HIGH-FREQUENCY HUMAN MOBILITY IN THREE AFRICAN COUNTRIES

Paul BLANCHARD — *Trinity College Dublin*

Douglas GOLLIN — *University of Oxford and CEPR*

Martina KIRCHBERGER — *Trinity College Dublin and CEPR*

September 2023

Abstract

This paper 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. Our data point to considerable mobility within the sample. The average smartphone user in our data ventures more than 10 km from home on 12-15% of the days when they are observed. On average, when we observe them away from home, our users are typically 35-50 km from home. The granular nature of our data allows us 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. We develop a conceptual framework that characterizes the role of visits for individuals and provides a number of testable predictions that are consistent with the movement patterns that we observe in the data. Although our 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.

We thank Neil Barsch, Neenu Vincent, and Sean Walsh for excellent research assistance. We are grateful to Safegraph for sharing the smartphone app location data and answering many questions. Many thanks to Costas Arkolakis, David Atkin, Josh Blumenstock, Paula Bustos, Dave Donaldson, Kevin Donovan, Jessie Handbury, Vernon Henderson, Clement Imbert, Gabriel Kreindler, David Lagakos, Guy Michaels, Joan Monras, Melanie Morten, Michael Peters, Adam Storeygard, David Weil, and participants at the STEG Themes 3&4 Joint Workshop, BREAD Conference on the Economics of Africa, the Online Spatial and Urban Seminar, the Cities and Development conference, the Virtual Meeting of the Urban Economics Association, the CRED Workshop on Regional and Urban Economics, Pompeu Fabra, University of Oxford, NYU Abu Dhabi, Berkeley ARE, Trinity College Dublin, DENeB Berlin, Dortmund, Jena, Regensburg, University College Cork, Sheffield, UC Riverside, the University of British Columbia, Yale University, and the Econ RSA webinar series for helpful comments and discussions. We are also grateful to Paddy Doyle for continuous support with Trinity College Dublin's Computing Cluster. Kirchberger gratefully acknowledges funding from the Provost Project Award Fund. Gollin acknowledges funding from the Oxford Martin School Program on Governing African Transitions. All potential errors are our own.

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 et al., 2014; Lagakos et al., 2022; Imbert and Papp, 2020). 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 et al., 2019; Kreindler and Miyauchi, 2021). 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 et al. (2012) or Akbar et al. (2018). 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.

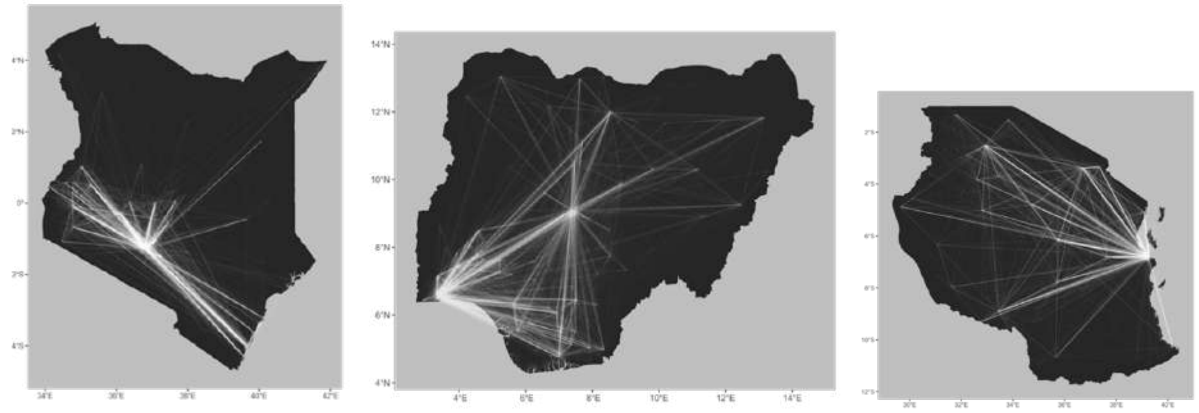
The paper makes three main contributions. First, since we study a new type of mobility, we start by defining a novel set of metrics for characterizing mobility across space related to frequency, spatial extent, and destination characteristics. Our metrics are parsimonious and easily interpretable across 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 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 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

[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 3.

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.

Figure 1: Mobility flows to cities.



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

Note: The brightness of lines reflects the counts of distinct visits to any city of more than 50,000, from every spatial grid cell outside the city perimeter.

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 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,

[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.

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 et al., 2020), to measure the effectiveness of social distancing (Mongey et al., 2021), social interactions (Couture et al., 2020) 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, 2021; Kreindler and Miyauchi, 2021). 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 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 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, 2001; Eckert and Peters, 2018; Lagakos et al., 2018). 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 be-

[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); 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); Jensen (2007)). Allen (2014) suggests that information frictions can compound spatial frictions.

tween cities, or between rural and urban locations, is unlikely to be one in which the costs 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 2 discusses the smartphone app data we use and how we define home locations. Section 3 focuses on sample selection and characterizes the sample. Section 4 presents our mobility indicators. Section 5 sketches our conceptual framework. Section 6 examines to what extent the data is consistent with the propositions coming out of our model. Section 7 concludes.

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

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

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

[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.

criteria. Unless specified otherwise, we use our high-confidence sample for our analysis.[6] We then carry out data cleaning procedures described in Appendix A.2.

Table 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 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: Sample and pings per user

	All		**High confidence**	
	Users	Pings ratio	Users	Pings ratio
	(1)	(2)	(3)	(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.)

Table 2 summarizes user-level temporal statistics for our high-confidence users considering three different measures. The first statistic that we consider is the 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 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

[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.

[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%.

[8] This differs from the pings ratio in Table 1 which simply summed over all pings in the data across all

Table 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.

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 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.

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 et al. (2014), Imbert and Papp (2020), Lagakos et al. (2018) or Blumenstock et al. (2019) using call detail records data.

