Human Mobility In Developing Countries: Evidence From Mobile Phone Data

A Thesis Submitted to Trinity College Dublin, the University of Dublin in Application for the Degree of Doctor of Philosophy

By

Paul Baptiste Blanchard

Supervised By

Martina Kirchberger

Dublin, 2024

Declaration, Online Access and the General Data Protection Regulation

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. Chapter 1 of this thesis is co-authored with Douglas Gollin and Martina Kirchberger. Chapter 3 is co-authored with Virginie Comblon, Flore Gubert, Erwan Le Quentrec, Anne-Sophie Robilliard and Stefania Rubrichi.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Paul Blanchard
September 15th, 2023

Summary

This thesis examines human mobility patterns in developing countries. It leverages different types of mobile phone data enabling the observation of individuals' movements and locations at spatio-temporal scales which are notoriously difficult to attain with traditional survey data. These movements are investigated in various contexts, including spatial inequalities and frictions, labor reallocation dynamics in the face of climate variability, and the interplay between different modalities of short-term mobility as mechanisms to access urban markets.

Chapter 1 uses smartphone app location data from three African countries over a one-year period to characterize patterns of high-frequency mobility. The data reveal the types of locations that people visit and the frequency with which they make trips. Overall, they point to considerable mobility within the sample. The average smartphone user in the data ventures more than 10 km from home on 10-15% of the days when they are observed. On average, when observed away from home, smartphone users are typically 35-50 km from home. The granular nature of the data allows to obtain insights into the specific destinations where people are observed when they are away from home. These include locations associated with shops and markets, government offices, and places offering a range of goods, services, and recreational venues. Big cities seem to be particularly important destinations, perhaps reflecting the range of amenities that they offer to visitors. A conceptual framework is developed to characterize the role of visits for individuals. It provides a number of testable predictions that are consistent with the movement patterns that are observed in the data. Although the sample of smartphone users is not representative of national populations, their mobility patterns offer novel insights into spatial frictions and the geographic patterns of economic activity.

Chapter 2 provides a suite of methodological tools to derive temporary migration statistics from mobile phone data. First, the chapter delves into well-known challenges surrounding cross-sectional representativeness. Additionally, it offers new insights on potential measurement errors and selection issues related to sample characteristics on the time dimension. Second, an enhanced temporary migra-

tion detection method based on a clustering technique is presented, wherein the preliminary identification of users' primary residence location allows for a better characterization of movement direction, i.e. departure versus return. Third, the chapter addresses challenges in creating time-disaggregated temporary migration statistics derived from individual trajectories, presenting a well-defined set of rules to mitigate them. Applying this methodological framework to three years of Call Detail Records (CDR) from Senegal, the results unveil a remarkable level of temporary movements and bring novel understandings about the magnitude, timing, spatial distribution and orientation of temporary migration trends in Senegal.

Chapter 3 investigates the impact of climate variability on the short-term spatial reallocation of individuals in a developing country setting. Centering on Senegal as a case study, the study scrutinizes the relationship between precipitation patterns during the rainy season (spanning June to October) and the temporary migration decisions for the remaining agricultural calendar. We exploit a multi-year mobile phone dataset in Senegal and draw on the methods developed in Chapter 2 to construct a uniquely granular temporary migration matrix. We combine the temporary migration dataset with satellite-based local precipitation measures, and exploit variations in the quality of rainy seasons over three years to identify the effect of climate variability on temporary migration decisions. First, we take advantage of the richness of our data to corroborate a recent explanation to a recurrent empirical puzzle in the migration literature, whereby local shocks are found to have significant impacts on local economic outcomes but only limited effect on out-migration rates from affected areas. Relying on our empirical setting, we confirm that failing to incorporate conditions at destinations when studying migration responses to local shocks can effectively yield misleading results. Second, we outline a simplified location choice model, within which climate variability is incorporated in the production environment. This conceptual framework serves to motivate the estimation of dyadic regressions for the analysis of the effect of rainy season conditions on temporary migration. Our main findings indicate that poorer rainfall conditions at a rural origin during the rainy season (June-October) impede temporary migration during harvest (September-November) but act as a push factor over the following off-season (February-May). We find evidence that these effects are more pronounced in rural locations exhibiting lower standards of living. On the other hand, the quality of the rainy season at a destination is found to be positively related to the level of attractiveness of that destination to temporary migrants.

Chapter 4 builds upon the novel insights on human mobility provided in Chapter 1, and examines the interplay between two predominant forms of transient

mobility toward cities in developing countries: visits and temporary migration. Drawing from the Senegalese mobile phone data utilized in previous chapters, this study taps into the unique amalgamation of high frequencies of observation with extended observation spans offered by CDR. Such a combination renders the CDR data especially apt for capturing individuals' mobility across various temporal scales. I rely on the methodology of Chapter 2 to measure temporary migration, and I develop a simple procedure for detecting visits within CDR trajectories. I provide an overview of visiting patterns that not only corroborates but also expands upon the main findings of Chapter 1, thanks to a larger rural coverage and longer periods of observation. During a year of observation, 83% of users embark on at least one visit to a city, while 17% temporarily migrate to an urban locale. Then, I investigate the relation between users' decisions to visit and temporarily migrate to cities via regression analyses. The results indicate that those who engage in urban temporary migrations exhibit 17.5 additional days of visits compared to their counterparts, which is almost entirely driven by supplementary visits made to their migration destination. Furthermore, I provide evidence highlighting the non-random nature of visits to migration destination; they display discernible patterns around the departure and return dates of temporary migration episodes. The mobility patterns are consistent with anticipatory and follow-up behaviors, wherein individuals willingly incur the costs of pre-migration visits to gain information about the destination, and often make subsequent visits in the weeks following their return. Finally, I capitalize on the simultaneous observation of both visits and temporary migration choices to shed light on cost differentials between these two mobility modalities. Findings from gravity regressions suggest that the inherent fixed costs of temporary migration to cities exceed those related to visits.

Acknowledgements

Pursuing a PhD is often perceived as a solitary journey. Yet, numerous individuals have contributed in their own ways to the elaboration of this thesis.

I am profoundly indebted to Martina Kirchberger for introducing me to this fascinating area of research, and the guidance provided throughout my PhD. I wish to extend my sincere gratitude to Nicola Fontana for his availability, constant dedication, and kindness that were of great help when they were most needed. I also thank Gaia Narciso for her kind support.

Had it not been for the encouragements and subtle nudges from remarkable characters I've had the privilege to meet, my journey would undoubtedly have taken a different direction. First of all, I wish to address my sincere gratitude to Doug Gollin: not only have you been a pivotal supporter throughout various stages of my academic and professional journey, but your wisdom and renowned kindness continue to inspire me. I wish to thank Oscar Ishizawa for offering me the numerous opportunities that eventually led to my decision to embark on the PhD journey. The trust you often placed in me contributed to build the self-confidence required to undertake this exercise. This applies to so many of my former colleagues, including Edmundo Murrugarra, Mathieu Lefebvre, and Juan Carlos Parra. I also thank Aline Coudouel for her sound and wise advice.

I wish to extend my sincere gratitude to Orange Sonatel and Orange Innovation for their invaluable contribution in providing the mobile phone data used in this thesis. Their generous data provision within the framework of the D4D challenge and the OPAL project has been instrumental to my research. I would like to extend heartfelt appreciation to Stefania Rubrichi for her unwavering commitment during countless hours of debugging and deciphering the intricacies of server glitches.

I am grateful to the incredible people I've met at TCD and who also contributed to make this experience worthwhile. Eugenia, Beren, Alexis, Fra, Max and all the others: thanks for making TRiSS such a fun and stimulating place to work at. Special thanks to Vince who provided invaluable support in the very last stretch.

My sincere thanks go to the entire DIAL office for welcoming me as a visiting researcher in Paris: Jeanne, Marin, Florence, Thomas, Marion, Alex, Antoine, Fatou, Mary, Kenneth, Camille, JN, Elise, Anne-Laure, Véro, Sandrine, and all those I'm

probably forgetting. My profound gratitude goes to Flore Gubert and Anne-Sophie Robilliard for placing their trust in me and providing the chance to contribute to their research projects. I would also like to thank Mathilde for sharing with me the long struggle of the final year, and the countless chocolate tablets that go with it.

On a personal note, I would like to express my heartfelt gratitude to my parents, my brother Olivier and my sister Sophie for their constant love, support, and faith in my abilities. To my best friend Max, for his support and invaluable friendship during these four years, and over the twenty years that preceded.

The Covid crisis obviously had a notable impact on my PhD journey, and some people extended their support when I most needed it to navigate these difficult times. Pyerre, thanks for the few months of improvised "colloc", it meant a lot. Sarah, thank you for welcoming me in Kigali when I critically needed some fresh air, thanks for your constant energy and your inspiring boundless optimism. I also owe a great deal of gratitude to my sister Sophie for welcoming me into her home when I needed it.

I feel incredibly lucky to be surrounded by so many individuals whom I can call my friends and have accompanied me throughout this journey: Ali, Rafa, Max, Belette, Thibaut, Chev, Romain, Zaz, Barb and so many others. I wish to thank Richard for his support and the inspiring conversations, Bruno for the good laughs and the hours upside down that perhaps sometimes cleared my head, and Julien for the good coffee without which any attempt to achieve this type of endeavor would be vain.

Lastly, but by no means least, I want to thank my partner Marion, for her endless patience and understanding throughout this lengthy and challenging process. Your unwavering support and kind love were the pillars that undeniably guided me to the moment of penning this concluding word.

Contents

List of Figures

List of Tables

Introduction

Human mobility is an integral component of development processes. From rural-urban migration driven by the promise of better economic opportunities to cross-border movements triggered by regional disparities, these movements reflect and influence the evolving economic landscape. In the developing world, where economies are in transition and the pace of urbanization is swift, understanding human mobility is crucial for several reasons. First, it informs us about the dynamics of labor markets, wages, and living conditions. These movements, in turn, impact wage equilibria, resource allocations, and productivity levels in both the origin and destination regions.

Second, mobility patterns are intricately linked with infrastructure development, urban planning, and housing policies. As people move to cities, they shape urban growth, leading to challenges related to infrastructure demand, housing shortages, and urban sprawl. Efficient policy planning, thus, requires insights into these mobility trends at different temporal scales.

Additionally, human mobility can be an adaptive response to environmental stresses, socio-political conflicts, and economic downturns, reflecting the resilience and adaptive capacities of populations. It also provides an opportunity for individuals to diversify income sources, potentially insulating households from localized economic shocks.

Yet, the dynamics of human mobility beyond permanent migratory movements in developing countries remain underexplored. This thesis aims to delve into the intricacies of subtler patterns of human mobility in the developing world, harnessing the potential of mobile phone data. It hopes to shed light on the economic implications, challenges, and opportunities associated with human mobility at previously understudied spatio-temporal dimensions. Through this exploration, it aspires to perhaps contribute valuable insights to policymakers, researchers, and development practitioners eager to harness the potential of human mobility for sustainable economic growth and development.

Chapter 1 examines a type of human mobility that has previously been difficult to capture in a developing context. Census data and standard surveys have

primarily allowed to capture migration between survey waves, but fail to capture movements that do not involve an individual's long-term relocation. On the other hand, recent studies have made strides in tracking temporary or seasonal migrations via specialized surveys – although their geographical scope remains limited. We know little, however, about human mobility over other temporal scales.

Meanwhile, a number of studies have highlighted the potential of newer sources of "big data" and digital traces to construct more granular measures of migration and commuting in low-income countries. Embracing these innovative approaches, our study taps into novel smartphone app location data in three African countries to examine what we term as "visits": the short-term movements of individuals from their home location to other locations. These movements, therefore, stand apart from routine commutes and do not imply a shift in the individual's permanent residence. For instance, short trips from rural to urban locations, or between cities differing in size, may enable individuals to access the amenities of large agglomerations without migrating to these locations.

Each entry in our dataset denotes an instance when a user's smartphone accesses the internet via an application,[1] providing a precise timestamp and a GPS coordinate location for each such use. We use these information to observe the movements and locations of over one million smartphone devices over an entire year across three large African countries: Nigeria, Kenya, and Tanzania.

First, we rely on secondary survey data sources to contextualize the demographics of smartphone owners in comparison to the broader population. Of course, smartphone users are not representative of the general population and our sample exhibits anticipated patterns of cross-sectional bias. Nevertheless, in the urban regions across our three studied countries, smartphone ownership manifests at noteworthy rates, ranging from 23% in Nigeria to a high of 51% in Kenya.

Second, we construct a set of mobility metrics that characterize the frequency of visits, the spatial extent of such movements, as well as the destination characteristics. We evaluate these measures within the three countries from which our data is sourced, and the results reveal a high degree of mobility among smartphone users. They are seen more than 10km away from their home location on about 10-15% of the days on which they are observed. Individuals residing in less densely populated regions tend to travel away from their homes more frequently than those living in urban centers, and travel longer distances when they depart from their home locales. Examining spatial transition matrices, we observe that in these countries, urban inhabitants visit towns and many rural villages. Correspondingly, some individuals originating from these villages travel to more prominent towns

[1]The data originates from a multitude of smartphone applications; however, we do know specifically which particular applications they are.

and urban centers. This suggests a robust network of connectivity across various geographical landscapes. The analysis of the composition of visitors to the main cities confirms those observations. The largest cities (Nairobi, Lagos, Dar Es Salaam) appear as mobility magnets and attract large numbers of non-urban dwellers. Moreover, secondary cities receive large inflows of visitors from these primate cities, as well as visitors from non-urban locales, but only modest flows originating from other secondary cities. This suggests that secondary cities are relatively substitutable for one another, whereas the major cities present distinctive amenities that are not available in other locations. We find that many users visit more than one city over the studied period, and we see users making multiple visits to the same city.

Third, we take advantage of the precise locations offered by smartphone data to identify the specific places that users frequent when they visit cities. We consider the six main cities across our three countries: Lagos, Abuja, Nairobi, Mombasa, Dar es Salaam and Dodoma. We link our data entries with Open Street Map polygons delineating buildings and areas with identified characteristics. Users are seen at a range of different places related to travel, administrative matters, health services, shops and markets, and commercial zones, thus reflecting the consumption of a wide array of amenities.

Finally, we develop a conceptual framework in which individuals decide to make city visits, taking into account an associated cost structure, and through which they consume an urban amenity. We derive a number of testable predictions, which we show are consistent with the mobility patterns observed in our data. For instance, the number of visits per person made from a smaller settlement to a larger one surpasses the number of visits made in the opposite direction. Moreover, the fraction of days users spend visiting a city depends on the distance between their home and the destination city, and thus follows a gravity equation. Lastly, controlling for distance, destinations with higher populations tend to be more appealing for prospective visitors than smaller cities.

This paper first contributes to a growing literature that has exploited digital traces to study human mobility and interactions. We add to this literature by providing a set of metrics allowing to describe a new type of mobility in a development context: visits. We study mobility patterns in three countries in sub-Saharan Africa, emphasizing inter-city movements in particular. Second, we relate to a literature that has used quantitative spatial models to study commuting, migration and labor markets. Our model focuses on visits and provide gravity-style expressions which are commonly found in this literature. Our study is connected to a large strand of literature that has documented significant and persistent spatial gaps in nominal wages and standards of living, as well as differences in productivity

across sectors in developing countries. These disparities are suggestive of significant frictions and market imperfections that constrain the movements of people and the flow of information, culminating in spatial and sectoral misallocation. We contribute to this discussion by analyzing the mobility patterns of individuals on a previously overlooked temporal dimension. The pervasiveness of visits across varied locales challenges the notion of insurmountable mobility costs.

More generally, our study has implications about the role of cities and the connectedness of locations in the context of spatial inequalities in developing countries. The widespread nature of visiting flows to urban areas suggest that cities provide benefits for a much larger set of people than their own residents and commuters. More importantly, the significant flows of visits originating from non-urban locales challenge the idea of a strict rural-urban divide, and could be interpreted as a mechanism by which people achieve partial urbanization.

Chapter 2 explores the use of mobile phone data to measure temporary migration within developing countries. Numerous studies have emphasized their significance in households' economic decisions, yet systematic inclusion of these temporary migrations in national statistical frameworks remains sparse, leaving them largely undocumented. In fact, such short-term movements are inherently difficult to measure due to recall biases and attrition, and require specialized and often costly survey instruments. Even when such surveys are deployed, they often adhere to strict definitions that miss relatively brief migration events of less than one to two months. These shorter durations, however, have been identified as constituting a substantial portion of temporary migration moves. Moreover, these surveys typically have a restricted geographical coverage, hindering a comprehensive analysis of temporary movements at national scales.

Concurrently, mobile phone data have emerged as a promising alternative, offering insights into human movements at broader geographic scales and with a refined spatio-temporal resolution. Nonetheless, deriving temporary migration statistics with mobile phone data poses several challenges. First, mobile phone users form a non-random subset of the population at large, which implies well-known issues of cross-sectional biases for the production of representative migration statistics. Moreover, patterns of phone usage align with socio-economic characteristics, potentially introducing additional biases when subsets of frequently observed users are selected for analysis. Second, the identification of migration events within raw trajectories from mobile phone data requires the definition of migration criteria and systematic algorithmic rules. Previous methods have predominantly used ad-hoc frequency-based methods, where a user's location over successive time frames is determined based on the most recurrent observation. Migration movements are

then flagged when a persistent shift in location is detected. Those methods present several drawbacks and limitations, including increased measurement errors on the timing and duration of migration events, an inability to detect short-duration migration episodes, and a deficiency in determining the directionality of movements. Third, periods of inactivity necessarily imply some degree of uncertainty in the exact start date, end date and, therefore, in the duration of migration events, which complicates the calculation of time-disaggregated migration measures . For instance, pinning down a migration departure to a particular time frame, such as a week or month, becomes uncertain if the user in question has not been observed for a significant span preceding that date.

In this chapter, I outline a set of methodological tools designed to address these various challenges, with the goal of generating robust temporary migration statistics. The first part of the chapter is dedicated to the analysis of mobile phone sample representativeness. I start by detailing systematic methodologies to effectively understand the non-random nature of mobile phone data samples. Statistics derived from secondary survey data allow to compare the characteristics of mobile phone users with those of a broader national population. This is informative about disparities between the at-large population and mobile phone users in general, but it does not pinpoint the cross-sectional biases inherent to a specific mobile phone dataset. In this respect, I also present simple metrics directly inferred from mobile phone data, allowing to effectively characterize the specific set of phone users in a dataset. Additionally, I delve into the temporal attributes of mobile phone data, focusing on the frequency with which users are observed and their observation span. First, I gauge the precision of the ensuing migration detection algorithm against those parameters. Most notably, I show that the detection rate declines sharply for users with a fraction of days with observations below 0.5. Second, I present a simple empirical test to investigate the existence of selection patterns on the time dimension, whereby users would exhibit observational gaps precisely when they temporarily migrate. I do not find evidence for such non-random attrition patterns in the sample of CDR from Senegal. Lastly, I quantify the influence of commonly employed filtering criteria on both sample size and observable cross-sectional, furnishing valuable guidance for selecting working subsets for the production of migration statistics.

Then, I build upon previous work to develop a temporary migration detection algorithm. The algorithm relies on a clustering method designed to identify periods of continuous presence of a user at a single location, allowing for idiosyncratic deviations corresponding to short visits at other locales. This segment-based approach has been showed to out-perform all versions of traditional frequency-based method. By examining user locations over an extended timeframe, I

also assign a primary residence to each individual. This step allows to clearly characterize the direction of migration flows, differentiating between departures from and returns to the designated home location.

Next, I develop a systematic methodology that facilitates the conversion of user-specific migration trajectories into time-disaggregated temporary migration statistics. I address the uncertainties arising from sampling irregularities and introduce a straightforward weighting strategy that compensates for spatial disparities in user distribution.

I illustrate the methodology with a large dataset of Call Detail Records (CDR) from Senegal. The temporary migration profile obtained from CDR data unveil a pronounced degree of short-term movements across the country. Nearly a third of the adult population engages in one or more temporary migrations of at least twenty days over a ten-month period, with a significant portion of these migrations estimated to last less than two months. The unique granularity and coverage of CDR data offer new insights into both the spatial and temporal dynamics of these movements. Rural areas exhibit a noticeably higher proportion of users undertaking temporary migration episodes compared to urban locales. Moreover, while primary cities, notably Dakar, account for a significant portion of the overall migration inflow, an equally substantial segment is directed towards rural regions. In particular, the data uncovers prominent short-distance, rural-to-rural movements. Furthermore, the data's three-year span distinctly brings to light marked seasonal trends. Large increases are consistently observed during the June-September period, coinciding with the rainy season. During this period, the count of temporary migrants roughly doubles compared to other times of the year.

This chapter makes a methodological contribution to the literature leveraging digital traces to study human mobility. While issues of cross-sectional biases in mobile phone data are well-established, I provide a streamlined suite of metrics and secondary data sources designed to systematically assess these biases. More importantly, this chapter elucidates previously unexplored facets of selection and measurement error associated with the temporal attributes of mobile phone data, i.e. with the frequency and length of observation of phone users. The temporary migration detection algorithm is largely inspired from previous work, but notably incorporates a primary home location estimation phase. This inclusion facilitates a more nuanced understanding of the directionality of temporary migration flows, distinguishing between departures from and returns to the designated home location. This research further enriches the migration literature centered on developing countries by offering a comprehensive account of temporary migration movements in Senegal, both at a national level and with an unparalleled spatial-temporal granularity.

Chapter 3 studies the impact of climate variability on temporary migration decisions in a developing context. The investigation centers on the specific case of Senegal, a country where a significant fraction of the population still lives in rural areas and heavily relies on precipitation patterns during a single rainy season. Rural households thus grapple with important year-on-year income fluctuations, occasionally facing negative shocks such as droughts. Past research has scrutinized various coping mechanisms, including consumption smoothing strategies, risk-sharing networks and informal insurance markets. Temporary migration constitutes another viable strategy in the face of income volatility. Yet, it remains largely understudied, mainly due to the lack of detailed data on these movements.

Building upon Chapter 2, we harness a multi-year CDR dataset to derive a highly granular pseudo-panel of temporary migration estimates in Senegal. This dataset is then paired with satellite-based local rainfall estimates, facilitating the observation of temporary migration choices across diverse periods of the year and under a range of rainy season conditions, notably after a pronounced drought event.

The reaction to variations in rainy season conditions in terms of temporary migration is not straightforward to predict. Anecdotal evidence suggests that historically nomadic West African populations are notably mobile. Adverse rainfall conditions could conceivably act as a push factor, prompting individuals to move temporarily from areas hampered by a shock to more productive areas. On the other hand, such income shocks may erode the financial capacity of affected households, exacerbating liquidity constraints and thus making migration unfeasible. Spatial frictions and market imperfections could further hinder households from resorting to temporary migration as a means to cope with income variability. Another perspective posits that temporary migration could be a standard element in livelihood strategies, specifically designed to mitigate income risks. If such strategies are well-devised and closely aligned with actual income distribution, the occurrence of a particular shock should not influence the decision to migrate.

We develop a simple temporary migration model which incorporates precipitations as an input to location-specific production functions. Individuals choose a location at each time period, where they provide labor. Movements come with a cost related to distance and individuals inherently favor their home location. This conceptual framework allows to derive a reduced-form expression relating rainy season conditions at an origin and a destination on the corresponding bilateral stock of temporary migrants for a specific time period.

First, we capitalize on the richness of our data to investigate an empirical

conundrum highlighted in the strand of literature studying migration responses to local shocks. Indeed, several studies have identified a detrimental effect of adverse shocks on economic outcomes such as income or employment. Yet, these impacts often do not often translate to migratory responses, prompting scholars to infer the presence of prohibitive migration costs. Recent work has suggested that this might stem from the bilateral nature of migration decisions. Failing to account for conditions at relevant destinations might introduce an omitted variable problem, especially when the shock exhibits significant spatial correlation. Taking advantage of our empirical setting, we estimate a conventional migration regression relating rainy season conditions at an origin to the (total) out-migration rate, and we compare the results to a dyadic regression that accounts for both rainfall conditions at origin and destination. The findings show critical differences, reinforcing the notion that traditional migration regressions should be interpreted with caution.

Second, we delve into the outcomes of dyadic regressions that offer insights into the impact of rainy season conditions at both the origin and destination on the bilateral stock of temporary migrants throughout various phases of the agricultural year, primarily focusing on rural locations. Controlling for precipitations at destination, poor rainfall conditions at origin are found to decrease temporary migration over the period immediately following the rainy season, and corresponding to the harvest season. This result indicates that temporary migration might be hindered by short-term liquidity constraints. However, we also find some evidence hinting that this result may be partly influenced by a non-linear effect, specifically the negative impact of excess rainfall. On the other hand, poor rainfall conditions at origin are associated with increased levels of temporary migration in the months following the main harvest season, and corresponding to the off-season. Surprisingly, these effects are primarily driven by rural-to-rural movements. Additionally, the effect of rainfall conditions at destination remains positive throughout the agricultural year. This suggests that destinations with relatively unfavorable conditions become less appealing to potential temporary migrants.

This chapter contributes to various strands of literature. It connects to an extensive body of research that has investigated households' coping strategies in the face of income variability, by studying temporary migration dynamics as a response to climate variability. In doing so, it concurrently enriches the climate migration literature. which has primarily focused on the impact of weather anomalies and climate trends on long-term permanent migration choices. It also resonates with a set of studies that have highlighted the benefits of and constraints to temporary migration amid income seasonality. Lastly, it supplements a growing collection of studies that integrate mobile phone data with environmental indicators to delve into human mobility patterns triggered by environmental shocks.

Chapter 4 builds upon the novel perspectives on human mobility presented in Chapter 1, examining the role of visits as a mechanism to access urban markets, relative to other forms non-permanent movements. I leverage the CDR dataset from Senegal to observe the mobility choices of a large number of phone users. The unique strength of CDRs, combining high observational frequencies with extended periods of user activity, allows me to simultaneously track both the visits and temporary migration decisions of individual phone subscribers. I draw on the methodology developed in Chapter 2 to gauge temporary migration and introduce a straightforward method to identify visits within CDR trajectories.

First, similar to Chapter 1, I document visiting patterns in Senegal over a period of one year. The higher coverage of rural areas, coupled with the heightened frequency and duration of observations, affords a more holistic view of these movements compared to the insights provided by smartphone app location data. I find that 83% of users make at least one visit to a city over the year. I can confidently estimate the frequency and duration of individual visits: the median visitor register a visit every 1.3 month, and each visit lasts for 1.5 days on average. The inclination to visit cities is consistently high, regardless of the place of origin, from large urban center to the most sparsely populated rural areas. However, as we transition to areas with lower population density, both the frequency of visits and the cumulative days spent in cities increase significantly. Intriguingly, this finding reinforces the hypothesis set forth in Chapter 1, suggesting that non-urban residents visit cities to consume a broadly defined urban amenity unavailable in their home location. By contrast, a mere 17% of users register a temporary migration event of at least 20 days.

Second, I conduct an empirical analysis to discern the relation between visits and temporary migration choices of individuals facing similar movement costs. In practical terms, I run regressions linking the total number of visits to cities by a user over a year with a binary variable indicating the occurrence of a temporary migration event within the same period. The results illuminate a positive association between these two forms of mobility: temporary migrants register an extra 17.5 days of urban visits per year compared to non-migrants. Subsequent analysis reveals that this association is almost entirely driven by supplementary visits undertaken by temporary migrants to their eventual migration destination. Delving deeper into the temporal dynamics of these supplementary visits in relation to the timing of temporary migration events, findings suggest that temporary migrants undertake more visits to their chosen destination in the weeks leading up to their actual migration. This pattern hints at the existence of anticipatory behaviors wherein prospective migrants would shoulder the affordable costs of visits in order to gain

information about the destination and mitigate potential risks of migration failure. The results also point to post-migration behaviors, indicating that temporary migrants frequently revisit their previous destinations after they have returned to their original location.

Finally, this chapter takes advantage of the joint observation of both visits and temporary migration to probe into potential disparities in the costs tied to each form of mobility. I assume that a transient move to a city, be it a visit or a temporary migration, is tied to a cost structure comprised of fixed costs specific to each mobility type, a distance-related cost, and a destination-specific cost related to the duration of stay. Incorporating this cost configuration into a rudimentary conceptual framework yields formulas for the frequency and duration of both visits and temporary migration episodes. Importantly, these relations underscore that the distance elasticity of mobility choices, which can be simply estimated with observed movements, does not directly correspond to the incremental cost of distance. Instead, it is inversely related to the fixed costs tied to the corresponding type of mobility. Drawing upon this conceptual foundation, I estimate gravity regressions. The findings corroborate the primary predictions of the model: while the frequency of both visits and temporary migration declines with distance, their duration increases, and the aggregate time spent exhibits a decrease. However, the disparities observed in the magnitude of these elasticities suggest that temporary migration is marked by elevated fixed costs, positioning visits as a comparatively affordable alternative.

The paper contributes to a strand of literature studying the causes, consequences and barriers to accessing urban markets via temporary migration. While previous studies have highlighted the lack of information about urban destinations as a significant deterrent to rural-urban temporary migration, this paper reveals that a considerable segment of non-migrants gain direct exposure to urban markets via regular visits. This work further resonates with existing research suggesting that substantial fixed costs, encompassing psychological barriers stemming from separation from familiar social networks and the apprehensions surrounding potential migration failures, serve as impediments to temporary migration. The paper closely align with Chapter 1 by confirming that visits are a comparatively more affordable alternative allowing individuals to access urban markets.

The thesis closes with a concluding section that summarizes the key insights from all four chapters and proposes potential directions for future research.

Chapter 1

High-Frequency Human Mobility in Three African Countries

Joint with Douglas Gollin and Martina Kirchberger.

Contribution

My contribution to this study includes cleaning and processing the data, building mobility indicators and datasets for regression analyses, performing analyses and producing figures and graphs, and co-writing the manuscript.

1.1 Introduction

Understanding human mobility patterns in low-income contexts has previously been limited by the lack of data. Census data and standard household surveys seek to capture migration flows between survey waves, but these data sources offer little information about movements that do not involve changes in an individual's home location. In a number of recent studies, survey instruments have been designed to measure temporary and seasonal migration flows in low-income countries (Bryan, Chowdhury, et al., 2014; Lagakos et al., 2023; Imbert and Papp, 2020a). For high-income economies, a few surveys provide detailed commuting data (e.g., the American Community Survey), but these normally miss non-work trips. Moreover, such surveys are not available for most low-income countries. Newer sources of "big data" have allowed researchers to construct more fine-grained measures to characterize migration and commuting behaviors for low-income economies (Blumenstock, 2012; Blumenstock, Chi, et al., 2022; Kreindler and Miyauchi, 2023). Migration inferred from such data is informative about human mobility over longer time periods, and commuting data offer insights into a specific type of daily travel. We know little, however, about human mobility within developing countries over other time scales.

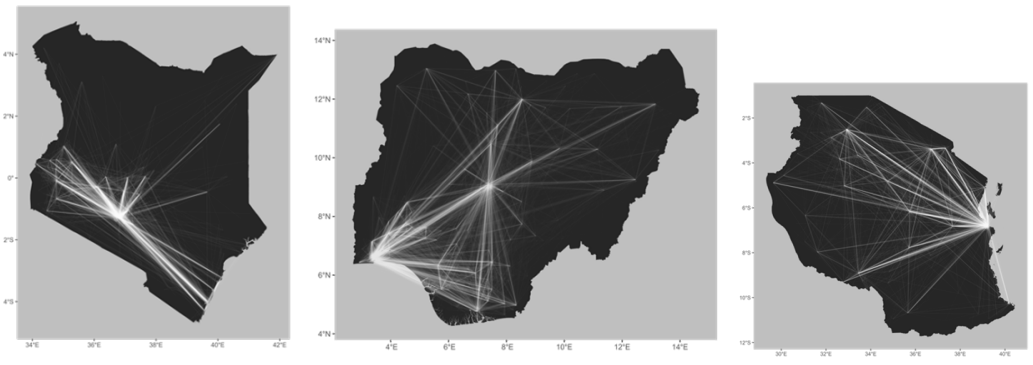
In this paper, we bring new data to the study of a type of mobility that has previously been difficult to capture. Specifically, we examine what might be characterized as "visits": the movement of people from their home locations to other locations, not necessarily for daily work. By using a new source of data and defining a novel set of metrics to measure phenomena that were previously difficult to characterize, we follow examples such as Henderson, Storeygard, et al. (2012) or Akbar et al. (2023). We find in our data that "visits" are in fact an important form of mobility. In a theoretical sense, trips between rural and urban locations (or between smaller cities and larger ones) may allow people to benefit from the amenities of large cities without migration. With short visits to cities, people from rural areas and small towns may be able to manage administrative and legal matters, enjoy consumption goods that are unavailable elsewhere, and perhaps also to purchase or consume market goods and services without having to pay costs to traders and middlemen. We know anecdotally that this kind of mobility is both important and ubiquitous; anyone who spends time at a bus station in Accra or Arusha can see first-hand the numbers of people in motion. But we have hitherto had little ability to quantify these flows or to understand their patterns.

To measure mobility, we use newly available, fine-grained, anonymized data on smartphone locations. Each observation in our data reflects an instance when a user's phone connects to the internet to use a particular app. For each such use, we observe the GPS location and the precise time. We use the data to map and categorize the movements of people and the connectedness of locations. Unique to our study is the scale at which we can study the phenomenon of short-term population movements. Our raw data covers more than one million smartphone devices over an entire year across three large African countries: Nigeria, Kenya, and Tanzania.[1] We are therefore able to present novel evidence on high-frequency mobility for large numbers of people, and at high spatial and temporal resolution. We show that this type of mobility is both substantial and prevalent.

The paper makes three main contributions. First, since we study a new type of mobility, we start by defining a novel set of metrics for characterizing mobility across space related to frequency, spatial extent, and destination characteristics. Our metrics are parsimonious and easily interpretable across different contexts, yet paint a rich picture of the extent of spatial mobility and the interconnectedness of locations. Second, we analyze these measures to provide insights into the patterns of human mobility within the three countries where our data originate. We can ask how frequently residents of a particular location pass through a given city or market centre; or how the composition of visitors to the capital differs from

[1] In the remainder of the paper we will refer to a device as a user. We recognize that this is an inexact equivalence: some users possess more than one device, and some devices are shared by multiple users. We address these issues in detail in Section 1.3.

Figure 1.1: Mobility flows to cities.



(a) Kenya (b) Nigeria (c) Tanzania

visitors to secondary cities. Within cities, we can examine the types of destinations where visitors are seen. Third, we develop a conceptual framework in which individuals decide what locations to visit. The framework delivers a number of testable propositions that, for example, relate the duration of visits or the distance travelled. To our knowledge, this is also one of the first papers using smartphone app location data in the context of low-income countries.

To provide a first glance at our data, Figure 1.1 shows visits from every spatial grid cell outside the city perimeter to any city of more than 50,000 residents in each of our study countries. The brightness of lines reflects the counts of distinct visits. It is immediately obvious that the largest cities in each of these countries draw in visitors from all over the populated areas of these countries, suggesting a strong connectedness of cities with their hinterlands. But it is also striking that there are many other lines linking secondary cities and other locations to one another.Our paper digs deeply into these connections and suggests a need to think in more nuanced ways about spatial frictions and patterns of mobility.

Capital cities of course attract disproportionate flows; but political centrality by itself is less of a driver than urban primacy; this is well illustrated in Tanzania's map, where Dar es Salaam (on the east coast) acts a clear magnet. By contrast, the capital, Dodoma (located towards the center of the country), is little different from other secondary cities in terms of incoming visitors. Our metrics allow us to quantify such patterns and to investigate the connectedness of locations at national scales.

The strength of our approach is that we are able to make clear and objective observations that match people to the locations they have visited, covering a large sample over a lengthy time period, without relying on recall data. These metrics can be easily applied in other contexts when similar data are available. Although the smartphone users whom we observe are in no way representative of the entire

population, we can characterize this set of people with reasonable accuracy. We interpret our results as broadly representative of mobility within the populations of smartphone users in each of our countries, and we develop a number of methods that allow us to characterize in great detail the similarities and differences that our sample shares with the general population of each of the three countries. Smartphone owners accounted for a significant fraction of the urban population in each of our three countries at the time period under study, ranging from 23 percent of the urban population in Nigeria to 51 percent in Kenya.[2] Given the virtual absence of data on this type of mobility for entire populations, we argue that our results represent a useful contribution. They provide insights into high-frequency mobility within a substantial fraction of the overall population – and a subset that is worthwhile and informative to study. While not the primary objective of this paper, the methods we develop to examine and characterize selection could easily be applied to similar digital trace data.

Our analysis finds striking evidence of a high degree of mobility within our samples for each of these three African countries. Our smartphone users are highly mobile. Users are seen more than 10km away from home on about one-sixth of the days on which they are observed. Residents from more sparsely populated areas are more frequently away from home than city center residents, and our users with rural home locations venture farther when they leave home. Spatial transition matrices show that towns and many villages in these countries appear to receive visits from urban dwellers, and in turn these villages generate travellers who venture to larger towns and cities. The networks of connectivity between different geographies are strong. This challenges, for instance, the notion that villages and towns in rural areas are effectively isolated; at least some (relatively prosperous) residents are maintaining regular connections to more densely populated locales.

Beyond these qualitative findings, we show that large cities exert a disproportionate influence: Nairobi, Lagos, and Dar es Salaam are powerful magnetic forces that pull in visitors from every corner of their countries, while secondary cities appear to be substitutes for each other. Finally, we show that high-frequency mobility follows specific patterns consistent with the propositions from our conceptual framework: first, the number of visits per person made from a smaller settlement to a larger one will exceed the number made in the opposite direction. Second, the fraction of days users spend visiting a city follows a gravity-style equation. Third, given a choice between visiting two equidistant locations, individuals more frequently visit the more populous destination.

This paper contributes to three main strands in the literature. First, our

[2] If "feature phones" are included (i.e., phones that have some limited ability to connect to particular apps), the numbers range from 36 percent of urban users in Tanzania to 63 percent in Kenya.

primary contribution is methodological, in proposing key metrics that allow us to characterize the extent of high-frequency mobility. Digital trace data, similar to ours, have been used, for example, to study the length of time that individuals spend with their families for Thanksgiving in the US (Chen and Rohla, 2018), to construct a measure of experienced segregation (Athey et al., 2021), to study the effect of chance meetings on knowledge spillovers in the Silicon Valley (Atkin, Chen, et al., 2022), to measure the effectiveness of social distancing (Mongey et al., 2021), social interactions (Couture et al., 2022) and the importance of travel along trip chains (Miyauchi et al., 2022). We add to this literature by focusing on three countries in sub-Saharan Africa and by looking at patterns of mobility across cities.

Second, we relate to a literature using quantitative spatial models (Monte et al., 2018; Owens et al., 2020; Ahlfeldt et al., 2015; Dingel and Tintelnot, 2023; Kreindler and Miyauchi, 2023). While our model focuses on visits, our conceptual framework also predicts a gravity-style equation, in flavor similar to the familiar gravity equations employed in this literature.

Third, our findings relate to a growing literature in economics that documents large gaps in nominal wages and productivity across sectors and in developing countries (Gollin, Lagakos, et al., 2014). There are similarly large gaps in living standards across space, with people in sparsely populated rural locations consistently worse off than those in dense urban settlements (Gollin, Kirchberger, et al., 2021). The persistence of these gaps raises the possibility that significant frictions and market imperfections limit the movements of people and information, leading to spatial and sectoral misallocation (Bryan and Morten, 2019; Brooks and Donovan, 2020; Caselli and Coleman II, 2001; Eckert and Peters, 2022; Lagakos et al., 2023). In contexts where spatial frictions are high, the allocation of factors across firms will tend to result in gaps in marginal products. Similarly, spatial frictions may lead to allocations such that marginal utilities are not equalized across consumers, and utility may not be equalized across people living in different locations. These static effects may also lead to dynamic impacts, as frictions move the economy away from a theoretically efficient benchmark.[3] By examining the frequency with which individuals move across space – from rural areas to towns and villages, or between cities – we inform this debate by assessing the potential salience of different frictions. For instance, a world in which people travel frequently between cities, or between rural and urban locations, is unlikely to be one in which the costs

[3]The importance of within-country spatial frictions in the movement of goods has been documented in recent work (e.g., Arkolakis et al. (2012), Costinot and Donaldson (2016), Atkin and Donaldson (2015), Donaldson and Hornbeck (2016), Donaldson (2018), and Allen and Arkolakis (2014)). This emerging literature has pointed out that spatial frictions have implications for patterns of specialization and exchange. An additional literature has documented the importance of spatial frictions as they relate to the flow of information (e.g., Aker (2010) and Jensen (2007)). Allen (2014) suggests that information frictions can compound spatial frictions.

of mobility are prohibitive.

Beyond the implications for spatial frictions, our analysis points to a number of interesting features of the data. First, the widespread prevalence of non-residents visiting cities suggests that urban areas generate benefits for a much broader set of people than their own residents and nearby commuters. Our data is consistent with a world in which people travel to cities from substantial distances – and with some frequency – to enjoy the benefits that cities provide. Second, we observe that 'visits' allow for some rural people (and the inhabitants of towns and small cities) to break down the rural-urban binary. Put differently, 'visits' allow people to achieve partial urbanization. In this sense, 'visiting' cities may substitute for migration, in the same way that rental markets allow people to solve the problems of lumpy capital purchases. The feasibility and (apparent) affordability of trips may represent an additional factor helping to explain the low rates of rural-urban migration, even in contexts where there are large differences in wages, productivity and living standards across space.[4] What is unambiguously clear in the data is the ubiquity of visits; this suggests that we should be cautious in treating rural and urban areas as entirely distinct; our data suggest that instead, they are connected by non-trivial flows of people. With the movements of people, it seems reasonable to imagine that there may also be corresponding flows of goods and information.

It would be interesting to compare what we observe in our three countries with a benchmark of high-frequency mobility patterns observed in higher-income countries where spatial frictions are less prevalent. Unfortunately, smartphone penetration rates across space within countries - and therefore the observed sample - would also be very different in these countries, making comparisons difficult to interpret. We therefore focus on analysing patterns within the three study countries.

This paper is structured as follows. Section 1.2 discusses the smartphone app data we use and how we define home locations. Section 1.3 focuses on sample selection and characterizes the sample. Section 1.4 presents our mobility indicators. Section 1.5 sketches our conceptual framework. Section 1.6 examines to what extent the data is consistent with the propositions coming out of our model. Section 1.7 concludes.

1.2 Smartphone app data

This paper draws primarily on smartphone app location data for three African countries: Kenya, Nigeria and Tanzania. We selected these countries based on data availability and on having a sufficiently high number of users in the sample. This

[4]There are of course many alternative interpretations of the frequency of trips.

section summarizes the main ways in which we process the raw data; for more detail, we refer the interested reader to Appendix 1.A.

Each observation in our data set (referred to hereafter as a "ping") represents an instance where a smartphone accesses the internet via a set of apps. Pings are sourced from a large number of apps that (with the user's permission) access location data. These apps include standard social, navigation, information and other apps, but we do not know precisely which apps, and we cannot associate specific pings with specific apps. Each ping comes from a device – i.e., a particular smartphone. For each ping we know the device identifier (i.e., a particular phone, rather than a SIM card), a timestamp and longitude/latitude coordinates of the current position, measured to an accuracy of approximately 10 meters. Each country dataset covers a period of one year between 2016 and 2018.[5]

In the remainder of the paper we refer to a device as a user, subject to the caveats already mentioned in Footnote 1 and discussed in further detail below. In this section we start by discussing how we assign home locations to users and outline how we identify and deal with irregularities in the data.

1.2.1 Home locations

We use two criteria to define home locations. First, we identify the modal 0.01-degree cell ($\approx 1.1km$ at the equator) in which the user is seen at night (between 7pm and 7am, local time). Second, we consider two additional restrictions: (a) that a user is observed for a minimum of 10 nights; and (b) that the user is at the inferred home location for at least 50% of the total nights when that user is observed anywhere. These two restrictions eliminate cases where the user is seen infrequently at night, or is seen frequently but at multiple locations. Given the central role home location plays in our analysis, we define our core sample – which we call the "high-confidence" sample – as users that satisfy both criteria. Unless specified otherwise, we use our high-confidence sample for our analysis.[6] We then carry out data cleaning procedures described in Appendix 1.A.2.

Table 1.1 shows the number of users and pings per user for our base sample of users and our high-confidence sample. Columns (1) and (2) show the number of users and average pings per user over the entire year, for those users who are observed at least once at night. The average is computed by summing over all pings and dividing by the number of users; for this sample we have on average slightly more than one ping per day per user. Columns (3) and (4) apply the two restrictions

[5]The precise time frame is 2016-12-01 to 2017-12-01 in Kenya and 2017-04-01 to 2018-04-01 in Nigeria and Tanzania. Note that these data come from before the period of the Covid-19 pandemic and do not reflect any of the subsequent lockdown restrictions.

[6]The distributions of home locations and patterns of mobility are very similar whether we use the base data or low-, medium-, and high-confidence samples.

Table 1.1: Sample and pings per user

	All		**High confidence**	
	Users (1)	Pings ratio (2)	Users (3)	Pings ratio (4)
Kenya	195,630	593	18,545	4,864
Nigeria	659,407	304	78,750	1,721
Tanzania	237,123	457	22,994	2,132
TOTAL	1,092,160	389	120,289	2,284

Note: Columns (1) and (2) show the total number of users per country and average pings per user. Columns (3) and (4) only use high-confidence users (users who are observed for a minimum of 10 nights and who are at the inferred home location for at least 50% of the total observed nights.)

to obtain our high-confidence sample. This yields a sample of just over 120,000 devices across the three countries, with an average of over 2,000 pings observed per user. Users in the high-confidence dataset are therefore seen on average 6 times per day, compared to users in the complete dataset who are seen on average slightly more than once per day.[7]

Table 1.2 summarizes user-level temporal statistics for our high-confidence users considering three different measures. The first statistic that we consider is the

Table 1.2: User-level temporal statistics by country

	Variable	**Mean**	**Median**	**Min**	**Max**
Kenya	Length of obs. (in days)	102.2	74.5	8.7	365.0
	Days seen	39.5	30.0	8.0	352.0
	Mean pings per day	99.1	9.0	1.0	20,665.4
Nigeria	Length of obs. (in days)	101.1	82.1	8.6	365.0
	Days seen	40.6	29.0	8.0	346.0
	Mean pings per day	40.2	12.9	1.0	9,585.8
Tanzania	Length of obs. (in days)	95.1	70.7	8.6	364.9
	Days seen	38.9	28.0	7.0	349.0
	Mean pings per day	51.6	10.7	1.0	14,765.6
TOTAL	Length of obs. (in days)	100.1	77.2	8.6	365.0
	Days seen	40.1	29.0	7.0	352.0
	Mean pings per day	51.4	11.8	1.0	20,665.4

Note: This table shows the duration over which we observe a user, the number of distinct days we observe a user, and mean pings per day, defined as the ratio of the total number of pings for a user divided by the number of distinct days she is seen.

duration over which we observe a particular user, defined as the number of days between the first and the last observation of that user. Second, we count the number of distinct days on which we see a particular user. The third statistic is the mean

[7] As is common with these types of data, there is a large variation in the number of pings across users, with about 59% of users having at most 20 pings in the initial sample. Our two conditions defining high-confidence users reduce the fraction of users with at most 20 pings to 0.3%.

number of pings per day per user. The mean number of pings per day is defined as the total number of pings for a user divided by the number of distinct days she is seen.[8] These statistics are roughly similar for the three countries. We see users on average over a span of about 100 days, on about 40 distinct days, and they have between 40 and 100 pings per day on average.[9] The relatively short time frame over which we observe individuals suggests that while the data is informative about the overall mobility of the population, it is not ideal for longer-term individual-level analysis, such as measuring the extent of seasonal or permanent migration.[10]

Similar to home locations, in Appendix Section 1.A.3 we have defined work locations as the modal 0.01-degree cell in which a user is observed between 9am and 6pm on weekdays, again imposing two restrictions: that (a) the user is observed for a minimum of 8 distinct weekdays and (b) is seen at the inferred work location for at least 50% of the total weekdays. We find that home and work locations are found within the same 0.01-degree cells for 80% of users, consistent with high rates of self-employment and short-distance commuting. We interpret this to mean that relatively few of the trips observed in our data are associated with daily commuting between home and work.

1.3 Selection

The key selection concern when using smartphone app location data is that we only capture individuals who own a smartphone. A further restriction affecting selection into our sample is that individuals require data credit on their phones, similar to requiring phone credit to make calls or send texts. On the other hand, as app usage is increasing through the use of messaging services (e.g., Facebook Messenger or WhatsApp), replacing "traditional" calling and texting, we are more likely to capture locations of individuals engaging in this kind of activity. Further, we are more likely to capture passive use of a mobile phone if a device connects to an app without the deliberate action of the holder of the device. This would make location detection more representative, in some sense, than relying on call and text events only which require a deliberate action. In terms of characteristics of the selected sample, we expect this to bias our sample towards richer, more educated and younger individuals.

Given these general concerns about selection, we seek to understand how our population of users compares to the broader populations of these three countries.

[8] This differs from the pings ratio in Table 1.1 which simply summed over all pings in the data across all users and divided by the number of users.

[9] The minimum number of days is less than 10 as some users are seen on 10 nights but have pings on fewer than 10 days.

[10] These are issues explored in Bryan, Chowdhury, et al. (2014), Imbert and Papp (2020a), Lagakos et al. (2023) or Blumenstock, Chi, et al. (2022) using call detail records data.

We proceed in three steps. First, we link users' locations with geo-coded population density data from WorldPop to understand how the home locations of users relate to the overall spatial distribution of population. Second, we draw on data from other nationally representative surveys – specifically, the ICT Access and Usage Surveys 2017-2018 – to examine differences between individuals who own a smartphone and those who do not. To the extent that our population of smartphone app users is typical of all smartphone owners, these survey data will tell us something about how our users compare to the broader national populations of their countries. Third, to measure how representative our users are, in terms of their home locations, we develop a methodology to match home locations with nationally representative micro-data from the Demographic and Health Surveys (DHS). This allows us to say something about whether the locations where our users live are typical or atypical.

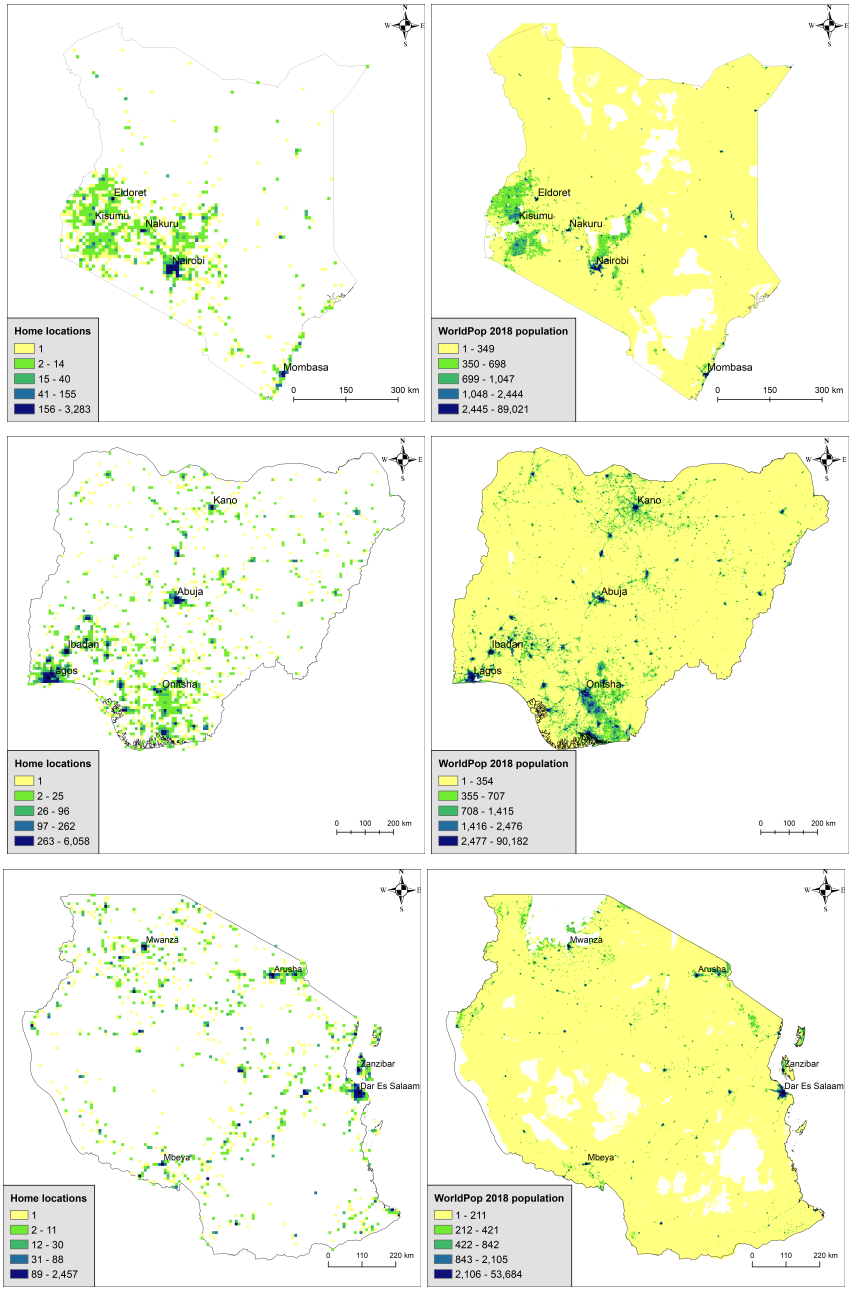
Figure 1.2 shows the distribution of home locations in the left panel and compares it with the population distribution in the right panel. Darker values indicate a higher number of users. Unsurprisingly, we observe a higher number of users in the main cities. However, the figure shows that coverage of users is broadly national, with users residing in fairly distant places as well as in the densest cities. In fact, we have users in all but three of the 115 regional capitals in the three countries we study. When looking within the three capital cities we find again that our users reside in locations spread out across these cities rather than being concentrated in a few rich neighborhoods.

To examine how representative home locations of our users are for different levels of population density, we extract the population density values at users' home locations using WorldPop population grids and we then infer the distribution of users across population density bins. The distribution of users is largely skewed to the right with around 70 percent of users falling in the two densest bins (see Figure 1.3).[11]

We compute three further metrics to measure the representativeness of our users across different levels of population density: first, we take all 10-km pixels in a country and regress the number of users in a pixel on population of the corresponding pixel. We find that the R-squared ranges between 0.36 in Kenya to 0.81 in Tanzania, depending on the source of the population density estimates. Second, we compare the rank in terms of the total number of users at the first administrative level in our three countries with the rank of the population. The bivariate correlation coefficients range between 0.29 in Nigeria and 0.7 in Tanzania.

[11]To be specific, we divide each country into gridcells and assign each gridcell an absolute population density based on WorldPop or other data. Using the national population data, we can divide the entire population into equal-sized bins based on the population density in which they live. This gives rise to a set of gridcells associated with each density decile. We can then identify each of our users with the population density and/or the density bin of their home location; e.g., we can speak of a user whose home location is in the third density decile.

Figure 1.2: Distribution of home locations and population.

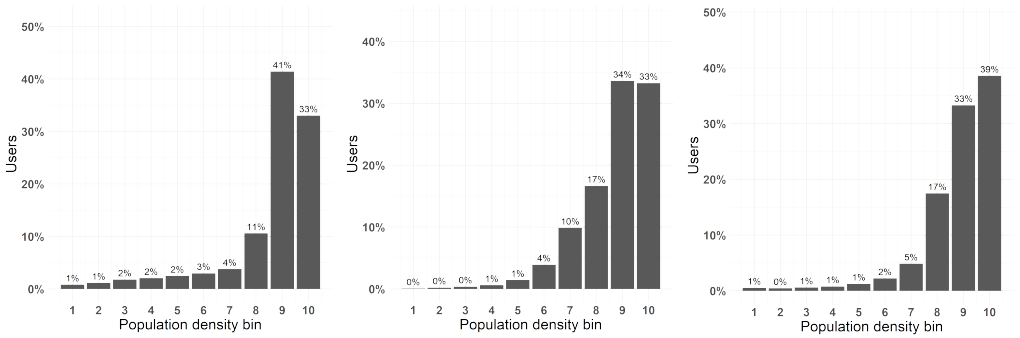


Note: This figure shows the distribution of home locations of users at a 10km resolution (on the left) and the distribution of the population at a 1km resolution (on the right).

Next, we compare the fraction of users located in cities of at least 200,000 people with the corresponding fraction of the population living in those cities.[12] In Nigeria, 86.1% of our users are found in cities of 200,000 people, whereas these are host to

[12]Our approach to defining urban peripheries is described in Appendix Section 1.B. Using 2018 as our base year, we identify 6, 39, and 10 cities of at least 200,000 people in Kenya, Nigeria, and Tanzania respectively.

Figure 1.3: Users by population density decile.



(a) Kenya (b) Nigeria (c) Tanzania

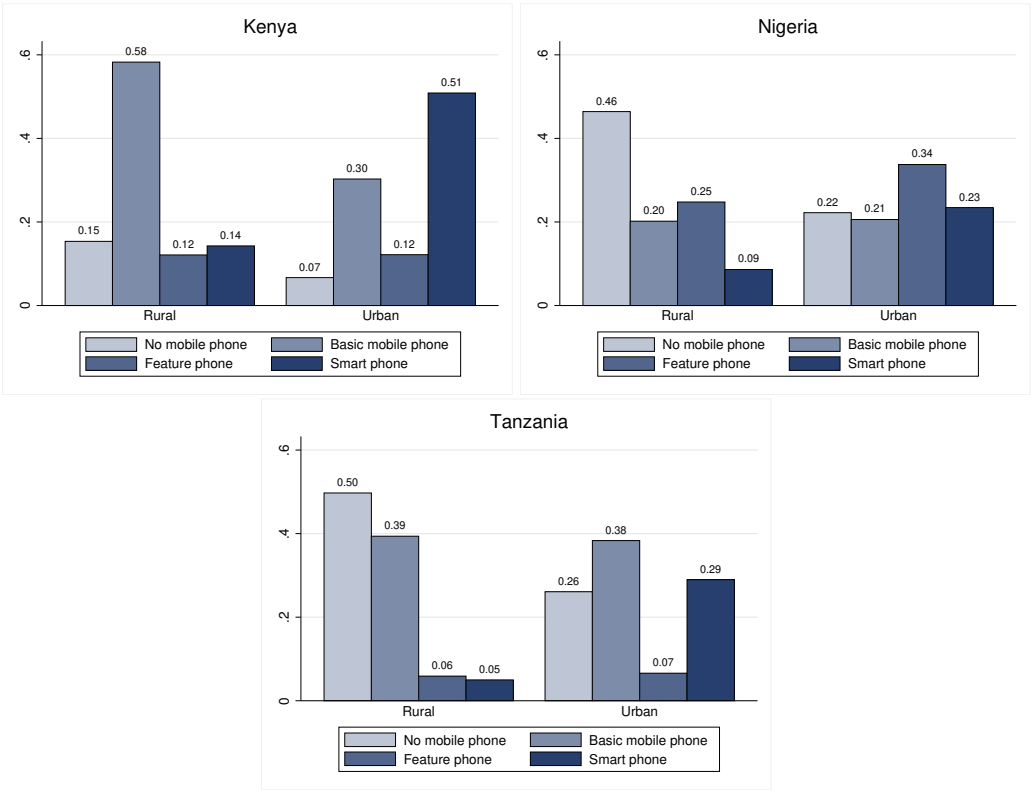
Note: This figure shows the distribution of users across population density deciles based on national population data so that each decile contains one tenth of the population (rather than one tenth of grid-cells).Appendix Figure 1.E.1 shows the same figure based on Landscan measures of density instead of WorldPop and also shows the sensitivity to our definition of the high-confidence sample.

only 20.5% of the population. Similar results are observed in Kenya and Tanzania where we find 75.9% and 68% of users in major cities that host 15.9% and 16.7% of the population respectively, which is indicative of an urban selection pattern. The urban tilt of our sample is unsurprising; we expect that smartphone users will be concentrated in cities.

To understand how this pattern is driven by differential device ownership rates across urban and rural areas, we use data from the ICT Access and Usage Survey 2017-2018 for Nigeria, Kenya and Tanzania. These surveys are nationally representative and have detailed questions on mobile phone ownership and usage, as well as individual and household characteristics. Overall, between 19 and 43 percent of the population have either a feature phone or a smartphone in our three countries.[13] Figure 1.4 shows ownership rates for different types of mobile phones, comparing rural and urban locations. Compared to rural areas, respondents in urban areas are unsurprisingly more likely to own a mobile phone, and the phone is likely to be more sophisticated. The figure shows that in all three countries, smartphone ownership is highest in urban areas, with rates between 23 and 51 percent. If we include feature phones, this increases the rate to between 50 and 60 percent. The proportion of individuals with a basic mobile phone ranges between 21 and 38 percent. Across the rural areas of our three countries, smartphone and feature phone ownership is highest in Nigeria, at 31 percent penetration, and lowest in Tanzania, with 11 percent. Figures 1.C.1-1.C.4 in Appendix 1.C examine ownership rates by gender, and explore how owners of different devices

[13] A "feature phone" is defined as one that has a small screen and some rudimentary internet access, but button-based data entry rather than touch screen. It is more complex than a "basic phone," which can only carry out simple calling and texting functions.

Figure 1.4: Device ownership by location.



Note: This figure shows device ownership rates for rural and urban respondents. All figures use the sample weights provided.

differ in terms of income, education, age and main source of income. In all three countries, women are less likely than men to own a mobile phone. While basic phone ownership rates are roughly equal between men and women, fewer women own a feature phone or a smartphone; still, smartphone ownership rates of women are between 11-20 percent in our three countries. Unsurprisingly, respondents with no mobile phones tend to have the lowest incomes and owners of smartphones tend to have the highest incomes. However, Figures 1.C.2 - 1.C.4 highlight that these distributions are not distinct. Appendix Table 1.C.1 shows the proportion of smartphone owners across different categories and compares this to the sample averages. The Table suggests that smartphone users are not just from one occupation (e.g., traders) but are represented across different types of economic activities.

Finally, the survey also asks respondents about their usage of a range of apps, including social networking apps and news, weather, trading, business, health and dating apps. Figure 1.C.5 shows that between 76 and 83 percent of smartphone owners report using an app weekly on their phones, and more than 55 percent use these apps daily, suggesting that selection due to differential usage patterns is likely less of a concern.

In a third step, we characterize the home locations of our users by drawing on

available data from Demographic and Health Surveys (DHS). The key challenges are how to link a relatively small number of DHS survey clusters (the total number of clusters ranges from 608 in Tanzania to 1,594 in Kenya) to a large number of home locations for our users, spread across the entire geography of our three countries.[14] For our analysis, we aim to match each user's home location to a nearby DHS cluster that might be considered comparable. We then compare these matched DHS clusters to the full DHS sample. Appendix 1.D provides details on how we link home locations of users with DHS clusters. Following this procedure, we are able to link 70% of our users in the high-confidence sample with at least one DHS cluster.

This matching exercise allows us to see whether the home locations of our users are atypical, relative to the nationally representative sampling frames that have yielded the DHS clusters. In other words, if we look at the set of DHS clusters where we find our users, we can ask whether this matched DHS sample looks statistically similar to the overall ("raw") set of DHS clusters. We carry out this analysis by conducting t-tests for equality of means between the raw DHS and matched DHS samples on a range of directly quantifiable household characteristics, such as whether the household has a constructed floor, walls, roof, overcrowding and access to public services such as electricity and tap piped water. Moreover, we produce results for rural and urban sub-samples separately to account for both the prevalence of urban users in our sample and the lower matching rate in low density areas, which together may lead to results being mainly driven by the urban component of the sample. We produce t-tests comparing our two weighted data sets, with bootstrapped standard errors robust to heteroskedasticity. The survey weights are used for the reference DHS sample, while those of the matched DHS sample correspond to the number of users each cluster is paired with.

Appendix Tables 1.E.1-1.E.3 show that we find statistically significant differences between the matched clusters and the raw DHS clusters. Our users live in locations that are not nationally representative. In particular, the DHS data show that individuals residing in matched clusters have smaller household size than that found in the nationally representative DHS sample. The matched clusters also have younger household heads with higher education levels, and better access to services and housing characteristics. Most of the differences are statistically significant. What we find, however, is that the absolute levels do not differ by large amounts; the differences between matched clusters and the raw DHS data are quantitatively small, especially *within* the rural and the urban samples.[15]

[14] Adding to the challenge is that the published locations for the DHS clusters are randomly displaced by a small amount in an effort to ensure data confidentiality (Perez-Heydrich et al., 2013).

[15] In almost two-thirds of rural and urban comparisons for these three categories of variables, the differences between the matched and unmatched clusters are less than 10 percent.

Our takeaway message from this analysis is that our population of users resides in more densely populated locations and is likely to be richer, more educated and younger. Within urban locations, smartphone users represent a significant fraction of the population. Given the selection biases here, we must be extremely cautious in generalizations about aggregate behavior. However, given the lack of data on the kind of mobility that we study in this paper, we feel that it is still worthwhile to study the mobility characteristics of our sample. While our samples are not nationally representative, they represent non-trivial sections of the population, and we can observe their behavior in rich detail.

To conclude this section of the paper, we return to the potential biases that we may have introduced by equating "devices" with "users". We also consider other potential challenges in working with our ping data. We acknowledge that distinct users may use the same device, and individual users might have multiple devices. Unfortunately we do not have data on the extent to which smartphones are shared among contacts. From the ICT Access and Usage Survey we know that between 20 and 35 percent who stated that they do not *own* a mobile phone say that they nevertheless *used* a mobile phone in the past three months. It is reasonable to assume that device sharing is likely to occur within households. If so, it would not affect the home locations we determined for our users, nor would it alter the characteristics of home locations we discussed.

Individuals could also have multiple phones or SIM cards. The latter problem is not a significant concern for us. Our data observe devices, rather than SIM cards; even when the SIM card is swapped, the device identifier remains the same, so our smartphone app data are unaffected. There is some reason for us to be concerned about users who own multiple devices. This would affect our results in the opposite way of device sharing, such that the movement data of these two-device-owners would get a higher weight in our mobility metric calculations. A possible additional complication would arise if a user maintains two devices, with each linked to a different location or set of locations. This would make a highly mobile user look artificially as though she does not move very much. For example, someone who commutes each week from home in a rural area to work in a big city, using a different device in each location, will appear as a relatively immobile individual. Unfortunately, we do not have information on the extent to which users own multiple devices, but given that smartphones are relatively expensive – and given the attachment that people feel to particular devices – it is likely to be a rather small number.

One other issue with the ping data is that, for many purposes, we may want to exclude incidental pings – such as those made by a person in transit. Someone traveling by road between two locations may appear to have 'visited' a location when

in fact she simply passed by in a bus or train. This requires distinguishing between locations that were deliberately visited and those that appear to be incidental. In particular, the use of navigation apps might skew the distribution of pings towards low density areas that users are simply passing through but not deliberately visiting. This is particularly relevant for our metrics that categorize destinations by their population density. In Appendix 1.A.4 we describe a filtering algorithm that we developed to identify transit pings. In general, we find that relatively small fractions of pings appear to be 'transit pings'. In the analysis that follows, where our descriptive statistics are most susceptible to being distorted by transit pings, we show the robustness of our results to removing transit pings.

Finally, we note that users may not leave their devices turned on at all times, they might not always have coverage, and they may not connect with apps during all of their travels (e.g., if data charges are high). This would lead to a systematic underestimation of the frequency of travel and the distance travelled. With all these caveats, however, we proceed to analyze the mobility data.

1.4 Quantifying mobility

In this section, we develop and implement a number of indicators to measure high-frequency mobility patterns. We consider mobility on two levels: the mobility of individual users across locations, and the connectedness of different locations through these individual movements. We characterize mobility at the user level on four key dimensions: frequency, spatial extent, densities and specific locations visited. Our preferred indicators in this respect are the fraction of days with mobility beyond 10km away from home (*frequency*), the average distance away from home (*spatial extent*), the distribution of (non-home) pings/users across population density categories (*densities visited*), and distinct cities visited.[16] We investigate how these vary across subsets of users residing in different population density categories – for which we use population density deciles as cutoff values to define these density bins. In characterizing the connectivity of locations, we quantify incoming and outgoing flows separately. We characterize incoming mobility flows by their size, with the number of distinct visitors during the period of observation, but also by the frequency of visits to the city, the distance travelled, and the population density at visitors' home locations. Similarly, we calculate the size of outgoing flows, i.e. the number of distinct residents seen outside the city during the period, the frequency of movements outside the city, their spatial extent and the population densities visited. In addition, we provide measures of mobility flows for pairs

[16] Appendix Figure 1.E.3 and Tables 1.E.4– 1.E.5 show days with mobility and mean distance away from home for the base, low-, medium-, and high confidence sets. We find that the observed patterns are very similar.

of cities. We examine the origin locations of visitors in the five largest cities in each of our three countries, and we also look at the top destinations visited by their residents. We disaggregate both the origin and destination locations into densities and summarize our data in the form of a spatial transition matrix to examine the connections between remote and dense areas. Finally, we define visits and present evidence on the type of locations visited: flows of visitors between specific locations, number of cities visited and destinations visited within cities.

We begin by considering the frequency with which people leave their home locations. Some initial notation is helpful. Let $x \in X$ denote a location, where X is a set of 0.01-degree resolution grid cells covering the country extent. For any given user i in the set of users I, we can partition X in two ways. First, we partition X into the home location and non-home locations. Let $d_i(x)$ denote the haversine distance to location x from the home location of user i.[17] Define the distance threshold $\bar{d}$ to be the limit of the home location. Then for user i, the set of locations such that $d_i(x) \leq \bar{d}$ defines a set of locations near home, H_i. Similarly, $\bar{H}_i = \{x \in X \mid d_i(x) > \bar{d}\}$ defines a set of locations away from home. For any user i, it is true that $H_i \cup \bar{H}_i = X$.

A second useful way to partition X for a given user i is into the subset of locations (typically a strict subset) where user i is observed with a ping and those where the user is not observed. We use Z_i to represent the set of locations where we observe a ping from i during the period of observation, and we in turn partition Z_i into those locations near i's home location - as defined by $\bar{d}$ - denoted Z_i^H and those that are considered away from home, denoted $Z_i^{\bar{H}}$. In addition, we denote by Z_{it} the set of locations where we observe a ping from i on any given day t and that we can partition into Z_{it}^H and $Z_{it}^{\bar{H}}$.

As a final notational preliminary, define an integer-valued function $p_i(x)$ that counts the number of pings for user i in each location $x \in X$. Clearly, $p_i(x) \geq 1$ for $x \in Z_i$, and $p_i(x) = 0$ elsewhere. Let $P_i = \sum_{x \in X} p_i(x)$ give the total number of pings for user i.

1.4.1 Frequency

As our first measure, we use the fraction of days a user is seen more than 10 km away from her home location (i.e., we set $\bar{d} = 10$km). Let M_{it} be a mobility indicator such that $M_{it} = 1$ on any day, t, if there is at least one ping observed for person i at a location away from home; i.e., $Z_{it}^{\bar{H}} \neq \emptyset$. Define $M_i = \sum_{t=1}^{365} M_{it}$ to be the number of days the user is seen more than 10 km away from her home location. Similarly, let

[17] Strictly speaking, we use the haversine distance between 2-decimal rounded latitude-longitude locations. This is equivalent to taking the haversine distance between the centroids of two narrowly defined grid cells.

T_{it} be a dummy indicating whether at least one ping is observed for person i at any location on day t; i.e., $T_{it} = 1$ if $Z_{it} \neq \emptyset$; and let $T_i = \sum_{t=1}^{365} T_{it}$ be the number of days over the period of study where at least one ping from user i is observed. Then we define the mobility frequency for user i as:

$$F_i = \frac{M_i}{T_i}. \tag{1.1}$$

In this expression, the numerator denotes the number of days with at least one ping 10 km away from home for user i, and the denominator gives the total number of days on which user i is observed (i.e., days with at least one ping). We find that the fraction of days on which users are more than 10km away from home ranges from 11.8 in Tanzania to 15.2 in Nigeria. A limitation of this metric is that it does not allow us to distinguish between users making a lot of short trips and those travelling less but spending more time at their destinations, something we consider in Section 1.4.4.

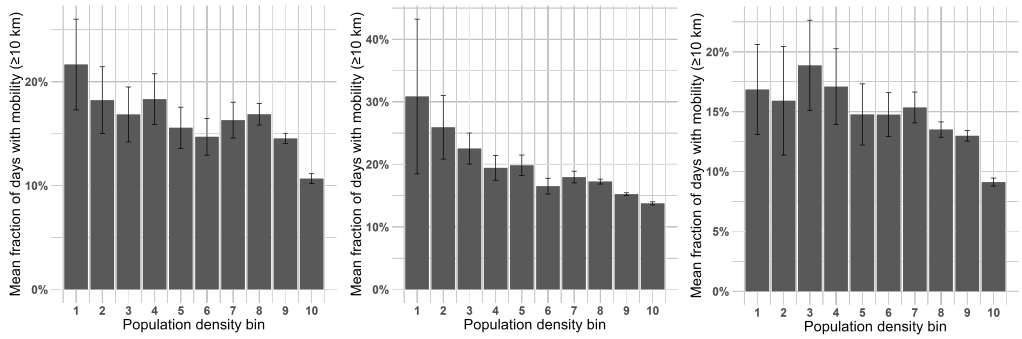
To translate this individual measure into a characteristic of a group of people, we average across the members of that group. For this, it is useful to define some groups of people. As noted above in Section 1.3, we assign each user to a population density bin, based on the characteristics of the user's home location. For instance, we consider the set of decile-bounded bins, $B = \{b_1, b_2, ..., b_{10}\}$, and we define the corresponding subsets of users $I_1, ..., I_{10}$. Let n_j denote the number of users assigned to bin b_j, i.e. the number of users in I_j. We then compute:

$$F^j = \frac{1}{n_j} \sum_{i \in I_j} F_i. \tag{1.2}$$

Figure 1.5 shows this frequency for all three countries, broken down by density bin. The pattern is consistent across countries: on roughly 12-15 percent of the days when we observe them, users appear beyond the 10 km radius from their home locations. There is a distinct pattern, too, in that those who live in the most densely populated areas are the least likely to be observed away from home. We also calculate the fraction of days with mobility beyond 20km and observe similar and even more marked patterns. One plausible interpretation is that those who live in relatively remote areas are likely to travel more frequently than those who live in towns and central cities. We cannot, of course, distinguish between the frequency of trips and the frequency with which users turn to their phones for information. It is possible that users are more likely (or less likely) to use their devices when they are travelling, compared to when they are home; and these patterns may differ for people whose home locations are in different bins of population density. Nevertheless, the data are suggestive both of a relatively high overall frequency of mobility and of differences between rural and urban residents.[18]

[18] As a robustness check, we reproduce Figure 1.5 with truncated means; that is, we discard values

Figure 1.5: Fraction of days with mobility beyond 10km by density bin.



(a) Kenya (b) Nigeria (c) Tanzania

Note: This figure shows the fraction of days on which a user is seen more than 10km away from their home location by density decile over the period of a year.

1.4.2 Spatial extent

We define the spatial extent of mobility for user i as the average distance between non-home pings and the home location. Note that for this metric, we take $\bar{d} = 0$ to define the sets of home locations and non-home locations, H_i and $\bar{H}_i$. As before, let $p_i(x)$ be the number of pings we observe for user i at location x. Then let $P_{iH} = \sum_{x \in H_i} p_i(x)$ and $P_{i\bar{H}} = \sum_{x \in \bar{H}_i} p_i(x)$; consistent with our notation above, the total number of pings observed for user i is simply $P_i = P_{iH} + P_{i\bar{H}}$. In simple terms, $P_{i\bar{H}}$ is the number of non-home pings of user i.

Given this, we can construct the spatial extent of user i's mobility, which is the average distance to each of her non-home pings. Thus:

$$S_i = \frac{1}{P_{i\bar{H}}} \sum_{x \in Z_{i\bar{H}}} d_i(x) p_i(x). \tag{1.3}$$

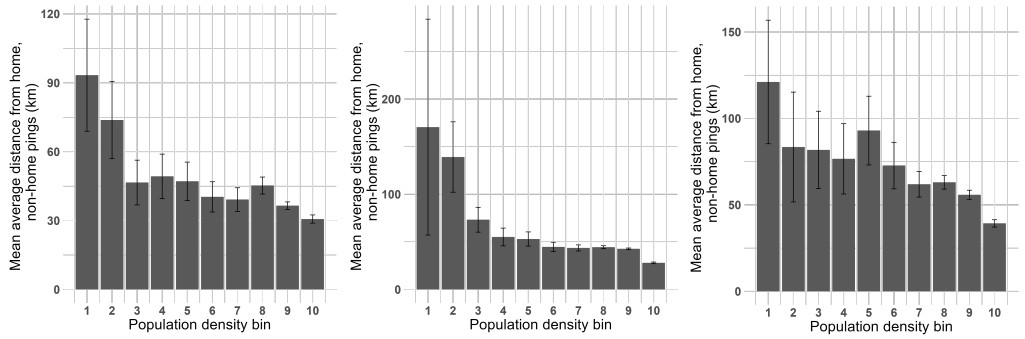
We find that the average distance of non-home pings ranges from 37.1 km in Kenya to 52.2 km in Tanzania. In extrapolating this measure to a group of people, we can once again take an average. For example, we can measure the average of our spatial extent measure for the individuals belonging to a population density bin b_j by simply averaging the individual values of S_i. Thus:

$$S^j = \frac{1}{n_j} \sum_{i \in I_j} S_i. \tag{1.4}$$

Figure 1.6 shows that non-home pings are not all highly local. In fact, the average distance – across countries and density bins – ranges from 30 km to above 100 km.

in the top 5 percentiles, to address the concern that the results could be driven by a small set of highly mobile users. We observe small decreases in the average fraction of days away in all density bins but no change in the overall pattern of decreasing frequency with population density.

Figure 1.6: Mean distance away from home by density bin.



(a) Kenya (b) Nigeria (c) Tanzania

Note: This figure shows the average distance from users' home locations of non-home pings by density decile over the period of a year.

As in Figure 1.5, we see a pattern across density bins suggesting that those in relatively sparsely populated areas seem to travel the farthest – in the sense that their average distance away from home (conditional on *being* away from home) is higher than for those in more densely populated locations. It is interesting that both the absolute distances and the relative patterns across density bins look quite similar across the three countries.

Taken together, Figures 1.5 and 1.6 seem suggestive of a pattern in which those from relatively remote areas travel more frequently and farther – possibly to get to towns and cities. To assess this conjecture, we next turn to the third dimension of mobility and construct a first measure that allows us to characterize locations visited by users in terms of population density.

1.4.3 Densities visited

Let $N(x)$ denote the population density at location x. Based on this, let $\tilde{N}(x)$ be an indicator mapping locations into density bins; in other words, $\tilde{N} : X \to B$. We consider the set of non-home locations pinged by person i, and we assign each ping to a density bin b_j. Then the fraction of pings in non-home locations by user i to locations in density bin b_j is given by:

$$v_{ij} = \frac{\sum_{x \in \{x \in \bar{H}_i : \tilde{N}(x) = b_j\}} p_i(x)}{P_{i\bar{H}}} \tag{1.5}$$

Once again, we summarize our measure at the level of each group I_o of users with home location in density bin of origin b_o by calculating the average fraction of non-home pings in each one of the 10 density bins of destination $(b_d)_{d \in [1;10]}$. Then our measure becomes:

$$V_{od} = \frac{1}{n_o} \sum_{i \in I_o} v_{id}. \tag{1.6}$$

From this, we construct an aggregate metric at the density bin level to describe the population densities visited at least once by users belonging to each density bin b_j. For each user $i \in I_j$ and each density bin b_k, we define p_{ik} as a dummy indicating whether user i ever visited a location in density bin b_k:

$$p_{ik} = \begin{cases} 1, & \text{if } \exists x \in \{x \in Z_i^{\overline{H}} | \tilde{N}(x) = b_k\} \\ 0, & \text{otherwise} \end{cases}$$

Then the fraction of users whose home location is in density bin b_j and who are seen at least once in a location belonging to population density bin b_k is:

$$\Delta_{jk} = \frac{\sum_{i \in I_j} p_{ik}}{n_j}. \tag{1.7}$$

Table 1.3 shows the results for the mobility measure Δ_{od} and thus provides more detail about the locations visited by people when they are away from their home location.[19] This table gives the fractions of users residing in a given density bin who are seen over the course of the observation span on at least one occasion in a non-home location within each of the ten density bins. For instance, this tells us that 6.7% of those Kenyans living in the most densely populated locations in the country were observed on at least one occasion during the year in a cell that falls within the *least* densely populated parts of the country. At the other end of the distribution, 32.1% of the users whose home locations are in the most sparsely populated areas of the country were observed at least once during the year in the most densely populated parts of the country. These results hold even after filtering out potential "transit pings" as discussed in Section 1.A.4 (for details, see Appendix Tables 1.E.7 and 1.E.8). Taken together, these tables offer a picture of highly mobile populations across all three countries, with people travelling both far (measured in terms of distance) and to locations that differ markedly from their home locations.

1.4.4 Specific locations visited

As an alternative to using density deciles for our analysis, we consider in Appendix Table 1.E.9 the "visitors" to the major cities of our three countries. A visitor is defined here as someone whom we observe in a city whose home location falls outside the city boundaries. We categorize visitors as those who are residents of other major cities in the same country, and then we also consider a group of "non-urban" visitors, who are those who live outside the boundaries of any city of more than 200,000 people.[20]

[19] Results for V_{od} (the average distribution of non-home pings across density bins) are shown in Appendix Table 1.E.6 for our three countries.

[20] See Appendix Section 1.B for the definition of city boundaries. The reference year for city-level population counts is 2018.

Table 1.3: Share of users by home bin-visited bin pair, no adjustment for transit pings.

		Home density bin									
		1	2	3	4	5	6	7	8	9	10
	1	72.3%	32.9%	15.1%	11.8%	11.9%	14.7%	13.3%	15.1%	9.5%	6.7%
	2	42.9%	61.4%	38.1%	26.9%	21%	17.5%	18.6%	20.9%	15%	11.4%
	3	25.9%	46.2%	55.5%	43.8%	35.8%	29.6%	28.8%	25.5%	19.4%	14.7%
	4	33.9%	34.2%	52.5%	56.6%	46.9%	39.4%	35.3%	29.6%	23.7%	17.8%
Visited	5	30.4%	25.9%	43%	52.2%	53.6%	49.3%	38.8%	35.7%	25.5%	18.9%
density	6	27.7%	27.2%	30.2%	47.1%	46.6%	55.5%	47.4%	38.3%	26.5%	19.7%
	7	26.8%	28.5%	35.5%	44.8%	45%	56.5%	57.9%	48.5%	35%	24.5%
	8	42%	44.9%	45.7%	56.9%	57.1%	60.8%	68.4%	69.7%	50.8%	36%
	9	55.4%	54.4%	53.6%	66%	65%	67.8%	72.1%	79.8%	89.8%	76%
	10	32.1%	36.1%	30.6%	41.4%	37.5%	40.9%	45.8%	51.7%	70%	88.6%

(a) Kenya

		Home density bin									
		1	2	3	4	5	6	7	8	9	10
	1	35.7%	19.6%	18.8%	6.8%	6.1%	3.8%	3.3%	3.1%	2.5%	1.5%
	2	23.8%	33.3%	35%	12.1%	12.6%	9.1%	6.8%	6.1%	5.3%	3.2%
	3	26.2%	29%	41.5%	32%	18.9%	13.1%	10.5%	8.8%	7.3%	4.7%
	4	31%	26.8%	45.3%	35.2%	32.6%	22.3%	15%	12%	11.2%	6.9%
Visited	5	23.8%	33.3%	43.6%	45.9%	51.3%	38.1%	27%	21%	20.1%	15.2%
density	6	33.3%	33.3%	37.6%	53.9%	60%	68.7%	45.8%	31.5%	26.8%	17.4%
	7	42.9%	55.8%	50.9%	52.7%	64%	69.9%	76.1%	56.1%	39.8%	25.5%
	8	71.4%	58.7%	54.7%	58.7%	61.5%	60%	72.8%	81.2%	63.7%	37.9%
	9	76.2%	61.6%	62.8%	62.6%	66.8%	64.1%	68.4%	81.2%	91.5%	64.7%
	10	42.9%	44.9%	43.2%	44.7%	50.2%	47.8%	46.9%	46.9%	61.9%	95.3%

(b) Nigeria

		Home density bin									
		1	2	3	4	5	6	7	8	9	10
	1	73.6%	33.8%	18.2%	15%	15.2%	10.3%	11.6%	9.5%	7.9%	4.4%
	2	18.7%	50%	40%	29.3%	22.6%	15.2%	14.9%	11.6%	9.1%	5.4%
	3	13.2%	39.7%	43.6%	38.8%	30%	20.1%	15.4%	13.7%	10.6%	6.3%
	4	14.3%	38.2%	40.9%	42.2%	39.2%	24.2%	20.7%	15.8%	12.3%	7.7%
Visited	5	16.5%	33.8%	42.7%	40.8%	36.9%	43.4%	27.2%	20.3%	14.3%	8.3%
density	6	19.8%	26.5%	35.5%	41.5%	46.5%	51.2%	42.2%	24.5%	17.7%	10.6%
	7	30.8%	38.2%	44.5%	46.9%	41.9%	54.2%	64.4%	42.6%	26.5%	15.8%
	8	42.9%	44.1%	50%	47.6%	51.2%	55%	62.6%	82.6%	56.9%	33.6%
	9	40.7%	51.5%	54.5%	48.3%	55.8%	59.1%	56.1%	68.7%	88.4%	66.2%
	10	40.7%	35.3%	31.8%	38.1%	32.7%	38.8%	39.7%	45%	64.5%	93.5%

(c) Tanzania

Note: These matrices show the proportion of users residing in home density bin i that are seen at least once in visited density bin j over the period of a year.

The data for all three countries show similar and interesting patterns. The largest city consistently has a large number of visitors defined as "non-urban", implying that these cities are magnets for travellers from the entire country. There are consistently large flows from secondary cities to these primate cities, but the proportions fall off sharply to more minor cities. In contrast, the secondary cities typically see large inflows of visitors from the primate cities, along with large inflows from non-urban areas. The flows across and between secondary cities are typically fairly modest, according to this metric. In Kenya, Eldoret has little that

Kisumu lacks, and vice versa – so even though these cities are less than 150 km apart, each accounts for less than 3% of the visitors in the other. The same patterns are seen in Nigeria and Tanzania. For Nigeria, to give another example, although visitors from Kano make up 10% of the documented visitors to Kaduna, relatively few of those visiting Kano are from Kaduna. In each city, far more visitors come from towns, villages, and rural areas (together characterized as "non-urban").[21] A striking feature of these tables is that the largest city is the leading destination for those living in almost all other cities – regardless of distance. Curiously, urban dwellers are also relatively likely to have been seen in non-urban areas. This is suggestive of the possibility that secondary cities are relatively substitutable for one another, but the largest cities (and perhaps also non-urban areas) offer benefits that are somehow distinct. This may reflect a lack of specialization and differentiation between secondary cities – an issue that has been raised previously in sub-Saharan Africa (see, for example, Henderson and Kriticos (2018)).

As a final step in our characterization of mobility, we examine in more detail the number of distinct visits individuals make as well as what type of amenities the data suggest people consume when making these visits. Appendix 1.A.5 provides the details on how we define visits. Figure 1.7 shows the distribution of users by number of cities that they visit (excluding the home cities of urban residents). The figure shows that a sizeable fraction of residents make visits to one or more cities other than their own during the period over which we observe them. Rural residents are again more likely to make a visit to a larger number of cities.

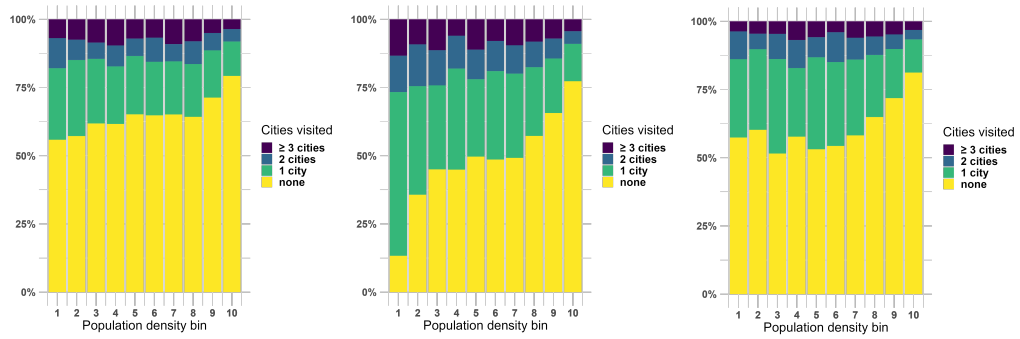
To what extent are visits to cities events that occur as an exception rather than journeys individuals embark on with some regularity? Figure 1.8 shows the average number of visits to cities users make, again by density decile. The data shows that users make multiple visits to non-home cities on average, further supporting the view that visits represent a technology to consume amenities on repeated occasions that these cities offer but home locations do not.

While we can not know the type of amenities that are consumed on visits nor the precise purpose of a visit to a particular location, in a final step we inspect the locations that visitors to cities are seen at. To investigate these destinations systematically, we link our ping locations with data from Open Street Map polygons for six cities, two from each country: Lagos, Abuja, Nairobi, Mombasa, Dar es Salaam and Dodoma.[22] We then pool all these pings and show the types of places visited for these six cities.

[21] We can similarly look at the destinations of those whose home locations are in the major cities of our three countries. For these urban dwellers, we can ask what proportion were seen during the year in other major cities and in non-urban areas. The results of this analysis are shown in Appendix Table 1.E.10.

[22] See Appendix 1.A.6 for details.

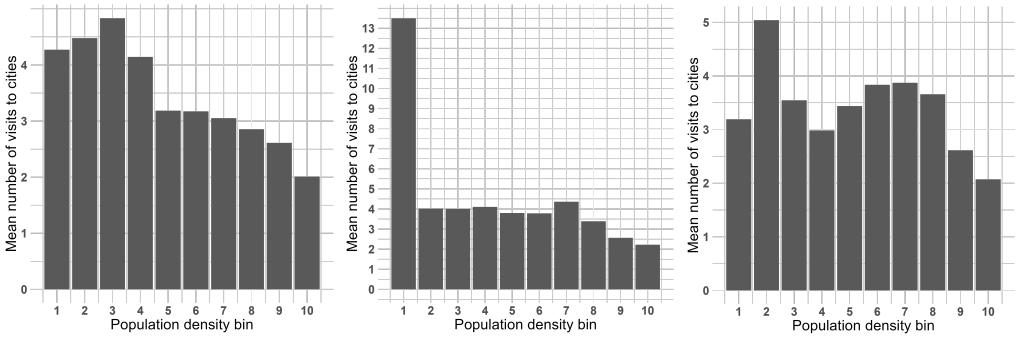
Figure 1.7: Distribution of users according to the number of cities visited, by population density bin.



(a) Kenya (b) Nigeria (c) Tanzania

Note: This figure shows for each decile the distribution of users who are never seen in a city, those who visit exactly one city, those seen in two cities and those visiting three or more cities. These counts exclude the home city in the case of urban residents.

Figure 1.8: Average number of visits to cities, by population density bin.



(a) Kenya (b) Nigeria (c) Tanzania

Note: This figure shows for each decile the average number of distinct visits to cities across users.

Overall, we match more than 80% of visitors to at least one polygon as shown in column (1) of Table 1.4.[23] The first two columns show the places visitors are seen at. We then split the sample of visitors into those from rural and urban areas, taking a threshold value of 300 people per square km.[24] For comparison, the final two columns show locations visited by residents of these cities. The table shows that about 80 percent of visitors are seen at residential locations, and about half of the visitors are seen while on a road or a roadside. Slightly more than one third of visitors are seen at locations related to travel (e.g., airports, train stations, hotels). About one out of three visitors is seen at shops and markets or retail

[23]Matching rates disaggregated by city are provided in Appendix Table 1.A.2.

