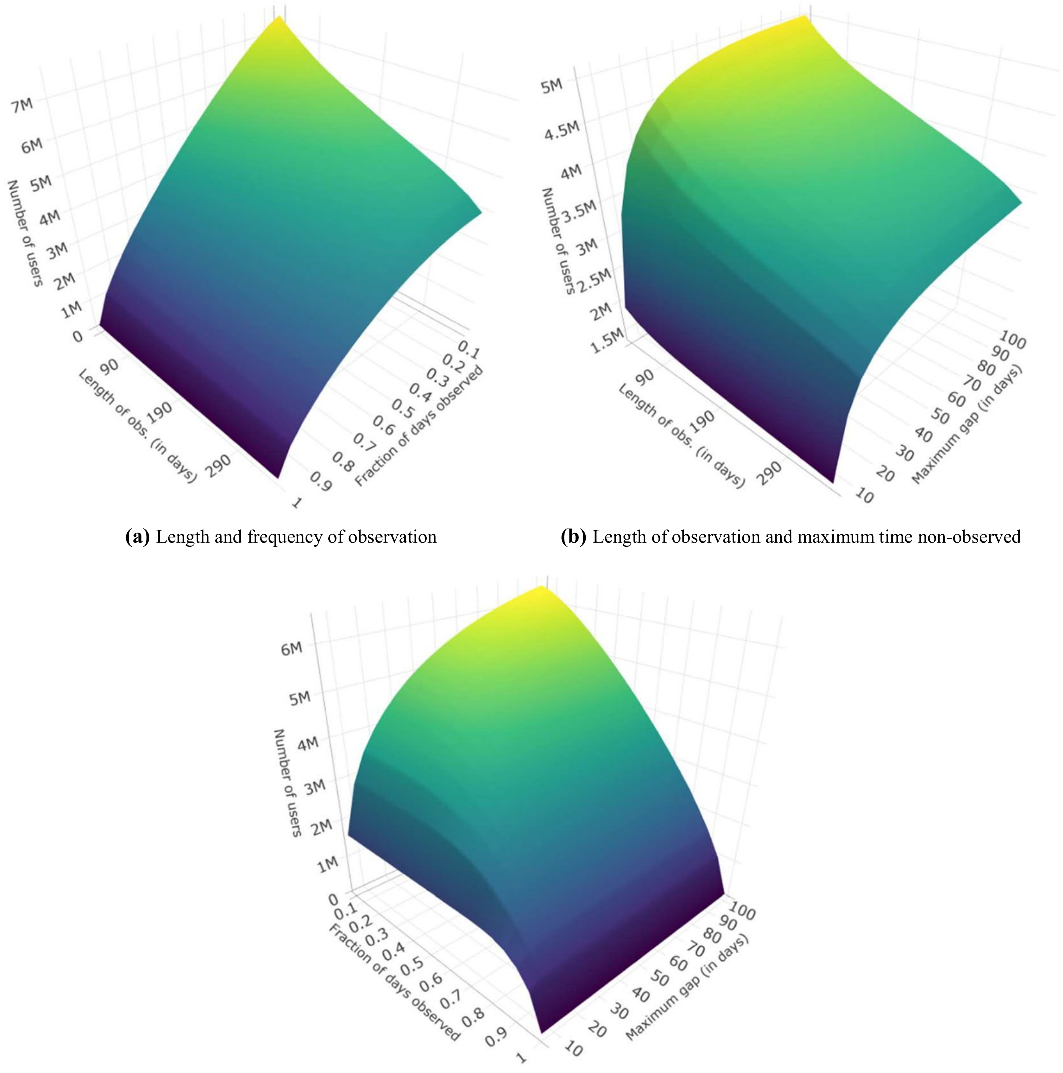
(a) Distribution of users across density bins

(b) Population-to-users ratio versus population density at the stratum-level

Fig. 13 Systematic bias of home locations toward denser areas. Panel (**a**) represents the distribution of phone users in the 2013 dataset across groups of cells defined based on population density deciles. Dakar is excluded from this analysis as it accounts for over 20% of the population and would cover the top two density bins. In any case, the selection towards Dakar is already clear in Fig. 12 (panel (**b**)). Panel (**b**) shows the ratio of the population aged 15 and above over the number of users at the stratum-level in the 2013 dataset against population density.

observed has a significant marginal impact on sample size while imposing a maximum observational gap lower than 10–15 days causes significant losses in sample size.