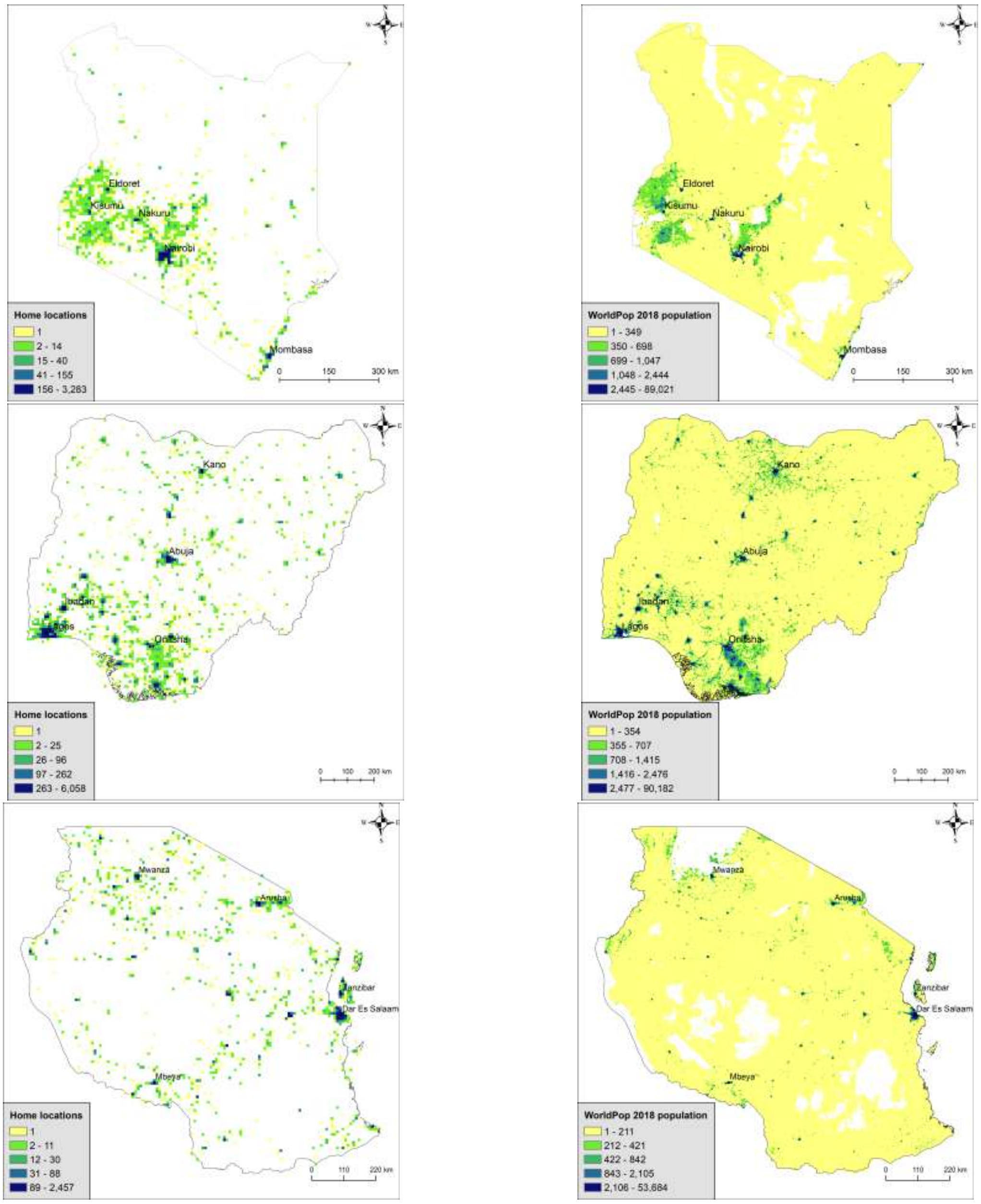
3. Selection

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. 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 microdata 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 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

Figure 2: Distribution of home locations and population.

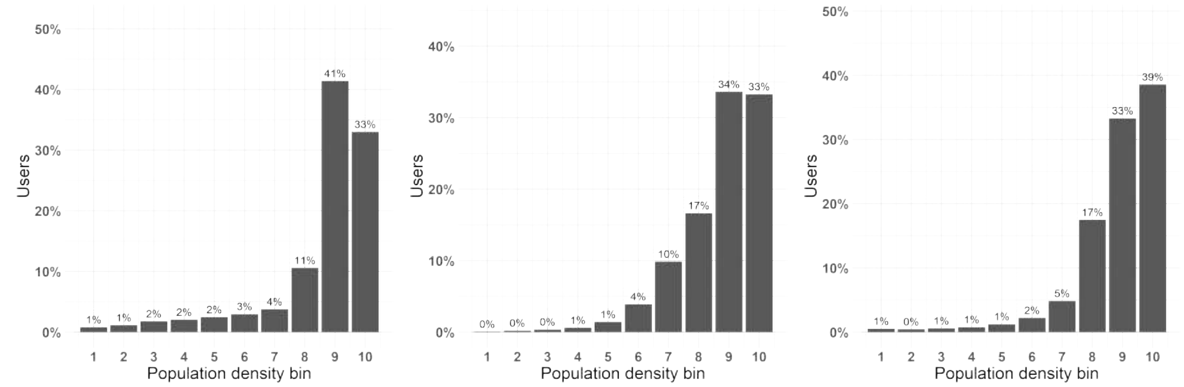


Note: This figure shows the distribution of home locations of users at a 10km resolution (on the left) and the distribution of the population at a 1km resolution (on the right).

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 3).[11] We compute three further metrics

Figure 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 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.

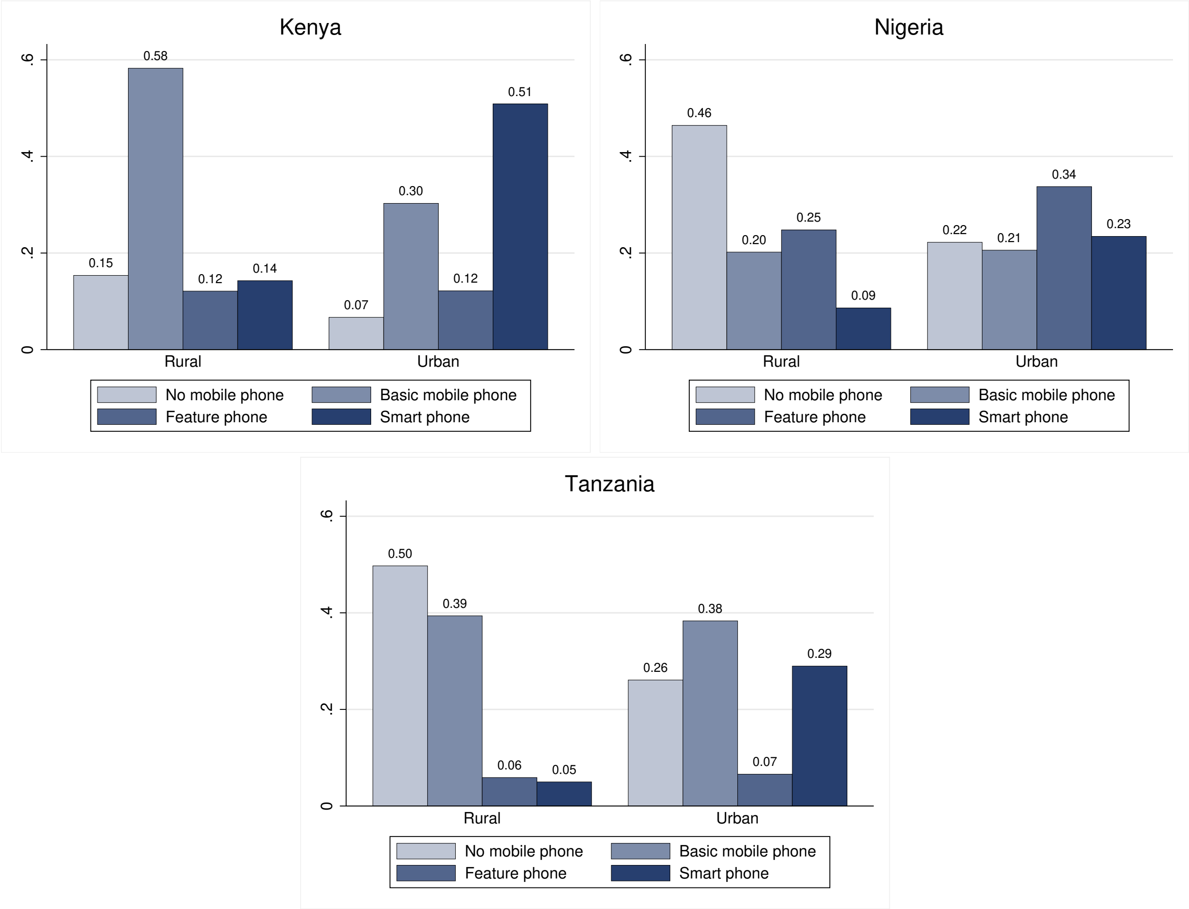
to measure the representativeness of our users across different levels of population density: first, we take all 10-km pixels in a country and regress the number of users in a pixel on population of the corresponding pixel. We find that the R-squared ranges between 0.36 in Kenya to 0.81 in Tanzania, 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. 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 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.