[24]We chose this threshold for comparability with other datasets; for example, in the Global Human Settlement Layer, most rural clusters have a density below 300 inhabitants per square km (Schiavina et al., 2022).

locations. About one out of five visitors is seen at a commercial or industrial zone. Slightly more than 10 percent of visitors are seen at recreational locations (e.g., stadium, cinema, nightclub, theatre). About 12 percent is seen at a location offering public goods and services (e.g., hospital, health centre, university, police station, government buildings). The *other* category includes military zones and urban agricultural areas. When disaggregating visitors by home population density, the main differences are that a lower proportion of rural visitors is seen at residential areas; they are more often seen at shops and markets, public goods and services, recreational locations, locations related to food and drinks (e.g., restaurants, bars, food courts, cafes) and places of worship (e.g., cathedrals, mosques, synagogues and churches).

When comparing residents with visitors, residents are, reassuringly, more often seen at residential locations as well as most of the other categories. This is unsurprising, since we observe them for longer times in these cities we are more likely to see them visiting one of these different types of locations. The only places we observe them less often than visitors are places related to travel.

Table 1.4: Distribution of users across places visited by density of origin.

	Visitors		**Visitors from Below 300**		**Visitors from Above 300**		**Residents**	
	Users	**% of users**	**Users**	**% of users**	**Users**	**% of users**	**Users**	**% of users**
Total	16,156	-	590	-	15,543	-	67,982	-
Total users matched with OSM	13,214	100.0%	438	100.0%	12,756	100.0%	60,432	100.0%
Residential	10,628	80.4%	288	65.8%	10,325	80.9%	54,633	90.4%
Roads and roadsides	6,815	51.6%	251	57.3%	6,560	51.4%	36,795	60.9%
Travel	4,825	36.5%	162	37.0%	4,652	36.5%	17,329	28.7%
Shops and markets	3,775	28.6%	159	36.3%	3,614	28.3%	30,076	49.8%
Commercial zone	2,835	21.5%	88	20.1%	2,745	21.5%	21,366	35.4%
Industrial zone	2,280	17.3%	78	17.8%	2,197	17.2%	21,445	35.5%
Public goods and services	1,540	11.7%	70	16.0%	1,469	11.5%	16,315	27.0%
Recreational	1,008	7.6%	45	10.3%	962	7.5%	10,516	17.4%
Other	733	5.5%	35	8.0%	696	5.5%	7,311	12.1%
Food and drinks	347	2.6%	16	3.7%	331	2.6%	3,893	6.4%
Worship	331	2.5%	16	3.7%	314	2.5%	4,733	7.8%

Note: This table links the locations of visitors to Lagos, Abuja, Nairobi, Mombasa, Dar es Salaam and Dodoma and residents of these cities with OSM data to show the type of locations visitors and residents are seen at.

This section has reported on a number of different measures of mobility. These measures point to some consistent stories. The smartphone users in our data represent a mobile population. On average, they are more than 10 km from home on about one-sixth of the days on which they are observed. Those in more sparsely populated areas are more frequently away from home than those who live in city center locations. When they venture from home, they frequently travel far; when we sight them away from home, they are on average between 35 and 50 km away.

Flows are not limited to inter-urban movements of city dwellers visiting other cities; on the contrary, the data show extensive movement across and between many different locations. Many users visit more than one city (other than their

home city) over the sample period, and we observe people making repeat visits to the same city. Users appear to consume a diverse range of amenities during their stays, taking advantage of opportunities for market visits, administrative tasks, health services, and more. We emphasize that these visits do not appear to reflect regular commuting, nor do they correspond to permanent or seasonal migration.

1.5 Conceptual framework

Having documented the patterns of mobility that we observe in the data, we now turn to a theoretical framework in which these mobility choices arise from optimizing behavior of individuals. We presume that individuals make choices about where to live, which destinations to visit (and how frequently and for what duration), along with the usual choices about consumption. We consider that individuals are operating within the context of spatially dispersed economies that are characterized by a range of mobility frictions. These frictions shape the equilibrium patterns of location choice and mobility.

Our theoretical structures are designed to correspond to the mobility patterns that we observe in the data. The evidence shows many individuals travelling from their home locations to visit other destinations, returning to their points of origin location. In our data, many of these visits are temporary; individuals return to the home location after each visit. But most of the visits we observe do not appear to be well characterized as commuting: they cover longer time periods and distances than one would expect from daily commutes. This is not to deny the significance of daily commuting in our three countries; but our model, like our data, focuses instead on the phenomenon of longer-duration and longer-distance visiting. We also note that our data do not allow us to observe permanent migration with any confidence, since we have only one year of data and observe individuals on average on 40 distinct days over a period of 100 days. Our theoretical framework leaves open the possibility of permanent migration but has little to say about it.

Our model draws on insights from models such as Miyauchi et al. (2022) or Redding and Turner (2015), but we simplify greatly in matters on which our data are silent. In particular we abstract from detailed modelling of housing costs, and we greatly simplify our treatment of labor markets and goods markets. This allows us to focus solely on the between-location visits that comprise our data. In comparison with Bryan and Morten (2019), we also abstract from modelling labor market matching and the corresponding implications for permanent or seasonal migration.

The model economy is defined spatially as consisting of a set of locations, X. As in our mobility metrics above, a particular location – corresponding approximately

to a grid cell in the data – can be denoted as $x \in X$. In our data, people are observed living at particular home locations. We consider that the initial allocation of individuals across home locations is historically determined but is sustained at present as a spatial equilibrium with frictions.

1.5.1 People

The economy is populated by a large number of people. Each person i has a home location, $h \in X$, which is the location in which the person lives and purchases consumption goods.

1.5.1.1 Preferences

Individuals have preferences over an agricultural good, a_i; a non-agricultural good, c_i; and a good q_i that can be characterized as location-specific amenities. Individuals also have additively separable idiosyncratic preferences over home locations; individual i receives utility $\psi_i(h)$ from living in home location h. These preferences over home locations capture a large range of unobserved dimensions of location characteristics that may differ across individuals, such as proximity to families and social networks, or local knowledge of customs and norms. This structure also rationalizes the initial distribution of population, in the sense that a spatial equilibrium holds essentially by construction. Thus, preferences are represented by the utility function $U_i = u(a_i, c_i, q_i) + \psi_i(h)$.

Note that the goods a_i and c_i are purchased in the home location at the prevailing prices in that location. When at home, individuals also consume the amenities produced in the home location. However, individuals may also consume the amenities produced in different locations. These are imperfect substitutes for one another, and individuals have a preference for variety in these location amenities. To consume the amenity of a different location, an individual must travel to that location for a "visit" of some minimum duration. (Without loss of generality, think of this as at least one day. In other words, simply passing through a location does not allow a person to experience the amenity.)

The quantity of the amenity consumed on a visit to a location depends on the duration of the visit. It also depends on the quantity of amenities that the location produces; as will be discussed below, different locations provide different levels of amenity to their visitors. Let θ_{ix} denote the fraction of time that person i spends in location x in the course of a year. Assume that location x produces amenities $y(x)$. Then $q_{ix} = \theta_{ix} y(x)$, where $0 \leq \theta_{ix} \leq 1$. Note that across locations, $\sum_x \theta_{ix} \leq 1$. (The inequality may hold strictly, since we exclude time spent in transit.) Over the course of the year, an individual thus aggregates location amenities based on

the time spent in different locations, according to a CES expression that allows for some preference for variety:

$$q_i = \left[\sum_x (q_{ix})^\rho\right]^{\frac{1}{\rho}} = \left[\sum_x (\theta_{ix} y(x))^\rho\right]^{\frac{1}{\rho}}.$$

1.5.1.2 Travel and the accumulation of location amenities

In what follows, we will assume that a visit to any particular location has a minimum time duration (e.g., one day), so as to avoid treating transit through a location as a visit. This implies that the fraction of time that individual i spends in location x will be the sum of time spent on some integer number of distinct blocks of time that the person makes to that location. We define each of these blocks of time as a visit. Let $V_{ix} \geq 0$ denote the number of distinct visits by person i to location x. (Without loss of generality, we can treat the home location as simply one of the locations $x \in X$.) Using v to index these visits, and letting θ_{ivx} denote the proportion of person i's time spent in location x on visit v, then:

$$\theta_{ix} = \sum_{v=1}^{V_{ix}} \theta_{ivx}$$

During a visit, the individual receives utility that reflects the duration of the visit and the quantity of amenities available in the destination, as discussed below. Longer visits generate higher utility, as do visits to locations with higher levels of amenities. Amenities accumulated from different locations are effectively varieties, and the utility structure allows for consumption to vary along both the extensive margin (number of different locations visited) and intensive margin (duration spent in particular locations).

Travel to a location is costly. When person i travels to location x, where $x \neq h$, three costs are incurred. The first is a fixed cost of making a trip – the cost of leaving home; this is denoted by λ. The second is a cost per unit of distance travelled from origin to destination. Finally, there is a cost per unit of time spent in x. In a slight abuse of notation, let D_{ix} represent the distance between the home location h of person i and location x, and let γ represent the unit cost of distance. Moreover, let τ_x denote the cost associated with time spent in location x. Then the cost faced by person i of a visit to location x of duration θ_{ivx} is: $\lambda + \gamma D_{ix} + \tau_x \theta_{ivx}^\alpha$, where $\alpha > 1$ to reflect the fact that longer visits are more costly, per unit of time, than shorter ones. (This assumption serves to motivate the possibility that an individual might make multiple visits to the same destination in the course of a year.)

The cost structure of travel seems complicated, but each of these costs has a corresponding real-world element. For instance, one could think of the fixed cost

as related to the monetary and non-monetary costs of planning a trip, while the distance cost is the bus fare. The increasing cost of visit duration is intended to capture the fact that a brief visit might involve only modest imposition on friends and relatives, while a longer visit requires a more substantial investment in room and board, not to mention higher costs associated with being absent from the home location. For instance, a shopkeeper from a small town can travel for two days at relatively low cost to a nearby city to visit family members and to source supplies. To be gone for two weeks, however, requires turning over management of the shop to an assistant, and it may require paying a higher price – either formally or informally – for room and board.

1.5.1.3 Budget constraint

Individuals supply one unit of labor inelastically to the labor market in their home location, and in return they receive a real wage $w(h)$ that is location-specific. They allocate this income to expenditures on the agricultural good, the non-agricultural good, and the costs of any trips that they make. The agricultural good and non-agricultural good have prices that are location-specific, $\pi_a(x)$ and $\pi_c(x)$. Wages and travel costs are denominated in a numeraire good. The amenities themselves are of course free to consume, but travel to non-home locations incurs the costs described above. This gives rise to a budget constraint for individual i that can be written as:

$$\pi_a(h)\, a_i + \pi_c(h)\, c_i + \sum_x \left[V_{ix}\left(\gamma D_{ix} + \lambda\right) + \sum_{v=1}^{V_{ix}} \tau_x \theta_{ivx}^{\alpha} \right] \leq w(h).$$

1.5.1.4 Individual's problem

The individual's problem is then well-defined. Taking her home location as given, she chooses the quantities of the consumption goods, a_i and c_i, and the number and duration of visits to each non-home location, V_{ix} and θ_{ivx} to maximize utility subject to the budget constraint above.

1.5.2 Geography

Let $N(x)$ be the population living within location x; in effect, this is a measure of population density. We will describe a location as populous if it has a population density $N(x) > \bar{n}$. We will go further and define a settlement (a term intended to include both towns and cities) to be a subset of populous locations $K \subset X$ that meets three criteria: (a) the locations form a contiguous spatial group within X; (b) for each location x in K, the density criterion is satisfied; and (c) the total

population of the settlement exceeds some threshold value for total population – i.e., $\sum_{x \in K} N(x) > \bar{N}$. There will necessarily be a finite set of settlements, which we denote as $\bar{K}$. For notational simplicity, let $N_1, N_2, ...N_{\bar{K}}$ denote the populations of the different settlements; furthermore, without loss of generality, we can order the indexing such that $N_1 < N_2 < ... < N_{\bar{K}}$. Note that not all people live in settlements; we define as "rural" those people who live in low-density locations, along with those living in clusters of density that do not meet the aggregate population threshold (e.g., small villages and communes).[25]

1.5.2.1 Location amenities

The amenity is a non-tradable public good (non-rival and non-excludable) that is consumed by people who live or visit a location. The amenity is produced with increasing returns to population size. In particular, for settlement k, $y(k) = AN_k^{\beta}$ gives the production quantity of this location amenity, where $\beta > 1$. It would be possible to define amenities produced at different rural locations in the same way, but for simplicity here, we will assume that all non-home rural locations produce an identical amenity, y_r, which is lower than the level produced in the smallest settlement; in other words, $y_r A\bar{N}^{\beta}$.

The structure of amenity production captures in a simple way that there are agglomeration effects in the provision of amenities, such that larger cities in general produce higher levels of amenities. This implies that the utility derived from a one-day visit to a large city is greater than that from a visit of identical length to a smaller city. However, working against that are the preference for variety and the role of distance. A nearby small city may be less costly to visit than a faraway city that is larger; and all else equal, individuals will be inclined to want to visit multiple locations. The duration of visits will reflect a balance between the fixed cost and distance cost of travel, on the one hand, and the increasing duration cost, on the other hand. Individuals will be likely to make multiple visits to the same destination when that location is relatively close (so that the distance cost is low). The duration of a visit will tend to be longer when the destination is far away.

1.5.3 Production

In what follows, we consider the simplest possible production arrangement for this economy. All rural areas produce the agricultural good, and all settlements produce the composite non-agricultural good. With no disutility from labor, each worker supplies one unit of labor inelastically. Each worker in a location produces

[25] In the data for our three countries, cities and towns are defined in a variety of different ways. Our formulation is a convenient one to use, and it is consistent with many standard approaches. However, none of our results depends on this particular way of defining or characterizing settlements.

one unit of the good, so $y_{ax} = N_x$ for every rural location, and $y_{cx} = N_x$ for every urban location. In the simplest specification, both goods are frictionlessly traded on a world market, with prices $\pi_a(x) = \pi_a^* \; \forall x$ and $\pi_c(x) = \pi_c^* \; \forall x$ determined exogenously to the model economy. This is obviously a strong simplification, particularly for the economies we are studying, but it allows us to focus on frictions to the mobility of people, consistent with our data. Note that an immediate implication of the production structure is that wages will differ in rural and urban regions, with $w_a = \pi_a^*$ and $w_c = \pi_c^*$.

1.5.4 Equilibrium

We focus on a short-run spatial equilibrium for this economy. The equilibrium is trivial, in the sense that there are few endogenous variables. Assume (not unrealistically) that the marginal value product of a worker in non-agriculture is higher than the marginal value product of a worker in agriculture; or in other words that $\pi_c^* > \pi_a^*$. With prices of the two tradable goods identical across locations, this immediately implies that real wages will be higher in urban areas than in rural areas; indeed, realized utility per unit of income will be higher in larger cities than in smaller cities, since larger cities are more productive in supplying amenities. This seemingly creates some potential for spatial gaps, but the equilibrium is sustained by a combination of differences in location-specific preferences and mobility frictions.

In a sense, the only interesting feature of the equilibrium is the endogenous optimization by individuals of the number, duration, and destination of visits. The structure of the problem gives rise to a number of predictions that can be tested against the data.

Proposition 1 *Assume for simplicity that $\tau_x = \bar{\tau} \; \forall x$. Define the number of visits from settlement k_1 to settlement k_2 as the sum of the number of visits by each individual living in any location within the boundary of k_1 to any location within the boundaries of k_2. Denote this number as $V_k(1,2)$. Then:*

$$N_{k_2} > N_{k_1} \Rightarrow \frac{V_k(1,2)}{N_{k_1}} > \frac{V_k(2,1)}{N_{k_2}}.$$

In other words, the number of visits per person made from the smaller settlement to the larger will exceed the number made in the opposite direction. This reflects the higher level of amenities produced in the larger settlement. The logic of this proposition is simple. Wages and prices are the same in both settlements; the distance and travel costs are also identical. But the utility value of visiting the more populous location is higher for an individual in the less populous location. The same logic will hold in general for visits from rural areas to settlements of different

size, but because rural wages are assumed to be lower, the overall prediction is ambiguous; it depends on the size of the income effect and the difference in wages. For the case where $\pi_c^* = \pi_a^*$, it certainly follows that rural people will visit settlements more frequently than town dwellers visit rural areas.

Lemma 1 *If an individual makes multiple visits to the same location, they will be of the same duration. This follows from the increasing cost with duration; the total cost is minimized by making all visits equal in duration.*

Proposition 2 *Building on Lemma 1, this tells us that for any two locations that are visited, there is a relationship between the settlement size (or rural status), the distance, the cost of spending time, and the duration of the visit. Visits to settlement $k_1$and k_2 will be related according to the non-linear relationship given by:*

$$\left(\frac{\theta_1 N_1^\beta}{\theta_2 N_2^\beta}\right)^{\rho-1} = \frac{\gamma D_1 + \lambda + \tau_1 \theta_1^\alpha}{\gamma D_2 + \lambda + \tau_2 \theta_2^\alpha}$$

This expression does not give neat closed-form relationships, but consider the simple case in which $\tau_1 = \tau_2 = \lambda = 0$; in other words, a situation in which the only costs of visits are the linear costs of distance. In this case, we can solve for the duration of a visit as a function of distance and city size:

$$\theta = \frac{(\xi \gamma D)^{\frac{1}{\rho-1}}}{A N^\beta}.$$

This in turn gives rise to an estimating equation in the form:[26]

$$\ln \theta = \delta_0 + \delta_1 \ln N + \delta_2 \ln D + \epsilon.$$

A more complete specification of the location-specific production function for amenities might include a set of observable and unobservable location characteristics; this would motivate an estimating equation in the same form, but including origin and destination fixed effects φ_o and ν_d, with the destination fixed effect subsuming the destination city size:

$$\ln \theta_{od} = \delta_0 + \delta_1 \ln D_{od} + \varphi_o + \nu_d + \epsilon_{od}. \tag{1.8}$$

We will explore this relationship further in the next section.

Proposition 3 *Given a choice between visiting two equidistant locations, an individual will be more likely to visit the more populous location, and/or to stay longer in the more populous location.*

[26]This equation is similar in flavor to a gravity equation coming out of quantitative spatial models developed by Ahlfeldt et al. (2015) and Kreindler and Miyauchi (2023).

This follows trivially from the fact that a visit to the more populous location delivers higher marginal utility because of the greater amenity value provided during a visit of the same length.

1.6 Empirical tests

We next explore to what extent our proposed conceptual framework is consistent with the mobility patterns that we observe in the data by examining each of the propositions.

1.6.1 Proposition 1

Proposition 1 states that the number of visits per person from a smaller settlement to a larger will be higher than the number made in the opposite direction. To test this proposition, we sum all visits of users between city pairs throughout the year.[27] We normalize the number of visits by the number of users with home locations in each city, reflecting the fact that we observe only a subset of the population. This gives a matrix where each entry corresponds to the proportion of residents in a particular origin city who are observed travelling to a given destination. We then determine which of the two cities is larger in population and compare the flows of visitors in each direction. We do this for all pairs and perform a simple pairwise t-test of the following null hypothesis

$$H_0 : \frac{V_k\,(1,2)}{N_{k_1}} = \frac{V_k\,(2,1)}{N_{k_2}} \tag{1.9}$$

where the proposition assumed that $N_{k_2} > N_{k_1}$ for any two settlements within one of our three countries. Table 1.5 presents the results from these tests. The table shows that in all cases the average number of visits per person from the smaller location to the larger exceeded the number made in the reverse direction. Given

Table 1.5: Number of visits between locations

	Kenya	Nigeria	Tanzania
$V_k(1,2)/N_{k_1}$	0.343	0.233	0.144
$V_k(2,1)/N_{k_2}$	0.056	0.037	0.033
H_a: $(V_k(1,2)/N_{k_1} - V_k(2,1)/N_{k_2}) > 0$	0.000	0.000	0.000
n	121	751	157

Note: This table tests Proposition 1 by conducting a paired t-test that compares the number of visits between locations of different sizes.

[27] As for the rest of the paper, we define city boundaries as described in Appendix Section 1.B. Here we consider the subset of cities above 50,000 inhabitants – based on 2018 WorldPop population estimates. We exclude visits that originate in non-urban locations.

that some of the location pairs might have small differences in populations, we also explore whether the distribution of visits becomes more distinct when we vary the difference between the origin and destination populations. Appendix Figure 1.E.4 shows that this is indeed the case. The same pattern holds true for Tanzania and Nigeria.

1.6.2 Proposition 2

Proposition 2 gives rise to a relationship between distance to the destination and the duration of visits. We now use our device-level data to estimate the equation (1.8)

$$\ln \theta_{od} = \delta_0 + \delta_1 \ln D_{od} + \varphi_o + \nu_d + \epsilon_{od}.$$

where θ_{od} represents the fraction of days a user residing in o spends in a particular city d, φ_o and ν_d are origin and destination fixed effects and D_{od} represents distance between the origin and the destination.

Table 1.6: Gravity model for inter-city mobility.

	Kenya	Nigeria	Tanzania
	(1)	(2)	(3)
ln (Distance)	-.049**	-.086***	-.051***
	(0.021)	(0.01)	(0.017)
Obs.	7201	40077	7032
R^2	0.115	0.107	0.111

Note: This table estimates equation (1.8). The dependent variable is the fraction of days a user residing in origin o spends in destination d. All models include origin and destination fixed effects. Reported standard errors are clustered at the user level. *, **, *** denote significance at 10%, 5% and 1% levels.

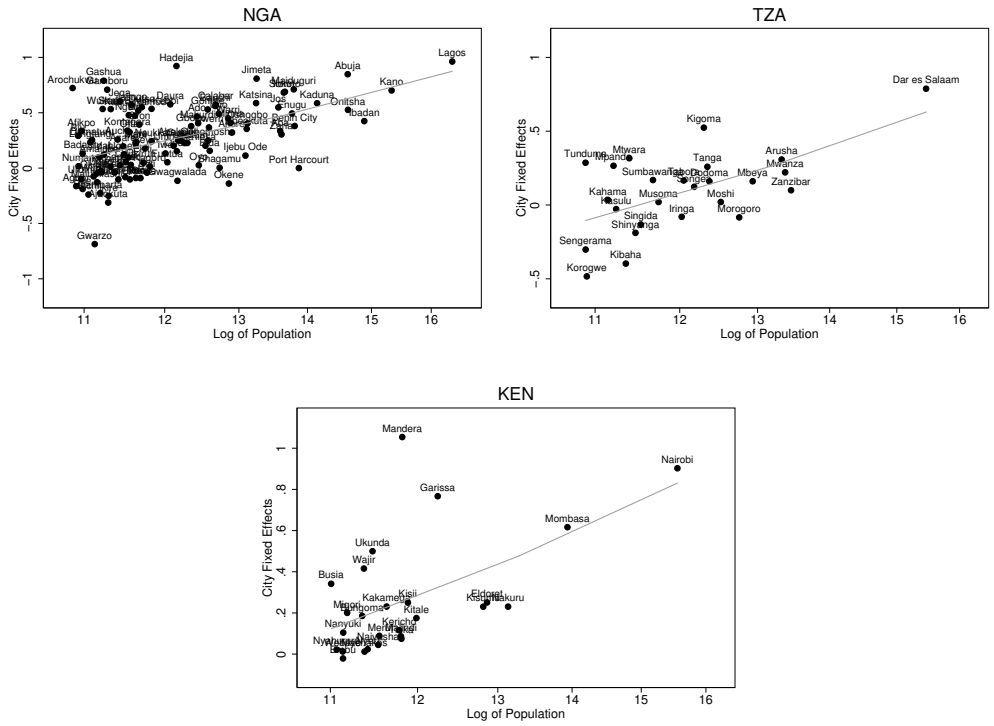
Origin fixed effects proxy for any observables or unobservables at the origin. Table 1.6 shows the results from estimating this relationship using all visits in our dataset, where we exclude visits that originate from rural areas. The table shows a clear negative relationship between distance and the fraction of days users spend visiting a city, after controlling for origin and destination fixed effects. The results are very similar when we use travel time instead of distance.[28] The negative coefficient on the distance variable is also a key empirical regularity found in standard gravity equations that regress a commuting or migration probability on the log of distance while controlling for origin and destination fixed effects.

[28]When we cluster standard errors at both the origin and destination the significance levels in Kenya and Tanzania drop to the 10 and 12 percent level, respectively.

1.6.3 Proposition 3

Proposition 3 states that holding distance constant, an individual will be more likely to visit a more populous destination and/or stay longer. To test this proposition, we extract the destination fixed effects that we estimated with equation 1.8 and examine their relationship with population. Figure 1.9 plots the city fixed effects against city size, where we use the smallest city in each of the countries as the omitted category. A few points are worth highlighting. First, the city fixed effects

Figure 1.9: Destination fixed effects and city size.



Note: This figure shows the city fixed effects $\hat{v_d}$ from equation (1.8) and log of population.

correlate significantly with city size. Second, the figures highlight that Lagos, Dar es Salaam and Nairobi are outliers in terms of city size; all have the highest city fixed effects, conditional on distances between city pairs and origin city fixed effects. The political capital Abuja is well above the predicted regression line, indicating that it receives more visits than its population size would predict. Other locations, like Zanzibar, receive fewer visits than predicted by their population size. The model suggests that Zanzibar, located on an island, clearly would receive more visitors than it does without this barrier.

Table 1.7 shows the results from the regressions of the city fixed effects on log population to investigate the relationship more formally. The table shows that

Table 1.7: Destination fixed effects and city size.

	Kenya	Nigeria	Tanzania
	(1)	(2)	(3)
ln (Population)	0.156***	0.146***	0.161***
	(0.024)	(0.021)	(0.042)
Obs.	26	105	25
R^2	0.313	0.268	0.374

Note: This table regresses the city fixed effects from equation (1.8) on log city city. Robust standard errors in parentheses. *, **, *** denote significance at 10%, 5% and 1% levels.

the destination fixed effect is significantly higher for more populous locations, suggesting that individuals are significantly more likely to spend a higher fraction of days in larger settlements. This underlines the magnetic forces large cities play.

1.7 Conclusion

Until now, most of our knowledge about human mobility in low-income countries has come from surveys that show migration flows between survey rounds. Often the surveys are several years apart or longer (e.g., decennial censuses). This means that mobility is only evident in these data sources over very long time horizons. The data from these surveys are useful and informative in thinking about certain types of population movements, but they tell us little about the ways in which individuals serve as links between different locations – potentially moving goods, ideas, information, and relationships.

In this paper, we use smartphone location data to show how individuals move between multiple locations, taking advantage of the different opportunities and amenities that are available, and presumably building and maintaining social networks. But individuals' movements also serve to construct networks of locations. The extent of mobility between locations serves as evidence of spatial integration. Our data provide a detailed look at one type of network of locations – a network based on human mobility. The paper builds on a recent literature that has used "big data" to study commuting, migration and travel along trip chains (Blumenstock, Chi, et al., 2022; Kreindler and Miyauchi, 2023; Miyauchi et al., 2022). Our contribution here is to focus on "visits", which turn out to be ubiquitous.

The data help us to improve our understanding of travel and mobility in African countries. Our smartphone users travel frequently and relatively far. Travel is not limited to peri-urban commuting, nor to migration (whether seasonal or permanent). Most of our sample consists of urban dwellers, and we observe many of them travelling to other cities – indeed, significant numbers travel to multiple cities other than their home cities. But perhaps surprisingly, our urban users

also travel to rural areas. For instance, some 20-40% of our urban Kenyan users are observed in locations that can be characterized as rural (i.e., locations in the bottom half of the population density distribution). Our research thus suggests that we should be cautious in imagining that the villages, towns, and small cities of sub-Saharan Africa are functionally cut off from large cities – or from each other. On the contrary, we see substantial flows of people in all directions. Smartphone ownership appears not to deter people from travelling; in that sense, smartphones do not appear to *substitute* for human mobility; we find that smartphones are often used by people when they are visiting non-home locations.

Our analysis benefits from the availability of new data sources that allow for a startling level of detail in observing mobility. Such data sources are increasingly available for low-income countries, as well as for rich countries. The Covid-19 pandemic saw similar data used to characterize the impact of lockdowns and other short-term questions. Our paper can be viewed as an illustration of the potential for using such data to address deeper questions about a range of issues in development. At the same time, the widespread availability of these data raises concerns about privacy and security. Our analysis has avoided mining the data to extract further information about individual users; we argue that there is much to learn from the data while respecting the anonymity and privacy of individuals.

The data clearly also embed some intrinsic limitations. One relates to the selection issues that make our sample unrepresentative. Although we filter out many "transit pings," we cannot fully determine which places people visit deliberately; we can only tell that people used their devices while they were in particular locations.[29] But we benefit from the large number of observations and the large number of users.

Our samples are clearly selected and are not representative of national populations. For the poorest people in our three countries, patterns of mobility may be very different from those we describe here. Even small monetary costs of mobility can be highly salient for the poor. Poverty is not the only barrier to mobility: people also face mobility barriers linked to gender, ethnicity, social class, age, and other dividing lines. Our data may also be atypical at the country level; we cannot extrapolate clearly from our three countries to other parts of sub-Saharan Africa, and certainly not to other parts of the developing world. Patterns of mobility and frictions may look very different in Latin America or Asia. Nevertheless, the methods that we develop in this paper illustrate the promise of new data sources. As such data become more widely available, there is potential to learn far more about spatial frictions, mobility, and the geographic patterns of human activity.

[29]The underlying distinction is itself somewhat unclear; it depends on the unobservable *intent* of the traveller, rather than on the characteristics of the locations or the trips.

Appendix 1.A Details on smartphone app data

1.A.1 Algorithm to identify home locations

The calculation of users' home locations plays a critical role in our analysis of high-frequency mobility patterns. First, home locations are often used as reference locations to observe mobility trajectories. Second, home locations are used to evaluate the spatial coverage of our sample by comparing the spatial distribution of users to the distribution of the population. Third, knowing where our users live helps us infer key information allowing to characterize them, e.g. by pairing users with DHS clusters. In our base sample, we define home locations as the most frequently observed 2-decimal rounded coordinates at night (between 7pm and 7am, local time). We consider that the likelihood of correct home location prediction increases with both the number of nights a user is seen and the fraction of these she is observed at the inferred home location. Therefore, we select a subset of users that are seen at least 10 nights, of which at least half are at their home location. We call this subset the "high-confidence" sample and use it as our core sample in the analysis of high-frequency mobility patterns throughout the paper. We also build medium- and low-confidence subsets that include users seen at least 8 and 5 nights respectively in order to evaluate the robustness of our results - the required fraction of nights seen at home is kept at 0.5. The corresponding sample sizes are given in Table 1.A.1.

Table 1.A.1: Number of users by subset and country

	Base	**High**	**Medium**	**Low**
Kenya	195,630	18,535	23,490	37,249
Nigeria	659,407	78,694	96,954	146,346
Tanzania	234,213	22,728	28,853	46,116
TOTAL	1,089,250	119,957	149,297	229,711

Note: This table shows the number of users in each subset by country. Unsurprisingly, the sample size decreases with the minimum number of observed nights imposed and nearly doubles between the high- and low-confidence subsets.

1.A.2 Construction of the base sample and data irregularities

Our initial samples have 317,420 users in Kenya, 958,207 users in Nigeria and 780,760 users in Tanzania. According to the methodology presented in Section 1.A.1, we cannot infer home locations for users never observed at night (7pm-7am) and 121,790, 297,895 and 173,886 users are thus removed in Kenya, Nigeria and Tanzania respectively. Moreover, in Nigeria, inferred home locations with equal latitude and longitude were deemed erroneous which resulted in 905 users being removed. In Tanzania, we identified a data sink of 372,661 users with an inferred

home location at (35.75;-6.18), which is located within the city of Dodoma. This represents 52% of the initial sample while we estimated the city of Dodoma to host 0.5% of the population.[30] We entirely remove users with home location coordinates at the data sink from the sample.

1.A.3 Algorithm to identify work locations

Similarly to home locations, we assign a work location as the modal 0.01-degree cell in which a user is observed between 9am and 6pm on weekdays. We again impose two restrictions: that (a) the user is observed for a minimum of 8 distinct weekdays and (b) is seen at the inferred work location for at least 50% of the total weekdays. Overall, nearly all users of the high-confidence set are seen for at least one weekday and 87,920 meet the confidence criteria for the identification of work location, which represents 73% of the high-confidence set. In this subset ("work subset"), home and work locations are found within the same 0.01-degree cells for 80% of users which is in line with high rates of self-employment and short-distance commuting.[31]

For those with distinct home and work cells, the median distance between home and work is about 4.4 km with again some differences between urban (4.5km) and non-urban (3km) users. Restricting our subset to users observed for a minimum of 10 days or considering a higher resolution (0.001-degree cells) for home and work locations imply only marginal changes to the results.

1.A.4 Algorithm to identify transit pings

To define transit pings we first define visits as sequences of successive pings located within a same 5-km grid cell. We infer the minimum duration of visits from the time elapsed between their first and last pings and classify these as a stay when they last more than some limit value T_{stay}. We choose a value for T_{stay} that corresponds to the amount of time required to drive through a 5km cell at 20 km/h. Other visits are then classified as transits when (i) there is no evidence of their duration being at least greater than T_{stay} and (ii) a speed value greater than 20km/h is observed for at least 25% of their pings. The second condition ensures that we are observing a user moving significantly faster than a walking pace.

[30] See Appendix Section 1.B for more details on the definition of city boundaries. We overlay 2018 WorldPop population map to estimate the population in Dodoma, as we do in other parts of the paper to estimate city sizes.

[31] For instance, in Tanzania, the LSMS data show that median travel time between home and work for urban wage workers is 30 minutes, which would normally correspond to about 2.5 km, assuming walking as the mode of transport. The numbers for the self-employed and for rural workers are substantially less. The fraction of users with identical home and work locations is higher in our data for the subset of non-urban residents (86%), consistent with lower fractions of commuters in small cities and rural areas.

More formally, for a user i, the sequence of successive pings is denoted $(a_1^i, ..., a_{P_i}^i)$ with P_i the total number of pings for user i. Each ping consists of a timestamp t_j^i (in seconds) and longitude/latitude coordinates $coord_j^i$. For each country, we can partition the country extent to resolve raw longitude/latitude coordinates and form a finite set of N locations $X = \{x_1, ..., x_N\}$. In this case, we use a 5-km resolution fishnet so that X is a set of 5km grid cells and we associate the sequence of pings $(a_1^i, ..., a_{P_i}^i)$ to the sequence of X-locations $(x_1^i, ..., x_{P_i}^i)$. We formally define a visit as a sequence of successive pings at one given location $x \in X$ where the time elapsed between two consecutive pings is lower than some parameter ϵ_{visit}.[32] For the m^{th} visit of user i, $v_m^i = (x_{j_m}^i, ..., x_{j'_m}^i)$, we define the visit minimum duration $T^{min}(v_m^i)$ as the time elapsed between the first and last pings of the visit, i.e. $T^{min}(v_m^i) = t_{j'_m}^i - t_{j_m}^i$. The visit maximum duration $T^{max}(v_m^i)$ is the time elapsed between the last ping of the preceding visit and the first ping of the following visit, i.e. $T^{max}(v_m^i) = t_{j'_m+1}^i - t_{j_m-1}^i$. $T^{min}(v_m^i)$ (resp. $T^{max}(v_m^i)$) represents a lower (resp. an upper) bound estimate of the actual amount of time spent at the corresponding location during visit v_m^i. Finally, we define the travelling speed at ping a_j^i, $speed_j^i$, as the ratio of the haversine distance to the preceding ping a_{j-1}^i over the corresponding time elapsed $t_j^i - t_{j-1}^i$, if $t_j^i - t_{j-1}^i \leq \epsilon_{speed}$. The value for ϵ_{speed} is typically small to ensure that the straight line between a_j^i and a_{j-1}^i is a good approximation for the user's trajectory between those two pings so that the estimated speed value reflects the actual travelling speed – here we set ϵ_{speed} to 30 seconds.

With these definitions in mind, we implement a filtering algorithm with the objective of identifying pings corresponding to users simply driving through some locations. First, we identify all visits for each user by setting ϵ_{visit} equal to 30 minutes. We classify a visit as a stay if its minimum duration is greater than some value T_{stay} corresponding to the time required to travel along the diagonal of a 5km cell at an average speed of 20km/h, i.e. T_{stay}=1,273 seconds.[33] Then, we classify a visit v_m^i as a transit visit if the following two criteria are met: (i) v_m^i is not a stay [34] and (ii) at least 25% of speed values are greater than 20km/h.[35] Visits that are neither stays nor transits are classified as undefined.

[32] ϵ_{visit} can be interpreted as the maximum amount of time of inactivity between two consecutive pings at the same location we are willing to tolerate before considering that the user may likely have visited other locations and returned to the initial location during said period of inactivity. Also, "isolated" pings, i.e. pings being at least ϵ_{visit} seconds away from both their preceding and following pings, are considered as single-ping visits.

[33] By considering the longest segment within a 5km cell and a speed value of 20km/h in the lower range of possible average driving speeds, we use a conservative value for the parameter T_{stay}.

[34] More formally, either $T^{max}(v_m^i) < T_{stay}$, or $T^{max}(v_m^i) \geq T_{stay}$ and $T^{min}(v_m^i) \leq T_{stay}$.

[35] We further impose that speed values are available for at least 80% of the pings in the visit to avoid misclassifying visits where there is a high uncertainty around the estimated proportion of pings with speed greater than 20km/h.

We apply this algorithm to the three countries. Overall, 11% of the pings in the high-confidence set are identified as transit pings while 70% are stay pings. Differences across individual countries are only modest. Since the estimated total fraction of transit pings can be largely influenced by a handful of major users, we also calculate the average fractions of transit, stay and undefined pings across users.[36] We find that, on average, only 2% of a user's pings are classified as transit – 48% are identifies as stay pings and the remaining 50% as undefined. The average fraction of transit pings is markedly lower than the total fraction and disparities between countries are also less pronounced, which together suggests that major users differ from other users in that they showcase a relatively larger fraction of pings sourced from navigation apps – or, at least, are relatively more observed when travelling.

1.A.5 Algorithm to identify visits

For the purpose of detecting distinct visits to cities, we consider the set of locations X as the set of cities defined by 3km-buffered GRUMP polygons[37] and its complement that we qualify as "non-urban" areas, such that their union forms the country extent. A visit of user i to a given (non-home) city c is broadly defined as a certain period of time spent by i in city c. Taking this to our smartphone data, the m^{th} visit of i to c, $v^i_{m,c}$, materializes as a sequence of pings $(a^i_{j_{m,c}}, ..., a^i_{j'_{m,c}})$ located within city c and reflecting a single stay of i to c. For each user i, we effectively observe successive locations but to the extent that we do not control the frequency of observation, we cannot always determine with absolute certainty the location of users between two consecutive pings. In particular, a higher duration between two consecutive pings in a visited city is associated with a greater uncertainty as to whether the user travelled to another location or returned home while unobserved. Also, we are willing to tolerate a higher inter-ping duration as the home-to-city distance increases as we can reasonably assume that the likelihood of a user making multiple trips decreases. We formalize these qualitative characterizations of distinct visits in a two-steps algorithm that we further describe below.

First, we detect sequences of consecutive pings at a visited city. In this first step, we use a rather conservative criterion and, for any given user i, we allow for a maximum inter-ping time ϵ^i_{visit} that corresponds to a return trip in straight line between the considered ping and the home location at a constant speed of 40 km/h. We introduce "home flags" that indicate when a user was observed back to her home location between two consecutive sequences of pings at a visited

[36] In Kenya, the top 100 users in the high-confidence set account for 56% of the total number of pings. In Nigeria and Tanzania, this ratio is estimated at 21% and 32% respectively.

[37] See Appendix Section 1.B. We calculate city-level population values by overlaying city polygons with 2018 World Population map and consider the subsets of cities above 50,000 inhabitants.

city. In fact, here we adopt a looser definition for home that we deem sufficient to consider that the user returned home between what therefore qualifies as two distinct visits: (i) the home city for urban residents and (ii) a 5-km buffer centered in the estimated home location for non-urban users. Second, we allow for some grouping of consecutive sequences of pings at the same visited city according to a set of well-defined rules: (i) consecutive sequences of pings at the same visited city within a single day are grouped to form a unique visit,[38] (ii) if the travel time between the visited city centroid and the home location is less than 2 hours, we group together sequences of pings that are less than 12 hours apart,[39] (iii) if the travel time between the visited city centroid and the home location is strictly beyond 2 hours, we group together sequences of pings that are less than 36 hours apart. With criterion (i), we allow for the possibility of commuters being observed early in the morning and late in the afternoon in their destination city. This is also relevant for visits to the closest cities where $\epsilon^i_{visit,c}$ is small and potentially leads to separate sequences of pings to a visited city on a given day when those are most likely part of the same visit. Criterion (ii) basically allows for users to spend a night in a nearby city and therefore be unobserved for that period of time. For instance, a sequence of pings in Nairobi ending at 9pm one night followed by another starting at 7am the day after from a user residing in Thika (approximately a 1h drive) will be considered as a single visit to Nairobi. Similarly, criterion (iii) allows for two nights away to more distant cities without being observed, i.e. it is sufficient to see the user at the visited city on one night and in the morning two days after to consider that we are observing the same visit.

Having identified sets of pings belonging to individual visits to cities, we then provide estimates for their duration. We define the lower-bound estimate for the duration of the m^{th} visit to city c for user i, $v^i_{m,c} = (a^i_{j_{m,c}}, ..., a^i_{j'_{m,c}})$, as the time elapsed between the first and last ping of the identified sequence $v^i_{m,c}$, $T^{min}(v^i_{m,c}) = t^i_{j'_{m,c}} - t^i_{j_{m,c}}$. The upper-bound estimate is the time elapsed between the pings preceding and following $v^i_{m,c}$, so $T^{max}(v^i_{m,c}) = t^i_{j'_{m,c}+1} - t^i_{j_{m,c}-1}$.

1.A.6 Algorithm to identify places visited within cities

We identify and characterize the places where visitors to cities are seen based on free and open source data from OpenStreetMap. Geographic elements are defined using mainly two data types. *Nodes* are points are typically used to map features

[38] Note that we still allow for multiple visits to a city in a single day in cases where the user is effectively observed in the home location vicinity.

[39] In this second step, we use a more precise estimate of the travel time between visited city and home location. Driving times are calculated using Google Maps API through the R *drive_time* function (*placement* package). Also, the time elapsed between two consecutive sequences is defined as the time between the last ping of the first sequence and the first ping of the second sequence.

considered without a size (e.g. road signs, wells, statues, electric poles). *Ways* are ordered lists of nodes that represent either a polyline (e.g. a road) or a polygon if they form a closed line. Metadata in the form of *tags* provide attribute information on map objects such as their type, their name or their unique identifier. OSM covers a vast array of mapable features, from buildings, to roads, to industrial or residential zones. For each city, we construct a shapefile of polygons defining places that we can easily characterize. By overlaying those polygons with visitors' ping locations, we are able to gain insights into the type of places our users visit and provide some characterization for the purpose of their trips. In what follows, we describe in full details the procedure we adopted to construct spatial datasets of places within cities from raw OSM data.

First, we create a standard categorization of places. Each category can be thought of as a set of places that reflect a distinguishable purpose. For instance, a user seen in residential areas is most likely visiting friends or relatives, whereas pings in commercial or industrial zones are rather indicative of an individual conducting business activities. Second, we map raw OSM features into those categories. OSM country extracts are downloaded from Geofabrik website (download.geofabrik.de).[40] Each country archive contains a set of files that classify OSM features into different layers. We primarily used six layers: places of interest, points of interest, buildings, places of worship, roads, and landuse.[41] The procedure used to process and assign features to our categories varies across layers depending on the nature of spatial objects (polygons versus points) and attribute information available. We describe below the method used to categorize raw features for each individual layer.

Places of interest. This layer contains polygon features with a well-defined "feature class" attribute with values that can easily be mapped into our categories.

Points of interest. Points of interest are point features (i.e. *nodes*), also with a feature class attribute. Many of those points actually define places which were not delineated and entered in as polygons, but are only associated with unique point locations that roughly correspond to the center of those hypothetical polygons. We approximate the extent of those places by simply transforming points to square polygons of $400m^2$ ($20m \times 20m$) and we incorporate these elements in our database.[42]

[40]The country extracts we used reflect the state of the OSM database at the date when the analysis was conducted, i.e. 21 September, 2021.

[41]Other layers include natural features, traffic-related objects, railways, waterways and water bodies. None of those contain features that are relevant to our categories (and which cannot be found in the layers that we use).

[42]We acknowledge this is a relatively crude approximation but it allows us to retain as many elements as possible with a minimal risk of overestimating the extent of places given the conservative area considered ($400m^2$). The resulting features are then assigned to categories of places based on the feature class attribute, using the same correspondence matrix as for places of interest.

Buildings. Buildings are polygon features with two useful attributes: "type" and "name". They do not have a feature class attribute but can be assigned to our categories by first using the type attribute;[43] However, most building features have a missing type value and cannot be categorized on that basis.[44] For those elements, we still attempt to assign a category by matching key words to the name attribute. For instance, one feature of the buildings layer for Nairobi has a missing type but a name value "Parklands primary school", which we assign to the education category based on the presence of the word "school".

Places of worship. This layer specifically gathers identifiable places of worship such as cathedrals, chapels, churches, mosques and synagogues. It is comprised of both polygons and points. Polygons are integrated as such in our dataset as elements of the "worship" category. As with points of interest, worship points are converted into squared polygons of $400m^2$ which are then added to the set of worship features.

Roads. The roads layer is comprised of a comprehensive set of polylines describing road networks. We convert those lines into road bands (i.e. road polygons) by applying a standard 12m buffer. These polygons are useful to identify pings that fall on major roads and clearly reflect a user moving around the city by car, bus or any other transport mode. In this respect, we only keep roads classified as "trunk", "primary" or "secondary". We acknowledge that misalignment and road width smaller than the imposed buffer may lead to mismatches between our polygons and the actual roads. We therefore label this category as "roads and roadsides" to account for the fact that our road bands may in fact overlap with sidewalks.

Landuse. The landuse layer contains features with a "landuse=*" tag in the OSM database. The value of the landuse tag is reported in a "feature class" attribute in the Geofabrik landuse layer. Landuse features typically map areas (e.g. an industrial zone or a residential neighborhood) rather than buildings but allow to usefully complement our dataset. In fact, other layers provide information that allow to precisely characterize places at the building-level, but they typically show large fractions of features with missing attributes that thus remain without a category assigned. This is especially true of the buildings layer that usually accounts for the bulk of features found in the Geofabrik archives.[45] While we acknowledge

[43]The type attribute in Geofabrik extracts simply corresponds to the value of the "building=*" tag in natives OSM elements.

[44]For instance, for the city of Nairobi in Kenya, the buildings layer has 109,730 features, of which 94% have missing type value. We get comparable proportions of missing values in other cities of our sample.

[45]Across the six cities that we consider in our analysis (Lagos, Abuja, Nairobi, Mombasa, Dar es Salaam, Dodoma), the fraction of features in the buildings layer that have neither a type nor a name attribute ranges from 76% (Dar es Salaam) to 99% (Mombasa).

landuse elements are second-best compared to a building-level information, we argue they still provide a useful characterization of places that users may visit. More importantly, they significantly increase the coverage of our final dataset and thus also increase the fraction of ping locations eventually matched to an OSM feature.

Some features are occasionally assigned several categories[46] and we force each feature to map to a unique category by establishing an order of precedence. The order of priority that we define follows a logic of ranking categories from the most general to the more specific. For instance, a user seen in a restaurant within a university campus is primarily considered as having visited the university; "education" takes precedence over "food and drinks". The complete list of categories (and sub-categories) by decreasing order is as follows: education, administration, justice, health, mobility, leisure, accommodation, sport, food and drinks, shops, markets, worship, commercial zone, industrial zone, residential. We acknowledge that this ranking is to some extent arbitrary although cases of multiple assignment are altogether fairly rare. For instance, in Lagos, only three such cases are found out of 8,839 categorized features. Also note some features appear in multiple layers and we make sure to remove duplicates that we identify via the unique OSM identifier assigned to each feature.

We then proceed with the characterization of locations visited by users. For each city, we consider the unique set of visitor-locations over which we superimpose the constructed OSM-based dataset of categorized places (see Table 1.A.2).

Table 1.A.2: Matching rates between OSM features and visitors' locations, by city.

City	Visitors	Visitors matched		Visitor-locations	Visitor-locations matched	
		N	%		*N*	%
Lagos	6,689	6,053	90%	965,076	642,304	66.6%
Abuja	4,086	3,293	80.6%	506,868	275,808	54.4%
Nairobi	1,583	1,090	68.9%	511,531	276,807	54.1%
Mombasa	954	587	61.5%	93,608	41,538	44.4%
Dar es Salaam	2,040	1,391	68.2%	503,085	198,976	39.6%
Dodoma	804	800	99.5%	77,823	77,064	99%
Total	16,156	13,214	81.8%	2,657,991	1,512,497	56.9%

Note: This table shows the matching rates between OSM features, visitors and locations visited for the cities we considered in our analysis: Lagos, Abuja, Nairobi, Mombasa, Dar es Salaam and Dodoma. We count 16,156 visitors to those six cities for a total of 2,657,991 unique visitor-locations, of which nearly 57% are matched to an OSM feature. Overall, 82% of visitors have locations matched to an OSM feature which means that, for 4 visitors out of 5, we are able to characterize some of the places he visited in the host city.

[46] For instance, a building feature may be categorized via its name attribute which can be something like "*Somename* restaurant & hotel". The words "restaurant" and "hotel" result in the feature being classified in both the "food and drinks" and "travel" categories.

Appendix 1.B Definition of city boundaries and regional capitals

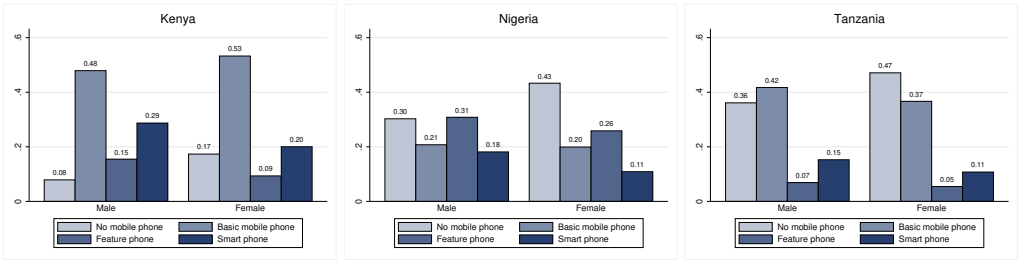
To define city boundaries, we use urban extents from the Global Rural-Urban mapping project v1.02 produced by Columbia University Center for International Earth Science Information Network (CIESIN). The original shapefile consists of polygons delineating urban settlements based on the point location of settlements, city-level population counts and 1995 DMSP-OLS nighttime lights to infer urban extents. Spatial extent for smaller settlements that do not emit detectable light are simply modelled with a buffer proportional to city size.[47] Given that most urban extents are based 1995 nighttime lights data, we apply a 3km buffer to GRUMP polygons to account for urban growth and better capture commuting zones. We overlay 2018 WorldPop population grids with GRUMP city polygons to obtain city-level population estimates and, for the sake of consistency, total population counts are also based on 2018 population grids. Cities which have boundaries less than 3km apart are merged. As a result, we find that there are 6, 39, and 10 cities of at least 200,000 people in Kenya, Nigeria and Tanzania respectively. Regional capitals are broadly understood as capital cities for subdivisions of the first administrative level.[48]

[47] The full documentation is available on the dedicated webpage https://sedac.ciesin.columbia.edu/data/set/grump-v1-urban-ext-polygons-rev02.

[48] More specifically, Kenya has 47 counties, Nigeria has 36 states and a Federal Capital Territory and there are 31 regions (or *mikoa*) in Tanzania. For the 19 regional capitals that have no boundaries defined in the GRUMP product, we overlay the ArcGIS labelled World Imagery basemap with our users' home location rasters and evaluate qualitatively whether some users are found within the built-up areas of the cities considered.

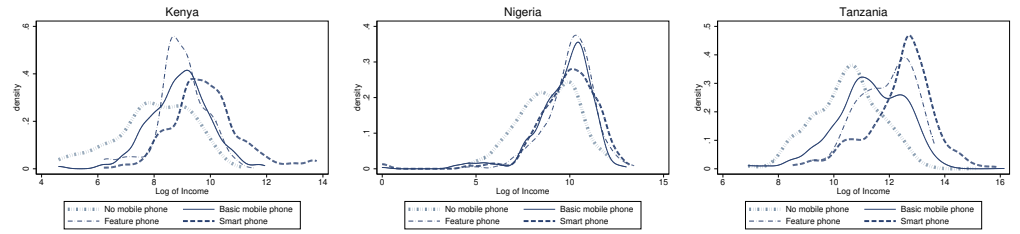
Appendix 1.C Sample selection: Comparing respondents by device ownership

Figure 1.C.1: Device ownership by gender.



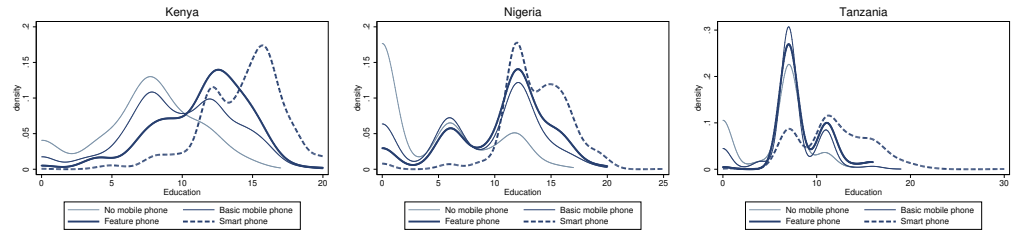
Note: This figure shows device ownership rates for female and male respondents. All figures use the sample weights provided.

Figure 1.C.2: Income and device ownership.



Note: This figure shows the distribution of income by device ownership. All figures use the sample weights provided. The figure shows that while there are differences in these distributions such that those with no mobile phone tend to have the lowest incomes, the distributions overlap across a large range of monthly incomes. This is particularly the case for individuals that have any type of mobile phone.

Figure 1.C.3: Education and device ownership.

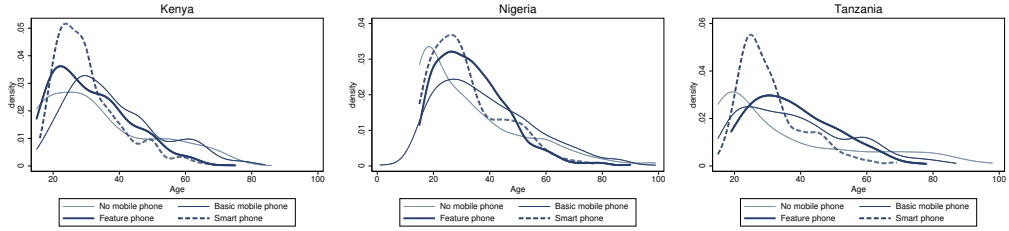


Note: This figure shows the distribution of education by device ownership. All figures use the sample weights provided. The figure highlights that these distributions are not distinct.

To further understand how smartphone users differ from the rest of the population and to interpret our data, the sectoral composition of smartphone users is relevant. The ICT Access and Usage Survey does not ask for the sector of employment, but does ask for income from different sources.[49] We use this information to

[49]The precise question is "How much income do you have every month in terms of ...?" If incomes are varying the interviewers are requested to ask for a typical amount.

Figure 1.C.4: Age and device ownership.



Note: This figure shows the distribution of age by device ownership. All figures use the sample weights provided. The figure highlights that these distributions are not distinct.

assign a main income source to each respondent in Table 1.C.1.[50]

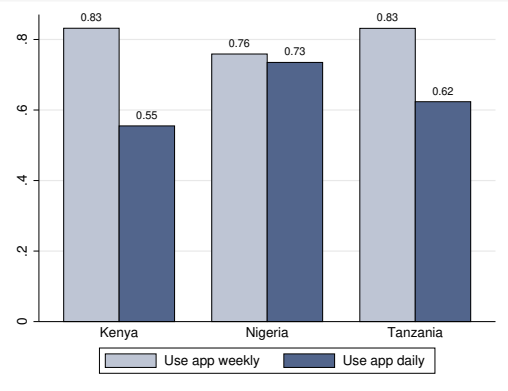
Table 1.C.1: Smartphone ownership and main source of income.

	Kenya		**Nigeria**		**Tanzania**	
	(1)	(2)	(3)	(4)	(5)	(6)
Rural						
Salary or wage	54.9	26.7	24.3	8.4	51.6	15.0
Agricultural produce/farming	9.9	34.0	8.7	25.3	18.6	34.2
Vending/trading	3.8	1.8	11.0	15.2		
Work you are doing at home	1.2	5.0	2.4	2.0	0.0	0.5
Income from your business	6.0	8.9	14.4	19.1	20.6	10.7
Property income/letting	0.0	0.1	0.0	0.4	0.0	0.4
Pension, social grant	0.0	1.4	3.3	0.9	1.7	0.6
Allowance	6.3	11.2	33.9	26.9	3.6	37.4
Scholarships	0.8	0.5	0.0	0.1		
Investments	6.6	4.5	1.9	0.3		
Other income	10.5	6.0	0.0	1.4	3.9	1.2
Urban						
Salary or wage	50.7	47.4	23.2	16.8	40.0	29.0
Agricultural produce/farming	1.6	4.1	0.4	2.3	0.8	6.5
Vending/trading	2.1	2.4	6.1	10.3	0.9	0.6
Work you are doing at home	2.7	2.3	0.4	2.6	0.3	0.7
Income from your business	18.3	18.6	29.1	26.5	16.9	18.0
Property income/letting	0.0	0.7	1.9	1.7	0.3	1.1
Pension, social grant	0.3	0.5	3.5	2.3	0.0	1.1
Allowance	21.9	20.5	33.0	34.7	40.3	41.6
Scholarships						
Investments	0.9	1.0	1.6	1.0	0.1	0.1
Other income	1.6	2.6	0.7	1.9	0.5	1.3

Note: This table shows the proportion of smartphone owners across different categories in columns (1), (3) and (5) and we compare this to the sample averages in columns (2), (4) and (6).

[50] About 1.7 percent of the sample report no income from any source, and 2.4 percent of the sample report equal amounts for two sectors. For respondents who reported to receive a pension, social grant, allowances, scholarships, investments or other income, we use the second source of income they report. We randomly allocate a main sector for respondents who report equal incomes from all other sources.

Figure 1.C.5: App usage of smartphone users.



Note: This figure shows the fraction of smartphone owners using apps weekly or daily. All figures use the sample weights provided. The figure illustrates that owners of smartphones use apps regularly.

Appendix 1.D Sample selection: Pairing users with DHS information

We link users' home locations with data from the most recently available Demographic and Health Survey (DHS) data to characterize areas where our users live: the 2014 standard DHS in Kenya, the 2018 standard DHS in Nigeria and the 2015-2016 standard DHS in Tanzania.[51] DHS data are geo-referenced at the cluster level and cluster coordinates are randomly displaced to maintain respondents' confidentiality.[52]

We first classify our users within urban and rural categories based on the overlay of users' home location with city polygons.[53] We then apply two criteria to associate each user with a set of DHS clusters. First, we select the set of DHS clusters located within a given distance from her home location (5km for urban users and 10km for rural users). This yields a set of DHS clusters that are comparable, in some sense, to the home location of our user. The number of these comparison clusters will be either zero or a strictly positive number of clusters. Not all these nearby clusters will offer valid comparisons, however. For example, a user at the outskirts of Dar es Salaam might be associated with a nearby rural cluster as well as a number of urban clusters. To ensure that we do not falsely assign an urban cluster as a comparison location for a rural user (or vice-versa), we add the second criterion that the cluster's average population density (calculated over a 5km buffer) must be within 25% of the average population density that we have computed for the user's home location. If this does not hold, we drop the DHS comparison cluster.

Following that methodology, we pair 70% of our users in the high-confidence sample with at least one DHS cluster (90% in Kenya, 66% in Nigeria, 72% in Tanzania). Some clusters are paired to more than one user so the matched DHS sample contains a number of duplicates. In practice, we construct a weighted subset of unique respondents within paired clusters, with weights being equal to the number of users each corresponding cluster is matched to. We call the subset of respondents within paired clusters the ''matched DHS'' sample.[54] Unsurprisingly, unmatched users are found in low density areas where the probability of selection in the DHS is lower by design - the average experienced density for unmatched users is estimated at 2,496 inh./km^2 against 8,835 inh./km^2 for users with at least

[51] More information on sampling design at https://dhsprogram.com/.

[52] Urban clusters are displaced by up to 2 kilometers and rural clusters by up to 5 kilometers with 1% of rural clusters being displaced up to 10 kilometers. The displacement is restricted such that clusters stay within the administrative 2 area where the survey was conducted.

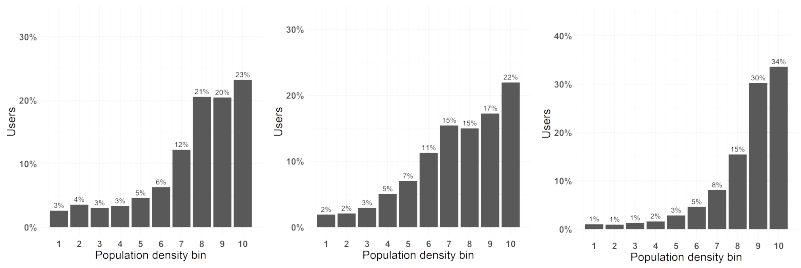
[53] See Appendix Section 1.B for details on the definition of city boundaries.

[54] Some clusters are paired to more than one user so the matched DHS sample contains a number of duplicates. It is in fact equivalent to the weighted subset of respondents in clusters paired to at least one user, with weights being equal to the number of users the corresponding cluster is matched to.

one paired cluster. In order to examine potential differences between our users and the population as a whole, we conduct t-tests for equality of means between the raw DHS and matched DHS samples on a range of household characteristics. We produce results for rural and urban sub-samples separately to account for both the prevalence of urban users in our sample and the lower matching rate in low density areas, which together may lead to results being mainly driven by the urban component of the sample. We produce t-tests comparing our two weighted data streams, with bootstrapped standard errors robust to heteroskedasticity. The survey weights are used for the reference DHS sample while those of the matched DHS sample correspond to the number of users each cluster is paired with.

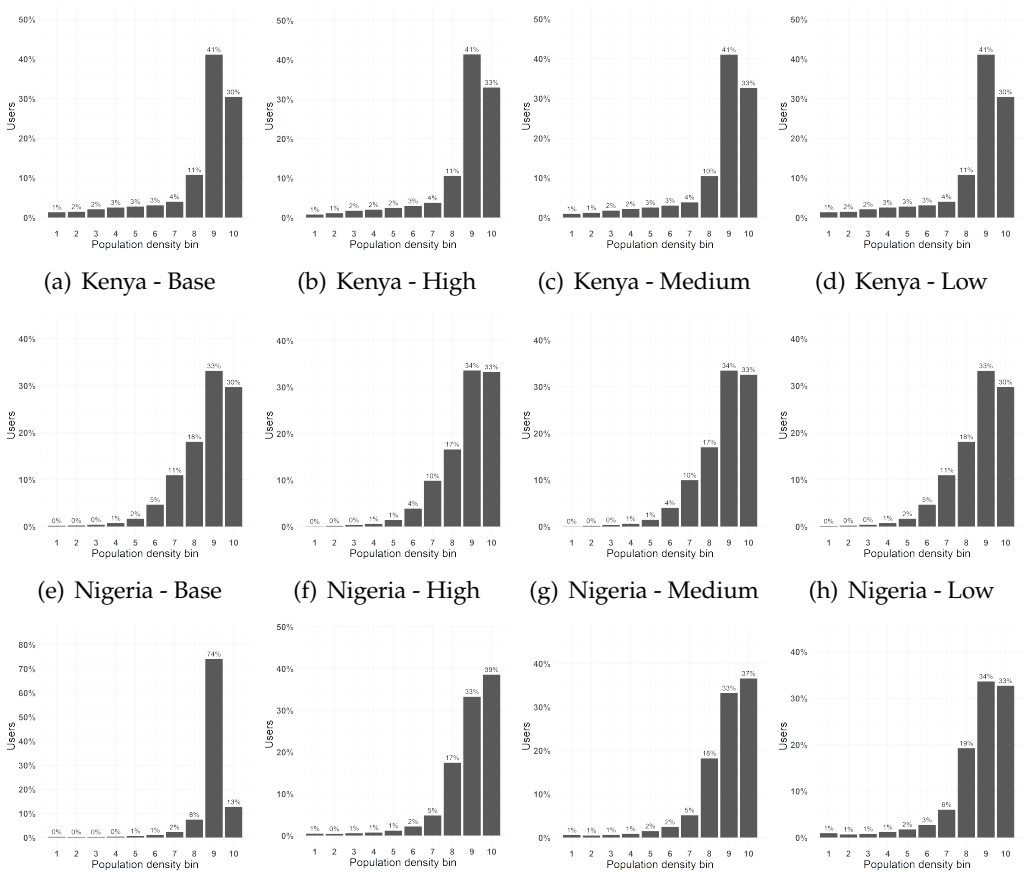
Appendix 1.E Additional tables and figures

Figure 1.E.1: Fraction of users by population density decile, Landscan.



(a) Kenya (b) Nigeria (c) Tanzania

Figure 1.E.2: Fraction of users by population density decile, WorldPop.



(a) Kenya - Base (b) Kenya - High (c) Kenya - Medium (d) Kenya - Low

(e) Nigeria - Base (f) Nigeria - High (g) Nigeria - Medium (h) Nigeria - Low

(i) Tanzania - Base (j) Tanzania - High (k) Tanzania - Medium (l) Tanzania - Low

Note: This figure shows the fraction of users by population density decile for the base, low-, medium, and high-confidence samples. Population data are taken from WorldPop.

Table 1.E.1: T-tests for equality of means between matched DHS and DHS samples, Kenya.

	Variable	DHS	Matched DHS	Difference	SE	p-value
All	Household size	3.99	3.08	-0.91	0.02	0.000***
	Age of HH head	42.93	37.29	-5.64	0.11	0.000***
	Education of HH head	8.00	10.32	2.33	0.03	0.000***
	Access to electricity	0.37	0.80	0.43	0.01	0.000***
	Radio	0.67	0.74	0.06	0.01	0.000***
	Television	0.35	0.64	0.29	0.01	0.000***
	Rooms per adult	0.66	0.66	0.00	0.00	0.522
	Access to piped water	0.44	0.79	0.35	0.01	0.000***
	Constructed floor	0.53	0.90	0.37	0.01	0.000***
	Constructed walls	0.64	0.92	0.28	0.01	0.000***
	Constructed roof	0.89	0.99	0.10	0.01	0.000***
Urban	Household size	3.28	3.02	-0.26	0.03	0.000***
	Age of HH head	38.60	36.82	-1.78	0.17	0.000***
	Education of HH head	9.90	10.46	0.56	0.05	0.000***
	Access to electricity	0.68	0.83	0.15	0.02	0.000***
	Radio	0.74	0.74	0.00	0.01	0.774
	Television	0.56	0.65	0.09	0.02	0.000***
	Rooms per adult	0.68	0.66	-0.02	0.01	0.001***
	Access to piped water	0.71	0.82	0.11	0.02	0.000***
	Constructed floor	0.82	0.92	0.10	0.01	0.000***
	Constructed walls	0.86	0.94	0.07	0.01	0.000***
	Constructed roof	0.98	0.99	0.01	0.00	0.002***
Rural	Household size	4.52	4.33	-0.19	0.02	0.000***
	Age of HH head	46.15	46.60	0.45	0.16	0.005***
	Education of HH head	6.58	7.58	0.99	0.04	0.000***
	Access to electricity	0.13	0.21	0.08	0.01	0.000***
	Radio	0.63	0.70	0.07	0.01	0.000***
	Television	0.19	0.25	0.07	0.01	0.000***
	Rooms per adult	0.64	0.67	0.03	0.00	0.000***
	Access to piped water	0.24	0.25	0.01	0.02	0.464
	Constructed floor	0.31	0.38	0.07	0.01	0.000***
	Constructed walls	0.46	0.46	0.00	0.02	0.949
	Constructed roof	0.82	0.93	0.11	0.01	0.000***

Note: This table compares the means between the overall DHS sample and the "Matched DHS" sample (DHS clusters with which we can match smartphone app users). We show a t-test that compares the two data sets, with bootstrapped standard errors robust to heteroskedasticity. Survey weights are used for the reference DHS sample, while those for the matched DHS sample correspond to the number of users each cluster is paired with.

Table 1.E.2: T-tests for equality of means between DHS and matched DHS samples, Nigeria.

	Variable	DHS	Matched DHS	Difference	SE	p-value
All	Household size	4.69	3.83	-0.86	0.02	0.000***
	Age of HH head	45.29	45.17	-0.12	0.12	0.344
	Education of HH head	7.43	11.52	4.10	0.04	0.000***
	Access to electricity	0.60	0.98	0.39	0.01	0.000***
	Radio	0.61	0.84	0.24	0.01	0.000***
	Television	0.49	0.90	0.41	0.01	0.000***
	Rooms per adult	0.74	0.65	-0.09	0.00	0.000***
	Access to piped water	0.11	0.14	0.03	0.01	0.003***
	Constructed floor	0.74	0.96	0.23	0.01	0.000***
	Constructed walls	0.84	1.00	0.16	0.01	0.000***
	Constructed roof	0.89	1.00	0.11	0.01	0.000***
Urban	Household size	4.44	3.83	-0.61	0.03	0.000***
	Age of HH head	45.21	45.18	-0.02	0.18	0.900
	Education of HH head	9.66	11.56	1.91	0.06	0.000***
	Access to electricity	0.88	0.99	0.11	0.01	0.000***
	Radio	0.72	0.85	0.13	0.01	0.000***
	Television	0.73	0.90	0.18	0.01	0.000***
	Rooms per adult	0.72	0.65	-0.08	0.01	0.000***
	Access to piped water	0.14	0.14	-0.01	0.01	0.572
	Constructed floor	0.89	0.96	0.08	0.01	0.000***
	Constructed walls	0.95	1.00	0.04	0.01	0.000***
	Constructed roof	0.98	1.00	0.02	0.00	0.000***
Rural	Household size	4.85	3.92	-0.93	0.03	0.000***
	Age of HH head	45.34	44.77	-0.57	0.16	0.000***
	Education of HH head	6.03	10.23	4.20	0.06	0.000***
	Access to electricity	0.42	0.84	0.42	0.02	0.000***
	Radio	0.54	0.67	0.14	0.01	0.000***
	Television	0.35	0.77	0.43	0.01	0.000***
	Rooms per adult	0.75	0.75	0.01	0.01	0.503
	Access to piped water	0.09	0.14	0.05	0.01	0.000***
	Constructed floor	0.64	0.96	0.32	0.01	0.000***
	Constructed walls	0.77	0.98	0.21	0.01	0.000***
	Constructed roof	0.83	0.99	0.16	0.01	0.000***

Note: This table compares the means between the overall DHS sample and the "Matched DHS" sample (DHS clusters with which we can match smartphone app users). We show a t-test that compares the two data sets, with bootstrapped standard errors robust to heteroskedasticity. Survey weights are used for the reference DHS sample, while those for the matched DHS sample correspond to the number of users each cluster is paired with.

Table 1.E.3: T-tests for equality of means between DHS and matched DHS samples, Tanzania.

	Variable	DHS	Matched DHS	Difference	SE	p-value
All	Household size	5.03	4.33	-0.70	0.04	0.000***
	Age of HH head	45.43	41.66	-3.77	0.22	0.000***
	Education of HH head	5.90	8.33	2.42	0.05	0.000***
	Access to electricity	0.23	0.78	0.55	0.02	0.000***
	Radio	0.52	0.66	0.14	0.01	0.000***
	Television	0.21	0.65	0.44	0.02	0.000***
	Rooms per adult	0.61	0.59	-0.02	0.00	0.000***
	Access to piped water	0.38	0.67	0.29	0.02	0.000***
	Constructed floor	0.44	0.95	0.51	0.02	0.000***
	Constructed walls	0.80	0.98	0.18	0.01	0.000***
	Constructed roof	0.75	0.99	0.24	0.01	0.000***
Urban	Household size	4.54	4.30	-0.24	0.07	0.001***
	Age of HH head	42.22	41.56	-0.67	0.37	0.073*
	Education of HH head	8.01	8.40	0.39	0.10	0.000***
	Access to electricity	0.63	0.80	0.17	0.03	0.000***
	Radio	0.65	0.66	0.01	0.02	0.462
	Television	0.52	0.67	0.14	0.03	0.000***
	Rooms per adult	0.62	0.59	-0.03	0.01	0.000***
	Access to piped water	0.67	0.67	0.00	0.04	0.980
	Constructed floor	0.87	0.96	0.09	0.02	0.000***
	Constructed walls	0.96	0.98	0.03	0.01	0.005***
	Constructed roof	0.97	0.99	0.02	0.01	0.002***
Rural	Household size	5.21	5.04	-0.16	0.05	0.002***
	Age of HH head	46.61	44.40	-2.21	0.27	0.000***
	Education of HH head	5.13	6.28	1.14	0.07	0.000***
	Access to electricity	0.08	0.31	0.23	0.02	0.000***
	Radio	0.47	0.59	0.12	0.01	0.000***
	Television	0.09	0.29	0.20	0.02	0.000***
	Rooms per adult	0.61	0.63	0.03	0.01	0.000***
	Access to piped water	0.27	0.54	0.27	0.03	0.000***
	Constructed floor	0.27	0.65	0.37	0.02	0.000***
	Constructed walls	0.73	0.85	0.12	0.02	0.000***
	Constructed roof	0.67	0.89	0.22	0.02	0.000***

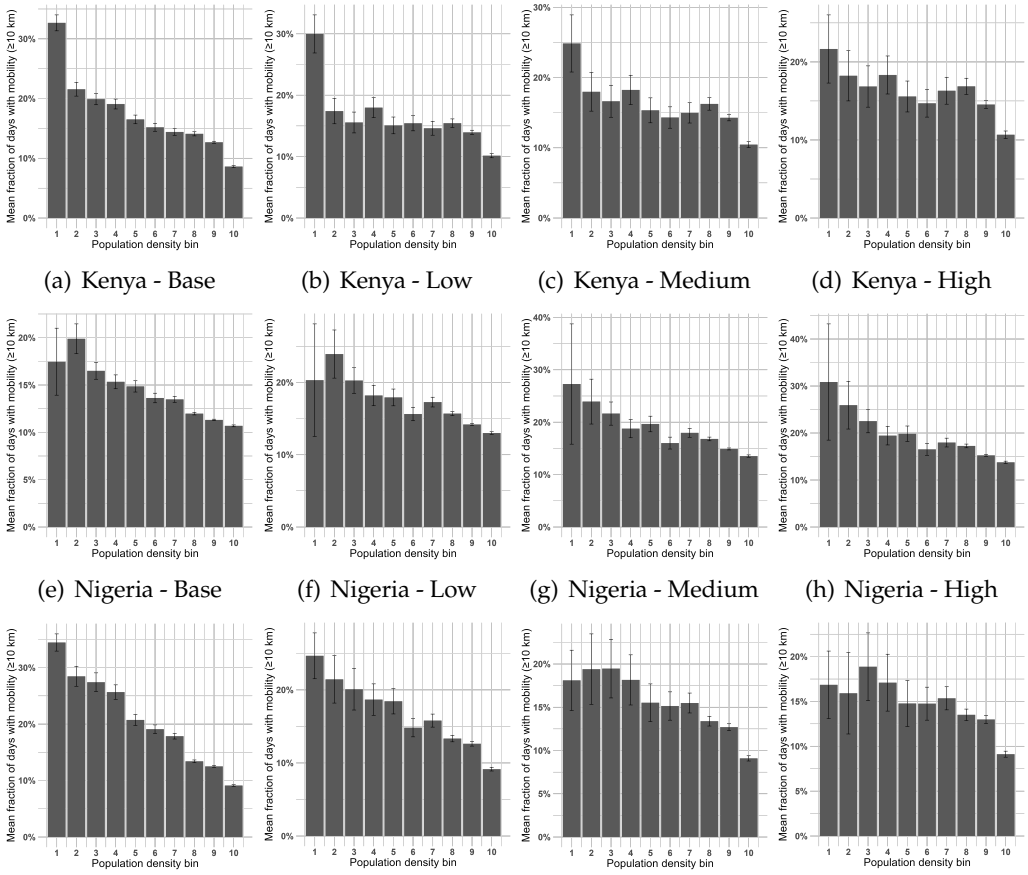
Note: This table compares the means between the overall DHS sample and the "Matched DHS" sample (DHS clusters with which we can match smartphone app users). We show a t-test that compares the two data sets, with bootstrapped standard errors robust to heteroskedasticity. Survey weights are used for the reference DHS sample, while those for the matched DHS sample correspond to the number of users each cluster is paired with.

Table 1.E.4: Mobility metrics for the high-confidence set and the overall sample.

	Kenya		Nigeria		Tanzania	
	Overall	High-confidence	Overall	High-confidence	Overall	High-confidence
Fraction of days with mobility >10km	0.13	0.14	0.11	0.15	0.13	0.12
Mean distance away from home	40.23	37.10	34.65	38.63	55.69	52.17

Note: This table shows the fraction of days with mobility > 10km and mean distance away from home for different samples.

Figure 1.E.3: Fraction of days with mobility beyond 10km by density bin, for all confidence sets.



(a) Kenya - Base (b) Kenya - Low (c) Kenya - Medium (d) Kenya - High

(e) Nigeria - Base (f) Nigeria - Low (g) Nigeria - Medium (h) Nigeria - High

(i) Tanzania - Base (j) Tanzania - Low (k) Tanzania - Medium (l) Tanzania - High

Note: This figure shows the fraction of days on which a user is seen more than 10km away from their home location by density decile over the period of a year.

Table 1.E.5: Mean fraction of days with mobility at 3 distance thresholds for 3 subsets, by country.

	Distance criterion	**HIGH**	**MED**	**LOW**
Kenya	0 km	39.8%	39.5%	38.8%
	10 km	13.8%	13.5%	13.2%
	20 km	7.2%	7.2%	7.3%
Nigeria	0 km	47%	46.7%	45.9%
	10 km	15.2%	14.9%	14.2%
	20 km	8.9%	8.7%	8.4%
Tanzania	0 km	42.7%	42.7%	43.1%
	10 km	11.8%	11.8%	12%
	20 km	7.3%	7.4%	7.8%

Note: This table shows the fraction of days with mobility for different thresholds and samples.

Table 1.E.6: Average distribution of pings across visited density bins by home density bin, transit pings included.

		Home density bin									
		1	2	3	4	5	6	7	8	9	10
	1	40.5%	7%	2.1%	1.2%	1.7%	1.4%	1.3%	1.6%	0.6%	0.3%
	2	8.5%	28.7%	13.2%	2.4%	1.7%	1.6%	1.3%	1.4%	0.7%	0.5%
	3	3.7%	8.5%	16.3%	9.3%	6%	3.1%	3%	2.1%	1.1%	0.6%
	4	3.4%	3.9%	13.5%	14.2%	10.7%	6.4%	3.9%	2.1%	1.3%	0.8%
Visited	5	6%	4.5%	8.5%	11.1%	12.7%	10.2%	5%	4.2%	1.7%	0.9%
density	6	3.4%	2.8%	3.9%	6.2%	9.5%	15.9%	8.2%	4.8%	1.9%	1.1%
	7	2.4%	1.7%	5.3%	7.3%	7.6%	12%	14.1%	8.5%	3.3%	1.9%
	8	7.7%	8.8%	10.2%	10.6%	13.9%	15.1%	19.1%	22.1%	8.5%	4.2%
	9	16.8%	23.3%	18.1%	28.1%	25.8%	25.6%	32.8%	39.4%	54%	37.7%
	10	7.6%	10.8%	8.9%	9.6%	10.4%	8.9%	11.4%	13.8%	26.8%	52%

(a) Kenya

		Home density bin									
		1	2	3	4	5	6	7	8	9	10
	1	7.2%	2.1%	2%	0.8%	0.4%	0.2%	0.1%	0.1%	0.1%	0.1%
	2	7.9%	10.9%	6.6%	1.5%	0.7%	0.4%	0.3%	0.2%	0.2%	0.1%
	3	3.2%	7.9%	10%	8.1%	1.6%	1%	0.5%	0.3%	0.2%	0.1%
	4	3.4%	4.1%	10.1%	6.5%	5.1%	2.6%	0.9%	0.5%	0.4%	0.3%
Visited	5	2.9%	5.1%	8.1%	10.2%	10.8%	5.7%	2.6%	1.4%	1%	0.6%
density	6	9.5%	4.4%	4.3%	10.7%	14.8%	21.4%	8.4%	3.4%	2%	1.2%
	7	6.1%	12.8%	11.4%	12.4%	15.6%	21.8%	26.6%	12%	4.9%	2.3%
	8	18.2%	15.6%	11.6%	13.5%	13.7%	13.7%	22.7%	30.5%	14.2%	4.8%
	9	29.4%	26%	25%	24.6%	25.5%	20.9%	26.4%	40.2%	56.9%	19.1%
	10	12.3%	11.1%	10.8%	11.7%	11.8%	12.4%	11.5%	11.3%	20.2%	71.5%

(b) Nigeria

		Home density bin									
		1	2	3	4	5	6	7	8	9	10
	1	41.3%	11.5%	2.3%	1.8%	2.3%	1%	1%	0.5%	0.3%	0.2%
	2	3.2%	17.7%	7.3%	5.5%	2.1%	2%	1.2%	0.6%	0.3%	0.1%
	3	1.5%	6.3%	12.9%	9%	8%	2%	1.6%	0.7%	0.3%	0.2%
	4	2.1%	8.2%	10.4%	12%	10.7%	3.8%	2.4%	0.9%	0.5%	0.3%
Visited	5	1.8%	6.1%	9.6%	9.3%	9.8%	8.7%	4.3%	1.7%	0.8%	0.3%
density	6	3.3%	1.3%	4.2%	11.7%	13.5%	16.9%	9.8%	2.4%	1.3%	0.6%
	7	3.2%	9.5%	6%	12.3%	9.6%	16.4%	25.2%	8.4%	3%	1.2%
	8	12.8%	12.4%	14.7%	13.1%	13.6%	16.7%	25.1%	40.2%	15%	4.8%
	9	13.7%	18.6%	20.4%	14.3%	21.4%	23%	19%	30.5%	50.6%	22.7%
	10	17.1%	8.4%	12.2%	11%	8.9%	9.5%	10.5%	14%	27.8%	69.6%

(c) Tanzania

Note: These matrices show the average fraction of non-home pings of users residing in home density bin i for visited density bin j over the period of a year.

Table 1.E.7: Average distribution of pings across visited density bin, by home density bin, transit pings excluded.

		Home density bin									
		1	2	3	4	5	6	7	8	9	10
	1	40%	7.1%	2.1%	1.2%	1.5%	1.4%	1.3%	1.5%	0.6%	0.3%
	2	8.2%	28.5%	13.2%	2.3%	1.6%	1.5%	1.3%	1.3%	0.7%	0.5%
	3	3.5%	8.6%	16.3%	9.1%	5.9%	3%	3%	2%	1.1%	0.6%
	4	3.3%	3.9%	13.3%	14.3%	10.8%	6.2%	3.8%	2%	1.2%	0.8%
Visited	5	6%	4.5%	8.5%	11.2%	12.3%	10.2%	4.9%	4.1%	1.7%	0.9%
density	6	3.4%	2.7%	3.7%	6.2%	9.6%	15.8%	8.2%	4.7%	1.9%	1.1%
	7	2.8%	1.6%	5.3%	7.2%	7.4%	11.8%	14.1%	8.4%	3.3%	1.9%
	8	7.4%	8.7%	10%	10.4%	13.7%	15.1%	18.9%	21.9%	8.5%	4.2%
	9	17.2%	23.6%	18.2%	28.5%	26.4%	26%	33.2%	40%	54.3%	37.9%
	10	8.1%	10.8%	9.2%	9.7%	10.7%	9%	11.4%	14%	26.9%	52.1%

(a) Kenya

		Home density bin									
		1	2	3	4	5	6	7	8	9	10
	1	7%	2.1%	1.9%	0.8%	0.4%	0.2%	0.1%	0.1%	0.1%	0%
	2	8.2%	10.8%	6.7%	1.5%	0.7%	0.4%	0.3%	0.2%	0.1%	0.1%
	3	3.3%	7.9%	10.1%	8.2%	1.6%	0.9%	0.5%	0.3%	0.2%	0.1%
	4	3.4%	4.1%	10.1%	6.7%	5.1%	2.6%	0.9%	0.5%	0.4%	0.2%
Visited	5	2.8%	5.1%	8.1%	9.9%	10.8%	5.6%	2.6%	1.4%	1%	0.5%
density	6	9.6%	4.4%	4.3%	10.5%	14.8%	21.4%	8.4%	3.4%	2%	1.2%
	7	6%	12.7%	11.4%	12.3%	15.5%	21.8%	26.6%	12.1%	4.8%	2.3%
	8	18.2%	15.6%	11.6%	13.6%	13.8%	13.8%	22.7%	30.6%	14.2%	4.8%
	9	29.3%	26.1%	24.8%	24.8%	25.5%	20.9%	26.4%	40.2%	57%	19.1%
	10	12.2%	11.1%	10.9%	11.8%	11.8%	12.4%	11.5%	11.3%	20.2%	71.7%

(b) Nigeria

		Home density bin									
		1	2	3	4	5	6	7	8	9	10
	1	41.5%	11.8%	2.3%	1.8%	2.3%	1%	1%	0.5%	0.3%	0.2%
	2	3.1%	17.3%	7.3%	5.5%	2.1%	2.1%	1.1%	0.5%	0.2%	0.1%
	3	1.5%	6.1%	13%	8.9%	7.9%	1.9%	1.5%	0.6%	0.3%	0.2%
	4	2%	8%	10.4%	12%	10.6%	3.8%	2.3%	0.7%	0.4%	0.3%
Visited	5	1.7%	6.3%	9.6%	9.4%	9.6%	8.6%	4.2%	1.6%	0.7%	0.3%
density	6	3.1%	1.1%	4.2%	11.6%	13.3%	16.7%	9.6%	2.2%	1.2%	0.5%
	7	3%	9.3%	5.9%	12.4%	9.5%	16.2%	25.1%	8.2%	2.8%	1.2%
	8	12.8%	12.8%	14.7%	13.1%	13.8%	16.5%	25.1%	40.3%	15%	4.7%
	9	14.1%	17.9%	20.5%	14.3%	22.1%	23.6%	19.3%	30.9%	51%	22.9%
	10	17.2%	9.6%	12.2%	11.1%	8.7%	9.6%	10.8%	14.4%	28%	69.8%

(c) Tanzania

Note: These matrices show the average fraction of non-home pings of users residing in home density bin i for visited density bin j over the period of a year, excluding transit pings.

Table 1.E.8: Share of users by home bin-visited bin pair, transit pings excluded.

		Home density bin									
		1	2	3	4	5	6	7	8	9	10
	1	71.2%	32.9%	14%	11.1%	11.4%	13.2%	12.1%	13.9%	8.7%	5.6%
	2	43.2%	60.8%	37.7%	24.9%	18.9%	17.1%	17.2%	19.3%	13.3%	9.6%
	3	25.2%	45.6%	55.1%	41.1%	34.9%	28.4%	26.5%	24.1%	17.6%	13.2%
	4	34.2%	32.9%	51.3%	56.6%	46.2%	37.7%	33.9%	27.5%	22.1%	16.5%
Visited	5	29.7%	25.3%	42.3%	51.5%	52.4%	48.3%	37.5%	34.3%	24.1%	17.8%
density	6	27%	24.7%	28.7%	46.1%	46.2%	54.6%	46.8%	37.3%	25.5%	18.4%
	7	27%	27.8%	34.7%	42.4%	43.8%	55.8%	57.7%	47.5%	34.1%	23.6%
	8	42.3%	44.3%	45.3%	55.9%	56.8%	60.8%	68.1%	69.3%	50.3%	35.4%
	9	55.9%	53.8%	53.6%	65.3%	65.1%	67.8%	72.1%	79.8%	89.8%	76%
	10	32.4%	36.1%	30.2%	41.4%	37.3%	40.1%	45.4%	51.3%	69.9%	88.6%

(a) Kenya

		Home density bin									
		1	2	3	4	5	6	7	8	9	10
	1	35.7%	18.8%	18.8%	6.3%	5.7%	3.5%	2.9%	2.7%	2.2%	1.3%
	2	23.8%	31.9%	35%	12.2%	12.1%	8.3%	6.2%	5.5%	4.6%	2.8%
	3	26.2%	29%	40.6%	31.6%	18%	12.5%	9.6%	8%	6.5%	4.2%
	4	31%	26.1%	44.9%	35%	31.7%	21.7%	14.1%	11.3%	10.4%	6.4%
Visited	5	23.8%	33.3%	42.7%	45.3%	50.6%	37.1%	26.2%	20.2%	19.4%	14.6%
density	6	33.3%	33.3%	36.8%	53%	59.5%	68.6%	45.2%	30.9%	26.2%	17%
	7	42.9%	55.8%	49.6%	52.8%	63.6%	69.9%	75.9%	55.9%	39.3%	25.1%
	8	71.4%	58%	54.3%	58.4%	61.1%	59.6%	72.5%	81%	63.4%	37.6%
	9	76.2%	61.6%	62.8%	62.5%	66.8%	63.9%	68.3%	81.1%	91.4%	64.5%
	10	42.9%	44.2%	41.9%	44.5%	49.9%	47.7%	46.7%	46.7%	61.7%	95.3%

(b) Nigeria

		Home density bin									
		1	2	3	4	5	6	7	8	9	10
	1	73.6%	33.8%	18.2%	14.3%	13.4%	8.5%	10.2%	7.9%	6.5%	3.6%
	2	18.7%	50%	39.1%	27.9%	21.2%	13.1%	13%	9.7%	7.1%	4.1%
	3	11%	38.2%	43.6%	38.8%	29%	18.3%	13.9%	11%	8.6%	5.1%
	4	13.2%	35.3%	40%	40.1%	37.8%	22.4%	19.3%	13.1%	10%	6.4%
Visited	5	16.5%	29.4%	42.7%	39.5%	36.4%	41.4%	25.6%	17.8%	12.2%	7%
density	6	18.7%	22.1%	35.5%	40.8%	45.2%	49.6%	41.1%	22.4%	15.9%	9.3%
	7	29.7%	38.2%	42.7%	46.3%	40.6%	53%	64%	40.8%	25.3%	14.9%
	8	42.9%	42.6%	50%	46.9%	50.2%	54.8%	61.9%	82.3%	56.5%	33.2%
	9	40.7%	50%	54.5%	48.3%	55.3%	58.9%	55.8%	68.5%	88.4%	66%
	10	39.6%	35.3%	30.9%	38.1%	31.8%	38.3%	39.5%	44.7%	64.2%	93.4%

(c) Tanzania

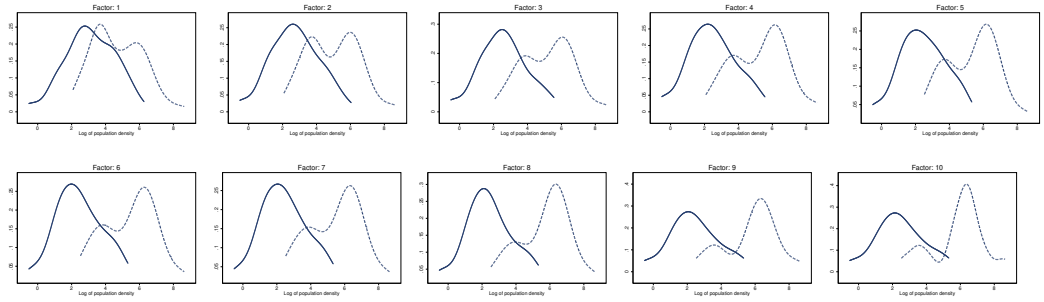
Note: These matrices show the proportion of users residing in home density bin i that are seen at least once in visited density bin j over the period of a year, transit pings excluded.

Table 1.E.9: Origin of visitors in top 5 cities.