Then, we evaluate the impact of filtering parameters on the pattern of selection toward denser areas documented above. Specifically, we first estimate the impact of filtering parameters on the bias toward Dakar in the



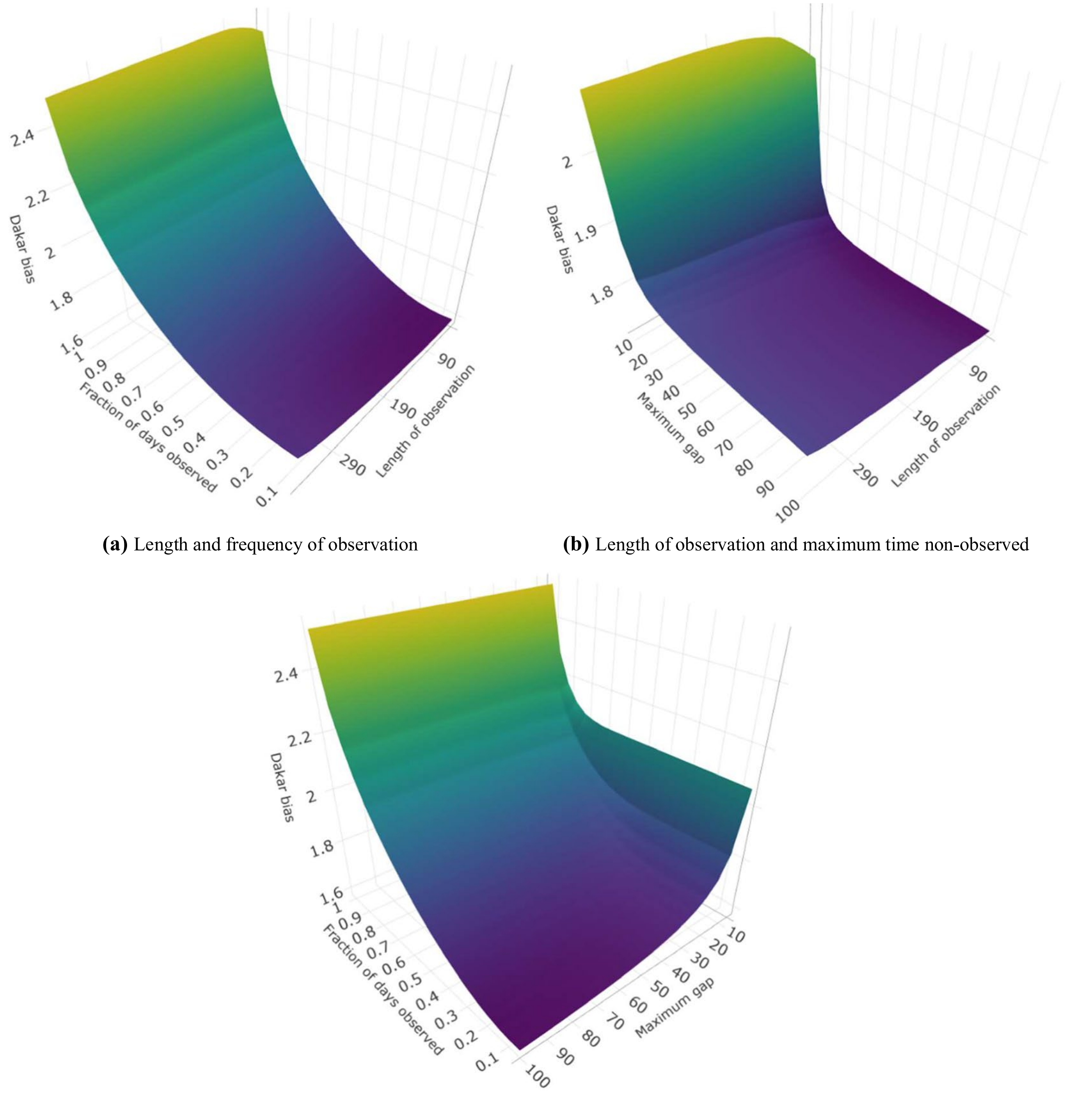
(a) Length and frequency of observation

(b) Length of observation and maximum time non-observed

(c) Frequency of observation and maximum time non-observed

Fig. 14 Impact of filtering parameters on the number of users left in a subset. Panel (**a**) shows three-dimensional surface representing the number of users in the 2013 dataset as a function of the minimal length and frequency of observation imposed, with the maximum time non-observed set to 100 days. Panel (**b**) represents the number of users as a function of the length of observation and the maximum time non-observed, setting the fraction of days with observations to 0.5. Panel (**c**) shows the number of users as a function of the fraction of days with observations and the maximum time non-observed, with the minimal length of observation set to 110 days.

sample composition. We define this bias for any given subset as the ratio of the fraction of users with an inferred home location in Dakar over the fraction of the population aged 15 and above effectively residing in Dakar. For example, a value of 2 would indicate that the sample contains twice as many users in Dakar than there would be if the sample had been randomly drawn