[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.

[12] Our approach to defining urban peripheries is described in Appendix Section 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.

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 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

Figure 4: Device ownership by location.



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

[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.

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 C.1-C.4 in the Appendix examine ownership rates by gender, and explore how owners of different devices 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 C.2 - C.4 highlight that these distributions are not distinct. Appendix Table 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 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 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

[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).

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 E.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]

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

[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.

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 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.

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 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

[16] Appendix Figure E.3 and Tables E.4– 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.

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.

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 $\bar{d} = 10\text{km}$). 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 T_{it} be a dummy indicating whether

[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.

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:

$$F_i = \frac{M_i}{T_i}. \tag{1}$$

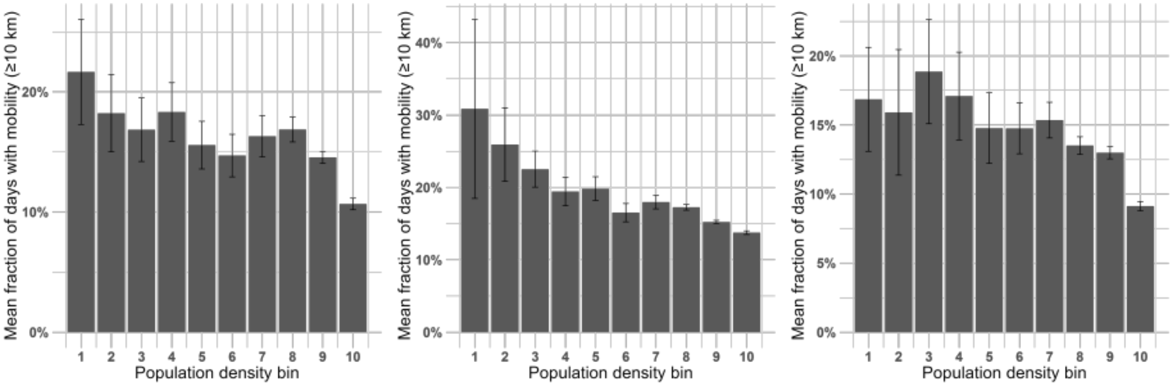
In this expression, the numerator denotes the number of days with at least one ping 10 km away from home for user i, and the denominator gives the total number of days on which user i is observed (i.e., days with at least one ping). We find that the fraction of days on which users are more than 10km 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 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 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:

$$F^j = \frac{1}{n_j} \sum_{i \in I_j} F_i. \tag{2}$$

Figure 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]

Figure 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.

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:

$$S_i = \frac{1}{P_{i\bar{H}}} \sum_{x \in Z_{i\bar{H}}} d_i(x) p_i(x). \tag{3}$$

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

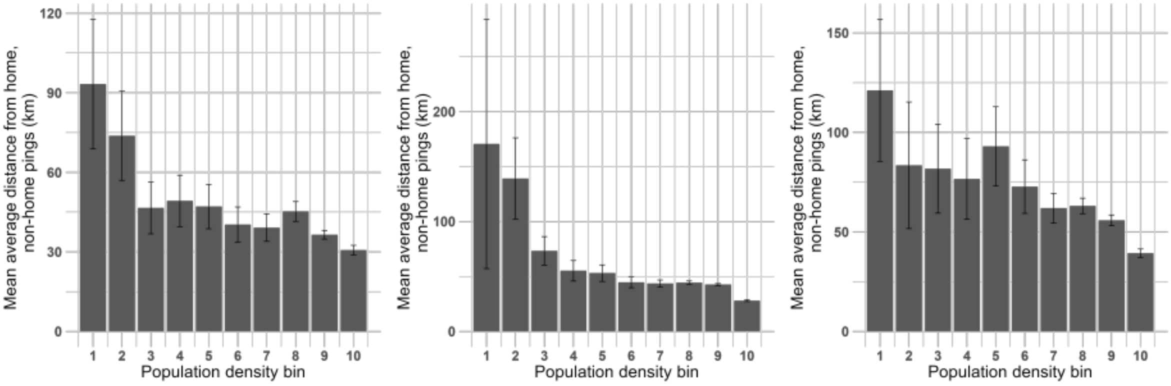
[18] As a robustness check, we reproduce Figure 5 with truncated means; that is, we discard values in the top 5 percentiles, to address the concern that the results could be driven by a small set of highly mobile users. We observe small decreases in the average fraction of days away in all density bins but no change in the overall pattern of decreasing frequency with population density.

the individuals belonging to a population density bin b_j by simply averaging the individual values of S_i. Thus:

$$S^j = \frac{1}{n_j} \sum_{i \in I_j} S_i. \tag{4}$$

Figure 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. As in Figure 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.

Figure 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.

Taken together, Figures 5 and 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.

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, $\tilde{N} : X \rightarrow B$. 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:

$$v_{ij} = \frac{\sum\limits_{x \in \{x \in \bar{H}_i : \tilde{N}(x) = b_j\}} p_i(x)}{P_{i\bar{H}}} \tag{5}$$

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:

$$V_{od} = \frac{1}{n_o} \sum_{i \in I_o} v_{id}. \tag{6}$$

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:

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

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:

$$\Delta_{jk} = \frac{\sum\limits_{i \in I_j} p_{ik}}{n_j}. \tag{7}$$

Table 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

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

Table 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 density	5	30.4%	25.9%	43%	52.2%	53.6%	49.3%	38.8%	35.7%	25.5%	18.9%
	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 density	5	23.8%	33.3%	43.6%	45.9%	51.3%	38.1%	27%	21%	20.1%	15.2%
	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 density	5	16.5%	33.8%	42.7%	40.8%	36.9%	43.4%	27.2%	20.3%	14.3%	8.3%
	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 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 A.4 (for details, see Appendix Tables E.7 and 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.