Kenya									
Nairobi (1,699 visitors)		Mombasa (953 visitors)		Nakuru (891 visitors)		Eldoret (448 visitors)		Kisumu (437 visitors)	
Origin	*Visitors*	*Origin*	*Visitors*	*Origin*	*Visitors*	*Origin*	*Visitors*	*Origin*	*Visitors*
Mombasa	20.2%	Nairobi	68.4%	Nairobi	62.5%	Nairobi	51.3%	Nairobi	57%
Nakuru	4.9%	Nakuru	1.5%	Eldoret	3.1%	Mombasa	3.3%	Mombasa	4.6%
Kisumu	4.1%	Kisumu	0.6%	Mombasa	2.9%	Kisumu	2.9%	Eldoret	2.3%
Eldoret	4.1%	Eldoret	0.5%	Kisumu	2%	Nakuru	2.2%	Nakuru	1.4%
Garissa	1.1%	Garissa	0.1%	Garissa	0.1%	-	-	-	-
Non-urban	65.6%	Non-urban	28.9%	Non-urban	29.3%	Non-urban	40.2%	Non-urban	34.8%
Nigeria									
Lagos (5,258 visitors)		Kano (807 visitors)		Ibadan (2,916 visitors)		Abuja (3,232 visitors)		Kaduna (1,296 visitors)	
Origin	*Visitors*	*Origin*	*Visitors*	*Origin*	*Visitors*	*Origin*	*Visitors*	*Origin*	*Visitors*
Abuja	21.9%	Abuja	43.5%	Lagos	68.7%	Lagos	47%	Abuja	54.9%
Ibadan	13.1%	Lagos	18.5%	Abuja	6.6%	Kaduna	8.8%	Lagos	12%
Abeokuta	7.4%	Kaduna	11%	Abeokuta	3.8%	Port Harc.	5.3%	Kano	10.3%
Shagamu	6.4%	Maiduguri	2.9%	Ilorin	2.9%	Kano	5.2%	Zaria	5.9%
Port Harc.	6.4%	Zaria	2.9%	Shagamu	2.7%	Jos	3.2%	Katsina	1.7%
Other urb.	5.7%	Other urb.	2.5%	Other urb.	2.4%	Other urb.	2.6%	Other urb.	1.2%
Non-urban	39.1%	Non-urban	18.8%	Non-urban	12.9%	Non-urban	27.9%	Non-urban	13.9%
Tanzania									
Dar Es Salaam (1,850 visitors)		Zanzibar (743 visitors)		Mwanza (704 visitors)		Arusha (859 visitors)		Mbeya (395 visitors)	
Origin	*Visitors*	*Origin*	*Visitors*	*Origin*	*Visitors*	*Origin*	*Visitors*	*Origin*	*Visitors*
Arusha	9.7%	Dar Es Sa.	53.3%	Dar Es Sa.	32.4%	Dar Es Sa.	39.5%	Dar Es Sa.	38.2%
Zanzibar	8.9%	Arusha	4%	Arusha	3.1%	Moshi	10.4%	Mwanza	2.8%
Mwanza	6.7%	Mwanza	0.8%	Dodoma	1.3%	Mwanza	3%	Arusha	2.3%
Morogoro	6%	Moshi	0.8%	Mbeya	0.9%	Dodoma	2.3%	Dodoma	1.8%
Dodoma	4.3%	Dodoma	0.8%	Moshi	0.7%	Zanzibar	2.2%	Morogoro	1.5%
Other urb.	3.5%	Other urb.	0.3%	Other urb.	0.6%	Other urb.	1.6%	Other urb.	0.8%
Non-urban	61%	Non-urban	40%	Non-urban	61.1%	Non-urban	41%	Non-urban	52.7%

Note: This table shows the origin of visitors for the five most populated cities. Origin and destination city boundaries are defined using 3km-buffered GRUMP polygons. Visitors are defined as being seen at least once in a location over the year. "Non-urban" refers to locations outside boundaries of cities with 200,000 or more residents. "Other urb." refers to all cities that are not in the top 5 origin cities.

Figure 1.E.4: Differences in flows between locations, Kenya



Note: This figure shows how the distributions of $\ln(V_k(1,2)/N_{k_1} * 1000)$ (dashed line) and $\ln(V_k(2,1)/N_{k_2} * 1000)$ (solid line) vary as we change the ratio of populations at origin and destination. We multiply the number of visits per resident by 1000 and take logs for expositional purposes.

Table 1.E.10: Top 5 destinations of residents from top 5 cities.

Kenya									
Nairobi (11,290 residents)		Mombasa (1,683 residents)		Nakuru (413 residents)		Eldoret (340 residents)		Kisumu (258 residents)	
Destination	*Residents*	*Destination*	*Residents*	*Destination*	*Residents*	*Destination*	*Residents*	*Destination*	*Residents*
Mombasa	5.8%	Nairobi	20.4%	Nairobi	20.1%	Nairobi	20.3%	Nairobi	27.1%
Nakuru	4.9%	Nakuru	1.5%	Mombasa	3.4%	Nakuru	8.2%	Nakuru	7%
Kisumu	2.2%	Kisumu	1.2%	Eldoret	2.4%	Kisumu	2.9%	Eldoret	5%
Eldoret	2%	Eldoret	0.9%	Kisumu	1.5%	Mombasa	1.5%	Mombasa	2.3%
Garissa	0.3%	Garissa	0.1%	Garissa	0.2%	Garissa	0.6%	-	-
Non-urban	31.4%	Non-urban	24.4%	Non-urban	37%	Non-urban	38.2%	Non-urban	51.9%
Nigeria									
Lagos (35,957 residents)		Kano (1,496 residents)		Ibadan (2,555 residents)		Abuja (7,988 residents)		Kaduna (1,303 residents)	
Destination	*Residents*	*Destination*	*Residents*	*Destination*	*Residents*	*Destination*	*Residents*	*Destination*	*Residents*
Shagamu	5.9%	Abuja	11.2%	Lagos	26.9%	Lagos	14.4%	Abuja	21.8%
Ibadan	5.6%	Kaduna	9%	Shagamu	9.2%	Kaduna	8.9%	Zaria	10.4%
Abuja	4.2%	Lagos	6.7%	Abeokuta	3.8%	Kano	4.4%	Kano	6.8%
Abeokuta	2.8%	Zaria	5.9%	Oshogbo	3.5%	Zaria	3%	Lagos	5.8%
Benin City	2.1%	Katsina	2.2%	Abuja	3.3%	Port Harc.	2.7%	Katsina	2.2%
Other urb.	14.1%	Other urb.	12.1%	Other urb.	20.8%	Other urb.	33.6%	Other urb.	19.1%
Non-urban	20.9%	Non-urban	21.9%	Non-urban	25.5%	Non-urban	32.2%	Non-urban	28.2%
Tanzania									
Dar Es Salaam (10,370 residents)		Zanzibar (832 residents)		Mwanza (963 residents)		Arusha (1,253 residents)		Mbeya (439 residents)	
Destination	*Residents*	*Destination*	*Residents*	*Destination*	*Residents*	*Destination*	*Residents*	*Destination*	*Residents*
Morogoro	4.9%	Dar Es Sa.	19.8%	Dar Es Sa.	12.9%	Moshi	14.9%	Dar Es Sa.	14.6%
Zanzibar	3.8%	Arusha	2.3%	Dodoma	3.6%	Dar Es Sa.	14.3%	Morogoro	3.4%
Dodoma	3.7%	Dodoma	1.4%	Arusha	2.7%	Dodoma	2.9%	Dodoma	3%
Arusha	3.3%	Tanga	1.3%	Morogoro	1.7%	Zanzibar	2.4%	Arusha	2.5%
Moshi	2.4%	Morogoro	1%	Moshi	1.3%	Mwanza	1.8%	Mwanza	1.4%
Other urb.	5.9%	Other urb.	0.7%	Other urb.	2.4%	Other urb.	3.9%	Other urb.	1.1%
Non-urban	26.4%	Non-urban	36.5%	Non-urban	37.8%	Non-urban	42.9%	Non-urban	36.4%

Note: This table shows the destinations of residents for the five most populated cities. Origin and destination city boundaries are defined using 3km-buffered GRUMP polygons. Visitors are defined as being seen at least once in a location over the year. "Non-urban" refers to locations outside boundaries of cities with 200,000 or more residents. "Other urb." refers to all cities that are not in the top 5 origin cities.

Chapter 2

Deriving Granular Temporary Migration Statistics from Mobile Phone Data

2.1 Introduction

A growing body of research has highlighted the importance of temporary migration within developing countries. These flows of internal movements have been found to be very common and to largely exceed permanent moves (Baker and Aina, 1995; Coffey et al., 2015; Delaunay et al., 2016). They have at first been portrayed as a sign of failure of rural livelihoods (Findley, 1994; Schareika, 1997) but have also been described more recently as a structural component within households' livelihood strategies. For instance, Coffey et al. (2015) identify temporary labor migration to cities in India as a long-term economic strategy whereby households enjoy short-term work opportunities during the agricultural off-season. Similarly, Delaunay et al. (2016) document the prominence of temporary movements and their role as a complementary source of income used to mitigate risks or cover education costs in rural Senegal. Additionally, Bryan, Chowdhury, et al. (2014) show that temporarily migrating to urban areas is an effective way for rural households to cope with seasonal famine in Bangladesh. Other targeted studies have also provided key insights into the interaction of temporary migration with informal insurance networks (Munshi and Rosenzweig, 2016; Meghir et al., 2019; Morten, 2019), and its consequences for origin and destination markets (Akram et al., 2017; Imbert and Papp, 2020b).

Despite its proven significance in economic decisions, temporary migration is seldom integrated in national statistical systems in a systematic way. Key studies cited above such as Bryan, Chowdhury, et al. (2014), Coffey et al. (2015), and Delaunay et al. (2016) or Imbert and Papp (2020b) rely on targeted surveys with

a limited geographic scope. Temporary migration patterns remain otherwise poorly documented at national scales – especially in sub-Saharan Africa. As a matter of fact, short-term movements are intrinsically difficult to measure (e.g. due to attrition and recall biases) and require specialized – and oftentimes costly – surveys (Lucas, 1997). 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. For example, the targeted survey conducted by Coffey et al. (2015) in rural North India reveals that half of temporary labor migration trips last for less than 30 days. They conclude that migration surveys such as India's National Sample Survey (NSS) that define temporary migration with high minimum duration thresholds – i.e. greater than one month – would significantly underestimate short-term movements.

In this respect, mobile phone data have emerged as a promising complement to traditional survey methods for measuring human mobility at larger scales, and with a spatio-temporal granularity allowing to capture subtler movements (González et al., 2008; Jurdak et al., 2015). As a result, a relatively recent body of research has exploited digital traces to infer various types of movements including commuting (Kreindler and Miyauchi, 2021; Miyauchi et al., 2021), short visits (see Chapter 1), internal migration (Blumenstock, 2012) and international migration (Ruktanonchai et al., 2018; Spyratos et al., 2019). The potential offered by these new mobility measures to deliver powerful ways of investigating a wide range of topics has been demonstrated in a number of studies. For instance, some have exploited mobile phone data to quantify the impact of human mobility flows on the transmission of infectious diseases (Wesolowski, Eagle, Tatem, et al., 2012; Wesolowski, Metcalf, et al., 2015; Wesolowski, Buckee, et al., 2016). Others have broadly focused on migration behaviors such as the link between short-term mobility and migration (Milusheva et al., 2017; Fiorio et al., 2017), the probability of urban migration in a developing context (Lavelle-hill et al., 2022), or the relationship between mobile phone activity and migration patterns (Hong et al., 2019). In some cases, Call Detail Records (CDR) data allow to describe connections between users and have therefore been employed to study the role of networks in both short-term mobility (Phithakkitnukoon et al., 2012) and migration decisions (Büchel et al., 2020; Blumenstock, Chi, et al., 2022). More recently, a number of studies have availed of the ever-increasing availability of GPS-accuracy location data from mobile phones to bring new perspectives on issues that are central for urban economists, such as agglomeration effects (Atkin, Chen, et al., 2022; Miyauchi et al., 2021), commuting and location choices (Kreindler and Miyauchi, 2021), connectedness of locations via short visits (Chapter 1), or segregation (Athey et al., 2020).

The production of migration statistics with mobile phone data poses several challenges. Firstly, mobile phone datasets do not result from a well-defined sampling strategy; the researcher does not have control over which individuals are observed and when they are observed. This raises critical questions about the representativeness of phone users of a broader population (Blumenstock and Eagle, 2010; Sinclair et al., 2023; Wesolowski, Eagle, Noor, et al., 2012, 2013) and the biases in mobility measures induced by irregularities in the frequency with which users are observed (Lu, Fang, et al., 2017). In this respect, previous studies have typically used subsets excluding infrequent phone users (Hankaew et al., 2019; Lai et al., 2019) but this potentially implies further selection given that phone usage is related to socio-economic characteristics (Blumenstock and Eagle, 2010), an issue that has been largely overlooked in the literature.

Secondly, migration events must be identified within user-level digital traces. This requires the definition of migration criteria and algorithmic rules. Other papers exploiting mobile phone data to measure migration have used ad hoc methods that are usually divided into two steps. User-level trajectories with heterogeneous sampling rates are regularized over the time dimension by defining a location for each user and time period, which is calculated with a frequency-based method as the modal location over that time period. Then, migration events are identified as a location change that persists over some number of consecutive time periods (Blumenstock, 2012; Hankaew et al., 2019; Lai et al., 2019; Zufiria et al., 2018). Those methods have three main limitations. First, the regularization of user's trajectories at a harmonized but coarser temporal resolution – e.g. by calculating monthly locations – necessarily causes some measurement error on the exact start and end dates of migration events and, therefore, 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.[1] Third, 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. This can be problematic, for instance, for the calculation of the number of individuals identified as being in migration for any given time period (i.e. the migration stock).

Thirdly, the unique granularity of mobile phone data offers the possibility to precisely describe fluctuations in temporary migration over time – e.g. seasonal

[1] For instance, an individual can be seen at his home location from the 1st, March to 16th, March of some year y, then temporarily migrate for 28 days from the 17th, March to 14th, April, and be back home from 15th, April 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.

movements. However, the production of time-disaggregated temporary migration measures poses a number of methodological issues. 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.

In this paper, I develop a thorough methodological framework that attempts to address these different challenges, with the objective of deriving granular temporary migration statistics from mobile phone metadata.

First, I propose systematic ways to characterize such (non-random) data samples in order to evaluate their degree of representativeness of a broader population. I make a clear distinction between two categories of complementary measures that allow to achieve this. Survey-based statistics inferred from secondary sources allow to compare the characteristics of mobile phone users and non-users. On the other hand, a set of sample-based metrics provide a characterization of the specific sample of mobile phone data at hand and allow to assess the magnitude of cross-sectional biases. Additionally, I analyze sample characteristics on the time dimension – i.e. the users' frequency and length of observation – to address two distinct issues; I evaluate whether these are fit for the purpose of measuring temporary migration and I investigate the existence of selection biases on the time dimension. In this respect, I propose a method to quantify the impact of observational parameters (i.e. the frequency and length of observation of phone users) on the performance of the proposed migration detection algorithm. Then, I implement an empirical test to identify potential patterns of non-random observational gaps, i.e. users being unobserved precisely when they are in migration. Finally, I discuss issues related to the selection of a subset of users with minimal observational requirements and I highlight the existing trade-off that must be made between migration measurement error on one side, and sample size and selection biases on the other.

Second, I build on Chi et al. (2020) to develop a temporary migration event detection algorithm that identifies migration spells within phone-based trajectories using a segment-based approach. Their method is based on the identification of contiguous sets of days at a primary location – called "location segments" – with a clustering technique, allowing for idiosyncratic deviations such as short observational gaps or short visits at a different location. Their approach is showed to out-perform all versions of traditional frequency-based methods. I also use a segment-based procedure for the detection of temporary migration events in CDR trajectories. However, an important addition relies on the estimation of a primary

residence location – which is defined in most cases as the most frequently observed location at night over a long time horizon – prior to detecting temporary migration events. This two-step procedure essentially allows to characterize the direction of migration flows by distinguishing between departures from and returns to a home location.[2]

Third, I develop a methodological approach that transforms user-level migration trajectories derived from mobile phone data into consistent migration statistics at various spatial and temporal scales. The paper specifically addresses the challenges posed by sampling irregularities for the disaggregation of migration measures on the time dimension. In this respect, I construct a set of rules allowing to optimize the amount of information provided in phone-derived trajectories for the construction of time-disaggregated flows, stocks and rates of temporary migration across locations.

Finally, I illustrate the methodology empirically with a large CDR dataset from Senegal. I present an exhaustive temporary migration profile of Senegal that exemplifies the detailed level of information that can be obtained on these types of movements from mobile phone data.

This paper primarily brings a methodological contribution to the literature on the use of digital traces to measure human mobility. Issues of selection on the cross-section are well-established in the literature (Blumenstock and Eagle, 2010; Lai et al., 2019; Wesolowski, Eagle, Noor, et al., 2012, 2013), and I discuss this aspect at length. However, the paper does not simply document cross-sectional biases and I propose a rectification method that partially account for those in the calculation of migration statistics. It is essentially based on a time-varying weighting scheme that neutralizes differences in the population-to-users ratio across spatial units. Also, and to the best of my knowledge, the quantitative assessment of observational requirements on the time dimension for the specific purpose of detecting temporary migration events is novel, as is the analysis of potential selection issues on the time dimension. Then, the detection of temporary migration events is largely inspired from the work of Chi et al. (2020) but an important contribution relies on the definition of a primary home location that allows to clearly identify the direction of the movements observed. Moreover, the method proposed for the construction of time-disaggregated migration statistics from user-level migration trajectories addresses issues that have been largely overlooked in the literature. Finally, I

[2] In Chi et al. (2020), migration events are defined as pairs of consecutive segments at distinct locations, so that a situation where a user moves from a location A to a location B and then from B to A results in the detection of two "migration events". A more precise characterization of the trajectory would be that we observe one migration period which materializes into two consecutive location changes. The procedure proposed in this paper essentially allows to determine that the user lives in A, migrates to B and then returns to his home location.

contribute to the broad strand of literature that aims to better understand migration phenomena in developing countries by providing a comprehensive temporary migration profile of Senegal.

The rest of the paper is structured as follows. Section 2.2 provides a brief description of the CDR dataset used throughout the paper to illustrate the proposed methods. Section 2.3 furnishes systematic methods, data and metrics to rigorously assess the degree of representativity of a typically non-random sample of digital traces. In section 2.4, I describe the algorithmic procedure used to detect temporary migration events in CDR trajectories and section 2.5 deals with the aggregation of user-level migration history into meaningful migration statistics disaggregated through time and space. In section 2.6, I derive a comprehensive temporary migration profile from CDR data in Senegal and section 2.7 concludes.

2.2 Data description

I use a dataset of anonymized CDR for Senegal that spans the period 2013-2015.[3] CDR are collected by telecommunication providers for billing purposes. Each record of a user i corresponds to an instance where a call[4] is made or received, and is associated with a set of attributes that typically includes: the (encrypted) phone number of the user, the (encrypted) phone number of the counterpart, the starting time of the call and an identifier of the antenna that processed i's call. [5]

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[6] and over 28.3 billion records, while the 2014-2015 dataset is comprised of 12,244,494 unique identifiers[7] for over 67 billion observations. With an estimated 5 million unique Sonatel customers

[3]To ensure privacy, records are in fact pseudonymized with phone numbers being replaced by unique identifiers.

[4]A "call" is used as a generic term to denote telecommunication transactions that can be of different types, typically a phone call or a text message.

[5]Other attributes are also available but not used in this particular study: the call type (call or text), the direction of the call (incoming or outgoing), the duration, whether the call is national or international, and the Type Allocation Code (TAC) of i's device, which is a unique 8-digit code that allows to identify a particular mobile phone model. Note the identifier of the antenna through which the call was routed for the counterpart is not provided in this dataset.

[6]To address concerns on the presence of bots and call centers in the sample, 102,313 identifiers that have over 100 records per day on average are removed – they account for a total of over 3.5 billion records.

[7]Similarly, this number excludes the 98,086 unique identifiers with over 100 records per day on average, which account for 5.4 billion records.

around that period[8], the large number of unique identifiers found in the sample cannot plausibly map into a unique set of individuals. This can be explained by at least three factors. First, some users may change their phone plan and purchase a new SIM card during the period of observation which would result in multiple identifiers in distinct periods of time being in fact associated with a single customer. Second, multiple SIM cards ownership is also common in Senegal and some individuals have several active phone plans.[9] In both cases, this may induce systematic biases in mobility measures to the extent that the timing of decisions to use a different SIM card may be non-random. More specifically, we may underestimate migration events if some users switch SIM cards precisely when they travel. I get back to this issue in more details in section 2.3.3. Third, the presence of foreigners using local SIM cards in Senegal – e.g. tourists or international migrants– and not captured by census population counts could virtually increase the number of users relative to the total population.

The georeferencing of antennas allows to reconstruct users' trajectories across space and over time. The spatial density of antennas combined with frequent phone usage[10] typically result in highly granular data that constitute an invaluable source of information to study human mobility. For the study period, the Sonatel network was comprised of 2,071 georeferenced Base Transceiver Stations (BTS).[11] The set of BTS point 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 pinging at the corresponding BTS. In other words, it is the approximate area covered by the station. BTS are not evenly distributed across the country and it is clear that the density of stations is higher in urban locations such as Dakar, Thiès or Touba (Figure 2.A.1). Since I do not focus on intra-city movements and to avoid systematic measurement errors[12], I group cells that belong to a single city in order to obtain a more balanced set of locations in terms of their size.[13] The resulting network of 916 cells showed in Figure 2.1 is a partition of the country extent and corresponds to the locations where users can be seen in the dataset.

[8]Source: author's calculations, see details in section 2.3.1

[9]In 2014, 44% of mobile phone users owned at least two SIM cards, and this fraction shows very little variation between Dakar, other urban areas and rural areas. Source: 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.

[10]In 2013, 76% of users report using their phone at least once a day.(Source: ibid 9)

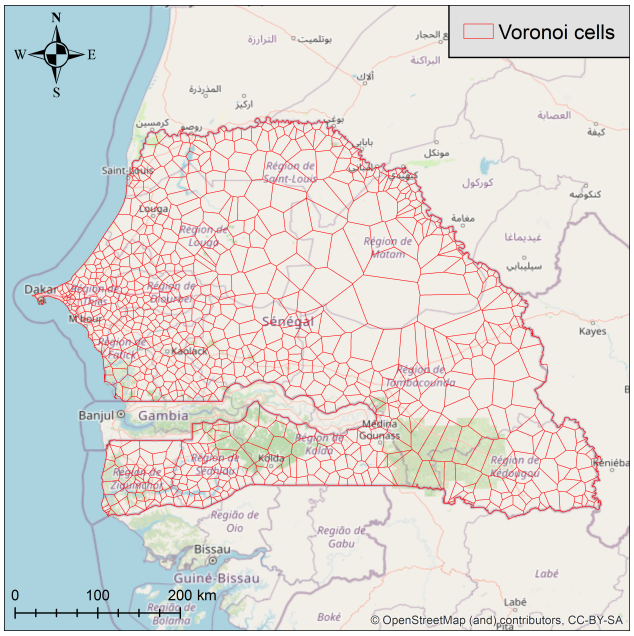
[11]A Base Transceiver Station can host multiple antennas at a single location.

[12]Since movements are essentially captured through consecutive calls at distinct stations, the likelihood of smaller movements remaining unobserved mechanically decreases with the density of BTS. In particular, short-distance moves in the more remote locations would be systematically underestimated compared to urban areas.

[13]Details are provided in appendix 2.A.

It is worth noting that those locations are not further aggregated, e.g. at the level of administrative units, as other papers have typically done (Blumenstock, 2012; Hankaew et al., 2019; Lai et al., 2019; Zufiria et al., 2018). Maintaining the high level of spatial granularity offered by CDR data allows to also capture short movements to nearby locations[14] and describe movements across a larger set of heterogeneous locations, e.g. between cities and rural areas with various levels of population density (see section 2.6). On the other hand, it increases the chances of erroneously identifying movements between adjacent cells in situations where users live near the border and connect to both BTS. Later in section 2.6, I make sure this phenomenon does not induce significant measurement errors by computing migration estimates that exclude movements between adjacent locations.

Figure 2.1: Voronoi cells defining locations.



2.3 Analyzing the representativeness of a mobile phone sample

The primary objective of the paper is to derive temporary migration statistics from digital traces and I therefore start by asking whether mobile phone operator billing logs provide sufficient information to achieve that goal.

[14] I later show in section 2.6 that, for instance, most of the rural-rural temporary migration movements are relatively short and mostly occur within regions.

First, unlike traditional survey data, mobile phone data do not rely on a well-defined sampling frame that would normally ensure that the sample is representative of a larger population, which usually allows to infer statistics on that population. In fact, mobile phone users have been found to statistically differ from the rest of the population along various dimensions such as wealth, age, gender, education, or the place of residence (Blumenstock and Eagle, 2010; Lai et al., 2019). To the extent that those likely correlate with migration choices, phone-derived migration estimates would necessarily differ from true migration flows within the overall population. A careful characterization of mobile phone data samples is therefore crucial in order to determine the sub-population they actually represent. Based on that, it is possible to simply derive statistics from the raw sample while being able to clearly define the sub-population those statistics refer to, or to implement sample rectification methods to account for identified patterns of selection and make the sample plausibly representative of a target population. In this section, systematic methods allowing to characterize users in mobile phone datasets are presented.

Second, users are only observed when making or receiving calls and I ask whether individual CDR trajectories carry enough information to allow for the identification of temporary migration events. Of course, observational requirements will depend on the type of mobility events one seeks to capture. For instance, measuring commuting requires multiple observations per day at different hours of the day while temporary migration is less demanding in terms of sampling frequency but asks for longer periods of observation. In any case, non-random observational gaps are a prime concern: do people use a different SIM card or change their phone usage patterns precisely when they migrate? Moreover, filtering procedures are often applied to select working samples of users with relatively more observations. This can create a selection issue since patterns of phone usage vary with individual characteristics (Blumenstock and Eagle, 2010).

Both aspects are discussed in the following sections and a systematic approach for characterizing a mobile phone data sample and precisely identifying selection issues and potential biases is provided.

2.3.1 Comparing the set of users with the population

First of all, it is useful to appreciate how the set of users at hand compares with the overall population. The unique set of users from any given telecommunication provider is essentially the subset of the population that both owns a phone and is a customer of that phone provider. This first macro-level characterization allows

to evaluate the size of the sample relative to the whole population and, most importantly, to identify specific segments of the population potentially missed in the data.

Estimates of mobile phone ownership rates at the individual- and household-level can be obtained from targeted surveys such as the ICT Access Surveys of Research ICT Africa and partners[15], or other surveys that include a question on mobile phone ownership. For instance, Demographic and Health Surveys (DHS) are available for almost every developing country worldwide and have a question on mobile phone ownership.[16]

Both an ICT Access survey (2017) and DHS data are available in the case of Senegal, as well as a national survey conducted in 2014 (*Enquête à l'écoute du Sénégal 2014*[17], henceforth referred to as the *2014 national survey*) that includes a module on mobile phone ownership and use. Based on the 2014 national survey, mobile phone ownership was estimated at 72% among the population aged over 18. Consistent with the increasing trend in phone penetration rates in the region for that period, I find comparable though slightly higher estimates for the year 2017 based on the ICT Access Survey and the 2017 continuous DHS, with rates evaluated at 78% and 76% respectively. Moreover, estimates disaggregated by age from the 2017 DHS reveal a steep increase in phone ownership within the 15-20 age category, from around 20% to 80% (Figure 2.B.1). Phone ownership can thus be reasonably assumed negligible for the population under 15, which is then almost entirely missed in the data although it accounts for 43% of the Senegalese population.[18] Migration measures derived from mobile phone data in Senegal are therefore primarily informative about movements of people aged 15 and above. The magnitude of the gap between phone-derived migration statistics and true values for the overall population will depend on migration behaviors within the missed population and the type of movements considered. For instance, children under 15 unlikely take part in seasonal migration movements whereas the prominence of fostering

[15]The 2017-2018 round covered nine African countries: Ghana, Kenya, Lesotho, Mozambique, Nigeria, Senegal, South Africa, Tanzania, and Uganda. Surveys were also conducted by RIA sister networks in six Asian countries (India, Indonesia, Pakistan, Bangladesh, Nepal and Cambodia) and five Latin American countries (Peru, Guatemala, Colombia, Argentina and Paraguay). A new round of the survey was conducted in eight African countries during 2022: Burkina Faso, Ethiopia, Kenya, Nigeria, Senegal, South Africa, Tanzania, and Uganda. More information is available at https://researchictafrica.net/.

[16]Both household data sets (since DHS phase IV) and individual data sets (since DHS phase VI) include a question on mobile phone ownership, allowing to estimate the fraction of households with at least one mobile phone and the proportion of individuals that own a phone, respectively.

[17]Source: 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

[18]Source: World Development Indicators, World Bank.

practices probably results in sizeable flows of mobility within this population.[19] I also calculate mobile phone ownership disaggregated along other dimensions such as sex, education, wealth or residence location. Some disparities in mobile phone ownership across various groups of individuals are found but no evidence of other identifiable subsets of the population being clearly missed in the data (see Figures 2.B.1-2.B.5).

In addition, the unique set of users from a single phone carrier only represents a subset of the universe of mobile phone users. In Senegal, and for the period of interest, over 88% of mobile phone users have a Sonatel SIM card and 77% report Sonatel as their main provider.[20] All in all, the unique set of Sonatel users represented around 36% of the population in 2013[21], i.e. over 5 million individuals.

2.3.2 Analyzing cross-sectional biases

Compared with traditional survey data sources, mobile phone datasets often provide information on a large fraction of the population, but their composition does not rely on a sampling protocol that would otherwise ensure the representativeness of a well-defined population. A first-order concern for the construction of phone-derived statistics is that phone users may differ from the general population along dimensions that possibly correlate with migration behaviors, thus inducing a selection issue.

In practice, phone logs from a particular provider seldom include personal information on users and we are usually limited in our ability to directly compare these particular users with the population as a whole. A second-best strategy therefore consists in comparing the population of mobile phone users with the at-large population (Blumenstock and Eagle, 2010) – assuming customers of a particular provider do not differ wildly from other phone users. To do so, secondary survey data sources that include information on mobile phone ownership – such as those mentioned in section 2.3.1 – constitute valuable sources of information. They allow to quantify differences along a variety of socio-economic dimensions between phone users and the population. I do this for Senegal using the 2017 DHS[22] men and women datasets. Results are showed in Table 2.1 where male

[19] In 2006-2007, it was estimated that 10% of children under 15 were being fostered and 14% of adults had been fostered in their childhood (Beck et al., 2015).

[20] Source: author's estimations based on micro-data from the 2014 national survey.

[21] This corresponds to the product of the fraction of the population aged 15 and above (57%) with the mobile phone ownership rate estimated in the 2014 national survey (72%) and the share of mobile phone users with a Sonatel SIM card (88%).

[22] 2017 is the closest year to the study period for which a DHS with a question on mobile phone ownership in individual questionnaires is available.

and female phone users are compared with the overall population of men and women respectively[23], for a few key characteristics: wealth, education, age, and zone of residence. Modest but statistically significant differences are observed between those two sub-populations, mainly within the women subset. Female phone users are found to be typically wealthier, more educated, older and more urban than the overall population of women. Male phone users, on the other hand, are slightly older and more urban than the general population of men, but are not statistically different in terms of wealth or education. Overall, the results are suggestive of a strong potential for mobile phone data to represent the (adult) male population – including the poorest – and a moderate tilt towards wealthier individuals among the female population. Although similar statistical comparisons cannot be made with respect to gender, I estimate that 54% of phone users are men whereas men only account for 48% of the population, so mobile phone data would tend to under-represent women.[24] Overall, the observed patterns of selection broadly corroborate previous findings of studies in Rwanda (Blumenstock and Eagle, 2010) and Kenya (Wesolowski, Eagle, Noor, et al., 2012) that documented differences between phone users and non-users with respect to population density, wealth and gender. Note that observed differences between phone users and the general population result from a combination of the relative size of the population of non-users and the magnitude of differences between users and non-users. As phone ownership rates are already high and keep increasing, the impact of the latter on representativeness becomes more and more minimal.

The barrier costs to owning a phone are a likely source of selection of wealthier individuals in the sub-population of phone owners. According to the ICT Access Survey, the high cost is the primary reason for not owning a phone and 65% of non-users declare not having a phone because they cannot afford one.[25] Mobile phone owners report spending FCFA 4,600 (approximately $9 USD) per month on voice, SMS and data, which represents 4% of the reported monthly income.

[23]Note individual-level datasets cover men and women of reproductive age, i.e. men aged between 15 and 59 and women aged between 15 and 49, so that in practice, I compare the characteristics of mobile phone users with the characteristics of the overall population in those age categories.

[24]As most DHS samples, the 2017 DHS sample in Senegal is primarily a stratified sample of households selected in two stages. All women of reproductive age (15-49) are administered an individual questionnaire and men aged between 15 and 59 are also interviewed for half of selected households. Normalized sampling weights from the women and men datasets cannot be combined in practice which prevents statistical inference on the overall population of men and women of reproductive age. That said, I estimate the share of phone users that are male by calculating ownership rates for men and women separately that I combine with the share of women in the general population.

[25]About 40% report the lack of access to electricity as a reason for not owning a phone and the third most frequently reported reason is that they do not know how to use a phone (36%).

Table 2.1: Comparison of phone users with the general population, 2017.

	Phone users (1)	**All** (2)	**Diff.** (1)-(2)
Women			
wealth: Richest dummy	0.285	0.237	0.048**
wealth: Richer dummy	0.236	0.210	0.026*
wealth: Middle dummy	0.192	0.191	0.001
wealth: Poorer dummy	0.163	0.184	-0.021*
wealth: Poorest dummy	0.123	0.177	-0.054***
Years of education	4.965	4.181	0.783***
Age	29.426	28.339	1.087***
Urban dummy	0.591	0.497	0.094***
Men			
wealth: Richest dummy	0.221	0.205	0.016
wealth: Richer dummy	0.208	0.199	0.009
wealth: Middle dummy	0.196	0.199	-0.003
wealth: Poorer dummy	0.205	0.210	-0.005
wealth: Poorest dummy	0.170	0.186	-0.017
Years of education	5.301	5.103	0.198
Age	32.228	30.497	1.73***
Urban dummy	0.561	0.531	0.03*

*p<0.1; **p<0.05; ***p<0.01

Notes: The first two columns provide mean values of the corresponding variables for the two groups. p-values are obtained with two-sample t-tests. Comparisons are conducted on women and men separately since DHS normalized weights from distinct datasets cannot be combined (ANSD/Sénégal and ICF, 2018). The "wealth" variable corresponds to the DHS zone-specific wealth index categorized into quintiles.

As mentioned above, the relevance of those comparisons relies on the assumption that phone users from the telecommunication company providing the data do not differ from the overall population of phone users. This is a minor concern in the particular case of Senegal given that the mobile telephony market was largely dominated by Sonatel, but it may matter in more fragmented markets. In any case, verifying that assumption remains quite difficult in practice as it requires survey data with information on both mobile phone ownership and the phone provider used. ICT Access Surveys do provide those information for the set of countries that they cover. I conduct a comparison exercise of Sonatel users' key characteristics with the overall population of users and find that Sonatel users are broadly comparable with the population of phone users[26] (Table 2.B.1).

Even though individual characteristics of users from a particular mobile phone

[26]The only remarkable difference is that Sonatel users are disproportionately more urban than other phone users (57.4% against 50.5%).

dataset cannot be directly compared with the general population, it is however possible to infer users' approximate home location and characterize them. Those characteristics can then be directly compared to those of the population at large. I propose three simple sample-specific metrics. First, I determine the zone of residence for each user in the sample (i.e. urban or rural) and define the "urban bias" as the ratio between the fraction of users living in urban areas with the level of urbanization in the general population. Second, I evaluate the distribution of users across space – e.g. across voronoi cells or administrative units – and calculate the degree of correlation with the population as a whole. This is particularly useful for appraising the spatial coverage of a sample and identify potential anomalies such as data sinks or holes. Third, a more granular version of the first metric can be obtained, similar to the metric proposed in chapter 1 (section 1.3). Locations can be ordered by population density and grouped into bins, where each group of locations accounts for a fixed fraction of the population. For instance, we can form ten bins going from the group of least dense locations to the group of densest locations, with each bin accounting for one tenth of the population. We can then assess the degree of selection with respect to population density in a mobile phone data sample by calculating the distribution of phone users' home location across those density categories. In the absence of selection, the fraction of users found in each bin should be comparable to the share of the population it hosts. To the extent population density correlates with a number of socio-economic variables such as poverty, access to basic services, or market access, the latter metric allows to indirectly appraise the degree of selection with respect to those indicators.

I estimate those three metrics on the 2014-2015 CDR dataset in Senegal.[27, 28] I find that 68.6% of users live in voronoi cells classified as urban whereas those only account for 49.3% of the population: the urban bias equal to 1.39 indicates that urban areas are clearly over-represented in the sample. Of course, this is consistent with higher phone ownership rates observed in urban areas and with the fact that Sonatel users are slightly more urban than other phone users.[29] Then, I count the number of users in each voronoi cell that I compare with the total population.[30]

[27] I consider a base sample where I impose minimal observational constraints that simply allow me to estimate a home location for each user. More specifically, I define a unique location for each night (6pm-8am) as the most frequently observed location during that night and the user's home location is then calculated as the most frequent location among all nights observed. Users in the base sample are those observed on at least 10 distinct nights and with at least half of the nights observed at the estimated home location.

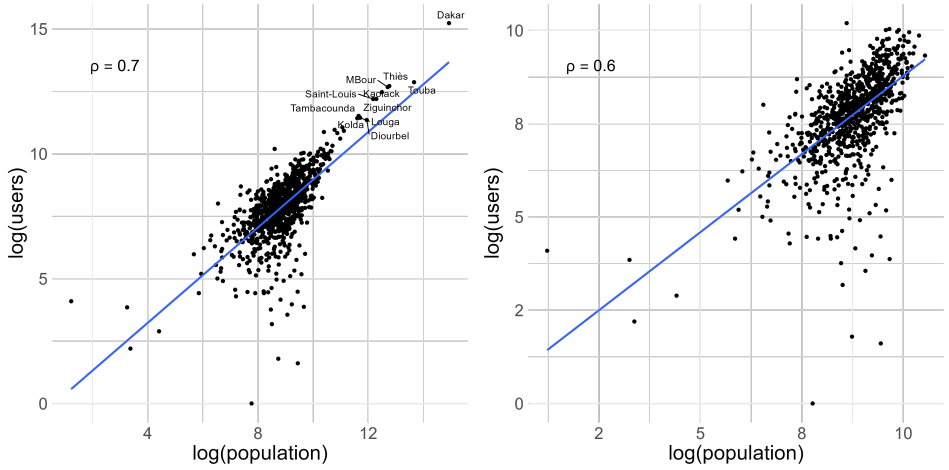
[28] Results of those estimations on the 2013 dataset are included in Appendix 2.B and show minimal differences.

[29] According to the 2014 national survey, 81% of urban residents own a mobile phone and this number drops to 62% in rural areas. Note that the urban-rural classification of the 2014 national survey does not necessarily match the definition of urban areas adopted here.

[30] Total population at the voronoi-level is obtained by overlaying voronoi polygons with the 2017

Results are showed in Figure 2.2(a) where a high degree of correlation (0.7) between the number of users and the population at the cell-level can be observed, although it may be driven by the most populated urban cells – in particular by Dakar.[31] Figure 2.2(b) shows the same result considering the subset of rural cells and the correlation is imperfect though remains relatively high (0.6). Finally, I estimate the third metric considering ten groups of voronoi cells ordered by population density and each accounting for 10% of the total population.[32] Results are showed in Figure 2.3(a) that confirms the existence of a selection pattern towards the densest locations. Interestingly, the tilt is far from dramatic and largely contained in the lower categories where fractions of users remain consistently over 7-8%. The pattern of selection is most likely driven by urban locations and I produce the same graph in Figure 2.3(b) considering the subset of rural locations to check whether that pattern also holds outside cities. The main conclusion is that it does not. The distribution is broadly balanced across the ten density categories so that the set of non-urban users equally represents individuals from locations of different densities, including the most remote areas.

Figure 2.2: Distribution of users across voronoi cells in the base sample.



(a) All cells (b) Rural cells

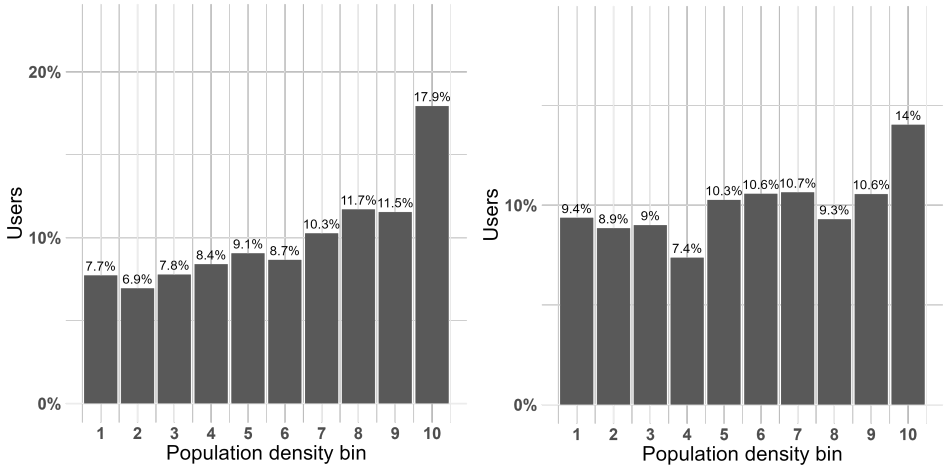
Note: The blue line represents a linear regression line between the (logged) number of users and the population at the voronoi-level.

100m-resolution gridded population product from the WorldPop Research Group (Qader et al., 2022).

[31] Figure 2.B.6 in Appendix 2.B shows the same result where the two most populated urban cells (i.e. Dakar and Touba) are excluded and the coefficient of correlation remains practically unchanged (0.68).

[32] I exclude Dakar as it accounts for 22% of the population and thus would itself cover the top two density bins. I still provide the graph that includes Dakar in Figure 2.B.7 in Appendix 2.B. In any case, the selection towards Dakar is already clear in the sample: 41% of the users have their estimated home location in Dakar whereas the corresponding cell only accounts for 22% of the population.

Figure 2.3: Distribution of users across population density categories in the base sample.



(a) All cells except Dakar

(b) Rural cells

2.3.3 Analyzing representativeness on the time dimension

Here, I examine how the characteristics of individual CDR trajectories on the time dimension can affect the ability to produce migration statistics. I discuss two separate aspects. First, I essentially ask whether users' frequency and length of observation allow to confidently capture temporary migration events. Then, I discuss a selection issue per se with the potential implications of non-random observational gaps.

For both the 2013 and 2014-2015 datasets, a large fraction of users are seen at least every other day on average and over a period of time spanning almost the entire year 2013 and the period 2014-2015 respectively. For instance, users in the 2013 (resp. 2014-2015) dataset have a median length of observation of 357 days (resp. 633 days) and a median fraction of days with at least one observation equal to 0.67 (resp. 0.59).[33] Although data coverage seems relatively high, it is important to consider the minimal sampling characteristics that are necessary to confidently identify temporary migration events. I propose to quantitatively assess the impact of both the length of observation and the fraction of days observed on the accuracy of the migration detection algorithm. This will allow me to clearly inform the choice of observational constraints imposed on the subset of users selected in the analysis.[34]

[33]Comprehensive summary statistics on sample characteristics for both datasets are provided in Table 2.B.3 and 2.B.4 in Appendix 2.B.

[34]Of course, the migration detection algorithm is also sensitive to its hyperparameters – e.g. the maximum gap between observations at destination within segments or the proportion of days at destination (see section 2.4). I do not conduct a sensitivity analysis with respect to those as this has been done in Chi et al. (2020) for an algorithm very similar to the one proposed in this paper.

To this end, I conduct a sensitivity analysis using a benchmark subset of users in the 2013 dataset who are seen over a period of at least 360 days and on at least 95% of days. The strict observational constraints allow me to consider the corresponding CDR trajectories as perfectly reflecting users' locations over time. I then simulate random sub-trajectories with lower length and frequency of observation, re-apply the migration detection model on those sub-trajectories and compare the outputs with those obtained on the full trajectories. All details and results are provided in appendix 2.C. As expected, longer periods of observation are associated with lower error rates in home location predictions while the level of accuracy in the detection of temporary migration events is primarily affected by the frequency of observation. Lengths of observation of at least 300 days allow to sustain rates of accurate home location predictions beyond 90%, even with low frequencies of observations. More importantly, the level of accuracy of the migration detection model starts to deteriorate quite sharply when the fraction of days observed falls below approximately 0.5. On the other hand, trajectories with fractions of days observed beyond 0.8 are associated with detection rates above 90%. In the 2013 dataset, over 2.1 million users are seen for a period of at least 300 days and on at least 80% of days, which suggests that a significant fraction of users meet minimal sampling requirements for the detection of temporary migration events.

Working with a subset of users satisfying strict observational constraints thus has the clear advantage of reducing measurement error. However, it obviously comes at a cost of a lower sample size and, more importantly, it may exacerbate selection bias on the cross-section since phone usage can vary with individual characteristics (Blumenstock and Eagle, 2010) that correlate with migration choices. I discuss those issues in section 2.3.4 where I define a "high-quality" subset of users from which temporary migration estimates are eventually derived. It is important to note that the results obtained from the sensitivity analysis conducted on this particular CDR dataset cannot necessarily be generalized to other contexts. Future applications of digital traces to research on human mobility should ideally apply the proposed method to estimate observational requirements in the specific study context.

The analysis above tends to confirm that a large number of mobile phone users satisfy sampling requirements for the production of accurate temporary migration measures. However, the existence for some users of extended periods of time without calls raises the issue of non-random attrition: people might use a different SIM card precisely when they travel away from home, e.g. to enjoy a

better coverage provided by a different telecommunication company at destination. In the ICT Access Survey, respondents with multiple SIM cards are asked about the reasons for switching SIM cards. Reception problems is one of the most frequently reported causes (40%), along with promotions and cheaper on-net calls. This potential selection on the time dimension can lead to systematic downward biases in temporary migration estimates. In the 2013 dataset, 17% of users observed for at least 300 days on at least 50% of days have an observational gap of 20 days or more. I propose an empirical test to evaluate whether observational gaps tend to coincide with migration events. I consider a random subset covering the full period 2013-2015 and comprised of users who are observed for a period of at least 360 days and on at least 50% of days. The low minimum frequency of observation allows for the occurrence of relatively long observational gaps that may plausibly coincide with migration events. For any time period t, a user can be at home or in migration, and may or may not be observed, so that the total number of users at time t can be written as:

$$N_t^{total} = N_t^{home} + N_t^{migrant} + \tilde{N}_t^{home} + \tilde{N}_t^{migrant} \tag{2.1}$$

Where N_t^{home} is the number of users observed and at home, $N_t^{migrant}$ the number of users observed and in migration, $\tilde{N}_t^{home}$ the number of users non-observed and at their home location at time t, and $\tilde{N}_t^{migrant}$ the number of users non-observed and in migration at time t.

Assuming that a fraction α of the total number of users in migration are systematically unobserved, I show that the total number of unobserved users at time t, $\tilde{N}_t$, is related to the observed number of migrants $N_t^{migrant}$:[35]

$$\tilde{N}_t = \beta_0 + \beta_1 N_t^{migrant} + \epsilon_t \tag{2.2}$$

where $\beta_1 = \frac{\alpha}{1-\alpha}$. I estimate equation (2.2) and full results are provided in appendix 2.D. I find a positive but not statistically significant relationship – in particular for the rural subset – suggesting that some observational gaps possibly coincide with migration events, although the extent of this phenomenon cannot be considered a major concern for the construction of migration statistics. It is important to note that the method relies on the assumption that migrants who switch SIM cards when they travel and those who do not both display comparable migration behaviors over time. Despite the limitations of the proposed method (see appendix 2.D), this is, to the best of my knowledge, the first attempt to evaluate the prominence of selection biases on the time dimension in mobile phone data for the production of migration measures. Nevertheless, the proposed method provides a valuable starting point for understanding the impact of selection biases on the time

[35] Demonstration in Appendix 2.D.

dimension in mobile phone data. Future research should aim to investigate the prominence of these biases more in depth, and, if need be, develop strategies for mitigating their effects.

2.3.4 Selecting a "high-quality" subset of users

Previous studies using mobile phone data to measure human mobility have typically considered subsets of users satisfying minimal observational constraints, e.g. a minimum number of days with some calls (Blumenstock, 2012; Hankaew et al., 2019; Lai et al., 2019).[36] Filtering out infrequently observed users and/or those observed for short periods of time has the obvious advantage of eliminating uncertainty on locations visited by users over time and reduces measurement error. On the other hand, higher observational constraints are associated with lower statistical power since they decrease sample size. More importantly, excluding users based on sampling characteristics may exacerbate selection biases on the cross-section since phone usage patterns can vary with individual characteristics (Blumenstock and Eagle, 2010) that potentially correlate with migration decisions. The choice of filtering parameters should therefore be the result of an informed trade-off between these costs and benefits.

Other papers have applied observational criteria aligned with their measurement objectives, but have disregarded the impact of those constraints on sample composition. Here, I propose to quantify both the benefits of higher observational constraints, i.e. lower measurement errors in migration estimates, as well as their associated costs, i.e. lower sample sizes and selection biases. The impact of observational constraints on migration measurement error is examined through the sensitivity analysis presented in section 2.3.3, which provides an empirical relationship between the two main filtering parameters, i.e. the length of observation and the fraction of days observed for a user, and the level of accuracy of the migration detection model (see Appendix 2.3.3).

To investigate the costs associated with observational constraints, I start by calculating the number of users left in the sample for different sets of filtering parameters. I first look at the impact on sample size of a joint constraint on both the length of observation and the fraction of days observed, varying between 30 and 360 days, and 0.05 and 1 respectively. Figure 2.4(a) represents the sample size as a function of these two parameters.[37] The length of observation has a limited impact

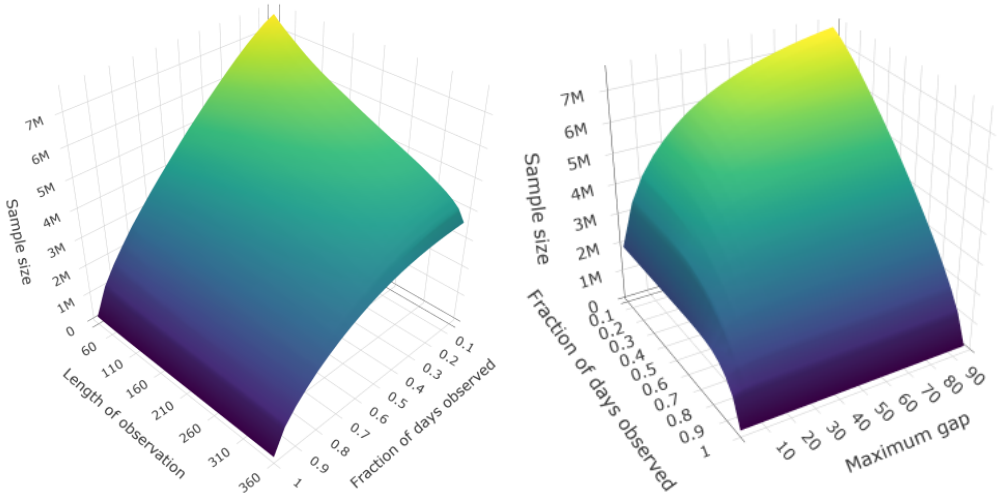
[36] It is also the method employed in chapter 1 using smartphone app location data (section 1.2).

[37] Two-dimensional versions of Figures 2.4(a) and 2.4(b) are provided in Appendix 2.B (Figure 2.B.10).

on sample size compared to the fraction of days observed. Setting high constraints on the fraction of days observed entails a large cost in terms of sample size: for a minimum length of observation set to 300 days, sample size decreases from 4.6 million to 1.4 million users when the minimum fraction of days imposed increases from 0.1 to 0.9. By contrast, setting the minimum fraction of days observed to 0.5 and increasing the minimum length of observation from 50 to 350 days results in a limited decrease in sample size, from 4.8 million to 3.4 million users. In addition, I introduce a third filtering parameter: the maximum observational gap tolerated for users in a subset.[38] Controlling for users' maximum duration of inactivity is a methodological approach that can be used to limit the presence of users with multiple SIM cards activated at different times during the period of observation. However, it is important to acknowledge that this approach does not address the non-random attrition problem – although it is still a useful approach for producing lower-bound estimates of temporary migration. I also look at the impact of this parameter on sample size. Figure 2.4(b) shows the sample size as a function of the minimum fraction of days observed and the maximum gap allowed. Both parameters appear to have a non-linear effect on sample size. In particular, the sample size sharply decreases for values of the maximum gap below 15-20 days.

[38] More specifically, for each user, I calculate the maximum time elapsed between two consecutive calls. I find a median maximum observational gap of 12 days in 2013 base sample and 25 days in 2014-2015.

Figure 2.4: Impact of filtering parameters on sample size, 2013.



(a) Obs. length and fraction of days (b) Maximum gap and fraction of days

Note: Panel (a) represents sample size as a function of minima length of observation and fraction of days observed imposed on users in the main sample, for a maximum gap parameter set to 100 days. Panel (b) represents sample size as a function of the minimum fraction of days and the maximum gap imposed on users in the main sample, for a minimal length of observation set to 30 days.

Finally, I evaluate the impact of the filtering procedure on selection biases in the cross-section by estimating the three sample-specific metrics proposed in section 2.3.2 on subsets with various levels of observational constraints. For the sake of conciseness, results on the sensitivity of sample composition biases[39] with respect to filtering parameters are left in appendix 2.B (Figures 2.B.11(a)-2.B.11(f)). The main conclusion is that the sample composition is primarily affected by the parameter setting the minimum frequency of observation. Higher fractions of days observed tend to exacerbate the pre-existing imbalances in the sample composition: they increase the bias toward Dakar, leave the bias for other cities practically unchanged, and markedly decrease the rural bias.

Similarly, I estimate the impact of filtering parameters on the degree of correlation between the number of users and the population across cells and show the results in Figures 2.B.12(a)-2.B.12(f). The correlation is primarily affected by the frequency of observation imposed, especially for the subset of rural cells. The non-linear shapes observed in Figures 2.B.12(e) and 2.B.12(f) indicate that it

[39]As explained in section 2.3.2, I define the "urban bias" as the ratio between the fraction of users found in urban cells with the level of urbanization in the general population. This metric is generalized to other subset of the population. For instance, I also evaluate the rural bias, which is the ratio of rural users over the rural population.

seriously deteriorates for values above approximately 0.8.

Finally, I evaluate how the minimum frequency of observation changes the distribution of users across population density categories. Figure 2.B.13(a) shows that higher fractions of days observed tend to exacerbate the selection with respect to population density by shifting the distribution of users toward denser categories. The pattern persists when considering the subset of rural locations (Figure 2.B.13(b)) but changes remain relatively modest. Filtering out users with less than 90% of days observed does not result in a dramatic loss of users for the most remote categories.

Overall, results on the costs of filtering parameters in terms of sample size and selection bias reveal that the frequency of observation is a critical factor. Higher lengths of observation have a comparatively low cost while they allow to improve the model accuracy by reducing errors in home location predictions, as seen in section 2.3.3. Weighing up these costs and the benefit for the model accuracy, I select a "high-quality" subset of users who are observed for a period covering at least 330 days on at least 80% of those days, allowing for observational gaps of at most 15 days. I also construct a "low-quality" subset of users with lower constraints: they are observed for a period of at least 250 days on at least 50% of days and have observational gaps of at most 25 days. The low-quality subset is supposedly less selected but may be associated with higher migration measurement errors. In section 2.6, I estimate the main migration measures on both subsets, which allows to evaluate the sensitivity of the results to filtering choices.

2.4 Migration event detection algorithm

By considering an individual's trajectory over distinct time horizons, three broad categories of human movements can be qualitatively defined. This is helpful to contextualize choices made in the construction of the proposed algorithm and clearly define the object of measurement. First, short-term mobility events such as daily commutes, short trips to cities or week-ends away are characterized by a short duration, typically a few days. 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. Third, permanent migration moves imply a long-term change in the usual place of residence. I refer to those different scales as the "micro", "meso", and "macro" scales, respectively, and to the time interval of the corresponding mobility events as micro-, meso-, and macro-segments. For any given individual observed over a long period of time, the sets of micro-, meso- and macro-segments form three

layers of mobility that necessarily overlap on the time dimension. For instance, an individual can have his usual place of residence in Dakar, for a period of three years that defines a macro-segment. Within this macro-segment, he temporarily migrates to Touba for a period of two months that defines a meso-segment. He spends two days in Saint-Louis while in Touba (i.e. a micro-segment within a meso-segment) and otherwise visits family in Thiès every other week-end (i.e. micro-segments within a macro-segment). Given the length of observation and the frequency with which phone users are observed, a raw CDR trajectory 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 macro-scale) is to develop algorithmic methods that smooth out noisy patterns created by movements at lower scales (e.g., at micro- and meso- scales).

I primarily focus on the detection of movements at a meso-scale, i.e. temporary migration movements. Empirical criteria on the duration of mobility events are required to clearly distinguish long micro-segments from short meso-segments, and long meso-segments from short macro-segments. In this respect, I define temporary migration events as meso-segments with a duration between 20 days and 180 days. The relatively low minimum duration threshold allows to capture short migration events, which are more common[40] and typically overlooked in survey data compared to longer-term migration spells. For instance, Coffey et al. (2015) conduct a specialized survey on rural-urban temporary migration in India and find that half of the events reported have a duration below 30 days. By contrast, the Indian National Sample Survey uses a minimum duration threshold of 30 days and thus misses a large fraction of short-term labor movements. On the other hand, the upper-bound duration is mainly constrained by sample characteristics. More specifically, it represents the longest temporary migration events that can be detected given the length of time that users are observed. As a simple heuristic, a migration event of a given duration can be detected in a CDR trajectory if the total length of observation is at least twice as long as the migration spell.[41] Since I consider users with a minimum length of observation of roughly a year, I thus define the upper-bound duration of temporary migration as 180 days (i.e. 6 months).

With these definitions in hand, I develop a mixed method that employs both frequency-based approaches and segment-based migration detection procedures

[40] Blumenstock (2012) finds a negative relationship between the minimum duration threshold and the rate of temporary migration estimated on a sample of CDR in Rwanda.

[41] Indeed, this is the limit over which it is possible to see 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 time at a distinct location as a temporary migration event.

inspired from Chi et al. (2020). Chi et al. (2020) focus on the detection of clusters of observations reflecting the continuous presence of a user at a particular location, within which they allow for idiosyncratic deviations corresponding to short-term trips away from the destination. Migration movements are then defined by the occurrence of two consecutive segments at distinct locations. I follow this approach for the detection of meso-segments in CDR trajectories. However, I add to their methodology by considering the trajectory of users at a macro-scale prior to detecting temporary migration events, which allows me to explicitly determine a primary residence location. This essentially allows to clearly characterize the direction of migration flows, from a home location to a destination (i.e. a departure) and from a destination back to a home location (i.e. a return). This is primarily useful for descriptive purposes. For instance, the total flow of movements can be decomposed into departing and returning flows, as I later show in section 2.6. Also, it allows to construct origin-destination migration matrices from which it is possible to identify net sending and net receiving areas.[42] More importantly, the identification of a home location in the measure of migration flows is crucial for migration models that incorporate home bias preferences in their setting to rationalize the patterns of movements observed (Imbert and Papp, 2020b; Monras, 2018).

For the sake of clarity and conciseness, I outline below the essential components of the methodology and I provide a more detailed description in Appendix 2.E.

First, hierarchical frequency-based methods are used to calculate the hourly, daily, and monthly locations[43] of users, as in Blumenstock, Chi, et al. (2022).

Then, I identify macro-segments with a segment detection procedure applied on monthly locations, which allows to define a primary home location for each user. Note that using monthly locations calculated with a frequency-based method offers a simple way to smooth out micro-segments. Consecutive months at a single location are grouped together, allowing for deviations reflecting the occurrence of temporary migration events. The groups that form periods of time longer than the specified upper-bound duration of meso-segments (i.e. 6 months) are identified as macro-segments. For the vast majority of users, this procedure yields only one

[42] By contrast, failing to account for the direction of observed flows in the description of migration patterns can be misleading. For instance, let's assume that ten users temporarily migrate from location A to location B while no user residing in B migrates to A. In aggregate terms, and ignoring the direction of flows, as many movements from A to B as from B to A are observed, and one would conclude that migration from A is comparable to migration from B.

[43] Details on the calculation of hourly, daily, and monthly locations are in Appendix 2.E. Note that daily – and therefore monthly – locations are primarily based on observations at night (6pm-8am) to avoid the influence of daytime movements such as commuting, and capture users' (temporary or permanent) places of residence. Filtering out diurnal activities is a standard practice in the literature (Vanhoof et al., 2018), although it has been showed to have little influence on measures of migration derived from mobile phone data (Blumenstock, 2012).

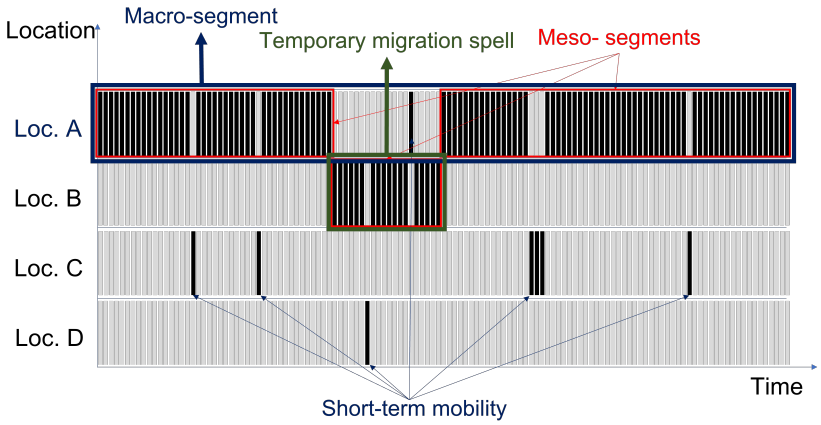
macro-segment that uniquely defines a user's home location. This corresponds to the dark thicker frame in Figure 2.5 which provides a schematic illustration of the overall migration detection procedure.[44] In this example, a unique macro-segment at location A defines the home location for the entire period of observation for this hypothetical user. Note that, when the period of observation of a user is longer than the maximum duration of meso-segments, multiple macro-segments at distinct locations may be detected. With the definitions adopted – and that can be flexibly adjusted – the user is considered as having moved permanently between the two locations, which implies a change in the primary home location.

Next, a similar segment detection algorithm is applied on daily locations. Meso-segments are identified as periods of time where a user is continuously seen at a single location, allowing for short-term idiosyncrasies (i.e., micro-segments). In the illustrative example of Figure 2.5, those correspond to the red frames. I define the observed duration of a meso-segment as the time elapsed between the start and end dates of the segment. Since some days before and after the segment may be unobserved, the actual duration may differ from the observed duration. Therefore, I also define the maximum duration of the meso-segment as the time elapsed between the observation immediately preceding the segment and the observation directly following the segment.

Finally, temporary migration events are identified by overlaying macro- and meso-segments: they correspond to meso-segments at non-home locations with a duration greater than a parameter τ^{temp} which, as aforementioned, I set to 20 days. As in Chi et al. (2020), I further impose a minimum proportion of days ϕ at destination in a meso-segment in order to limit cases where a segment would capture frequent movements between multiple locations rather than a temporary migration event at a single location.

[44] Note that the representation of a trajectory extracted from CDR data is inspired from Figure 1 in Chi et al. (2020).

Figure 2.5: Illustration of the migration detection procedure.



Note: Black bars represent the illustrative trajectory of a hypothetical user across four locations: A, B, C and D. The dark thick frame indicates the detection of a macro-segment covering the entire period of observation, which defines the user's home location (A). Red frames describe meso-segments detected with the clustering procedure on daily locations. The green frame designate the only meso-segment detected at a non-home location (B) and thus classified as a temporary migration segment.

One important limitation of the methodology is worth highlighting. As mentioned above, the maximum duration of temporary migration events that can be identified as such is effectively dictated by the length of observation of users. Consider a user usually residing in a location A who decides to temporarily migrate to a location B from February to November of some year y. If the user is only observed from January to December in year y, the algorithm will not be able to correctly identify the February-November period as a migration spell to location B. Instead, it will consider that location B is the user's home location over the period of observation. Of course, this remains consistent with the definition of temporary migration that I provide since the upper-bound duration is precisely adjusted for that purpose. But one will have to bear in mind that long-term events which could still qualify as temporary can only be identified with longer periods of observation.

2.5 From user-level migration history to migration statistics

2.5.1 Weighting scheme

In section 2.3.2, the analysis of the distribution of users relative to the population uncovers some degree of selection towards denser locations: urban areas tend to be over-represented compared to other locations. Moreover, the distribution of users across rural locations roughly aligns with the population distribution; however, this association is far from perfect. Such discrepancies pose significant challenges

in generating accurate aggregate statistics to the extent that migration behaviors also vary with population density. For instance, Dakar, while conspicuously over-represented in the sample, is also linked to a comparatively lower propensity to migrate, as we will see in section 2.6. Consequently, a non-adjusted sample-based estimation of the national-level migration rate would invariably lead to systematic underestimation of the actual rate.

To remedy this issue, I implement a simple weighting scheme that neutralizes differences in the users-to-population ratio across locations and allows to derive meaningful statistics at national and sub-national levels. The design of the weighting scheme must be consistent with the targeted granularity of the final migration statistics. For the migration profile presented in section 2.6, I calculate weights at the level of individual cities and for rural sub-stratas within each third-level administrative unit.[45] For each weighting unit, the weight corresponds to the ratio between the population and the number of users observed, i.e. it represents the number of individuals that any particular user represents in the population. Weights are therefore lower for relatively over-represented locations (e.g., cities) and higher for relatively under-represented areas. Moreover, I allow for weights to vary over time to account for the fact that, in any given location, the number of users actually observed differs across time units so that the composition of the sample might slightly vary over time. For any location ℓ and time period t, the value of the weight $w_{\ell t}$ is then:

$$w_{\ell t} = \frac{pop_{\ell}}{N_{\ell t}^{usersobs.}} \tag{2.3}$$

Where pop_{ℓ} is the total population represented in the data[46] in location ℓ and $N_{\ell t}^{usersobs.}$ is the total number of users residing in ℓ effectively observed during time period t.

It is worth noting that the rectification method operates as if users were randomly drawn from the population at the level of weighting units. This means that the weights simply correct for irregularities in the number of users selected relative to the population. It is important to acknowledge that other forces drive the selection mechanisms at play, as highlighted in section 2.3.2. For instance, the CDR sample in Senegal slightly over-represents men and relatively wealthier individuals within the women subset. A limitation of the rectification method is that it does not account for these biases. Note that basic information on users' characteristics (e.g. gender,

[45] Four categories of rural locations are defined at the country-level based on population density, as explained in appendix 2.A.

[46] For instance, the mobile phone dataset used in this paper is assumed to represent at best the population over 15 (see section 2.3.2).

wealth) would allow to improve the proposed method by introducing a socio-demographic component to the weighting scheme. Nevertheless, there is some evidence to suggest that local mobility patterns do not exhibit significant disparities between phone users and non-users. For example, Wesolowski, Eagle, Noor, et al. (2013) use a sample of CDR from Kenya in 2008-2009 and show that correcting CDR-based mobility estimates to account for phone ownership disparities among income groups results in only marginal differences when compared to the original, non-adjusted estimates at a local level. Moreover, the survey data utilized in section 2.3 do not allow to directly compare the level of mobility of phone users and non-users, but the observable socio-economic characteristics along which I compare them do not yield large differences. On the other hand, the notable disparities observed in the spatial distribution of users within the CDR sample, particularly between Dakar and other locations, are readily apparent and can be addressed with relative ease. In any case, many studies in the literature have mostly bypassed considerations of CDR sample composition when constructing mobility measures. Therefore, I contend that the introduced weighting scheme offers a notable enhancement for generating nearly representative migration statistics from a selectively biased set of digital traces.

To illustrate the significance of the rectification method, I compare migration rates drawn from both the weighted and unweighted versions of the 2013 high-quality subset. The raw data suggests that 24.4% of users embarked on at least one migration during 2013. In contrast, the weighted sample pinpoints a higher migration rate of 32.6%[47], that is an 8.2 p.p. difference. This clearly supports the notion that migration statistics derived from uncorrected mobile phone datasets may be prone to inaccuracies.

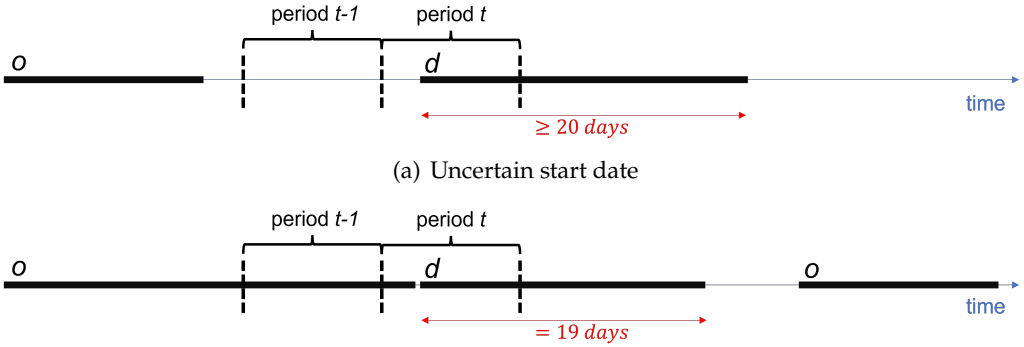
2.5.2 Regularizing unbalanced user-level trajectories

The migration detection model furnishes individual user location histories in the form of successive meso-segments. These meso-segments represent heterogeneous trajectories. To derive migration statistics, these diverse trajectories are aggregated 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, and the migration stock which corresponds to the number of users residing in o being in migration at destination d during t.

[47]These numbers consider migration events of at least 20 days. A user is considered a migrant if he has at least one migration event during the year 2013.

The calculation of migration flows may appear straightforward at first glance. 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, i returned from d to o at time t if a migration segment at d ended during t. However, the identification of migration departures and returns can be ambiguous because users are not necessarily observed every day and observational gaps imply some degree of uncertainty on the start date, end date, and ultimately the duration of meso-segments. Let us consider two illustrative examples. In Figure 2.6(a), user i residing in o has a migration segment at destination d which starts within period t. However, i is unobserved in period $t-1$, making it uncertain on which specific period the actual migration departure occurred. In Figure 2.6(b), the start date of the segment at destination d certainly 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 observation gap following the segment indicates that its actual duration may possibly be greater than 20 days, in which case user i should be considered as having departed for migration at time t.

Figure 2.6: Uncertainty in the identification of migration departures.



(a) Uncertain start date

(b) Uncertain segment duration

Similarly, the migration status of user i for a time period t – i.e. whether or not i is in migration at time t – can be ambiguous. Using the second example above (Figure 2.6(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 may have an actual duration greater than 20 days, in which case i should be considered as being in migration at time t. Also, defining a migration status for a time period t relies on a normative and somehow arbitrary choice: the extent of overlap required between a migration segment and the time period t to consider the corresponding user as being in migration during that specific time window.

I consider all possible configurations in the trajectory of users that give rise to ambiguous cases such as those presented in the examples of Figure 2.6. I implement an exhaustive set of algorithmic rules that resolve them. In Appendix 2.F, I provide details on each of those configurations along with illustrative diagrams similar to those of Figure 2.6.

Some of the key parameters used in the aggregation algorithm are worth highlighting. For the identification of migration departures and returns, I introduce a tolerance parameter ϵ^{tol}. Considering again the example in Figure 2.6(a), ϵ^{tol} represents the maximum time unobserved before the start of period t that we are willing to tolerate to still consider that user i departed for migration at d during period t. Additionally, I use a "certainty" parameter Σ to denote the minimum overlap required between a migration segment and a time period t to define the corresponding user as being in migration at time t. In the migration statistics disaggregated by half-month that I present in section 2.6, I set ϵ^{tol} and Σ equal to 7 days and 8 days respectively.[48] Finally, I define two sets of migration estimates associated with different levels of confidence with respect to the duration of meso-segments detected. High-confidence migration events correspond to migration segments with an observed duration greater than the minimum duration τ^{temp}, which is set to 20 days. On the other hand, low-confidence migration events also include segments with an observed duration lower than 20 days but a maximum duration greater than 20 days. Comparing high- and low-confidence estimates allows to appreciate the impact of observational gaps on the degree of uncertainty in time-disaggregated migration measures derived from CDR data.

Finally, estimating time-disaggregated migration rates requires some measures of the actual number of users observed at any given time period t, i.e. the denominator of such rates. As for the measure of migration flows and stock, the existence of observational gaps may imply some variations in the number of users actually observed over time. Broadly speaking, a user i is defined as observed at time t for a given migration measure (i.e. departures, returns, or stock) if the amount of information on the user's location during and around t allows to determine with certainty 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. Again, I consider all possible configurations in the trajectory of users and identify all cases where users are considered as unobserved for measures of migration

[48] By setting $\Sigma = 8\ days$, I simply impose that the overlap represents at least half the time unit since half-months have a duration of at most 16 days.

departures, returns and stock. In Appendix 2.G, I provide details on each of those cases along with illustrative diagrams that facilitate the understanding of algorithmic rules implemented. It is important to note 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^{temp}, the tolerance parameter ϵ^{tol} and the certainty parameter Σ.

2.6 Senegal temporary migration profile

In this section, a comprehensive set of phone-derived migration measures based on the methodology described in sections 2.4 and 2.5 are presented. Unless otherwise specified, results showed correspond to the high-confidence migration estimates derived from the high-quality weighted sample; they are therefore scaled to the population above 15.

First of all, the data reveal that temporary migration is very common in Senegal. I estimate that 4.3 million migration events of at least 20 days occurred over the year 2013. Of course, this number decreases with the duration threshold considered: 2.9 million events of at least 30 days, 1.2 million events of at least 60 days and 0.5 million events of at least 90 days. This trend is consistent with the findings of Blumenstock (2012). At the extensive margin, 33% of the adult population – approximately 2.6 million individuals – engage in one or more migrations of at least 20 days in 2013. Similarly, the migration rate decreases with the migration duration considered, from 26% for events of 30 days or more to 7% when considering events of at least 90 days.[49]

To illustrate the impact of the weighting scheme on migration estimates, I derive the same results from the raw (i.e. unweighted) sample and show the results in Table 2.H.1 in Appendix 2.H. I find substantial discrepancies in the estimated migration rates. In the raw dataset, 24.4% of users have a migration events of at least 20 days, which is 8.2 p.p. below the weighted estimate. Differences persist when considering other migration duration values; e.g. 10% of users have a migration events of at least 60 days whereas the weighted estimate is nearly 40% higher (13.9%). This is simply due to the fact that urban users are generally both over-represented in the sample and show a relatively lower propensity to migrate. Correcting the sample

[49]The construction of similar aggregate statistics for the period 2014-2015 is subject to some important limitations. In the 2014-2015 high-quality subset, not all users are observed over the entire two-year period, which creates a missing data bias; some users are seen over only one year and may not be classified as migrants although they migrated outside their period of observation. This necessarily induces a systematic downward bias in migration estimates. However, it is still possible to recover the total number of events over the period 2014-2015 by aggregating the number of migration departures estimated by half-month.

Table 2.2: Migration statistics at the national level, 2013.

	Migration events	**Migrants**	**Migration rate**
$\geq$ *20 days*	4,276,706	2,568,976	32.6%
$\geq$ *30 days*	2,874,507	2,037,406	25.8%
$\geq$ *60 days*	1,200,775	1,092,802	13.9%
$\geq$ *90 days*	528,388	520,205	6.6%

Note: Aggregate statistics for the year 2013 are based on a weighted sample where weights are equal to the ratio of the (adult) population over the total number of users observed in 2013 for each weighting unit.

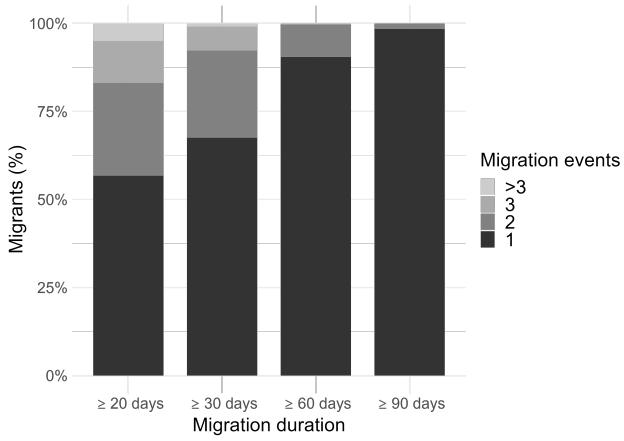
so it reflects more closely the actual composition of the population has a major impact on phone-derived migration estimates. Statistics derived from uncorrected samples can be largely misleading when the degree of selection is high; they should thus be treated with caution. Moreover, I compare low- and high-confidence migration estimations in Table 2.H.2 and find only marginal differences in both the number of migration events and the migration rates. However, I reproduce the results in Table 2.H.3 using the low-quality subset and intervals between low- and high-confidence estimates are in that case larger. This is consistent with the fact that relaxing observational constraints leads to higher uncertainty in migration estimates (see section 2.3.3).

Then, I describe the characteristics of temporary migration events at the intensive margin in Senegal, i.e. the the frequency with which individuals engage in temporary migration moves and their duration. First, the majority of those who migrate only migrate once per year. Figure 2.7 shows the distribution of migrants according to their observed number of migration episodes over the entire year 2013. Considering events of at least 20 days, 57% of temporary migrants engage in only one migration during the year. This number naturally increases with the minimum duration considered. For instance, considering events of 60 days and more only, I find that 90% of migrants are seen migrating only once. Second, the median duration of temporary migration episodes calculated on the full universe of events detected over the period 2013-2015 is estimated at 38 days.[50] The distribution of the duration of migration events is clearly skewed to the right; the average duration is estimated at 50 days for the period 2013-2015. A non-negligible fraction of migration episodes last for several months: 28% have a duration of at least 2

[50]More specifically, this is the weighted median of the observed duration of detected events based on the weighted sample as described in section 2.5.1. I also calculate the weighted median maximum duration and the result is practically unchanged (39 days). Similarly, I derive the same statistic from the 2013 and 2014-2015 datasets separately and numbers obtained are almost identical (i.e. 39 days and 38 days respectively), indicating that the final result is not driven by one period or the other.

months and 12% have a duration greater than 3 months. Overall, variations in the observed duration of temporary moves remain relatively modest; over 70% of migration spells last for less than 2 months.

Figure 2.7: Number of migration events conditional on being a migrant, 2013.

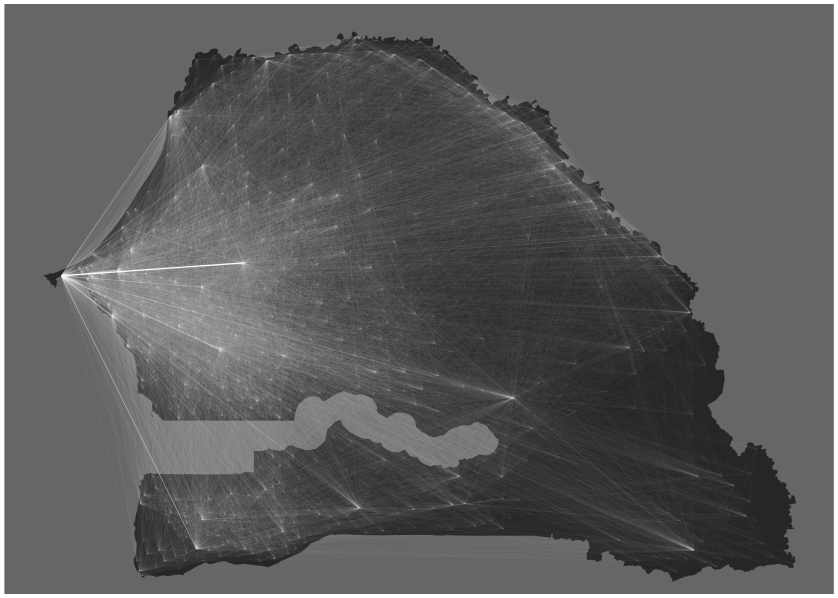


Next, I harness the fine-grained detail of the phone-derived migration estimates to characterize the spatial patterns of temporary migration movements in Senegal. The map of Figure 2.8 represents total migration flows for all pairs of locations over the year 2013. Overall, migration flows are largely widespread across the territory and the data reveal a remarkable level of mobility. Unsurprisingly, big cities such as Dakar, Touba, Ziguinchor, Thiès or Kaolack seem to be sending and/or receiving large numbers of migrants to and from all around the country. More specifically, important flows are observed between Dakar and those cities. However, a remarkable number of migration flows also involve rural locations.

Then, I aggregate the movements observed in the map of Figure 2.8 across relevant groups of locations. In Figure 2.9, I consider three sub-urban (Dakar, primary cities, secondary cities) and four sub-rural categories (very dense, dense, remote, very remote) and aggregate migration flows by pair of origin-destination category.[51] Unsurprisingly, urban locations appear as net receivers of temporary migrations while rural locations are net senders. In particular, Dakar alone attracts a relatively large fraction of temporary migration flows from all categories (25%). On the other hand, a clear majority of migration flows originate from rural areas as these account for 65% of the total flow. Interestingly, and perhaps at odds with the common narrative, a significant fraction of the rural-out flow is actually directed to other rural areas; rural-to-rural movements account for one third of the total flow. Importantly, I find that those movements are primarily local. The median

[51] Details on the construction of urban and rural categories are provided in appendix 2.A.

Figure 2.8: Total migration flows by origin-destination pair, 2013.



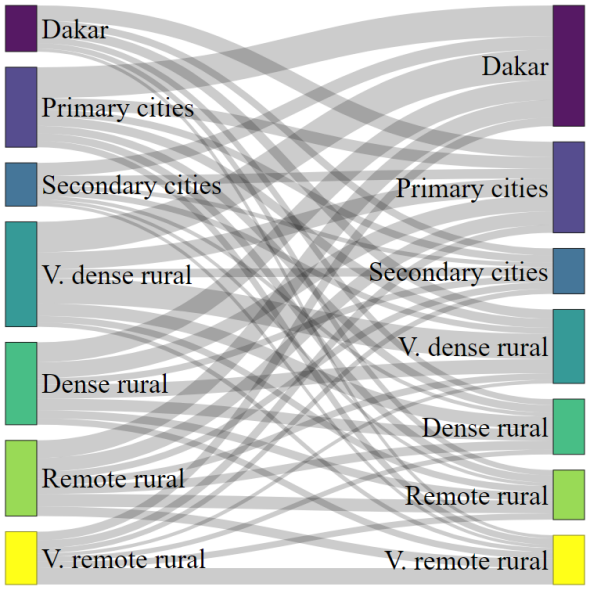
Note: Each white segment on the map represents the total number of migration events of at least 20 days between locations A and B at both ends of the segment, i.e. the number of migrations from A to B and from B and A. The brightness and width of the segments represents the magnitude of pairwise flows.

distance travelled by rural residents to other rural locations is 39km.[52] By contrast, the median distance travelled to migrate to urban locations is 214km, and is even larger in the case of Dakar (340km). Urban residents do not exhibit such differences and travel comparable median distances to migrate to rural and urban locations – 172km and 178 km respectively.

Figure 2.8 and 2.9 allow to describe the spatial distribution of temporary migration flows in absolute terms. I also investigate how the propensity to migrate varies across the categories considered in Figure 2.9. For each of these categories, I estimate the fraction of individuals with at least one migration event of 20 days or more to any destination, to a rural location and to a city. The results, shown in Figure 2.10, indicate that the propensity to migrate increases as we move to categories representing more remote locations. Only 15% of Dakar residents engage in a temporary migration of at least 20 days whereas this number consistently remains above 40% in all rural sub-categories. The observed trend seems driven by

[52]Distances between locations correspond to the distance travelled by car based on the Open Source Routing Machine (OSRM) tool(http://project-osrm.org/) that uses OpenStreetMap data. First, I calculate for each rural migrant the median distance travelled across all migration events with a rural destination. Then, I calculate the (weighted) median of this user-level metric across all rural users with at least one migration event to another rural location. Therefore, the estimated median is in fact a "median of a median". This measure is overall less sensitive to extreme values than a simple average.

Figure 2.9: Migration flows between urban and rural areas, 2013.

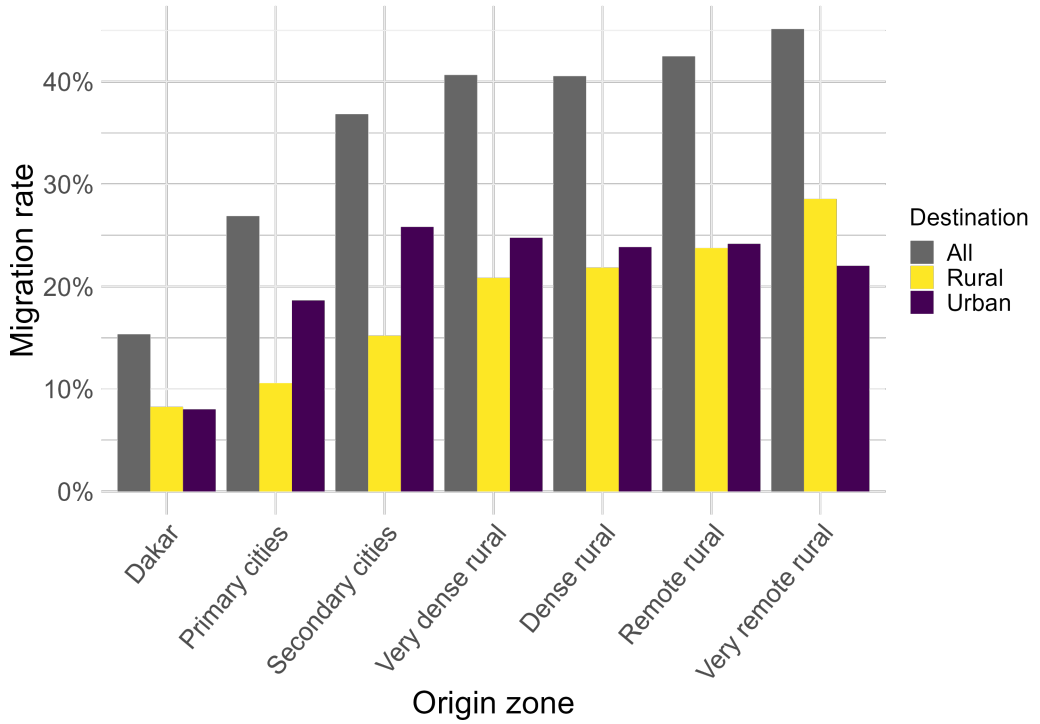


Note: Migration departures by origin-destination zones are aggregated over the period February-November 2013, considering migration events of 20 days or more. Those counts are based on the weighted high-quality subset. Origin zones are on the left-hand side of the graph and destination zones are on the right. Then, for each origin-destination pair, the grey bar represents the total number of temporary migration departures from origin to destination. The definition of rural and urban sub-classes is given in Appendix 2.A.

migrations to rural areas. The propensity to migrate to a rural destination gradually increases from 8% in Dakar to 29% for the category comprised of the most remote rural locations. On the other hand, the migration rate to urban locations peaks at 26% for secondary cities and then slightly decreases as we move along rural sub-categories, although it remains high at around 20-25%.

As illustrated above, phone-based migration estimates allow to precisely characterize the spatial distribution of temporary migration flows across origin and destination locations that can be flexibly defined. Mobile phone data can also uncover the temporal distribution of short-term moves at a level of granularity that could hardly be achieved with traditional survey instruments. For instance, Figure 2.11 shows the evolution of the stock of migrants (black line) over time for the period 2013-2015, as well as the underlying departing (green line) and returning (red line) flows. First of all, seasonal patterns clearly emerge. For all three years, a very steep increase in the stock of migrants is consistently observed, starting in June until reaching a peak in August-September. The magnitude of this increase is striking; e.g. the number of migrants more than doubles between the lowest (first half of June) and highest (second half of September) points in 2013. In absolute terms, this roughly represents an additional 470,000 migrants.

Figure 2.10: Migration rate by origin zone, 2013.



Note: The graph represents the rate of temporary migration by sub-zone of origin and by destination zone. Origin zones are represented as categories on the x-axis while bars are associated with colors each reflecting a particular destination category (all destinations in grey, rural in yellow, urban in blue). For each origin category, each bar gives the fraction of users with at least one temporary migration event to the corresponding destination category over the period February-November 2013. Estimates are based on the weighted high-quality subset.

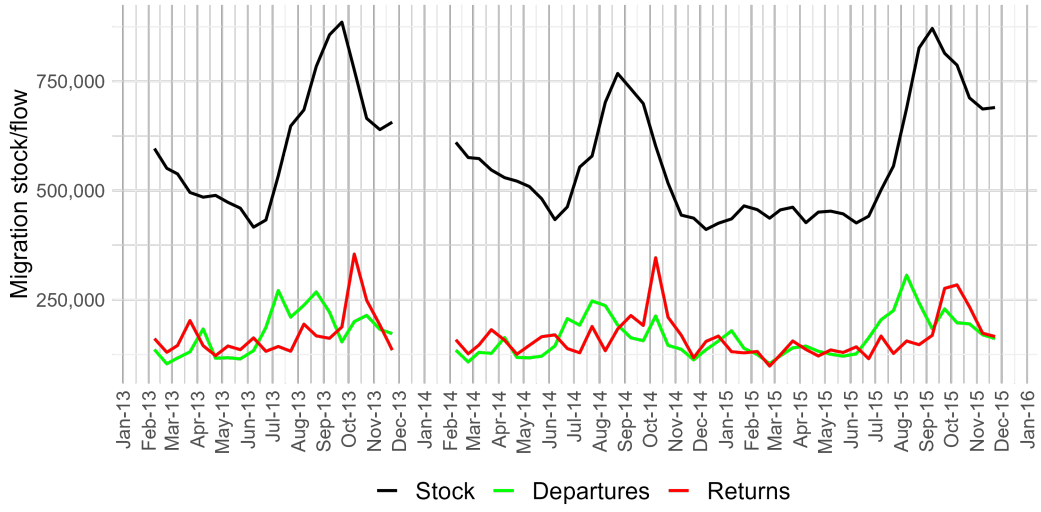
Interestingly, the seasonal trend observed goes against the common perception that short-term movements predominantly occur during the off-season (January to June). However, it is consistent with the study of Delaunay et al. (2016) who highlight that significant increases in school attendance rates have mechanically induced a concentration of the temporary migration moves of the youngest around the main holiday period (June-September). Boys exempted from agricultural tasks typically leave for a few weeks to find an informal job or work in a factory, whereas girls usually work as domestic employees. Extra revenues earned allow to cover school expenses for the migrant but also those of his or her siblings. Note that large returning flows around the end of September exactly coincide with the back-to-school period.

To investigate further the composition of the stock of migrants overtime, I disaggregate it by both origin and destination zone and present the results in Figure

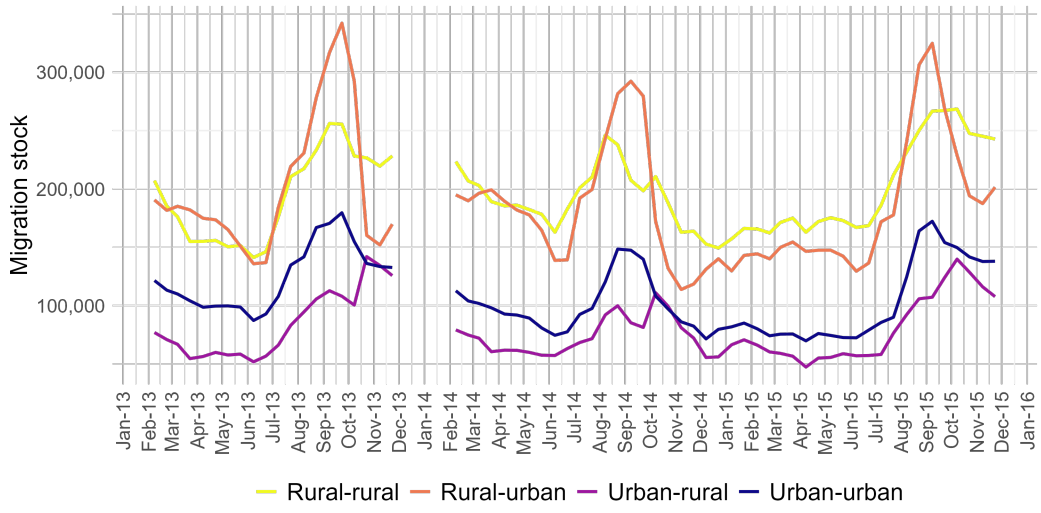
2.11(b). Consistent with findings in Figure 2.9, migrants originating from rural areas dominate the stock of migrants over the study period (yellow and orange lines). Interestingly, the increasing pattern from June to September observed at the national-level holds across all sub-categories of the total stock, but seems largely driven by the rural-to-urban component. The latter is systematically followed by a sharp decline in October. By contrast, the rural-to-rural stock of migrants also increases from the start of the rainy season but is usually sustained at a relatively high level in October-November.