from the target population. In the spirit of Fig. 14, we represent in Fig. 15 the value of this bias as a function of any pair of constraints, holding the third parameter fixed. Figure 15 (panel (**a**)–(**b**)) clearly shows that selecting users with a higher length of observation has practically no impact on the bias toward Dakar. In contrast, Fig. 15 (panel (**a**)) and (panel (**c**)) reveal that augmenting the minimum frequency of observation exacerbates the bias. For instance, regardless of the minimum length of observation, the bias increases from about 1.6 for a minimum fraction of days of 0.1 to 2-2.2 when this constraint is raised to 0.9 (Fig. 15, panel (**a**)). Similar to the impact of filtering parameters on sample size (Fig. 14), reducing the maximum time unobserved increases the bias only for values below a threshold of about 10–15 days (Fig. 15, panel (**b**)–(**c**)). Notably, this impact diminishes significantly when higher values for the minimal frequency of observation are considered, specifically for a minimum fraction of days observed around 0.8 and above (Fig. 15, panel (**c**)). Secondly, we assess the degree of selection induced by filtering parameters on the composition of the rest of the sample. To do this, we calculate the distribution of phone users across groups of cells defined based on population density deciles (as in Fig. 13, panel (**a**)), for different values of filtering parameters. Figure 16 (panel