4.4. Specific locations visited

As an alternative to using density deciles for our analysis, we consider in Appendix Table 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]

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

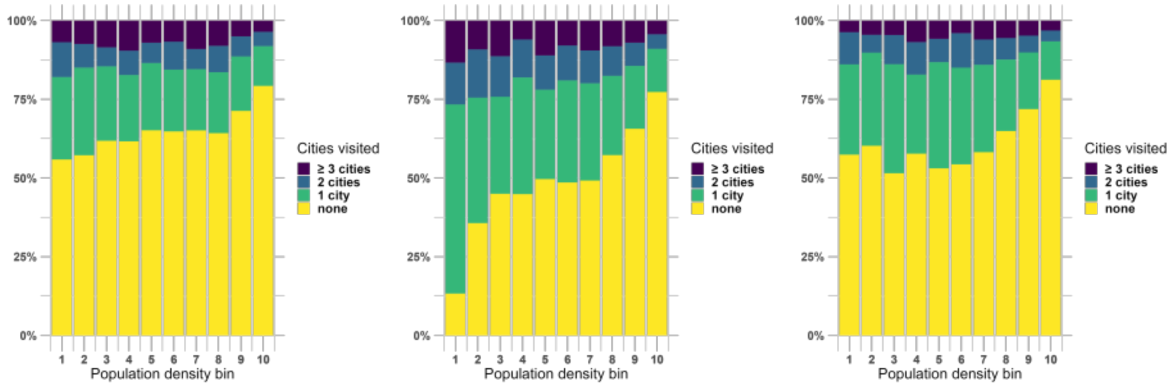
[20] See Appendix Section B for the definition of city boundaries. The reference year for city-level population counts is 2018.

[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 E.10.

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 A.5 provides the details on how we define visits. Figure 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 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.

Figure 7: Distribution of users according to the number of cities visited, by population density bin.

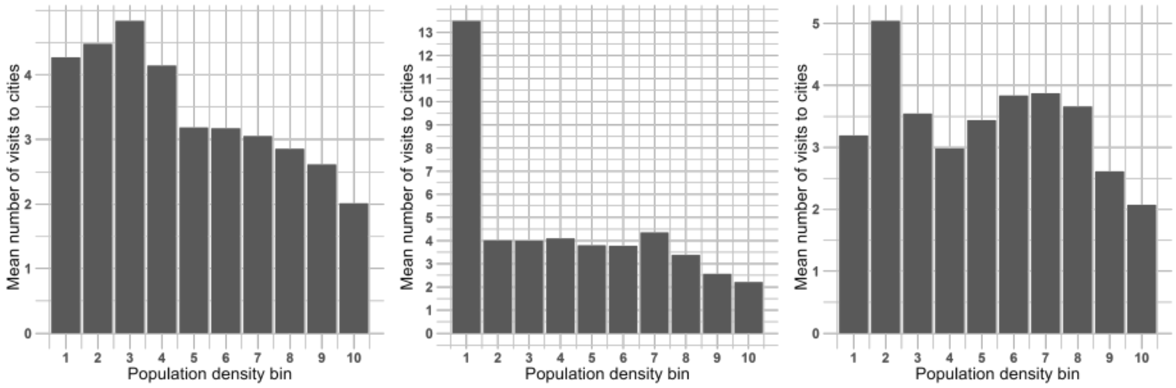


(a) Kenya (b) Nigeria (c) Tanzania

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.

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

Figure 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.

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.

Overall, we match more than 80% of visitors to at least one polygon as shown in column (1) of Table 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 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,

[22] See Appendix A.6 for details.

[23] Matching rates disaggregated by city are provided in Appendix Table 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).

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 roads and roadsides and places related to travel.

Table 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.

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.

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.

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:

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

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_{ix} 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.

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:

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

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.

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]

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

[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.

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.

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 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 $w_c = \pi_c^*$.

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:

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

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:

$$\ln \theta_{od} = \delta_0 + \delta_1 \ln D_{od} + \varphi_o + \nu_d + \epsilon_{od}. \tag{8}$$

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.*

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.

[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 (2021).

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.

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

$$H_0 : \frac{V_k(1,2)}{N_{k_1}} = \frac{V_k(2,1)}{N_{k_2}} \tag{9}$$

where the proposition assumed that $N_{k_2} > N_{k_1}$ for any two settlements within one of our three countries. Table 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 that some of the location pairs might have small differences in populations, we also explore whether the distribution of

Table 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 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.

visits becomes more distinct when we vary the difference between the origin and destination populations. Appendix Figure E.4 shows that this is indeed the case. The same pattern holds true for Tanzania and Nigeria.

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 (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. Origin fixed effects proxy for any observables or unobservables at the origin. Table 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.