Even though the motives behind the observed movements cannot be unequivocally identified, the patterns are suggestive of some interpretations. Broadly speaking, the systematic temporary reallocation of labour from the rural sector to the urban sector during the rainy season (June-October) points to the existence of income diversification strategies. Excess labor in the agricultural sector may push some households to send temporary migrants to urban areas where their marginal productivity is higher. Note that a further disaggregation of the rural-out stock of migrants by zone of destination reveals that the rural-to-urban flow during the rainy season is primarily directed toward Dakar, and to some extent to other primary cities (see Figure 2.H.1). Moreover, the sustained levels of migration to rural areas in October-November support the idea that people – including from urban locations (magenta line in Figure 2.11(b)) – temporarily move to agricultural areas to enjoy harvest job opportunities at that time of year.

Figure 2.11: Temporary migration patterns over the period 2013-2015.



(a) Stock, departures and returns at the national-level



(b) Stock by origin and destination zone

Note: All migration measures consider events of 20 days or more and are calculated on the weighted high-quality subset. Panel (a) represents the total stock of temporary migrants at the national-level (black line) by half-month over the period 2013-2015. The underlying flows of departures (green line) and returns (red line) from which the stock dynamic results are also showed. In panel (b), the migration stock is decomposed by origin-destination zones. The definition of rural and urban zones corresponds to the definition given in Appendix 2.A.

2.7 Conclusion

In this study, I provide a methodological framework for deriving temporary migration statistics from digital traces data such as Call Detail Records. Cross-sectional biases in mobile phone data are a well-known issue and I start with a careful analysis of the degree of representativity. I demonstrate that useful statistical comparisons of mobile phone owners with the population at large can be derived from nationally representative surveys containing a question on mobile phone ownership. Demographic and Health Surveys are a prime example of a secondary data source that can be tapped into for this exercise, and the proposed analysis could easily be replicated in other contexts given the DHS program covers almost the entire developing world. Conclusions in the context of Senegal are in line with previous research (e.g. Blumenstock and Eagle (2010)): mobile phone users are more predominantly male, more urban, and tend to be wealthier – although this is exclusively true for female users. I also construct a series of metrics allowing to evaluate cross-sectional biases in a specific mobile phone dataset. Such metrics primarily rely on the estimation of a home location and compare the spatial distribution of users with the population as a whole.

The paper also addresses the question of minimum sampling requirements at the user-level for the detection of temporary migration events. A sensitivity analysis suggests that shorter lengths of observation are associated with a lower accuracy in home location prediction, while the level of accuracy in the detection of migration events is primarily affected by the frequency of observation. Period of observation greater than 300 days with over 80% of days observed are associated with a 90% accuracy in home location predictions and a 90% rate of detection of migration events. Although the external validity of these results cannot be assumed, the proposed methodology could easily be applied on other mobile phone data samples and for other types of movements (e.g. short visits or commuting) to generate context-specific sampling requirements.

Moreover, the possibility exists that individuals use a different SIM card when migrating to a different location and the paper investigates this potential issue of non-random attrition. Selection biases on the time dimension at the user-level has been largely overlooked in the literature, although they may lead to systematic downward biases in phone-derived mobility measures. A simple statistical test applied to the three-year CDR dataset in Senegal does not support the existence of non-random attrition. However, the proposed method relies on strong assumptions about aggregate phone usage behaviors and future work could focus on investigating this issue more closely.

The migration detection algorithm presented in this paper is clearly inspired from the approach developed in Chi et al. (2020), where it is proved to outperform

previous ad hoc methods. An important addition is the estimation of a primary residence location prior to detecting location changes classified as migratory movements. This allows to clearly identify the direction of these flows that may either be departures from or returns to the primary residence. A central contribution of the paper is then to aggregate the user-level migration trajectories obtained from this procedure into regularized migration statistics disaggregated across space and time. I show how sampling irregularities imply some degree of uncertainty around the start date, end date and duration of detected migration events and complicate this aggregation exercise. The proposed algorithm treats the dozens of configurations that can be encountered when attempting to classify an observed segment start date as a migration departure during a given time unit, an end date as a migration return, or when trying to determine the migration status of a user for a given time period.

The weighting scheme forms another crucial component in the construction of meaningful aggregate statistics. The representativeness analysis highlights selection biases that make the sample composition clearly distinct from the population at large. As is typical with CDR data, denser areas are found to be systematically over-represented compared with more remote locations. The proposed weighting scheme therefore allows to neutralize differences in the population-to-users ratio across locations. In addition, its time-varying feature accounts for the fact that the actual number of users observed at each location may change over time; some users exit the sample while others are simply temporarily unobserved. I find large differences between weighted and non-weighted migration estimates, which I explain by the fact that over-represented groups are usually associated with a lower propensity to migrate. Since weights allow to mimic a sample composition that is comparable to the at-large population, weighted estimates plausibly depict a more realistic picture of actual temporary migration patterns. Nonetheless, a validation exercise with ground truth survey data would be useful to confirm this fact. Additionally, an important limitation of this weighting scheme is that it does not correct for biases along socio-economic dimensions such as wealth, age, or gender.

The application of the proposed methodology to three years of mobile phone data brings new insights into temporary migration patterns in Senegal. Mobile phone data unveil the ubiquity of temporary migration movements in the country. In 2013, one third of the population engaged in a migration episode of at least 20 days. A large majority of this flow originates from rural areas while urban areas – in particular Dakar – appear as net receivers of temporary migrants. Perhaps surprisingly, the phone-derived migration measures shed light on relatively important and short-distance rural-to-rural flows. In relative terms, the propensity

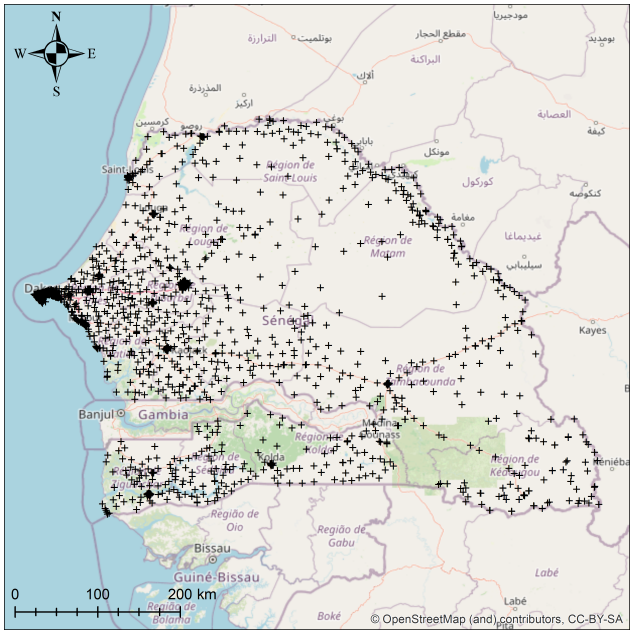
to engage in temporary migration increases as we move to more remote locales. For the year 2013, rural categories showcase migration rates consistently above 40% while only 15% of Dakar residents engaged in a temporary migration. Finally, time-disaggregated estimates reveal marked seasonal trends. While common narratives usually highlight the importance of off-season (January-June) movements, phone-derived migration measures rather suggest that the bulk of temporary moves occur during the rainy season (June-October).

The potential of mobile phone data as a complement to traditional surveys for the measure of subtler human movements is now well established. Applications in the fields of epidemiology, disaster risk management, or urban development have paved the road to an integrated use of such big data sources to support decision-making. However, a number of technical challenges remain and the lack of systematic approaches to construct robust mobility indicators constitutes an important barrier to a wider use of mobile phone data for policy making. This paper gathers a set of methodological tools that will hopefully support future research efforts involving the derivation of meaningful migration statistics from digital traces.

Appendix 2.A Voronoi tesselation

A Base Transceiver Station (BTS) describes an equipment that is responsible for the reception and transmission of radio signals from mobile phones. Each CDR reports the BTS that processed the corresponding user's call – in the form of a unique BTS identifier –, which is almost always the closest BTS in the network. This means that the exact location of the user is contained in the polygon formed by the set of points that are closer to that BTS than to any other station. The geometric transformation that converts a network of point coordinates to a set of contiguous such polygons is called a Voronoi tesselation. We apply a tesselation on the SONATEL network of BTS in order to define two-dimensional locations visited by users that we can then characterize in terms of population, weather conditions and so on. For the period 2013-2015, the SONATEL network had 2,071 BTS covering the entire country of Senegal (Figure 2.A.1). In what follows, we describe the three-step tesselation procedure that we apply to obtain the final set of Voronoi cells used throughout the paper.

Figure 2.A.1: Base transceiver stations.

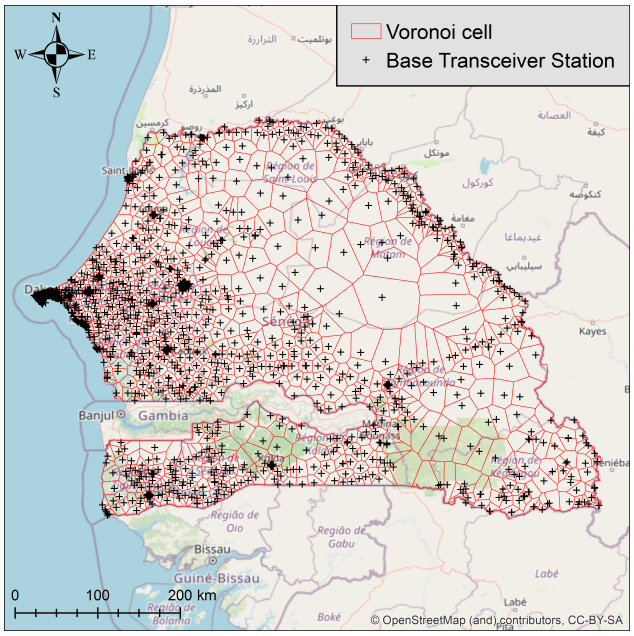


Step 1: Simple Voronoi tesselation

We first proceed with a Voronoi tesselation on the raw network of BTS. Results are showed in Figure 2.A.2. Telecommunications providers make network infrastructure deployment choices consistent with expected demand for mobile phone services. Of course, this results in an uneven distribution of BTS that is broadly in line with the population distribution. For instance, the density of BTS

is markedly higher in the main cities compared to the rest of the country. The variation in Voronoi cell sizes is consequently high; the 10% smallest cells are less than 0.23km^2 while the 10% largest cells are more than 291km^2. This means that measurement errors in users' location decrease with the density of stations, as well as the likelihood of small movements remaining unobserved. To achieve a better balance in cell sizes and because we do not focus on intra-city mobility, we implement a procedure that allows to group (small) Voronoi cells within what we identify as urban locations.

Figure 2.A.2: Simple Voronoi tesselation.



Step 2: Group cells within cities

City polygons are defined based on the GHS Settlement Model 2015 product (GHS-SMOD) from the Joint Research Center (JRC). The GHS-SMOD layers classify 1km^2 grid cells into settlement typologies ranging from rural to urban centre via a logic of cell clusters population size, population and built-up area densities.[53] The GHS-SMOD data package also comes with a vector global dataset of urban centre boundaries that we use to delineate 33 Senegalese cities.

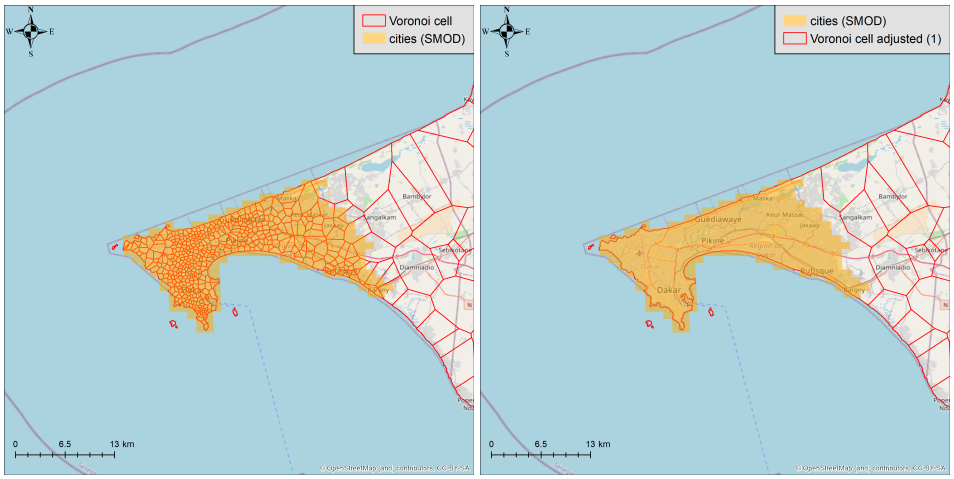
Voronoi cells intersecting a city polygon are assigned the corresponding city identifier. All cells assigned to a given city are grouped together to form a set of 33 city cells.[54]. We illustrate this procedure in Figure 2.A.3 for the case of Dakar. As a result of step 2, the number cells decreases from 1,666 to 919. The corresponding

[53]For more details, see the corresponding JRC webpage.

[54]The cell corresponding to the Gorée Island is manually aggregated to Dakar polygon, and another cell clearly overlapping Touba urban extent is assigned to that city.

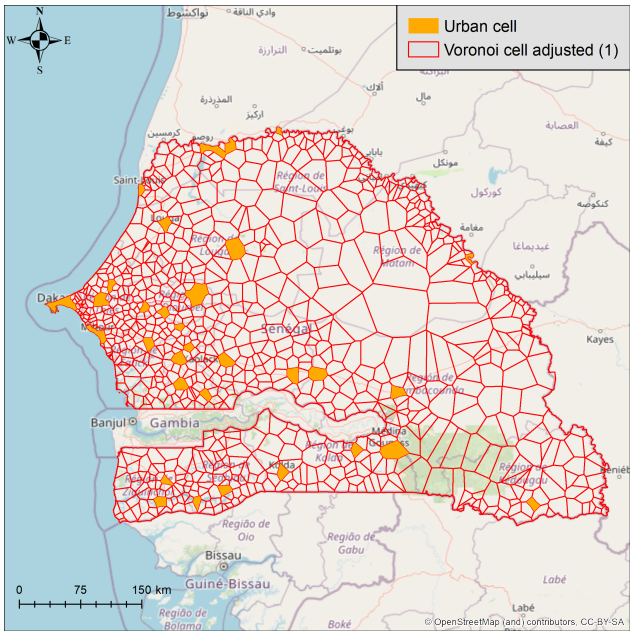
adjusted Voronoi diagram for the entire country is provided in Figure 2.A.4 where urban cells are showed in orange.

Figure 2.A.3: Grouping of urban cells, Dakar.



(a) Initial Voronoi diagram (b) Adjusted Voronoi diagram (1)

Figure 2.A.4: Adjusted Voronoi diagram (1), Senegal.

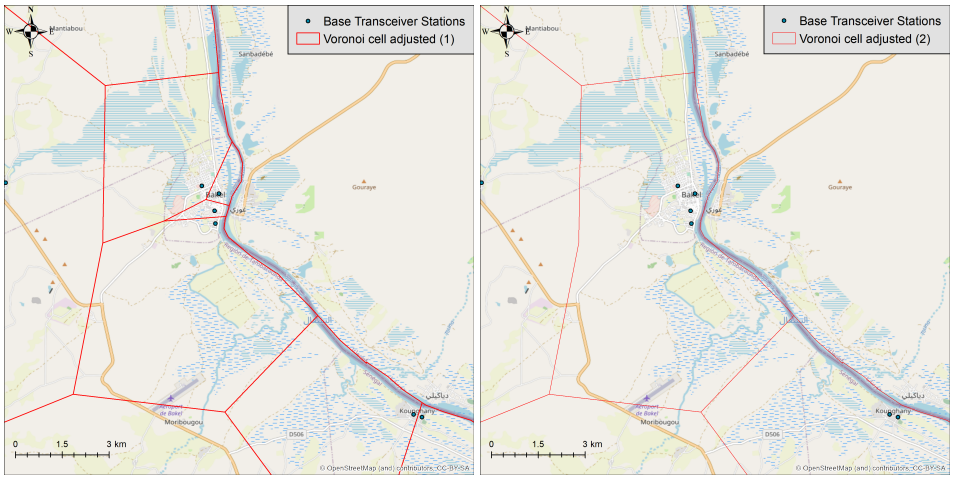


Step 3: Group cells within secondary urban areas

Some secondary urban areas are not captured by the GHS-SMOD product and clusters of small cells still remain after step 2. We thus detect clusters of BTS that are less than 2km apart and merge the corresponding Voronoi cells. We illustrate this

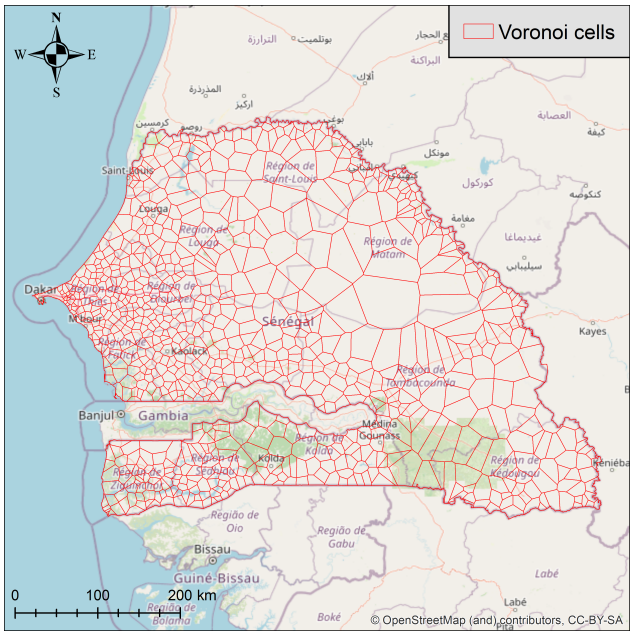
in Figure 2.A.5 for the city of Bakel. The final adjusted Voronoi diagram showed in Figure 2.A.6 has a total 916 cells.

Figure 2.A.5: Grouping cells within secondary urban areas, Bakel.



(a) Adjusted Voronoi diagram (1) (b) Adjusted Voronoi diagram (2)

Figure 2.A.6: Final voronoi diagram, Senegal.



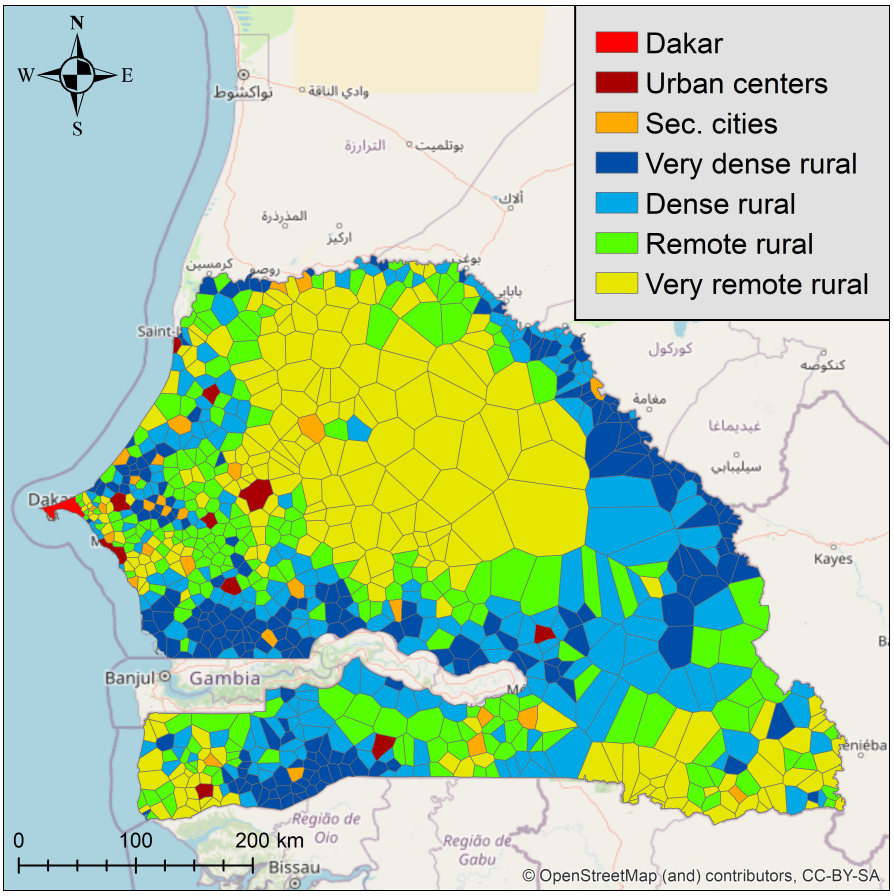
Finally, we classify those 916 cells into urban and rural locations with population-based criteria. The population of each cell is determined by overlaying the final voronoi diagram with a 100m-resolution gridded population product from the WorldPop Research Group (Qader et al., 2022). We define 6 categories:

- "urban centers" correspond to cells with over 100,000 inhabitants.

- "Secondary cities" have a population between 25,000 and 100,000 and a population density of at least 300 inh./km^2
- Other cells are categorized as "rural" and we further divide those into 4 groups of equal size based on population density. We label those groups as "Very dense rural", "Dense rural", "Remote rural" and "Very remote rural" respectively.

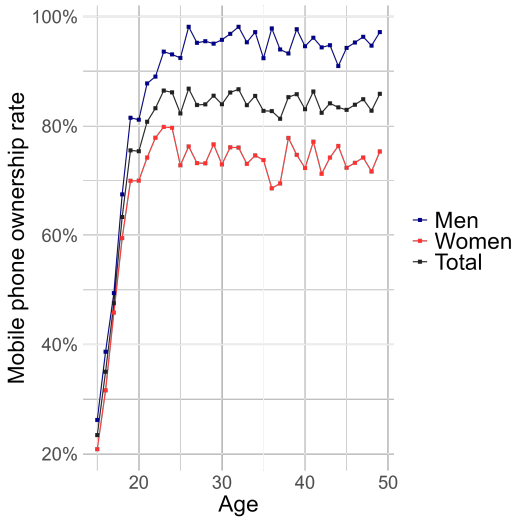
Urban centres (11 cells) and secondary cities (28 cells) together form the sub-group of urban cells while the remaining 877 cells form the rural sub-group. A map representation of this classification is provided in the map of Figure 2.A.7.

Figure 2.A.7: Urban-rural classification of voronoi cells.



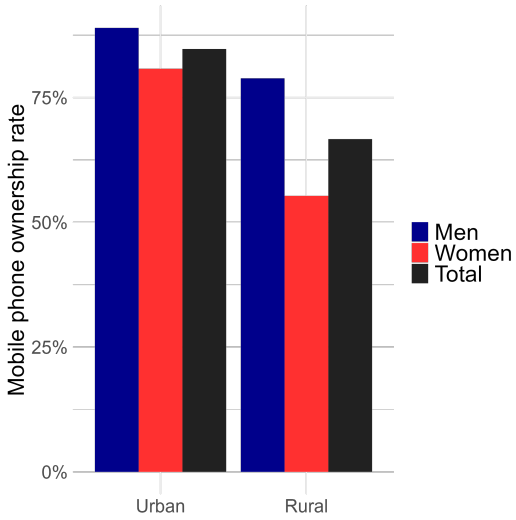
Appendix 2.B Sample representativeness: additional material

Figure 2.B.1: Mobile phone ownership by age and gender.



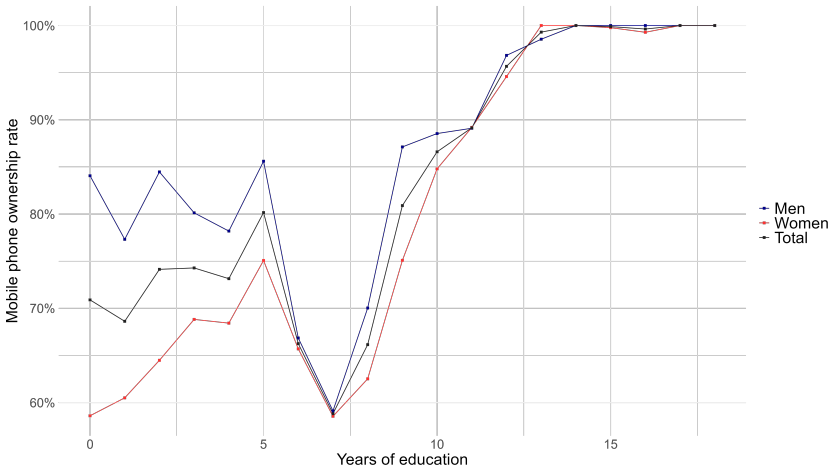
Note: Author's estimations based on the 2017 continuous DHS.

Figure 2.B.2: Mobile phone ownership by zone and gender.



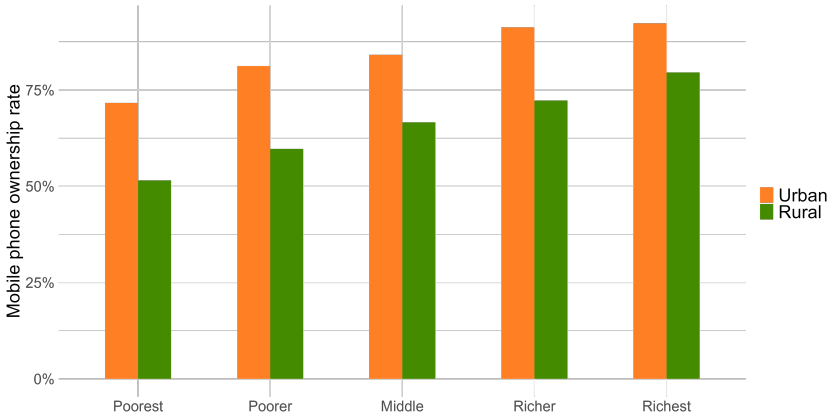
Note: Author's estimations based on the 2017 continuous DHS.

Figure 2.B.3: Mobile phone ownership by years of education and gender.



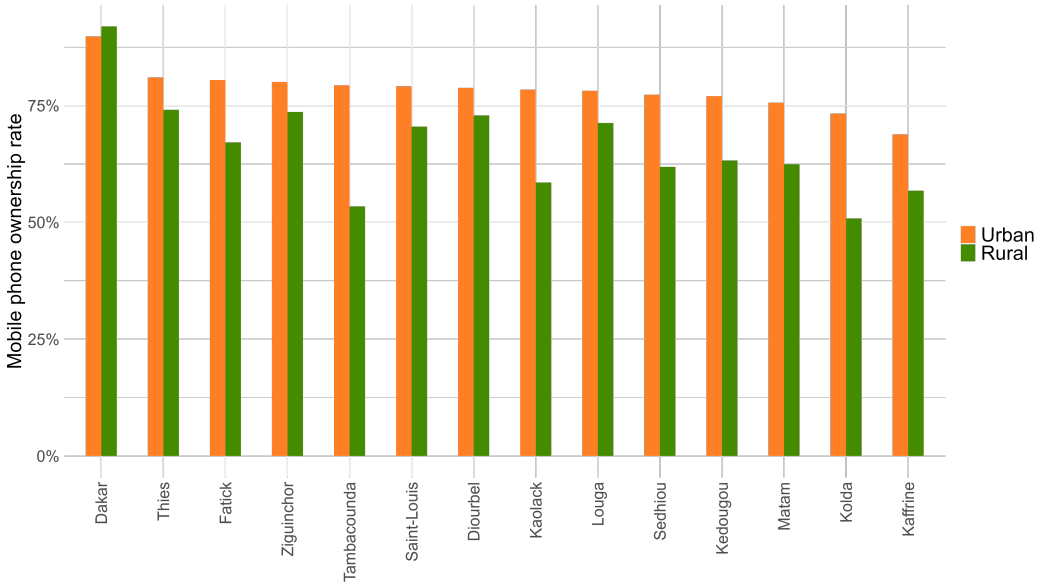
Note: Author's estimations based on the 2017 continuous DHS.

Figure 2.B.4: Mobile phone ownership by wealth category.



Note: Author's estimations based on the 2017 continuous DHS. The wealth classification is based on the DHS rural- and urban-specific wealth indices, which are composite measures of a household's cumulative living standards.

Figure 2.B.5: Mobile phone ownership by region and zone.



Note: Author's estimations based on the 2017 continuous DHS.

Table 2.B.1: Comparison of Sonatel users with the overall population of phone users.

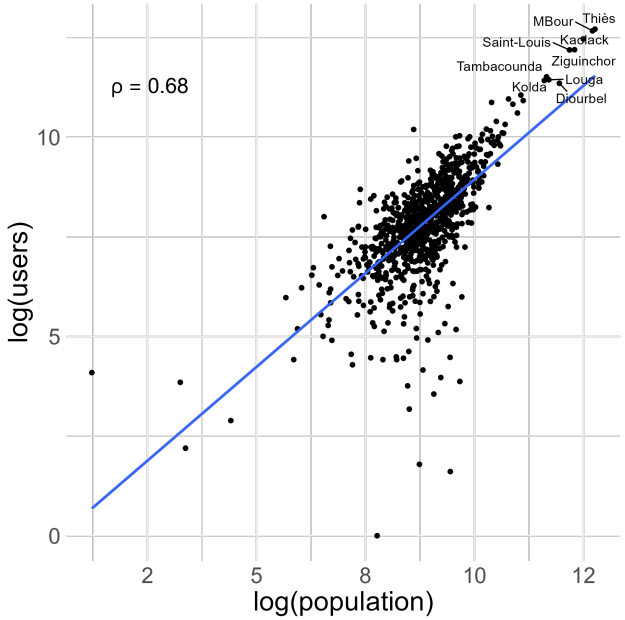
	Sonatel users	**All users**	**Diff.**
Male dummy	0.559	0.554	0.005
Age	37.237	36.975	0.262
Years of education	6.785	6.101	0.685***
Urban dummy	0.574	0.505	0.07***
Has electricity	0.910	0.897	0.013
Has piped water	0.869	0.847	0.022
Has a fridge	0.447	0.415	0.033**
Has a radio	0.710	0.726	-0.016
Has a 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 2.B.2: Rural-urban composition of the base CDR sample.

	Urban (%)	**Rural (%)**
Population	49.3%	50.7%
CDR- 2013	70.5%	29.5%
CDR - 2014/2015	68.6%	31.4%

Note: The definition of urban and rural locations is consistent with that provided in Appendix 2.A.

Figure 2.B.6: Distribution of users across voronoi cells in the base sample excluding Dakar and Touba, 2014-2015.



Note: The blue line represents a linear regression line between the (logged) number of users and the population.

Table 2.B.3: Summary statistics on the base sample characteristics, 2013.

	mean	sd	min	q10	q20	q30	q40	q50	q60	q70	q80	q90	max
Length of obs. (in days)	283.3	111.4	9	85	168	251	320	357	364	365	365	365	365
Distinct days obs.	184.3	119.6	7	30	53	84	126	176	230	280	322	351	365
Fraction of days obs.	0.63	0.28	0.02	0.22	0.34	0.46	0.57	0.67	0.77	0.85	0.92	0.98	1
Number of records	3165	4488	13	242	444	716	1077	1554	2200	3129	4639	7856	36496
Records/days ratio	14.7	13.7	1	4.8	6.1	7.3	8.5	10.1	12.1	14.9	19.7	29.9	100
Maximum time unobs. (in days)	30.4	45.6	0.3	2.3	3.8	5.6	8.1	12.1	18.5	28.9	46.2	80.9	356.9
Number of nights obs.	142.6	106.5	10	20	35	56	85	120	160	205	254	308	366
Fraction of nights obs.	0.49	0.27	0.03	0.13	0.22	0.3	0.38	0.47	0.56	0.66	0.77	0.88	1.11

Table 2.B.4: Summary statistics on the base sample characteristics, 20142-2015.

	mean	sd	min	q10	q20	q30	q40	q50	q60	q70	q80	q90	max
Length of obs. (in days)	525.4	233.1	9	140	272	402	526	633	712	729	730	730	730
Distinct days obs.	306.3	234.5	7	38	71	113	168	243	345	465	581	671	723
Fraction of days obs.	0.57	0.29	0.01	0.15	0.25	0.36	0.48	0.59	0.69	0.79	0.88	0.95	1
Number of records	5733.4	8428.3	12	361	699	1156	1778	2628	3821	5584	8491	14618	72297
Records/days ratio	15.9	13.4	1	5.7	7	8.4	9.8	11.5	13.7	16.8	21.6	31.4	100
Maximum time unobs. (in days)	81.9	127.7	0.2	4	6.9	10.7	16.1	25.1	39.9	66	115	280.3	720.3
Number of nights obs.	232.4	200.3	10	25	47	76	114	166	236	323	429	557	729
Fraction of nights obs.	0.43	0.27	0.01	0.09	0.16	0.23	0.31	0.4	0.49	0.59	0.7	0.83	1.11

Figure 2.B.7: Distribution of users across population density categories in the base sample including Dakar, 2014-2015.

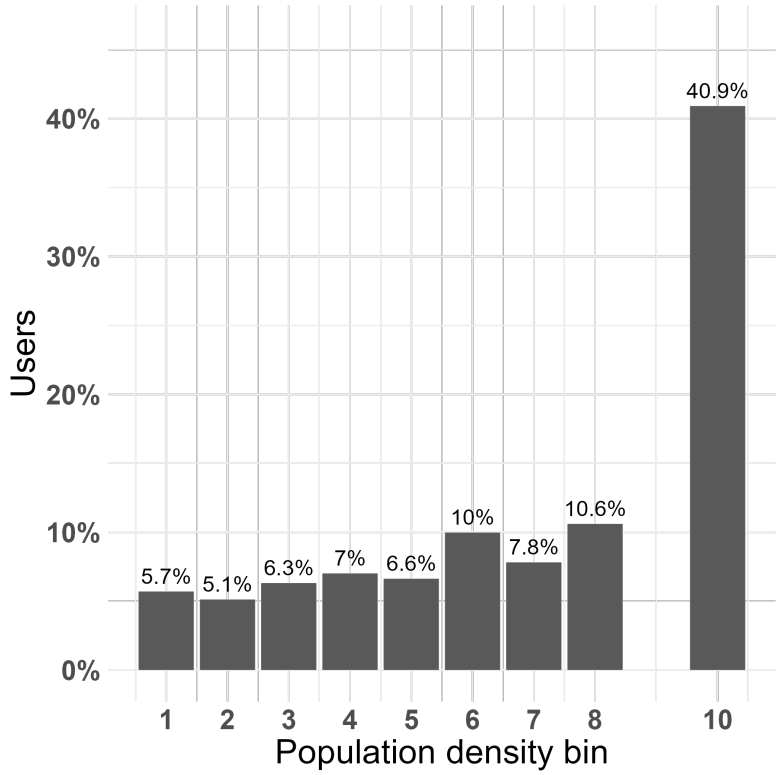
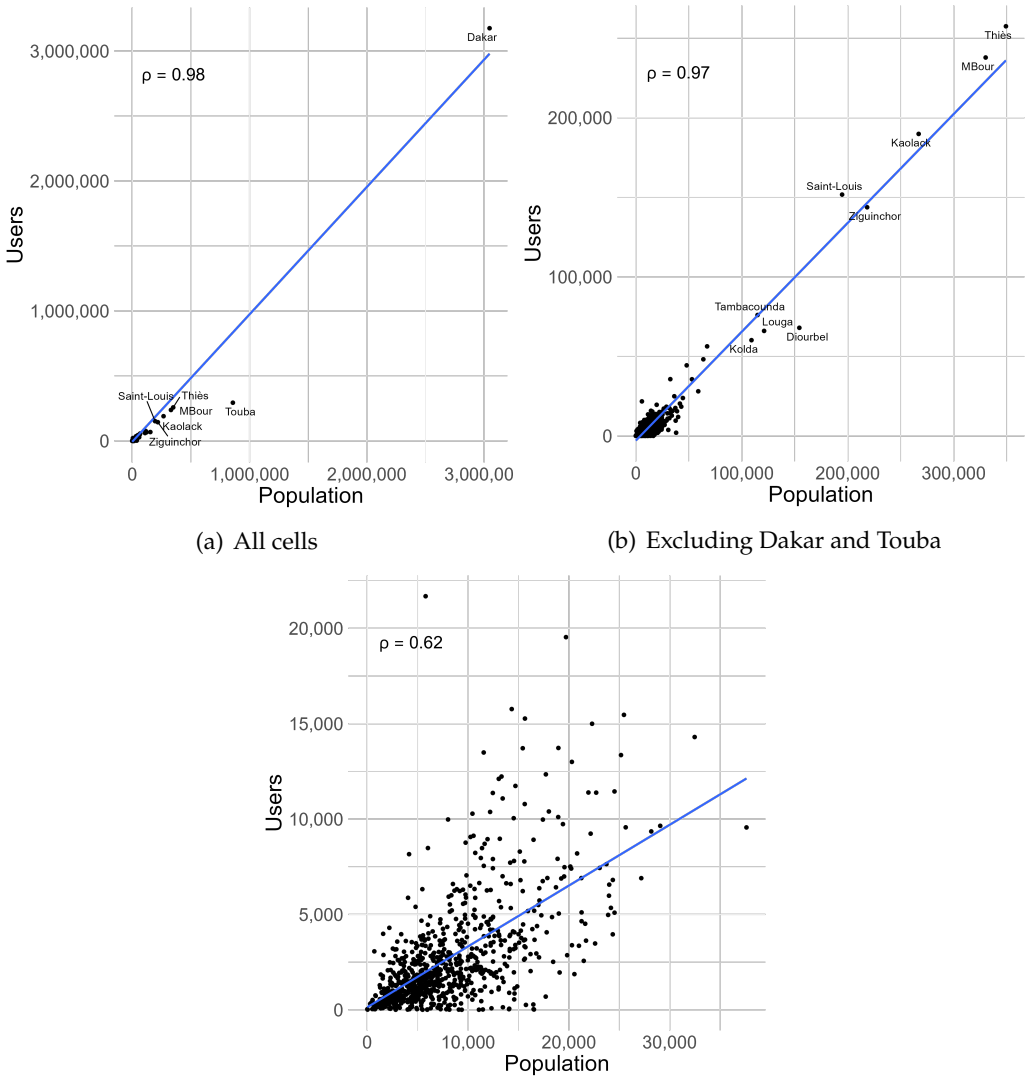


Figure 2.B.8: Distribution of users across voronoi cells in the base sample, 2013.



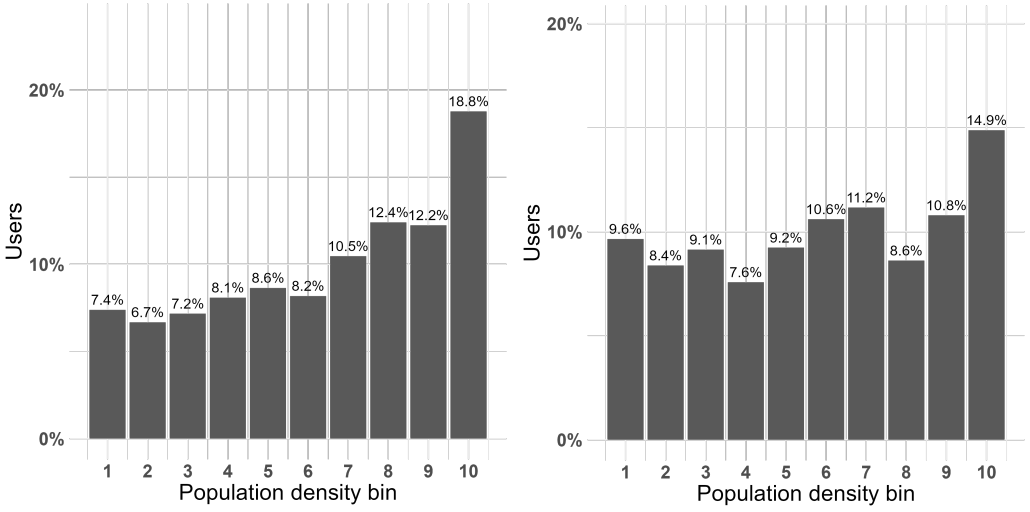
(a) All cells

(b) Excluding Dakar and Touba

(c) Rural cells

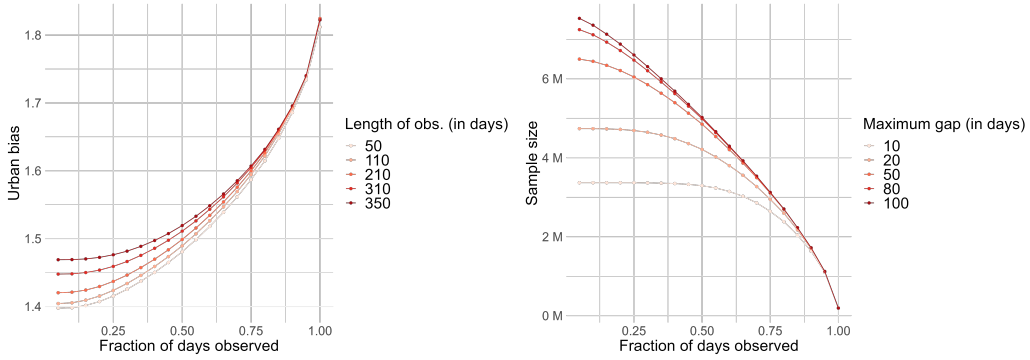
Note: The blue line represents a linear regression line between the number of users and the population at the voronoi-level.

Figure 2.B.9: Distribution of users across population density categories in the base sample, 2013.



(a) All cells except Dakar (b) Rural cells

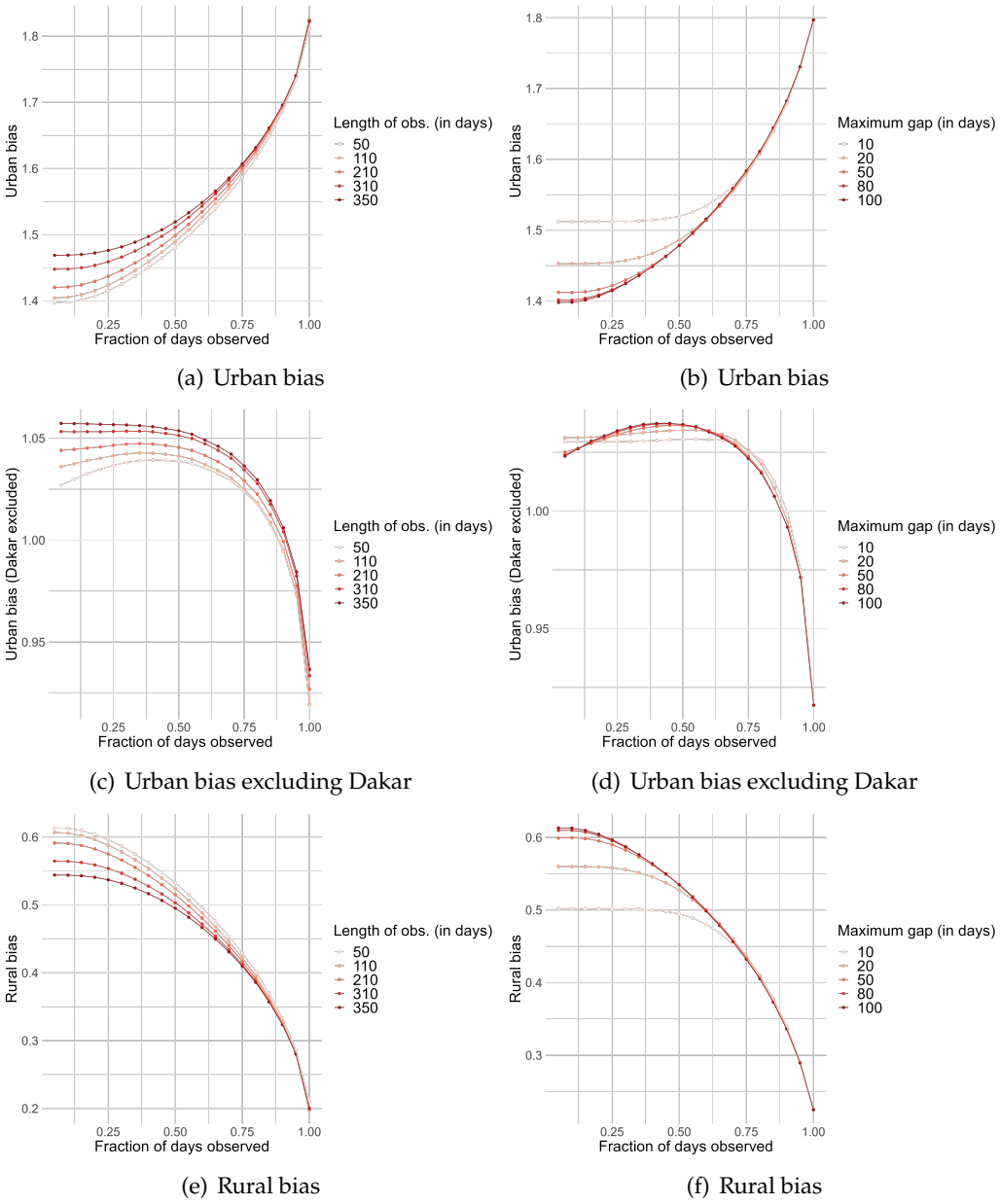
Figure 2.B.10: Impact of filtering parameters on sample size, 2013.



(a) Obs. length and fraction of days (b) Maximum gap and fraction of days

Note:Panel (a) represents sample size as a function of the fraction of days observed imposed on users in the main sample, for a maximum gap parameter set to 100 days and different values for the minimum length of observation. Panel (b) represents sample size as a function of the minimum fraction of days, for a minimal length of observation set to 30 days and different values for the maximum observational gap allowed.

Figure 2.B.11: Impact of filtering parameters on urban and rural biases, 2013.



(a) Urban bias

(b) Urban bias

(c) Urban bias excluding Dakar

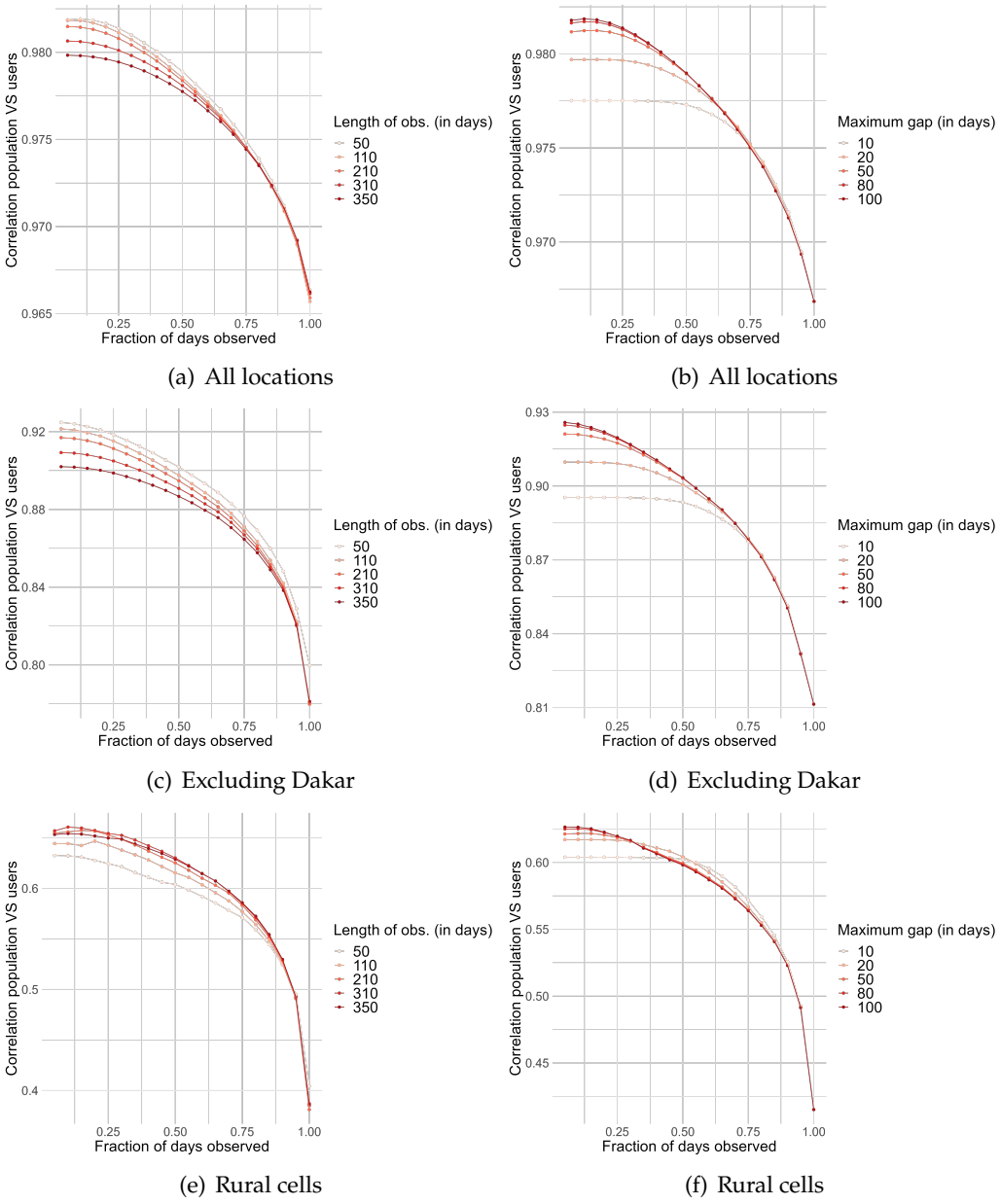
(d) Urban bias excluding Dakar

(e) Rural bias

(f) Rural bias

Note: Panel (a) represents the urban bias as a function of the fraction of days observed imposed on users in the main sample, for a maximum gap parameter set to 100 days and different values for the minimum length of observation. Panel (b) represents the urban bias as a function of the minimum fraction of days, for a minimal length of observation set to 30 days and different values for the maximum observational gap allowed. Panel (c) and (d) show the same results excluding Dakar, and panel (e) and panel (f) represent the rural bias.

Figure 2.B.12: Impact of filtering parameters on the correlation between users and population across locations, 2013.



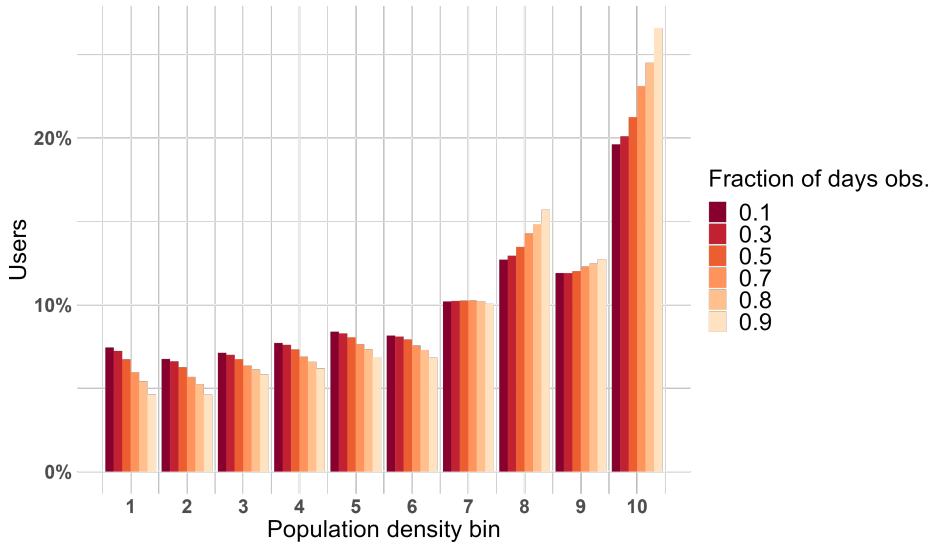
(a) All locations (b) All locations

(c) Excluding Dakar (d) Excluding Dakar

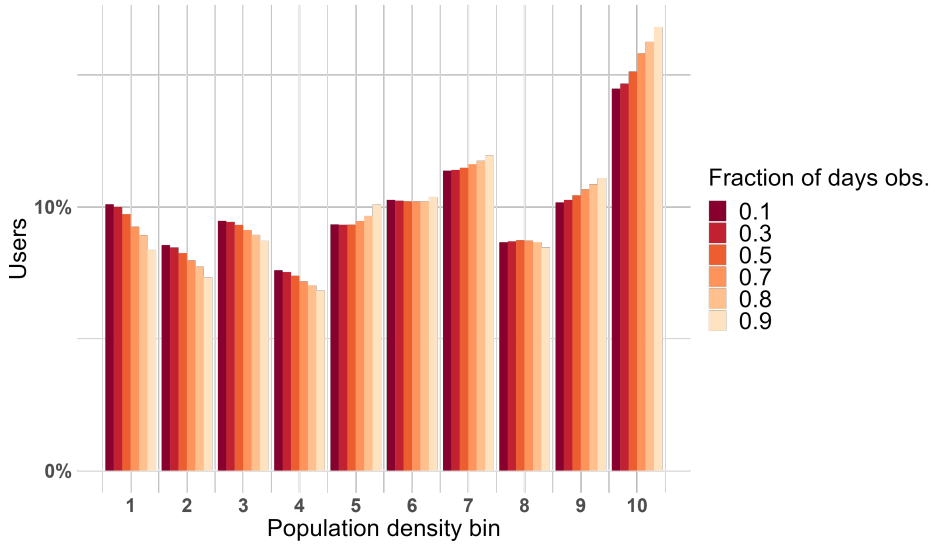
(e) Rural cells (f) Rural cells

Note: Panel (a) represents the correlation between population and users across cell locations as a function of the fraction of days observed imposed on users in the main sample, for a maximum gap parameter set to 100 days and different values for the minimum length of observation. Panel (b) represents the same correlation as a function of the minimum fraction of days, for a minimal length of observation set to 30 days and different values for the maximum observational gap allowed. Panel (c) and (d) show the same results excluding Dakar, and panel (e) and panel (f) represent the correlation between population and users across the subset of rural cells.

Figure 2.B.13: Impact of the minimal fraction of days observed on the distribution of users across population density categories.



(a) All cells except Dakar



(b) Rural cells

Note: Panel (a) represents the distribution of users across density categories for different values of the minimal fraction of days observed imposed on users, and setting the minimal observation length to 310 days and the maximum observational gap to 100 days. Density categories correspond to groups of locations that account for 10% of the population, excluding Dakar. Panel (b) presents the same results considering the subset of rural cells.

Appendix 2.C Sensitivity analysis of temporary migration detection accuracy to observational characteristics

As illustrated in summary statistics in Tables 2.B.3 and 2.B.4, sampling characteristics vary significantly across users. Digital traces are oftentimes characterized by high levels of attrition – users leaving (and entering) the sample over the period covered – and irregular sampling issues; mobile phone users do not necessarily make and/or receive calls on a regular basis. Observational requirements for the measure of human mobility necessarily depend on the type of movements one seeks to identify. For instance, measuring long-term changes in the place of residence requires long periods of observation (e.g. several years) with modest sampling frequencies, whereas capturing commuting movements asks for high sampling frequencies (e.g. multiple observations per day) with observation periods that can be relatively short. 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) identify a home location and (ii) detect temporary changes in the usual place of residence. I 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, I 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.

To do this, I consider a benchmark subset of users in the 2013 dataset that meet stringent observational constraints: they are observed for at least 360 days and on at least 95% of days.[55] Moreover, I select users with a unique home location identified and with at least one migration event of at least 20 days detected. I randomly select 10,000 users that meet those criteria.[56] The strict observational constraints imposed on this subset allow to reasonably consider the migration detection outputs as reflecting (i) the actual home locations of users and (ii) their actual temporary migration moves. To test the sensitivity of the model accuracy with respect to users' sampling characteristics, I consider random sub-trajectories of users in this benchmark sample satisfying different sets of observational constraints

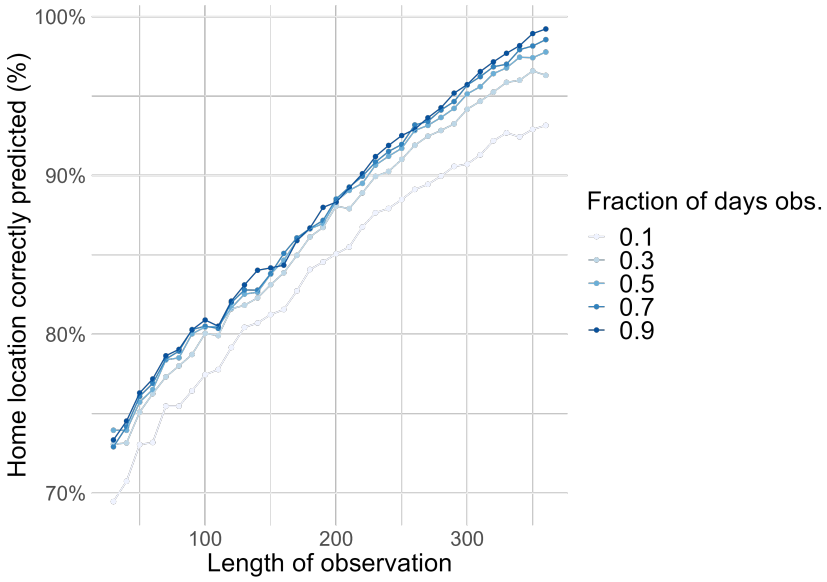
[55]The definition of daily locations from raw CDR trajectories is provided in section 2.E. It essentially corresponds to the modal location at night (6pm-8pm) when observations at night are available, and to the daytime modal location otherwise.

[56]I find a total of 195,070 users satisfying those constraints.

and I re-estimate the detection model and compare the outputs with those obtained with the full trajectories.

First, I 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.[57] For each set of parameters (Δ, Ω), I simply define the model accuracy as the fraction of users with a correctly predicted home location. Figure 2.C.1 shows estimates of the model accuracy for length 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; e.g. 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 almost linearly with the length of observation. For $\Omega = 0.9$, it increases from 73% to 99% when considering lengths of observation of 30 and 360 days respectively.

Figure 2.C.1: Model accuracy for home location predictions.



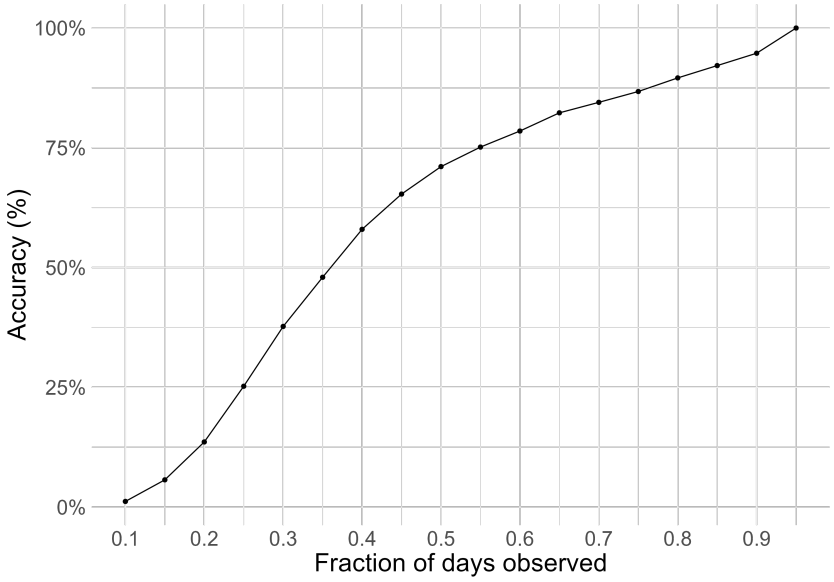
Second, I focus on the impact of Ω on the accuracy of the migration detection model, holding Δ fixed.[58] Here, I define the model accuracy for any given value of Ω as the fraction of real migration segments (i.e. those detected in the benchmark subset) that are effectively identified in sub-trajectories of density Ω. Removing

[57] Since I consider users with a unique home location, the latter is simply defined as the modal daily location over the period of observation.

[58] 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.

observations from a full trajectory can lead to migration events being still detected although with slightly different start and end dates. I therefore consider that a real migration segment is identified in a sub-trajectory of density Ω if a migration segment overlapping at least half real migration segment is detected in this sub-trajectory. Results for Δ set to 360 days and Ω varying from 0.1 to 0.95 are provided in Figure 2.C.2. Unsurprisingly, the frequency of observation has a significant impact on the accuracy of the migration detection model: from 95% with a density of 0.9, it decreases to as low as 1% 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 0.35. On the other hand, densities greater than 0.8 allow to sustain a high level of accuracy, i.e. beyond 90%.

Figure 2.C.2: Model accuracy for migration event detection.



Appendix 2.D An empirical test for non-random attrition

Non-random observational gaps correspond to extended periods of time during which a user is in migration but use a different SIM card and is thus not observed in the data. This potential phenomenon is problematic for two reasons. First, working with subsets of users with high sampling frequencies de facto tends to exclude those with significant observational gaps. To the extent that observational gaps may coincide with migration events, the resulting subset is mechanically biased on the cross-section via the exclusion of a relatively more mobile segment of the population. Second, when such users are still maintained in the sample, they tend to bias migration rates downward by inflating the denominator while being wrongly classified as non-migrants.

To test for the existence of non-random observational gaps, I assume that, for any time period t, a constant fraction α of the total number of users in migration is systematically unobserved:

$$\frac{\tilde{N}_t^{migrant}}{N_t^{migrant} + \tilde{N}_t^{migrant}} = \alpha \tag{2.4}$$

Where $\tilde{N}_t^{migrant}$ is the number of users in migration not observed at time t and $N_t^{migrant}$ the number of users observed in migration, so that $N_t^{migrant} + \tilde{N}_t^{migrant}$ is the total number of users in migration. Re-arranging terms in equation 2.4, $\tilde{N}_t^{migrant}$ can be expressed as:

$$\tilde{N}_t^{migrant} = \frac{\alpha}{1-\alpha} N_t^{migrant} \tag{2.5}$$

I further assume that the number of users not observed and not in migration, $\tilde{N}_t^{home}$, is orthogonal to the observed number of migrants and can be written as $\tilde{N}_t^{home} = \beta_0 + \epsilon_t$, where ϵ_t is a random error term. Adding $\tilde{N}_t^{home}$ on both sides of equation 2.5:

$$\tilde{N}_t = \beta_0 + \beta_1 N_t^{migrant} + \epsilon_t \tag{2.6}$$

Where $\tilde{N}_t = \tilde{N}_t^{home} + \tilde{N}_t^{migrant}$ is the total number of users not observed at time t and $\beta_1 = \frac{\alpha}{1-\alpha}$.

I empirically estimate equation 2.6 using a subset of users with relatively low observational constraints that naturally allow for some users to showcase periods of inactivity. I select 10,000 users in the 2013 dataset and 10,000 users in the 2014-2015 dataset, who are seen for a period of at least 360 days and on at least 50% of days. Users are randomly selected in third-level administrative units and within urban and rural strata, so that the distribution of this subset is broadly in line with the

population. The migration detection algorithm is applied to infer the number of migrants by half-month over the period 2013-2015. Next, I calculate the number of users "not observed" by half-month. I define a user as being not observed during half-month h if an observation gap of at least 20 days – i.e. equivalent to the minimum duration imposed for my definition of temporary migration – within the user's period of observation overlaps h on at least 8 days. I exploit temporal variations in the number of users in migration and users unobserved to estimate equation 2.6 with OLS. Regression results are presented in Table 2.D.1 and illustrated graphically with scatter plots in Figure 2.D.1(a) and 2.D.1(b). Column (1) shows the results of the OLS estimation considering the total number of users not observed at the national-level as a dependent variable. The coefficient is positive but not statistically significant. Columns (3) and (5) show estimates that consider the subset of rural and urban users respectively. Neither coefficients are statistically significant but both are again positive, with the magnitude of the rural coefficient being roughly ten times larger than the urban coefficient. Columns (2), (4) and (6) show the same results introducing the squared number of migrants observed as an independent variable, in an attempt to relax the assumption of a constant fraction of users in migration being unobserved (parameter α). Estimated coefficients remain statistically non-significant.

These results support the idea that non-random observational gaps are most likely not a major concern for the construction of migration statistics with CDR data. Even though results are not statistically significant, if anything, they are still suggestive of the existence of a positive relationship between the number of observed migrants and users unobserved in the rural subset. They do not exclude the possibility that some observational gaps effectively coincide with some users being in migration at a destination where they use a different SIM card. They simply indicate that the extent of this phenomenon does not constitute a clear impediment to the construction of migration statistics. It is important to highlight that conclusions drawn from the analysis above rest on the assumption that users who switch their SIM card when they travel and those who do not are somehow comparable. More specifically, it is implicitly assumed that the probabilities of being in migration within these two sub-populations follow comparable trends over time. Yet, it may well be that users in migration at specific times of the year are, for instance, more likely to switch SIM cards when they travel. As it does not constitute the core of this paper, I leave it to future research to investigate this issue more in depth.

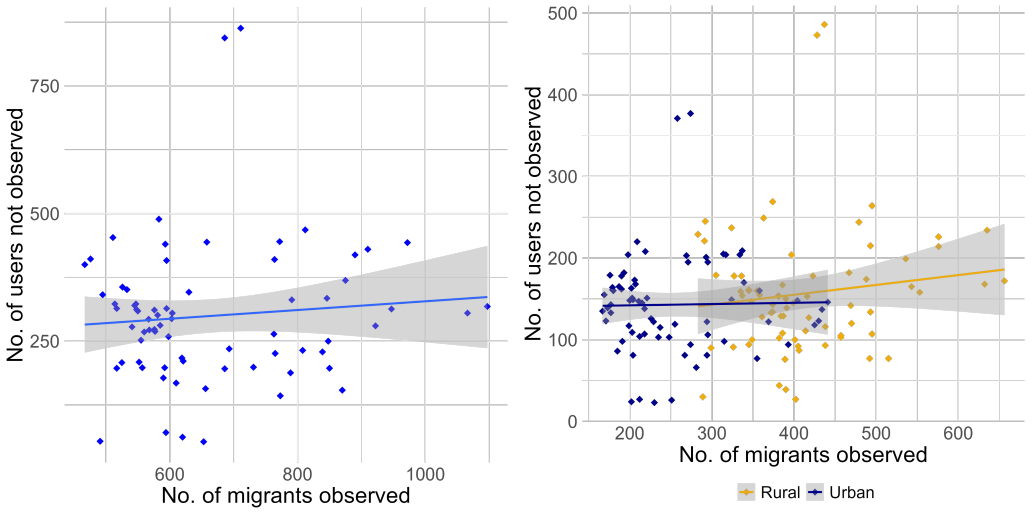
Table 2.D.1: Regression testing the existence of non-random observational gaps.

	Total		Rural		Urban	
	(1)	(2)	(3)	(4)	(5)	(6)
No. of migrants obs.	0.086	−0.067	0.121	−0.629	0.016	0.572
	(0.101)	(1.401)	(0.108)	(1.234)	(0.080)	(1.038)
No. of migrants obs.^{2}		0.0001		0.001		−0.001
		(0.001)		(0.001)		(0.002)
Observations	72	72	72	72	72	72
R^2	0.008	0.009	0.017	0.027	0.0004	0.008
Adjusted R^2	−0.006	−0.020	0.003	−0.002	−0.014	−0.021

*p<0.1; **p<0.05; ***p<0.01

Note: Each observation represents a half-month over the period 2013-2015. Each column shows results of a regression of the number of users not observed on the number of observed migrants. In columns (1) and (2), I consider the total numbers at the national-level while columns (3)-(4) and and columns (5)-(6) show estimations that consider sub-totals for rural and urban locations respectively. Standard errors in parentheses are robust to heteroskedasticity and autocorrelation. They are derived from the Newey-West HAC estimator of the variance-covariance matrix.

Figure 2.D.1: Number of users not observed against the number of migrants observed.



(a) National (b) By origin zone

Note: Each point represents a couple (number of migrants observed,number of users not observed) for a half-month of the period 2013-2015.

Appendix 2.E Migration detection algorithm

The temporary migration detection algorithm proceeds in three stages. First, implement an algorithm that smooths out short-term events and temporary migrations from estimated monthly locations in order to detect macro-segments of home locations. Those represent periods of time of at least 6 months during which a user is consistently seen at a single location that we consider his usual place of residence (i.e. his home location), and within which short-term movements and temporary migrations can occur. Second, we identify meso-segments within daily location trajectories by smoothing out short-term mobility events only. They correspond to periods of time over which a user is primarily observed at a single location that may or may not be his home location, allowing for short-term movements. We eventually overlay the corresponding macro- and meso-trajectories to find temporary migration events, which are defined as meso-segments of at least 20 days at non-home locations. We provide below full methodological details on each of those stages.

2.E.1 First stage: macro-segment detection

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, which I denote by $\mathcal{L} = (\ell_k)_{k\in[1;L]}$, with L the total number of locations. In the present case, $\mathcal{L}$ is the set of voronoi cells introduced in section 2.2. The raw CDR trajectory of a user i is denoted by $(x^i_{t_1}, x^i_{t_2}, ..., x^i_{t_{T_i}})$, where each $x^i_t \in \mathcal{L}$ represents i's observed location at timestamp t. T_i is i's total number of CDR.

I implement a hierarchical frequency-based method as in Blumenstock, Chi, et al. (2022) to determine monthly locations. For a user i, the hourly location $x^i_{h,d}$ for an hour h of day d is defined as the most frequently visited location during that one-hour time interval, which I denote h_d:

$$x^i_{h,d} = mode\Big\{ x^i_t \,\Big|\, t \in (t_1, ..., t_{T_i}),\ t \in h_d \Big\} \tag{2.7}$$

Hourly locations are then aggregated up to daily locations, which are calculated as the most frequent hourly location. As is customary in the literature, night hours between 6pm and 8am are preferred to determine daily locations 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. To limit the loss of information induced by this filtering procedure, I also calculate daily locations based on daytime

hourly location between 8am and 6pm and assign those values to those user-days that do not have observations at night.[59] I denote the set of night hours for day d as $\mathcal{N}_d = \left\{ (h,d) \,\middle|\, (h,d) \in \{(18,d-1),...,(23,d-1)\} \cup \{(0,d),...,(7,d)\} \right\}$ and the set of daytime hours is $\overline{\mathcal{N}_d}$. Then, $\mathcal{D}_i = \{d_1^i,...,d_{D_i}^i\}$ is the set of D_i observed days for user i so that the daily location of user i on any day $d \in \mathcal{D}_i$ is given by:

$$x_d^i = \begin{cases} mode\left\{ x_{h,d}^i \,\middle|\, (h,d) \in \mathcal{N}_d \right\}, & \text{if } \left\{ x_{h,d}^i \,\middle|\, (h,d) \in \mathcal{N}_d \right\} \neq \varnothing \\ mode\left\{ x_{h,d}^i \,\middle|\, (h,d) \in \overline{\mathcal{N}_d} \right\}, & \text{otherwise} \end{cases} \tag{2.8}$$

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[60]. Note that frequency-based monthly location estimates naturally smooth out short-term mobility events.

Then, a segment detection algorithm is applied to the monthly location dataset to identify macro-segments of home locations. The algorithm proceeds in four steps:

i. Preliminary unique home location estimation:
In a preliminary step, 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:

$$\overline{home_i} = mode\left\{ x_d^i \,\middle|\, d \in \mathcal{D}_i \right\} \tag{2.9}$$

ii. Detect contiguous monthly locations:
Consecutive months at the same location are grouped together, allowing for within-group observation gaps of at most ϵ_{gap}^{macro} months. ϵ_{gap}^{macro} is set such that no permanent change in the home location could occur during unobserved periods. I define a permanent change in the place of residence as a migration of at least τ^{home}, which I set to 6 months, and I therefore choose ϵ_{gap}^{macro} also equal to 6 months.[61]

iii. Merge monthly location groups:
Groups of months at a single location are then merged when they are separated by other groups that account for a total duration that is strictly less than ϵ_{gap}^{macro}

[59] Blumenstock (2012) shows that restricting the sample to locations at night as virtually no impact on the temporary migration measures he derives from CDR data in Rwanda.

[60] The monthly location for user-months with 9 days observed or less thus show up as missing in the final dataset.

[61] Note that, in practice, observation gaps are generally much shorter given the observational constraints considered.

months. This step essentially allows to group home stays that are separated by potential temporary migration spells.

iv. Resolve overlap:
Next, the overlap between merged groups that may result from the previous step (Chi et al., 2020) is resolved. First, merged groups with a duration strictly lower than τ^{home} months are removed: as per the definition adopted, they cannot be home macro-segments. For two consecutive overlapping groups, overlapping months are simply assigned to the longest group. Start and end dates of merged groups are updated accordingly and merged groups with a duration strictly lower than τ^{home} months are removed. To address rare cases of multiple overlaps, this procedure is iterated until no overlapping groups are left. For each user, the merged groups left 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^{home}, the vast majority of users ends up with only one macro-segment detected. Those users are assigned the default unique home location determined in the first step of the macro-segment detection procedure. For other users with at least two macro-segments detected, a monthly home location dataset is produced.

2.E.2 Second stage: meso-segment detection

A comparable approach is used to detect meso-segments. 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} days. Small values of ϵ_{gap}^{meso} may fail to smooth out short-mobility events while larger values are associated with potentially large overlap between segments. I rely on Chi et al. (2020) to determine a reasonable value for ϵ_{gap}^{meso} and set it to the optimal value of 7 days they infer from a cross-validation exercise.

ii. Merge daily location groups:
Groups of daily locations are then merged when they are less than ϵ_{gap}^{meso} days apart. For each user, we obtain a set of intermediary meso-segments. As in Chi et al. (2020), I filter out meso-segments with a proportion of days at destination lower than some parameter ϕ, that I set equal to 0.5.[62]

[62]This value is roughly in line with the optimal value found in Chi et al. (2020).

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. 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. This process is iterated until no overlap is left.

2.E.3 Identification of temporary migration events

Temporary migration events are defined as meso-segments of at least τ^{temp} days, at a destination that is not the user's home location. Three attributes are therefore required for a meso-segment to be classified as a temporary migration episode, a home stay or a visit to a non-home location: location, duration and home location for the period covered by the segment.

Meso-segment locations are directly obtained as an output of the meso-segment detection procedure. Then, for most users that have a unique home location (i.e. no permanent migration detected), the definition of a home location at the level of meso-segments is straightforward. For other users with multiple home locations across the period of observation, each meso-segment is assigned a home location by overlaying the macro- and meso- trajectories. If a meso-segment is entirely covered by a macro-segment, it is assigned the corresponding home location. If a meso-segment overlaps between two macro-segments, it is assigned the home location of the macro-segment with the largest overlap. Finally, we calculate lower-bound and upper-bound estimates for a meso-segment duration. For any segment S_i of user i, the lower-bound duration $minDuration(S_i)$, that is referred to as the "observed duration", corresponds to the time elapsed between the start and end dates of the segment. The upper-bound duration $maxDuration(S_i)$, referred to as the "maximum duration", is the time elapsed between the observed day just preceding the segment and the observed day directly following S_i. The relative gap between the lower- and upper-bound duration estimates represents the uncertainty in the meso-segment duration measure. For instance, a set of contiguous meso-segments will result in observed duration estimates being exactly equal to maximum duration estimates, so that the level of uncertainty will be null. In general terms, users with a higher frequency of observation are associated with lower uncertainty in the actual start and end dates of location meso-segments, and therefore in the estimation of their duration.

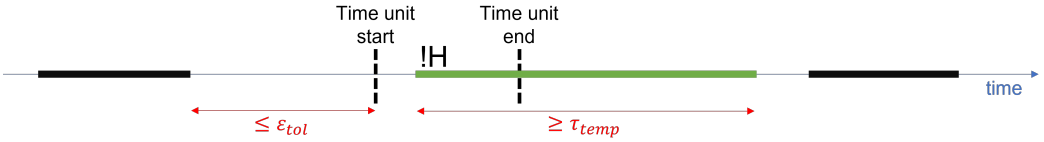
Appendix 2.F Algorithmic rules to aggregate user-level trajectories

The following diagrams illustrate the algorithmic rules used to identify migration departures, migration returns and the status of migration for a given time unit, for both high- and low-confidence estimates. An explanatory note below each diagram provides a description of the corresponding configuration and the criteria applied. Thick segments along the time arrow reflect meso-segments for a hypothetical user while empty spaces correspond to observational gaps. Segment locations are indicated at the top left of segments, using some common notations throughout all diagrams. "H" denotes the home location, "!H" any non-home location and "not !H" simply means "any location that is not !H". When no location is specified, it is assumed that the corresponding segment could be at any location.

It is important to note that some configurations are somehow redundant from a logical perspective, but need to be treated separately in the algorithm. All cases are still presented in this appendix with the intent to facilitate the understanding of the code for anyone wishing to reproduce or simply use it.

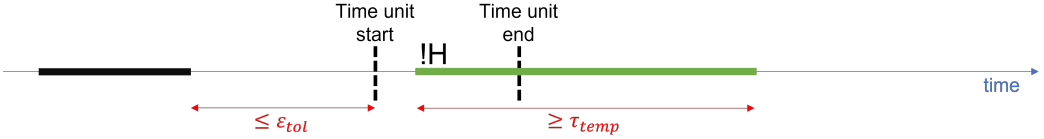
2.F.1 Identifying migration departures: high-confidence

Diagram 2.F.1: High-confidence migration departure: case 1



Note: The green segment is a migration segment with a start date within the time unit. The observation gap between the time unit start date and the observation preceding the green segment is lower than the tolerance parameter $\epsilon^t ol$. As a result, the start date of the green segment is counted as a migration departure in the high-confidence estimation.

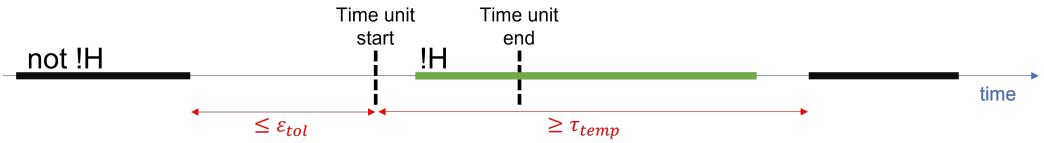
Diagram 2.F.2: High-confidence migration departure: case 2



Note: This is the same configuration as in diagram 2.F.1, but the user exits the sample at the end of the green segment.

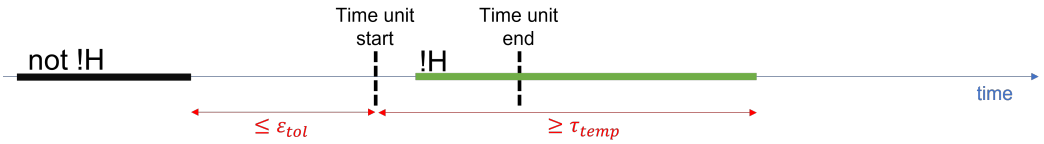
2.F.2 Identifying migration departures: low-confidence

Diagram 2.F.3: Low-confidence migration departure: case 1



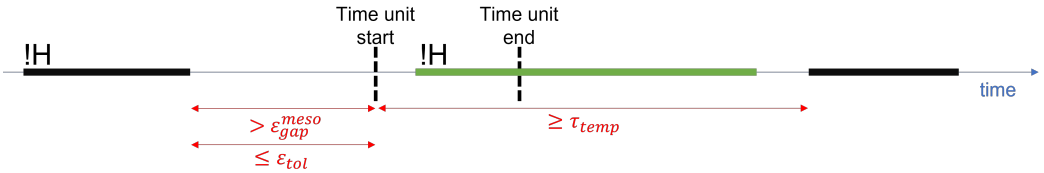
Note: The green segment at !H does not necessarily have an observed duration greater than τ^{temp}. The maximum duration possible to consider the segment started during the time unit is the time elapsed between the time unit start date and the day preceding the start of the following segment. When this is greater than τ^{temp} and the tolerance criterion is not exceeded, this configuration results in one additional migration departure in the low-confidence estimate.

Diagram 2.F.4: Low-confidence migration departure: case 2



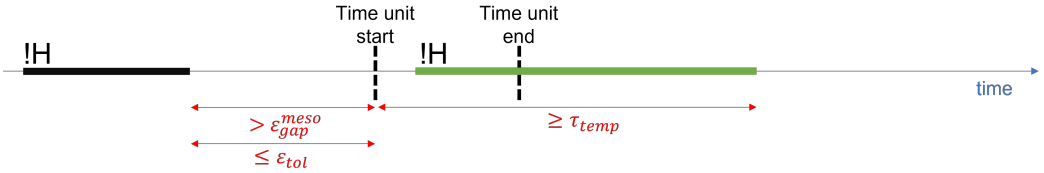
Note: This is the same configuration as in diagram 2.F.3, although the user exits the sample at the end of the green segment. The maximum duration possible to consider the segment started during the time unit is the time elapsed between the time unit start date and the observed end date of the segment. If this is greater than $\tau^{t}emp$ and the tolerance criterion is not exceeded, this configuration results in one additional migration departure in the low-confidence estimate.

Diagram 2.F.5: Low-confidence migration departure: case 3



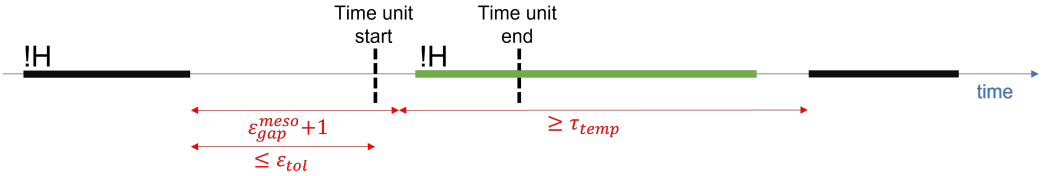
Note: This is the same configuration as in diagram 2.F.3, although the segment preceding the green segment is at the same location !H. In this case, the green segment cannot have started less than $\epsilon_{gap}^{meso} + 1$ days after that segment. The diagram presents a situation where the gap between the end of the preceding segment and the first day of the time unit is strictly larger than ϵ_{gap}^{meso}, so that the green segment may have started on the first day of the time unit. The situation is then equivalent to that described in diagram 2.F.3.

Diagram 2.F.6: Low-confidence migration departure: case 4



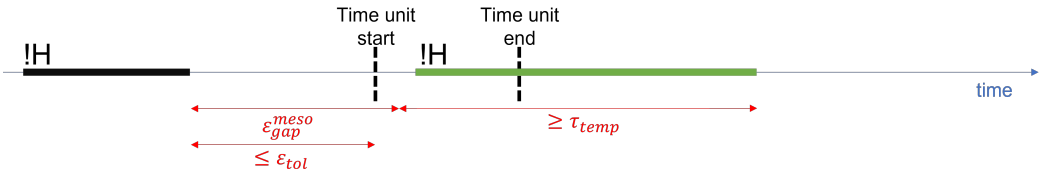
Note: This is the same configuration as in diagram 2.F.5, but the user exits the sample at the end of the green segment. The maximum duration is therefore lower as the maximum end date coincides with the observed end date.

Diagram 2.F.7: Low-confidence migration departure: case 5



Note: This is the same configuration as in diagram 2.F.5, although the date corresponding to $\epsilon_{gap}^{meso} + 1$ days after the end of the preceding segment falls within the time unit. The maximum duration of the green segment is therefore slightly lower because the minimum start date possible for the green segment is greater than the first day of the time unit.

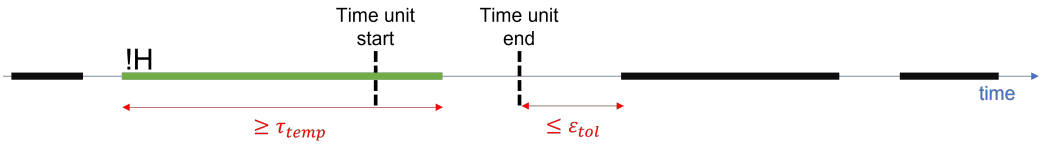
Diagram 2.F.8: Low-confidence migration departure: case 6



Note: This is the same configuration as in diagram 2.F.7, but the user exits the sample at the end of the green segment. The maximum duration is therefore lower as the maximum end date coincides with the observed end date.

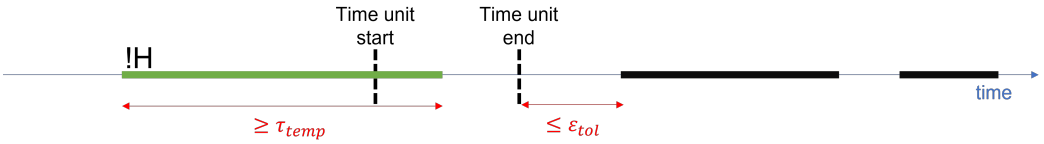
2.F.3 Identifying migration returns: high-confidence

Diagram 2.F.9: High-confidence migration return: case 1



Note: The green segment is at a non-home location and has a duration greater than τ^{temp}: it is a migration segment. The observed end date falls within the time unit but the observation gap following the segment indicates that the user may have actually returned after the time unit. Since the time elapsed between the end of the time unit and the day preceding the following segment is less than the tolerance criterion ϵ^{tol}, the user is considered to have returned during the time unit in the high-confidence estimate.

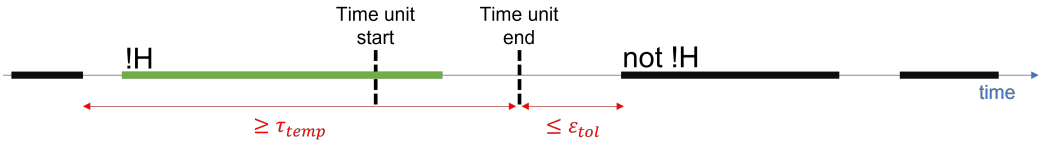
Diagram 2.F.10: High-confidence migration return: case 2



Note: This is the same configuration as in diagram 2.F.9, but the user is never observed before the green segment.

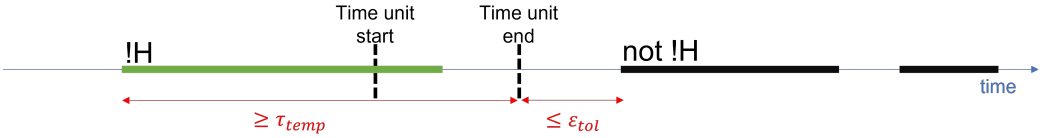
2.F.4 Identifying migration returns: low-confidence

Diagram 2.F.11: Low-confidence migration return: case 1



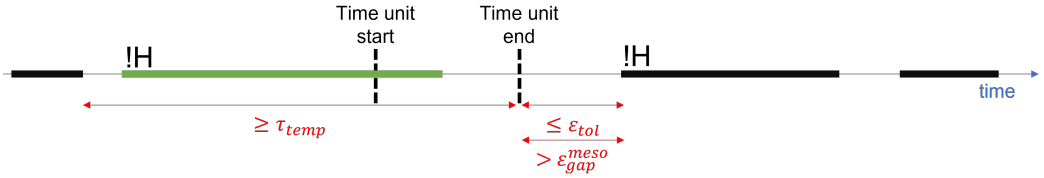
Note: The green segment at !H does not necessarily have an observed duration greater than τ^{temp}. The maximum duration possible to consider the segment ended during the time unit is the time elapsed between day following the end date of the preceding segment and the time unit end date. When this is greater than τ^{temp} and the tolerance criterion is not exceeded, this configuration results in one additional migration return in the low-confidence estimate.

Diagram 2.F.12: Low-confidence migration return: case 2



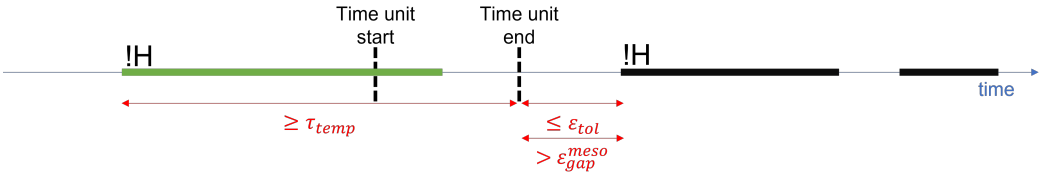
Note: This is the same configuration as in diagram 2.F.11, although the user is never observed before the green segment. The maximum duration possible to consider the segment ended during the time unit is the time elapsed between the observed start date of the green segment and the time unit end date. If this is greater than $\tau^t emp$ and the tolerance criterion is not exceeded, this configuration results in one additional migration return in the low-confidence estimate.

Diagram 2.F.13: Low-confidence migration return: case 3



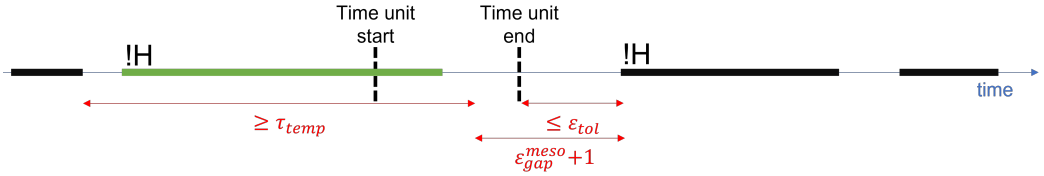
Note: This is the same configuration as in diagram 2.F.11, although the segment following the green segment is at the same location !H. In this case, the green segment cannot have ended less than $\epsilon_{gap}^{meso} + 1$ days before that segment. The diagram presents a situation where the gap between the start of the following segment and the last day of the time unit is strictly larger than ϵ_{gap}^{meso}, so that the green segment may have lasted up until the very last day of the time unit. The situation is then equivalent to that described in diagram 2.F.11.

Diagram 2.F.14: Low-confidence migration return: case 4



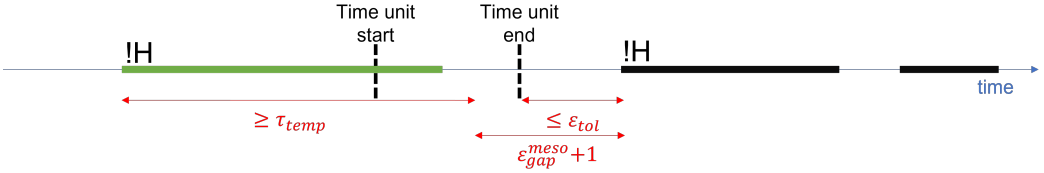
Note: This is the same configuration as in diagram 2.F.13, but the user is never observed before the green segment. The maximum duration is therefore lower as the minimum start date coincides with the observed start date. The situation is otherwise equivalent to that of diagram 2.F.13.

Diagram 2.F.15: Low-confidence migration return: case 5



Note: This is the same configuration as in diagram 2.F.13, although the date corresponding to $\epsilon_{gap}^{meso} + 1$ days before the start of the following segment falls within the time unit. The maximum duration of the green segment is therefore slightly lower because the maximum end date possible for the green segment precedes the last day of the time unit. The situation is otherwise equivalent to that of diagram 2.F.13.

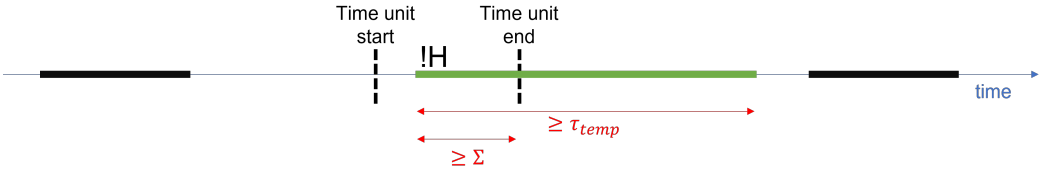
Diagram 2.F.16: Low-confidence migration return: case 6



Note: This is the same configuration as in diagram 2.F.15, but the user is never observed before the green segment. The maximum duration is slightly lower because the minimum start date coincides with the observed start date. The situation is otherwise equivalent to that of diagram 2.F.15.