(a) Length and frequency of observation

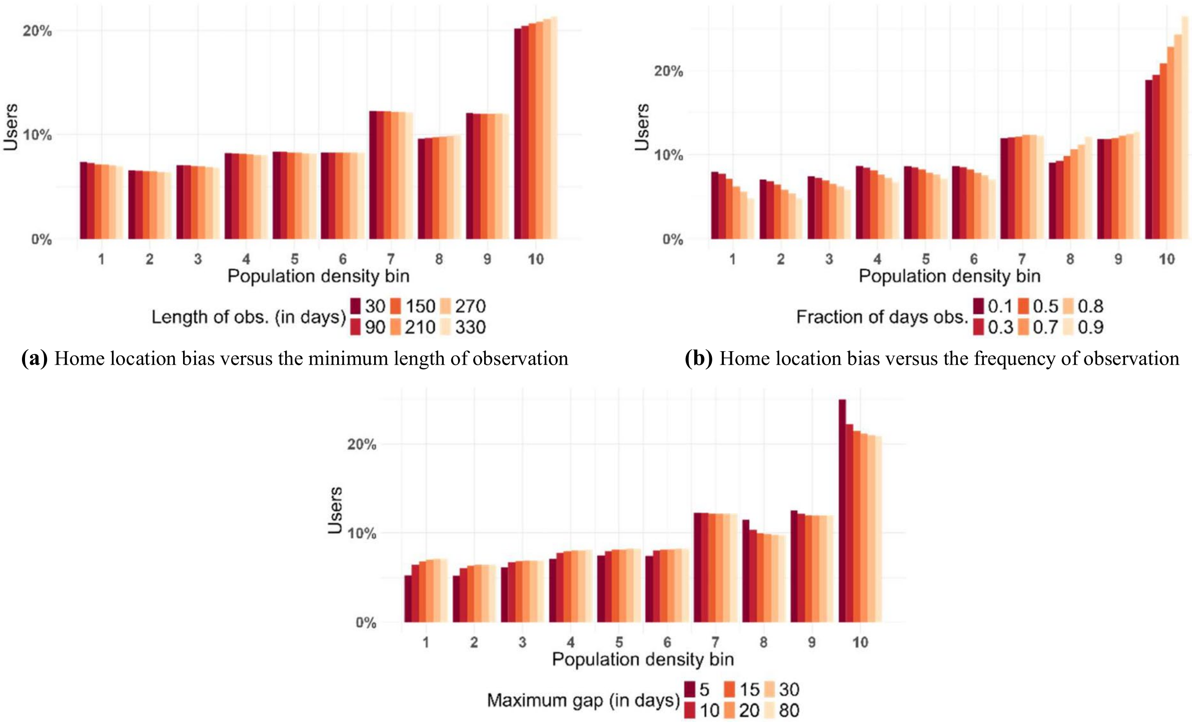
(b) Length of observation and maximum time non-observed

(c) Frequency of observation and maximum time non-observed

Fig. 15 Impact of filtering parameters on the bias of home locations toward Dakar. "Dakar bias" is defined for any given subset of users as the ratio between the fraction of users in the subset residing in Dakar and the fraction of individuals effectively living in Dakar in the target population. Panel (**a**) shows a three-dimensional surface representing Dakar bias in subsets of the 2013 dataset as a function of the minimum length and frequency of observation imposed, with the maximum time non-observed set to 100 days. Panel (**b**) represents Dakar bias as a function of the length of observation and the maximum time non-observed, setting the fraction of days with observations to 0.5. Panel (**c**) shows Dakar bias as a function of the fraction of days with observations and the maximum time non-observed, with the minimal length of observation set to 110 days.

(**a**)–(**c**)) shows how these distributions vary with the minimum length of observation, the minimum frequency of observation, and the maximum time non-observed, respectively. Once again, increasing the minimum length of observation has minimal impact on the distribution of users across density bins (Fig. 16, panel (**a**)), and reducing the maximum time non-observed allowed slightly exacerbates the bias toward denser areas only for values around 10 days and below. Finally, the frequency of observation is the main parameter influencing the bias in the distribution of users across density bins. As illustrated in Fig. 16 (panel (**b**)), increasing the minimum fraction of days observed to select a subset magnifies the tilt toward categories of denser cells. For instance, increasing the fraction of days observed from 0.1 to 0.9 decreases the fraction of users in the category of least dense cells (bin 1) from 8% to 5% and increases the share of users in the densest areas (bin 10) from 19% to 26%.

In summary, the cost of the filtering procedure in terms of reduced sample size and increased selection is primarily driven by the frequency of observation parameter, which is also the main determinant of the migration detection model accuracy. Moreover, it is worth noting that the impact on selection is clear but largely contained. Imposing a high minimum fraction of days with observations induces further distortions toward Dakar and denser areas, but the resulting subsets still provide wide coverage with consistent fractions of users found in the