Table 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

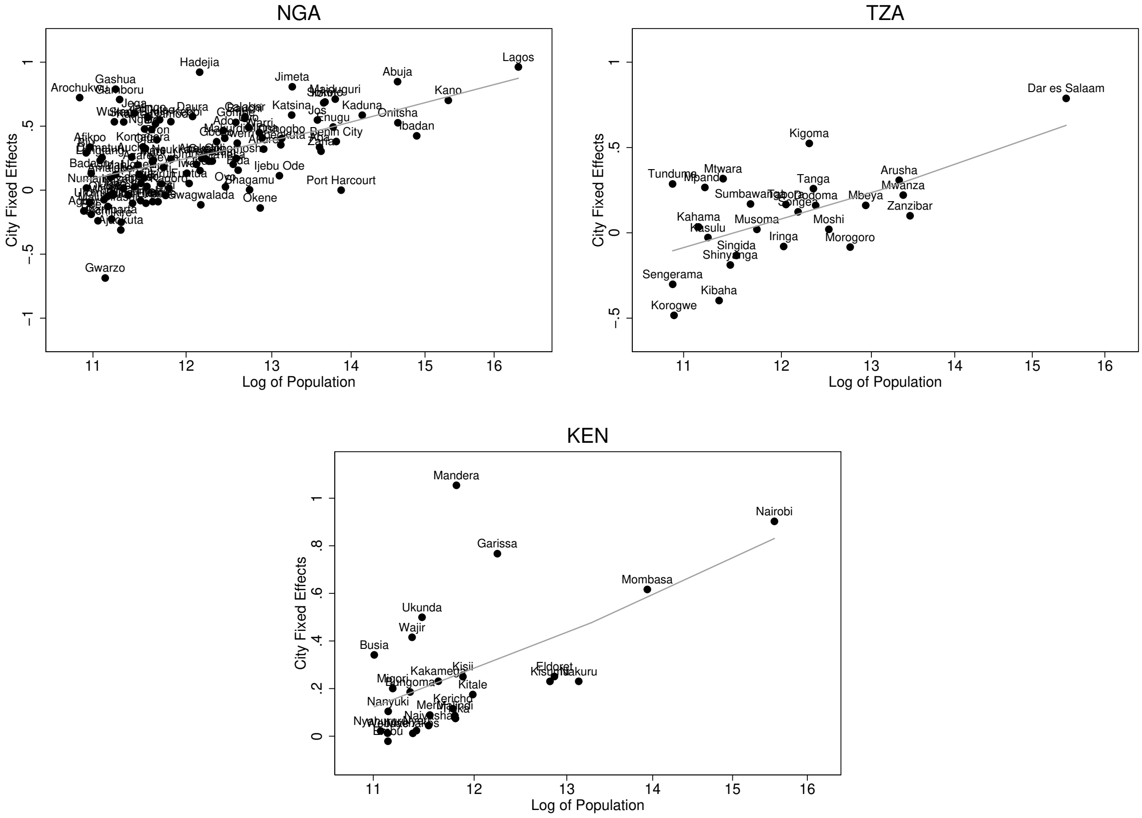
Note: This table estimates equation (8). The dependent variable is the fraction of days a user residing in origin o spends in destination d. All models include origin and destination fixed effects. Reported standard errors are clustered at the user level. *, **, *** denote significance at 10%, 5% and 1% levels.

[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.

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 8 and examine their relationship with population. Figure 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 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.

Figure 9: Destination fixed effects and city size.



Note: This figure shows the city fixed effects $\hat{\nu}_d$ from equation (8) and log of population.

Table 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 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.

Table 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 (8) on log city city. Robust standard errors in parentheses. *, **, *** denote significance at 10%, 5% and 1% levels.

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 et al., 2019; Kreindler and Miyauchi, 2021; 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

[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.

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 difficult 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.

References

AHLFELDT, G. M., S. J. REDDING, D. M. STURM, AND N. WOLF (2015): "The economics of density: Evidence from the Berlin Wall," *Econometrica*, 83, 2127–2189.

AKBAR, P. A., V. COUTURE, G. DURANTON, AND A. STOREYGARD (2018): "Mobility and congestion in urban India," Tech. rep., National Bureau of Economic Research.

AKER, J. C. (2010): "Information from Markets Near and Far: Mobile Phones and Agricultural Markets in Niger," *American Economic Journal: Applied Economics*, 2, 46–59.

ALLEN, T. (2014): "Information Frictions in Trade," *Econometrica*, 82, 2041–2083.

ALLEN, T. AND C. ARKOLAKIS (2014): "Trade and the Topography of the Spatial Economy," *Quarterly Journal of Economics*, 129, 1085–1140.

ARKOLAKIS, C., A. COSTINOT, AND A. RODRÍGUEZ-CLARE (2012): "New Trade Models, Same Old Gains?" *American Economic Review*, 102, 94–130.

ATHEY, S., B. FERGUSON, M. GENTZKOW, AND T. SCHMIDT (2021): "Estimating experienced racial segregation in US cities using large-scale GPS data," *Proceedings of the National Academy of Sciences*, 118, e2026160118.

ATKIN, D., K. CHEN, AND A. POPOV (2020): "The Returns to Serendipity: Knowledge Spillovers in Silicon Valley," Unpublished working paper.

ATKIN, D. AND D. DONALDSON (2015): "Who's Getting Globalized? The Size and Implications of Intra-national Trade Costs," Working Paper 21439, National Bureau of Economic Research.

BLUMENSTOCK, J. E., G. CHI, AND X. TAN (2019): "Migration and the value of social networks," .

BROOKS, W. AND K. DONOVAN (2020): "Eliminating Uncertainty in Market access: The Impact of New Bridges in Rural Nicaragua," *Econometrica*, 88, 1965–1997.

BRYAN, G., S. CHOWDHURY, AND A. M. MOBARAK (2014): "Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh," *Econometrica*, 82, 1671–1748.

BRYAN, G. AND M. MORTEN (2019): "The Aggregate Productivity Effects of Internal Migration: Evidence from Indonesia," *Journal of Political Economy*, 127, 2229–2268.

CASELLI, F. AND W. J. COLEMAN (2001): "The U.S. Structural Transformation and Regional

Convergence: A Reinterpretation," *Journal of Political Economy*, 109, 584–616.

CHEN, M. K. AND R. ROHLA (2018): "The Effect of Partisanship and Political Advertising on Close Family Ties," *Science*, 360, 1020–1024.

COSTINOT, A. AND D. 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.

COUTURE, V., J. I. DINGEL, A. E. GREEN, J. HANDBURY, AND K. R. WILLIAMS (2020): "Measuring Movement and Social Contact with Smartphone Data: A Real-Time Application to COVID-19," Working Paper 27560, National Bureau of Economic Research.

DINGEL, J. I. AND F. TINTELNOT (2021): "Spatial Economics for Granular Settings," .

DONALDSON, D. (2018): "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure," *American Economic Review*, 108, 899–934.

DONALDSON, D. AND R. HORNBECK (2016): "Railroads and American Economic Growth: A "Market Access" Approach," *Quarterly Journal of Economics*, 131, 799–858.

ECKERT, F. AND M. PETERS (2018): "Spatial Structural Change," Unpublished Manuscript.

GOLLIN, D., M. KIRCHBERGER, AND D. LAGAKOS (2021): "Do Urban Wage Premia Reflect Lower Amenities? Evidence from Africa," *Journal of Urban Economics*, 121, 103301.

GOLLIN, D., D. LAGAKOS, AND M. E. WAUGH (2014): "The Agricultural Productivity Gap," *Quarterly Journal of Economics*, 129, 939–993.

HENDERSON, J. V. AND S. KRITICOS (2018): "The Development of the African System of Cities," *Annual Review of Economics*, 10, 287–314.

HENDERSON, V., A. STOREYGARD, AND D. WEIL (2012): "Measuring Economic Growth from Outer Space," *American Economic Review*, 102, 994–1028.

IMBERT, C. AND J. PAPP (2020): "Costs and Benefits of Rural-urban Migration: Evidence from India," *Journal of Development Economics*, 146, 102473.

JENSEN, R. (2007): "The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector," *The Quarterly Journal of Economics*, 122, 879–924.

KREINDLER, G. E. AND Y. MIYAUCHI (2021): "Measuring Commuting and Economic Activity inside Cities with Cell Phone Records," Unpublished Manuscript.