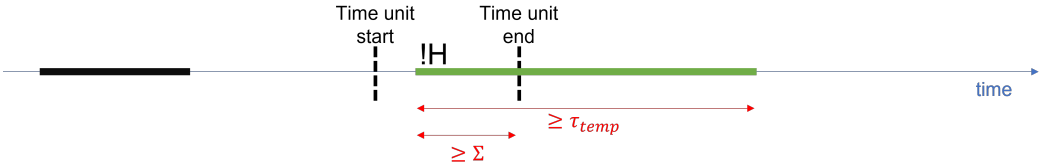
2.F.5 Identifying migration status: high-confidence

Diagram 2.F.17: High-confidence migration status: case 1



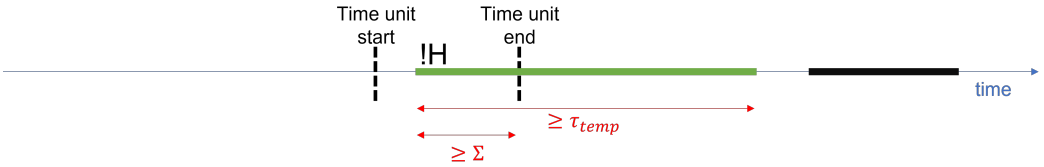
Note: The green segment is at a non-home location and has a duration greater than τ^{temp}: it is a migration segment. It overlaps on the right of the time unit for a duration of at least Σ days. The user is therefore considered as being in migration during the time unit in the high-confidence estimate.

Diagram 2.F.18: High-confidence migration status: case 2



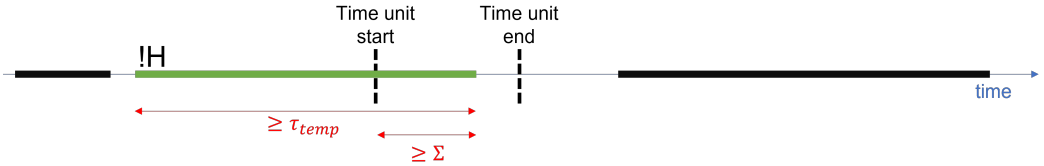
Note: This is the same configuration as in diagram 2.F.17, but the user exits the sample at the end of the green segment. The situation is otherwise equivalent to that of diagram 2.F.17.

Diagram 2.F.19: High-confidence migration status: case 3



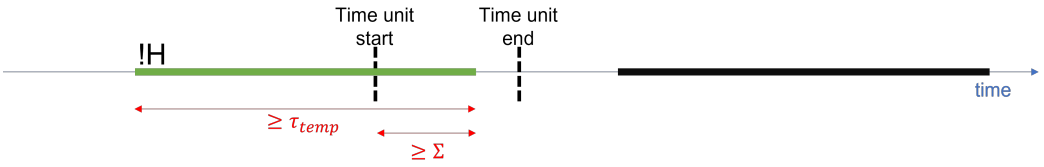
Note: This is the same configuration as in diagram 2.F.17, but the user is never observed before the green segment. The situation is otherwise equivalent to that of diagram 2.F.17.

Diagram 2.F.20: High-confidence migration status: case 4



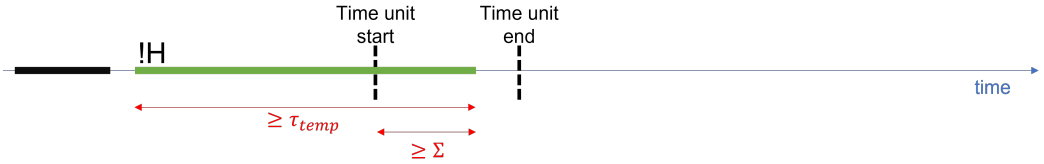
Note: The green segment is at a non-home location and has a duration greater than τ^{temp}: it is a migration segment. It overlaps on the left of the time unit for a duration of at least Σ days. The user is therefore considered as being in migration during the time unit in the high-confidence estimate.

Diagram 2.F.21: High-confidence migration status: case 5



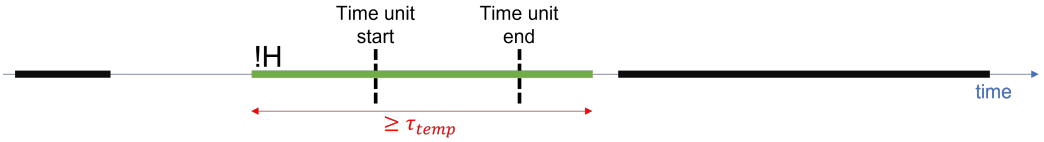
Note: This is the same configuration as in diagram 2.F.20, but the user is never observed before the green segment. The situation is otherwise equivalent to that of diagram 2.F.20.

Diagram 2.F.22: High-confidence migration status: case 6



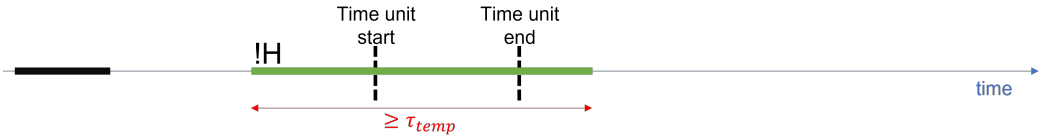
Note: This is the same configuration as in diagram 2.F.20, but the user exits the sample at the end of the green segment. The situation is otherwise equivalent to that of diagram 2.F.20.

Diagram 2.F.23: High-confidence migration status: case 7



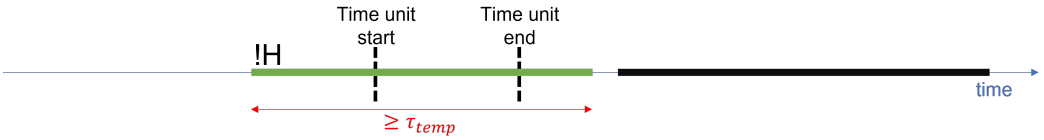
Note: The green segment is at a non-home location and has a duration greater than τ^{temp}: it is a migration segment. It covers the entire time unit so the user is considered as being in migration during the time unit in the high-confidence estimate. Note that Σ is necessarily set at a value that is lower than the duration of time units considered.

Diagram 2.F.24: High-confidence migration status: case 8



Note: This is the same configuration as in diagram 2.F.23, but the user exits the sample at the end of the green segment. The situation is otherwise equivalent to that of diagram 2.F.23.

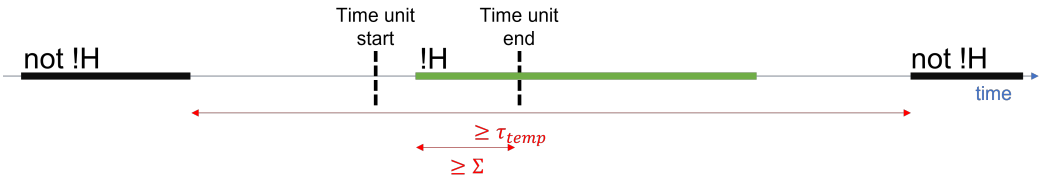
Diagram 2.F.25: High-confidence migration status: case 9



Note: This is the same configuration as in diagram 2.F.23, but the user is never observed before the green segment. The situation is otherwise equivalent to that of diagram 2.F.23.

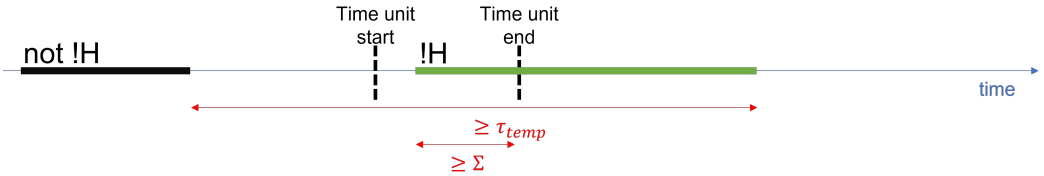
2.F.6 Identifying migration status: low-confidence

Diagram 2.F.26: Low-confidence migration status: case 1



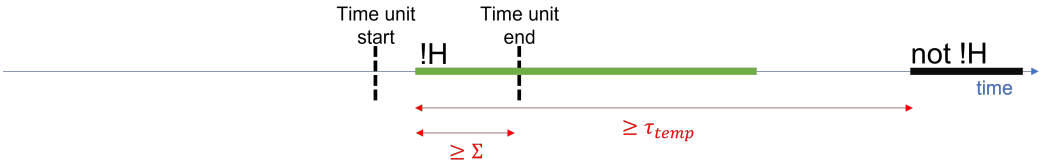
Note: The green segment at !H does not necessarily have an observed duration greater than τ^{temp}. The maximum duration is the time elapsed between the day following the previous segment and the date preceding the following segment. When this is greater than τ^{temp} and the green segment overlaps with the time unit on the right for at least Σ days, the user is considered as being in migration during the time unit in the low-confidence estimate.

Diagram 2.F.27: Low-confidence migration status: case 2



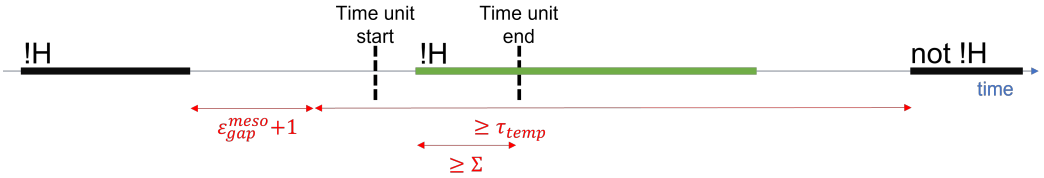
Note: This is the same configuration as in diagram 2.F.26, but the user exits the sample at the end of the green segment. In the absence of information about the user's location after the green segment, the maximum duration is limited to the time elapsed between the day following the previous segment and the observed end date. The situation is otherwise equivalent to that of diagram 2.F.26.

Diagram 2.F.28: Low-confidence migration status: case 3



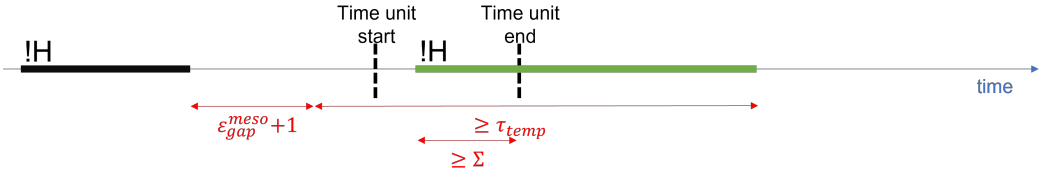
Note: This is the same configuration as in diagram 2.F.26, but the user is never observed before the green segment. In the absence of information about the user's location before the green segment, the maximum duration is limited to the time elapsed between the observed start date and the day preceding the following segment. The situation is otherwise equivalent to that of diagram 2.F.26.

Diagram 2.F.29: Low-confidence migration status: case 4



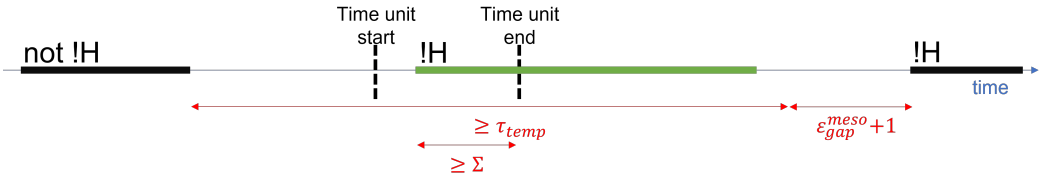
Note: This is the same configuration as in diagram 2.F.26, but the segment preceding the green segment is at the same location !H. In this case, the green segment cannot have started less than $\epsilon_{gap}^{meso}+1$ days after that segment. The maximum duration is then the time elapsed between the minimum start date of the green segment (i.e. $\epsilon_{gap}^{meso}+1$ days after the end of the preceding segment) and the day preceding the first day of the following segment. When this is larger than τ^{temp} and the green segment overlaps with the time unit on the right on at least Σ days, the user is considered as being in migration during the time unit in the low-confidence estimate.

Diagram 2.F.30: Low-confidence migration status: case 5



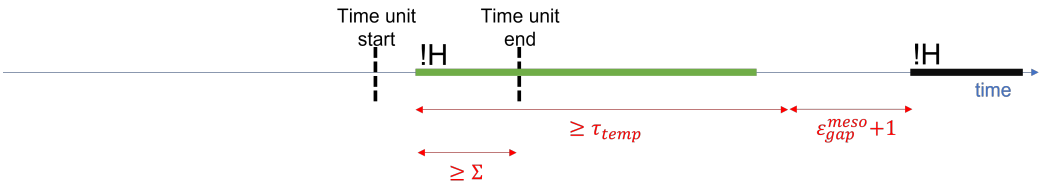
Note: This is the same configuration as in diagram 2.F.29, but the user exits the sample after the green segment. In the absence of information about the user's location after the end of the green segment, the maximum end date possible is considered to coincide with the observed end date. The maximum duration is then the time elapsed between the minimum start date of the green segment (i.e. $\epsilon_{gap}^{meso}+1$ days after the end of the preceding segment) and the observed end date. When this is larger than τ^{temp} and the green segment overlaps with the time unit on the right on at least Σ days, the user is considered as being in migration during the time unit in the low-confidence estimate.

Diagram 2.F.31: Low-confidence migration status: case 6



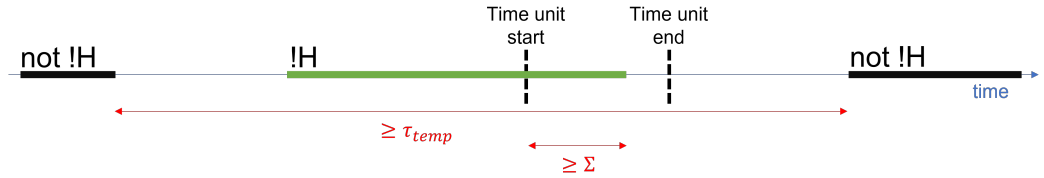
Note: This is the same configuration as in diagram 2.F.26, but the segment following the green segment is at the same location !H. In this case, the green segment cannot have ended less than $\epsilon_{gap}^{meso}+1$ days before that segment. The maximum duration is then the time elapsed between the day following the last day of the preceding segment and the maximum end date of the green segment (i.e. $\epsilon_{gap}^{meso}+1$ days before the start of the following segment). When this is larger than τ^{temp} and the green segment overlaps with the time unit on the right on at least Σ days, the user is considered as being in migration during the time unit in the low-confidence estimate.

Diagram 2.F.32: Low-confidence migration status: case 7



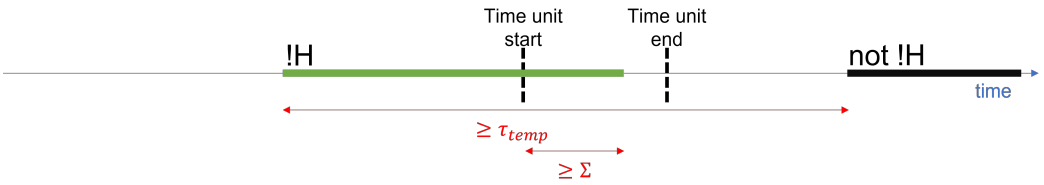
Note: This is the same configuration as in diagram 2.F.31, but the user is never seen before the green segment. In the absence of information about the user's location before the green segment started, the minimum start date possible is considered to coincide with the observed start date. The maximum duration is then the time elapsed between the observed start date and the maximum end date of the green segment (i.e. $\epsilon_{gap}^{meso} + 1$ days before the start of the following segment). When this is larger than τ^{temp} and the green segment overlaps with the time unit on the right on at least Σ days, the user is considered as being in migration during the time unit in the low-confidence estimate.

Diagram 2.F.33: Low-confidence migration status: case 8



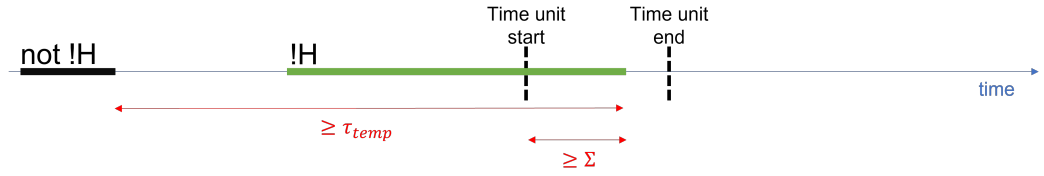
Note: This configuration is exactly equivalent to diagram 2.F.26 with the green segment overlapping on the left of the time unit.

Diagram 2.F.34: Low-confidence migration status: case 9



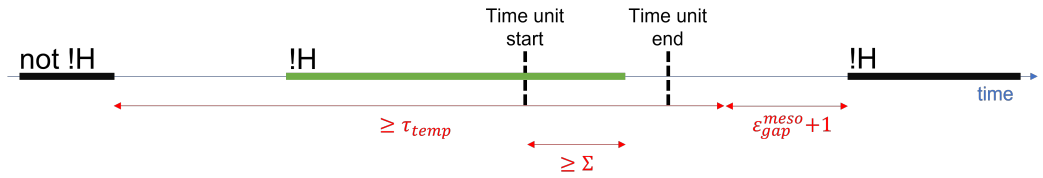
Note: This configuration is exactly equivalent to diagram 2.F.28 with the green segment overlapping on the left of the time unit.

Diagram 2.F.35: Low-confidence migration status: case 10



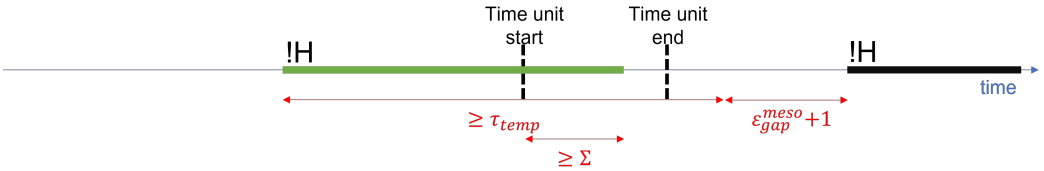
Note: This configuration is exactly equivalent to diagram 2.F.27 with the green segment overlapping on the left of the time unit.

Diagram 2.F.36: Low-confidence migration status: case 11



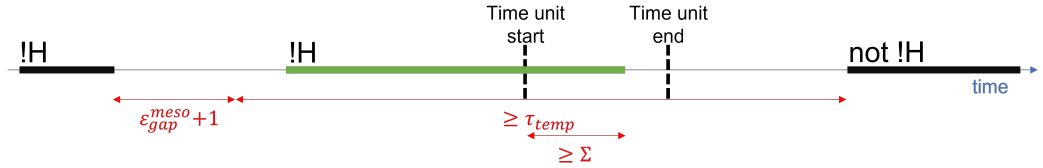
Note: This configuration is exactly equivalent to diagram 2.F.31 with the green segment overlapping on the left of the time unit.

Diagram 2.F.37: Low-confidence migration status: case 12



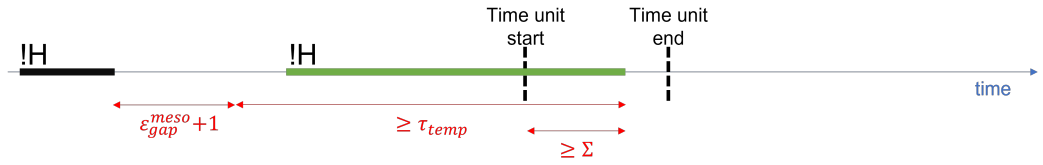
Note: This configuration is exactly equivalent to diagram 2.F.32 with the green segment overlapping on the left of the time unit.

Diagram 2.F.38: Low-confidence migration status: case 13



Note: This configuration is exactly equivalent to diagram 2.F.29 with the green segment overlapping on the left of the time unit.

Diagram 2.F.39: Low-confidence migration status: case 14

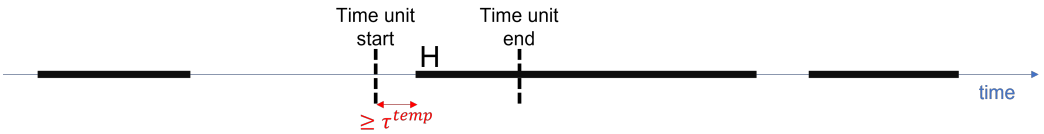


Note: This configuration is exactly equivalent to diagram 2.F.30 with the green segment overlapping on the left of the time unit.

Appendix 2.G Algorithmic rules to count the observation status of users by time unit

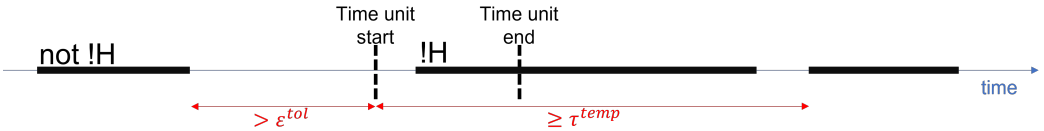
2.G.1 Identifying observation status for migration departure

Diagram 2.G.1: Observation status for migration departure: case 1



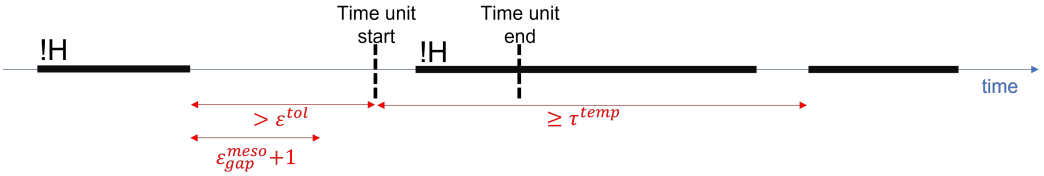
Note: An observation gap overlaps with the time unit on the left and is followed by a home segment. If the gap between the time unit start date and the home segment start date is larger than the parameter τ^{temp}, a migration segment may have started during the time unit without being observed. The user is considered as being not observed when calculating the number of migration departures during that time unit.

Diagram 2.G.2: Observation status for migration departure: case 2



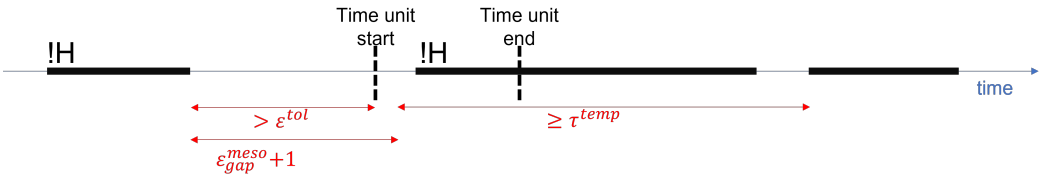
Note: An observation gap overlaps with the time unit on the left and is assumed to be strictly smaller than τ^{temp} – we are back to case 1 of diagram 2.G.1 otherwise. It is followed by a non-home segment at location !H, and preceded by a segment at a location that is not !H. The non-home segment may be a migration segment with a start date within the time unit if the time elapsed between the time unit start date and the day preceding the following segment is greater than τ^{temp}. When the tolerance criterion is exceeded, i.e. the time elapsed between the start of the observation gap and the day preceding the start of the time unit exceeds the parameter $\epsilon^t ol$, the uncertainty around the actual start date of the segment at the non-home location is considered as too large and the user is not counted as being observed for the calculation of migration departures during that time unit. Note that if the time elapsed between the time unit start date and the day preceding the following segment is strictly less than τ^{temp}, the non-home segment cannot be a migration segment and the user is then considered as observed and not having departed for migration during the time unit.

Diagram 2.G.3: Observation status for migration departure: case 3



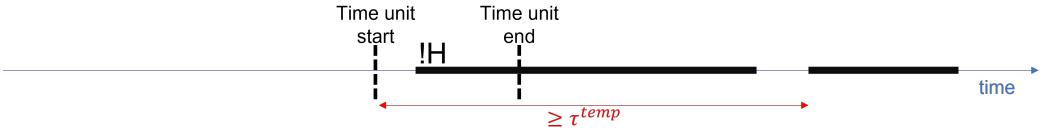
Note: This configuration is the same as in diagram 2.G.2, but the segment preceding the observation gap is at the same location !H as the segment following the gap. The observation gap is necessarily strictly larger than ϵ_{gap}^{meso}, otherwise both segments would be have been merged in the migration detection procedure. In this case, the non-home segment overlapping with the time unit cannot have started earlier than $\epsilon_{gap}^{meso} + 1$ days after the previous segment. When this minimum start date falls outside the time unit, the maximum possible duration of that non-home segment to still be considered as having potentially started during the time unit is the time elapsed between the time unit start date and the day preceding the first day of the following segment. When this is larger than τ^{temp}, the possibility exists that a migration event started during the time unit. However, if the tolerance criterion is exceeded, i.e. the time elapsed between the start of the observation gap and the day preceding the start of the time unit exceeds the parameter ϵ^{tol}, the uncertainty around the actual start date of the segment at the non-home location is considered as too large and the user is not counted as being observed for the calculation of migration departures during that time unit.

Diagram 2.G.4: Observation status for migration departure: case 4



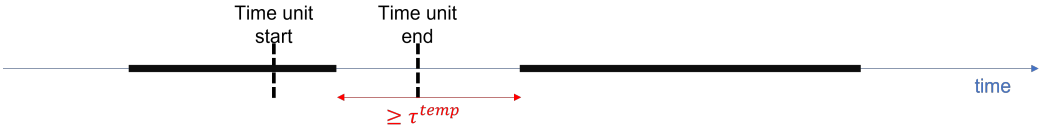
Note: This configuration is the same as in diagram 2.G.3 for the case when the date corresponding to $\epsilon_{gap}^{meso} + 1$ days after the previous segment falls within the time unit. The maximum possible duration of the non-home segment to still be considered as having potentially started during the time unit is now the time elapsed between this date – instead of the time unit start date– and the day preceding the first day of the following segment. The situation is then the same as in diagram 2.G.3.

Diagram 2.G.5: Observation status for migration departure: case 5



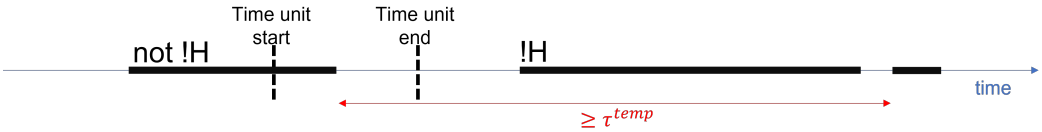
Note: This configuration is the same as in diagrams 2.G.2-2.G.4, but the user is never observed before the non-home segment overlapping with the time unit. In the absence of further information about the user's location, it is assumed that the minimum start date of the non-home segment to consider it started during the time unit coincides with the first day of the time unit. When the maximum duration, i.e the time elapsed between this minimum start date and the day preceding the following segment, is greater than τ^{temp}, the possibility exists that a migration occurred and started during the time unit. Since the user is not observed before the non-home segment, the uncertainty around the actual start date is virtually infinite and the user is not counted as being observed for the calculation of migration departures during that time unit.

Diagram 2.G.6: Observation status for migration departure: case 6



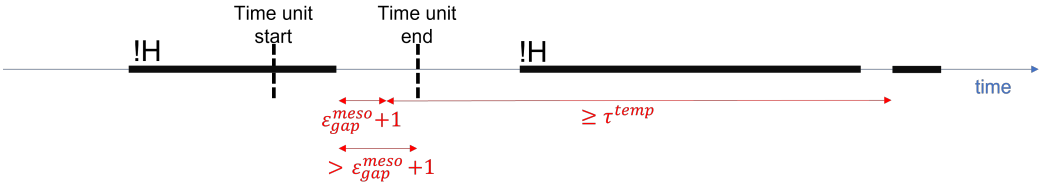
Note: An observation gap overlaps with the time unit on the right. Regardless of the locations of the preceding and following segments, if the gap is greater than τ^{temp} then a migration event could have started during the time unit without being observed. The user is thus considered as being not observed when calculating the number of migration departures during that time unit.

Diagram 2.G.7: Observation status for migration departure: case 7



Note: An observation gap overlaps with the time unit on the right and is assumed to be strictly smaller than τ^{temp} – we are back to case 6 of diagram 2.G.6 otherwise. It is followed by a non-home segment at location !H, and preceded by a segment at a location that is not !H. The non-home segment may be a migration segment with a start date within the time unit if the time elapsed between the observation gap start date and the day preceding the following segment is greater than τ^{temp}. In that case, it is impossible to determine whether the user effectively departed for migration during the time unit or not. The user is thus considered as being not observed when calculating the number of migration departures during that time unit. Conversely, if this maximum duration was strictly less than τ^{temp}, the observation gaps could not be concealing a migration event starting during the time unit and the user would be considered as effectively observed and not having departed for migration.

Diagram 2.G.8: Observation status for migration departure: case 8

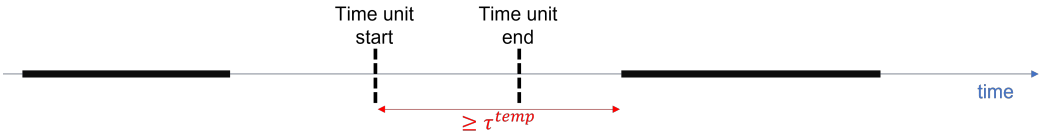


Note: This is the same configuration as in diagram 2.G.7, although the segment preceding the observation gap is at the same location !H as the segment following the gap. Two conditions allow for the possibility that the segment following the gap started during the time unit without being observed. First, the minimum start date has to fall within the time unit; the minimum start date is $\epsilon_{gap}^{meso} + 2$ days after the preceding segment. Second, the maximum duration, defined as the time elapsed between the minimum start date and the day preceding the following segment, has to be greater than τ^{temp}. When both conditions are met, the degree of uncertainty is such that the user is considered as being not observed when calculating the number of migration departures during that time unit.

Diagram 2.G.9: Observation status for migration departure: case 9

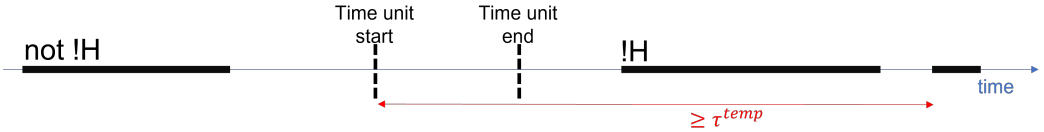
Note: An unbounded observation gap overlaps with the time unit on the right; the user exits the sample. The possibility exists that the user departed for migration during the unobserved period on the right of the time unit. Without further information on the user's location after that, the user is considered as being not observed when calculating the number of migration departures during that time unit.

Diagram 2.G.10: Observation status for migration departure: case 10



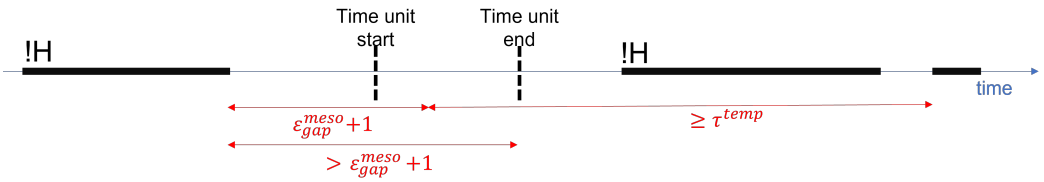
Note: An observation gap fully covers the time unit. Regardless of the locations of the preceding and following segments, if the time elapsed between the time unit start date and the end of the observation gap is greater than τ^{temp} then a migration event could have started during the time unit without being observed. The user is thus considered as being not observed when calculating the number of migration departures during that time unit.

Diagram 2.G.11: Observation status for migration departure: case 11



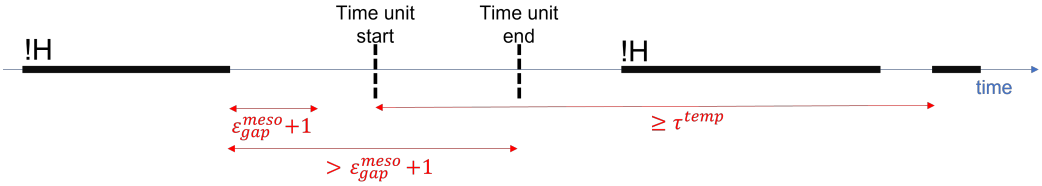
Note: An observation gap fully covers the time unit and the fraction of the gap that starts on the first day of the time unit is assumed strictly smaller than τ^{temp} – we are back to case 10 of diagram 2.G.10 otherwise. It is followed by a non-home segment at location !H, and preceded by a segment at a location that is not !H. The non-home segment may be a migration segment with a start date within the time unit if the time elapsed between the time unit start date and the day preceding the following segment is greater than τ^{temp}. In that case, it is impossible to determine whether the user effectively departed for migration during the time unit or not. The user is thus considered as being not observed when calculating the number of migration departures during that time unit. Conversely, if this maximum duration was strictly less than τ^{temp}, the observation gaps could not be concealing a migration event starting during the time unit and the user would be considered as effectively observed and not having departed for migration.

Diagram 2.G.12: Observation status for migration departure: case 12



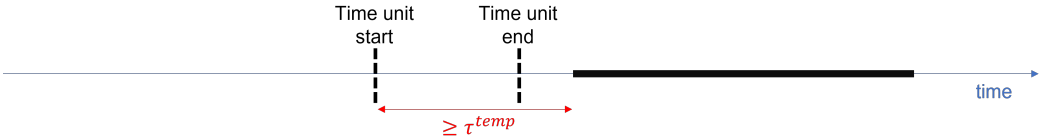
Note: This is the same configuration as in diagram 2.G.11, although the segment preceding the observation gap is at the same location !H as the segment following the gap. Two conditions allow for the possibility that the segment following the gap started during the time unit without being observed. First, the minimum start date cannot fall after the end of the time unit; the minimum start date is $\epsilon_{gap}^{meso} + 2$ days after the preceding segment. This diagram shows the case when the minimum start date falls within the time unit. Second, the maximum duration, defined as the time elapsed between the minimum start date and the day preceding the following segment, has to be greater than τ^{temp}. When both conditions are met, the degree of uncertainty is such that the user is considered as being not observed when calculating the number of migration departures during that time unit.

Diagram 2.G.13: Observation status for migration departure: case 13



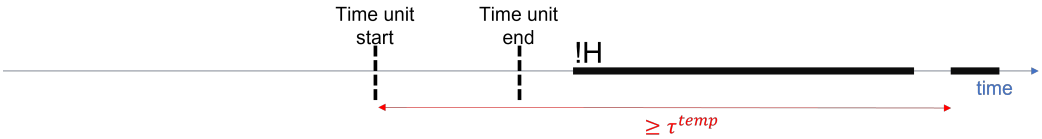
Note: This is the same configuration as in diagram 2.G.12 above, for the case where the minimum start strictly precedes the start of the time unit. In this case, the maximum duration that allows for the possibility that the non-home segment actually conceals a migration event that started during the time unit is the time elapsed between the time unit start date and the day preceding the following segment. Again, when this is greater than τ^{temp}, the user is considered as being not observed when calculating the number of migration departures during that time unit.

Diagram 2.G.14: Observation status for migration departure: case 14



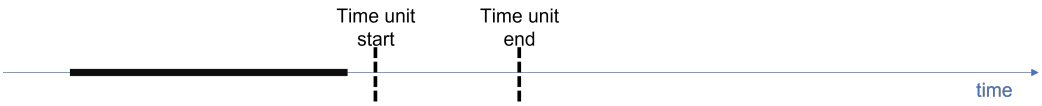
Note: This is the same configuration as in case 10 described in diagram 2.G.10, for the case where the user is never observed before the time unit. The criterion to classify the user as not being observed for migration departure for that time unit remains unchanged: the time elapsed between the time unit start date and the day preceding the first segment is greater than τ^{temp}.

Diagram 2.G.15: Observation status for migration departure: case 15



Note: This is the same configuration as in case 14 and where it is implicitly assumed that the time elapsed between the time unit start date and the day preceding the first segment is strictly less than τ^{temp}. The first segment is at a non-home location !H. When the time elapsed between the time unit start date and the day preceding the start date of the segment following the non-home segment is greater than τ^{temp}, the possibility exists that a migration departure effectively occurred during the time unit without being observed. The user is thus considered as being not observed when calculating the number of migration departures during that time unit.

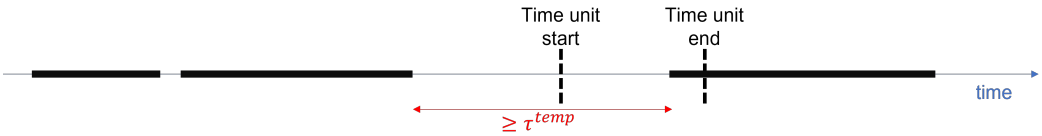
Diagram 2.G.16: Observation status for migration departure: case 16



Note: The configuration is the same as in case 9 described in diagram 2.G.9, but the observation gap fully covers the time unit. The criterion remains unchanged and in this situation, the user is considered as being not observed when calculating the number of migration departures during that time unit.

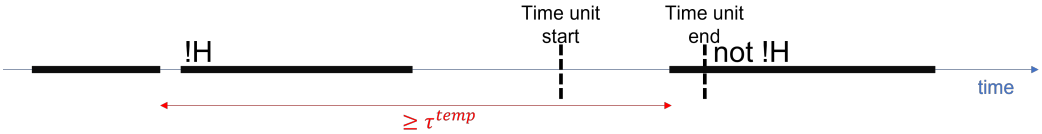
2.G.2 Identifying observation status for migration return

Diagram 2.G.17: Observation status for migration returns: case 1



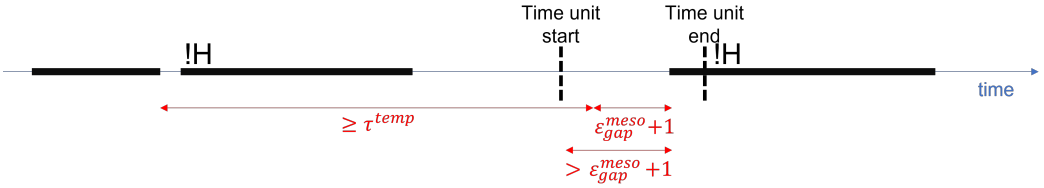
Note: An observation gap overlaps with the time unit on the left. Regardless of the locations of the preceding and following segments, if the gap is greater than τ^{temp} then a migration event could have ended during the time unit without being observed. The user is thus considered as being not observed when calculating the number of migration returns during that time unit.

Diagram 2.G.18: Observation status for migration returns: case 2



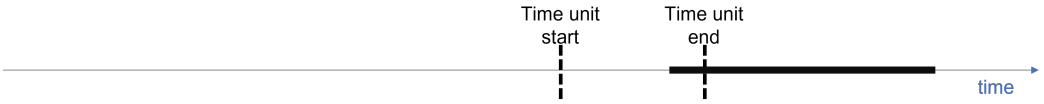
Note: An observation gap overlaps with the time unit on the left and is assumed to be strictly smaller than τ^{temp} – we are back to case 1 of diagram 2.G.17 otherwise. It is preceded by a non-home segment at location !H, and followed by a segment at a location that is not !H. The non-home segment may be a migration segment with an end date within the time unit if the time elapsed between the day following the preceding segment and the end of the observation gap is greater than τ^{temp}. In that case, it is impossible to determine whether the user effectively returned from migration during the time unit or not. The user is thus considered as being not observed when calculating the number of migration returns during that time unit. Conversely, if this maximum duration was strictly less than τ^{temp}, the observation gaps could not be concealing a migration event ending during the time unit and the user would be considered as effectively observed and not having returned from migration.

Diagram 2.G.19: Observation status for migration returns: case 3



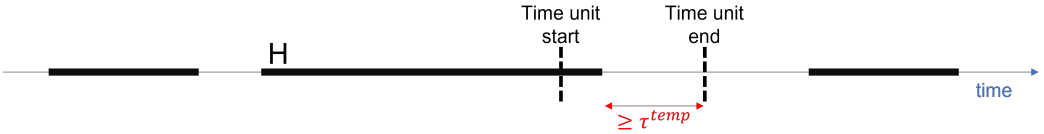
Note: This is the same configuration as in diagram 2.G.18, although the segment following the observation gap is at the same location !H as the segment preceding the gap. Two conditions allow for the possibility that the non-home segment preceding the gap reflects a migration event that ended during the time unit without being observed. First, the maximum end date has to fall within the time unit; the maximum end date is $\epsilon_{gap}^{meso} + 2$ days before the following segment. Second, the maximum duration, defined as the time elapsed between the day following the preceding segment and the maximum end date, has to be greater than τ^{temp}. When both conditions are met, the degree of uncertainty is such that the user is considered as being not observed when calculating the number of migration returns during that time unit.

Diagram 2.G.20: Observation status for migration returns: case 4



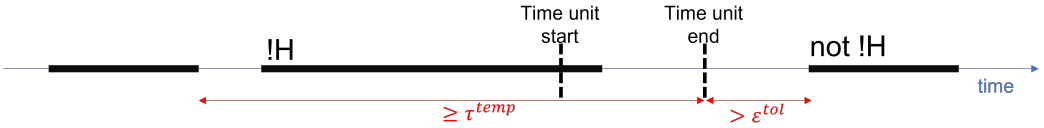
Note: An unbounded observation gap overlaps with the time unit on the left; the user enters the sample during the time unit. The possibility exists that the user returned from migration during the unobserved period on the left of the time unit. Without further information on the user's location before the first segment observed, the user is considered as being not observed when calculating the number of migration returns during that time unit.

Diagram 2.G.21: Observation status for migration returns: case 5



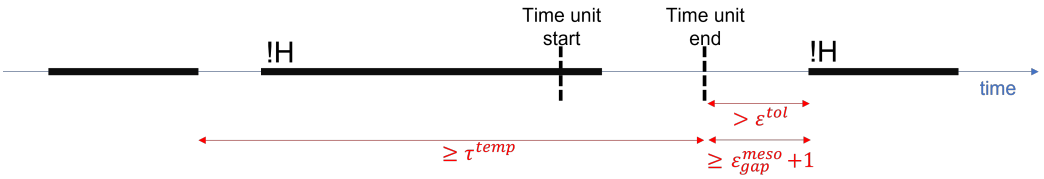
Note: An observation gap overlaps with the time unit on the right and is preceded by a home segment. If the gap between the observation gap start date and the time unit end date is larger than the parameter τ^{temp}, a migration segment may have occurred and ended during this portion of the time unit, without being observed. The user is considered as being not observed when calculating the number of migration returns during that time unit.

Diagram 2.G.22: Observation status for migration returns: case 6



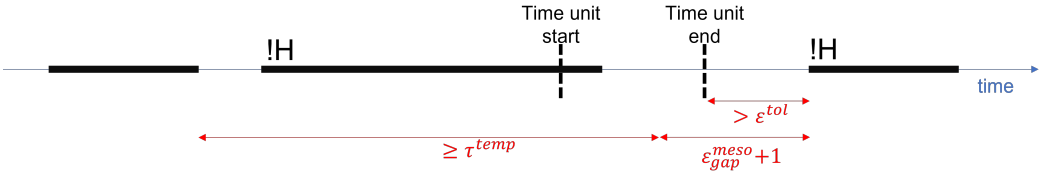
Note: An observation gap overlaps with the time unit on the right and the fraction of the gap ending at the end of the time unit is assumed strictly smaller than τ^{temp} – we are back to case 5 of diagram 2.G.21 otherwise. It is preceded by a non-home segment at location !H, and followed by a segment at a location that is not !H. The non-home segment may be a migration segment ending within the time unit if the time elapsed between the day following the preceding segment and the time unit end date is greater than τ^{temp}. When the tolerance criterion is exceeded, i.e. the time elapsed between the day following the time unit end date and the day preceding the first day of the following segment exceeds the parameter $\epsilon^t ol$, the uncertainty around the actual end date of the segment at the non-home location is considered as too large and the user is not counted as being observed for the calculation of migration returns during that time unit. Note that if the maximum duration considered is strictly less than τ^{temp}, the non-home segment cannot be a migration segment and the user is then considered as observed and not having returned from migration during the time unit.

Diagram 2.G.23: Observation status for migration returns: case 7



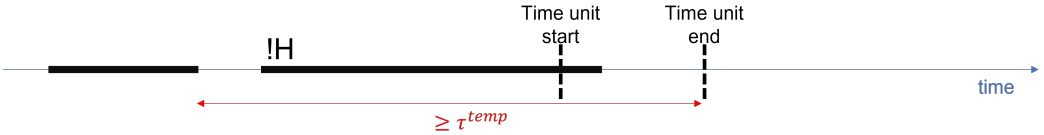
Note: This is the same configuration as in case 6 described in diagram 2.G.22, although the segment following the observation gap is at the same location !H as the segment preceding the gap. The non-home segment preceding the gap cannot have ended later than $\epsilon^{meso}_{gap} + 2$ days before the following non-home segment started – they would have been merged by the detection algorithm otherwise. This maximum end date falls either within or after the time unit. Case 7 deals with the latter configuration whereas case 8 in the following diagram (2.G.24) considers the former. In this case, the maximum end date possible for the first non-home segment to consider it ended during the time unit coincides with the time unit end date. Two conditions allow for the possibility that the non-home segment preceding the gap reflects a migration event that ended during the time unit without being observed. First, the maximum duration, i.e. the time elapsed between the day following the last day of the previous segment and the maximum end date, is greater than τ^{temp}. Second, the tolerance criterion is exceeded: the time elapsed between the maximum end date and the day preceding the first day of the following segment is strictly greater than ϵ^{tol}. When both conditions are met, the degree of uncertainty is such that the user is considered as being not observed when calculating the number of migration returns during that time unit.

Diagram 2.G.24: Observation status for migration returns: case 8



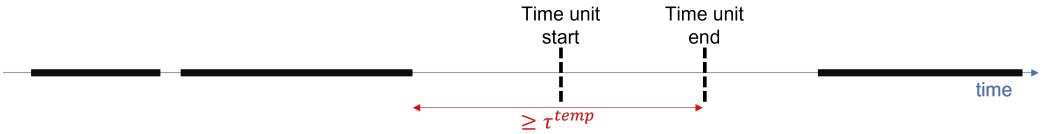
Note: This is the same configuration as in case 7 above, for the case where the maximum end date falls within the time unit. In this case the maximum duration is the time elapsed between the day following the last day of the previous segment and the maximum end date, which is strictly lower than the time unit end date. The criteria to define the user as not being observed are then equivalent.

Diagram 2.G.25: Observation status for migration returns: case 9



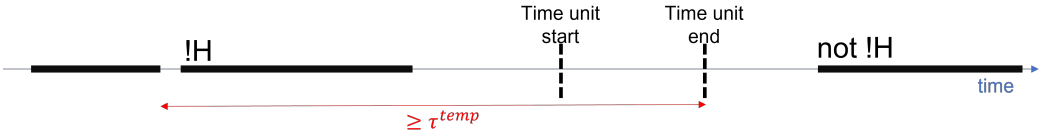
Note: This configuration is the same as in diagrams 2.G.22-2.G.24, but the user exits the sample after the non-home segment overlapping with the time unit. In the absence of further information about the user's location, it is assumed that the maximum end date of the non-home segment to consider it ended during the time unit coincides with the last day of the time unit. When the maximum duration, i.e the time elapsed between the day following the preceding segment and this maximum end date, is greater than τ^{temp}, the possibility exists that the observation gaps conceal a migration event that ended during the time unit. Since the user is not observed after the non-home segment, the uncertainty around the actual end date is virtually infinite (i.e. the tolerance criterion is exceeded) and the user is not counted as being observed for the calculation of migration returns during that time unit.

Diagram 2.G.26: Observation status for migration returns: case 10



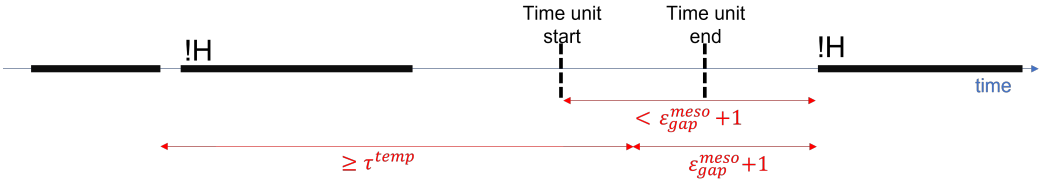
Note: An observation gap fully covers the time unit. Regardless of the locations of the preceding and following segments, if the time elapsed between the observation gap start date and the time unit end date is greater than τ^{temp}, then a migration event could have occurred and ended during the time unit without being observed. The user is thus considered as being not observed when calculating the number of migration returns during that time unit.

Diagram 2.G.27: Observation status for migration returns: case 11



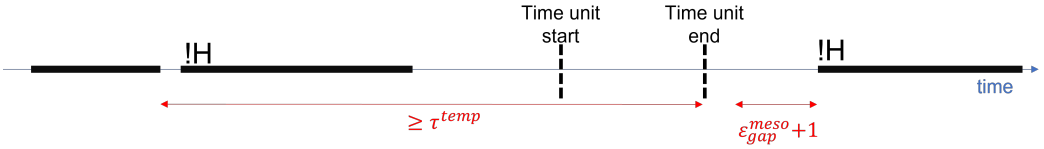
Note: An observation gap fully covers the time unit and the left-hand side portion of the gap that ends on the last day of the time unit is assumed strictly smaller than τ^{temp} – we are back to case 10 of diagram 2.G.26 otherwise. It is preceded by a non-home segment at location !H, and followed by a segment at a location that is not !H. The non-home segment may be a migration segment with an end date within the time unit if the time elapsed between the day following the preceding segment and the time unit end date is greater than τ^{temp}. In that case, it is impossible to determine whether the user effectively returned from migration during the time unit or not. The user is thus considered as being not observed when calculating the number of migration returns during that time unit. Conversely, if this maximum duration was strictly less than τ^{temp}, the observation gaps could not be concealing a migration event ending during the time unit and the user would be considered as effectively observed and not having returned from migration.

Diagram 2.G.28: Observation status for migration returns: case 12



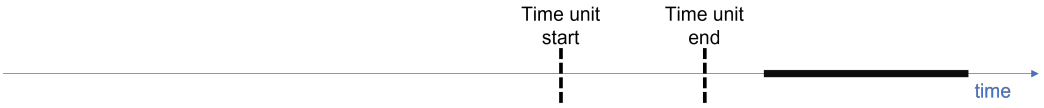
Note: This is the same configuration as in diagram 2.G.27, although the segment following the observation gap is at the same location !H as the segment preceding the gap. Two conditions allow for the possibility that the segment preceding the gap ended during the time unit without being observed. First, the maximum end date cannot fall before the start of the time unit; the maximum end date is $\epsilon_{gap}^{meso} + 2$ days before the following segment. In other words, and as showed on the diagram, the gap between the time unit start date and the end of the observation gap has to be strictly lower than $\epsilon_{gap}^{meso} + 1$ days. Then, the maximum end date falls either within or strictly after the time unit. The present case treats the former while case 13 below considers the latter. In this case, the maximum duration is defined as the time elapsed between the day following the preceding segment and the maximum end date. The second condition then implies that this duration has to be greater than τ^{temp}. When both conditions are met, the degree of uncertainty is such that the user is considered as being not observed when calculating the number of migration returns during that time unit.

Diagram 2.G.29: Observation status for migration returns: case 13



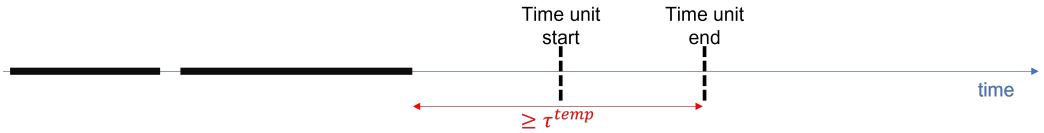
Note: This is the same configuration as in case 12 above, for the case where the maximum end date falls after the time unit. In this case the maximum duration considered is the time elapsed between the day following the last day of the previous segment and the last day of the time unit. The criteria to define the user as not being observed are then equivalent.

Diagram 2.G.30: Observation status for migration returns: case 14



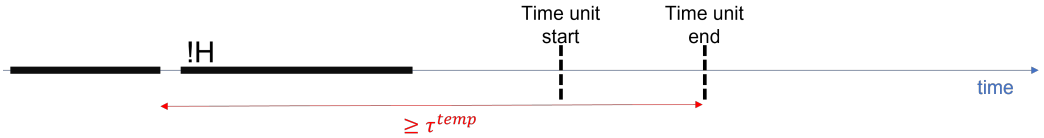
Note: An unbounded observation gap fully covers the time unit; the user is never observed before and during the time unit. This is the most straightforward case where a user is considered as being not observed for the calculation of migration returns during that time unit.

Diagram 2.G.31: Observation status for migration returns: case 15



Note: The user exits the sample before the time unit start date. Regardless of the last segment location, when the time elapsed between the day following the last day observed and the end of time unit is greater than τ^{temp}, the possibility exists that a migration event ending within the time unit occurred without being observed. The user is thus considered as being not observed for the calculation of migration returns during that time unit.

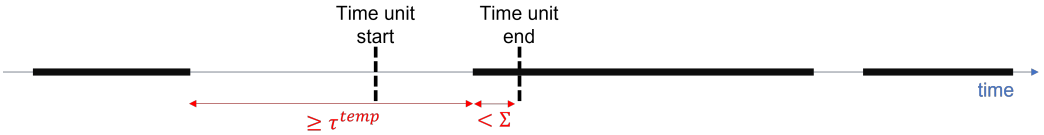
Diagram 2.G.32: Observation status for migration returns: case 16



Note: This is the same configuration as in case 15, but it is implicitly assumed that the time elapsed between the day following the last day observed and the end of time unit is strictly less than τ^{temp} – we are back to case 15 of diagram 2.G.31 otherwise. The last segment is at a non-home location !H. It may have ended during the time unit if the time elapsed between the day following the last day of the preceding segment and the end of the time unit is greater than τ^{temp}. Since the user is not observed after the non-home segment, the uncertainty around the actual end date is virtually infinite (i.e. the tolerance criterion is exceeded) and the user is not counted as being observed for the calculation of migration returns during that time unit.

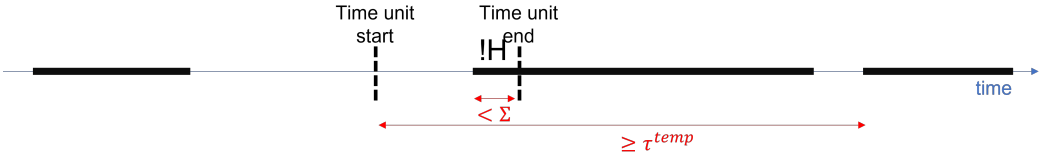
2.G.3 Identifying observation status for migration status

Diagram 2.G.33: Observation status for migration status: case 1



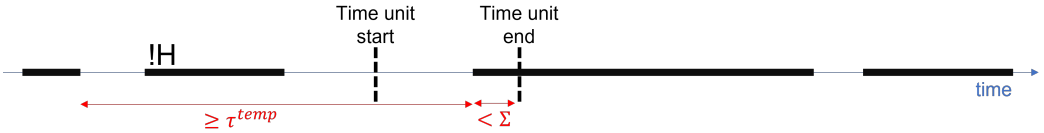
Note: An observation gap overlaps with the time unit on the left. The segment following the gap overlaps with the gap for a duration that is strictly lower than Σ days: the certainty criterion is not satisfied with respect to the only segment overlapping with the time unit. If the observation gap is greater than τ^{temp}, the possibility exists that a migration event overlapping with the time unit on at least Σ days occurred during the observation gap. The user is thus considered as being not observed for the calculation of the migration stock during that time unit. Note that when the observation gap is strictly less than τ^{temp}, the non-observation conditions depend on the characteristics of the preceding and following segments. The corresponding configurations are considered in the following diagrams.

Diagram 2.G.34: Observation status for migration status: case 2



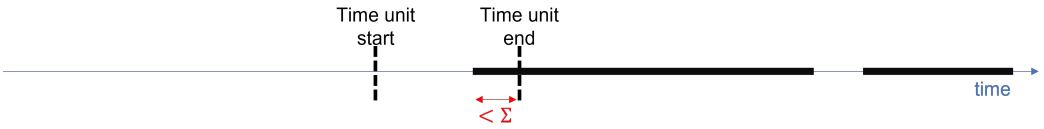
Note: This case is equivalent to case 1 described in diagram 2.G.33 but implicitly assumes that the observation gap is strictly less than τ^{temp} – otherwise we are back to case 1. The segment following the gap is at a non-home location !H and overlaps with the time unit for a duration strictly lower than Σ days. If the time elapsed between the first day of the time unit and the day preceding the first day of the following segment is greater than τ^{temp}, the possibility exists that the observed segment conceals a migration episode overlapping with the time unit on at least Σ days. Given this uncertainty, the user is considered as being not observed for the calculation of the migration stock during that time unit. Note that, conversely, if the duration considered is strictly lower than τ^{temp}, then we are certain that no migration event overlapping with the time unit occurred, and the user is considered as observed and not in migration.

Diagram 2.G.35: Observation status for migration status: case 3



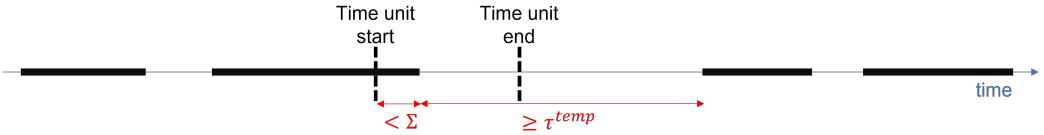
Note: This is the same configuration as in case 2 above, where the non-observation conditions with respect to the characteristics of the segment preceding the gap are considered. Similarly, the preceding segment is at a non-home location !H and the following segment still overlaps with the time unit for less than Σ days. If the time elapsed between the day following the last day of the preceding segment and the observation gap end date is greater than $\tau_t emp$, the possibility exists that the observed non-home segments conceals a migration episode overlapping with the time unit on at least Σ days. Given this uncertainty, the user is considered as being not observed for the calculation of the migration stock during that time unit.

Diagram 2.G.36: Observation status for migration status: case 4



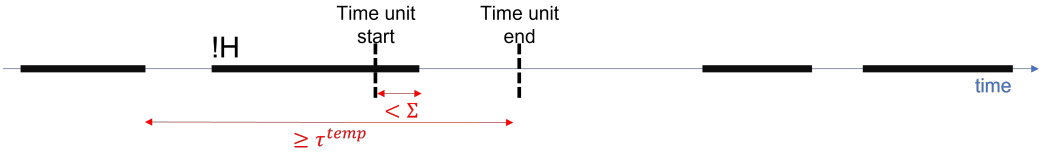
Note: The user enters the sample during the time unit. If the entry date is such that the first segment overlaps with the time unit on strictly less than Σ days, the possibility exists that a migration event overlapping with at least the first Σ days of the time unit occurred without being observed. The user is thus considered as being not observed for the calculation of the migration stock during that time unit.

Diagram 2.G.37: Observation status for migration status: case 5



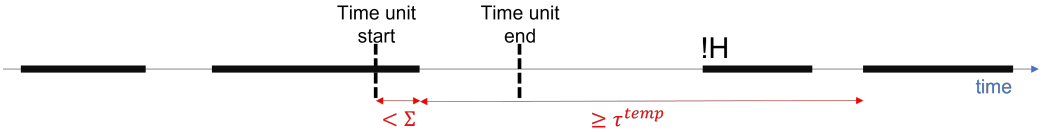
Note: An observation gap overlaps with the time unit on the right. The segment preceding the gap overlaps with the gap for a duration that is strictly lower than Σ days: the certainty criterion is not satisfied with respect to the only segment overlapping with the time unit. If the observation gap is greater than τ^{temp}, the possibility exists that a migration event overlapping with at least the last Σ days of the time unit occurred during the observation gap. The user is thus considered as being not observed for the calculation of the migration stock during that time unit. Note that when the observation gap is strictly less than τ^{temp}, the non-observation conditions depend on the characteristics of the preceding and following segments. The corresponding configurations are considered in the following diagrams.

Diagram 2.G.38: Observation status for migration status: case 6



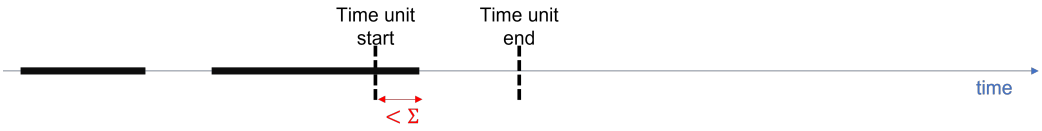
Note: This case is equivalent to case 5 described in diagram 2.G.37 but implicitly assumes that the observation gap is strictly less than τ^{temp} – otherwise we are back to case 5. The segment preceding the gap is at a non-home location !H and overlaps with the time unit for a duration strictly lower than Σ days. If the time elapsed between the day following the last day of the preceding segment and the time unit end date is greater than τ^{temp}, the possibility exists that the observed segment conceals a migration episode overlapping with the time unit on at least Σ days. Given this uncertainty, the user is considered as being not observed for the calculation of the migration stock during that time unit. Note that, conversely, if the duration considered is strictly lower than τ^{temp}, then we are certain that no migration event overlapping with the time unit occurred, and the user is considered as observed and not in migration.

Diagram 2.G.39: Observation status for migration status: case 7



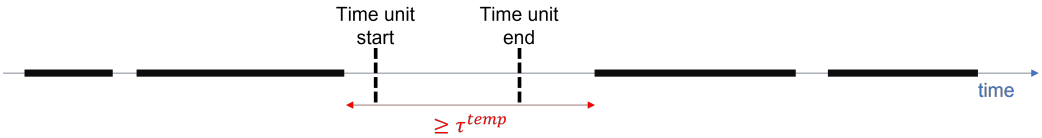
Note: This is the same configuration as in case 6 above, where the non-observation conditions with respect to the characteristics of the segment following the gap are considered. Similarly, the following segment is at a non-home location !H and the preceding segment still overlaps with the time unit for less than Σ days. If the time elapsed between the observation gap start date and the day preceding the first day of the following segment is greater than $\tau_t emp$, the possibility exists that the observed non-home segments conceals a migration episode overlapping with the time unit on at least Σ days. Given this uncertainty, the user is considered as being not observed for the calculation of the migration stock during that time unit.

Diagram 2.G.40: Observation status for migration status: case 8



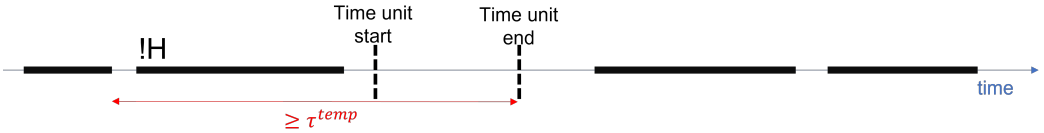
Note: The user exits the sample during the time unit. If the exit date is such that the last segment overlaps with the time unit on strictly less than Σ days, the possibility exists that a migration event overlapping with at least the first Σ days of the time unit occurred without being observed. The user is thus considered as being not observed for the calculation of the migration stock during that time unit.

Diagram 2.G.41: Observation status for migration status: case 9



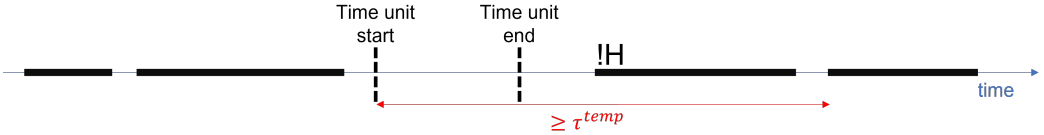
Note: An observation gap fully covers the time unit. When this gap is larger than τ^{temp} days, the possibility exists that a migration event overlapping with the time unit on at least Σ days occurred during the observation gap. The user is thus considered as being not observed for the calculation of the migration stock during that time unit. Note that when the observation gap is strictly less than τ^{temp}, the non-observation conditions depend on the characteristics of the preceding and following segments. The corresponding configurations are considered in the following diagrams.

Diagram 2.G.42: Observation status for migration status: case 10



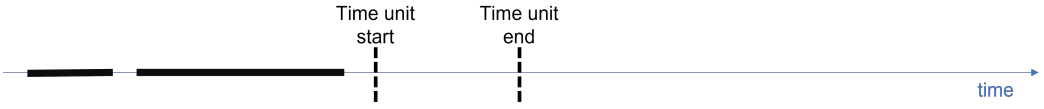
Note: This case is equivalent to case 9 described in diagram 2.G.41 but implicitly assumes that the observation gap is strictly less than τ^{temp} – otherwise we are back to case 9. The segment preceding the gap is at a non-home location !H – and does not overlap with the time unit. If the time elapsed between the day following the last day of the preceding segment and the time unit end date is greater than τ^{temp}, the possibility exists that the observed segment conceals a migration episode overlapping with the time unit on at least Σ days. Given this uncertainty, the user is considered as being not observed for the calculation of the migration stock during that time unit. Note that, conversely, if the duration considered is strictly lower than τ^{temp}, then we are certain that no migration event overlapping with the time unit occurred, and the user is considered as observed and not in migration.

Diagram 2.G.43: Observation status for migration status: case 11



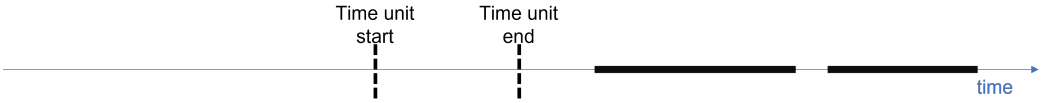
Note: This is the same configuration as in case 10 above, where the non-observation conditions with respect to the characteristics of the segment following the gap are considered. Similarly, the following segment is at a non-home location !H and the preceding segment does not overlap with the time unit. If the time elapsed between the time unit start date and the day preceding the first day of the following segment is greater than $\tau_t emp$, the possibility exists that the observed non-home segments conceals a migration episode overlapping with the time unit on at least Σ days. Given this uncertainty, the user is considered as being not observed for the calculation of the migration stock during that time unit.

Diagram 2.G.44: Observation status for migration status: case 12



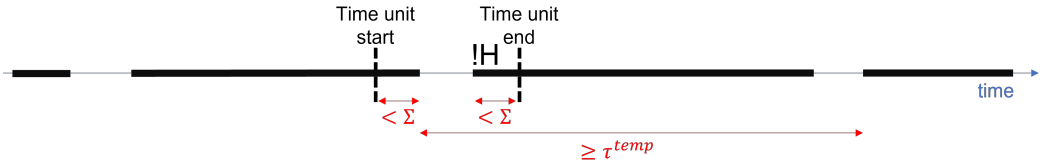
Note: The user exits the sample before the time unit. There is complete uncertainty about the user's migration status for that time unit and he is considered as no observed.

Diagram 2.G.45: Observation status for migration status: case 13



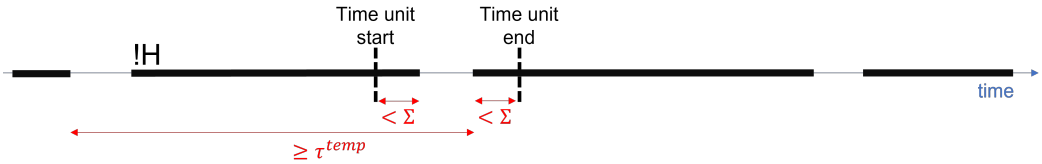
Note: The user enters the sample only after the time unit. There is complete uncertainty about the user's migration status for that time unit and he is considered as no observed.

Diagram 2.G.46: Observation status for migration status: case 14



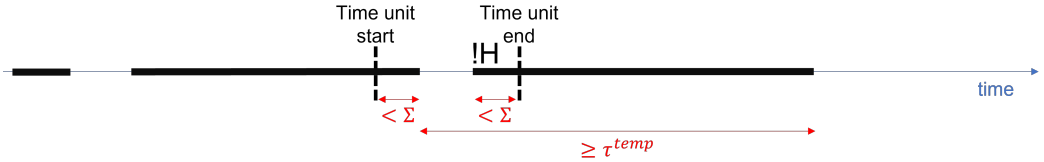
Note: The observation gap is strictly within the time unit. Both the preceding and following segments overlap with the time unit on strictly less than Σ days. In this case, the following segment is at a non-home location !H. If the maximum duration of that segment is greater than τ^{temp}, the possibility exists that it conceals a migration event overlapping with the time unit on at least Σ days. Given this uncertainty, the user is considered as being not observed for the calculation of the migration stock during that time unit.

Diagram 2.G.47: Observation status for migration status: case 15



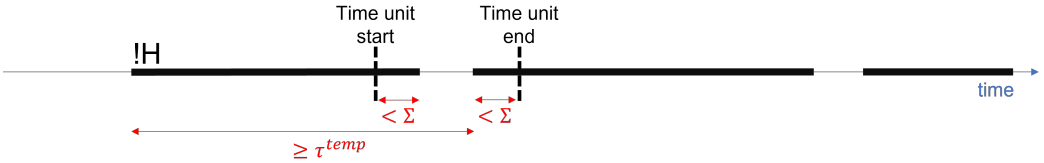
Note: The observation gap is strictly within the time unit. Both the preceding and following segments overlap with the time unit on strictly less than Σ days. In this case, the preceding segment is at a non-home location !H. If the maximum duration of that segment is greater than τ^{temp}, the possibility exists that it conceals a migration event overlapping with the time unit on at least Σ days. Given this uncertainty, the user is considered as being not observed for the calculation of the migration stock during that time unit..

Diagram 2.G.48: Observation status for migration status: case 16



Note: The observation gap is strictly within the time unit. Both the preceding and following segments overlap with the time unit on strictly less than Σ days. In this case, the following segment is at a non-home location !H and the user exits the sample at the end of this segment. Compared with case 14, the maximum end date is therefore considered to coincide with the segment end date and the maximum duration is shorter. Similarly, if it is greater than τ^{temp}, the possibility exists that it conceals a migration event overlapping with the time unit on at least Σ days. Given this uncertainty, the user is considered as being not observed for the calculation of the migration stock during that time unit.

Diagram 2.G.49: Observation status for migration status: case 17



Note: The observation gap is strictly within the time unit. Both the preceding and following segments overlap with the time unit on strictly less than Σ days. In this case, the preceding segment is at a non-home location !H and the user actually enters the sample on the start date of this segment. Compared with case 15, the minimum start date is therefore considered to coincide with the segment start date and the maximum duration is shorter. If the maximum duration of that segment is greater than τ^{temp}, the possibility exists that it conceals a migration event overlapping with the time unit on at least Σ days. Given this uncertainty, the user is considered as being not observed for the calculation of the migration stock during that time unit..

Appendix 2.H Senegal temporary migration profile: additional material

Table 2.H.1: Migration statistics at the national level derived from the unweighted sample, 2013.

	Migration events	**Migrants**	**Migration rate**
≥ *20 days*	782,296	487,300	24.4%
≥ *30 days*	527,382	381,031	19.1%
≥ *60 days*	217,501	199,237	10.0%
≥ *90 days*	91,729	90,910	4.6%

Note: Numbers showed in the table are based on the raw high-quality subset for the year 2013.

Table 2.H.2: Comparison of low- and high-confidence migration estimates at the national-level, high-quality subset.

	Migration events		**Migrants**		**Migration rate**	
	High-conf.	**Low-conf.**	**High-conf.**	**Low-conf.**	**High-conf.**	**Low-conf.**
≥ *20 days*	4,276,706	4,406,711	2,568,976	2,607,922	32.6%	33.1%
≥ *30 days*	2,874,507	2,929,497	2,037,406	2,058,398	25.8%	26.1%
≥ *60 days*	1,200,775	1,218,813	1,092,802	1,105,612	13.9%	14.0%
≥ *90 days*	528,388	534,849	520,205	526,268	6.6%	6.7%

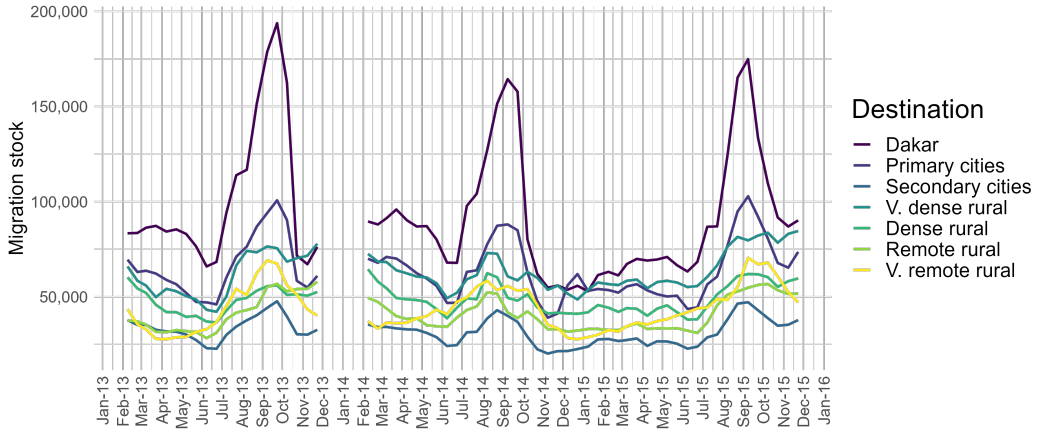
Note: The table shows aggregated statistics at the national-level for the year 2013 based on the 2013 high-quality weighted sample. Two estimates are compared for each migration metric: one considers all high-confidence migration events and the other is based on the low-confidence migration events. High-confidence events have an observed duration greater than the minimum duration considered (i.e. 20, 30, 60 or 90 days) whereas low-confidence events only have their maximum duration exceeding the minimum duration threshold.

Table 2.H.3: Comparison of low- and high-confidence migration estimates at the national-level, low-quality subset.

	Migration events		**Migrants**		**Migration rate**	
	High-conf.	**Low-conf.**	**High-conf.**	**Low-conf.**	**High-conf.**	**Low-conf.**
≥ *20 days*	4,270,546	5,178,167	2,612,010	2,847,314	33.1%	36.1%
≥ *30 days*	2,826,887	3,189,877	2,047,743	2,172,571	26.0%	27.5%
≥ *60 days*	1,101,779	1,178,350	1,020,863	1,079,132	12.9%	13.7%
≥ *90 days*	456,429	481,924	452,060	476,911	5.7%	6.0%

Note: The table shows aggregated statistics at the national-level for the year 2013 based on the 2013 low-quality weighted sample. Two estimates are compared for each migration metric: one considers all high-confidence migration events and the other is based on the low-confidence migration events. High-confidence events have an observed duration greater than the minimum duration considered (i.e. 20, 30, 60 or 90 days) whereas low-confidence events only have their maximum duration exceeding the minimum duration threshold.

Figure 2.H.1: Rural-out migration stock disaggregated by destination zone, 2013-2015.



Chapter 3

Temporary migration response to climate variability: New Evidence From Three Years of Mobile Phone Data

Joint with Virginie Comblon, Flore Gubert, Erwan Le Quentrec, Anne-Sophie Robilliard and Stefania Rubrichi.

Contribution

My contribution to this study includes conceptualization, data processing, dataset construction, performing the analyses and writing the manuscript.

3.1 Introduction

Within the Sahel region, rural livelihoods are largely dominated by agropastoral activities that are profoundly reliant on rainfall. Specifically, in the context of Senegal, which is the focus of this chapter, a significant 74% of rural households engage in some form of agriculture. Of this segment, 86% rely on rainfed agriculture, while a mere 5% have access to irrigation. As a result, rural households face significant year-on-year volatility in production and income. To navigate these uncertainties, rural households implement a range of coping mechanisms. These strategies encompass the adoption of modern agricultural technologies (Dercon and Christiaensen, 2011), consumption smoothing via savings and credit markets (Basu and Wong, 2015), the use of buffer stocks (Fafchamps, Udry, et al., 1998; Chaudhuri and Paxson, 2021) or the reliance on risk-sharing networks (Fafchamps and Gubert, 2007; Morten, 2019). Alternatively, individuals might opt to offer their labor in different markets by engaging in temporary migration strategies. Yet, this phenomenon has received less scholarly attention, largely due to the scarcity of data on such short-term movements.

Over the past two decades, Senegal has faced four major drought episodes, including three that occurred after 2010 (specifically in 2011, 2014, 2018). Taken together, these recent events have reportedly impacted over 1.8 million individuals.[1] Yet, it remains ambiguous whether, and to what extent, the income disruptions stemming from these climatic shocks influenced individuals' choices to temporarily relocate. Temporary migration may be used as an ex post response to compensate for income losses caused by a negative shock, effectively acting as a push factor. Nonetheless, migration comes with its inherent costs, and these income setbacks can also inhibit migration by intensifying liquidity constraints. In parallel, various frictions and market imperfections could hinder an efficient short-term reallocation of labor across locations, particularly from areas impacted by the shock to those that remain unaffected. On the other hand, temporary migration might be proactively incorporated ex ante within livelihood strategies, as a means to diversify income sources and mitigate potential risks. Under this perspective, adverse shocks would not directly influence migration choices, given that such movements are conceived precisely as a long-term strategy to manage income fluctuations. Lastly, adverse climatic events might diminish the temporary migration of individuals who are not directly impacted, due to decreased productivity in potential destination locations.

In other contexts, empirical research has mostly pointed towards a positive relationship between environmental shocks and migration. More specifically, slow onset weather events such as drought conditions or heat waves have been found to significantly increase migration movements in Asia and South America (Gray and Mueller, 2012b; Dallmann and Millock, 2016; Thiede et al., 2016; Call et al., 2017), but also in African contexts (Henry et al., 2004; Gray and Mueller, 2012a; Mastrorillo et al., 2016; Mueller et al., 2020). However, only a handful of quantitative studies have focused on drought-induced migrations in the Sahel. Findley (1994) showed that the 1980's drought in Mali caused a doubling of short-term internal migrations, with a more pronounced effect for the poorest households. More recently, Defrance et al. (2023) combined census data and high-resolution drought indices and found that drought conditions were associated with positive net flows of rural-to-urban migrants over the 1987-2009 period in Mali. Similarly, Henry et al. (2004) show that men are more likely to migrate domestically in response to drought conditions in Burkina Faso.

Empirical findings nonetheless reveal that the climate-migration relation is complex, multidimensional, and inherently context-specific and therefore at odds with the monolithic narrative positing that climate variability unambiguously leads to more migration for the most vulnerable. Weather shocks typically decrease the utility of staying for affected households – for instance, through crop failure or land

[1] EM-DAT, CRED / UCLouvain, Brussels, Belgium – www.emdat.be

degradation – and effectively create an incentive to use temporary migration as a coping strategy. But they also strengthen liquidity constraints to a point where investing in migration strategies may no longer be a feasible option. Some studies thus showed that climatic events can hamper migration in specific contexts and for particular population sub-groups (e.g. women, poorest) (Gray and Mueller, 2012a; Henry et al., 2004; Hirvonen, 2016; Mueller et al., 2020). Moreover, most longitudinal studies shed light on the heterogeneity of the effect of climatic events on migration with respect to household and individual characteristics such as wealth, land ownership, gender, but also across countries.

A major impediment to advancing the understanding of short-term migration responses to climate variability in sub-Saharan Africa has been the lack of detailed and reliable data. First, standard survey instruments imply measurement errors such as recall bias and attrition. More importantly, they are limited in their ability to capture finer mobility patterns such as seasonal, circular, temporary or high-frequency migrations that nonetheless often prevail (Coffey et al., 2015). In this respect, exploiting mobile phone data has emerged as a promising alternative to study these subtler human movements in developing contexts (Blumenstock, 2012; Hong et al., 2019; Lai et al., 2019; Demissie et al., 2019). However, a limited number of studies have used mobile phone data sources to track environmentally-induced human mobility. Lu, Wrathall, et al. (2016) focus on mobility patterns in the aftermath of 2013 Cyclone Mahasen in Bangladesh while Lu, Bengtsson, et al. (2012) unveiled population movements after the 2010 earthquake in Haiti. No empirical research using mobile phone data has been conducted for the study of human mobility triggered by slow onset events, or even more broadly to investigate environmentally-induced migrations in sub-Saharan Africa. As the penetration rate of mobile phone subscriptions has significantly progressed since the 2010's with coverage beyond 100% in most Sahelian countries[2], so has the relevance of using mobile phone data for the study of internal migrations in the region.

In this study, we delve into the influence of rainy season conditions on temporary migration decisions in Senegal. Harnessing three years of mobile phone metadata, we capture temporary migration movements at a national scale with a degree of spatial and temporal detail unattainable with conventional datasets. This enables us to precisely characterize the timing, duration and direction of temporary migration flows across a set of over 900 locations, which is roughly equivalent to twice the number of administrative divisions at the highest level in Senegal.[3] We use the migration estimates produced in chapter 2, which provides a rich benchmark

[2]Source: World Development Indicators, The World Bank.

[3]There are 433 communes in Senegal (fourth administrative level).

description of temporary migration patterns in Senegal. Temporary migration is ubiquitous in the country: our data allow to identify over 4 million events estimated in 2013, involving a third of the adult population. Two thirds of migration flows originate from rural areas and are primarily directed to cities. Rural-to-rural movements still represents a significant fraction of the total flow and are mostly local. Moreover, the unique temporal granularity of the data allows to unveil clear seasonal patterns: the stock of temporary migrants typically doubles between June and September and sharply decreases in October.

We develop a simple temporary migration model in which a single sector produces using labor and precipitations. Production functions are location-specific which allows to account for spatial differences in the sensitivity of local economies to rainfall conditions. Location choices are modelled within a nested logit structure in which individuals have home bias preferences. In this model, rainy season conditions are viewed as a local annual agricultural productivity draw that determines the trajectory of local wages over the agricultural year. Differences in the spatio-temporal distribution of rainfall across locations in distinct years is expected to induce some level of labor reallocation to places that are temporarily relatively more productive. Within-year reallocation dynamics are also possible given the differences in the agricultural calendar and economic structure that exist across locations. The temporal granularity of our data ultimately facilitates a meticulous exploration of these dynamics. We derive a simplified version of the model that provides some intuitions on the dynamics of short-term labor reallocation with climate variability, and lays the foundation for the ensuing empirical analysis.

We combine a granular phone-derived pseudo-panel of temporary migration estimates with satellite-based measures of the quality of rainy seasons. The dataset covers the period 2013-2015, during which a succession of good and bad rainy seasons is observed. In a first approach, we aggregate migration across destinations for each origin location and we estimate a conventional migration regression. Namely, we focus on the identification of the effect of rainy season conditions at origin on the probability of out-migration to any destination during the agricultural year. In essence, this is comparable to other identification exercises that have been conducted on this topic using survey data that usually do not include information on migration destinations. We find that relatively poorer conditions immediately lead to lower temporary migration in relative terms during the harvest season (October-November), and that the effect persists until the month of April the following year.

As underscored in Borusyak, Dix-Carneiro, et al. (2022), this kind of estimation suffers from an omitted-variable bias, especially when conditions in potential destinations correlate with those in the origin. Our observations affirm this phe-

nomenon in our context. Consequently, we estimate dyadic regressions, focusing on bilateral temporary migration rates across origin-destination pairs. This enables us to incorporate the role of rainfall conditions at destination in shaping migration decisions. With this identification strategy, and consistent with our first approach, we find a positive effect of precipitations on temporary migration during harvest. Yet, factoring in conditions at the destination reshapes our understanding of the effect of rainy season conditions at the origin on migration patterns during the following off-season (January-June). Relatively poorer conditions at origin lead to a higher propensity to temporarily migrate during that period. On the other hand, conditions at destination are found to be positively linked to their level of attractiveness. Finally, a heterogeneity analysis reveals that the observed negative effect of precipitations at origin on temporary migration during the off-season is more pronounced in locations with lower standards of living, suggesting that poorer households are more likely to resort to temporary migration as a coping strategy in times of hardship.

This study is primarily connected to a large strand of literature that has investigated the mechanisms by which individuals cope with high income variability in developing countries (Basu and Wong, 2015; Chaudhuri and Paxson, 2021; Dercon and Christiaensen, 2011; Fafchamps, Udry, et al., 1998; Fafchamps and Gubert, 2007; Morten, 2019; Townsend, 1994; Udry, 1994). It adds to this literature by studying temporary migration in a West African country characterized by a high vulnerability to frequent rainfall shocks. It further contributes to the climate migration literature, which focuses on the influence of environmental factors on migration decisions. Most studies have offered important insights into the long-term migration responses to weather anomalies in various developing contexts. Notable examples include Defrance et al. (2023) who use multiple census waves in Mali to investigate permanent migration responses to drought conditions, Dallmann and Millock (2016) who conduct a similar study in India, or Marchiori et al. (2012) who study the impact of weather anomalies on rural-urban migration in sub-Saharan Africa. This research augments the existing body of work by shedding light on the immediate temporary migration reactions to climate shocks, a temporal scale that has often been neglected due to the challenges associated with tracking such movements. Further, it complements a set of studies that investigated the benefits and barriers to temporary migration in the face of income seasonality (Bryan, Chowdhury, et al., 2014; Imbert and Papp, 2020a). Finally, it enriches a collection of studies, including Lu, Bengtsson, et al. (2012) and Lu, Wrathall, et al. (2016), which have combined mobile phone data with environmental signals to advance our understanding of mobility responses to environmental shocks. To the best of

our knowledge, this is the first study to exploit mobile phone data to identify the impact of slow onset events on temporary migration decisions within a developing setting.

In section 3.2, we present a simple model of temporary migration with climate variability, understood empirically as inter-year variations in the quality of rainy seasons. In section 3.3, we describe the sample of mobile phone data and we summarize the method used to produce measures of temporary migration in Senegal for the period 2013-2015. Section 3.4 shows the results of the empirical analysis on the effect of rainy season conditions on temporary migration and section 3.5 concludes.

3.2 Simple model of temporary migration with climate variability

3.2.1 Framework

The economy is comprised of N individuals that inelastically provide one unit of labor in each time period t. At the beginning of each period t, each individual i chooses a location $m_{i,t} \in \mathcal{M} = \{m_1, ..., m_M\}$ where he provides labor for a wage $w_{m,t}$. The home location of individual i is denoted by $h_i \in \mathcal{M}$, and is fixed throughout the entire period of study. This means that we consider any choice $m_{i,t} \neq h_i$ as a temporary relocation for individual i that does not affect the definition of his usual place of residence h_i.

3.2.2 Production function

We assume a one-sector economy in which firms produce using natural capital available at time t, $K_{m,t}$, and labor $N_{m,t}$ at each location m. In our context, $K_{m,t}$ represents the amount of accumulated precipitations since the start of the rainy season and is provided exogenously at no cost. The production process is equivalent to a representative firm producing according to a Constant Returns to Scale (CRS) Cobb-Douglas production function:

$$Q_{m,t} = \Gamma_{m,s(t)} K_{m,t}^{\alpha_m} N_{m,t}^{1-\alpha_m} \tag{3.1}$$

$s(t)$ is a season index that represents the time of the year at time t. For instance, when t represents periods of time corresponding to half-months, $s(t)$ can take on 24 values to denote the first half of January, the second half of January, the first half of February, and so on. $\Gamma_{m,s(t)}$ is thus a location- and season-specific TFP that allows to capture differences in the temporal trajectories of local productivity across locations. In other words, it captures seasonal patterns in the production and allows

for those to differ across locations. For instance, rural locations that primarily rely on rainfed agriculture will be productive from the start of the rainy season (June) to the harvest season (October-December); the demand for agricultural labor will be relatively high. On the other hand, urban locations generally experience lower levels of seasonality and $\Gamma_{m,s(t)}$ is expected to exhibit lower variations over a typical year.

3.2.3 Labor market

We assume that the labor market is perfectly competitive and for each period t and location m, the representative firm observes $K_{m,t}$ and chooses $N_{m,t}$ that maximizes profits. This allows to determine an expression of $w_{m,t}$ that defines the inverse labor demand curve:

$$w_{m,t} = (1 - \alpha_m)\Gamma_{m,s(t)} \left(\frac{K_{m,t}}{N_{m,t}} \right)^{\alpha_m} \tag{3.2}$$

The inverse labor demand elasticity from the labor demand curve is thus given by:

$$\frac{\partial \ln w_{m,t}}{\partial \ln N_{m,t}} = -\alpha_m \tag{3.3}$$

Since $-\alpha_m < 0$, the labor market is a source of congestion: an increase in labor is associated with lower wages, which in turn decreases the value of location m.

Similarly, the elasticity of wages with respect to precipitations $K_{m,t}$ is $\alpha_m > 0$, so that higher precipitations lead to higher wages that make locations relatively more attractive.

3.2.4 Utility maximization

Individuals' preferences are Cobb-Douglas over a tradable good c_1 and housing c_2. Utility also depends on the level of amenities $A_{m'}$ at the chosen location m', and we assume iceberg migrations costs $\tau_{m,m'}$ for an individual that moved from m to m' ($\tau_{m,m} = 1$). We assume free trade so that the price of c_1 is homogeneous and normalized to 1. The price of housing at location m' is denoted by $p_{m'}$ and we assume that the sector is non-competitive, e.g. $p_{m'}$ is fixed and does not vary with $N_{m',t}$.[4] This allows to simplify the environment with preferences that depend only

[4]Competition in the housing sector would imply additional congestion forces on the housing market, which is not the focus of our paper. In fact, we assume sticky local prices given the short duration of the movements considered. Namely, local markets do not adjust to short-term changes in local demand. However, future work could consider a version of the model that relaxes this assumption.

on the consumption of the tradable good:

$$u_{m,m'}(c_1) = \frac{A_{m'} c_1}{\tau_{m,m'}} \tag{3.4}$$

Conditional on moving from location m to location m' at time t and assuming no savings, individual i maximizes utility subject to the budget constraint $w_{m,t}$. So the indirect utility associated with choosing m' can be directly expressed as:

$$V_{m,m',t} = \frac{A_{m'} w_{m',t}}{\tau_{m,m'}} \tag{3.5}$$

3.2.5 Location choice

We assume that, for any individual i residing in location h, location choices are consistent with a random utility model that depends on the indirect utility associated with the location choice and an idiosyncratic taste shock:

$$v^i_{m,m',t} = \ln\ V_{m,m',t} + \epsilon^i_{m',t} \tag{3.6}$$

The idiosyncratic taste shock follows a nested logit structure. The choice of location is modelled as a two-step decision process where the individual first decides whether to be at home or in migration (i.e. at a non-home location) in the upper nest, and then picks a specific location in the lower nest conditional on choosing to be away from home. As in Monras (2018) and Imbert and Papp (2020b), among others, this structure allows to account for home biased preferences. Namely, it allows to make the home location a special place that individuals disproportionately prefer. The absence of home biased preferences would otherwise describe a world in which migration rates would be substantially higher that what is observed in practice (Imbert and Papp, 2020b).

The inverse of the scale parameters associated with the upper and lower nests are denoted by θ and $\bar{\theta}$. $\mathcal{M} = \{m_1, ..., m_M\}$ is the universal choice set and we define $\overline{\mathcal{M}^i} = \{m_1, ..., m_M\} \backslash \{h_i\}$, the choice set in the lower nest corresponding to the set of non-home locations for individual i. With this structure, the probability of choosing a location $m_j \in \overline{\mathcal{M}^i}$ can be expressed as the product of the marginal probability of choosing $\overline{\mathcal{M}^i}$ (i.e. being in migration) with the conditional probability of choosing $m_j \in \overline{\mathcal{M}^i}$.

In the lower model, the conditional probability of choosing a location $m_j \in \overline{\mathcal{M}^i}$ takes the form of a logit model:

$$Pr(m_{i,t} = m_j | \overline{\mathcal{M}^i}, m_{i,t-1}) = \frac{V^{1/\bar{\theta}}_{m_{i,t-1},m_j,t}}{\sum_{k \in \overline{\mathcal{M}^i}} V^{1/\bar{\theta}}_{m_{i,t-1},m_k,t}} \tag{3.7}$$

In the upper model, the marginal probability of choosing to be away from home also follows a logit model:

$$Pr(\overline{\mathcal{M}^i}|m_{i,t-1}) = \frac{V_{\bar{h}}^{1/\theta}}{V_{m_{i,t-1},h,t}^{1/\theta} + V_{\bar{h}}^{1/\theta}} \tag{3.8}$$

Where $\ln V_{\bar{h}} = \bar{\theta} \ln \sum_{k \in \overline{\mathcal{M}^i}} V_{m_{i,t-1},m_k,t}^{1/\bar{\theta}}$.