(a) Home location bias versus the minimum length of observation

(b) Home location bias versus the frequency of observation

(c) Home location bias versus the maximum time non-observed

Fig. 16 Impact of filtering parameters on the bias of home locations toward denser areas. Panel (**a**) represents the distribution of non-Dakar phone users across population density bins, for subsets of the 2013 dataset associated with different values of the minimum length of observation, and a minimum fraction of days observed and maximum time non-observed fixed and set equal to 0.5 and 100 days respectively. As in Fig. 13 (panel (**a**)), each bin is a group of cells with similar population density that account for 10% of the non-Dakar population aged 15 and above. Similarly, panel (**b**) shows the distribution of non-Dakar phone users across population density bins, for subsets of the 2013 dataset associated with different values of the minimum fraction of days observed, and a length of observation and maximum time non-observed equal to 210 days and 100 days respectively. Panel (**c**) shows the distribution of non-Dakar phone users across population density bins, for subsets of the 2013 dataset associated with different values of the maximum time non-observed, and a length of observation and a fraction of days observed equal to 210 days and 0.5 respectively.

most remote locations. Consequently, the minimum fraction of days observed of 0.8 used to construct *subset A* is viewed as a credible tradeoff allowing to achieve a high level of accuracy while avoiding a significant reduction in sample size and an unreasonable distortion in sample composition. The value used for the construction of *subset B* (0.5) is thought of as resulting from a tradeoff assigning relatively more weight to the cost on selection and less to the accuracy of the migration detection model. On the other hand, the constraint on the length of observation remains quite stringent for both subsets since it has only a limited impact on sample size and selection. Finally, we do not consider values below 15 days for the maximum time non-observed given the potentially large impact implied on sample size and selection. Also, with temporary migration events being defined with a minimum duration of 20 days, constraining observational gaps to a maximum of 15–25 days is in fact sufficient to avoid cases of non-random attrition where users would be non-observed precisely while in migration.

It is essential to note that the distortion in sample composition caused by the filtering procedure is more concerning for unweighted estimates. The observed pattern of selection leads to a small over-representation of individuals in cities and denser areas and therefore tends to move the sample away from the target population of mobile phone users. On the other hand, weighted estimates systematically address discrepancies between the distribution of users' home locations in a given subset and the distribution of the target population.

Furthermore, it is crucial to acknowledge that without information on users' socio-economic characteristics, we cannot fully assess the impact of filtering parameters on the validity of the *local representativeness* assumption. Future research could delve into this aspect. Nevertheless, given the modest selection patterns observed in the distribution of home locations, there are reasons to believe that selection at a local level induced by the filtering procedure remains limited.

Usage Notes

The temporary migration estimates in the dataset can be aggregated to coarser spatial and temporal resolutions, but users should be aware of certain limitations. Absolute migration flows (e.g., the number of departures and returns) can be summed without restriction. For instance, to find the total number of migration departures in Senegal for 2013, one can simply add up the departures across all origin-destination pairs and half-months for that year. However, computing a departure rate is not feasible since the denominator – i.e. the number of users observed – cannot be directly derived from the dataset. This calculation would require knowing the number of unique users observed throughout the entire year 2013, specifically those with CDR trajectories allowing to detect any migration movements if they effectively occurred. Then, migration stock estimates can be aggregated spatially without restriction for any given half-month period. For example, the total stock of temporary migrants to Dakar in the first half of August 2013 is obtained by adding up the stock of migrants to Dakar across all origin locations for that particular half-month. Nonetheless, aggregating migration stocks over time periods longer than the minimum duration defined for temporary migration events may not always be meaningful. Alternative metrics could be considered to provide temporary migration measures over extended periods of time. For instance, one could calculate the number of unique users with at least one migration event of at least 20 days having occurred during 2013. Such metrics may be included in future versions of the dataset.

Code availability

The full code allowing to process raw mobile phone data and produce a final temporary migration dataset at the desired level of granularity is made available on GitHub at https://github.com/blanchap/TempMigration_SciData. A README file containing additional information on the content of each script and how to properly execute them is provided in the GitHub repository.