LAGAKOS, D., A. M. MOBARAK, AND M. E. WAUGH (2018): "The Welfare Effects of Encouraging Rural-Urban Migration," Working Paper 24193, National Bureau of Economic Research.

LAGAKOS, D., M. MOBARAK, AND M. E. WAUGH (2022): "The Welfare Effects of Encouraging Rural-Urban Migration," Federal Reserve Bank of Minneapolis.

MIYAUCHI, Y., K. NAKAJIMA, AND S. REDDING (2022): "The Economics of Spatial Mobility: Theory and Evidence Using Smartphone Data," Unpublished Manuscript.

MONGEY, S., L. PILOSSOPH, AND A. WEINBERG (2021): "Which workers bear the burden of social distancing?" *The Journal of Economic Inequality*, 19, 509–526.

MONTE, F., S. J. REDDING, AND E. ROSSI-HANSBERG (2018): "Commuting, Migration, and Local Employment Elasticities," *American Economic Review*, 108, 3855–90.

OWENS, RAYMOND, I., E. ROSSI-HANSBERG, AND P.-D. SARTE (2020): "Rethinking Detroit," *American Economic Journal: Economic Policy*, 12, 258–305.

PEREZ-HEYDRICH, C., J. L. WARREN, C. R. BURGERT, AND M. E. EMCH (2013): "Guidelines on the Use of DHS GPS Data," Tech. rep., Demographic and Health Surveys.

REDDING, S. J. AND M. A. TURNER (2015): "Transportation Costs and the Spatial Organization of Economic Activity," *Handbook of regional and urban economics*, 5, 1339–1398.

SCHIAVINA, M., M. MELCHIORRI, M. PESARESI, P. POLITIS, S. FREIRE, L. MAFFENINI, P. FLORIO, D. EHRLICH, K. GOCH, P. 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.

Online Appendix

Table of Contents

A. Details on smartphone app data

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 help 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 A.1.

Table 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.

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 2, 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.

[30] See Appendix Section 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.

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.

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.

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)

[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.

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.

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

[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 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.

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 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

[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.

the time elapsed between the pings preceding and following $v^i_{m,c}$, so $T^{max}(v^i_{m,c}) = t^i_{j'_{m,c}+1} - t^i_{j_{m,c}-1}$.

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 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.

[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).

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]

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

[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.

[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.

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 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

[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).

[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.

consider the unique set of visitor-locations over which we superimpose the constructed OSM-based dataset of categorized places (see Table A.2).

Table 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.

B. Definition of city boundaries and regional capitals

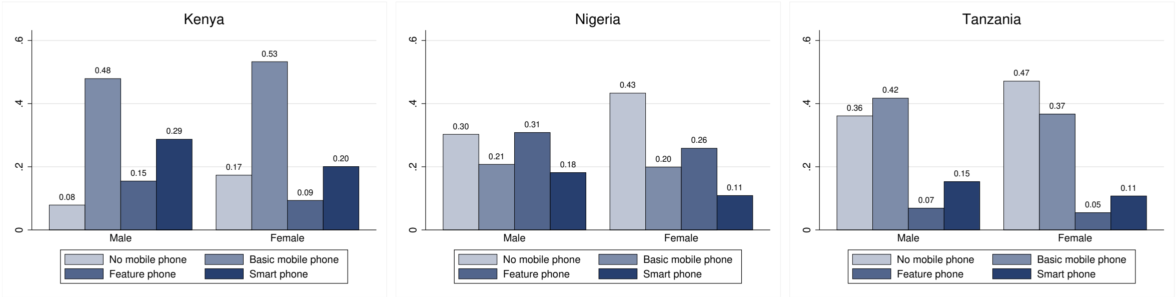
To define city boundaries, we use urban extents from the Global Rural-Urban mapping project v1.02 produced by Columbia University Center for International Earth Science Information Network (CIESIN). The original shapefile consists of polygons delineating urban settlements based on the point location 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. Cities' boundaries are defined according to GRUMP 3km-buffered

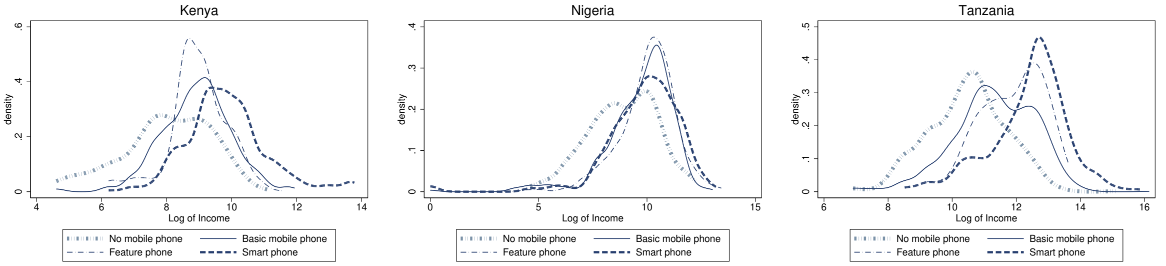
C. Sample selection: Comparing respondents by device ownership

Figure C.1: Device ownership by gender.



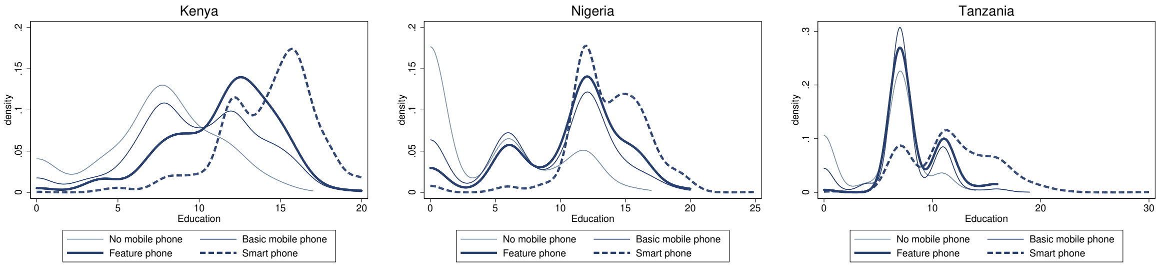
Note: This figure shows device ownership rates for female and male respondents. All figures use the sample weights provided.

Figure 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.

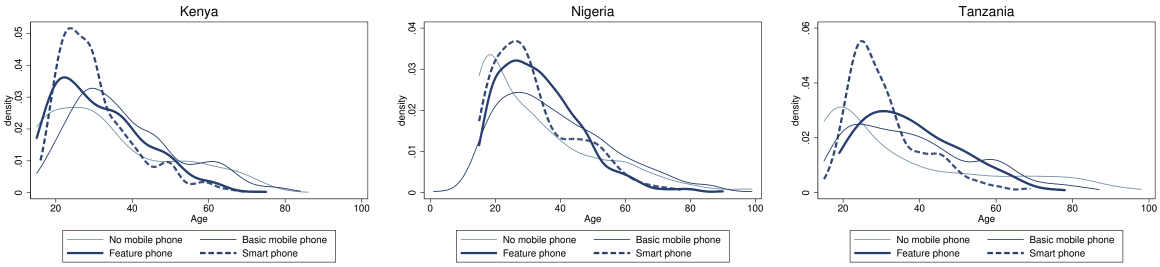
Figure C.3: Education and device ownership.