The probability of being in migration at destination m_j conditional on the location at $t-1$, $m_{i,t-1}$, is therefore given by:[5]

$$Pr(m_{i,t} = m_j|m_{i,t-1}) = \frac{V_{\bar{h}}^{1/\theta}}{V_{m_{i,t-1},h,t}^{1/\theta} + V_{\bar{h}}^{1/\theta}} \frac{V_{m_{i,t-1},m_j,t}^{1/\bar{\theta}}}{\sum_{k \in \overline{\mathcal{M}^i}} V_{m_{i,t-1},m_k,t}^{1/\bar{\theta}}} \tag{3.10}$$

We make further assumptions allowing to arrive at a simplified version of (3.10) that supports the empirical analysis of section 3.4, while retaining the relevant features allowing to rationalize the mechanisms at play. In particular, we assume no mobility frictions and no congestion on the labor market, which allows to obtain a closed-form expression for the log probability of being in migration at location m_j at time t for an individual residing in h (see proof in appendix 3.A):

$$\ln Pr(m_{i,t} = m_j) = \frac{\alpha_{m_j}}{\bar{\theta}} \ln K_{m_j,t} - \frac{\alpha_h}{\theta} \ln K_{h,t} + \frac{1}{\bar{\theta}} \ln \tilde{\Gamma}_{m_j,s(t)} - \frac{1}{\theta} \ln \tilde{\Gamma}_{h,s(t)} + C \tag{3.11}$$

Equation (3.11) provides simple intuitions about labor reallocation dynamics in the presence of inter-year differences in rainy season conditions. Conditional on a particular time of year $s(t)$, the elasticity of the bilateral stock of temporary migrants between an origin o and a destination d at time t with respect to conditions at origin $K_{o,t}$ is $-\frac{\alpha_o}{\theta} \leq 0$. The elasticity with respect to conditions at destination is $\frac{\alpha_d}{\bar{\theta}} \geq 0$. The magnitude of the elasticity with respect to conditions at origin (resp. at destination) thus depends on the sensitivity of the production at origin (resp. at destination) to rainfall conditions. Higher values of α_o (resp. α_d) imply larger effects on local wages at origin (resp. destination), which in turn affect the value of the origin (resp. destination) location. For instance, all else equal, rural areas with a dominant agricultural sector using a larger share of natural capital will show a stronger sensitivity of the propensity to out-migrate with respect to local

[5]The probability of being at the home location h is simply:

$$Pr(m_{i,t} = h|m_{i,t-1}) = \frac{V_{m_{i,t-1},h,t}^{1/\theta}}{V_{m_{i,t-1},h,t}^{1/\theta} + V_{\bar{h}}^{1/\theta}} \tag{3.9}$$

conditions. On the other hand, the elasticity of the bilateral stock with respect to local conditions at origin (resp. destination) also naturally depends on the elasticity of substitution in the upper (resp. lower) model, $\frac{1}{\theta}$ (resp. $\frac{1}{\bar{\theta}}$). This parameter defines the elasticity between staying at the home location or moving to a different place and lower values of θ therefore increase the elasticity of the bilateral stock with respect to local conditions at origin. Similarly, lower values of $\bar{\theta}$ result in individuals being more sensitive to conditions at destination.

3.3 Data description

3.3.1 Phone-derived temporary migration estimates

The temporary migration measures utilized in this chapter rely on the work of chapter 2. As a reminder, CDR are mobile phone metadata containing information on phone transactions such as calls and text messages. Each user is associated with a unique identifier[6] and billing logs provide the timestamp and approximate location of each call made or received by the user, which allows to reconstruct trajectories through space and time. Then, a two-step algorithm is applied to identify temporary migration events in raw user-level CDR trajectories. First, each user is assigned a home location. Second, "meso-segments" are defined as continuous periods of time over which a user is consistently seen at a single location and are detected with a clustering technique. Temporary migration events are then identified as meso-segments at a non-home location with a duration of at least 20 days.

In this chapter, we construct a pseudo-panel of temporary migration estimates at the (origin * destination * time) level. The spatial unit of analysis is the voronoi cell, considering the set of 916 cells constructed in chapter 2.[7] Each time unit corresponds to a "half-month", which is defined as the periods going from the 1^{st} to the 15^{th}, and from the 16^{th} to the end of each month. Each year is thus comprised of 24 half-months. The final dataset covers the period 2013-2015 and therefore has a total of 60,412,032 observations. The main outcome of interest is the migration stock rate, which we define for each origin o, destination d and time period t as the fraction of users residing in o who are in temporary migration at destination d during t. By the law of large number, this corresponds to an estimator for the probability that an individual i residing in o chooses to be in

[6] More precisely, each unique identifier in the CDR data represents a SIM card.

[7] As a reminder, cells are defined in a way that tends to balance their size across the sample. Also, each city corresponds to a unique cell and 39 cells are effectively classified as urban.

location d at time t, i.e. the left-hand side of equation (3.11) presented in section 3.2.5.

Several features make this phone-derived migration matrix particularly germane to the purpose of the present study.

First, CDR-based measures from previous studies have focused on migration flows, i.e. on movements from one location to another over some periods of time.[8] By contrast, the methodology of chapter 2 defines a home location for each user and departing flows can thus be distinguished from returning flows. At any given time, we are able to unambiguously characterize the observed location of a user as his home location or a non-home location. This is an essential point in the environment of the model presented in section 3.2, where home bias preferences embedded in the idiosyncratic taste shock make the home location a special place that individuals disproportionately prefer.

Second, the high level of spatial disaggregation allows to define rainy season conditions at a local level, which better reflects the conditions effectively experienced by individuals. On the other hand, the spatial resolution of survey-based migration measures is usually limited to administrative levels. Moreover, the available spatial granularity allows to capture short movements that are missed in migration estimates considering larger spatial units, although they may be far from negligible. For instance, as seen in chapter 2, rural-to-rural movements represent a large fraction of the total rural-out flow observed and they predominantly occur within regions.

Third, the ability to precisely capture migration destinations is a major advantage over survey measures. This allows us to characterize the conditions at destination locations and investigate their impact on the decision to out-migrate. As later explained in section 3.4.2, failing to account for rainy season conditions at destinations would actually induce a serious threat to the identification of the effect of conditions at origin.

Fourth, the high temporal resolution of migration estimates and the multi-year period of observation allow to identify the effect of rainy season conditions while controlling for seasonality. In other words, we are able to exploit variations in rainy season conditions within locations and for specific periods of the year – more specifically, within half-months. This is particularly crucial for two reasons. Firstly, temporary migration patterns in Senegal are highly seasonal (see chapter 2) so that any effect caused by climate variability would be difficult to identify using intra-year variations. Moreover, using within-location variations over time rather than cross-sectional variations is more suitable for the identification of the effect of rainy season conditions which typically vary more over time than across space

[8] See, for instance, Blumenstock (2012), Chi et al. (2020), and Lai et al. (2019).

(Blanchard et al., 2023; Hill and Porter, 2017).

3.3.2 Rainy season conditions

We quantify the quality of rainy seasons based on precipitations received locally. We use the Climate Hazard Center's CHIRPS-2.0 gridded precipitations product, which provides daily precipitations at a 0.05° spatial resolution for the period 1981 to present (Funk et al., 2015). Cell-level rainfall values are calculated with interpolation techniques allowing to blend gauge station data with high-resolution satellite-based precipitation estimates.[9] The performance of CHIRPS for monitoring drought conditions has been demonstrated in a variety of contexts (Aadhar and Mishra, 2017; Guo et al., 2017; Mianabadi et al., 2022; Najjuma et al., 2021; Pandey et al., 2021; Sandeep et al., 2021).

For each voronoi cell, precipitations are aggregated spatially in a 10km buffer centered on the cell centroid. This allows to obtain a homogeneous geographic basis for the definition of rainy season conditions across cells, despite differences in their shape and size. Daily cell-level precipitations are then aggregated over relevant time windows to define the observed quality of the rainy season for any given time period t. We get back to this in more details in section 3.4.

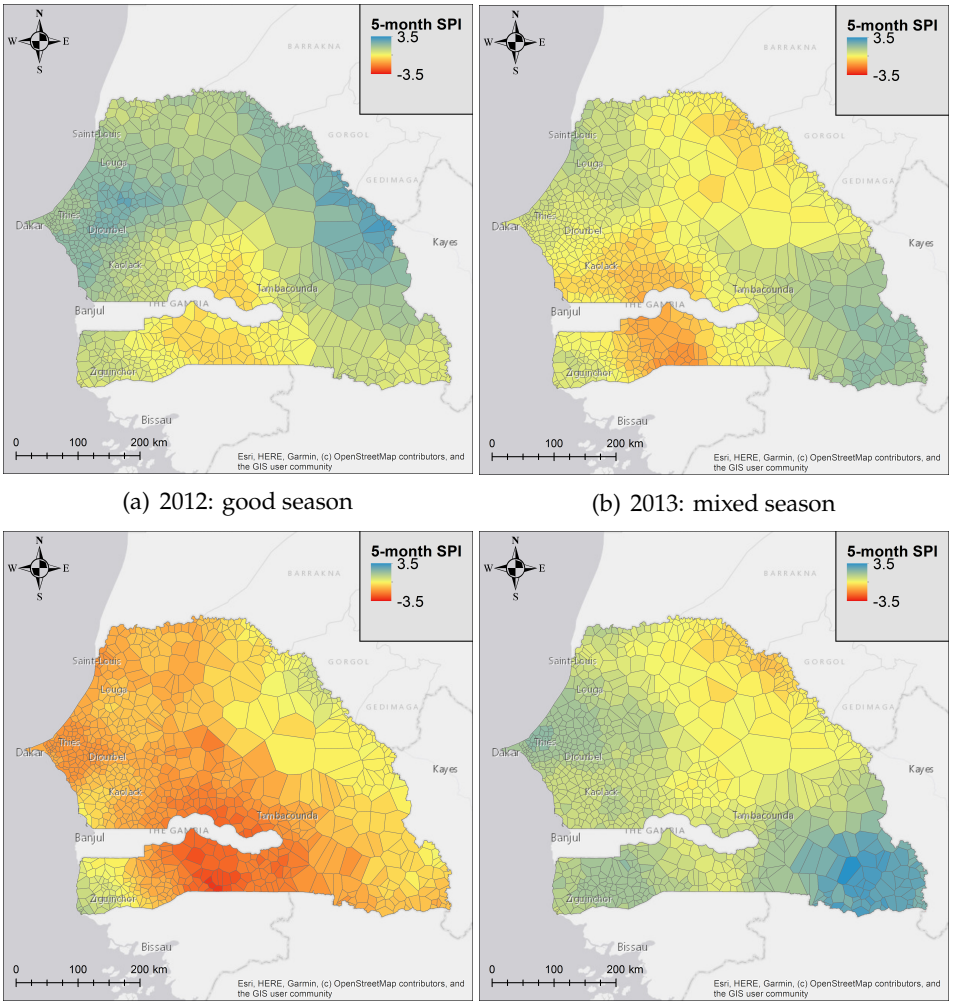
Important variations in the quality of rainy seasons are observed over the study period (2013-2015). For descriptive purposes, we quantify the quality of rainy seasons over the period of interest with precipitation anomalies based on the Standardized Precipitation Index (SPI), which measures the normalized distance of local precipitation estimates to long-term means (McKee et al., 1993). The SPI has the advantage of providing comparable measures across areas with distinct rainfall regimes. We show maps of precipitation anomalies calculated over the entire rainy season (June-October) for the years 2012 to 2015[10] in Figure 3.1. Only localized rainfall deficits were observed in 2013, while 2014 was clearly a drought year where the vast majority of locations experienced lower-than-average precipitations. In 2012 and 2015, however, precipitations were in excess almost everywhere in the country. The 2014 drought is of particular interest in our empirical setting as it provides an opportunity to observe migration behaviors for a realization of rainfall

[9]The number of gauge stations available varies significantly over time. Monthly maps of stations used in CHIRPS estimates in Senegal can be consulted using the following link:http://data.chc.ucsb.edu/products/CHIRPS-2.0/diagnostics/chirps-n-stations_byCountry/Senegal/

[10]We also show a map for the 2012 season since the model assumes that migration choices observed in early 2013 before the 2013 rainy season are influenced by the quality of the most recent rainy season, i.e. 2012.

conditions at the lower end of the distribution.

Figure 3.1: 5-month precipitation anomalies over the rainy season, 2012-2015.



(a) 2012: good season

(b) 2013: mixed season

(c) 2014: drought

(d) 2015: good season

Note: Historical precipitation values at the voronoi-level are estimated with gridded precipitation data from the CHIRPS2.0 product of the Climate Hazard Group. SPI values are calculated based on historical values of accumulated precipitations for the June-October period since 1981.

3.4 The effect of rainy season conditions on temporary migration

3.4.1 Context and empirical setting

More than half of the population in Senegal resides in rural areas where livelihood means mainly rely on subsistence agriculture and livestock. Economic activities are therefore highly dependent on precipitations received during the rainy season, between June and October. Agricultural productivity is also characterized by pronounced seasonal fluctuations as a result, especially since irrigation access remains marginal. Indeed, 87% of agricultural households practice rainfed agriculture[11] and the few irrigated lands concentrate in the northern part of the country along the Senegal River Valley (see map of Figure 3.B.2 in Appendix 3.B). The annual aggregate demand for agricultural labor peaks around the period going from June to August for soil preparation, sowing and planting, and from September to December for the main harvesting season. Off-season agricultural activities during the January-May period are limited and mostly confined to the Senegal River Valley zone. In that area, access to irrigation is more widespread and households also practice flood recession agriculture (see maps of Figures 3.B.2-3.B.4), which allows to grow crops outside of the usual rainy season period.

As a result, for locations dominated by rainfed agriculture, rainy season conditions affect agricultural productivity from the start of the agricultural campaign in June to the end of the main harvesting period in December. On the other hand, locations with higher access to irrigation and counter-season production are more likely affected during the off-season, between January and May. By contrast, the quality of rainy seasons is expected to have only a limited impact on the level of productivity in urban areas where agricultural activities are typically marginal. Overall, spatial disparities in the rainy season quality, differences in the prominence and resilience of the agricultural sector, and distinct agricultural calendars across locations constitute the key forces driving dynamic differences in labor productivity across areas and over time, creating incentives for individuals to temporarily relocate where they are most productive.

As outlined in the model of section 3.2, migration decisions are also dampened by the existence of movement costs and idiosyncratic preferences for the home location. Also, labor reallocation at the equilibrium is potentially shaped by congestion forces. For instance, the labor market can be a source of congestion: an increase in

[11] Source: RGPHAE 2013, Agence Nationale de la Statistique et de la Démographie (ANSD) de la République du Sénégal, www.ansd.sn

labor via migration to a particular location is associated with lower wages, which in turn decrease the value of that location. Similarly, congestion effects on a broadly defined housing market can increase the cost of staying at a particular destination. We do not explicitly investigate those in this paper and rather focus the scope of our empirical analysis on the direct effect of rainy season conditions on temporary migration choices. Whether local population dynamics induced by temporary migration movements lead to market adjustments remains an open question that we leave for future research.

Over the medium term, we observe important inter-year variations in the amount and spatio-temporal distribution of rainfalls, with the occasional occurrence of very good years (i.e. higher-than-average precipitations) and drought years with significant rainfall deficits that can be localized or more extensive. Our empirical analysis exploits the exogenous variations in the quality of rainy seasons within locations and over the period 2012-2015. The temporal granularity of the data allows to precisely investigate the timing of the effect of rainy season conditions on temporary migration choices.

The use of temporary migration as a strategy to cope with variations in the quality of rainy seasons is likely mediated by the availability of other coping mechanisms. These include, for instance, efforts to reduce the level of risk – for instance, through technology adoption (Dercon and Christiaensen, 2011) – or consumption smoothing strategies via savings/credit markets (Basu and Wong, 2015), buffer stocks (Fafchamps, Udry, et al., 1998; Chaudhuri and Paxson, 2021) or risk-sharing networks (Fafchamps and Gubert, 2007). Of course, a key limitation in the use of mobile phone data is that they do not provide this type of information about users. We are therefore not in a position to study the interaction effect of the availability of alternative coping strategies at the individual-level with the rainy season quality on temporary migration choices. Similarly, our data do not offer the possibility to control for individual-level characteristics known as key determinants of the propensity to migrate such age, sex or income sources. As a second-best strategy, we match the origin and destination locations of our temporary migration estimates with spatially disaggregated information from secondary data sources, which allows us to perform some heterogeneity analyses.

3.4.2 Conventional migration equation

We start by naively estimating the conventional migration equation that generally relates a local change in population to a local shock. As pointed out in Borusyak, Dix-Carneiro, et al. (2022), a plethora of studies have considered this equation to

investigate the impact of local shocks on labor reallocation dynamics and have arrived at varying, and sometimes puzzling conclusions. Local shocks are found to be associated with significant effects on economic outcomes such as wage or employment, but imply only limited migratory responses.[12] This has led researchers to conclude that migration costs are simply prohibitive and individuals unresponsive to local shocks. Instead, Borusyak, Dix-Carneiro, et al. (2022) suggest that most interpretations of the conventional migration equation are flawed since conditions at relevant alternatives are simply ignored. In situations where shocks at potential destinations are correlated with the local shock at origin, this creates a problematic case of omitted variable bias. This is clearly the case in our empirical setting where the spatial correlation in local rainy season conditions is obvious in the maps of Figure 3.1. Therefore, this setting offers the opportunity to evaluate the seriousness of this issue, since the granularity of our migration estimates allows to incorporate rainy season conditions at destination in our specifications. Also, this is a natural starting point in the analysis and similar estimations have been conducted in other studies investigating migration responses to drought conditions using survey data (Defrance et al., 2023; Henry et al., 2004). In practice, those surveys provide information on the location of sending areas but are usually more limited in their ability to capture destination locations.

Our identification strategy uses within-location variations in the quality of rainy seasons over the multiple years observed in our sample in order to identify their marginal effect on the rate of temporary out-migration. More specifically, we observe at each time period t and for each origin o the fraction of users from o that are in migration during t. We use granular units of analysis on the time dimension, i.e. half-months. For each origin location o, we are thus able to observe variations in the quality of rainy seasons over multiple years for each half-month of the year. This allows to disaggregate the effect of rainy season conditions on temporary migration over the agricultural year at a relatively fine temporal scale. Any given time unit t is uniquely defined by a year y and the half-month of the year within which t falls, indexed by a parameter s. The index s thus takes on values ranging from 1 (the first half of January) to 24 (the second half of December).

Consistent with the model presented in section 3.2, location choices at time t are based on the information on the quality of the rainy season available when the decision is made. We denote this treatment variable by $x_{o,t}$ or, equivalently, $x_{o,s,y}$. In practice, and consistent with equation (3.11), this corresponds to the logged precipitations accumulated between the start of the most recent rainy season and

[12] See Borusyak, Dix-Carneiro, et al. (2022) for a review.

$t-1$. The start of the rainy season is set to the first half of June and ends on the second half of October and therefore lasts for 5 months, i.e. 10 half-months. We formally define $x_{o,s,y}$ as a function of the precipitations variable $rain^{\Delta}_{o,s,y}$, which represents the amount of precipitations accumulated in location o at time $t=(s,y)$ over the past Δ half-months. For instance, $rain^{2}_{o,12,2013}$ denotes the precipitations accumulated over a period of 2 half-months ending at the 12$^{\text{th}}$ half-month of 2013 (the second half of June), i.e. the total precipitations received in June 2013 at location o. With these notations, $x_{o,s,y}$ can be expressed as[13]:

$$x_{o,s,y} = \begin{cases} \ln rain^{10}_{o,20,y-1} & if \quad 1 \leq s \leq 11 \\ \ln rain^{s-11}_{o,s-1,y} & if \quad 12 \leq s \leq 20 \\ \ln rain^{10}_{o,20,y} & if \quad 21 \leq s \leq 24 \end{cases} \tag{3.12}$$

For each time period $t=(s,y)$ and origin o, we observe the fraction of users residing in o who are in temporary migration – to any destination. We thus estimate a simplified version of equation (3.8) with a Poisson Pseudo-Maximum Likelihood (PPML) estimator:

$$P_{o,s,y} = exp(\beta_0 + \sum_{k=1}^{24}(\beta_k x_{o,s,y} \times \alpha_s^k) + \gamma_{o,s} + \eta_{o,s,y}) \tag{3.13}$$

$P_{o,s,y}$ is the total out-migration rate from origin o in half-month s of year y. $x_{o,s,y}$ is the observed rainy season quality in o for half-month s of year y, as defined by equation (3.12). Each coefficient of interest β_k represents the elasticity of out-migration with respect to precipitations at the origin location.[14] α_s^k is a dummy equal to 1 when $k=s$. $\gamma_{o,s}$ is an (origin*half-month) fixed effect that acts as a location-specific control for seasonality. The panel structure with three years of observation together with the granularity of our migration matrix on both the

[13]For the first 11 half-months of a year y, the treatment corresponds to precipitations accumulated over the most recent rainy season, which is the 10-half-month period that ended on the second half of October ($s=20$) in year $y-1$. Similarly, in November and December of year y ($21 \leq s \leq 24$), $x_{o,s,y}$ is the precipitations accumulated over the rainy season that just ended, which is the 10-half-month period that ended on the second half of October ($s=20$) in year y. Finally, for any half-month between the second half-month of the rainy season ($s=12$) and the last one ($s=20$), $x_{o,s,y}$ corresponds to the precipitations accumulated from the start of the rainy season ($s=11$) to the preceding time unit. The length of the period over which the anomaly is calculated therefore varies from 1 half-month to 10 half-months.

[14]An alternative definition for $x_{o,s,y}$ could be used based on the precipitation anomaly defined by the SPI, using the time windows considered in equation (3.12). More specifically, we can replace the logged precipitations by the (non-logged) SPI-based rainfall anomalies and interpret β_k as a semi-elasticity. We prefer to use the logged precipitations since it simplifies the interpretation of estimated coefficients. However, we occasionally consider an SPI-based definition of $x_{o,s,y}$ because it enters linearly in our specification which turns out more convenient in some cases, for instance when testing the existence of non-linearities.

space and time dimensions allow to consider this restrictive set of fixed effects.[15] Note that in this general specification the coefficient of interest varies by half-month of the year but is fixed across locations. We later explore the heterogeneity of the effect across locations by interacting the treatment $x_{o,s,y}$ with categorical variables defining different groups of cells, e.g. rural and urban locations.

Results are presented in Figure 3.2.[16] Standard errors are clustered at the (origin*half-month)-level and error bars on the graph show the 95% confidence intervals. We observe a positive and significant effect for the September-November period, that corresponds to the start of the harvest season. A 10% increase in precipitations at a location is associated with an increase in the out-migration rate from that location of between 3% (in September) and 7% (in November) on average. The effect seems to persist until the month of March the following year before fading away at the end of the agricultural year. Figure 3.C.1 shows results of the same estimation including an urban dummy interaction that allows for heterogeneous effects for rural and urban locations respectively. Given that they dominate the set of locations considered (877 out of 916), rural locations are unsurprisingly found to be largely driving the patterns observed in Figure 3.2. Urban areas seem to also be responding in a similar way over the September-November period, and coefficients for the rest of the agricultural year are positive but imprecisely estimated. The positive relationship between the rainy season quality at origin and the out-migration rate contradicts the predictions of our model. Poorer rainfall conditions are expected to act as a push factor for temporary migration via a decrease in the local productivity but are found to deter migration instead. However, the results could be supporting a liquidity constraint narrative in which people directly affected by poor agricultural outcomes are not able to bear the cost of migration – or cannot hedge against the risk of migration failure.[17]

As mentioned above, a critical limitation of this estimation is that it fails to

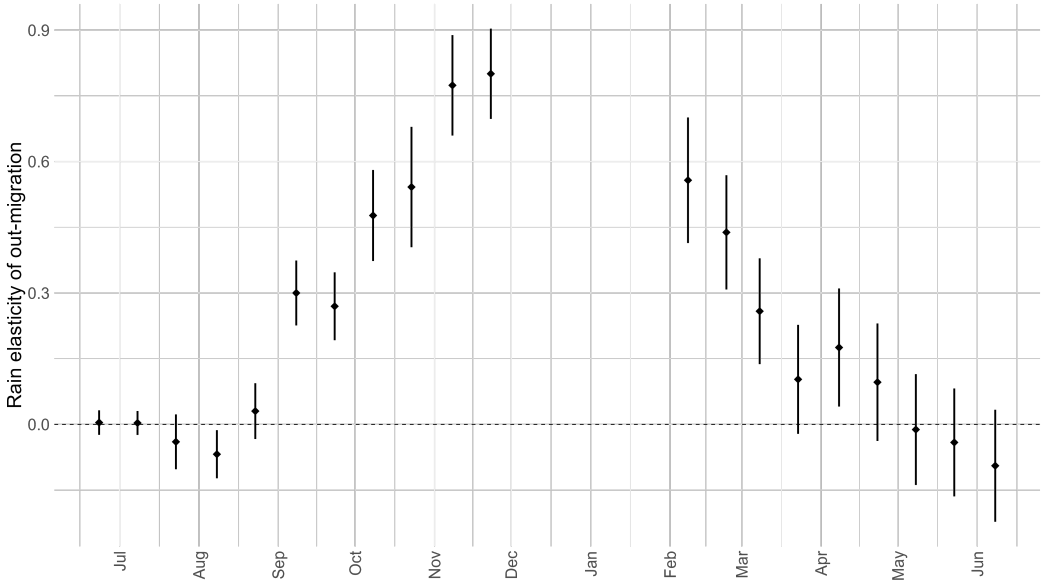
[15] We essentially work with 916 locations observed over 3 years that we decompose into 72 time periods. This represents a matrix of over 60 million observations. Aggregating over destinations to get total migration rates by origin and half-month still leaves us with nearly 66,000 observations.

[16] Note that no coefficients are estimated for the months of December and January. This is because temporary migration estimates in January 2013, December 2013, January 2014 and December 2015 are associated with large measurement errors and considered as missing in the final dataset. Indeed, the 2013-2015 raw CDR dataset is split into two sub-periods: 2013 and 2014-2015. This means that users observed in the 2013 subset cannot be identified in the 2014-2015 subset, and conversely. This creates censoring effects at each end of both subsets. As a result, we are not able to infer an approximate duration for mobility events that start before, or end after the period of observation. For instance, if a user migrates for 30 days from December 15, 2012 to January 14, 2013, we will only be able to observe that he is at a non-home location for at least 14 days from January 1, 2013 to January 14, 2013, which is insufficient to classify this mobility event as a temporary migration move.

[17] Note that a more elaborate migration cost structure would be required to include this feature in the conceptual framework proposed in section 3.2.

account for characteristics on the supply side of the migration market. More specifically, it ignores rainy season conditions at potential destinations. An alternative explanation of the results of Figure 3.2 therefore relies on the high degree of spatial correlation in the quality of the rainy season. The destinations where individuals usually migrate to find agricultural employment were most likely also affected by poor rainfall conditions, so that our measures of conditions at origin may incidentally capture conditions at relevant destinations. It thus becomes unclear whether our results describe a situation in which poorer conditions lead to a lower propensity to migrate through tighter liquidity constraints at origin or lower productivity at destinations that decrease their value.

Figure 3.2: Elasticity of out-migration estimated by half-month over the agricultural year.



Note: Each point estimate represents the average elasticity of out-migration with respect to precipitations at origin. Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*half-month)-level.

3.4.3 Effect of rainy season conditions at origin and destination on bilateral temporary migration rates

Here, we use the bilateral migration estimates, $P_{o,d,s,y}$, to estimate the model described by equation (3.11). The richness of our data allows to precisely identify the locations of origin and destination of the temporary migration movements observed. As a result, we are able to estimate the effect of rainy season conditions at destination on bilateral rates of temporary migration to that particular destination. In other words, we can identify the impact of the quality of the rainy season on the

level of attractiveness of locations to potential temporary migrants. We estimate the following model with a PPML estimator:

$$P_{o,d,s,y} = exp(\beta_0 + \sum_{k=1}^{24}(\beta_{1,k} x_{o,s,y} \times \alpha_s^k + \beta_{2,k} x_{d,s,y} \times \alpha_s^k) + \gamma_{o,d,s} + \eta_{o,d,s,y}) \quad (3.14)$$

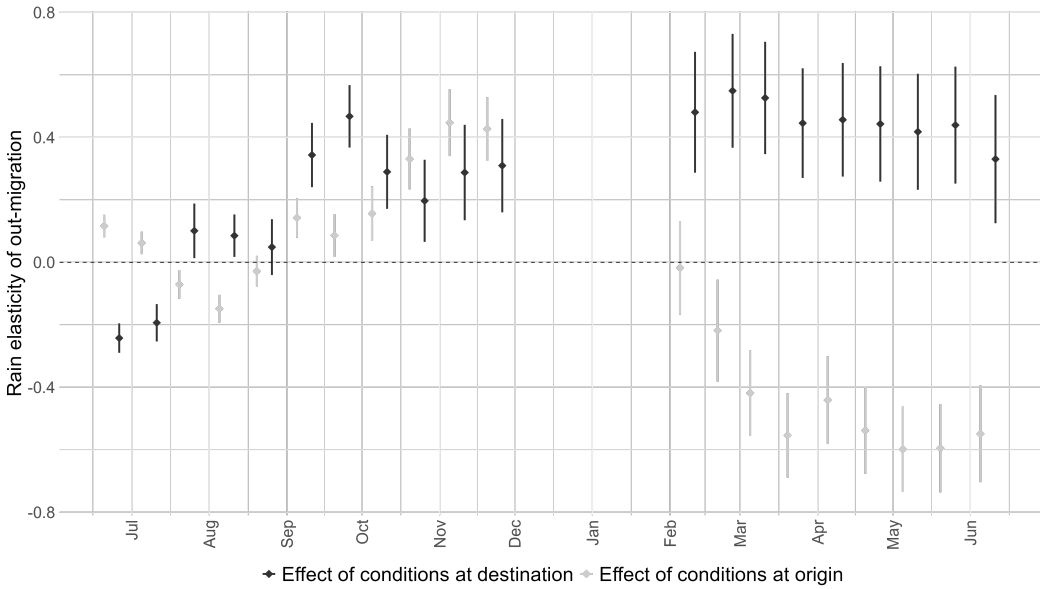
$P_{o,d,s,y}$ is the fraction of residents in o that are in migration at destination d during half-month s of year y. $\gamma_{o,d,s}$ is an (origin*destination*half-month) fixed effect. Note this is a more restrictive specification compared to equation (3.11) derived from the model, where fixed effects at the (origin*half-month)- and (destination*half-month)-level enter the expression separately. Here, we use the variations within each pair of origin-destination locations over multiple years to identify the effects of rainy season conditions at origin and destination on the bilateral migration rate, for each half-month of the year. Although mobility frictions were ignored in section 3.2 in order to simplify the algebra, it is important to note that the fixed effect term that we consider actually absorbs any gravity effect related to the distance between o and d.[18]

We estimate equation (3.14) allowing for heterogeneous effects of the rainy season quality at origin and destination on the bilateral migration stock, for rural and urban locations. More specifically, we interact $x_{o,s,y}$ with a categorical variable indicating whether o is a rural or an urban location and, similarly, $x_{d,s,y}$ is interacted with a variable indicating whether d is classified as rural or urban. Four coefficients are thus estimated for each half-month k: $\beta_{1,k}^{rural}$, $\beta_{1,k}^{urban}$, $\beta_{2,k}^{rural}$ and $\beta_{2,k}^{urban}$. We focus first on the effect of rainy season conditions at origin and destination for rural locations, i.e. on the coefficients $\beta_{1,k}^{rural}$ and $\beta_{2,k}^{rural}$. Results are shown in Figure 3.3 where light dots represent the effect of conditions at origin ($\beta_{1,k}^{rural}$) on the bilateral stock of temporary migration, and dark dots the effect of conditions at destination ($\beta_{2,k}^{rural}$). Rainy season conditions at origin have a positive and significant effect over the period September-November (i.e., the harvest period), with an elasticity reaching 0.4 in November. This is consistent with the results from the estimation of the conventional migration equation (Figure 3.2) in the previous section, although the magnitude of the effect is roughly halved. The effect of conditions at destination over the same period is also found to be positive and significant. On average, a 10% increase in precipitations at a rural location is associated with an increase in bilateral stocks of temporary migrants from other locations to that destination

[18]We also estimate a different version of equation (3.14) which is closer to the expression yielded by the model in equation (3.11), where we keep separate (origin*half-month) and (destination*half-month) fixed effects but explicitly control for the cost of distance between origin and destination. To do this, we calculate the travel time by car between each pair of cell centroids, $\tau_{o,d}$, via the Open Source Routing Machine (OSRM) project(http://project-osrm.org/) that uses OpenStreetMap data. We provide the results of the main estimation for rural locations in Figure 3.C.2 in Appendix 3.C.2, and find that they are qualitatively unchanged compared to the results shown in 3.3.

by 2 to 4%. Overall, these results suggest that part of the effect of rainy season conditions at origin on the out-migration rate during the harvest season that is captured in the conventional migration equation is in fact biased by the effect of rainy season conditions at destinations. Interestingly, effects of conditions at origin and destination for that period of the year are comparable in size.

Figure 3.3: Elasticity of the bilateral migration stock over time with respect to conditions at origin and destination, rural locations.



Note: Each point estimate represents an average elasticity of the bilateral out-migration stock with respect to precipitations at origin (light dots) and destination (dark dots) in rural locations. All coefficients are obtained with a PPML estimation in a single regression of the bilateral stock of temporary migrants between an origin location and a destination, against the logged precipitations at origin interacted with an urban origin dummy and a half-month indicator, and the logged precipitations at destination interacted with an urban destination dummy and a half-month indicator. This graph shows the coefficients associated with urban origin and urban destination dummy interaction values equal to 0 (i.e. rural). Temporary migration bilateral rates are based on migration events of at least 20 days. Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

This important finding supports the idea introduced in our model: both local conditions at origin and destination participate in shaping temporary migration decisions during the harvest period. The direction of the effect of conditions at destination is consistent with our model. Relatively poorer rainfall conditions at a given rural location are associated with a decrease in its level of attractiveness as a migration destination. This supports the mechanism whereby lower precipitations cause a decrease in local productivity, which implies lower wages and eventually a decrease in the value of that destination. This effect persists for the rest of the period: the degree of attractiveness of locations between February and June is

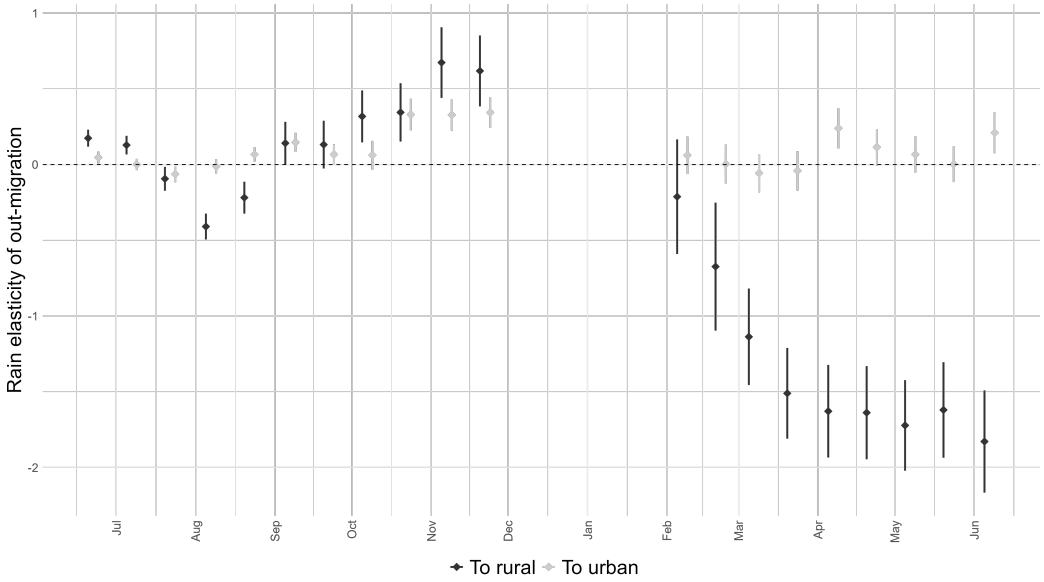
positively linked with the quality of the most recent rainy season, with an elasticity that stabilizes at around 0.4-0.5.

However, similar to the conventional migration equation, the estimated effect of conditions at origin over the September-November period remains rather puzzling. Even considering bilateral migration measures and controlling for rainy season conditions at destination, rainfall conditions at origin are found to be positively associated with the bilateral out-migration rate, although our model predicts an effect in the opposite direction. Instead, the results could be suggestive of the existence of an increase in liquidity constraints with the worsening of rainfall conditions, rendering migration options temporarily unfeasible. In fact, several empirical studies have found evidence supporting the notion that climatic events could effectively decrease migration movements (Gray and Mueller, 2012b; Henry et al., 2004; Hirvonen, 2016; Mueller et al., 2020). An alternative explanation could rely on the existence of non-linear effects of precipitations on local productivity. More specifically, excess rainfall could in fact have a negative impact on agricultural production and thus act as a push factor. We find evidence supporting this hypothesis in robustness checks that further investigate non-linearities in the precipitation-migration relationship and that we present later in this section.

On the other hand, the pattern of results observed during the off-season, from February to early June, are more in line with expectations derived from the conceptual framework. From the second half of February onward, the effect of rainy season conditions at origin is negative and precisely estimated: on average, for a rural origin location, a 10% increase in precipitations accumulated over the past rainy season leads to a decrease in bilateral out-migration stock rates of between 4% and 6% during the March-June period. To investigate this effect further, we adopt a distinct specification where precipitations at the origin are interacted with both an urban indicator that signifies if the origin o is an urban or a rural location, and another urban indicator determining if the destination d is urban or rural. This enables us to inspect whether the observed effects of rainy season conditions at a rural origin are driven by temporary migration movements specifically directed to other rural destinations or, conversely, to urban areas. Results are presented in Figure 3.4. Perhaps surprisingly, the effect seems entirely driven by temporary migration movements to other rural locations and we find no response of temporary migration choices from rural locations to cities. Since rural-to-rural movements were previously identified as primarily short-distance, we test whether these dynamics could explain the observed patterns of results. We thus re-estimate our model excluding pairs of adjacent locations, but the coefficients remain practically

unchanged (see Figure 3.C.11 in appendix).

Figure 3.4: Elasticity of the bilateral migration stock over time with respect to conditions at a rural origin, by zone of destination.



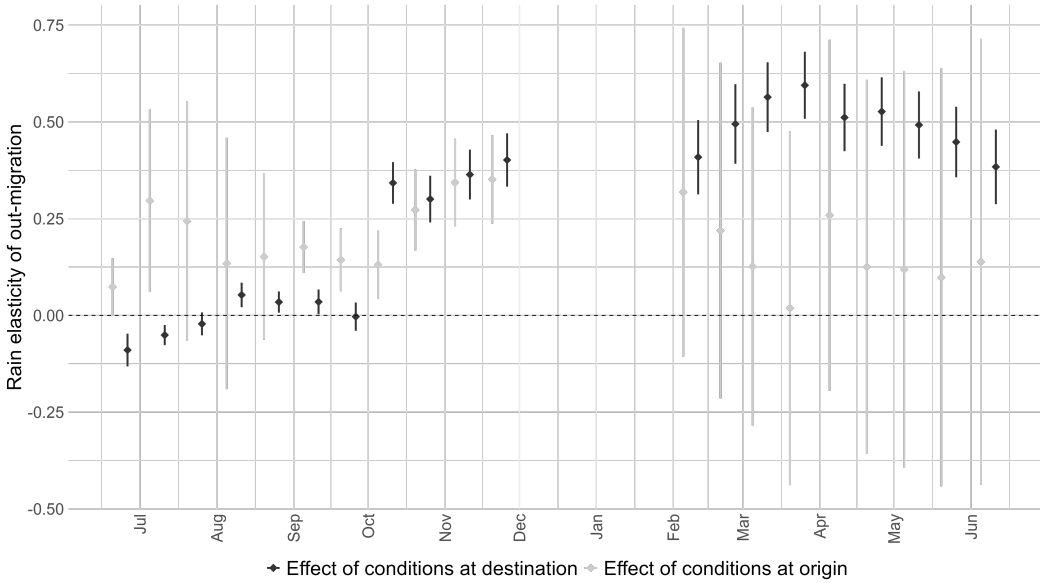
Note: Each point estimate represents an average elasticity of the bilateral out-migration stock with respect to precipitations at a rural origin. Light dots represent the effect on the movements toward urban locations, whereas dark dots show the effect on temporary migration to other rural areas. All coefficients are obtained with a PPML estimation in a single regression of the bilateral stock of temporary migrants between an origin location and a destination, against the logged precipitations at origin interacted with an urban origin dummy, an urban destination dummy, and a half-month indicator, and the logged precipitations at destination interacted with an urban destination dummy and a half-month indicator. This graph shows the coefficients associated with the urban origin dummy interaction value equal to 0 (i.e. rural). Temporary migration bilateral rates are based on migration events of at least 20 days. Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

Moreover, the effect of conditions at destination observed for the harvest period persists during the off-season: the degree of attractiveness of locations between February and June is positively linked with the quality of the most recent rainy season, with an elasticity that stabilizes at around 0.4-0.5. Consistent with the model predictions, rainy season conditions at origin and destination create opposite forces for the decision to temporarily migrate during the off-season: poorer conditions at origin act as a push factor while similar conditions at destination decrease the value of migrating to that destination. This result clearly highlights the value of phone-derived, highly granular, migration estimates to identify the effects of interest. The ability to control for conditions at destination changes the conclusion on the effect of rainy season conditions at origin on the propensity to out-migrate compared to the results of section 3.4.2. This indicates that we should be cautious

in interpreting results from conventional migration regression estimations that ignore conditions at destination.

In the same vein, Figure 3.5 shows the effect of rainy season conditions at origin and destination on the bilateral stock of temporary migrants for urban locations. Conditions at origin have a positive and significant effect on the bilateral out-migration rate over the September-November period, and the pattern is surprisingly similar to the one observed for rural locations in Figure 3.3. This is rather unexpected given that the urban sector is generally marked by lower agricultural activities, making urban residents less sensitive to local rainfall conditions compared to their rural counterparts. However, and perhaps reassuringly, estimated coefficients are statistically non-significant for the rest of the agricultural year. On the other hand, we find a positive, significant and persistent effect of the rainy season quality at urban destinations, from the first-half of October and until the month of June the following year. The result is again quite surprising: it is indicative of urban areas being relatively more attractive to temporary migrants following a relatively good rainy season. We do not have a clear-cut explanation for this finding but we propose at least two assumptions. The first one somehow relates to a measurement problem. Most of the cells classified as urban strictly include the city extent and also encompass the outskirts of that city where agricultural land may in fact dominate – this is especially true of smaller, secondary cities. In short, we do not exclude the possibility that a fraction of temporary migrants seen at urban destinations actually find employment in the agricultural sector. The second assumption has to do with home and work location choices of temporary migrants. The algorithm used to determine daily locations is primarily based on observations at night so that migration destinations are representative of where individuals stay, but not necessarily where they work. It may well be that temporary migrants spend nights in a city where they more likely have connections, and commute to some locations at the periphery during the day to supply labor in the agricultural sector. Future work could examine those mechanisms more closely.

Figure 3.5: Elasticity of the bilateral migration stock over time with respect to conditions at origin and destination, urban locations.



Note: Each point estimate represents an average elasticity of the bilateral out-migration stock with respect to precipitations at origin (light dots) and destination (dark dots) in urban locations. All coefficients are obtained with a PPML estimation in a single regression of the bilateral stock of temporary migrants between an origin location and a destination, against the logged precipitations at origin interacted with an urban origin dummy and a half-month indicator, and the logged precipitations at destination interacted with an urban destination dummy and a half-month indicator. This graphs shows the coefficients associated with urban origin and urban destination dummy interaction values equal to 1 (i.e. urban). Temporary migration bilateral rates are based on migration events of at least 20 days. Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

We perform a variety of robustness checks. First, we replicate the results of Figure 3.3 and Figure 3.5 considering higher minimum duration thresholds for the definition of temporary migration events: 30, 60 and 90 days, respectively. We show the results in Figures 3.C.3-3.C.8 in Appendix 3.C.3. The patterns of results remain largely consistent across the board, which suggests that the main results in Figures 3.3 and 3.5 are not driven by migration events of very short duration.

Second, the movements between pairs of adjacent locations may be subject to measurement errors. Users residing close to the border between two locations can be erroneously identified has having moved although calls were simply re-routed by the network from one antenna to the other in order to load balance. Consequently, we re-estimate equation (3.14) excluding all adjacent pairs of locations from the dataset. Assuming that identified movements between adjacent pairs of locations do occur in reality, this estimation concurrently allows to verify whether our results are driven by short-haul movements. We show the results in Figures 3.C.9-3.C.10 in

Appendix 3.C.4 and find practically no difference with the results obtained when adjacent cells are included.[19]

Third, PPML estimators of the parameters of interest, $\beta_{1,k}$ and $\beta_{2,k}$, correspond to an average elasticity of the bilateral out-migration stock across all origin-destination pairs in the sample. Since all pairs of locations enter the estimation with the same weight, the estimated effect could actually differ from the overall population average effect because origin locations vary in population size. For instance, the population average effect could be non-significant in a situation where our results are exclusively driven by the most sparsely populated cells. We provide evidence against this hypothesis by estimating our model with heterogeneous effects across groups of origin-destination pairs with different population size at origin. Figure 3.C.12 in Appendix 3.C.5 shows the results of the estimation on rural locations, considering five groups of origin locations of equal size based on population quintiles. Reassuringly, each rural group exhibits a pattern of results that is comparable with the one obtained in Figure 3.3. In any case, there is no indication that the magnitude of the estimated coefficients correlates with the population at origin.

Lastly, in Appendix 3.C.6, we provide a complementary analysis investigating the existence of non-linearities in the relationship between precipitations at origin and destination rural locations, and the bilateral out-migration rate. In particular, we consider the possibility that excess rainfall may have a negative impact on the agricultural sector, which could distort the interpretation of the linear models estimated above. More generally, we check whether estimated effects could be particularly driven by precipitation values at the lower or higher end of the distribution. We find that conditions of excess rainfall are a strong determinant for the estimated positive effect of precipitations at origin on temporary migration during the harvest period, compared to drought conditions. In short, considering the median precipitations as a reference scenario, excess precipitations at origin lead to a large increase in out-migration, whereas the decrease in out-migration induced by a rainfall deficit of the equivalent magnitude is comparatively small. We find no evidence of non-linearities in the effect of rainy season conditions at origin on migration during the off-season. On the other hand, the positive effect of precipitations at destination on the bilateral migration rate during harvest observed in Figure 3.3 seems to be majorly driven by drought conditions. Rainfall deficits at a rural destination are associated with a decrease in out-migration rate to that destination whereas excess precipitations do not imply significant differences in migratory movements compared to a scenario of normal conditions. Similar to the

[19] One exception is worth noting: the magnitude of the positive elasticity of temporary migration during the off-season with respect to precipitations at rural destinations is nearly halved.

effect of conditions at origin, we do not find evidence of non-linearities in the effect of rainy season conditions at destination on the bilateral migration rate during the off-season.

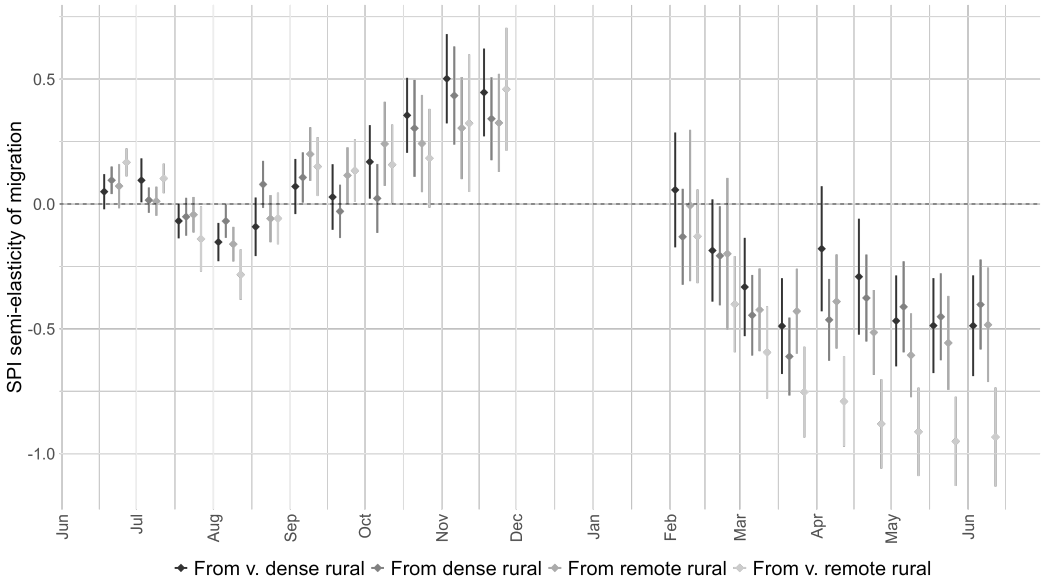
3.4.4 Heterogeneity across rural locations

In this section, we exclusively focus on rural areas and we further explore the heterogeneity of the effect of rainy season conditions at origin and destination across different groups of rural locations.

The literature suggests that poverty may well play a key role in prompting migration movements. For instance, Findley (1994) found that the 1983-1985 drought episode in Mali induced an increase in short-cycle migrants from poorer families characterized by lower incomes and fewer remittances. Indeed, poorer households that are closer to the subsistence level and with fewer coping strategies available to them could be more likely to resort to migration in times of hardship. Although our mobile phone data do not provide socio-economic information about users, we manage to explore this idea in our context by using secondary data sources allowing to characterize the large set of rural locations in our sample.

First, we overlay our rural cells with the 100m-resolution gridded population product from the WorldPop Research Group (Qader et al., 2022) to estimate the population density of each location. We use population density as a crude proxy measure of the local standards of living and we investigate whether the effect of rainy season conditions at origin is indeed more pronounced in rural locations of lower density. We divide the set of rural locations into four groups of equal size, ranging from the set of the most sparsely populated rural cells to the most densely populated cells. We estimate equation (3.14) interacting the rainy season conditions variable $x_{o,s,y}$ with a rural group indicator. Results are presented in Figure 3.6. The pattern of results is similar across the four groups and no clear difference can be observed for the first part of the agricultural year (June-December). However, and in line with the assumption outlined above, the negative effect during the off-season is clearly more pronounced for the group of least densely populated cells from the second half of March onward.

Figure 3.6: Elasticity of the bilateral migration stock over time with respect to conditions at origin, rural locations, by population density.

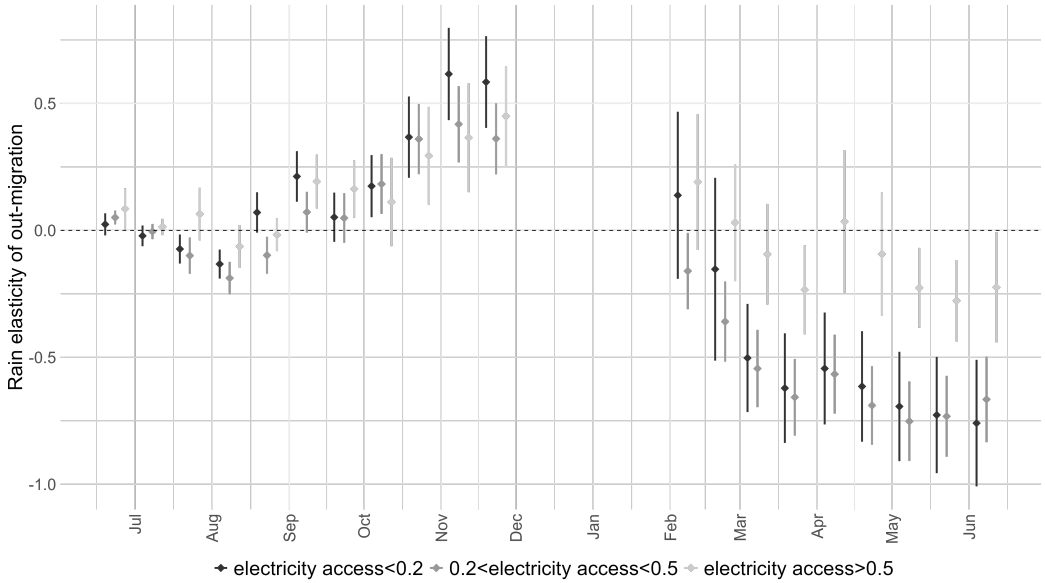


Note: Each point estimate represents an average elasticity of the bilateral out-migration stock with respect to precipitations at origin. Rural cells are overlaid with the 2017 100m-resolution gridded population product from the WorldPop Research Group (Qader et al., 2022) to determine the cell-level population density. Cells are grouped by population density quartile, e.g. the first quartile represents the 25% least densely populated rural cells. The effect of conditions at origin is allowed to vary across these groups and estimated coefficients for each group are represented on the graph with a distinct color. Lighter colors correspond to groups of cells with a lower population density. Temporary migration rates are based on migration events of at least 20 days. Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

To confirm this result, we also consider an alternative, survey-based, proxy measure of the local standards of living. We estimate the local rate of access to electricity based on a 10% extract of the 2013 census retrieved from the Integrated Public Use Microdata Series (IPUMS) database (Minnesota Population Center, 2020). We form three groups of rural cells based on the fraction of households which have access to electricity: less than 20%, between 20% and 50%, and above 50%.[20] We again estimate the same regression interacting the rainy season conditions variable $x_{o,s,y}$ with a categorical variable indicating the group of electricity access. The results are shown in Figure 3.7 and corroborate the findings based on population density. The effect of rainy season conditions at origin on the bilateral temporary migration stock is markedly more pronounced for the groups of rural cells with a lower access to electricity.

[20]These groups represent 45%, 45% and 10% of the total set of rural cells. A map of the estimated electricity rate by third-level administrative unit is provided in Figure 3.B.6.

Figure 3.7: Elasticity of the bilateral migration stock over time with respect to conditions at origin, rural locations, by level of access to electricity.



Note: Each point estimate represents an average elasticity of the bilateral out-migration stock with respect to precipitations at origin. Rural cells are categorized in three groups based on the fraction of households with access to electricity: less than 20%, between 20% and 50%, and above 50%. The effect of conditions at origin is allowed to vary across these groups and estimated coefficients for each group are represented on the graph with a distinct color. Lighter colors correspond to groups of cells with a higher rate of electricity access. Temporary migration rates are based on migration events of at least 20 days. Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

3.5 Conclusion

In this study, we develop a simple model of temporary migration in the context of year-on-year variations in the quality of rainy seasons. The simplified version of the model points to a simple intuition. Relatively poorer conditions at origin decrease local wages and act as a push factor. On the other hand, poorer conditions at destination decrease wages at destination and affect the value of that location, thus creating a disincentive to temporarily migrate to that destination.

By combining granular temporary migration estimates derived from a multi-year mobile phone dataset with satellite-based measures of the quality of the rainy season, we are able to identify and characterize the effect of rainy season conditions at origin and destination on the bilateral temporary migration rate. Our results depict a rather complex story that goes beyond a simple narrative where climate stresses would increase or decrease temporary migration. Our regression estimations clearly reveal that both rainy season conditions at origin and destination

contribute to shaping temporary migration decisions between any two locations. First, precipitations levels at a rural destination over a rainy season have a positive effect on bilateral stocks of temporary migrants headed that destination, both during the harvest season and the off-season. In other terms, a negative rainfall shock at a destination diminishes its attractiveness to potential temporary migrants for the remaining agricultural year, consistent with the predictions of our model. However, rainfall during the rainy season at the origin exhibits divergent effect across the harvest season and the off-season. In the immediate aftermath of the negative rainfall shock, i.e. during the rainy season, individuals show a lower propensity to out-migrate (factoring in conditions at destination). Yet, in the subsequent off-season, notably between March and June, they exhibit a higher inclination to out-migrate. Interestingly, this effect is entirely driven by rural-to-rural temporary migration dynamics, and we do not find evidence supporting the idea that affected rural households would react to a negative income shock by sending temporary migrants to cities during the off-season. Lastly, the heterogeneity analysis tends to show that the result is particularly driven by locations with lower standards of living, and arguably higher levels of poverty.

The paper undeniably sheds new light on the temporary migration responses to the year-on-year variability in rainfall conditions. However, along with these findings arise several questions, hopefully setting the stage for further research endeavors. Firstly, the contrasting impact of conditions at the origin during the harvest and off-season remains puzzling. One plausible explanation is the presence of a liquidity constraint hindering temporary migration immediately following a poor rainy season. Under this premise, individuals might need several months to gather the necessary resources enabling them to undertake migration during the subsequent off-season. Moreover, the fact that the off-season response is predominantly steered by rural-to-rural movements is also quite surprising. One would not typically anticipate rural destinations to offer alternative employment prospects, particularly during that specific period. Future research could investigate the motivations behind those movements. A plausible hypothesis is that these rural-to-rural migrants might be reaching out for short-term assistance from kin in neighboring villages, thereby alleviating the burden on their originating households and leveraging communal support and resource sharing at their destinations. Note that this idea could be further explored with the CDR data, by examining whether these migrants specifically move to locations in times of hardship where they already maintain a robust social network. Finally, while our empirical analysis provides average elasticities of bilateral temporary migration with respect to rainfall conditions at origin and destination, it does not quantify the overarching net aggregate impact of rainy season conditions. We leave such quantification

exercise to future work.

Appendix 3.A Model simplification

First, the dependence of $Pr(m_{i,t} = m_j|m_{i,t-1})$ on the previous location $m_{i,t-1}$ is induced by mobility frictions, which are governed by $\tau_{m_{i,t-1},m_t}$. We assume that distance between origin and destination does not play a role ($\tau_{m_{i,t-1},m_t} = 1$). While including this feature is relevant for the prediction of bilateral flows at the equilibrium, it plays only a limited role in determining the response of temporary migration to inter-year variations in local precipitations (Monras, 2018).

$$\begin{aligned}\ln Pr(m_{i,t} = m_j) = \frac{1}{\bar{\theta}}\left[\ln\tilde{\Gamma}_{m_j,s(t)} + \alpha_{m_j}\ln K_{m_j,t} - \alpha_{m_j}\ln N_{m_j,t}\right] \\ - \ln\left[\left(\tilde{\Gamma}_{h,s(t)}\left(\frac{K_{h,t}}{N_{h,t}}\right)^{\alpha_h}\right)^{1/\theta} + \left(\sum_{k\in\overline{\mathcal{M}^i}}\tilde{\Gamma}_{m_k,s(t)}\left(\frac{K_{m_k,t}}{N_{m_k,t}}\right)^{\alpha_{m_k}}\right)^{\bar{\theta}/\theta}\right] \\ + \left(\frac{\bar{\theta}}{\theta} - 1\right)\ln\left[\sum_{k\in\overline{\mathcal{M}^i}}\left(\tilde{\Gamma}_{m_k,s(t)}\left(\frac{K_{m_k,t}}{N_{m_k,t}}\right)^{\alpha_{m_k}}\right)^{1/\bar{\theta}}\right]\end{aligned} \tag{3.15}$$

Where $\tilde{\Gamma}_{m,s(t)} = A_m\Gamma_{m,s(t)}$.

Second, we assume that $\frac{1}{\theta} \approx 0$[21] which allows to simplify the second term of (3.15):

$$\begin{aligned}\ln Pr(m_{i,t} = m_j) = \frac{1}{\bar{\theta}}\left[\ln\tilde{\Gamma}_{m_j,s(t)} + \alpha_{m_j}\ln K_{m_j,t} - \alpha_{m_j}\ln N_{m_j,t}\right] \\ - \frac{1}{\theta}\left[\ln\tilde{\Gamma}_{h,s(t)} + \alpha_h\ln K_{h,t} - \alpha_h\ln N_{h,t}\right] - \ln 2 \\ + \left(\frac{\bar{\theta}}{\theta} - 1\right)\ln\left[\sum_{k\in\overline{\mathcal{M}^i}}\left(\tilde{\Gamma}_{m_k,s(t)}\left(\frac{K_{m_k,t}}{N_{m_k,t}}\right)^{\alpha_{m_k}}\right)^{1/\bar{\theta}}\right]\end{aligned} \tag{3.16}$$

Then, the third term imply second-order effects of the values of non-home locations on the probability to be in m_j at time t and is assumed constant, so that equation (3.16) becomes:

$$\ln Pr(m_{i,t} = m_j) = \frac{1}{\bar{\theta}}\left[\ln\tilde{\Gamma}_{m_j,s(t)} + \alpha_{m_j}\ln K_{m_j,t} - \alpha_{m_j}\ln N_{m_j,t}\right] - \frac{1}{\theta}\left[\ln\tilde{\Gamma}_{h,s(t)} + \alpha_h\ln K_{h,t} - \alpha_h\ln N_{h,t}\right] + C \tag{3.17}$$

Where $C = \left(\frac{\bar{\theta}}{\theta} - 1\right)\ln\left[\sum_{k\in\overline{\mathcal{M}^i}}\left(\tilde{\Gamma}_{m_k,s(t)}\left(\frac{K_{m_k,t}}{N_{m_k,t}}\right)^{\alpha_{m_k}}\right)^{1/\bar{\theta}}\right] - ln2$.

[21] This hypothesis is verified in Monras (2018) in a setting similar to ours.

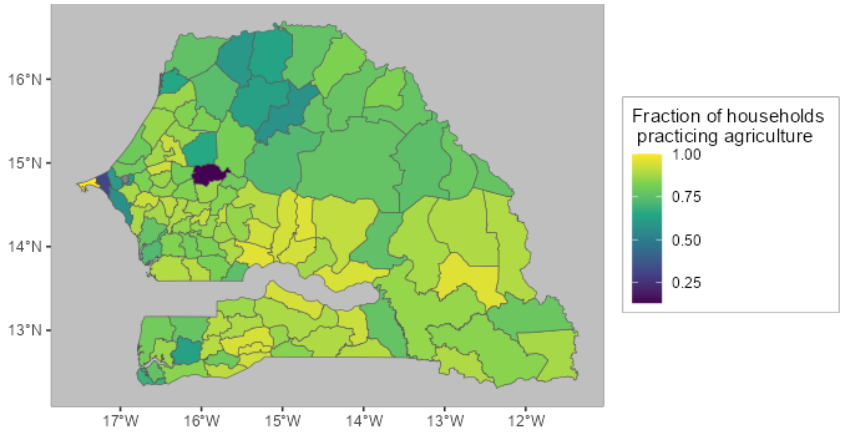
In the simplified version of the model presented in this paper, we ignore congestion effects on the labor market at origin and destination, i.e. $\frac{\partial \ln Pr(m_{i,t}=m_j)}{\partial \ln N_{m_j,t}} \approx 0$ and $\frac{\partial \ln Pr(m_{i,t}=m_j)}{\partial \ln N_{h,t}} \approx 0$. We thus obtain the following simplified reduced form:

$$\ln Pr(m_{i,t} = m_j) = \frac{\alpha_{m_j}}{\bar{\theta}} \ln K_{m_j,t} - \frac{\alpha_h}{\theta} \ln K_{h,t} + \frac{1}{\bar{\theta}} \ln \tilde{\Gamma}_{m_j,s(t)} - \frac{1}{\theta} \ln \tilde{\Gamma}_{h,s(t)} + C \quad (3.18)$$

Appendix 3.B Spatial distribution of agricultural activities

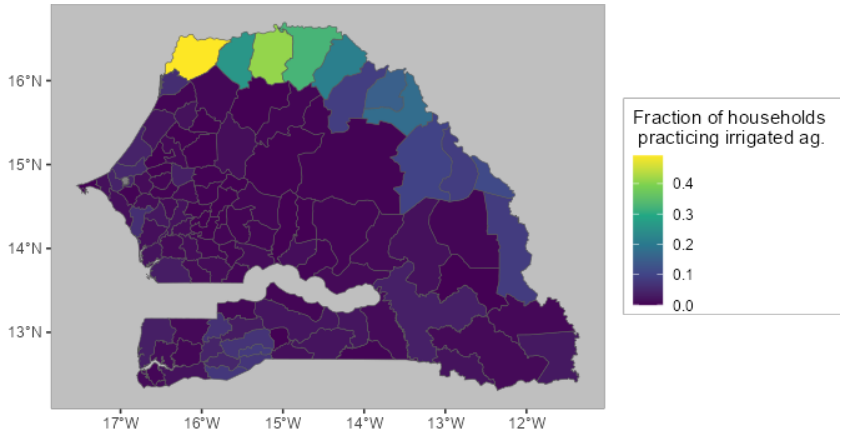
All the maps below are based on a 10% extract of the 2013 census in Senegal retrieved from the Integrated Public Use Microdata Series (IPUMS) database (Minnesota Population Center, 2020). The authors wish to acknowledge the statistical office that provided the underlying data making this research possible: National Agency of Statistics and Demography, Senegal.

Figure 3.B.1: Fraction of households with at least one member practicing agriculture, by arrondissement.



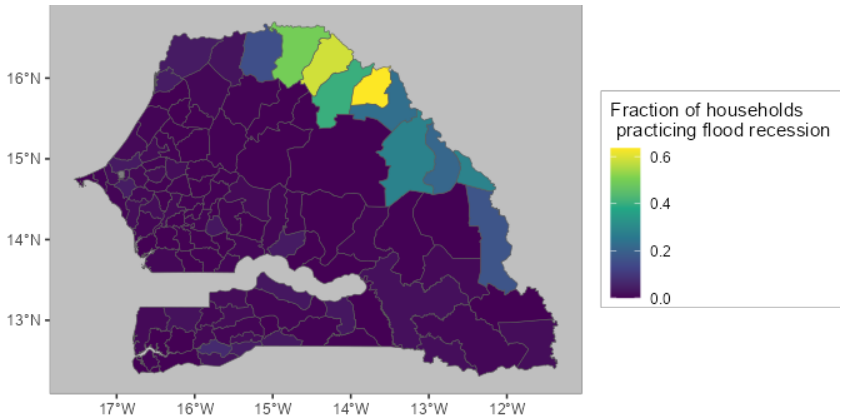
Note: Estimates are based on a 10% extract of the 2013 census retrieved from the Integrated Public Use Microdata Series (IPUMS) database.

Figure 3.B.2: Fraction of households with at least one member practicing irrigated agriculture, by arrondissement.



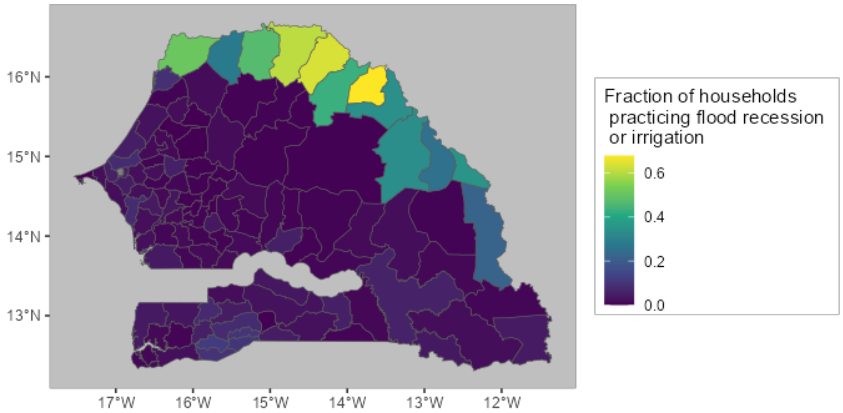
Note: Estimates are based on a 10% extract of the 2013 census retrieved from the Integrated Public Use Microdata Series (IPUMS) database.

Figure 3.B.3: Fraction of households with at least one member practicing flood recession agriculture, by arrondissement.



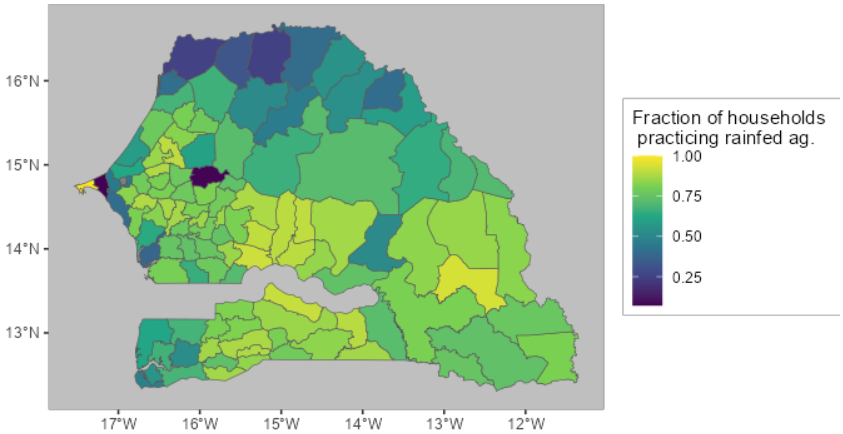
Note: Estimates are based on a 10% extract of the 2013 census retrieved from the Integrated Public Use Microdata Series (IPUMS) database.

Figure 3.B.4: Fraction of households with at least one member practicing irrigated or flood recession agriculture, by arrondissement.



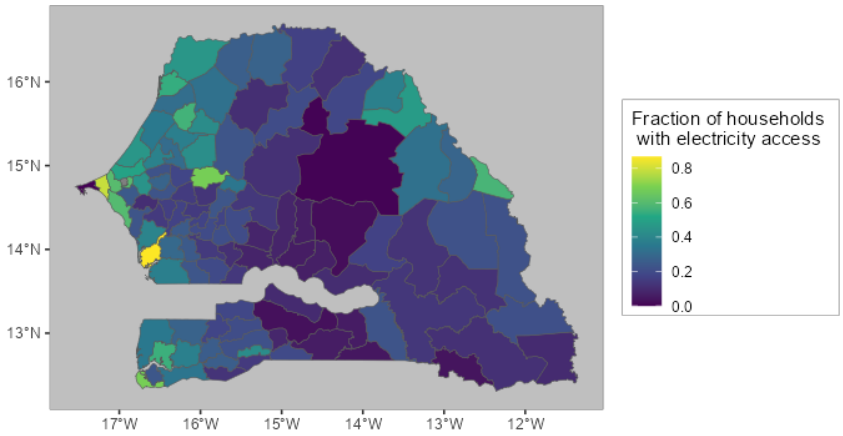
Note: Estimates are based on a 10% extract of the 2013 census retrieved from the Integrated Public Use Microdata Series (IPUMS) database.

Figure 3.B.5: Fraction of households with at least one member practicing rainfed agriculture, by arrondissement.



Note: Estimates are based on a 10% extract of the 2013 census retrieved from the Integrated Public Use Microdata Series (IPUMS) database.

Figure 3.B.6: Fraction of households with electricity access, by arrondissement.

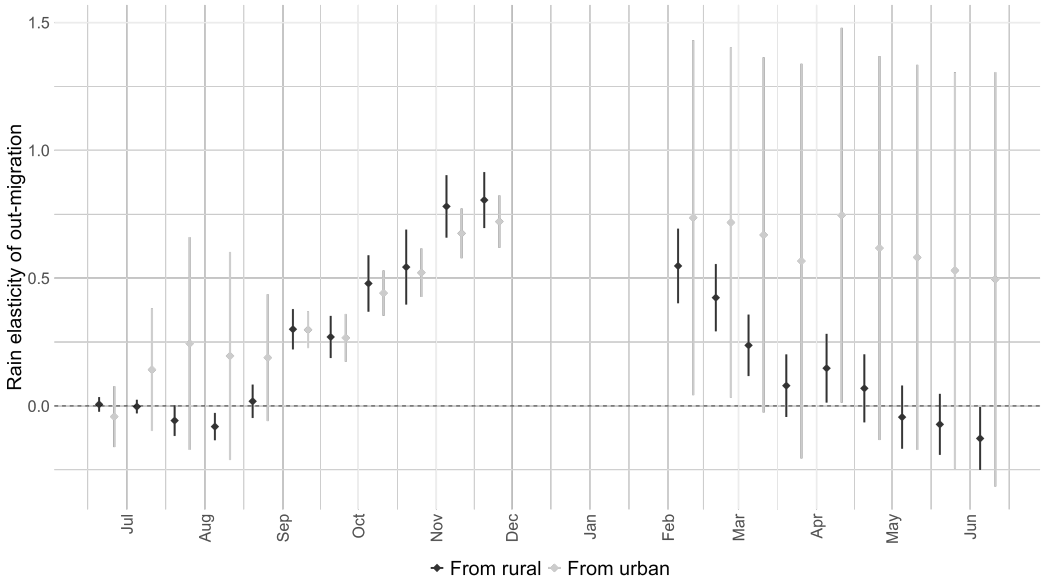


Note: Estimates are based on a 10% extract of the 2013 census retrieved from the Integrated Public Use Microdata Series (IPUMS) database.

Appendix 3.C The effect of rainy season conditions on temporary migration: additional results

3.C.1 Conventional migration equation, effect by zone of origin

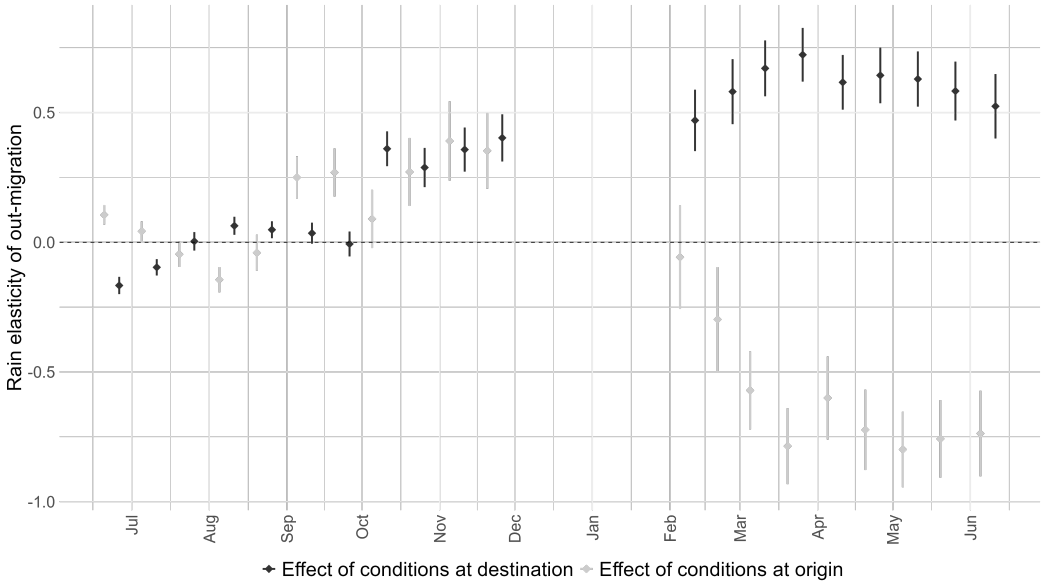
Figure 3.C.1: Elasticity of out-migration estimated by half-month over the agricultural year, by zone of origin.



Note: Each point estimate represents the average elasticity of out-migration with respect to precipitations at origin. The set of estimated elasticities for the subsets of rural and urban cells are shown in different colors. Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*half-month)-level.

3.C.2 Dyadic regression estimations with origin and destination fixed effects, controlling for distance

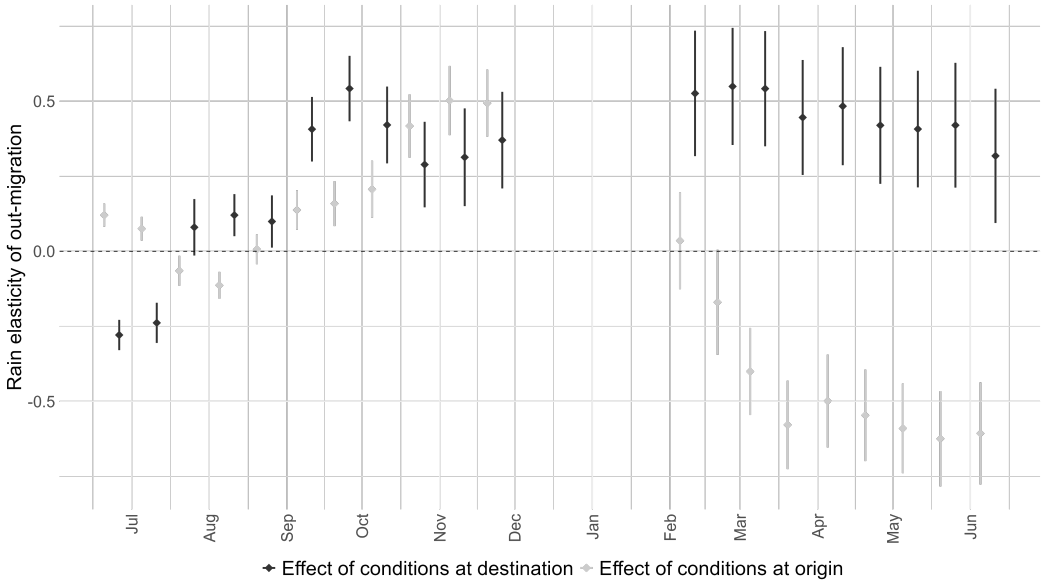
Figure 3.C.2: Elasticity of the bilateral migration stock over time with respect to conditions at origin and destination, rural locations.



Note: Each point estimate represents an average elasticity of the bilateral out-migration stock with respect to precipitations at origin (light dots) and destination (dark dots) in rural locations. All coefficients are obtained with a PPML estimation in a single regression of the bilateral stock of temporary migrants between an origin location and a destination, against the logged precipitations at origin interacted with an urban origin dummy and a half-month indicator, and the logged precipitations at destination interacted with an urban destination dummy and a half-month indicator. Contrary to results showed in Figure 3.3, this specification considers separate fixed effect terms at the origin- and destination-level, and concurrently controls for the travel time between origin and destination. This graph shows the coefficients associated with urban origin and urban destination dummy interaction values equal to 0 (i.e. rural). Temporary migration bilateral rates are based on migration events of at least 20 days. Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

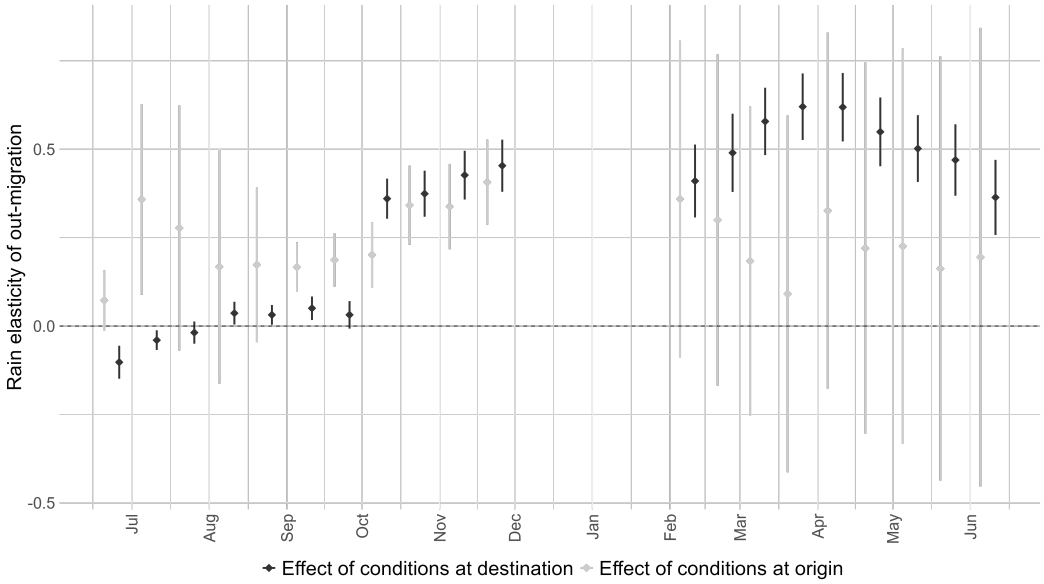
3.C.3 Dyadic regression estimations with different migration duration thresholds

Figure 3.C.3: Elasticity of the bilateral migration stock over time with respect to conditions at origin and destination, rural locations, migration events of at least 30 days.



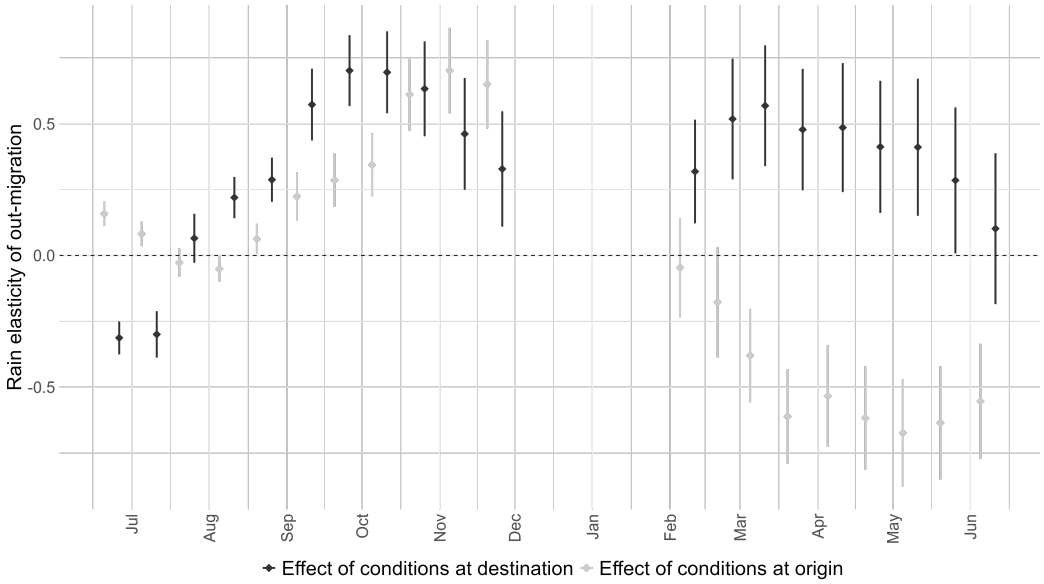
Note: Each point estimate represents the average elasticity of the bilateral out-migration stock with respect to precipitations at origin (light dots) and destination (dark dots). Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

Figure 3.C.4: Elasticity of the bilateral migration stock over time with respect to conditions at origin and destination, urban locations, migration events of at least 30 days.



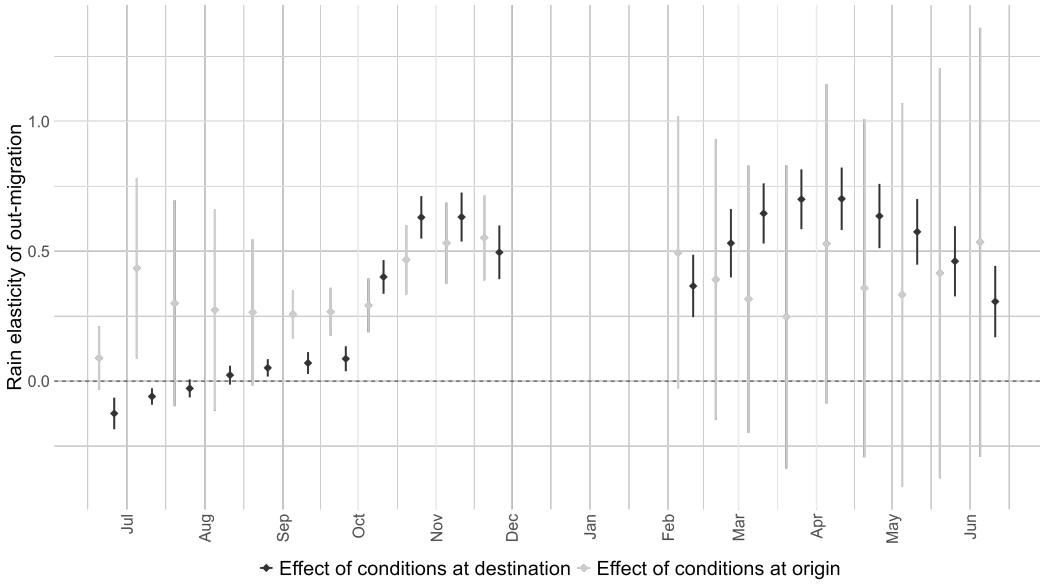
Note: Each point estimate represents the average elasticity of the bilateral out-migration stock with respect to precipitations at origin (light dots) and destination (dark dots). Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

Figure 3.C.5: Elasticity of the bilateral migration stock over time with respect to conditions at origin and destination, rural locations, migration events of at least 60 days.



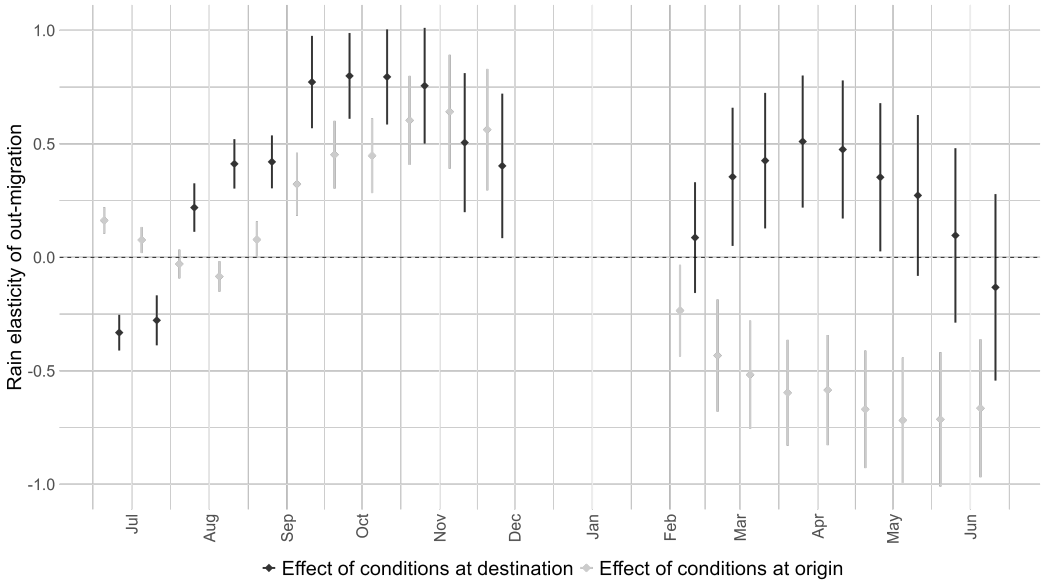
Note: Each point estimate represents the average elasticity of the bilateral out-migration stock with respect to precipitations at origin (light dots) and destination (dark dots). Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

Figure 3.C.6: Elasticity of the bilateral migration stock over time with respect to conditions at origin and destination, urban locations, migration events of at least 60 days.



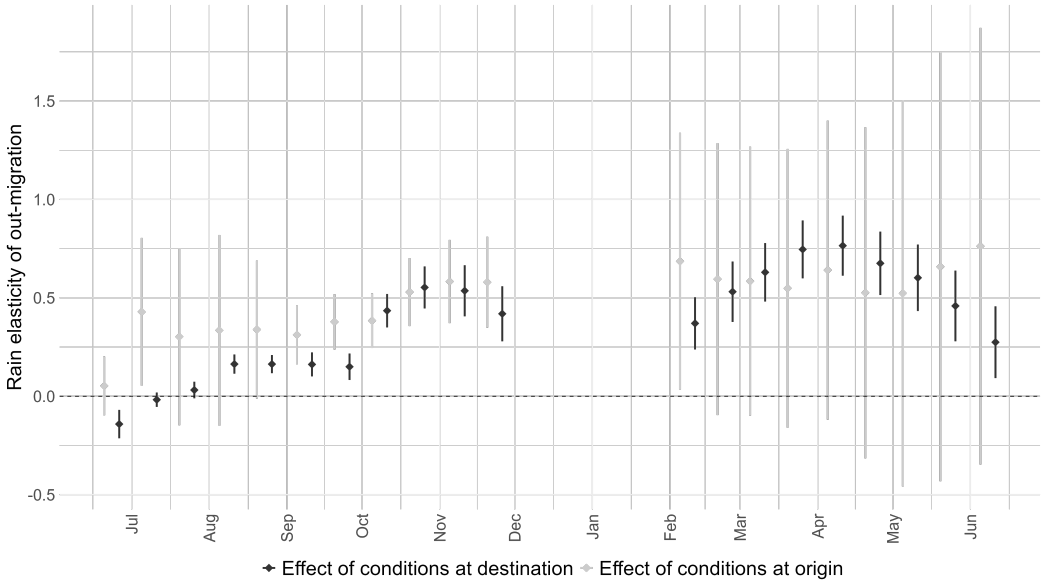
Note: Each point estimate represents the average elasticity of the bilateral out-migration stock with respect to precipitations at origin (light dots) and destination (dark dots). Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

Figure 3.C.7: Elasticity of the bilateral migration stock over time with respect to conditions at origin and destination, rural locations, migration events of at least 90 days.



Note: Each point estimate represents the average elasticity of the bilateral out-migration stock with respect to precipitations at origin (light dots) and destination (dark dots). Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

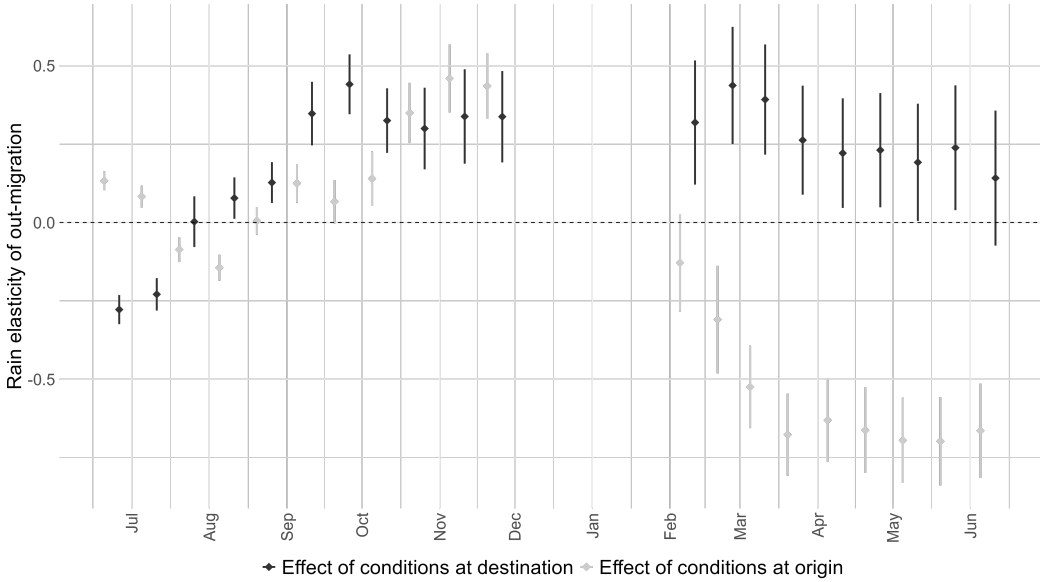
Figure 3.C.8: Elasticity of the bilateral migration stock over time with respect to conditions at origin and destination, urban locations, migration events of at least 30 days.