Received: 17 June 2024; Accepted: 10 February 2025;
Published online: 20 June 2025

References

1. Young, A. Inequality, The Urban-Rural Gap, And Migration. *Quarterly Journal of Economics* 1727–1785, https://doi.org/10.1093/qje/qjt025 (2013).
2. Bryan, G. & Morten, M. The aggregate productivity effects of internal migration: Evidence from indonesia. *Journal of Political Economy* **127**, 2229–2268, https://doi.org/10.1086/701810 (2019).
3. Baker, J. & Aina, T. A.The Migration Experience in Africa (Nordic Africa Institute, 1995).
4. Coffey, D., Papp, J. & Spears, D. Short-Term Labor Migration from Rural North India: Evidence from New Survey Data. *Population Research and Policy Review* **34**, 361–380, https://doi.org/10.1007/s11113-014-9349-2 (2015).
5. Delaunay, V. *et al.* La migration temporaire des jeunes au Sénégal: Un facteur de résilience des sociétés rurales sahéliennes? *Afrique Contemporaine* **259**, 75–94, https://doi.org/10.3917/afco.259.0075 (2016).
6. Findley, S. E. Does drought increase migration? A study of migration from rural Mali during the 1983-1985 drought. *International Migration Review* **28**, 539–553, https://doi.org/10.2307/2546820 (1994).
7. Schareika, N. Arid Ways : Cultural Understandings of Insecurity in Fulbe Society, Central Mali. *Nomadic Peoples* **1**, 120–125 (1997).
8. Bryan, G., Chowdhury, S. & Mobarak, A. M. Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh. *Econometrica* **82**, 1671–1748, https://doi.org/10.3982/ecta10489 (2014).
9. Lucas, R. E. Internal migration in developing countries. vol. 1 of *Handbook of Population and Family Economics*, 721–798, https://doi.org/10.1016/S1574-003X(97)80005-0 (Elsevier, 1997).
10. Mauro, G., Luca, M., Longa, A., Lepri, B. & Pappalardo, L. Generating mobility networks with generative adversarial networks. *EPJ Data Science* **11**, 58, https://doi.org/10.1140/epjds/s13688-022-00372-4 (2022).
11. González, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779–782, https://doi.org/10.1038/nature06958 (2008).
12. Jurdak, R. *et al.* Understanding human mobility from twitter. *PLOS ONE* **10**, 1–16, https://doi.org/10.1371/journal.pone.0131469 (2015).
13. Hughes, C. *et al.* Inferring migrations: traditional methods and new approaches based on mobile phone, social media, and other big data: feasibility study on inferring (labour) mobility and migration in the european union from big data and social media data, https://data.europa.eu/doi/10.2767/61617 (2016).
14. Zagheni, E., Weber, I. & Gummadi, K. Leveraging facebook's advertising platform to monitor stocks of migrants. *Population and Development Review* **43**, https://doi.org/10.1111/padr.12102 (2017).
15. Fiorio, L. *et al.* Using twitter data to estimate the relationship between short-term mobility and long-term migration. 103–110, https://doi.org/10.1145/3091478.3091496 (2017).
16. Mazzoli, M. *et al.* Migrant mobility flows characterized with digital data. *PLOS ONE* **15**, https://doi.org/10.1371/journal.pone.0230264 (2020).
17. Blumenstock, J. E. Inferring Patterns of Internal Migration from Mobile Phone Call Records : Evidence from Rwanda. *Information Technology for Development* **18**, 107–125, https://doi.org/10.1080/02681102.2011.643209 (2012).
18. Zufiria, P. J. *et al.* Identifying seasonal mobility profiles from anonymized and aggregated mobile phone data. Application in food security. *PLoS ONE* **13**, https://doi.org/10.1371/journal.pone.0195714 (2018).
19. Hankaew, S. *et al.* Inferring and Modeling Migration Flows Using Mobile Phone Network Data. *IEEE Access* **7**, 164746–164758, https://doi.org/10.1109/ACCESS.2019.2952911 (2019).
20. Lai, S. *et al.* Exploring the use of mobile phone data for national migration statistics. *Palgrave Communications* **5**, https://doi.org/10.1057/s41599-019-0242-9 (2019).
21. Chi, G., Lin, F., Chi, G. & Blumenstock, J. A general approach to detecting migration events in digital trace data. *PLoS ONE* **15**, https://doi.org/10.1371/journal.pone.0239408 (2020).
22. République du Sénégal, Ministère de l'économie, des finances et du plan, Agence National de la Statistique et de la Démographie (ANSD). Enquête à l'écoute du sénégal 2014 (2015).
23. Pesaresi, M., Florczyk, A., Schiavina, M., Melchiorri, M. & Maffenini, L. GHS-SMOD R2019A - GHS settlement layers, updated and refined REGIO model 2014 in application to GHS-BUILT R2018A and GHS-POP R2019A, multitemporal (1975-1990-2000-2015) - OBSOLETE RELEASE. European Commission, Joint Research Centre (JRC) [Dataset], https://doi.org/10.2905/42E8BE89-54FF-464E-BE7B-BF9E64DA5218.