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.

polygons (more details in Appendix Section B). 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.

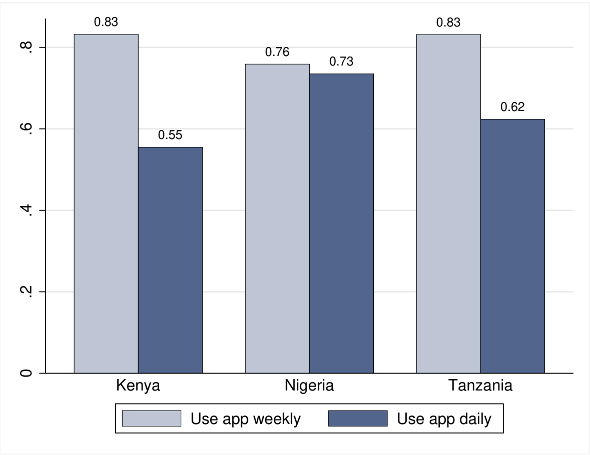
Figure C.4: Age and device ownership.



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.

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

Figure 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.

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

[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.

[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.

Table 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).

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

[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 B for details on the definition of city boundaries.

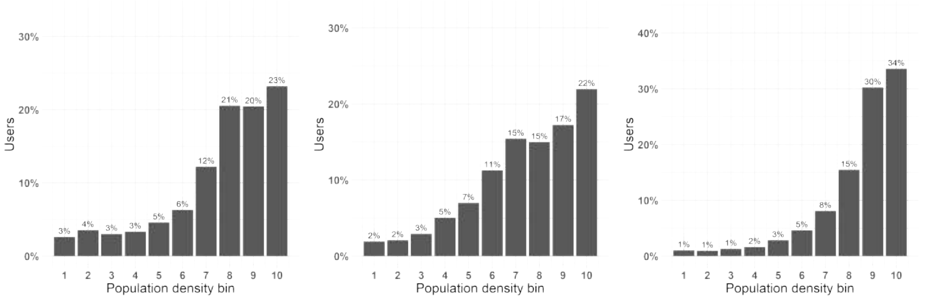
number of these comparison clusters will be either zero or a strictly positive number of clusters. Not all these nearby clusters will offer valid comparisons, however. For example, a user at the outskirts of Dar es Salaam might be associated with a nearby rural cluster as well as a number of urban clusters. 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 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.

[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.

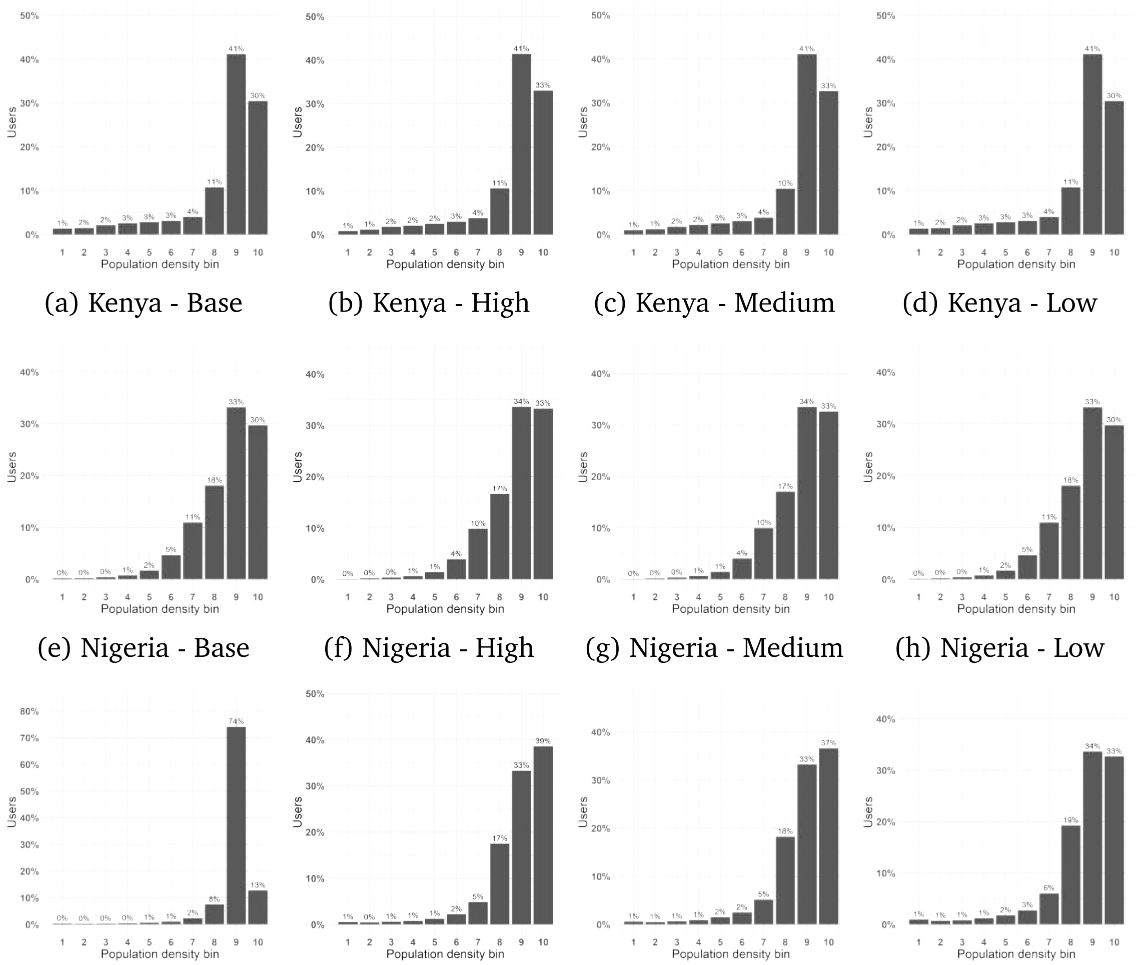
E. Additional tables and figures

Figure E.1: Users by population density decile, Landscan.



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

Figure E.2: Fraction of users by population density deciles.



(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.