Note: Each point estimate represents the average elasticity of the bilateral out-migration stock with respect to precipitations at origin (light dots) and destination (dark dots). Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

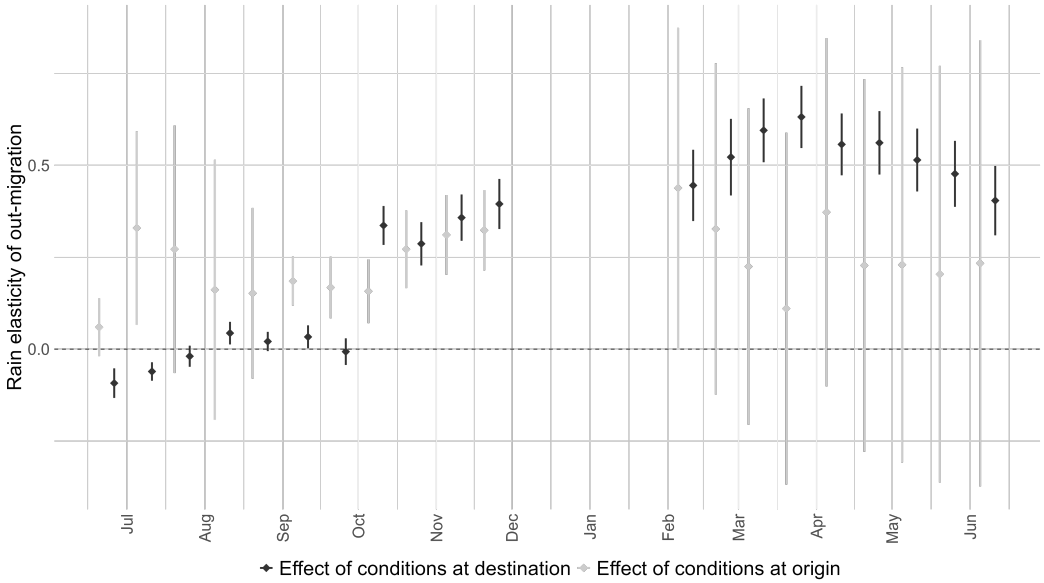
3.C.4 Dyadic regression estimations, excluding pairs of adjacent cells

Figure 3.C.9: Elasticity of the bilateral migration stock over time with respect to conditions at origin and destination, rural locations, excluding pairs of adjacent cells.



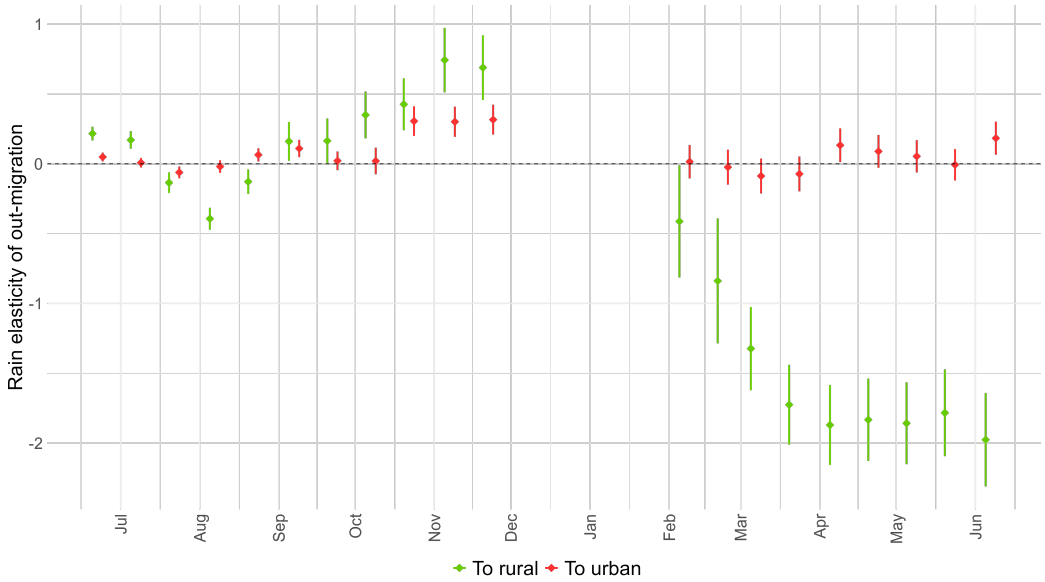
Note: Each point estimate represents the average elasticity of the bilateral out-migration stock with respect to precipitations at origin (light dots) and destination (dark dots). A total of 5,998 pairs of adjacent cells are excluded from the dataset. Temporary migration rates are based on migration events of at least 20 days. Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

Figure 3.C.10: Elasticity of the bilateral migration stock over time with respect to conditions at origin and destination, urban locations, excluding pairs of adjacent cells.



Note: Each point estimate represents the average elasticity of the bilateral out-migration stock with respect to precipitations at origin (light dots) and destination (dark dots). A total of 5,998 pairs of adjacent cells are excluded from the dataset. Temporary migration rates are based on migration events of at least 20 days. Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

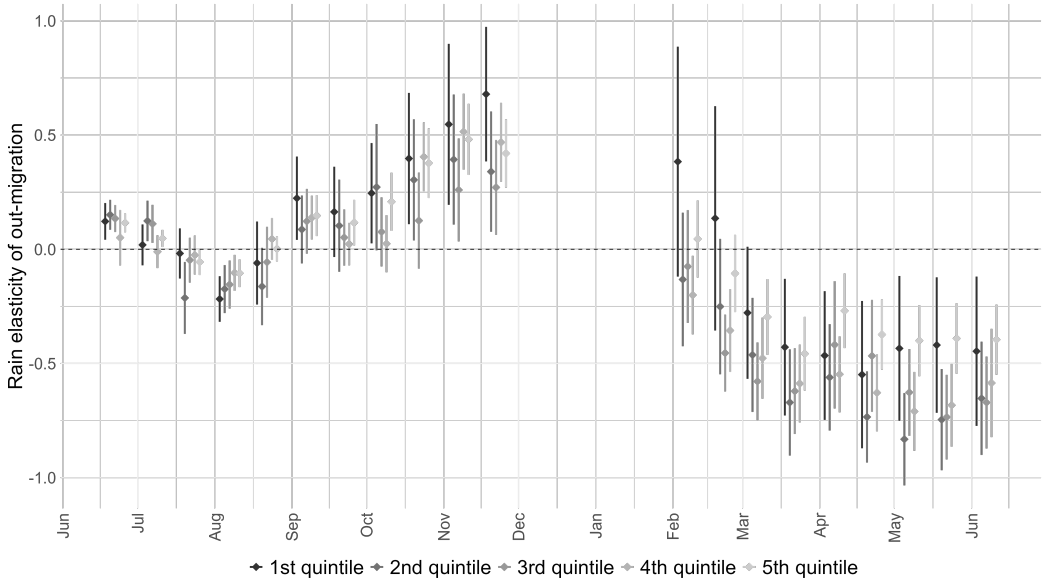
Figure 3.C.11: Elasticity of the bilateral migration stock over time with respect to conditions at origin and destination, urban locations, excluding pairs of adjacent cells.



Note: Each point estimate represents an average elasticity of the bilateral out-migration stock with respect to precipitations at a rural origin. Red dots represent the effect on the movements toward urban locations, whereas green dots show the effect on temporary migration to other rural areas. All coefficients are obtained with a PPML estimation in a single regression of the bilateral stock of temporary migrants between an origin location and a destination, against the logged precipitations at origin interacted with an urban origin dummy, an urban destination dummy, and a half-month indicator, and the logged precipitations at destination interacted with an urban destination dummy and a half-month indicator. This graph shows the coefficients associated with the urban origin dummy interaction value equal to 0 (i.e. rural). A total of 5,998 pairs of adjacent cells are excluded from the dataset. Temporary migration bilateral rates are based on migration events of at least 20 days. Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level..

3.C.5 Dyadic regression estimation with heterogeneity by cell population size

Figure 3.C.12: Elasticity of the bilateral migration stock over time with respect to conditions at origin, rural locations, by cell population size.



Note: Each point estimate represents an average elasticity of the bilateral out-migration stock with respect to precipitations at origin. Rural cells are overlaid with the 2017 100m-resolution gridded population product from the WorldPop Research Group (Qader et al., 2022) to determine the cell-level population. Cells are grouped by population quintile, e.g. the first quintile represents the 20% least populated rural cells. The effect of conditions at origin is allowed to vary across these groups and estimated coefficients for each group are represented on the graph with a distinct color. Darker colors correspond to groups of cells with a lower population. Temporary migration rates are based on migration events of at least 20 days. Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

3.C.6 Dyadic regression estimation, non-linearities

As a complement to the results presented in Figure 3.3, we consider an alternative specification that allows to investigate the existence of non-linearities in the relationships between the bilateral out-migration rate between an origin and a destination location, and the rainfall conditions at origin and destination respectively. As explained in footnote 14, an alternative definition for the rainy season variables $x_{o,s,y}$ and $x_{d,s,y}$ in equation (3.14) can use the anomaly of precipitations as defined by the Standardized Precipitation Index (see section 3.3.2). The main advantage of the SPI is that it provides measures of precipitation anomalies that are comparable across locations characterized by distinct rainfall regimes. As a result, the SPI represents a relative measure of local precipitations and can thus enter linearly in the definition of $x_{o,s,y}$, i.e. in a non-logged form. Then, we can simply evaluate the

existence of non-linear patterns by considering discrete versions of $x_{o,s,y}$ and $x_{d,s,y}$ for the estimation of equation (3.14).

Therefore, we define $x_{o,s,y}$ and $x_{d,s,y}$ as SPI values over time windows consistent with those described in section 3.4.2. Positive values indicate above-median precipitation values and, conversely, negative values reflect a level of precipitations that is below the historical median for the location considered. We define a categorical variable $x^{cat}_{o,s,y}$ with five categories (and similarly for $x_{d,s,y}$):

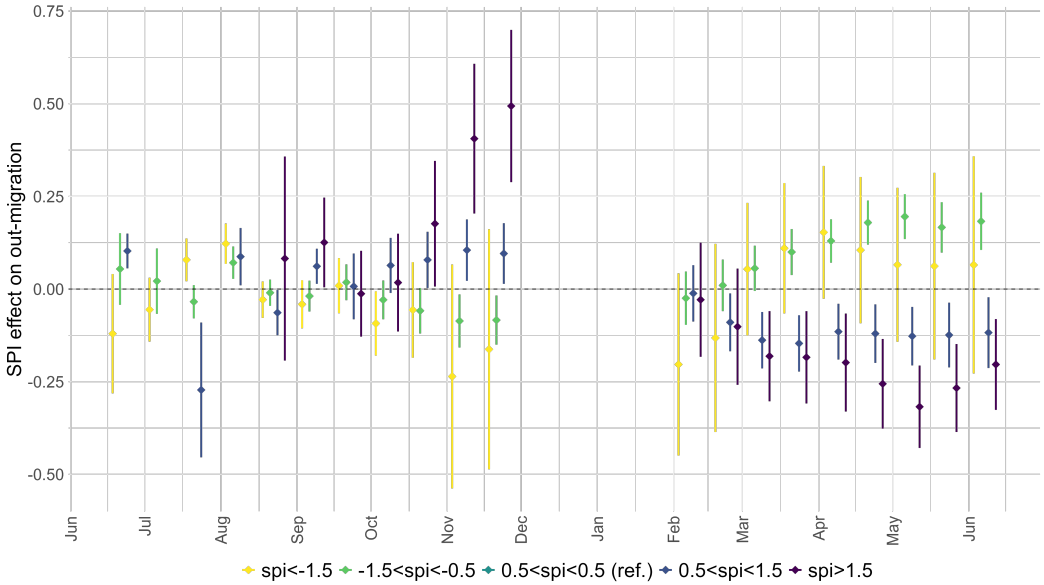
$$x^{cat}_{o,s,y} = \begin{cases} 1 & if \quad x_{o,s,y} < -1.5 \\ 2 & if \quad -1.5 \leq x_{o,s,y} \leq -0.5 \\ 3 & if \quad -0.5 \leq x_{o,s,y} \leq 0.5 \\ 4 & if \quad 0.5 \leq x_{o,s,y} \leq 1.5 \\ 5 & if \quad x_{o,s,y} > 1.5 \end{cases} \tag{3.19}$$

We estimate equation (3.14) with a PPML estimator, replacing $x_{o,s,y}$ and $x_{d,s,y}$ by the categorical variables $x^{cat}_{o,s,y}$ and $x^{cat}_{d,s,y}$ and taking the third category (i.e. precipitation conditions close to normal) as reference. We again focus on the effect of conditions at rural origin and rural destination locations. Figure 3.C.13 presents the estimated effects of conditions at origin. Each estimated coefficient represents the difference between the average outcome for SPI values in a given SPI category and the average outcome in normal conditions (i.e. $x^{cat}_{o,s,y} = 3$). During the harvest period, in October and November, estimated coefficients for the higher categories (4 and 5) are positive and those associated with the lower categories (1 and 2) are negative. This is consistent with the positive linear relationship found in Figure 3.3. However, the coefficients associated with the highest category ($spi > 1.5$) are close to 0.5 in November, and thus particularly large in absolute terms compared to coefficients associated with the lowest category which remain around -0.1. This indicates that the positive relationship observed in Figure 3.3 is particularly driven by above-median precipitations at origin acting as a push factor for out-migration, rather than drought conditions being an impeding factor. For the rest of the agricultural year, no clear evidence of non-linearities can be distinguished.

Similarly, Figure 3.C.14 show the estimated effects of rainy season conditions at destination on the bilateral stock of temporary migrants. Looking first at the harvest period, it seems clear that above-median precipitations do not lead to higher migration compared to a scenario of median precipitations (i.e. "normal conditions"). If anything, the coefficients associated with the highest SPI category are negative (but non-significant at 5% percent level) in November, indicating that excess rainfall could have a negative effect on the degree of attractiveness of rural locations. However, the positive relationship identified in 3.3 is clearly driven

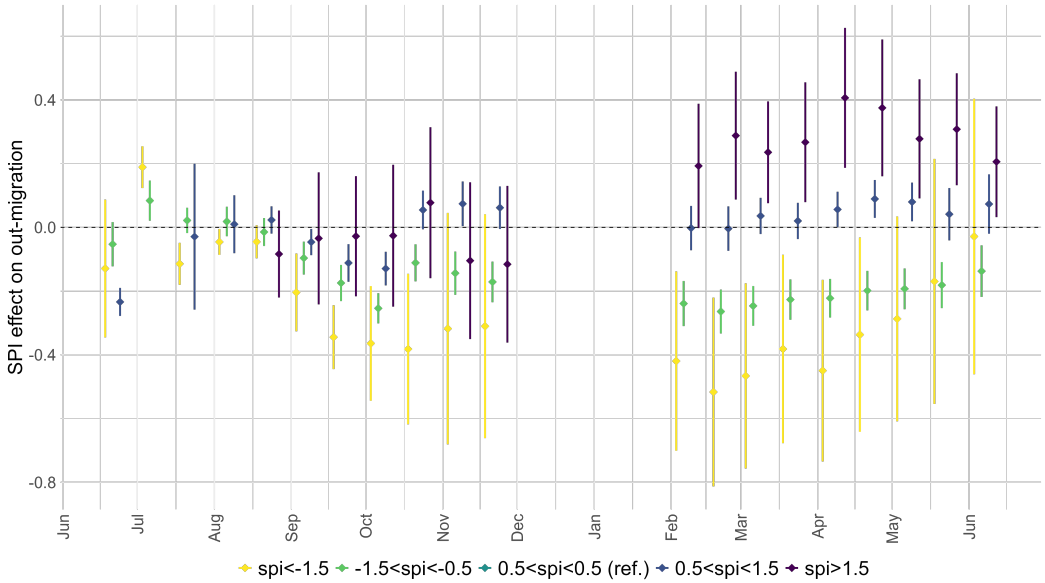
by rainfall deficits at a destination leading to lower bilateral rates of migration to that destination. The results do not show evidence of non-linear patterns for the February-to-June off-season period.

Figure 3.C.13: Effect of the categorized SPI at origin on the bilateral migration stock over time, rural locations.



Note: The dependent variable is the bilateral stock of temporary migrants between an origin and a destination at some half-month period. Each point estimate represents the difference between the average outcome for SPI values in a given SPI category compared to the average outcome in normal conditions (i.e. $x^{cat}_{o,s,y} = 3$). Each color represents a distinct SPI category, with darker colors corresponding to lower SPI values. Temporary migration rates are based on migration events of at least 20 days. Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

Figure 3.C.14: Effect of the categorized SPI at destination on the bilateral migration stock over time, rural locations.



Note: The dependent variable is the bilateral stock of temporary migrants between an origin and a destination at some half-month period. Each point estimate represents the difference between the average outcome for SPI values in a given SPI category compared to the average outcome in normal conditions (i.e. $x^{cat}_{o,s,y} = 3$). Each color represents a distinct SPI category, with darker colors corresponding to lower SPI values. Temporary migration rates are based on migration events of at least 20 days. Vertical bars represent 95% confidence intervals based on standard errors clustered at the (origin*destination*half-month)-level.

Chapter 4

Short visits and temporary migration to cities in Senegal

4.1 Introduction

In the developing world, individuals move across space and away from their primary residence location in pursuit of better employment prospects or access goods, services and amenities that their home location may not offer. Urban centers emerge as particularly enticing destinations, attributed to their diverse array of amenities, higher levels of productivity, and the well-documented advantages conferred by agglomeration economies (Duranton and Puga, 2004; Duranton, 2015; Henderson, 2010; Rosenthal and Strange, 2004). Permanent migration has received considerable scholarly attention: rural-to-urban movements are recognized as integral to structural transformation processes in developing countries (Beegle et al., 2011; Brauw et al., 2014; Young, 2013; Munshi and Rosenzweig, 2016; Hamory et al., 2020; Garriga et al., 2023). However, they have been found to be surprisingly low in view of the persistent sectoral and spatial gaps in productivity and wealth (Gollin, Lagakos, et al., 2014). On the other hand, non-permanent forms of mobility occupy a seminal position in delineating spatial equilibria, offering the possibility to access non-home markets without incurring the cost of a permanent relocation. Such movements are particularly common in developing countries, where they are often incorporated into livelihood strategies. For instance, temporary migration is typically used as a way to diversify income sources or respond to fluctuations in local productivity[1] (Bryan, Chowdhury, et al., 2014; Coffey et al., 2015; Morten, 2019; Imbert and Papp, 2020b,a). Moreover, recent advances in the use of big data have enabled researchers to uncover new and subtler patterns of human movements (Blumenstock, 2012; Williams et al., 2015; Demissie et al., 2019). In this respect, in Chapter 1, we leverage smartphone app location data in three

[1]We provide another example in Chapter 3, where we study temporary migration responses to variations in the quality of rainy seasons in Senegal.

African countries to highlight the prominence of short and relatively frequent "visits" to large agglomerations. Such movements distinctly diverge from both daily commuting patterns and migratory behaviors, but their exact nature and purpose are still largely unknown. In particular, the role of visits relative to other forms of non-permanent movements aiming at accessing urban markets is yet to be ascertained.

Building upon the novel perspectives on human mobility delineated in Chapter 1, this chapter further delves into the study of visits to cities and examines the distinctions and interconnections with other transient mobility patterns in Senegal, focusing specifically on temporary migration. To achieve this, I harness the potential of mobile phone metadata, specifically Call Detail Records (CDR). These records allow to reconstruct individuals' trajectories across space and over time at a highly granular level. The uniqueness of CDR data lies in its amalgamation of high-frequency observations with extensive activity periods. This allows for a concurrent examination of both visits and temporary migration decisions at an individual level, spanning multiple years, and encompassing a vast demographic of phone users on a national scale. I draw on Chapter 2 to construct user-level metrics of temporary migration movements to cities, characterizing such movements as continuous periods ranging from 20 to 180 days wherein a user is continuously seen at a location distinct from his primary residence. Then, I devise measures of visits to urban locales, conceptualized as concise temporal intervals, generally spanning a few days and not exceeding a 20-day duration, at locations other than the user's home location. The fine spatial resolution afforded by the CDR data facilitates the observation of movements spanning across over 900 locations in Senegal, including 39 cities of varied magnitudes.[2]

Transient mobility patterns can conceivably encapsulate a myriad of endeavors ranging from the consumption of goods, services or leisure activities, to the production processes, business operations, or even participation in cultural or religious events. The mobile phone data utilized in this research do not proffer a direct glimpse into the activities pursued by individuals when they either visit or engage in temporary migration to cities. Nonetheless, I postulate that a simultaneous examination of individual choices pertaining to visits and temporary migration can shed light on the comparative roles of these two mobility patterns. In this context, the diverse activities from which both visitors and migrants derive utility in urban settings are collectively designated under the umbrella term of "amenity consumption".

[2]See Appendix 2.A in Chapter 2 for more details on the delineation of locations and urban locales from Base Transceiver Stations (BTS).

Admittedly, this term may appear broad in its scope,[3] yet it offers a pertinent concept to elucidate the observed interplay between visits and temporary migration to cities.

One possibility is that the motivations behind visits and temporary migration are fundamentally distinct, so that the demand for visits is simply unrelated to the demand for longer-term temporary movements. In short, amenities accessed through visits might be mutually exclusive with those sought via temporary migration. However, the existence of some overlap in visit-based and migration-based amenities would raise the possibility that visits represent a hitherto disregarded alternative technology to longer-term movements. The choice of one mode of transient mobility over another could subsequently hinge on differentials in the encompassing costs associated with each movement. Conversely, visits and temporary migrations might operate synergistically if the underlying amenities they seek to access are complementary. For instance, visits could be a precursor to migration, enabling individuals to acquire insights regarding potential employment opportunities and facilitate requisite preparations ahead of a more sustained migration.

The detection of temporary migration episodes in CDR trajectories is based on the methodology developed in Chapter 2. A segment-based approach akin to a clustering procedure allows to first determine a home location for each user, and then to identify continuous periods of at least 20 days during which a user is persistently observed at a non-home location. Such time intervals denote instances of temporary migration episodes. On the other hand, visits are defined as short stays, not exceeding 19 days, at a location that differs from the current place of residence.[4] In practice, I use a simple frequency-based method[5] to determine the trajectory of successive daytime and nighttime locations for each user. Then, I identify clusters of consecutive observations at a single location – which I call "micro-segments" – allowing for some observational gaps.[6] I define visits as micro-segments of at most 19 days at a location that differs from the current place of residence.[7] I apply

[3]Moreover, aggregating heterogeneous activities that may or may not have a direct monetary valuation is notoriously challenging (Su, 2022).

[4]The current place of residence can either be the primary home location or a temporary residence location if a user is in migration at some destination. This means that the primary home location can be a visit destination if a user is seen visiting his home location for a short period of time while in migration.

[5]A frequency-based method defines a user's location over some period of time by the most frequently observed location over that period.

[6]Some degree of tolerance toward missing data is required in order to avoid overestimating the number of visits.

[7]For the sake of consistency, the maximum duration defining visits is set just below the minimum duration parameter used to define temporary migration events. In other terms, a visit exceeding the maximum duration threshold is in fact identified as a temporary migration event in the migration detection procedure. Empirically, the preponderance of visits discerned in the dataset spans merely

this method to a large sample of CDR in Senegal to construct a dataset of observed visits and temporary migration choices at the individual-destination-time level, which can be flexibly aggregated at different spatio-temporal scales.

Similar to Chapter 1, I start by documenting the visiting patterns in Senegal inferred from CDR data. Owing to the expansive rural coverage provided by CDR data and the extended observation spans of phone users,[8] in contrast to smartphone app location data, I can better elucidate visiting flows across a diverse array of locales, encompassing rural-to-urban movements, and capture visits occurring on a less frequent basis. This arguably serves to augment the analysis of Chapter 1, which predominantly concentrates on visits between cities. Consistent with the findings of Chapter 1, the CDR data reveal a significant degree of mobility toward cities in Senegal. Over an observational window of approximately one year, an overwhelming majority – almost nine in ten users – register a visit to a city at least once. The median visitor makes one visit to a city every 1.3 months and each visit lasts for 1.5 days on average. While the propensity to visit cities exhibits a marginal decline with population density, it consistently remains elevated across various origin zones. Notably, there exist pronounced disparities in the frequency of visits to cities: a median visitor from Dakar undertakes an inter-city visit roughly every 86 days, in stark contrast to their counterparts from the most sparsely populated rural zones, who are observed in a city approximately every 18 days. Consequently, the cumulative duration dedicated to city visits diverges considerably based on the zone of origin: while Dakar residents average 12 days annually, those originating from the remotest rural precincts spend an average of 33 days per year.

Then, a comparison of the pervasiveness of visits and temporary migration to cities helps to fully appreciate the significance of short-term movements. Considering a large sample of users observed over an entire year, I estimate that 17% temporarily migrate to a city whereas 82% are seen visiting a city. Moreover, despite migration events being much longer than visits by definition, the aggregate time spent by individuals visiting cities is notably higher than the time spent migrating to cities: on average, phone users spend 17 days per year visiting cities against 11 days on temporary migration. Of course, this results from the striking prominence of visiting behaviors compared to migration, combined with the high frequency with which they effectively visit cities. Still, in aggregate terms, the data paint a picture of a situation in which visits, rather than longer-term movements,

a few days, with only an inconsequential proportion of prolonged visits potentially subject to ambiguous classification.

[8] See section 2.3 in Chapter 2 for a thorough description of the coverage and representativeness of the Senegal CDR dataset utilized in this chapter.

constitute the primary medium through which individuals experience non-home urban locations first hand.

The first part of the empirical analysis concentrates on deciphering the interrelation between visits and temporary migration. I discuss how this exploration can shed light on the relationship between amenities consumed during those distinct mobility events: are they acting as substitutes, complements, or are they wholly independent? Holding the cost of distance fixed, I probe whether an individual's visiting choices bears any association with their decision to temporarily migrate: ceteris paribus, do temporary migrants visit cities more or less than their counterparts? In practical terms, I analyze mobility decisions of users aggregated over a one-year period and I estimate regression models that relate visiting to migration choices, with the inclusion of origin fixed effects. First, I show that, for individuals facing comparable costs related to the distance to potential destinations, a positive association exists between visits and temporary migration to cities. Simply put, users who temporarily migrate to one or more cities also make more visits and, overall, spend more time in cities. Second, I consider more restrictive fixed effects at the origin-destination level to show that this observed relationship is predominantly propelled by temporary migrants visiting their migration destinations more abundantly.

These results do not have a clear causal interpretation and I discuss the various mechanisms that would be consistent with them. Notably, the results challenge the rudimentary hypothesis postulating visits as potential substitutes for extended temporary relocations. However, they lend credence to the idea of an inherent complementarity between those two types of mobility. To test that assumption, I consider the temporal dimension of visiting choices with respect to the timing of temporary migration events. A pivotal conclusion from this analysis is the non-random nature of visits to migration destinations with respect to the departure and return dates of temporary migration episodes. On average, and relative to other time periods when they are observed, prospective migrants exhibit a higher propensity to visit their imminent migration destinations in the weeks leading up to their departure. This could be reflecting the existence of prospective behaviors, wherein individuals willingly shoulder the cost of short visits in order to gain information about the destination, thereby mitigating the risk of migration failure. Furthermore, the analysis also highlights an increased likelihood of these migrants visiting their previous migration destinations in the weeks immediately following their return, in comparison to other periods within the observation timeline. While this finding presents a more enigmatic narrative, it potentially underscores follow-

up behaviors. Such behaviors may entail individuals revisiting cities where they had previously migrated to conclude pending tasks, procure pending payments, or engage with acquaintances established during their migration, among other conceivable motivations.

In the second part of the empirical analysis, I capitalize on the simultaneous observation of visits and temporary migration to cities with the objective of elucidating differentials in the costs intrinsic to each mobility modality. I develop a simple conceptual framework wherein a temporary move to a city, be it a visit or a temporary migration, is underpinned by a cost structure akin to that posited in the theoretical environment presented in Chapter 1. Each mobility event entails a fixed cost that is specific to the mobility modality, a movement cost implied by the distance between origin and destination, and a destination-specific cost associated with the duration of stay. In contrast to Chapter 1, which ultimately centers solely on the cost of movement related to distance to derive, for example, a gravity-style relation between visits and the geographical span between origin and destination, I retain the comprehensive cost framework. I derive expressions governing the frequency, duration and aggregate time allocation for both visiting cities and temporarily migrating to urban locales. The model elucidates that the distance elasticity of these mobility indicators, which can be inferred from observed mobility choices, is not simply related to the marginal cost of distance (i.e. the bus fare). Instead, it exhibits a negative relationship with the fixed cost of mobility. I estimate gravity regressions using the metrics of visits and temporary migration derived from the CDR data in Senegal. The results are reassuringly consistent with the basic predictions of the model: both for visits and temporary migrations, the frequency and cumulative time allocation to a destination exhibit a negative relationship with the distance from the origin, while the duration depends positively on the same. More importantly, the observed patterns of visits and temporary migration yield significant differences in the magnitude of the estimated distance elasticities. These disparities align with the assertion that the fixed costs associated with temporary migration exceed those affiliated with visits.

The paper contributes to the literature focusing on the causes, consequences and barriers to accessing urban markets in developing countries through temporary movements (Bryan, Chowdhury, et al., 2014; Morten, 2019; Imbert and Papp, 2020b,a), including in the context of Senegal (Delaunay et al., 2016; Lalou and Delaunay, 2017). It is, for instance, connected to Bryan, Chowdhury, et al. (2014) which studies the impact of an intervention subsidizing bus tickets to the city for rural villagers in Bangladesh. The authors argue that the lack of information and the

risk of unemployment at destination prevent rural workers from taking advantage of the large benefits of seasonal migration. The large inflows of visits to cities observed in the data presented in this paper – especially from rural zones – suggests that individuals do not completely ignore the characteristics of urban markets: they experience them first hand and on a regular basis. However, the existence of a risk premium associated with the uncertainty of employment is consistent with the relatively high fixed costs of temporary migration identified in this study. More generally, this finding is more in line with Imbert and Papp (2020a) and Lagakos et al. (2023) where harsh living conditions at destination, uncertainty and non-monetary fixed costs are identified as the main barrier to temporary migration, as opposed to the cost of living at destination or the pure cost of movement (i.e. the bus fare). Furthermore, Blumenstock, Chi, et al. (2022) harness mobile phone data to study the relationship between the migratory decisions and the typology of social networks. Their findings suggest an augmented propensity to migrate to locations where one's social network is larger and more interconnected. This study complements such findings, highlighting that individuals also undertake preliminary visits to potential migration destinations. Moreover, the paper contributes to the burgeoning literature that employs digital footprints to study human mobility (Blumenstock, 2012; Williams et al., 2015; Demissie et al., 2019). While other papers have usually focused on a single type of mobility, I underscore the potential of CDR data in concurrently gauging individual mobility decisions across varied time horizons, and I explore interconnections between visits and temporary migration to cities in a devleoping context.

The rest of the paper is organized as follows. Section 4.2 provides a brief description of the mobile phone data and outlines the methodology employed to quantify visits utilizing CDR data. Section 4.3 documents visiting patterns in Senegal with phone-derived mobility measures, and compares these with observed temporary migration movements. Section 4.4 examines the interplay between visits and temporary migration via regression analyses. In section 4.5, gravity regressions are used to uncover differences in the various costs tied to visits versus temporary migration. Limitations and avenues for future research are explored in Section 4.6, while Section 4.7 provides concluding remarks.

4.2 Data and mobility measurement

4.2.1 Data description

Digital traces constitute a powerful source of information, enabling the quantification of human movements across diverse temporal scales. Each movement type

inherently demands specific observational prerequisites, encompassing factors such as frequency and duration of observation, as well as spatial granularity. For instance, the detection of long-term movements, like permanent migration, may not necessitate a high observational frequency, but require observations over several years. Conversely, capturing daily commuting patterns mandates multiple observations per day at distinct hours of the day. Within this framework, CDR emerge as a versatile tool. They combine high frequencies of observation with long periods of observation that can go up to multiple years. On the other hand, the spatial resolution obtained from CDR data typically surpasses that associated with high administrative levels, rendering them apt for capturing long-distance movements as well as short-haul trips. Of course, coverage is confined to specific countries and CDR data are therefore best suited for delineating internal movements.[9]

I utilize three years of CDR data in Senegal, spanning the period 2013-2015, to quantify visits and temporary migration to cities. The characteristics of the dataset are extensively described in Chapter 2 (section 2.2). For the purpose of this study, I consider the same "high-quality" subset of users as in Chapter 2, who are observed for at least 330 days, on at least 80% of days, and with a maximum period unobserved of at most 15 days. Note that these observational constraints result from a trade-off between the accuracy of temporary migration measures and the reduced sample size and degree of selection that they induce. While the measurement of visits might technically mandate even more stringent observational criteria, for the sake of consistency and to mitigate further selection biases, I employ the high-quality subset for gauging both visits and temporary migrations. The presence of days without observations in users' trajectories unequivocally implies that individual-level measures of visits aggregated over time serve as lower-bound estimates.

Mobile phone users in the CDR dataset form a non-random sample of the Senegalese population. In Chapter 2, I discussed at length the various issues of representativeness inherent to the use of mobile phone data in general, and of this dataset in particular. As a reminder, I found that the population of mobile phone users with an active subscription to the telecommunication company providing the data (Sonatel) covers a large fraction of the population above 15 (63%). Mobile phone users are identified as being relatively more urban, predominantly more male and wealthier, even though the magnitude of these biases is generally small. Lastly, the comparison of the spatial distribution of users in the dataset with

[9]Other types of digital traces have been used to capture movements across countries. See, for instance, the work of Spyratos et al. (2019) or Ruktanonchai et al. (2018).

the population at large shows a pattern of selection to urban areas. However, it is largely driven by Dakar alone, and the distribution of the rest of the sample is broadly in line with the whole population, including for the most remote locations.

I observe the movements of users across 916 locations, of which 39 are identified as cities. In some analyses throughout the paper, I distinguish between large urban centers with a population above 100,000, i.e. primary cities, and other smaller, secondary cities. I also consider four groups of rural locations based on population density, which are labelled as very dense, dense, remote and very remote rural locations.[10]

Overall, the use of CDR data appears as particularly germane to the purpose of this study. They allow to observe the locations of a large number of individuals over relatively long periods of time, and with a frequency conducive to discerning both short-term visits and protracted temporary migration moves. Temporary migration events are identified based on the methodology outlined in Chapter 2, considering a minimum duration of 20 days (see section 2.4). The methodology for pinpointing visits is delineated in the subsequent section.

4.2.2 Measuring visits with CDR

Consistent with the terminologies introduced in Chapter 2, I define a user's meso-location at time t as the temporary migration destination if the user is in migration at time t, and as the place of residence otherwise. Then, visits are defined as short continuous blocks of time at a location that does not coincide with the prevailing meso-segment location.[11]

The detection of visits is based on a simple algorithmic procedure applied to the trajectory of consecutive daytime (8am-6pm) and nighttime (6pm-8am) locations.[12] The algorithm identifies clusters of consecutive half-days at a single location, which I refer to as "micro-segments". I permit some degree of tolerance towards missing data by allowing for observational gaps of at most ϵ_{gap}^{micro} within micro-segments. This consideration is pivotal to preclude the algorithm from inflated estimations of visits, which could result from misinterpreting a single stay punctuated with

[10] Appendix section 2.A in Chapter 2 provide details on the construction of those locations from georeferenced phone towers, and the definition of primary and secondary cities and rural sub-categories.

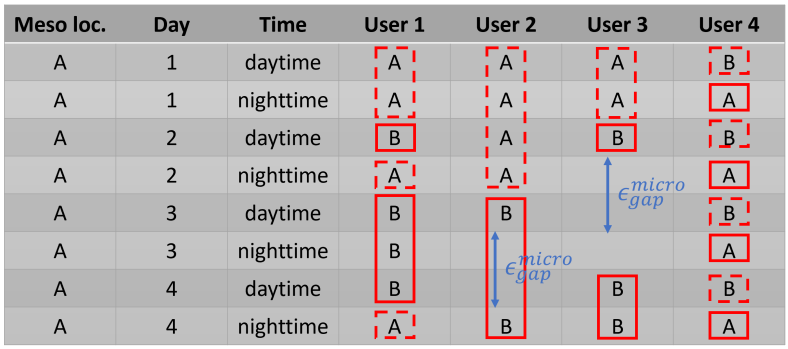
[11] Note that since the meso-segment location can be a temporary migration destination, the home location can be a visit destination if a user is seen visiting his home location while in migration.

[12] Daytime and nighttime locations are calculated as the modal hourly locations within the corresponding half-day, themselves defined as the modal observed location for the corresponding hour.

missing data points as multiple, distinct visits. The parameter ϵ_{gap}^{micro} essentially represents the unobserved duration under which it is deemed highly unlikely for the user to have visited an alternative location. Here, ϵ_{gap}^{micro} is set to two observations, i.e. one day and one night unobserved.

Some illustrative examples are provided in Figure 4.1 where four users have location A as meso-location and are observed at daytime and nighttime over four consecutive days. All red frames represent the micro-segments detected in those illustrative trajectories. $User$ 1 has five micro-segments, among which three are at the meso-location A (dashed frames). The other two are at a distinct location B (plain frames) and are effectively identified as visits. $User$ 2 has two consecutive missing observations (nighttime of day 3 and daytime of day 4), which creates an observational gap that is still below the tolerance threshold ϵ_{gap}^{micro} so that observations at location B at daytime of day 3 and nighttime of day 4 are still grouped within a single micro-segment. On the other hand, $user$ 3 has three consecutive missing observations, which results in an observational gap larger than ϵ_{gap}^{micro} so that the observation at location B at daytime of day 2 and those at the same location on day 4 form two distinct visits.

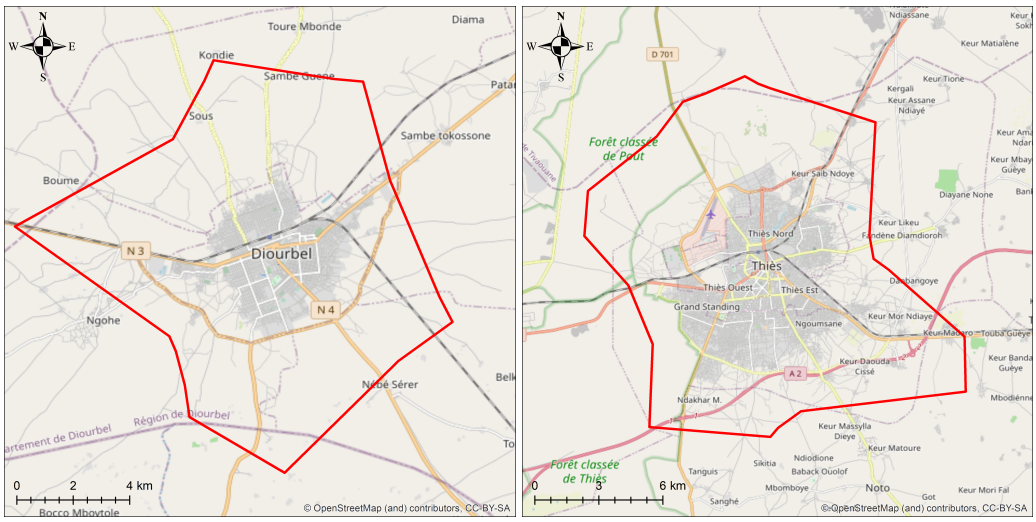
Figure 4.1: Illustrative example of visit detection.



In Figure 4.1, the trajectory of $user$ 4 is that of a typical commuter: he is systematically seen at his current primary location (A) at nighttime but at a distinct location B during daytime. Such mobility behaviors are theoretically captured by the algorithm and eventually qualified as (frequent) visits to location B, although visits and commuting remain conceptually distinct. Commuting movements can be considered a sub-type of visits characterized by systematic backwards and forwards movements between two locations – especially on week days – that typically reflect work (daytime) and home (nighttime) location choices. On the other hand, visits are understood as intermittent – but potentially frequent – stays of one or more days that serve a purpose that may or may not be work-related. Since this paper focuses exclusively on visits, the ability to distinguish commuting patterns from the rest of detected visits is essential. Two elements can help address

concerns about the contamination of the resulting measure of visits by commuting trips. First, commuting is primarily an urban phenomenon and the city polygons I construct from voronoi urban cells are qualitatively larger than actual city extents and plausibly encompass commuting zones (see, for example, the maps of Figure 4.2). Commuting patterns within a city or between a city and its periphery are therefore likely invisible to the detection algorithm since they take place within a single location. Second, commuting is relatively less common in developing countries, as outlined in Chapter 1 of this dissertation.

Figure 4.2: Examples of city polygons derived from phone tower locations.



(a) Diourbel (b) Thiès

4.3 Patterns of visits and temporary migration to cities in Senegal

This section provides a descriptive overview of visiting patterns to cities in Senegal. Unless stated otherwise, all statistics reported are based on a large sample of phone users observed over the one-year period going from February 2014 to January 2015. Details on the construction of this subset of users, labelled as *Subset 1*, are provided in appendix 4.A.1. It is only worth noting that the spatial distribution of the sample reflects the distribution of the population as a whole. Consequently, even though selection may exist at a local level, aggregate statistics inferred from this subset are at least unaffected by biases that typically result from the overrepresentation of some areas in the initial CDR dataset. Additionally, since visiting patterns could be influenced by the occurrence of events that were specific to that period of time (e.g. floods, conflicts, economic downturn...), I reproduce the results presented in this

section though with a different subset of users in the 2013 dataset.[13] Results are provided in appendix 4.C and, reassuringly, show very similar patterns.

First, I calculate simple mobility metrics to describe the propensity to visit cities, and the frequency and duration of visits over the course of one year. The overwhelming majority of users are seen visiting a city at least once; 83.4% of users visit at least one city. More specifically, 32% visit Dakar, 67% are seen in other primary cities with a population above 100,000[14], and 57% visit secondary cities.[15] On average, visitors travel to 3.4 distinct cities and nearly four out of five visitors are seen in more than one city over their period of observation. Moreover, disparities across zones of origin are limited and, for instance, 85% of rural residents are seen visiting a city with a comparable fraction of urban dwellers visiting at least one city (82%). Full results showing the fraction of users visiting urban destinations broken down by zone of origin are provided in Table 4.B.1.

For each visitor, I define the frequency of visits as a return period that represents the average length of time between two consecutive visits. I calculate the return period of visits as the total number of days observed divided by the number of visits. The median visitor makes one visit to a city every 1.3 months (40 days). However, the frequency of visits varies greatly across individuals and the population of visitors is comprised of both highly mobile individuals and occasional visitors. For instance, a quarter of visitors travel to a city at least every 15 days while at the other end of the distribution, 10% of visitors make at most 1.6 visits per year. Then, I calculate the median frequency of visits across distinct zones of origin representing groups of locations with different levels of urbanization. I form seven zones that include three urban zones (Dakar, primary cities, secondary cities) and four rural sub-categories each comprised of one quarter of rural locations and going from the most dense to the least dense rural areas. Interestingly, the median frequency of visits clearly increases as we move to more remote locations. The median visitor from Dakar makes one visit every 88 days on average, while users residing in primary and secondary cities show a higher frequency of visits with medians of 59 days and 34 days respectively. The median rural visitor makes one visit every 24 days on average and the return period decreases

[13]This subset corresponds to the dataset labelled as *Subset 3*, which construction is detailed in Appendix section 4.A.3.

[14]A total of 10 cities are classified as primary cities: Touba, Thiès, Mbour, Kaolack, Ziguinchor, Saint-Louis, Diourbel, Louga, Tambacounda, Kolda.

[15]The relatively low propensity to visit Dakar could be driven by the large fraction of users who actually live in Dakar and do not visit that city – by definition. I thus calculate the fraction of non-Dakar residents who visit Dakar and find a higher proportion of 47%.

from 31 days for the most dense rural zone to 18 days for the least dense rural zone.[16]

I then calculate the observed duration of a visit as the time elapsed between the start and end dates of the visit, and derive statistics from the universe of visits detected in the sample. Results are reported in Table 4.B.4. On average, visits to cities last for 1.5 days and rarely exceed a few days (the 9^{th} decile is equal to 3 days). The mean duration of visits is more extended when Dakar is the destination, averaging 2.7 days. In contrast, visits to primary cities average 1.4 days, and those to secondary cities are relatively shorter, averaging 0.9 days. Note that the observed duration of a visit is considered a lower-bound estimate since users are not necessarily observed right before and after the observed start and end dates. Therefore, I also define a maximum duration of visits as the time elapsed between the observation immediately preceding the visit start date and the observation following the visit end date. As expected, statistics on the maximum duration of visits yield higher numbers but differences remain small (see Table 4.B.5). For instance, the average maximum duration is estimated at 1.9 days – against an average observed duration of 1.5 days.

Next, I estimate for each visitor the total number of days spent visiting urban locations over the course of a year.[17] This allows to appraise the aggregate time allocation to cities that results from the frequency of visits and the time spent at destination during each visit. On average, visitors accumulate 23 days of visits to (non-home) cities per year. For a significant share of users, the total number of days spent visiting cities over a year is qualitatively comparable to the typical duration of temporary migration events detected in the data. For instance, 25% of users accumulate a total of at least 30 days of visits to cities per year. Consistent with the spatial disparities observed in the frequency of visits, I find that the average total number of days of visits per year varies significantly by zone of origin, from 12.6 days for Dakar residents to 33 days for visitors residing in the most remote rural locations.[18]

Finally, I estimate the total number of visits to cities disaggregated over time in order to gain new insights into the dynamics of visiting patterns over a typical

[16] Full tables of summary statistics on the frequency of visits to urban destinations at the national-level and by zone of origin are provided in Tables 4.B.2-4.B.3.

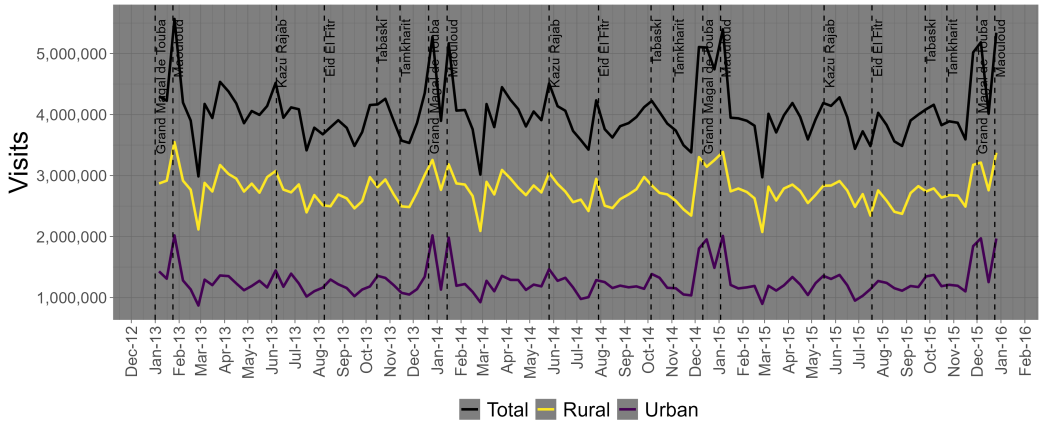
[17] To account for differences in the total number of days observed across users, I normalize the observed number of days of visits to cities to the number of days of visits per year by multiplying it by 365 divided by the total number of days observed. Note that such adjustments are only marginal given that all users are observed throughout almost the whole period under consideration.

[18] Full results on the total number of days of visits to urban destinations at the national-level and by zone of origin are provided in Tables 4.B.6-4.B.7.

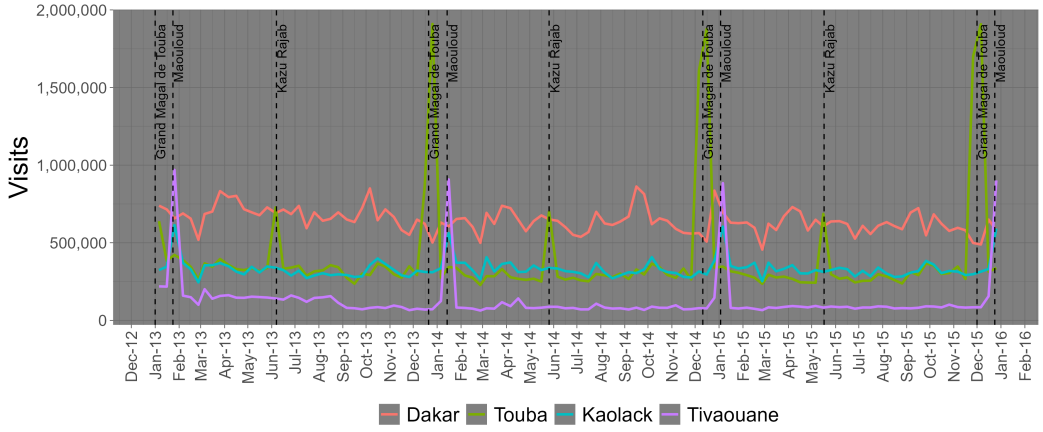
year. Because the mobile phone dataset at hand covers multiple years (2013-2015), it is possible to investigate the existence of seasonal and other forms of systematic patterns. I therefore use another large sample of users that spans the entire period 2013-2015.[19] Figure 4.3(a) shows the total number of visits by dekad – i.e. for each period of ten days – for the period 2013-2015. The trajectory of visits over the three-year period does not showcase any obvious time trend and variations around the long-term average are generally modest. Religious events usually go hand in hand with pilgrimages towards specific locations where festivities are held, or short-term visits of individuals joining their family for celebrations. They could therefore be a potential driver of the short-term movements observed in the data. In Figure 4.3(a), vertical dashed lines indicate the dates of the main religious events in Senegal and are found to often coincide with visiting peaks. The two most important religious festivals – the *Grand Magal* of Touba and *Maouloud* (or *Mawlid*) – traditionally imply pilgrimages to specific places of celebration and are effectively associated with marked increase in the flow of visits. Figure 4.3(b) shows visits over time to these particular destinations and allows to confirm that visits to places of celebration largely drive the observed significant deviations from the long-term average. The city of Touba, which hosts the annual *Grand Magal* pilgrimage of the Senegalese Mouride order, experiences a significant increase in the amount of visits received during the event, with nearly 2 million visits compared to about 300,000 per dekad the rest of the time. This corroborates anecdotal evidence asserting that the *Grand Magal* is the most unmissable religious event in Senegal where millions of individuals travel to Touba to join the celebration. Similarly, Tivaouane and Kaolack are the two main pilgrimage destinations for the *Maouloud* festival which celebrates the day when the prophet Muhammad was born. Both cities show clear visiting peaks every year at the date of the event. Major religious events therefore drive punctual increases in visits. However, they certainly do not account for the overall sustained level of visiting flows observed throughout the period.

[19] Details on the construction of this sample is provided in appendix 4.A.

Figure 4.3: Visits to cities by ten-day time interval, 2013-2015.



(a) Total and by origin zone



(b) By destination city

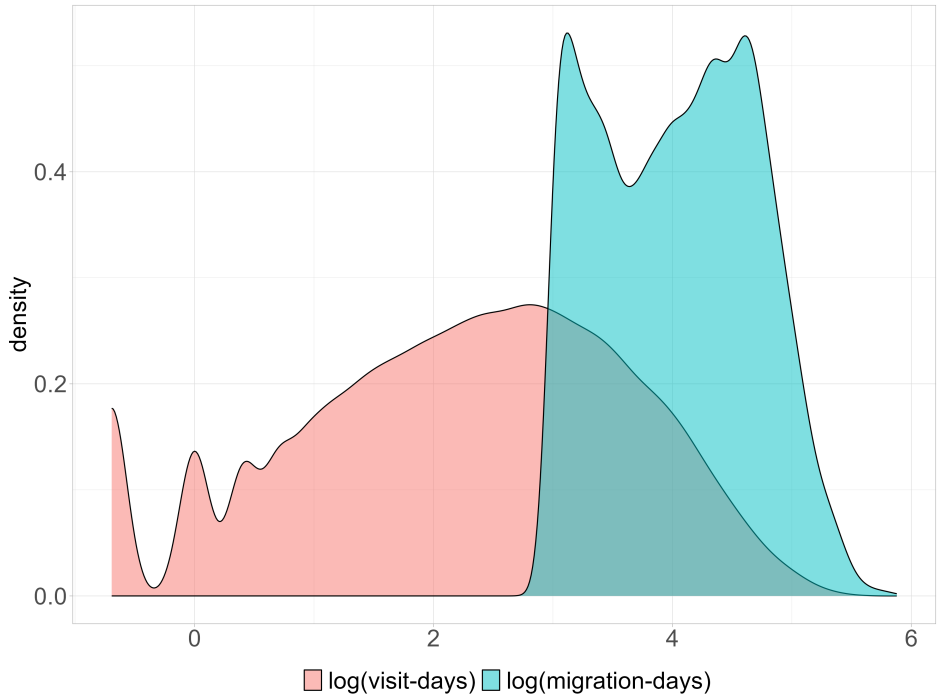
Note: in Figure (a), the black line represents the total number of visits made by all users to all cities per dekad for the period January 2013-December 2015. The yellow and blue lines give the sub-totals for visitors originating from rural and urban areas, respectively. Numbers are scaled up to the total population above 15 through weights defined at the third administrative level and for each sub-rural category and individual city as the population-to-users ratio. Similarly, Figure (b) provides the total number of visits made per dekad to specific cities. The vertical dashed lines represent the dates of the main religious events in Senegal for the period considered.

Next, I abstain from furnishing a similarly exhaustive description of temporary migration movements in Senegal, given that such a detailed account is already presented in Chapter 2 (section 2.6). Rather, I provide key indicators allowing to appreciate the magnitude and characteristics of temporary migration to cities relative to visiting flows. Employing the identical dataset that spans from February 2014 to January 2015 (*Subset 1*), the proportion of users who engaged in a temporary migration event to a city is estimated at 17%. While this percentage pales in comparison to the proportion of users making visits to cities (82%), it is evident

that both modalities of transient mobility concurrently manifest within the studied population.

Furthermore, within the observed sample, mobile phone users register an average of 17 days of visits annually, contrasted with 11 days of temporary migration. At a cursory glance, this suggests that visits are predominantly the mechanism through which individuals engage with cities outside their home locales. Notably, a substantial segment of these visitors commit an amount of time to city visits comparable to the typical durations exhibited by temporary migration episodes as recorded in the data. Figure 4.4 depicts the density distribution of total days visitors spend in cities annually versus the cumulative days temporary migrants allocate to migrations. A significant overlap between these two distributions emerges, indicating that a notable fraction of visitors allocate as much time to city visits as certain migrants do for their migration stints to analogous destinations.

Figure 4.4: Density of users' total time spent visiting and migrating to cities.



Note: The red curve represents the density of individuals' total time spent visiting cities over a year, excluding non-visitors. The blue curve is the density of the total time spent in migration to cities over a year, excluding non-migrants. Both variables are logged-transformed for presentational purposes. Calculations are based on the sample of 100,000 users observed over the period February 2014-January 2015.

Thus, what reasons could explain the decision of certain individuals to distribute their time in cities across multiple short visits, while others opt for extended stays

amounting to a similar duration? One plausible rationale could be inherent distinctions in the activities undertaken by individuals when they either visit or temporarily migrate to urban areas. In essence, visit-based and migration-based urban amenities could be fundamentally divergent. Alternatively, shorter visits might serve as feasible substitutes for more extended stays, and the disparate cost structures associated with each mobility mode could elucidate the observed movement patterns.

4.4 The empirical relationship between visits and temporary migration to cities

As previously illustrated, individuals spend time in (non-home) cities via regular visits as well as through episodes of temporary migration. The simultaneous presence of these two mobility modalities prompts inquiries concerning the relationship between the amenities accessed during visits and those sought during temporary migrations. Specifically, do these amenities function as substitutes, complements, or operate independently of one another?

Urban amenities consumed via visits and temporary migration are substitutes if the consumption of one competes against the other. In that case, the intrinsic attributes of the visit-based and migration-based urban amenities would have considerable overlaps. For example, some producers in the agricultural sector may choose to frequently visit relatively small urban markets for their sales, while others may find it preferable to travel for several weeks to a larger city with a bulkier stock.

Visit-based and migration-based urban amenities are perceived as complements if they are used in conjunction with one another. For instance, people contemplating a temporary relocation to a city may choose to visit that city in order to gather information and acquaint themselves with prospective opportunities. This specific hypothesis is delved into with greater detail in section 4.4.2.

Finally, urban amenities consumed via visits and temporary migration are independent if they are neither substitutes nor complements: the demand for one has no incidence on the demand for the other. In that scenario, amenities consumed via visits and temporary migration address clearly distinct needs. For instance, individuals may visit urban locations for the sole purpose of consuming amenities which are unavailable at home, such as health and administrative services, or to source or sell products. On the other hand, they migrate to cities to find employment in the non-agricultural sector where they temporarily enjoy a higher marginal productivity (e.g. due to seasonal fluctuations in agricultural productivity), or as a way to diversify income sources.

The actual nature of the interplay between visits and temporary migration is most likely a combination of those scenarios. In this section, drawing upon the concurrent observation of individual-level visit and temporary migration choices derived from CDR data, I aim to discern which scenario aligns more closely with the observed mobility behaviors.

4.4.1 Do temporary migrants make more visits to cities?

I start by examining visits and temporary migration aggregated over a year and across all cities for the identical sample of users considered in section 4.3 (i.e. *Subset 1*), and I investigate the direction of the relationship between them. In simple terms, I ask whether users who are seen temporarily migrating to a city make comparatively more or less visits to cities. To that end, I estimate the following regression model:

$$V_{io} = \beta_{1,0} + \beta_{1,1}\mathbb{1}(M_{io} > 0) + \delta_o + \epsilon_{io} \tag{4.1}$$

where V_{io} represents some measure of visits to cities made over a year for individual i residing in location o, which I define as either the total number of visits to cities, or the total number of days of visits to cities. Similarly, M_{io} is the total number of temporary migration episodes or the total number of days spent in migration to cities, so that $\mathbb{1}(M_{io} > 0)$ is a dummy equal to 1 if individual i engages in at least one temporary migration to a city over the period considered. δ_o controls for the time-invariant unobserved characteristics of location o that influence the amount of visits to cities. In particular, it captures the movement costs related to the distance to cities shared by all residents in o.

The estimation of the parameter of interest $\beta_{1,1}$ relies on within-location variations in mobility choices among individuals, where users face the same geographic costs of movement to cities.The primary objective of this estimation is discerning the direction of the relationship between visits and temporary migration , rather than positing a causal linkage – it does not assert that a migration occurrence directly spurs an increase or decrease in visits. Variations in temporary migration choices across individuals cannot be considered as exogenous a priori; confounders may influence both the propensity to visit and to temporarily migrate to cities. Nonetheless, the sign of $\beta_{1,1}$ is still informative about the relation between visits and temporary migration, and by extension, between the intrinsic amenities associated with these mobility events.

It is important to note that drawing conclusions about the nature of the relationship between visit- and migration-based urban amenities via equation (4.1)

supposes some important assumptions about the functional relation between observed mobility events and urban amenity consumption. Here, the proposed mobility measures are simply viewed as a proxy for the consumption of an urban amenity. The amount of amenity consumed is proportional to the number of visits (resp. migration episodes) or the total time spent visiting (resp. migrating to) cities. Moreover, the amenity consumed per mobility event or per unit of time spent at destination is independent of the destination city. In other words, for each individual, I aggregate all the visits made to cities during the study period without assigning weights to the various cities visited. While a reasonable starting point in the analysis, I acknowledge this may only be a coarse approximation for the actual value derived from those trips. For example, one day of visit to Dakar likely provides a higher amenity than one day of visit to the small and landlocked city of Matam. I get back to the issue of quantifying the heterogeneous "value of cities" in section 4.5 but, for now, keeping those limitations in mind, I proceed with the estimation of equation (4.1).

Equation (4.1) is estimated with OLS and results are presented in Table 4.1. Aggregate visits and temporary migration are found to be positively related. All else equal, temporary migrants are associated with 6.5 additional visits per year on average compared to non-migrants, as showcased in column (1). The result still holds when considering the total time spent visiting cities as a dependent variable (column (3)). On average, temporary migrants spend an extra 17.5 days visiting cities compared to non-migrants. This is equivalent to the average time spent by users on visits to cities and is thus considered as quantitatively large. In Appendix 4.D, I explore the heterogeneity of this relationship across groups of origin locations. The coefficient of interest remains positive and statistically significant when considering rural and urban categories, primary cities, secondary cities and sub-rural zones of different density, but also across regions (Tables 4.D.1-4.D.3).[20]

As mentioned above, the positive association between aggregate visits and temporary migration cannot be causally interpreted. In fact, at least three distinct mechanisms could be consistent with a positive relationship between visits and temporary migration, and entail different interpretations about the substitutability, complementarity or independence of visit- and migration-based urban amenities.

First, temporary migration could cause an increase in visits to the migration destination in a situation where individuals make short trips to the future migration

[20]It should only be noted that the magnitude of the coefficient is found to be slightly larger for urban compared to rural areas. There are also differences across regions and the estimated average additional number of visits made by temporary migrants varies from 3.9 (in Matam) to 9.9 (Dakar).

destination, for instance, to expand their social network or obtain information about potential work opportunities. Consequently, and to be more specific, it is the intention to migrate that leads to more visits. In that case, urban amenities associated with visits and temporary migration can be at least considered as "one-way" complements; temporary migration is used in conjunction with visits but the converse is not necessarily true. Alternatively, migration episodes could entail a learning process about the characteristics of the destination, making it a desirable location that is subsequently visited as a result. In this scenario, the implication for the relation between the two types of urban amenities is less clear. They may be independent if visits simply result from gaining information about the location while in migration. As in the interpretation above, they may also be one-way complements although in the opposite direction; individuals may need to use post-migration visits in order to complete an unfinished task in relation to the activity carried out during the migration episode. I explore these ideas more in depth in section 4.4.2 where I look at migrants' distribution of visits over time with respect to the timing of migration.

Conversely, visits could be considered as causing migration if some exogenous factor induces visits during which individuals learn about the destination. They then decide to temporarily migrate to that location precisely because they hold information about that place. In that case, the amenities underlying visits and temporary migration would most likely be independent. For instance, people may visit locations to visit friends and relatives and opportunistically learn about temporary work opportunities. This would imply an alternative interpretation on the role of individuals' social network in determining migration choices compared to previous studies, in which network members actively and specifically support migration processes by providing relevant information about the destination, or acting as a safety net for migrant (Munshi, 2003, 2014; Blumenstock, Chi, et al., 2022). Furthermore, religious festivals could also be seen as exogenously causing occasional visits during which individuals gain information about the host city, and which in turn could increase the propensity to migrate to that specific destination. For example, the *Grand Magal* that takes place in the city of Touba attracts millions of visitors each year (see section 4.3). If the proposed mechanism is true, then Touba should appear as a particularly attractive location for temporary migrants, ceteris paribus.

Thirdly, a confounding factor could be positively related to the demand for both visits and migration, leading to a spurious association. For instance, individuals' income could be a primary determinant of mobility choices, with stronger liquidity constraints impeding the demand for both types of mobility. Moreover, the size and topology of an individual's social network at destination could influence

both the decisions to visit and temporarily migrate, without visits having any causal impact on migration – and vice versa. If visits and temporary migration are entirely determined by confounding factors, then the demand for one has no incidence on the demand for the other and the underlying amenities are considered as independent.

In any case, the positive correlation between visits and temporary movements contradicts the idea that these two types of mobility could substitute for each other within a set of individuals sharing comparable costs of movement.

Table 4.1: Relationship between aggregate visits and temporary migration, controlling for origin fixed effects.

	No. of visits (1)	No. of visit-days (2)
Migrant dummy	6.545***	17.48***
	(0.4586)	(0.4960)
Observations	113,452	113,452
Pseudo R^2	0.03049	0.02718
Origin FE	✓	✓

Note: Each observation represents a user with mobility measures aggregated over the period of observation and across all destination cities. Column (1) shows the PPML estimation of a regression of the total number of visits to cities on a migration dummy equal to 1 if the user has at least one temporary migration event to any city. Column (2) shows the same estimation considering the total time spent visiting cities as a dependent variable. Standard errors are clustered by origin location.

The estimation of equation (4.1) highlights that, on average, temporary migrants make more visits to cities – and also spend more time visiting cities. However, it does not allow to say anything about the specific destinations which are marginally more visited. For that reason, I also estimate a version of equation (4.1) allowing to investigate the link between visits and temporary migration to a particular destination. I simply consider individual-level mobility measures disaggregated by destination city, and I replace origin fixed effects with origin-destination fixed effects:

$$V_{iod} = \beta_{2,0} + \beta_{2,1}\mathbb{1}(M_{iod} > 0) + \delta_{od} + \epsilon_{iod} \tag{4.2}$$

where V_{iod} is the total of visits made over a year to destination d, by individual i residing in location o. $\mathbb{1}(M_{iod} > 0)$ is a dummy equal to 1 if individual i engages in at least one temporary migration to city d. δ_{od} represents origin-destination fixed effects that absorb time-invariant characteristics of the origin-destination pair $o - d$ that influence the level of mobility between o and d. In particular, it controls for the cost associated with the geographic distance between o and d.

I estimate equation (4.2) with OLS and show the results in Table 4.2. The coefficient associated with the temporary migration dummy is again positive and statistically significant. On average, individuals who temporarily migrate to a city make 5.8 additional visits and spend 14.5 extra days to that particular city, compared to those who do not temporarily migrate to that destination but reside in the same origin location. Again, I dig into the heterogeneity of the results across groups of origin locations and find remarkably similar effects, except across regions where some differences appear – although all coefficients remain positive and significant (Tables 4.D.4-4.D.6).

Table 4.2: Relationship between visits and temporary migration controlling for origin-destination fixed effects.

	No. of visits (1)	No. of visit-days (2)
Migration dummy	5.841***	14.49***
	(0.2303)	(0.3303)
Observations	4,424,628	4,424,628
Pseudo R^2	0.07478	0.06605
Origin-destination FE	✓	✓

Note: Each observation represents a user-destination couple with mobility measures aggregated over the period of observation. Column (1) shows the PPML estimation of a regression of the total number of visits to a destination on a migration dummy equal to 1 if the user has at least one temporary migration event to that destination. Column (2) shows the same estimation considering the total time spent visiting the destination as a dependent variable. Standard errors are clustered by origin-destination pair.

Interestingly, the magnitude of the estimated coefficients is slightly smaller though largely comparable to those obtained in the estimation of equation (4.1). This strongly indicates that the extra visits made by temporary migrants to their

migration destination account for a large fraction of the total additional visits to any city that the estimation of equation (4.1) suggests they make. I test that assumption formally by comparing the aggregate visits that migrants make to cities other than their migration destination to the aggregate visits that non-migrants make to those cities, i.e. also excluding the visits to that migration destination. To do so, for each pair of origin location o and migration destination d, I form a group of users comprised of all non-migrants residing in o and all users with a temporary migration event to destination d.[21] Then, for each user, I aggregate visits across all cities except d and I estimate a regression model similar to equation (4.1) with OLS, although with fixed effects at the level of (origin * migration destination) groups and considering the constructed measure of visits that excludes the visits made to a migration destination. Results are showed in Table 4.3. In column (1), the regression of the number of visits on the migrant dummy yields a positive but non-significant coefficient. Column (2) indicates a positive and significant association between the total time spent visiting cities and being a migrant, but the coefficient is quantitatively small (1.4) compared to the one obtained in the estimation of equation (4.2) (14.5). Those results tend to suggest that visiting patterns of temporary migrants do not differ from those of non-migrants when considering the subset of cities that excludes the migration destination.

[21] Note that non-migrants residing in o therefore appear in as many groups as the number of migration destinations observed among temporary migrants from o. Also, for simplicity, I consider only migrants with a unique migration destination, which account for 90% of all temporary migrants in my sample.

Table 4.3: Relationship between temporary migration and aggregate visits to non-migration destinations

	No. of visits (1)	No. of visit-days (2)
Migrant dummy	0.1043	1.426***
	(0.3028)	(0.2107)
Observations	1,708,926	1,708,926
Pseudo R^2	0.01857	0.01067
Adjusted R^2	0.14587	0.08544
Origin-destination group FE	✓	✓

Note: Each observation represents a user. A group of user is formed for each origin-destination pair that has at least one temporary migrant to the destination, and is comprised of all temporary migrants to that destination (and that destination only) and non-migrants. Mobility measures are then aggregated over the period of observation and across cities, excluding the destination. Column (1) shows the PPML estimation of a regression of the total number of visits to cities on a migration dummy identifies temporary migrants in each origin-destination group. Column (2) shows the same estimation considering the total time spent visiting cities as a dependent variable. Standard errors are clustered by origin-destination group.

Moreover, the origin-destination fixed effect term in equation (4.2) effectively allows to absorb the expected impact of the cost of distance on the level of visits, but it is expected that distance also mediates the magnitude of the coefficient of interest itself. All else equal, increasing the distance between origin and destination locations should decrease the average additional visits made by temporary migrants. I estimate a model that includes an interaction between the migration dummy and the logged distance between origin and destination. I also interact the migration dummy with a categorical variable that allows for heterogeneity across three destination classes: Dakar, primary cities and secondary cities. Results are provided in Table 4.4. Reassuringly, the coefficients associated with the interaction term between the migration dummy and distance is negative and statistically significant; the average extra number of visits to the migration destination – or time spent visiting the destination – decreases with the travel distance between origin and destination. Still, an alternative estimation considering a categorized version of the distance variable reveals that the effect persists even when origin and destination locations are more than 200km apart (see Table 4.D.7). Then, net of the interaction effect of distance, the estimated coefficients associated with the

interaction between the migration dummy and the destination category show only small differences. If anything, the relationship between visits and migration is slightly stronger when the destination is a secondary city and the lowest coefficients are obtained for Dakar.

Table 4.4: Relationship between visits and temporary migration controlling for origin-destination fixed effects, heterogeneity with respect to destination.

	No. of visits (1)	No. of visit-days (2)
Migration dummy × To Dakar	24.93***	36.67***
	(1.367)	(1.411)
Migration dummy × To primary cities	26.36***	39.35***
	(1.266)	(1.391)
Migration dummy × To sec. cities	29.85***	40.49***
	(1.294)	(1.289)
Migration dummy × log(distance)	-4.117***	-4.793***
	(0.2376)	(0.2512)
Observations	4,369,673	4,369,673
Pseudo R^2	0.07674	0.06834
Origin-destination FE	✓	✓

Note: Each observation represents a user-destination couple with mobility measures aggregated over the period of observation. Column (1) shows the PPML estimation of a regression of the total number of visits to a destination on a migration dummy equal to 1 if the user has at least one temporary migration event to that destination, interacted with the logged travel distance by road between origin and destination, and another interaction with the urban category to which the destination belongs (Dakar, other primary cities, or secondary cities). Column (2) shows the same estimation considering the total time spent visiting the destination as a dependent variable.Standard errors are clustered by origin-destination pair.

4.4.2 The dynamics of visits before and after migration events

The empirical analyses of section 4.4.1 underline the existence of a strong positive relationship between visits and temporary migration. On average, individuals who are seen temporarily migrating to a city over a year also make more visits to cities during that period of time compared to non-migrants. This relationship is almost entirely driven by additional visits made to the migration destination alone. I have presented a number of underlying mechanisms that could be consistent with this result, with some innately suggesting a temporal pattern for these supplementary

visits. For example, the idea that visits could be a precursor to migration would naturally imply a surge in visits shortly before migration commences, ceteris paribus. In this section, I aim to delve deeper into the temporal patterns of visits made by temporary migrants surrounding their migration events to discern which proposed mechanisms carry more empirical weight.

To do this, I consider temporary migrants from the *Subset 5* defined in Appendix 4.A.5, in which users have observations spanning at least from February 1, 2014, to September 30, 2015. I build an individual-destination-time panel that provides measures of visits as well as dummy variables indicating the occurrence of a migration departure or return. The extended observation period facilitates the observation of temporary migration events occurring over an entire year (from June 2014 to May 2015), as well as the visiting decisions both in the four months leading up to migration departures and in the four months following the returns from migration. This ensures that the subsequent analysis of visits dynamics before and after migration events is not influenced by seasonal factors. Utilizing the detailed temporal insights afforded by CDR data, I derive mobility metrics in ten-day intervals, or dekads. Then, I investigate whether within-individual variations in visits over time to a particular destination are related to the timing of the migration departure and return to that destination. Formally, I estimate the following regression model:

$$\begin{aligned}\mathbb{1}(V_{iodt} > 0) = &\sum_{\tau=2}^{T-1} \alpha_\tau m^{depart}_{i,o,d,t+\tau} + \alpha_{T^+} m^{depart}_{i,o,d,T^+} + \sum_{\tau=1}^{T-1} \beta_\tau m^{return}_{i,o,d,t-\tau} + \beta_{T^-} m^{return}_{i,o,d,T^-} \\ &+ \Delta_{id} + \Gamma_{dt} + u_{iodt}\end{aligned} \tag{4.3}$$

where $\mathbb{1}(V_{iodt} > 0)$ is a dummy equal to 1 if individual i residing in o visits destination d during time period t. $m^{depart}_{i,o,d,t}$ is a dummy equal to 1 if i migrates to d at time t so that the first term on the right-hand side represents leads of the migration departure up to $T-1$ time periods after t. Note that the first lead $m^{depart}_{i,o,d,t+1}$ is omitted as it is used as the base category. The time periods beyond $t+T-1$ are binned and the variable m^{depart}_{i,o,d,T^+} is thus a dummy equal to 1 if i moves to d on $t+T$ or any period after that. Similarly, $m^{return}_{i,o,d,t}$ is a dummy equal to 1 if i migrates to d at time t and the third term corresponds to lags of the migration return up to $T-1$ time periods before t. The time periods preceding time $t-(T-1)$ are binned and m^{return}_{i,o,d,T^-} is defined as a dummy equal to 1 if i returns from migration on $t-T$ or any period before that. Δ_{id} represents individual-destination fixed effects and Γ_{dt} controls for destination-time-specific unobservable factors that influence the likelihood of visits in the same way for all individuals. For instance, Γ_{dt} absorbs the

impact of religious festivals which can induce large influxes of visits to some cities.

For any given τ, the coefficient α_τ (resp. β_τ) is interpreted as the expected deviation from the individual-level average of the probability of visit to the migration destination τ periods before (resp. after) a migration departure (resp. return), compared to the deviation one period before departure, and net of destination-time fixed effects. Thus, as specified, the model does not allow for a straightforward interpretation in absolute terms of the differential in the probability of visit τ periods before a migration departure.[22] However, it is sufficient to simply test whether the timing of visits to a migration destination is non-random with respect to the timing of the migration event itself.

I estimate equation (4.3) as a linear probability model with OLS on the sample of temporary migrants from *Subset 5*,[23] considering $T = 9$ dekads. Results are presented in Figure 4.5. I also estimate a logit model and the results are qualitatively unchanged (see Figure 4.E.1 in Appendix 4.E). Some interesting patterns emerge. First, there is a clear increase in the mean probability of visit at the future destination two and three dekads before departure, relative to the dekad just preceding departure. Second, the average probability of visit for time periods beyond 3 months (i.e. 9 dekads) prior to departure is relatively lower – it is actually the lowest estimate. Third, a symmetrical effect is observed two and three dekads following the return from migration: temporary migrants have a relatively higher propensity to re-visit the location where they migrated. The effect decays over time but persists on the fourth and fifth dekads after the return. Finally, the probability of visit during time periods that are beyond 3 months after the return is also relatively lower on average.

The central lesson of the results is that, from a statistical perspective, the timing of temporary migrants' visits to their migration destination is non-random with respect to the migration departure and return dates. Putting these results together with those of section 4.4 provides evidence that the additional visits that temporary migrants make to their migration destination compared to non-

[22] The event study design that would potentially allow to do so would substantially complicate the estimation. It would essentially imply a difference-in-differences setting with staggered treatments where traditional two-way fixed effects estimators have been showed to exhibit serious biases (Goodman-Bacon, 2021; Chaisemartin and D'Haultfœuille, 2020), including for event studies (Sun and Abraham, 2021). The various remedies described in the literature such as those proposed by Sun and Abraham (2021) or Borusyak, Jaravel, et al. (2021) are difficult – if not impossible – to apply in the present setting, most notably because treated individuals (i.e. migrants) do not have untreated observations. This is because visits are a priori related to temporary migration events for both periods pre-departure and post-return; in fact, results from the estimation of the model I propose do not allow to reject these hypotheses. This, for instance, makes the "imputation" method proposed by Borusyak, Jaravel, et al. (2021) impractical.

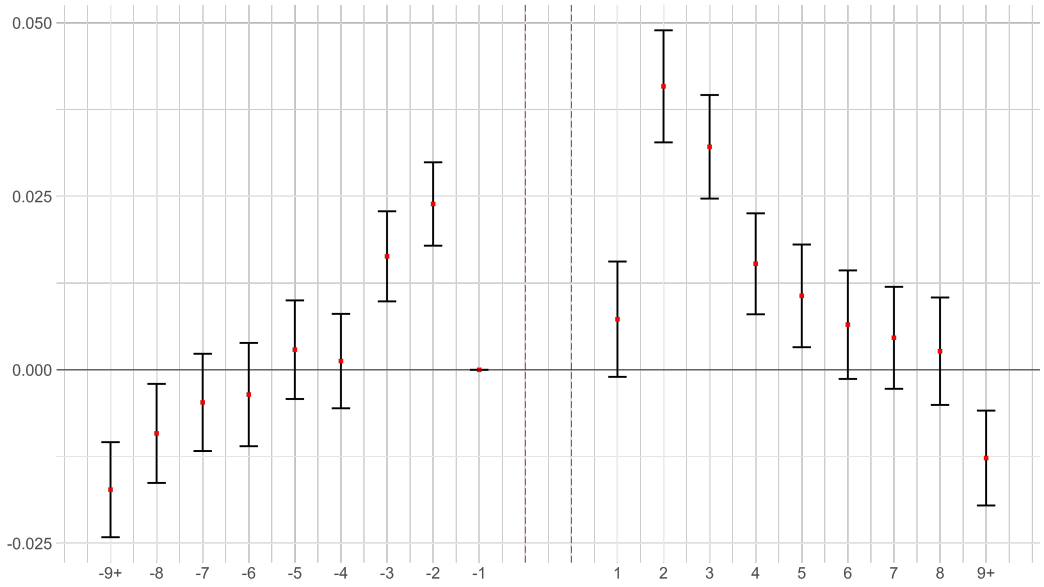
[23] See details in Appendix section 4.A.5

migrants are primarily undertaken in the month preceding departure and the month following return.

A plausible interpretation of the relatively higher propensity to visit the future destination in the dekads preceding a migration departure is that individuals adopt prospective behaviors. They are willing to bear the costs of visits in order to get more information about a potential destination and perhaps with the objective of reducing the risk of migration failure (Harris and Todaro, 1970), e.g. by learning about work opportunities, negotiating contracts, and so on. Visits could then be viewed as a potential technology allowing to reduce spatial frictions caused by imperfect information. Moreover, this result could complement the existing evidence on the role of social networks in shaping migration decisions. Individuals do not only need contacts at destination to provide for social capital and information, they also exhibit some demand for experiencing the destination first-hand via short trips before migrating. Then, a natural follow-up analysis would look at the possible link between pre-migration visits and social network dynamics. Do prospective migrants effectively make new contacts at destination during visits? Although this could be investigated with the mobile phone data employed in this study, I leave it to future work to potentially tackle this question.

The relative increase in the probability of visits in the dekads following a migration return is less easily interpreted. Broadly speaking, it points to the existence of follow-up behaviors according to which individuals re-visit the destination where they migrated, although the possible reasons behind such movements are arguably unclear. The possibility exists that temporary migrants tend to return to their migration destination to finish unfinished business, to collect a payment, or simply to visit the new connections they made while in migration.

Figure 4.5: Relative distribution of migrants' visits to destination before and after a migration event, linear model.



Note: The two vertical dashed red lines in the center of the graph represent the migration departure (left) and return (right). The x-axis represents the relative time with respect to the migration departure and return. Note that the gap between migration departure and return theoretically coincides with the migration duration but is normalized to one dekad for representation purposes. Red dots on the left-hand side correspond to OLS estimates of $\alpha_2, ..., \alpha_8, \alpha_{9+}$ and those on the right-hand side are estimates of $\beta_1, ..., \beta_8, \beta_{9-}$. Coefficients are estimated on a sample of 30,523 unique temporary migrants observed over the period 2014-2015, resulting in a grand total of 1,358,184 observations. The error bars show the 95% confidence intervals based on two-way clustered standard errors at the individual-destination and dekad-destination levels.

4.5 Comparative gravity estimates for visits and temporary migration

The empirical analyses in section 4.4 essentially probe to understand the relationship between visits and temporary migration choices of individuals facing comparable mobility costs. In this section, I focus more specifically on the role of movement costs in shaping these distinct mobility decisions.

4.5.1 A simple conceptual framework

I start by introducing a simplified conceptual framework that considers a mobility cost structure and essentially models an individual's endogenous mobility decision in terms of the frequency and duration of movements over some period of time. I derive expressions relating measurable choices of visits or temporary migration – i.e. the frequency of movements, the duration of mobility events and the total time

spent at destination – with the corresponding cost parameters. This allows me to provide useful intuitions about the role of unobserved fixed costs of mobility in shaping the observable relationship between distance and mobility decisions.

Assume a two-location environment in which individuals residing in a location o are endowed with some budget w, which they use to visit or temporarily migrate to a location d. The distance between o and d is denoted by D_{od}. Spending some amount of time θ at destination d – either visiting or in migration – is associated with the consumption of an amenity, which value corresponds to the product of θ with a location-specific parameter equal to $A_v(d)$ if the individual visits d, and $A_m(d)$ if he is in migration to d.

The movement cost structure is similar to the one introduced in Chapter 1. It is the sum of a fixed cost, denoted as λ, which differs between visits (λ_v) and temporary migration (λ_m), a distance-related variable cost scaled by D_{od} and parametrized by γ, and a duration-associated cost dictated by the destination-specific parameter κ_d: $\lambda + \gamma D_{od} + \kappa_d \theta^\alpha$, where $\alpha > 1$.[24] The parameter κ_d primarily reflects the marginal costs of living and accommodation at the destination d. Therefore, the cost of a visit lasting θ_v is given by $\lambda_v + \gamma D_{od} + \kappa_d \theta_v^\alpha$, whereas the cost of a temporary migration of duration θ_m is $\lambda_m + \gamma D_{od} + \kappa_d \theta_m^\alpha$. The differential costs between visits and temporary migrations are primarily anchored in the fixed costs of movement, which encompass a spectrum of both monetary and intangible considerations linked to relocating from one's home location. For instance, this might encapsulate the psychological implications of detachment from one's social circle, the degree of strategic planning ensuring sustained economic activity at home, or the potential costs imposed on relatives at destination. Moreover, this cost paradigm can also be perceived as reflecting the apprehension or perceived risk associated with a suboptimal outcome from a mobility event – an aspect that is conceivably magnified in the context of temporary migration.

Note that, similar to lemma 1 in Chapter 1, given a total amount of time spent visiting or migrating to destination d through ν mobility events, the imposed cost structure implies that a cost-minimizing behavior will result in all events having the same duration.

The utility derived from an individual mobility event is assumed to be a function of the overall amenity consumption, exhibiting diminishing returns governed by a parameter $\beta < 1$.[25] For instance, the utility stemming from a visit of duration θ_v to

[24] As noted in Chapter 1, the increasing marginal duration-associated cost of mobility is introduced to motivate the possibility that an individual might make multiple visits to a single destination.

[25] Note that the principle of diminishing returns to the consumption of amenities accessed through

a designated destination d is expressed as $(\theta_v A_v(d))^\beta$, and similarly, a temporary migration spanning a duration θ_m is associated with a utility $(\theta_m A_m(d))^\beta$. I do not include the consumption of other items, such as a tradable and/or a non-tradable good. This is arguably a strong assumption but one that allows to focus entirely on mobility choices that can be captured with CDR data, with respect to distance D_{od} and cost parameters. Other types of consumption would greatly complicate the algebra and also call for strong assumptions about the functional form for the utility function. Yet, they would not influence the shape of the relationship of interest between mobility choices and the distance between origin and destination. For simplicity, I also ignore the potential interactions between visits and migration choices and the model is silent about the degree of substitutability or complementarity between amenities consumed via visits and migration. This is an issue that I explored empirically in section 4.4 but which is not tackled conceptually by the present model. In short, I do not consider a unifying utility function that allows for the joint consumption of amenities via both visits and migration. Instead, I solve a utility maximization problem considering two distinct scenarios. First, I consider a situation in which an individual can only visit destination d and I solve the utility maximization problem that optimizes the number ν_v and duration θ_v of visits to d:

$$\begin{aligned} \underset{\nu_v, \theta_v}{\arg\max} \quad & \nu_v [\theta_v A_v(d)]^\beta \\ \text{s.t.} \quad & \nu_v (\lambda_v + \gamma D_{od} + \kappa_d \theta_v^\alpha) \leq w \end{aligned} \tag{4.4}$$

A necessary condition for the existence of some mobility immediately follows from the budget constraint by considering the cost of one visit of some minimum duration. For instance, θ_v is expressed in days and the duration of visits is measured in blocks of time equivalent to one day, so that the minimum value for θ_v is equal to 1. Then, the individual makes at least one visit if he can at least afford one visit of one day, i.e. if $w \geq \lambda_v + \gamma D_{od} + \kappa_d$.

Second, I examine the utility maximization problem of an individual who can only consume amenities in d by temporarily migrating to d:

$$\begin{aligned} \underset{\nu_m, \theta_m}{\arg\max} \quad & \nu_m [\theta_m A_m(d)]^\beta \\ \text{s.t.} \quad & \nu_m (\lambda_m + \gamma D_{od} + \kappa_d \theta_m^\alpha) \leq w \end{aligned} \tag{4.5}$$

where ν_m is the number of migration events and θ_m their duration. Here, the condition for movement depends on the minimum duration defining a temporary migration event τ^{temp}: $w \geq \lambda_m + \gamma_{od} + \kappa_d [\tau^{temp}]^\alpha$.

visits (or temporary migrations) serves as an auxiliary hypothesis to rationalize the phenomenon of individuals undertaking multiple excursions to an identical destination. In the conceptual framework elucidated in Chapter 1, this dimension is addressed through the duration-linked cost parameter, α.

A simple intuition already emerges from the movement conditions derived from this simple framework. The condition for seeing any migration is rendered more stringent by the fact that it imposes a larger cost associated with the minimum duration τ_{temp}. More importantly, the distinct fixed costs λ_m and λ_v may imply significant differences in the propensity to visit versus migrate to d. Interestingly, the data are consistent with a scenario in which λ_m is greater than λ_v since the fraction of individuals visiting cities is found to be much larger than the proportion of individuals migrating to cities.

Conditional on movement conditions being satisfied, solving the utility maximization problems above yields some expressions for the optimal number of visits v_v^* and duration θ_v^* (proof in Appendix 4.F):

$$v_v^* = \left(1 - \frac{\beta}{\alpha}\right) \frac{w}{\lambda_v + \gamma D_{od}} \tag{4.6}$$

$$\theta_v^* = \left(\frac{\lambda_v + \gamma D_{od}}{\kappa_d \left(\frac{\alpha}{\beta} - 1\right)}\right)^{\frac{1}{\alpha}} \tag{4.7}$$

Equivalent expressions are obtained for temporary migration choices v_m^* and θ_m^*, simply replacing λ_v by λ_m:

$$v_m^* = \left(1 - \frac{\beta}{\alpha}\right) \frac{w}{\lambda_m + \gamma D_{od}} \tag{4.8}$$

$$\theta_m^* = \left(\frac{\lambda_m + \gamma D_{od}}{\kappa_d \left(\frac{\alpha}{\beta} - 1\right)}\right)^{\frac{1}{\alpha}} \tag{4.9}$$

Taking logs on equations (4.6) to (4.9) and deriving with respect to the log of distance allows to obtain expressions for the distance elasticity of the frequency and duration of visits and temporary migrations, as well as of the total time spent on visits and temporary migration:

$$\frac{\partial \ln \nu_v^*}{\partial \ln D_{od}} = -\frac{\gamma D_{od}}{\lambda_v + \gamma D_{od}} \tag{4.10}$$

$$\frac{\partial \ln \theta_v^*}{\partial \ln D_{od}} = \frac{\gamma D_{od}}{\alpha(\lambda_v + \gamma D_{od})} \tag{4.11}$$

$$\frac{\partial \ln \nu_v^* \theta_v^*}{\partial \ln D_{od}} = -\left(1 - \frac{1}{\alpha}\right) \frac{\gamma D_{od}}{\lambda_v + \gamma D_{od}} \tag{4.12}$$

$$\frac{\partial \ln \nu_m^*}{\partial \ln D_{od}} = -\frac{\gamma D_{od}}{\lambda_m + \gamma D_{od}} \tag{4.13}$$

$$\frac{\partial \ln \theta_m^*}{\partial \ln D_{od}} = \frac{\gamma D_{od}}{\alpha(\lambda_m + \gamma D_{od})} \tag{4.14}$$

$$\frac{\partial \ln \nu_m^* \theta_m^*}{\partial \ln D_{od}} = -\left(1 - \frac{1}{\alpha}\right) \frac{\gamma D_{od}}{\lambda_m + \gamma D_{od}} \tag{4.15}$$

Equations (4.10) and (4.13) indicate that the number of visits and migration events is negatively related to the distance to destination. On the other hand, equations (4.11) and (4.14) show that the duration of visits and migration events is positively related to distance. Individuals will thus tend to make fewer but longer visits to more distant cities. The net effect on the total time on mobility is determined by equations (4.12) and (4.15): with $\alpha > 1$, the total time spent visiting or migrating to the destination decreases with distance. Therefore, the increased duration is not expected to compensate for the lower frequency with which individuals travel to more distant locales.

The expressions for the frequency and duration of visits and temporary migration in equations (4.6) to (4.9) do not necessarily provide convenient closed-form relations, but serve to highlight the importance of fixed costs in determining the sensitivity of mobility choices to the distance between origin and destination. Importantly, the distance elasticity of mobility does not merely appear as proportional to the marginal cost of distance. The various elasticities established in equations (4.10) to (4.15) depend negatively on the fixed costs λ_v and λ_m. This implies that the discernable distance elasticity of a given type of mobility, as perceived through CDR-based mobility measures, also carries information about the associated fixed costs, which cannot be directly observed. More specifically, the model anticipates that should temporary migration be characterized by elevated fixed costs relative to visits (i.e. $\lambda_m > \lambda_v$), then its distance elasticity will be lower than that of visits in absolute terms.

Contrary to the model introduced in Chapter 1, the conceptual framework presented here does not aim at rationalizing the observed mobility levels. Rather, it employs the same cost structure for both visits and temporary migration, within a streamlined environment, to highlight the relation between observable distance elasticities for visits and temporary migration to the intrinsic fixed costs of each mobility type. While the analysis in Chapter 1 ultimately focuses on the distance cost of mobility, neglecting other costs to derive tractable relationships between inter-city visiting flows and distance or city size, the model considered in this chapter upholds the comprehensive cost framework. This approach allows to relate the distance elasticity of mobility with the underlying fixed costs of mobility, offering an indirect way to uncovering differences in the fixed costs associated with visits and temporary migration.

4.5.2 Gravity estimates

As in section 4.4, I consider the mobility choices aggregated over a year for a large sample of users observed over the period going from February 2014 To January 2015 (*Subset 1*). I estimate standard gravity equations in order to test the main predictions implied by the conceptual framework of section 4.5.1. First, I run a Poisson Pseudo-Maximum likelihood (PPML) estimation of a gravity equation relating an individual i's frequency of visits or temporary migration to a destination d over a year, to the distance between i's residence location o and d, denoted D_{od}:

$$v_{iod} = exp(a_1 \ln D_{od} + \delta_o + \psi_d + \epsilon_{iod}) \tag{4.16}$$

where v_{iod} is either the total number visits or the number of temporary migrations made by i to d over the course of one year. δ_o and ψ_d are origin and destination fixed effects respectively.

Then, a similar log-log model relating the mean duration of visits or temporary migration, θ_{iod}, to the distance D_{od} is estimated with OLS on the subset of visitors and migrants to destination d respectively:

$$\ln \theta_{iod} = a_2 \ln D_{od} + \delta_o + \psi_d + \epsilon_{iod} \tag{4.17}$$

Finally, I run a PPML estimation of a gravity equation considering the total time spent visiting or migrating as a dependent variable, i.e. the product of v_{iod} and θ_{iod}:

$$v_{iod}\theta_{iod} = exp(a_3 \ln D_{od} + \delta_o + \psi_d + \epsilon_{iod}) \tag{4.18}$$

Results are presented in Table 4.5. Note that the distance between origin and destination locations corresponds to the actual travel distance calculated via

the Open Source Routing Machine (OSRM) project[26] that uses OpenStreetMap data.[27] Column (1) and (2) show the results of the PPML estimation of equation (4.16) considering the frequency of visits and temporary migration respectively. Consistent with the model prediction, both coefficients are negative and statistically significant. Moreover, the magnitude of the distance elasticity of the frequency of visits is practically twice as large as the elasticity of the frequency of temporary migration, in absolute terms. According to the conclusions drawn from the conceptual framework in section 4.5.1, this is consistent with the fixed cost of temporary migration being larger than the fixed cost associated with visits (i.e. $\lambda_m > \lambda_v$).