24. Blumenstock, J. & Eagle, N. Mobile divides: Gender, socioeconomic status, and mobile phone use in rwanda. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, ICTD '10, https://doi.org/10.1145/2369220.2369225 (Association for Computing Machinery, New York, NY, USA, 2010).
25. Blanchard, P., Gollin, D. & Kirchberger, M. Perpetual Motion: High-Frequency Human Mobility in Three African Countries. Trinity Economics paper tep0823, Trinity College Dublin, Department of Economics (2023).
26. Blumenstock, J. E., Chi, G. & Tan, X. Migration and the Value of Social Networks. *The Review of Economic Studies* **rdad113**, https://doi.org/10.1093/restud/rdad113 (2023).
27. Vanhoof, M., Reis, F., Ploetz, T. & Smoreda, Z. Assessing the quality of home detection from mobile phone data for official statistics. *Journal of Official Statistics* **34**, 935–960, https://doi.org/10.2478/jos-2018-0046 (2018).
28. Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W. & Buckee, C. O. Heterogeneous mobile phone ownership and usage patterns in kenya. *PLOS ONE* **7**, 1–6, https://doi.org/10.1371/journal.pone.0035319 (2012).
29. Agence Nationale de la Statistique et de la Démographie - ANSD/Sénégal, & ICF. Sénégal : Enquête démographique et de santé continue - eds-continue 2017 Rockville, Maryland, USA : ANSD et ICF (2018).
30. Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W. & Buckee, C. O. The impact of biases in mobile phone ownership on estimates of human mobility. *J R Soc Interface* **10**, https://doi.org/10.1098/rsif.2012.0986 (2013).
31. Felbo, B., Sundsøy, P., Pentland, A., Lehmann, S. & Montjoye, Y.-A. Using deep learning to predict demographics from mobile phone metadata. In *International Conference on Representation Learning (ICLR) 2016 Workshop* (2015).
32. Blanchard, P. & Rubrichi, S. A highly granular temporary migration dataset derived from mobile phone data in Senegal *figshare* https://doi.org/10.6084/m9.figshare.28023170 (2024).
33. Qader, S. *et al*. Census disaggregated gridded population estimates for senegal (2020), version 1.0 (2022).
34. Agence Nationale de la Statistique et de la Démographie (ANSD) de la République du Sénégal. RGPHAE 2013 www.ansd.sn (2014).
35. Research ICT Africa. RIA ICT Access Survey 2017-2018 [Senegal], https://doi.org/10.25828/kcsd-nb04 (2018). Version 1.1. Cape Town: RIA [producer], 2020. Cape Town: DataFirst [distributor], 2020.
36. Minnesota Population Center. Integrated public use microdata series, international: Version 7.3 [senegal 2013 census] (2020). Minneapolis, MN: IPUMS.

Acknowledgements

This study was partly conducted within the OPAL project. Mobile phone data were provided by Sonatel (Senegal) as part of the D4D Challenge and the OPAL project.

Author contributions

P.B. and S.R. designed the research. P.B. and S.R. analysed the data. P.B. performed the statistical analysis. P.B. wrote the Article. All authors edited and approved the final version of the Article.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04599-4.

Correspondence and requests for materials should be addressed to P.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025