Table 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 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***

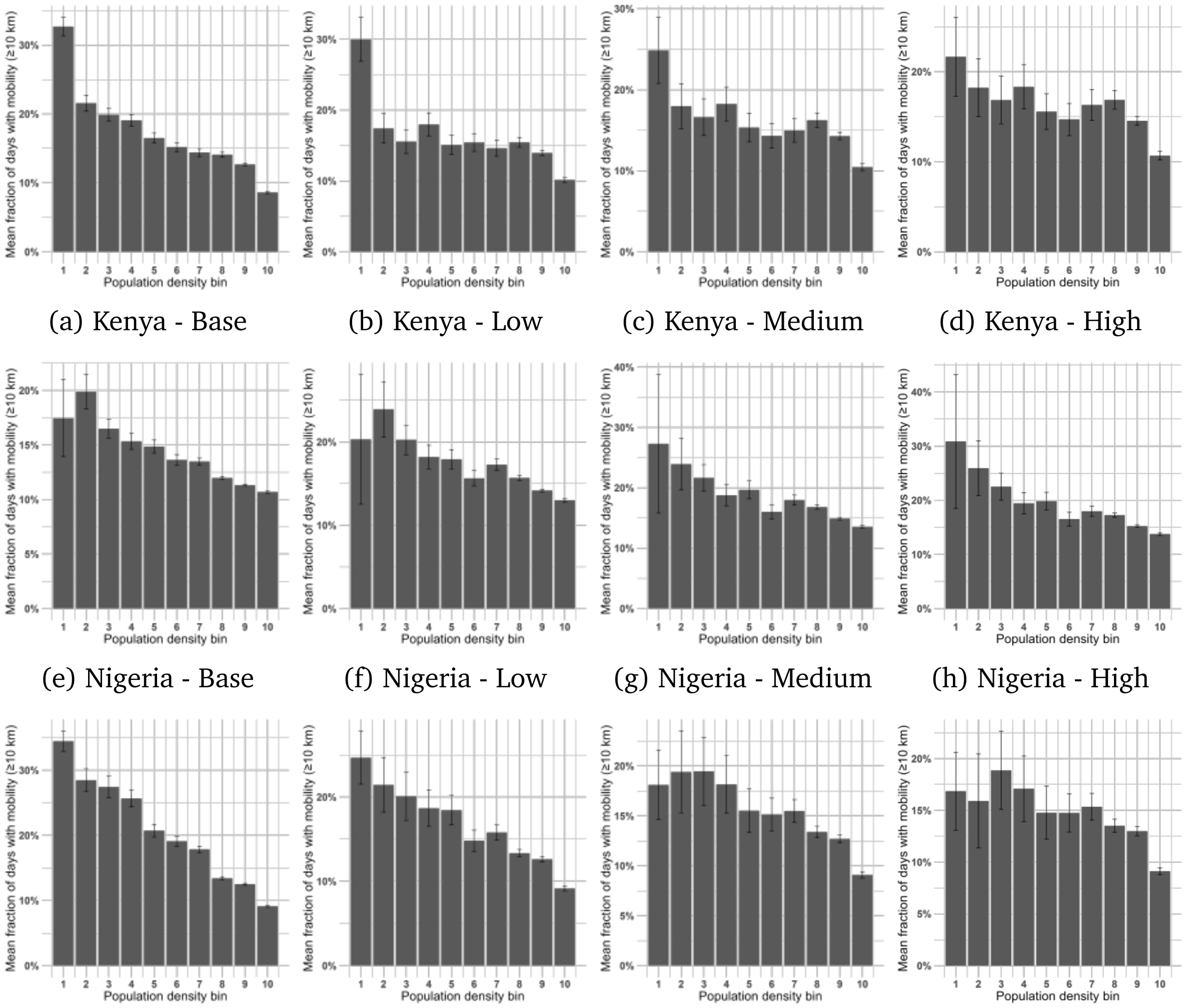
Note: This table compares the means between the overall DHS sample and the "Matched DHS" sample (DHS clusters with which we can match smartphone app users). We show a t-test that compares the two data sets, with bootstrapped standard errors robust to heteroskedasticity. Survey weights are used for the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

Table 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.

Figure 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 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.

Table 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 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
Visited density	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%
	5	6%	4.5%	8.5%	11.1%	12.7%	10.2%	5%	4.2%	1.7%	0.9%
	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
Visited density	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%
	5	2.9%	5.1%	8.1%	10.2%	10.8%	5.7%	2.6%	1.4%	1%	0.6%
	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
Visited density	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%
	5	1.8%	6.1%	9.6%	9.3%	9.8%	8.7%	4.3%	1.7%	0.8%	0.3%
	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 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 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
Visited density	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%
	5	29.7%	25.3%	42.3%	51.5%	52.4%	48.3%	37.5%	34.3%	24.1%	17.8%
	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
Visited density	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%
	5	23.8%	33.3%	42.7%	45.3%	50.6%	37.1%	26.2%	20.2%	19.4%	14.6%
	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
Visited density	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%
	5	16.5%	29.4%	42.7%	39.5%	36.4%	41.4%	25.6%	17.8%	12.2%	7%
	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 E.9: Origin of visitors in top 5 cities.

Kenya									
Nairobi		Mombasa		Nakuru		Eldoret		Kisumu	
(1,699 visitors)		(953 visitors)		(891 visitors)		(448 visitors)		(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		Kano		Ibadan		Abuja		Kaduna	
(5,258 visitors)		(807 visitors)		(2,916 visitors)		(3,232 visitors)		(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		Zanzibar		Mwanza		Arusha		Mbeya	
(1,850 visitors)		(743 visitors)		(704 visitors)		(859 visitors)		(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%

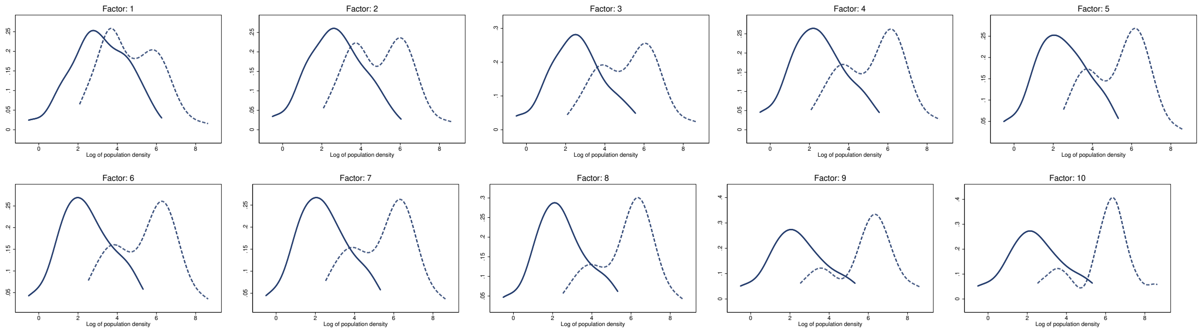
Note: This table shows the origin of visitors for the five most populated cities. Origin and destination city boundaries are defined using 3km-buffered GRUMP polygons. Visitors are defined as being seen at least once in a location over the year. "Non-urban" refers to locations outside boundaries of cities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

Table 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.

Figure 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.