Estimations of equation (4.17) for the duration of visits and temporary migration events are reported in columns (3) and (4). As expected, the distance elasticity of visits' duration is positive and statistically significant, and suggests that a doubling in distance is associated with a 14% increase in the mean duration of visits on average. The distance elasticity of temporary migration duration is also positive and significant. The size of the estimated effect is much smaller in the temporary migration case, which again supports the hypothesis of a higher fixed cost for temporary migration compared to visits.

Finally, the estimated distance elasticities of the total time spent visiting and migrating are also consistent with expectations (columns (5) and (6)). Both coefficients are negative and significant and the elasticity of visits is larger than the elasticity of temporary migration in absolute terms. Consistent with the results of columns (1) and (3), the magnitude of the impact of distance on the total time spent on visits is marginally lower than the effect on the frequency of visits: when distance increases, the time loss induced by a reduction in the number of visits is partially compensated by longer visits. In the same vein, the distance elasticity of the time spent migrating is reduced compared to the elasticity with respect to the frequency of temporary migrations, although by a small margin given the relatively low impact of distance on the duration of migration events.

[26] Details can be found at http://project-osrm.org/. Note that I access the OSRM API via the *osrm* R package.

[27] Robustness checks considering other distance metrics are provided in appendix 4.G and show comparable results (Table 4.G.1).

Table 4.5: Gravity estimates for the frequency, duration and total time of visits and temporary migration.

	Frequency		**Duration**		**Total days**	
	Visits	Migration	Visits	Migration	Visits	Migration
	(1)	(2)	(3)	(4)	(5)	(6)
	PPML	PPML	OLS	OLS	PPML	PPML
log(distance)	-2.147***	-1.199***	0.1434***	0.0333***	-1.844***	-1.180***
	(0.0845)	(0.1150)	(0.0230)	(0.0090)	(0.0966)	(0.1140)
Observations	4,360,430	4,341,983	319,382	20,723	4,360,430	4,341,983
Pseudo R^2	0.51911	0.26740	0.13265	0.03170	0.44319	0.30160
Origin FE	✓	✓	✓	✓	✓	✓
Destination FE	✓	✓	✓	✓	✓	✓

Note: Estimations are based on a subset of 100,000 users from the high-quality sample, observed over the period February 2014-January 2015. Each observation corresponds to a user-destination couple and the mobility metrics are aggregated over the entire one-year period. Each column indicates a separate regression considering a distinct dependent variable. Columns (1) and (2) show PPML estimates from a regression of the number of visits and the number of temporary migration events observed over a year. Columns (3) and (4) present OLS estimates from a regression of the logged mean duration of visits and the logged mean duration of temporary migration events, considering the subset of user-destination pairs with at least one visit and one migration respectively. Columns (5) and (6) show PPML estimates from a regression of the time spent visiting and migrating over a year. The distance used corresponds to the travel distance by road. Standard errors are two-way clustered by origin and destination. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

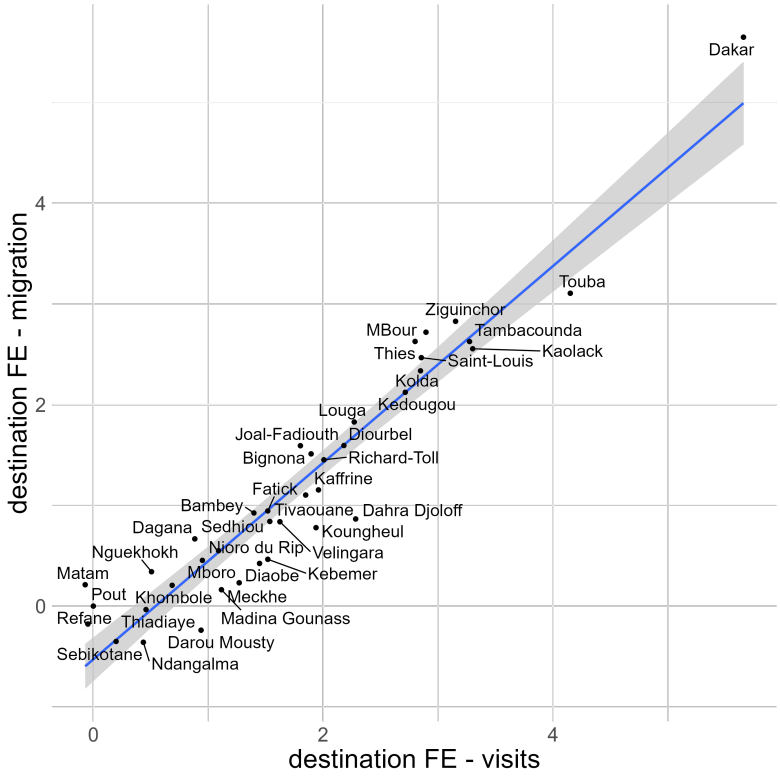
Some additional tests allow to appreciate the robustness of the results to the distance metric used and the time window considered. I reproduce the results of Table 4.5 using the great circle distance as well as the travel time by car[28] between origin and destination, and find that the results are practically unchanged (see Table 4.G.1 in Appendix 4.G). Also, I estimate the gravity equations considering mobility choices aggregated over distinct time windows, i.e. a window covering the one-year period going from October 2014 to September 2015 and another spanning the period February 2013 to November 2013. Results are again remarkably comparable to those obtained in Table 4.5 (see Table 4.G.2). Finally, I address potential concerns about the influence of high-frequency visitors originating from the outskirts of cities (i.e. commuters) on the results by considering another set of estimations where pairs of adjacent locations are excluded from the sample. The results are again largely consistent with estimations that include pairs of adjacent cells.

Finally, the destination fixed effects included in gravity estimations of Table 4.5

[28]The travel time by car is also calculated via the OSRM project.

have a useful interpretation that allows to complement the analyses of section 4.4. For instance, considering the estimation in column (5), each destination-specific coefficient represents the average time spent visiting the corresponding city, net of the effect of movement costs and of the average propensity to spend time in cities specific to each origin location. In other words, destination fixed effects reflect the level of attractiveness of cities for visitors from other locations ceteris paribus. Similarly, destination fixed effects from the estimation in column (6) give the relative level of attractiveness of cities for temporary migrants. As a result, the comparison of destination fixed effects derived from these estimations is informative about potential differences between urban amenities consumed via visits and temporary migration. Do some cities specifically attract visitors while others attract temporary migrants? This would indicate that visits are motivated by the consumption of amenities that some cities have and others do not, and likewise for temporary migration. Figure 4.6 plots the migration destination fixed effects against the visits destination fixed effects and tends to suggest this is not the case. All else equal, cities that attract visitors attract temporary migrants in the same proportions, compared to other cities. I find no evidence of some cities exhibiting a high destination fixed effect for visits and a relatively low destination fixed effect for migration, or conversely.

Figure 4.6: Attractiveness of cities to visitors and temporary migrants.



Note: Each observation represents a destination city. Destination fixed effects for visits are extracted from the regression given in column (5) of Table 4.5 where the total days on visits is regressed on the logged distance between origin and destination, origin fixed effects, and destination fixed effects. Destination fixed effects for temporary migration are likewise extracted from the estimation showed in column (6) of Table 4.5. The blue line represents a linear regression line between destination fixed effects for visits and migration, with the grey area showing the 95% confidence interval.

4.6 Discussion

The use of highly granular mobility measures derived from mobile phone data proves itself insightful to better understand the interplay between various types of temporary movements. The analyses of section 4.4 do not provide evidence in support of the hypothesis that frequent visits to cities could be used in substitution of longer-term but more costly movements that only some individuals could afford. Instead, the results suggest that a positive relationship exists between visits and temporary migration and is driven by temporary migrants making prospective and follow-up visits before and after a migration spell. The phone-derived mobility measures also allow to gain new insights into the differential costs associated with

visits and temporary migration, respectively. However, a number of limitations that could motivate future research are worth highlighting.

Observing patterns of movements only represents an indirect method for the characterization of urban amenities consumed during visits and temporary migration episodes. An undeniable advantage with the use of mobile phone data is the granularity with which human mobility can be measured, but it comes at a cost of a very limited amount of other information about users. In particular, it is not possible to clearly identify what individuals actually do when they travel to a city, the type of activities they conduct, or the type of goods and services they consume. Thus, the mobile phone data used in the analysis allow to identify novel questions about human mobility behaviors – such as the possible existence of prospective visiting behaviors among temporary migrants – but complementary survey-based data would be crucially needed to unequivocally confirm the actual motives underlying those movements.

As emphasized in section 4.4, the absence of further socio-economic information on users seriously complicates the elaboration of compelling causal claims about the relationship between visits and temporary migration. Yet, analyzing within-individual variations of visits over time with respect to the timing of migration provides evidence in support of the assumption that visits could be a precursor to migration. In this respect, a complementary analysis using mobile phone data could be envisaged to further explore this mechanism. Given that CDR records also indicate the identifier of the counterpart of the call, it is possible to examine whether users' social network is affected by observed visits at destination in a way that would confirm the prospective behavior hypothesis. For instance, do future migrants get new contacts at destination when they visit it prior to migrating?[29]

4.7 Conclusion

The paper explores the interplay between two distinct forms of temporary movements to cities: visits and temporary migration. I exploit a unique mobile phone dataset in Senegal that allows to capture both types of mobility for a large sample of users and at a high level of granularity. Visiting flows to cities are strikingly high and 80-90% of users are seen visiting at least one city over a year. Their significance

[29] In their study about the role of social network typology on migration, Blumenstock, Chi, et al. (2022) find no evidence of such anticipatory behaviors in the context of Rwanda over the period 2005-2009. However, one would be inclined to revisit this result in the context of Senegal given the observed patterns of visits around migration events.

raises questions about the role they play as a way to access urban amenities, relative to other types of movements. By comparing visiting and temporary migration choices of phone users, I am able to identify plausible ways in which those types of movements are interrelated and I discuss implications for the nexus between amenities consumed during visits versus temporary migration. I find no evidence supporting the idea that visits could substitute for longer-term temporary movements. The data do not allow to conclude that the nature of urban amenities consumed during visits and temporary migration somewhat overlap. Instead, the empirical analysis uncovers a positive relationship between visits and temporary migration to cities. This relationship is almost entirely driven by the additional visits that temporary migrants make at their migration destination. The paper shows that the user-level distribution of these visits over time are non-random with respect to the timing of migration. In particular, the results point to the possibility that visits could be a precursor to migration. Then, I analyze differences between visits and temporary migration from the perspective of movement costs associated with each type of mobility. A simple conceptual framework underlines the role of fixed costs in determining observed elasticities of visits and temporary migration with respect to distance. The estimation of gravity regressions show patterns supporting the idea that the fixed cost of visits is low compared to the fixed cost of migration.

Appendix 4.A Construction of subsets of phone users

I construct various subsets of phone users utilized throughout the paper, with the objective of streamlining individual-level data analysis and expediting computational procedures, while preserving a substantial sample size and ensuring comprehensive coverage. In what follows, I provide details about the construction of each of those datasets. While *Subset 1* serves as the principal dataset for the majority of this paper's analyses, *Subset 2* and *Subset 3* are principally curated to facilitate robustness tests. Meanwhile, *Subset 4* is designed to cover the entire 2013-2015 period and produce a complete time series of visits indicators over that period (see section 4.3).

4.A.1 Subset 1: 100,000 users from February 2014 to January 2015

First, users of the high-quality subset, which is defined in Chapter 2 (section 2.3.4), are selected from the raw 2014-2015 dataset. As a reminder, these users are characterized by their presence over a minimum duration of 330 days, with observations on at least 80% of those days and a maximal observational gap not exceeding 15 days.

Second, from the high-quality subset, users whose observation span encompasses the one-year period from February 1, 2014, to January 31, 2015, are selected.

Thirdly, from this refined subset, a sample of 100,000 users is randomly selected, according to a sampling scheme that follows the idea of the weighting scheme developed in Chapter 2 (section 2.5.1). Each of the 916 voronoi locations is associated with a third-level administrative unit based on a maximum population criterion.[30] On the other hand, voronoi cells are categorized into five distinct strata, consisting of one urban category and four rural sub-categories. The urban classification encompasses the 39 cells identified as city cells. The four rural divisions are delineated based on population density, spanning from the most sparsely populated rural cells to the densest rural locations, as detailed in Chapter 2 (section 2.A). Each urban cell serves as an individual sampling unit, while the rural sampling units are demarcated by a specific rural stratum within each administrative unit, resulting in a total of 329 defined sampling units. Within each sampling unit, the number of users selected is proportional to its population. It is worth noting that the population values for each cell are ascertained by juxtaposing the 2017 WorldPop population grid with the voronoi cells. A minimum of 100 users is imposed in each sampling unit. However, in instances where the total users within a sampling unit fall short of the number determined by the sampling scheme,

[30] Essentially, in instances where voronoi cells overlap multiple administrative units, the unit with the maximum population within the overlapping areas is selected.

all users from that particular unit are incorporated. Given these two adjustments, the final subset does not precisely encompass 100,000 users; instead, it comprises a total of 103,704 users.

4.A.2 Subset 2: 100,000 users from October 2014 to September 2015

This subset is defined based on the same criteria presented above, considering users from the 2014-2015 high-quality subset whose period of observation spans the one-year period from October 1, 2014, to September 30, 2015.

4.A.3 Subset 3: 100,000 users from January 2013 to December 2013

Similarly, a subset comprising 100,000 users is extracted from the 2013 high-quality subset, adhering to the same sampling scheme exposed above. Given that the 2013 subset spans precisely a year, no additional filters related to the users' observation duration are implemented.

4.A.4 Subset 4: 300,000 users from January 2013 to December 2015

Subset 4 is tailored to encompass data from the entire 2013-2015 span, with the aim of calculating a temporal series of visits indicators throughout this period. I simply use *Subset 3* to cover the year 2013, and I select a random sample of 200,000 users in the 2014-2015 dataset, following again the sampling scheme delineated above in section 4.A.1.

4.A.5 Subset 5: 200,000 users from February 2013 to September 2015

Subset 5 is tailored for the analysis conducted in section 4.4.2. It consists of a random selection of 200,000 users from the 2014-2015 high-quality subset, in line with the sampling framework detailed in section 4.A.1. A stringent observational span criterion is imposed, including only users with records spanning at least from February 1, 2014, to September 30, 2015. This expansive timeframe allows to observe temporary migration events over an entire year, from June 2014 to May 2015, along with the visiting choices both in the four months leading up to migration departures and in the four months subsequent to returns from migration.

Appendix 4.B Tables of summary statistics on visits over the period February 2014-January 2015

Table 4.B.1: Fraction of users visiting urban destinations, by zone of origin.

Origin zone	To any city	To Dakar	To Primary cities	To secondary cities
Urban	81.7%	32.4%	66.9%	57.2%
Dakar	77.6%	0.0%	67.9%	54.5%
Primary cities	84.3%	59.2%	61.7%	59.9%
Secondary cities	86.5%	56.7%	77.3%	57.8%
Rural	85.1%	40.1%	70.5%	65.6%
Very dense rural	81.9%	39.6%	66.8%	61.7%
Dense rural	86.4%	40.4%	73.2%	64.4%
Remote rural	88.6%	43.3%	75.5%	69.0%
Very remote rural	84.5%	35.5%	66.6%	71.1%
Total	83.4%	36.3%	68.7%	61.4%

Note: The table shows the fractions of users with at least one visit to any city (column (1)), to Dakar (column (2)), to a primary city (column (3)) and to a secondary city (column (4)), broken down by zone of origin. Estimations are based on a random sample of 100,000 users in the high-quality subset observed over the period February 2014-January 2015 (see details in appendix 4.A).

Table 4.B.2: Return period of visits to urban destinations (in days).

	mean	25% quantile	median	75% quantile	90% quantile
To any city	84	15	40	105	224
To Dakar	165	56	122	309	361
To primary cities	133	32	83	186	357
To secondary cities	153	37	107	282	368

Note: For each user-destination, the frequency of visits is calculated as a return period that corresponds to the number of days observed divided by the number of visits, i.e. the average length of time between consecutive visits to that destination. The table shows the mean and quantiles of the frequency of visits, broken down by urban destination. Estimations are based on a random sample of 100,000 users in the high-quality subset observed over the period February 2014-January 2015 (see details in appendix 4.A).

Table 4.B.3: Median frequency of visits to urban destinations, by zone of origin.

Origin zone	To any city	To Dakar	To Primary cities	To secondary cities
Urban	64	115	119	163
Dakar	88	-	128	178
Primary cities	59	122	133	161
Secondary cities	34	90	73	117
Rural	24	128	56	70
Very dense rural	31	137	68	78
Dense rural	24	130	51	82
Remote rural	20	120	48	64
Very remote rural	18	145	52	45
National	40	122	83	107

Note: For each user-destination, the frequency of visits is calculated as a return period that corresponds to the number of days observed divided by the number of visits, i.e. the average length of time between consecutive visits to that destination. The table shows the median frequency of visits among users identified as visitors, broken down by origin zone and urban destination. Estimations are based on a random sample of 100,000 users in the high-quality subset observed over the period February 2014-January 2015 (see details in appendix 4.A).

Table 4.B.4: Observed duration of visits to urban destinations (in days).

	mean	1st quartile	median	3rd quartile	90% quantile
To any city	1.47	0.64	1.00	1.75	2.97
To Dakar	2.71	0.75	1.67	3.50	6.25
To primary cities	1.41	0.50	0.90	1.67	3.00
To secondary cities	0.92	0.50	0.56	1.00	1.67

Note: The observed duration of a visit is defined as the time elapsed between the start and end dates of the visit. For each user, I calculate the average observed duration of visits to each urban destination category. The table shows the mean and quantiles of the user-level average duration, broken down by destination category. Estimations are based on a random sample of 100,000 users in the high-quality subset observed over the period February 2014-January 2015 (see details in appendix 4.A).

Table 4.B.5: Maximum duration of visits to urban destinations (in days).

	mean	**1st quartile**	**median**	**3rd quartile**	**90% quantile**
To any city	1.93	0.97	1.45	2.35	3.75
To Dakar	3.28	1.00	2.20	4.50	7.50
To primary cities	1.84	0.81	1.28	2.25	3.75
To secondary cities	1.30	0.58	1.00	1.50	2.50

Note: The maximum duration of a visit is defined as the time elapsed between the observation preceding the visit start date and the observation following the visit end date. For each user, I calculate the average maximum duration of visits to each urban destination category. The table shows the mean and quantiles of the user-level average duration, broken down by destination category. Estimations are based on a random sample of 100,000 users in the high-quality subset observed over the period February 2014-January 2015 (see details in appendix 4.A).

Table 4.B.6: Statistics on the number of days of visit to urban destinations.

	mean	**25% quantile**	**median**	**75% quantile**	**90% quantile**
To any city	23.4	3.9	11.9	30.8	61.7
To Dakar	13.1	2.1	6.4	16.4	33.3
To primary cities	13.2	1.8	5.0	14.8	35.8
To secondary cities	9.2	1.0	2.6	8.6	25.2

Note: For each user, the total number of days of visit to a city is calculated as the sum across the distinct visits to that city of the observed duration of those visits. Then, the total number of visit-days by city is summed across all cities, all primary cities, and all secondary cities. The table shows the mean and quantiles across all users of the total number of visit-days to all cities, to Dakar, to primary cities, and to secondary cities. Note that since those measures utilize the observed duration rather than the maximum duration, they shall be considered lower-bound estimates for the total number of visit-days. Estimations are based on a random sample of 100,000 users in the high-quality subset observed over the period February 2014-January 2015 (see details in appendix 4.A).

Table 4.B.7: Average number of days of visits to urban destinations, by zone of origin.

Origin zone	To any city	To Dakar	To Primary cities	To secondary cities
Urban	17.1	12.5	9.6	6.2
Dakar	12.6	-	9.8	5.7
Primary cities	18.4	11.8	8.1	5.8
Secondary cities	25.5	14.2	12.0	8.3
Rural	29.3	13.6	16.6	11.8
Very dense rural	26.1	13.2	13.8	11.4
Dense rural	29.4	13.2	18.1	10.6
Remote rural	31.1	14.5	17.8	11.3
Very remote rural	33.1	13.7	18.3	15.4
National	23.4	13.1	13.2	9.2

Note: For each user, the total number of days of visit to a city is calculated as the sum across the distinct visits to that city of the observed duration of those visits. Then, the total number of visit-days is summed across all cities, all primary cities, and all secondary cities. The table shows the average across users of the total number of visit-days to all cities, to Dakar, to primary cities, and to secondary cities, broken down by zone of origin. Note that since those measures utilize the observed duration rather than the maximum duration, they shall be considered lower-bound estimates for the total number of visit-days. Estimations are based on a random sample of 100,000 users in the high-quality subset observed over the period February 2014-January 2015 (see details in appendix 4.A).

Appendix 4.C Tables of summary statistics on visits in 2013

Table 4.C.1: Fraction of users visiting urban destinations, by zone of origin.

Origin zone	To any city	To Dakar	To Primary cities	To secondary cities
Urban	84.7%	35.2%	69.8%	59.6%
Dakar	79.9%	0.0%	70.0%	56.2%
Primary cities	88.0%	64.2%	64.8%	63.2%
Secondary cities	90.3%	62.0%	81.9%	60.6%
Rural	89.3%	46.1%	74.5%	69.5%
Very dense rural	85.4%	44.5%	70.0%	64.8%
Dense rural	90.1%	46.3%	76.0%	67.6%
Remote rural	93.1%	49.5%	80.8%	73.4%
Very remote rural	91.3%	44.6%	72.9%	77.5%
Total	87.1%	40.7%	72.2%	64.6%

Note: The table shows the fractions of users with at least one visit to any city (column (1)), to Dakar (column (2)), to a primary city (column (3)) and to a secondary city (column (4)), broken down by zone of origin. Estimations are based on a random sample of 100,000 users in the 2013 high-quality subset (see details in appendix 4.A).

Table 4.C.2: Return period of visits to urban destinations (in days).

	mean	**25% quantile**	**median**	**75% quantile**	**90% quantile**
To any city	81	15	40	105	183
To Dakar	164	56	120	317	353
To primary cities	128	32	84	178	345
To secondary cities	150	39	111	310	354

Note: For each user-destination, the frequency of visits is calculated as a return period that corresponds to the number of days observed divided by the number of visits, i.e. the average length of time between consecutive visits to that destination. The table shows the mean and quantiles of the frequency of visits, broken down by urban destination. Estimations are based on a random sample of 100,000 users in the 2013 high-quality subset (see details in appendix 4.A).

Table 4.C.3: Median frequency of visits to urban destinations, by zone of origin.

Origin zone	**To any city**	**To Dakar**	**To Primary cities**	**To secondary cities**
Urban	61	115	116	166
Dakar	86	-	120	176
Primary cities	55	118	121	163
Secondary cities	34	106	70	117
Rural	24	150	56	72
Very dense rural	32	158	70	83
Dense rural	25	150	52	84
Remote rural	21	121	47	68
Very remote rural	18	121	52	49
National	40	120	84	111

Note: For each user-destination, the frequency of visits is calculated as a return period that corresponds to the number of days observed divided by the number of visits, i.e. the average length of time between consecutive visits to that destination. The table shows the median frequency of visits among users identified as visitors, broken down by origin zone and urban destination. Estimations are based on a random sample of 100,000 users in the 2013 high-quality subset (see details in appendix 4.A).

Table 4.C.4: Observed duration of visits to urban destinations (in days).

	mean	**1st quartile**	**median**	**3rd quartile**	**90% quantile**
To any city	1.45	0.64	1.00	1.74	2.89
To Dakar	2.62	0.70	1.50	3.43	6.17
To primary cities	1.39	0.50	0.88	1.62	2.92
To secondary cities	0.92	0.50	0.56	1.00	1.67

Note: The observed duration of a visit is defined as the time elapsed between the start and end dates of the visit. For each user, I calculate the average observed duration of visits to each urban destination category. The table shows the mean and quantiles of the user-level average duration, broken down by destination category. Estimations are based on a random sample of 100,000 users in the 2013 high-quality subset (see details in appendix 4.A).

Table 4.C.5: Maximum duration of visits to urban destinations (in days).

	mean	**1st quartile**	**median**	**3rd quartile**	**90% quantile**
To any city	1.84	0.94	1.38	2.23	3.50
To Dakar	3.03	1.00	2.00	4.00	7.00
To primary cities	1.76	0.78	1.25	2.10	3.50
To secondary cities	1.27	0.56	0.97	1.50	2.47

Note: The maximum duration of a visit is defined as the time elapsed between the observation preceding the visit start date and the observation following the visit end date. For each user, I calculate the average maximum duration of visits to each urban destination category. The table shows the mean and quantiles of the user-level average duration, broken down by destination category. Estimations are based on a random sample of 100,000 users in the 2013 high-quality subset (see details in appendix 4.A).

Table 4.C.6: Statistics on the number of days of visit to urban destinations.

	mean	**25% quantile**	**median**	**75% quantile**	**90% quantile**
To any city	23.0	4.0	12.0	30.4	60.0
To Dakar	12.3	2.1	6.0	15.5	30.9
To primary cities	12.9	1.8	5.1	14.6	34.9
To secondary cities	8.8	1.0	2.6	8.3	23.9

Note: For each user, the total number of days of visit to a city is calculated as the sum across the distinct visits to that city of the observed duration of those visits. Then, the total number of visit-days by city is summed across all cities, all primary cities, and all secondary cities. The table shows the mean and quantiles across all users of the total number of visit-days to all cities, to Dakar, to primary cities, and to secondary cities. Note that since those measures utilize the observed duration rather than the maximum duration, they shall be considered lower-bound estimates for the total number of visit-days. Estimations are based on a random sample of 100,000 users in the 2013 high-quality subset (see details in appendix 4.A).

Table 4.C.7: Average number of days of visits to urban destinations, by zone of origin.

Origin zone	**To any city**	**To Dakar**	**To Primary cities**	**To secondary cities**
Urban	16.9	12.0	9.4	6.0
Dakar	11.8	-	9.2	5.3
Primary cities	18.2	11.4	7.8	5.7
Secondary cities	26.8	13.5	13.2	8.3
Rural	28.6	12.5	16.0	11.3
Very dense rural	25.2	12.2	13.3	10.5
Dense rural	28.5	12.0	17.4	10.2
Remote rural	30.6	13.3	17.3	10.8
Very remote rural	32.8	12.6	17.5	14.9
National	23.0	12.3	12.9	8.8

Note: For each user, the total number of days of visit to a city is calculated as the sum across the distinct visits to that city of the observed duration of those visits. Then, the total number of visit-days is summed across all cities, all primary cities, and all secondary cities. The table shows the average across users of the total number of visit-days to all cities, to Dakar, to primary cities, and to secondary cities, broken down by zone of origin. Note that since those measures utilize the observed duration rather than the maximum duration, they shall be considered lower-bound estimates for the total number of visit-days. Estimations are based on a random sample of 100,000 users in the 2013 high-quality subset (see details in appendix 4.A).

Appendix 4.D The empirical relationship between visits and temporary migration to cities: additional results

Table 4.D.1: Relationship between aggregate visits and temporary migration controlling for origin fixed effects, with heterogeneity by zone of origin.

	No. of visits (1)	No. of visit-days (2)
Migration dummy × rural origin	5.422***	16.61***
	(0.3287)	(0.3933)
Migration dummy × urban origin	8.213***	18.76***
	(0.5786)	(0.7799)
Observations	113,452	113,452
Pseudo R^2	0.03054	0.02721
Origin FE	✓	✓

Note: Each observation represents a user with mobility measures aggregated over the period of observation and across all destination cities. Column (1) shows the PPML estimation of a regression of the total number of visits to cities on a migration dummy equal to 1 if the user has at least one temporary migration event to any city, interacted with categorical variable indicating the zone to which the origin location belongs. Column (2) shows the same estimation considering the total time spent visiting cities as a dependent variable. Standard errors are clustered by origin location.

Table 4.D.2: Relationship between aggregate visits and temporary migration controlling for origin fixed effects, with heterogeneity by sub-zone of origin.

	No. of visits (1)	No. of visit-days (2)
Migration dummy × Primary cities	8.125***	18.69***
	(0.6864)	(0.9654)
Migration dummy × Sec. cities	8.520***	19.02***
	(1.188)	(1.164)
Migration dummy × V. dense rural	5.087***	15.04***
	(0.5696)	(0.7514)
Migration dummy × Dense rural	5.638***	16.91***
	(0.5606)	(0.7249)
Migration dummy × V. remote rural	5.318***	17.79***
	(0.7571)	(0.7472)
Migration dummy × Remote rural	5.967***	17.88***
	(0.8029)	(0.8381)
Observations	113,452	113,452
Pseudo R^2	0.03054	0.02724
Origin FE	✓	✓

Note: Each observation represents a user with mobility measures aggregated over the period of observation and across all destination cities. Column (1) shows the PPML estimation of a regression of the total number of visits to cities on a migration dummy equal to 1 if the user has at least one temporary migration event to any city, interacted with categorical variable indicating the sub-zone to which the origin location belongs. Column (2) shows the same estimation considering the total time spent visiting cities as a dependent variable. Standard errors are clustered by origin location.

Table 4.D.3: Relationship between aggregate visits and temporary migration controlling for origin fixed effects, with heterogeneity by region of origin.

	No. of visits (1)	No. of visit-days (2)
Migration dummy × Dakar	9.877***	20.25***
	(0.3943)	(0.2452)
Migration dummy × Diourbel	4.672**	17.51***
	(1.938)	(1.288)
Migration dummy × Fatick	3.982***	14.99***
	(1.017)	(1.324)
Migration dummy × Kaffrine	7.424***	19.59***
	(0.8975)	(1.152)
Migration dummy × Kaolack	4.099***	15.48***
	(0.7078)	(0.8895)
Migration dummy × Kédougou	7.934***	22.32***
	(1.117)	(1.611)
Migration dummy × Kolda	6.938***	21.12***
	(0.7068)	(1.508)
Migration dummy × Louga	6.689***	18.57***
	(0.8547)	(0.9332)
Migration dummy × Matam	3.893***	9.844***
	(0.5322)	(1.056)
Migration dummy × Saint-Louis	5.468***	15.28***
	(0.9408)	(1.237)
Migration dummy × Sédhiou	4.548***	13.43***
	(0.4336)	(0.7188)
Migration dummy × Tambacounda	6.830***	17.30***
	(0.6786)	(1.071)
Migration dummy × Thiès	9.579***	20.71***
	(0.9912)	(0.8259)
Migration dummy × Ziguinchor	4.789***	13.25***
	(0.5817)	(1.203)
Observations	113,452	113,452
Pseudo R^2	0.03063	0.02747
Origin FE	✓	✓

Note: Each observation represents a user with mobility measures aggregated over the period of observation and across all destination cities. Column (1) shows the PPML estimation of a regression of the total number of visits to cities on a migration dummy equal to 1 if the user has at least one temporary migration event to any city, interacted with categorical variable indicating the region to which the origin location belongs. Column (2) shows the same estimation considering the total time spent visiting cities as a dependent variable. Standard errors are clustered by origin location.

Table 4.D.4: Relationship between visits and temporary migration controlling for origin-destination fixed effects, with heterogeneity by zone of origin.

	No. of visits (1)	No. of visit-days (2)
Migration dummy × rural origin	5.754***	14.07***
	(0.2275)	(0.2723)
Migration dummy × urban origin	5.981***	15.17***
	(0.4828)	(0.7400)
Observations	4,424,628	4,424,628
Pseudo R^2	0.07478	0.06607
Origin-destination FE	✓	✓

Note: Each observation represents a user-destination couple with mobility measures aggregated over the period of observation. Column (1) shows the PPML estimation of a regression of the total number of visits to a destination on a migration dummy equal to 1 if the user has at least one temporary migration event to that destination, interacted with categorical variable indicating the zone to which the origin location belongs. Column (2) shows the same estimation considering the total time spent visiting the destination as a dependent variable. Standard errors are clustered by origin-destination pair.

Table 4.D.5: Relationship between visits and temporary migration controlling for origin-destination fixed effects, with heterogeneity by sub-zone of origin.

	No. of visits (1)	No. of visit-days (2)
Migration dummy × Primary cities	5.655***	15.11***
	(0.6914)	(1.171)
Migration dummy × Sec. cities	7.024***	15.34***
	(0.9365)	(0.9019)
Migration dummy × V. dense rural	5.257***	12.77***
	(0.4106)	(0.5381)
Migration dummy × Dense rural	5.322***	13.53***
	(0.3965)	(0.5404)
Migration dummy × V. remote rural	6.410***	15.80***
	(0.4653)	(0.6007)
Migration dummy × Remote rural	6.661***	15.30***
	(0.5240)	(0.5579)
Observations	4,424,628	4,424,628
Pseudo R^2	0.07481	0.06615
Origin-destination FE	✓	✓

Note: Each observation represents a user-destination couple with mobility measures aggregated over the period of observation. Column (1) shows the PPML estimation of a regression of the total number of visits to a destination on a migration dummy equal to 1 if the user has at least one temporary migration event to that destination, interacted with categorical variable indicating the sub-zone to which the origin location belongs. Column (2) shows the same estimation considering the total time spent visiting the destination as a dependent variable. Standard errors are clustered by origin-destination pair.

Table 4.D.6: Relationship between visits and temporary migration controlling for origin-destination fixed effects, with heterogeneity by region of origin.

	No. of visits (1)	No. of visit-days (2)
Migration dummy × Dakar	7.842***	17.87***
	(0.9771)	(1.657)
Migration dummy × Diourbel	6.045***	15.46***
	(0.8971)	(0.7413)
Migration dummy × Fatick	5.153***	14.14***
	(0.4879)	(0.7397)
Migration dummy × Kaffrine	5.808***	14.88***
	(0.6615)	(0.7665)
Migration dummy × Kaolack	5.131***	13.41***
	(0.6689)	(0.8321)
Migration dummy × Kédougou	5.373***	17.04***
	(0.5903)	(1.215)
Migration dummy × Kolda	4.132***	14.02***
	(0.5445)	(0.9514)
Migration dummy × Louga	6.609***	15.63***
	(0.8033)	(0.8394)
Migration dummy × Matam	2.224***	7.808***
	(0.3982)	(0.6521)
Migration dummy × Saint-Louis	5.197***	13.84***
	(0.6756)	(0.7846)
Migration dummy × Sédhiou	3.069***	9.332***
	(0.3507)	(0.6592)
Migration dummy × Tambacounda	3.940***	12.30***
	(0.4923)	(0.7584)
Migration dummy × Thiès	9.758***	18.26***
	(0.8568)	(0.6271)
Migration dummy × Ziguinchor	3.412***	9.866***
	(0.6612)	(1.115)
Observations	4,424,628	4,424,628
Pseudo R^2	0.07518	0.06678
Origin-destination FE	✓	✓

Note: Each observation represents a user-destination couple with mobility measures aggregated over the period of observation. Column (1) shows the PPML estimation of a regression of the total number of visits to a destination on a migration dummy equal to 1 if the user has at least one temporary migration event to that destination, interacted with categorical variable indicating the region to which the origin location belongs. Column (2) shows the same estimation considering the total time spent visiting the destination as a dependent variable. Standard errors are clustered by origin-destination pair.

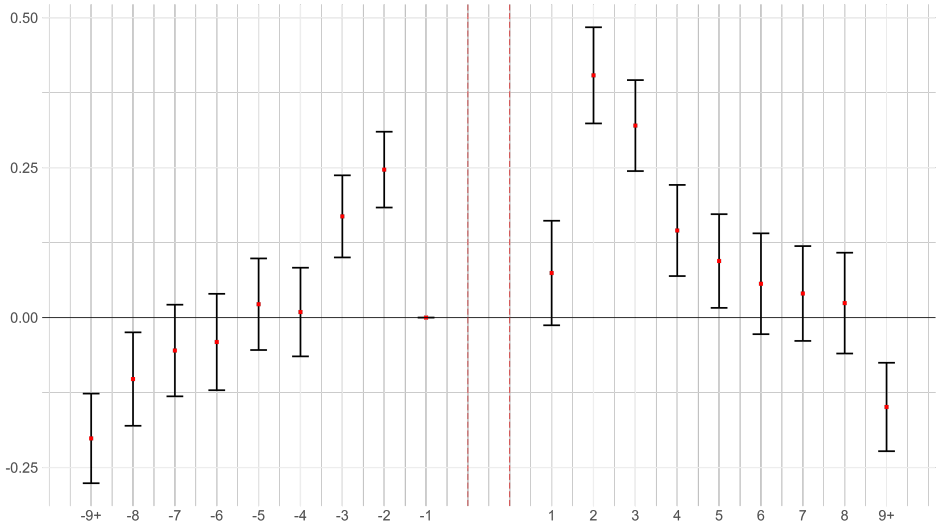
Table 4.D.7: Relationship between visits and temporary migration controlling for origin-destination fixed effects, with heterogeneity according to the distance to destination.

	No. of visits (1)	No. of visit-days (2)
Migration dummy × distance≤50	15.95***	23.81***
	(0.7553)	(0.6322)
Migration dummy × distance≤100	7.966***	19.28***
	(0.5222)	(0.8100)
Migration dummy × distance≤200	4.511***	14.19***
	(0.2838)	(0.4316)
Migration dummy × distance>200	2.469***	9.592***
	(0.1012)	(0.2771)
Observations	4,424,628	4,424,628
Pseudo R^2	0.07657	0.06816
Origin-destination FE	✓	✓

Note: Each observation represents a user-destination couple with mobility measures aggregated over the period of observation. Column (1) shows the PPML estimation of a regression of the total number of visits to a destination on a migration dummy equal to 1 if the user has at least one temporary migration event to that destination, interacted with a categorized distance variable. Column (2) shows the same estimation considering the total time spent visiting the destination as a dependent variable. Standard errors are clustered by origin-destination pair.

Appendix 4.E The dynamics of visits around migration events: additional results

Figure 4.E.1: Relative distribution of migrants' visits to destination before and after a migration event, logit model.



Note: The two vertical dashed red lines in the center of the graph represent the migration departure (left) and return (right). The x-axis represents the relative time with respect to the migration departure and return. Note that the gap between migration departure and return theoretically coincides with the migration duration but is normalized to one dekad for representation purposes. Red dots on the left-hand side correspond to estimates of $\alpha_2, ..., \alpha_8, \alpha_{9+}$ from a logit model, and those on the right-hand side are estimates of $\beta_1, ..., \beta_8, \beta_{9-}$. Coefficients are estimated on a sample of 30,523 unique temporary migrants observed over the period 2014-2015, resulting in a grand total of 1,358,184 observations. The error bars show the 95% confidence intervals based on two-way clustered standard errors at the individual-destination and dekad-destination levels.

Appendix 4.F Conceptual framework: proofs

I first consider the utility maximization problem associated with visits choices to a destination d for an individual residing in location o:

$$\begin{aligned} \underset{\nu_v, \theta_v}{\arg\max} \quad & \nu_v [\theta_v A_v(d)]^\beta \\ \text{s.t.} \quad & \nu_v (\lambda_v + \gamma D_{od} + \kappa_d \theta_v^\alpha) \leq w \end{aligned} \tag{4.19}$$

If the condition for movement is not satisfied (i.e. $w < \lambda_v + \gamma D_{od} + \kappa_d$), it follows trivially that $\nu_v = \theta_v = 0$. Otherwise, with the Lagrange multiplier denoted as L, the Lagrangian optimization problem can be written as:

$$\mathcal{L}(\nu_v, \theta_v) = \nu_v [\theta_v A_v(d)]^\beta + L[w - \nu_v (\lambda_v + \gamma D_{od} + \kappa_d \theta_v^\alpha)] \tag{4.20}$$

To simplify the derivation, I treat ν_v as a real number and I then calculate the partial derivatives of $\mathcal{L}(\nu_v, \theta_v)$ with respect to ν_v and θ_v:

$$\frac{\partial \mathcal{L}}{\partial \theta_v} = \beta \nu_v A_v(d)^\beta \theta_v^{\beta-1} - L\nu_v \kappa_d \alpha \theta_v^{\alpha-1} \tag{4.21}$$

$$\frac{\partial \mathcal{L}}{\partial \nu_v} = [\theta_v A_v(d)]^\beta - L(\lambda_v + \gamma D_{od} + \kappa_d \theta_v^\alpha) \tag{4.22}$$

The optimal choice (ν_v^*, θ_v^*) satisfies the condition:

$$\beta \nu_v^* A_v(d)^\beta \theta_v^{*\beta-1} - L\nu_v^* \kappa_d \alpha \theta_v^{*\alpha-1} = 0 \tag{4.23}$$

$$[\theta_v^* A_v(d)]^\beta - L(\lambda_v + \gamma D_{od} + \kappa_d \theta_v^{*\alpha}) = 0 \tag{4.24}$$

Multiplying (4.23) by θ_v^*, dividing by ν_v^*, and rearranging terms yields an expression that depends only on θ_v^* and exogenous parameters:

$$\theta_v^{*\beta} = \frac{L\kappa_d \alpha}{\beta A_v(d)^\beta} \tag{4.25}$$

Plugging in (4.25) into (4.24) allows to obtain the following expression for :

$$\theta_v^* = \left(\frac{\lambda_v + \gamma D_{od}}{\kappa_d \left(\frac{\alpha}{\beta} - 1 \right)} \right)^{\frac{1}{\alpha}} \tag{4.26}$$

Then, plugging in (4.26) into the budget constraint yields an expression for ν_v^*:[31]

$$\nu_v^* = \left(1 - \frac{\beta}{\alpha} \right) \frac{w}{\lambda_v + \gamma D_{od}} \tag{4.27}$$

Then, the accumulated time of visits $\nu_v^* \theta_v^*$ is obtained by multiplying (4.26) and (4.27):

$$\nu_v^* \theta_v^* = \frac{1}{\kappa_d^{\frac{1}{\alpha}}} \frac{\beta}{\alpha} \left(\frac{\alpha}{\beta} - 1 \right)^{1-\frac{1}{\alpha}} \frac{w}{(\lambda_v + \gamma D_{od})^{1-\frac{1}{\alpha}}} \tag{4.28}$$

[31] The budget constraint is satisfied with equality since an individual can always keep the number of visits fixed and adjust the duration θ_v.

Appendix 4.G Gravity estimates: robustness checks

Table 4.G.1: Gravity estimations for the period February 2014-January 2015, different distance metrics.

	Frequency		**Duration**		**Total days**	
	Visits	Migration	Visits	Migration	Visits	Migration
	(1)	(2)	(3)	(4)	(5)	(6)
	PPML	PPML	OLS	OLS	PPML	PPML
Panel A: Great circle distance						
log(distance)	-2.195***	-1.221***	0.1443***	0.0346***	-1.879***	-1.201***
	(0.0937)	(0.1166)	(0.0242)	(0.0091)	(0.1053)	(0.1151)
Observations	4,360,430	4,341,983	319,382	20,723	4,360,430	4,341,983
Pseudo R^2	0.51712	0.26651	0.13239	0.03172	0.44095	0.30060
Panel B: Travel distance						
log(distance)	-2.147***	-1.199***	0.1434***	0.0333***	-1.844***	-1.180***
	(0.0845)	(0.1150)	(0.0230)	(0.0090)	(0.0966)	(0.1140)
Observations	4,360,430	4,341,983	319,382	20,723	4,360,430	4,341,983
Pseudo R^2	0.51911	0.26740	0.13265	0.03170	0.44319	0.30160
Panel C: Travel time						
log(travel time)	-2.375***	-1.254***	0.1630***	0.0340***	-1.990***	-1.234***
	(0.1186)	(0.1369)	(0.0263)	(0.0095)	(0.1284)	(0.1372)
Observations	4,360,430	4,341,983	319,382	20,723	4,360,430	4,341,983
Pseudo R^2	0.50777	0.26458	0.13348	0.03164	0.43282	0.29862
Origin FE	✓	✓	✓	✓	✓	✓
Destination FE	✓	✓	✓	✓	✓	✓

Note: Estimations are based on a subset of 100,000 users from the high-quality sample, observed over the period February 2014-January 2015. Each observation corresponds to a user-destination couple and mobility metrics are aggregated over the entire period. Each column in each panel indicates a separate regression. Columns (1) and (2) show PPML estimates from a regression of the number of visits and the number of temporary migration events observed over a year. Columns (3) and (4) present OLS estimates from a regression of the logged mean duration of visits and the logged mean duration of temporary migration events, considering the subset of user-destination pairs with at least one visit and one migration respectively. Columns (5) and (6) show PPML estimates from a regression of the time spent visiting and migrating over a year. Panel A, B and C show results considering different measures of distance between origin and destination: the great circle, the travel distance by road and the travel time by car, respectively. Standard errors are two-way clustered by origin and destination. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 4.G.2: Gravity equations estimated on different time windows.

	Frequency		**Duration**		**Total days**	
	Visits	Migration	Visits	Migration	Visits	Migration
	(1)	(2)	(3)	(4)	(5)	(6)
	PPML	PPML	OLS	OLS	PPML	PPML
Panel A: February 2013 - November 2013						
log(Travel distance)	-2.133***	-1.250***	0.1339***	0.0330***	-1.848***	-1.219***
	(0.0808)	(0.1046)	(0.0211)	(0.0077)	(0.0936)	(0.1058)
Observations	3,986,201	3,978,557	263,847	21,240	3,986,201	3,978,557
Pseudo R^2	0.52039	0.28357	0.11645	0.03232	0.44756	0.33047
Panel B: February 2014 - January 2015						
log(Travel distance)	-2.147***	-1.199***	0.1434***	0.0333***	-1.844***	-1.180***
	(0.0845)	(0.1150)	(0.0230)	(0.0090)	(0.0966)	(0.1140)
Observations	4,360,430	4,341,983	319,382	20,723	4,360,430	4,341,983
Pseudo R^2	0.51911	0.26740	0.13265	0.03170	0.44319	0.30160
Panel C: October 2014 - September 2015						
log(Travel distance)	-2.145***	-1.155***	0.1493***	-0.0174*	-1.829***	-1.210***
	(0.0826)	(0.1180)	(0.0243)	(0.0100)	(0.0958)	(0.1232)
Observations	4,364,473	4,351,642	322,625	21,282	4,364,473	4,351,642
Pseudo R^2	0.51682	0.26470	0.13397	0.02370	0.43634	0.29746
Origin FE	✓	✓	✓	✓	✓	✓
Destination FE	✓	✓	✓	✓	✓	✓

Note: Estimations are based on subsets of 100,000 users from the high-quality sample, for different time periods. Each observation corresponds to a user-destination couple and mobility metrics are aggregated over the entire period considered. Each column in each panel indicates a separate regression. Columns (1) and (2) show PPML estimates from a regression of the number of visits and the number of temporary migration events observed over a year. Columns (3) and (4) present OLS estimates from a regression of the logged mean duration of visits and the logged mean duration of temporary migration events, considering the subset of user-destination pairs with at least one visit and one migration respectively. Columns (5) and (6) show PPML estimates from a regression of the time spent visiting and migrating over a year. Panel A shows results for the period February 2013 - November 2013, panel B considers the period February 2014 - January 2015 and panel C presents results with users observed over the period October 2014 - September 2015. Standard errors are two-way clustered by origin and destination. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 4.G.3: Gravity equations estimated on the period February 2014-January 2015, excluding pairs of adjacent locations.

	Frequency		**Duration**		**Total days**	
	Visits	Migration	Visits	Migration	Visits	Migration
	(1)	(2)	(3)	(4)	(5)	(6)
	PPML	PPML	OLS	OLS	PPML	PPML
log(distance)	-2.054***	-1.114***	0.1664***	0.0423***	-1.732***	-1.073***
	(0.1008)	(0.1260)	(0.0272)	(0.0103)	(0.1087)	(0.1215)
Observations	4,340,190	4,318,346	303,812	19,674	4,340,190	4,318,346
Pseudo R^2	0.43149	0.26612	0.12992	0.03286	0.39569	0.29976
Origin FE	✓	✓	✓	✓	✓	✓
Destination FE	✓	✓	✓	✓	✓	✓

Note: Estimations are based on a subset of 100,000 users from the high-quality sample, observed over the period February 2014-January 2015. Each observation corresponds to a user-destination couple and mobility metrics are aggregated over the entire period. Each column indicates a separate regression. Columns (1) and (2) show PPML estimates from a regression of the number of visits and the number of temporary migration events observed over a year. Columns (3) and (4) present OLS estimates from a regression of the logged mean duration of visits and the logged mean duration of temporary migration events, considering the subset of user-destination pairs with at least one visit and one migration respectively. Columns (5) and (6) show PPML estimates from a regression of the time spent visiting and migrating over a year. User-destination pairs where the residence location of the user is adjacent to the destination city are excluded. Standard errors are two-way clustered by origin and destination. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Conclusion

This thesis presented four distinct chapters, each delving into different facets of human mobility within developing countries, utilizing mobile phone data as a foundational source of information. The data demonstrated significant potential in identifying previously overlooked short-term mobility and offering enhanced granularity for better-known migratory movements. In the concluding chapter, I summarize the main findings of each chapter and suggest some avenues for future research.

Chapter 1 tapped into smartphone app location data in three African countries to unveil insights on distinctive short-term movements, which we labelled as "visits". These movements stand apart from daily commutes and more prolonged migration events, and are ubiquitous in the three countries studied. Smartphone users are frequently seen more than 10km away from their estimated home location, on about 10-15% of the days when they are observed. When these individuals venture out, they travel fairly significant distances and explore a diverse array of locales, spanning various population densities. The data's precision notably spotlights individual trips to urban areas. Large urban centers appear especially appealing, attracting visitors from across the nation. Furthermore, a substantial portion of these urban visitors come from areas categorized as non-urban. By juxtaposing the smartphone ping locations of urban visitors with data from Open Street Map, we were also able to characterize the specific locations that these visitors frequent in the destination cities. Their presence spans across a myriad of locations, from travel-associated venues like airports and hotels to administrative establishments, shopping areas, markets, and commercial districts. We then developed a conceptual framework aimed at rationalizing the patterns of visits observed in the data. In this model, individuals make visits to cities where they consume a broadly defined urban amenity. The model yields a number of testable predictions that are consistent with the movements captured by our smartphone data. For instance, the number of visits per person made from a smaller settlement to a larger one will exceed the number made in the opposite direction. Also, the fraction of days users spend visiting a city follows a gravity-style equation.

Lastly, given a choice between visiting two equidistant locations, individuals more frequently visit the more populous destination.

The significant patterns of visits to urban areas observed in our data suggest that cities offer benefits to a broader demographic than merely their inhabitants and daily commuters. Crucially, the substantial volume of visits emanating from non-urban areas calls into question the conventional notion of a rigid rural-urban dichotomy. Instead, visits could represent a potential conduit for individuals to achieve partial urbanization.

Future research might explore the role of visiting flows as a determining factor for spatial equilibria. An unresolved question is whether the consumption of urban amenities through visits can explain a portion of the observed spatial gaps. Moreover, while we characterized the specific places frequented by urban visitors, a clear limitation of the mobile phone data we exploit is its lack of insight into the underlying motives for these visits. Through these transient city visits, rural and small-town residents may address administrative matters, avail of consumption options otherwise inaccessible in their locales, and potentially even access market goods and services, sidestepping additional costs often imposed by intermediaries or traders. To truly unpack the characteristics of these identified movements, integrating survey data would be invaluable.

Chapter 2 presented a set of methodological tools to derive temporary migration statistics from mobile phone data. First, the chapter focused on systematic methods and data for the characterization of mobile phone data samples, with the overarching objective of generating migration statistics. I addressed well-known issues of cross-sectional selection, relying upon both secondary survey data and simple metrics directly derived from a mobile phone sample. Results on a sample of CDR data from Senegal reveal modest differences between mobile phone users and the overall population. Mobile phone users tend to be predominantly male and urban. However, there are virtually no disparities in phone ownership rates across income brackets among males, and only minor variances within the female sub-population. Moreover, I also shed light on potential measurement and selection issues linked to temporal sample characteristics. Crucially, I quantified the impact of the frequency and length of observation of users on the accuracy of the subsequent temporary migration detection algorithm. The length of observation primarily influences the home detection process, while the detection rate of migration episodes starts to decline notably when users are observed on less than 50% of days within their observation span. The results provide guidance on observational requirements for capturing temporary migration events with mobile phone data. They also inform the choice of filtering parameters when selecting working subsets of users for the

production of mobility metrics. In this respect, I also evaluated the implications of stringent filtering parameters, considering both the reduction in sample size and the exacerbation of selection biases. As a result, I highlighted the existing trade-off that must be made between migration measurement error on one side, and sample size and selection biases on the other. With this in mind, I constructed a high-quality subset of users with observational characteristics deemed necessary and sufficient for the measure of temporary migration.

Second, the chapter built upon previous work to develop a temporary migration detection algorithm adopting a segment-based approach. While previous approaches have focused on identifying flows of migration through persistent location changes, I incorporated the identification of a primary home location to facilitate a clearer characterization of movement direction, distinguishing between departures and returns.

Third, the chapter navigated the complexities of generating time-specific temporary migration statistics from individual migration trajectories, and introduced specific rules to streamline this process. Additionally, a weighting scheme was designed to neutralize the observed differences in the users-to-population ratio across locations. This approach was demonstrated to attenuate sample composition biases, such as when the less mobile but disproportionately represented urban phone users might artificially reduce migration estimates.

Finally, I applied this methodology to the CDR dataset from Senegal and presented a comprehensive temporary migration profile. The granularity of the CDR-based migration measures offered new insights into the temporary migration movements in Senegal. Considering migration episodes of at least 20 days, it is estimated that over a third of the adult population engages in at least one migration per year. Two-thirds of migration events originate from rural areas, and approximately half of the temporary migration inflow is directed to rural locations. In particular, the data uncover a large proportion of short-distance rural-to-rural movements. Moreover, the data provides a unique temporal insight into the dynamics of temporary migration. While common narratives usually highlight the importance of off-season (January-June) movements, CDR-derived migration measures rather suggest that the bulk of temporary moves occur during the rainy season (June-October).

Future research could conceivably enhance this study by conducting validation exercises against survey-based temporary migration estimations. A crucial aspect of this validation would involve assessing the efficacy of the weighting scheme in addressing sample composition biases. In this respect, future work could focus on utilizing data from a population observatory recording the short-term movements of the local population in the community of Niakhar in the Fatick region to carry

out a validation exercise, at least at a local scale.

Chapter 3 delved into the temporary migration responses to climate variability in Senegal. It drew on the methodology introduced in Chapter 2 to construct a granular pseudo-panel of temporary migration estimates. This dataset was combined with satellite-based precipitation estimations, allowing to observe temporary migration choices across an extensive array of locations under diverse rainfall scenarios. We developed a simple temporary migration model in which precipitations are incorporated in location-specific production functions. The location choices of individuals at each time period are modelled within a nested logit structure in which individuals have home bias preferences. The expression derived from this conceptual framework points to a simple intuition. Poorer rainfall conditions at origin have a negative impact on local production, decreasing wages and thus increasing the propensity to out-migrate. Similarly, a poorer rainy season quality at a destination is likewise associated with a negative productivity shock leading to a decline in wages and, therefore, a lower propensity to migrate to that destination.

The empirical analysis focused first on a well-known puzzle in the literature studying migration responses to local shocks. Several studies have found a negative effect of adverse shocks on local economic outcomes. However, these impacts often do not often translate into migratory responses, and researchers have responded to these findings by assuming that migration costs are high and individuals unresponsive to local shocks. Recent work has suggested that the conventional migration regression used to investigate these issues is in fact misspecified, due to the bilateral nature of location choices. Individuals opt for a location change based not only on conditions at their current location but also on conditions at potential alternative locations, which are omitted in conventional estimations. Taking advantage of our empirical setting, we estimated a conventional migration regression of the out-migration rate from a location on the rainy season quality observed at that location of origin. We compared the results to a dyadic regression relating the bilateral stock of temporary migrants between an origin and a destination, and the rainfall conditions at origin and destination. The findings show important disparities, confirming that traditional migration regressions could yield misleading results. In particular, poorer rainfall conditions are found to decrease the propensity to out-migrate during the off-season within the conventional regression setting, while the opposite effect is found using a dyadic regression that accounts for rainfall conditions at destination.

The rest of the chapter analyzed results from dyadic regressions, focusing particularly on rural locations. Poorer rainfall conditions at origin during the rainy season were found to have opposite effects on temporary migration during

the harvest season (September-November) and the off-season (February-May the following year). A 10% decrease in precipitations at the origin is associated with an average 2-4% decrease in a bilateral stock of temporary migrants originating from that location during harvest. Conversely, this decline in local precipitations during the rainy season leads to a 4-6% surge in the temporary migration stock during the following off-season. One possible explanation is the presence of a prevailing liquidity constraint hindering temporary migration immediately following a poor rainy season. Then, individuals might require several months to accumulate the necessary resources allowing them to initiate migration during the following off-season. Moreover, and perhaps surprisingly, our findings indicate that these effects were largely driven by rural-to-rural movements. In particular, we found no evidence of drought conditions triggering notable outflows of individuals from affected areas. Finally, our heterogeneity analysis revealed that the effects identified seem exacerbated in locations associated with lower standards of living.

The model, although relatively simple, provides a clear framework to integrate climate variability into temporary migration decisions. Future research endeavor could expand on this foundation, incorporating elements that better mirror the climate migration responses observed in the data. One such potential enhancement could involve the integration of a more sophisticated cost structure, enabling the depiction of stringent liquidity constraints. However, to validate this causal mechanism, one would need socio-economic information about users, which is notably absent in CDR data. One potential avenue could then be to draw on previous research that has demonstrated the possibility to infer an approximate socio-economic status directly from individuals' observable phone usage patterns. In the same vein, further exploration of CDR data could delve into understanding the influence of social networks on temporary migration decisions following a climate shocks. It would be insightful to examine whether a temporary migrant's choice of destination is influenced not just by the rainfall conditions at that destination, but also by the presence of a strong social network.

Chapter 4 expanded upon the novel insights on human mobility provided in Chapter 1. Specifically, it examined the interplay between visits and temporary migration decisions. I relied upon the uniqueness of CDR data that combine high-frequency observations and extensive activity periods, allowing to simultaneously observe the visits and temporary migration choices for a large sample of phone users in Senegal.

I started by documenting the visiting patterns toward cities in Senegal. The results highlighted a striking degree of mobility, with an estimated 83% of phone users having visited a city over the course of a year. Moreover, the data demonstrated

a distinct gradient where individuals from the most sparsely populated areas exhibit a higher propensity to undertake urban visits compared to those from denser areas. The median visitor makes one visit to a city every 1.3 months and each visit lasts for 1.5 days on average. Individuals originating from the remotest locations spend an average of 33 days per year on urban visits, while at the other end of the distribution, Dakar residents average 12 days of visits to other cities annually. The CDR data employed for this analysis arguably augment the results presented in Chapter 1 by providing measures of visits for a larger rural base, over longer observation spans, and based on a higher frequency of observation. The results broadly corroborate the conclusions of Chapter 1, highlighting the ubiquity of short-term movements to cities which are distinct from commuting and migration. On the other hand, drawing upon the method developed in Chapter 2, I estimated that 17% of phone users in the sample undertake a temporary migration. Interestingly, a notable fraction of visitors allocate as much time to city visits as certain migrants do for their migration spells to analogous destinations.

Regression analyses revealed a positive relationship between visits and temporary migration choices, among individuals facing comparable mobility costs. On average, over a period of one year and controlling for origin fixed effects, those undertaking a temporary migration spell to a city make 17.5 additional days of urban visits. Furthermore, I showed that this association is predominantly driven by temporary migrants visiting their migration destinations more abundantly than their non-migrant counterparts. I further investigated the temporal dimension of these supplemental visits with respect to the timing of temporary migration events. The results hint at the existence of anticipatory behaviors. Temporary migrants exhibit a higher probability of visit to their prospective migration destination in the weeks preceding departure, relative to other time periods where they are observed. This is interpreted as suggestive evidence that temporary migrants accept to bear the cost of visits in order to gain information about a destination and mitigate the risks of migration failure. Similarly, those individuals revisit their migration destination after returning to their home location, possibly to conclude pending tasks, procure pending payments, or engage with acquaintances established during their migration.

The second part of the chapter capitalize on the simultaneous observation of visits and temporary migration choices to investigate cost differentials between these two mobility types. A simple conceptual framework introduces a mobility cost structure that comprises a fixed cost specific to each mobility type (i.e. visits or temporary migration), the cost related to the distance between an origin and a destination (i.e. the bus fare), and a destination-specific cost associated with the stay duration. Within this framework, I derived expressions for the frequency and

duration of both visits and temporary migration for an individual. I demonstrated that retaining the full cost structure – contrary to Chapter 1 which ultimately considers the sole distance-related cost – implies that the distance elasticity of mobility choices is not straightforwardly related to the marginal cost of distance. Rather, it shows a negative relationship with the fixed cost of mobility. The model laid the foundations for the estimation of gravity regressions yielding distance elasticities of the frequency, duration, and accumulated time spent in cities, through visits and temporary migration respectively. Reassuringly, the sign of these elasticities were found to be coherent with the model predictions. More importantly, the set of estimated elasticities for visits showed notable disparities with those obtained from observed temporary migration movements. The results support the idea that the fixed costs associated with temporary migration exceed those affiliated with visits.

Bibliography

Aadhar, Saran and Vimal Mishra (2017). "High-resolution near real-time drought monitoring in South Asia". *Sci Data* 4, p. 170145. DOI: 10.1038/sdata.2017.145.

Ahlfeldt, Gabriel M, Stephen J Redding, Daniel M Sturm, and Nikolaus Wolf (2015). "The economics of density: Evidence from the Berlin Wall". *Econometrica* 83.6, pp. 2127–2189.

Akbar, Prottoy, Victor Couture, Gilles Duranton, and Adam Storeygard (2023). "Mobility and Congestion in Urban India". *American Economic Review* 113.4, pp. 1083–1111. DOI: 10.1257/aer.20181662.

Aker, Jenny C (2010). "Information from Markets Near and Far: Mobile Phones and Agricultural Markets in Niger". *American Economic Journal: Applied Economics* 2.3, pp. 46–59.

Akram, Agha Ali, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak (2017). *Effects of Emigration on Rural Labour Markets*. NBER Working Paper Series 23929. National Bureau of Economic Research.

Allen, Treb (2014). "Information Frictions in Trade". *Econometrica* 82.6, pp. 2041–2083.

Allen, Treb and Costas Arkolakis (2014). "Trade and the Topography of the Spatial Economy". *Quarterly Journal of Economics* 129.3, pp. 1085–1140.

ANSD/Sénégal and ICF (2018). *Sénégal: Enquête Démoggraphique et de Santé Continue (EDS-Continue) 2017*. Dakar, Sénégal: ANSD and ICF.
Available at http://dhsprogram.com/pubs/pdf/FR345/FR345.pdf.

Arkolakis, Costas, Arnaud Costinot, and Andrés Rodríguez-Clare (2012). "New Trade Models, Same Old Gains?" *American Economic Review* 102.1, pp. 94–130.

Athey, Susan, Billy Ferguson, Matthew Gentzkow, and Tobias Schmidt (2021). "Estimating experienced racial segregation in US cities using large-scale GPS data". *Proceedings of the National Academy of Sciences* 118.46, e2026160118.

Athey, Susan, Billy A Ferguson, Matthew Gentzkow, and Tobias Schmidt (2020). *Experienced Segregation*. Working Paper 27572. National Bureau of Economic Research. DOI: 10.3386/w27572.

Atkin, David, M. Keith Chen, and Anton Popov (2022). *The Returns to Face-to-Face Interactions: Knowledge Spillovers in Silicon Valley*. Working Paper 30147. National Bureau of Economic Research. DOI: 10.3386/w30147.

Atkin, David and Dave Donaldson (2015). *Who's Getting Globalized? The Size and Implications of Intra-national Trade Costs*. Working Paper 21439. National Bureau of Economic Research. DOI: 10.3386/w21439.

Baker, Jonathan and Tade Akin Aina (1995). *The Migration Experience in Africa*. Nordic Africa Institute.

Basu, Karna and Maisy Wong (2015). "Evaluating seasonal food storage and credit programs in east Indonesia". *Journal of Development Economics* 115, pp. 200–216. DOI: 10.1016/j.jdeveco.2015.02.001.

Beck, Simon, Philippe De Vreyer, Sylvie Lambert, Karine Marazyan, and Abla Safir (2015). "Child Fostering in Senegal". *Journal of Comparative Family Studies* 46.1, pp. 57–73.

Beegle, Kathleen, Joachim De Weerdt, and Stefan Dercon (2011). "Migration and Economic Mobility in Tanzania: Evidence from a Tracking Survey". *The Review of Economics and Statistics* 93.3, pp. 1010–1033.

Blanchard, Paul, Oscar Anil Ishizawa Escudero, Thibaut Humbert, and Rafael Van Der Borght (2023). *Struggling with the Rain : Weather Variability and Food Insecurity Forecasting in Mauritania*. Policy Research Working Paper Series 10457. The World Bank.

Blumenstock, Joshua, Guanghua Chi, and Xu Tan (2022). *Migration and the Value of Social Networks*. Working Paper 60. IZA Institute of Labor Economics.

Blumenstock, Joshua and Nathan Eagle (2010). "Mobile Divides: Gender, Socioeconomic Status, and Mobile Phone Use in Rwanda". *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*. ICTD '10. London, United Kingdom: Association for Computing Machinery. ISBN: 9781450307871. DOI: 10.1145/2369220.2369225.

Blumenstock, Joshua E. (2012). "Inferring Patterns of Internal Migration from Mobile Phone Call Records : Evidence from Rwanda". *Information Technology for Development* 18.2, pp. 107–125. DOI: 10.1080/02681102.2011.643209.

Borusyak, Kirill, Rafael Dix-Carneiro, and Brian Kovak (2022). "Understanding Migration Responses to Local Shocks". *Unpublished work*.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess (2021). "Revisiting Event Study Designs: Robust and Efficient Estimation". *Unpublished manuscript*.

Brauw, Alan de, Valerie Mueller, and Hak Lim Lee (2014). "The Role of Rural–Urban Migration in the Structural Transformation of Sub-Saharan Africa". *World Development* 63. Economic Transformation in Africa, pp. 33–42. DOI: https://doi.org/10.1016/j.worlddev.2013.10.013.

Brooks, Wyatt and Kevin Donovan (2020). "Eliminating Uncertainty in Market access: The Impact of New Bridges in Rural Nicaragua". *Econometrica* 88.5, pp. 1965–1997.

Bryan, Gharad, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak (2014). "Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh". *Econometrica* 82.5, pp. 1671–1748. DOI: 10.3982/ecta10489.

Bryan, Gharad and Melanie Morten (2019). "The Aggregate Productivity Effects of Internal Migration: Evidence from Indonesia". *Journal of Political Economy* 127.5, pp. 2229–2268.

Büchel, Konstantin, Maximilian V. Ehrlich, Diego Puga, and Elisabet Viladecans-Marsal (2020). "Calling from the outside: The role of networks in residential mobility". *Journal of Urban Economics* 119. DOI: 10.1016/j.jue.2020.103277.

Call, Maia A., Clark Gray, Mohammad Yunus, and Michael Emch (2017). "Disruption, not displacement: Environmental variability and temporary migration in Bangladesh". *Global Environmental Change* 46.September, pp. 157–165. DOI: 10.1016/j.gloenvcha.2017.08.008.

Caselli, Francesco and Wilbur John Coleman II (2001). "The US Structural Transformation and Regional Convergence: A Reinterpretation". *Journal of political Economy* 109.3, pp. 584–616.

Chaisemartin, Clément de and Xavier D'Haultfœuille (2020). "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects". *American Economic Review* 110.9, pp. 2964–96. DOI: 10.1257/aer.20181169.

Chaudhuri, Subham and Christina Paxson (2021). *Smoothing consumption under income seasonality : Buffer Stocks vs. Credit Markets*. Discussion Paper 0102-54. Columbia University.

Chen, M Keith and Ryne Rohla (2018). "The Effect of Partisanship and Political Advertising on Close Family Ties". *Science* 360.6392, pp. 1020–1024.

Chi, Guanghua, Fengyang Lin, Guangqing Chi, and Joshua Blumenstock (2020). "A general approach to detecting migration events in digital trace data". *PLoS ONE* 15.10. DOI: 10.1371/journal.pone.0239408.

Coffey, Diane, John Papp, and Dean Spears (2015). "Short-Term Labor Migration from Rural North India: Evidence from New Survey Data". *Population Research and Policy Review* 34.3, pp. 361–380. DOI: 10.1007/s1.

Costinot, Arnaud and Dave Donaldson (2016). *How Large Are the Gains from Economic Integration? Theory and Evidence from U.S. Agriculture, 1880-1997*. Working Paper 22946. National Bureau of Economic Research. DOI: 10.3386/w22946.

Couture, Victor, Jonathan I. Dingel, Allison Green, Jessie Handbury, and Kevin R. Williams (2022). "JUE Insight: Measuring movement and social contact

with smartphone data: a real-time application to COVID-19". *Journal of Urban Economics* 127, p. 103328. DOI: 10.1016/j.jue.2021.103328.

Dallmann, Ingrid and Katrin Millock (2016). "Climate Variability and Internal Migration: A Test on Indian Inter-State Migration".

Defrance, Dimitri, Esther Delesalle, and Flore Gubert (2023). "Migration response to drought in Mali. An analysis using panel data on Malian localities over the 1987-2009 period". *Environment and Development Economics* 28.2, pp. 171–190. DOI: 10.1017/S1355770X22000183.

Delaunay, Valérie, Emmanuelle Engeli, Régine Franzetti, Guillaume Golay, Aurore Moullet, and Claudine Sauvain-Dugerdil (2016). "La migration temporaire des jeunes au Sénégal: Un facteur de résilience des sociétés rurales sahéliennes?" *Afrique Contemporaine* 259.3, pp. 75–94. DOI: 10.3917/afco.259.0075.

Demissie, Merkebe Getachew, Santi Phithakkitnukoon, Lina Kattan, and Ali Farhan (2019). "Understanding human mobility patterns in a developing country using mobile phone data". *Data Science Journal* 18.1, pp. 1–13. DOI: 10.5334/dsj-2019-001.

Dercon, Stefan and Luc Christiaensen (2011). "Consumption risk, technology adoption and poverty traps: Evidence from Ethiopia". *Journal of Development Economics* 96.2, pp. 159–173. DOI: 10.1016/j.jdeveco.2010.08.003.

Dingel, Jonathan I. and Felix Tintelnot (2023). "Spatial Economics for Granular Settings".

Donaldson, Dave (2018). "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure". *American Economic Review* 108.4-5, pp. 899–934.

Donaldson, Dave and Richard Hornbeck (2016). "Railroads and American Economic Growth: A "Market Access" Approach". *Quarterly Journal of Economics* 131.2, pp. 799–858.

Duranton, Gilles (2015). "Growing through cities in developing countries". *World Bank Research Observer* 30.1, pp. 39–73. DOI: 10.1093/wbro/lku006.

Duranton, Gilles and Diego Puga (2004). "Chapter 48 - Micro-Foundations of Urban Agglomeration Economies". *Cities and Geography*. Ed. by J. Vernon Henderson and Jacques-François Thisse. Vol. 4. Handbook of Regional and Urban Economics. Elsevier, pp. 2063–2117. DOI: https://doi.org/10.1016/S1574-0080(04)80005-1.

Eckert, Fabian and Michael Peters (2022). *Spatial Structural Change*. Working Paper 30489. National Bureau of Economic Research. DOI: 10.3386/w30489.

Fafchamps, Marcel and Flore Gubert (2007). "The formation of risk sharing networks". *Journal of Development Economics* 83.2, pp. 326–350. DOI: 10.1016/j.jdeveco.2006.05.005.

Fafchamps, Marcel, Christopher Udry, and Katherine Czukas (1998). "Drought and saving in West Africa: are livestock a buffer stock?" *Journal of Development Economics* 55, pp. 273–305.

Findley, S. E. (1994). "Does drought increase migration? A study of migration from rural Mali during the 1983-1985 drought". *International Migration Review* 28.3, pp. 539–553. DOI: 10.2307/2546820.

Fiorio, Lee, Guy Abel, Jixuan Cai, Emilio Zagheni, Ingmar Weber, and Guillermo Vinué (2017). "Using Twitter Data to Estimate the Relationship between Short-term Mobility and Long-term Migration". *WebSci '17: Proceedings of the 2017 ACM on Web Science Conference*, pp. 103–110.

Funk, Chris, Pete Peterson, Martin Landsfeld, Diego Pedreros, James Verdin, Shraddhanand Shukla, Gregory Husak, James Rowland, Laura Harrison, Andrew Hoell, and Joel Michaelsen (2015). "The climate hazards infrared precipitation with stations - A new environmental record for monitoring extremes". *Scientific Data* 2, pp. 1–21. DOI: 10.1038/sdata.2015.66.

Garriga, Carlos, Aaron Hedlund, Yang Tang, and Ping Wang (2023). "Rural-Urban Migration, Structural Transformation, and Housing Markets in China". *American Economic Journal: Macroeconomics* 15.2, pp. 413–40. DOI: 10.1257/mac.20160142.

Gollin, Douglas, Martina Kirchberger, and David Lagakos (2021). "Do Urban Wage Premia Reflect Lower Amenities? Evidence from Africa". *Journal of Urban Economics* 121, p. 103301. DOI: https://doi.org/10.1016/j.jue.2020.103301.

Gollin, Douglas, David Lagakos, and Michael E. Waugh (2014). "The Agricultural Productivity Gap". *Quarterly Journal of Economics* 129.2, pp. 939–993.

González, Marta C., César A. Hidalgo, and Albert-László Barabási (2008). "Understanding individual human mobility patterns". *Nature* 453, pp. 779–782.

Goodman-Bacon, Andrew (2021). "Difference-in-differences with variation in treatment timing". *Journal of Econometrics* 225.2, pp. 254–277.

Gray, Clark and Valerie Mueller (2012a). "Drought and Population Mobility in Rural Ethiopia". *World Development* 40.1, pp. 134–145. DOI: 10.1016/j.worlddev.2011.05.023.

Gray, Clark L. and Valerie Mueller (2012b). "Natural disasters and population mobility in Bangladesh". *Proceedings of the National Academy of Sciences of the United States of America* 109.16, pp. 6000–6005. DOI: 10.1073/pnas.1115944109.

Guo, Hao, Anming Bao, Tie Liu, Felix Ndayisaba, Daming He, Alishir Kurban, and Philippe De Maeyer (2017). "Meteorological drought analysis in the Lower Mekong Basin using satellite-based long-term CHIRPS product". *Sustainability* 9.6, p. 901. DOI: 10.3390/su9060901.

Hamory, Joan, Marieke Kleemans, Nicholas Y Li, and Edward Miguel (2020). "Reevaluating Agricultural Productivity Gaps with Longitudinal Microdata".

Journal of the European Economic Association 19.3, pp. 1522–1555. DOI: 10.1093/jeea/jvaa043.

Hankaew, Soranan, Santi Phithakkitnukoon, Merkebe Getachew Demissie, Lina Kattan, Zbigniew Smoreda, and Carlo Ratti (2019). "Inferring and Modeling Migration Flows Using Mobile Phone Network Data". *IEEE Access* 7, pp. 164746–164758. DOI: 10.1109/ACCESS.2019.2952911.

Harris, John R and Michael P Todaro (1970). "Migration , Unemployment and Developmnent : A Two-Sector Analysis". *The American Economic Review* 60.1, pp. 126–142.

Henderson, J Vernon and Sebastian Kriticos (2018). "The Development of the African System of Cities". *Annual Review of Economics* 10, pp. 287–314.

Henderson, J. Vernon (2010). "Cities and Development". *Journal of Regional Science* 50.1, pp. 515–540. DOI: https://doi.org/10.1111/j.1467-9787.2009.00636.x.

Henderson, Vernon, Adam Storeygard, and David Weil (2012). "Measuring Economic Growth from Outer Space". *American Economic Review* 102.2, pp. 994–1028.

Henry, Sabine, Bruno Schoumaker, and Cris Beauchemin (2004). "The impact of rainfall on the first out-migration: A multi-level event-history analysis in Burkina Faso". *Population and Environment* 25.5, pp. 423–460. DOI: 10.1023/B:POEN.0000036928.17696.e8.

Hill, Ruth and Catherine Porter (2017). "Vulnerability to Drought and Food Price Shocks: Evidence from Ethiopia". *World Development* 96, pp. 65–77. DOI: https://doi.org/10.1016/j.worlddev.2017.02.025.

Hirvonen, Kalle (2016). "Temperature Changes, Household Consumption, and Internal Migration: Evidence from Tanzania". *American Journal of Agricultural Economics* 98.4, pp. 1230–1249. DOI: 10.1093/ajae/aaw042.

Hong, Lingzi, Jiahui Wu, Enrique Frias-Martinez, Andrés Villarreal, and Vanessa Frias-Martinez (2019). "Characterization of internal migrant behavior in the immediate post-migration period using cell phone traces". *ACM International Conference Proceeding Series*. January. ISBN: 9781450361224. DOI: 10.1145/3287098.3287119.

Imbert, Clément and John Papp (2020a). "Costs and benefits of rural-urban migration: Evidence from India". *Journal of Development Economics* 146. DOI: https://doi.org/10.1016/j.jdeveco.2020.102473.

— (2020b). "Short-term Migration, Rural Public Works, and Urban Labor Markets: Evidence from India". *Journal of the European Economic Association* 18.2, pp. 927–963.

Jensen, Robert (2007). "The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector". *The Quarterly Journal of Economics* 122.3, pp. 879–924.

Jurdak, Raja, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth (2015). "Understanding Human Mobility from Twitter". *PLOS ONE* 10.7, pp. 1–16. DOI: 10.1371/journal.pone.0131469.

Kreindler, Gabriel and Yuhei Miyauchi (2021). *Measuring Commuting and Economic Activity Inside Cities with Cell Phone Records*. Working Paper 28516. National Bureau of Economic Research.

Kreindler, Gabriel E. and Yuhei Miyauchi (2023). "Measuring Commuting and Economic Activity inside Cities with Cell Phone Records". *The Review of Economics and Statistics* 105.4, pp. 899–909. DOI: 10.1162/rest_a_01085.

Lagakos, David, Ahmed Mushfiq Mobarak, and Michael E Waugh (2023). "The Welfare Effects of Encouraging Rural–Urban Migration". *Econometrica* 91 (3), pp. 803–837. DOI: https://doi.org/10.3982/ECTA15962.

Lai, Shengjie, Elisabeth zu Erbach-Schoenberg, Carla Pezzulo, Nick W. Ruktanonchai, Alessandro Sorichetta, Jessica Steele, Tracey Li, Claire A. Dooley, and Andrew J. Tatem (2019). "Exploring the use of mobile phone data for national migration statistics". *Palgrave Communications* 5.1. DOI: 10.1057/s41599-019-0242-9.

Lalou, Richard and Valérie Delaunay (2017). "Seasonal migration and climate change in rural Senegal: a form of adaptation or failure to adapt ?" *Rural societies in the face of climatic and environmental changes in West Africa*. Ed. by Benjamin Sultan, Richard Lalou, Mouftaou Amadou Sanni, Amadou Oumarou, and Mame Arame Soumaré. Marseille: IRD, pp. 269–293.

Lavelle-hill, Rosa, John Harvey, Gavin Smith, Anjali Mazumder, Madeleine Ellis, Kelefa Mwantimwa, and James Goulding (2022). "Using mobile money data and call detail records to explore the risks of urban migration in Tanzania". *EPJ Data Science* 11.28. DOI: 10.1140/epjds/s13688-022-00340-y.

Lu, Shiwei, Zhixiang Fang, Xirui Zhang, Shih-Lung Shaw, Ling Yin, Zhiyuan Zhao, and Xiping Yang (2017). "Understanding the Representativeness of Mobile Phone Location Data in Characterizing Human Mobility Indicators". *ISPRS International Journal of Geo-Information* 6.1. DOI: 10.3390/ijgi6010007.

Lu, Xin, Linus Bengtsson, and Petter Holme (2012). "Predictability of population displacement after the 2010 Haiti earthquake". *Proceedings of the National Academy of Sciences of the United States of America* 109.29, pp. 11576–11581. DOI: 10.1073/pnas.1203882109.

Lu, Xin, David J. Wrathall, Pål Roe Sundsøy, Md Nadiruzzaman, Erik Wetter, Asif Iqbal, Taimur Qureshi, Andrew Tatem, Geoffrey Canright, Kenth Engø-Monsen,

and Linus Bengtsson (2016). "Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh". *Global Environmental Change* 38, pp. 1–7. DOI: 10.1016/j.gloenvcha.2016.02.002.

Lucas, Robert E.B. (1997). "Internal migration in developing countries". Vol. 1. Handbook of Population and Family Economics. Elsevier, pp. 721–798. DOI: https://doi.org/10.1016/S1574-003X(97)80005-0.

Marchiori, Luca, Jean François Maystadt, and Ingmar Schumacher (2012). "The impact of weather anomalies on migration in sub-Saharan Africa". *Journal of Environmental Economics and Management* 63.3, pp. 355–374. DOI: 10.1016/j.jeem.2012.02.001.

Mastrorillo, Marina, Rachel Licker, Pratikshya Bohra-Mishra, Giorgio Fagiolo, Lyndon D. Estes, and Michael Oppenheimer (2016). "The influence of climate variability on internal migration flows in South Africa". *Global Environmental Change* 39, pp. 155–169. DOI: 10.1016/j.gloenvcha.2016.04.014.

McKee, Thomas B., J. Doesken Nolan, and John Kleist (1993). "The relationship of drought frequency and duration with time scales". *Eight conference on applied climatology, American Meteorological Society*, pp. 179–184. DOI: 10.1002/jso.23002.

Meghir, Costas, Ahmed Mushfiq Mobarak, Corina D Mommaerts, and Melanie Morten (2019). *Migration and Informal Insurance: Evidence from a Randomized Controlled Trial and a Structural Model*. Working Paper 26082. National Bureau of Economic Research. DOI: 10.3386/w26082.

Mianabadi, Ameneh, Khosro Salari, and Yavar Pourmohamad (2022). "Drought monitoring using the long-term CHIRPS precipitation over Southeastern Iran". *Applied Water Science* 12.8, pp. 1–13. DOI: 10.1007/s13201-022-01705-4.

Milusheva, Sveta, Elisabeth Zu Erbach-Schoenberg, Linus Bengtsson, Erik Wetter, and Andy Tatem (2017). "Understanding the Relationship between Short and Long Term Mobility". *AFD Research Paper Series*. No. 2017-69, June.

Minnesota Population Center (2020). *Integrated Public Use Microdata Series, International: Version 7.3 [Senegal 2013 census]*. Minneapolis, MN: IPUMS.

Miyauchi, Yuhei, Kentaro Nakajima, and Stephen J Redding (2021). *Consumption Access and Agglomeration: Evidence from Smartphone Data*. Discussion Paper 1745. Centre for Economic Performance.

Miyauchi, Yuhei, Kentaro Nakajima, and Steve Redding (2022). *The Economics of Spatial Mobility: Theory and Evidence Using Smartphone Data*. Working Paper 295. Griswold Center for Economic Policy Studies.

Mongey, Simon, Laura Pilossoph, and Alexander Weinberg (2021). "Which workers bear the burden of social distancing?" *Journal of Economic Inequality* 19.3, pp. 509–526. DOI: 10.1007/s10888-021-09487-6.

Monras, Joan (2018). *Economic Shocks and Internal Migration*. Discussion Paper 12977. CEPR.

Monte, Ferdinando, Stephen J. Redding, and Esteban Rossi-Hansberg (2018). "Commuting, Migration, and Local Employment Elasticities". *American Economic Review* 108.12, pp. 3855–90. DOI: 10.1257/aer.20151507.

Morten, Melanie (2019). "Temporary migration and endogenous risk sharing in village India". *Journal of Political Economy* 127.1.

Mueller, Valerie, Clark Gray, and Douglas Hopping (2020). "Climate-Induced migration and unemployment in middle-income Africa". *Global Environmental Change* 65, pp. 102–183. DOI: 10.1016/j.gloenvcha.2020.102183.

Munshi, Kaivan (2003). "Networks in the Modern Economy: Mexican Migrants in the U. S. Labor Market". *The Quarterly Journal of Economics* 118.2, pp. 549–599.

— (2014). "Community Networks and the Process of Development". *Journal of Economic Perspectives* 28.4, pp. 49–76. DOI: 10.1257/jep.28.4.49.

Munshi, Kaivan and Mark Rosenzweig (2016). "Networks and Misallocation: Insurance, Migration, and the Rural-Urban Wage Gap". *American Economic Review* 106.1, pp. 46–98.

Najjuma, Mabel, Alex Nimusiima, Geoffrey Sabiiti, and Ronald Opio (2021). "Characterization of Historical and Future Drought in Central Uganda Using CHIRPS Rainfall and RACMO22T Model Data". *International Journal of Agriculture and Forestry* 11.1, pp. 9–15. DOI: 10.5923/j.ijaf.20211101.02.

Owens Raymond, III, Esteban Rossi-Hansberg, and Pierre-Daniel Sarte (2020). "Rethinking Detroit". *American Economic Journal: Economic Policy* 12.2, pp. 258–305. DOI: 10.1257/pol.20180651.

Pandey, Varsha, Prashant K. Srivastava, Sudhir K. Singh, George P. Petropoulos, and Rajesh Kumar Mall (2021). "Drought Identification and Trend Analysis Using Long-Term CHIRPS Satellite Precipitation Product in Bundelkhand, India". *Sustainability* 13.3, p. 1042. DOI: 10.3390/su13031042.

Perez-Heydrich, Carolina, Joshua L. Warren, Clara R. Burgert, and Michael E. Emch (2013). *Guidelines on the Use of DHS GPS Data*. Tech. rep. Demographic and Health Surveys.

Phithakkitnukoon, Santi, Zbigniew Smoreda, and Patrick Olivier (2012). "Socio-geography of human mobility: A study using longitudinal mobile phone data". *PLoS ONE* 7.6. DOI: 10.1371/journal.pone.0039253.

Qader, Sarchil, Thomas Abbott, Emily Boytinck, Mathias Kuepie, Attila Lazar, and Andrew Tatem (2022). *Census disaggregated gridded population estimates for Senegal (2020), version 1.0*.

Redding, Stephen J and Matthew A Turner (2015). "Transportation Costs and the Spatial Organization of Economic Activity". *Handbook of regional and urban economics* 5, pp. 1339–1398.

Rosenthal, Stuart S. and William C. Strange (2004). "Chapter 49 - Evidence on the Nature and Sources of Agglomeration Economies". *Cities and Geography*. Ed. by J. Vernon Henderson and Jacques-François Thisse. Vol. 4. Handbook of Regional and Urban Economics. Elsevier, pp. 2119–2171. DOI: https://doi.org/10.1016/S1574-0080(04)80006-3.

Ruktanonchai, Nick Warren, Corrine Warren Ruktanonchai, Jessica Rhona Floyd, and Andrew J. Tatem (2018). "Using Google Location History data to quantify fine-scale human mobility". *International Journal of Health Geographics* 17.1. DOI: 10.1186/s12942-018-0150-z.

Sandeep, P., G. P. Obi Reddy, R. Jegankumar, and K. C. Arun Kumar (2021). "Monitoring of agricultural drought in semi-arid ecosystem of Peninsular India through indices derived from time-series CHIRPS and MODIS datasets". *Ecological Indicators* 121, p. 107033. DOI: 10.1016/j.ecolind.2020.107033.

Schareika, Nikolaus (1997). "Arid Ways : Cultural Understandings of Insecurity in Fulbe Society , Central Mali". *Nomadic Peoples* 1.2, pp. 120–125.

Schiavina, Marcello, Michele Melchiorri, Martino Pesaresi, Panagiotis Politis, S Freire, Luca Maffenini, Pietro Florio, Daniele Ehrlich, Katarzyna Goch, Pierpaolo Tommasi, et al. (2022). *GHSL Data Package 2022*. Publications Office of the European Union, Luxembourg, 2022, ISBN 978-92-76-53071-8, doi:10.2760/19817, JRC 129516.

Sinclair, Michael, Saeed Maadi, Qunshan Zhao, Jinhyun Hong, Andrea Ghermandi, and Nick Bailey (2023). "Assessing the socio-demographic representativeness of mobile phone application data". *Applied Geography* 158, p. 102997. DOI: https://doi.org/10.1016/j.apgeog.2023.102997.

Spyratos, Spyridon, Michele Vespe, Fabrizio Natale, Ingmar Weber, Emilio Zagheni, and Marzia Rango (2019). "Quantifying international human mobility patterns using Facebook Network data". *PLoS ONE* 14.10. DOI: 10.1371/journal.pone.0224134.

Su, Yichen (2022). "Measuring the Value of Urban Consumption Amenities: A Time-Use Approach". *Journal of Urban Economics* 132, p. 103495. DOI: https://doi.org/10.1016/j.jue.2022.103495.

Sun, Liyang and Sarah Abraham (2021). "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects". *Journal of Econometrics* 225.2, pp. 175–199. DOI: 10.1016/j.jeconom.2020.09.006.

Thiede, Brian, Clark Gray, and Valerie Mueller (2016). "Climate variability and inter-provincial migration in South America, 1970–2011". *Global Environmental Change* 41, pp. 228–240. DOI: 10.1016/j.gloenvcha.2016.10.005.

Townsend, Robert M. (1994). "Risk and Insurance in Village India". *Econometrica* 62.3, pp. 539–591. DOI: https://doi.org/10.2307/2951659.

Udry, Christopher (1994). "Risk and insurance in a rural credit market: An empirical investigation in Northern Nigeria". *The Review of Economic Studies* 61.3, pp. 495–526. DOI: 10.2307/2297901.

Vanhoof, Maarten, Fernando Reis, Thomas Ploetz, and Zbigniew Smoreda (2018). "Assessing the quality of home detection from mobile phone data for official statistics". *Journal of Official Statistics* 34.4, pp. 935–960. DOI: 10.2478/jos-2018-0046.

Wesolowski, Amy, Caroline O. Buckee, Kenth Engø-Monsen, and C. J.E. Metcalf (2016). "Connecting mobility to infectious diseases: The promise and limits of mobile phone data". *The Journal of Infectious Diseases* 214.S4, S414–S420. DOI: 10.1093/infdis/jiw273.

Wesolowski, Amy, Nathan Eagle, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee (2012). "Heterogeneous Mobile Phone Ownership and Usage Patterns in Kenya". *PLOS ONE* 7.4, pp. 1–6. DOI: 10.1371/journal.pone.0035319.

— (2013). "The impact of biases in mobile phone ownership on estimates of human mobility". *J R Soc Interface* 10.81. DOI: 10.1098/rsif.2012.0986.

Wesolowski, Amy, Nathan Eagle, Andrew J. Tatem, David L. Smith, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee (2012). "Quantifying the impact of human mobility on malaria". *Science* 338.6104, pp. 267–270. DOI: 10.1126/science.1223467.

Wesolowski, Amy, C. J.E. Metcalf, Nathan Eagle, Janeth Kombich, Bryan T. Grenfell, Ottar N. Bjørnstad, Justin Lessler, Andrew J. Tatem, and Caroline O. Buckee (2015). "Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data". *Proceedings of the National Academy of Sciences of the United States of America* 112.35, pp. 11114–11119. DOI: 10.1073/pnas.1423542112.

Williams, Nathalie E., Timothy A. Thomas, Matthew Dunbar, Nathan Eagle, and Adrian Dobra (2015). "Measures of human mobility using mobile phone records enhanced with GIS data". *PLoS ONE* 10.7. DOI: 10.1371/journal.pone.0133630.

Young, Alwyn (2013). "Inequality, The Urban-Rural Gap, And Migration". *Quarterly Journal of Economics*, pp. 1727–1785.

Zufiria, Pedro J., David Pastor-Escuredo, Luis Úbeda-Medina, Miguel A. Hernandez-Medina, Iker Barriales-Valbuena, Alfredo J. Morales, Damien C. Jacques, Wilfred Nkwambi, M. Bamba Diop, John Quinn, Paula Hidalgo-Sanchís, and Miguel Luengo-Oroz (2018). "Identifying seasonal mobility profiles from anonymized and aggregated mobile phone data. Application in food security". *PLoS ONE* 13.4.

This thesis was typeset using the LaTeX typesetting system created by Leslie Lamport and Donald Knuth, and the memoir class. The body text is set 11pt with Palatino designed by Hermann Zapf, which includes italics and small caps. Other fonts include Monospace from Young Ryu's TX Fonts family, Sans from TeX User Group Poland's Latin Modern family, and Helvetica by Max Miedinger and Eduard Hoffmann.