DISCUSSION PAPER SERIES

DP16661

Perpetual Motion: Human Mobility and Spatial Frictions in Three African Countries

Douglas Gollin, Martina Kirchberger and Paul Blanchard

DEVELOPMENT ECONOMICS

MACROECONOMICS AND GROWTH

ISSN 0265-8003

Perpetual Motion: Human Mobility and Spatial Frictions in Three African Countries

Douglas Gollin, Martina Kirchberger and Paul Blanchard

Discussion Paper DP16661
Published 21 October 2021
Submitted 11 October 2021

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

Perpetual Motion: Human Mobility and Spatial Frictions in Three African Countries

Abstract

Recent literature points to the importance in developing countries of spatial and sectoral gaps in wages and living standards. These gaps seemingly imply frictions to human mobility. In this paper, we present new evidence on mobility within three African countries. We use a novel data source that provides highly detailed location data on more than one million smartphone devices across three large African countries for an entire year. This allows us to examine high-frequency mobility patterns for a subset of people for whom we can determine home locations confidently. The data offer insights into patterns of mobility, with corresponding implications for the nature and extent of mobility frictions. In particular, our data point to the ubiquity of relatively high-frequency journeys within countries -- i.e., visits. We observe that many users travel to relatively distant within-country locations, with big cities acting as particularly attractive magnets. 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. The analysis suggests that distance-linked mobility costs are not so high as to discourage travel. In fact, travel enables users to benefit from the opportunities that large cities provide, without having to incur the costs of relocation. The frictions sustaining spatial and sectoral gaps thus seem to reflect deep fixed costs associated with the dislocations of migration and sectoral change, rather than the direct costs of movement.

JEL Classification: O1, O12, O18, R12

Keywords: spatial frictions, mobility, cities, Sub-Saharan Africa

Douglas Gollin - douglas.gollin@qeh.ox.ac.uk
Oxford University and CEPR

Martina Kirchberger - martina.kirchberger@tcd.ie
Trinity College Dublin

Paul Blanchard - blanchap@tcd.ie
Trinity College Dublin

Acknowledgements
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, Dave Donaldson, Kevin Donovan, Vernon Henderson, Clement Imbert, Gabriel Kreindler, David Lagakos, 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, Trinity College Dublin, DENeB Berlin,Dortmund, Jena, Regensburg, University College Cork, Sheffield, UC Riverside,

the University of British Columbia, 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.

Perpetual Motion: Human Mobility and Spatial Frictions in Three African Countries

Paul Blanchard — *Trinity College Dublin*

Douglas Gollin — *University of Oxford*

Martina Kirchberger — *Trinity College Dublin*

October 2021

Abstract

Recent literature points to the importance in developing countries of spatial and sectoral gaps in wages and living standards. These gaps seemingly imply frictions to human mobility. In this paper, we present new evidence on mobility within three African countries. We use a novel data source that provides highly detailed location data on more than one million smartphone devices across three large African countries for an entire year. This allows us to examine high-frequency mobility patterns for a subset of people for whom we can determine home locations confidently. The data offer insights into patterns of mobility, with corresponding implications for the nature and extent of mobility frictions. In particular, our data point to the ubiquity of relatively high-frequency journeys within countries – i.e., visits. We observe that many users travel to relatively distant within-country locations, with big cities acting as particularly attractive magnets. 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. The analysis suggests that distance-linked mobility costs are not so high as to discourage travel. In fact, travel enables users to benefit from the opportunities that large cities provide, without having to incur the costs of relocation. The frictions sustaining spatial and sectoral gaps thus seem to reflect deep fixed costs associated with the dislocations of migration and sectoral change, rather than the direct costs of movement.

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, Dave Donaldson, Kevin Donovan, Vernon Henderson, Clement Imbert, Gabriel Kreindler, David Lagakos, 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, Trinity College Dublin, DENeB Berlin,Dortmund, Jena, Regensburg, University College Cork, Sheffield, UC Riverside, the University of British Columbia, 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

In most developing countries, there are large gaps in nominal wages and productivity across sectors (Gollin, Lagakos, and Waugh, 2014). There are similarly large gaps in living standards across space, with people in sparsely populated rural locations consistently worse off than those in dense urban settlements (Gollin, Kirchberger, and Lagakos, 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; Lagakos, Mobarak, and Waugh, 2018). This paper aims to advance our understanding of sectoral and spatial gaps by documenting and analyzing high-frequency mobility patterns within three low-income African economies.

Understanding human mobility patterns in low-income contexts has previously been limited by the lack of data. Census data or standard household surveys (e.g., those carried out in collaboration with the World Bank's program of Living Standards Measurement Surveys) typically measure migration flows between survey waves, but there is typically little data about movements that do not involve changes in an individual's home location. A few surveys provide detailed commuting data for high-income countries (e.g., the American Community Survey), but these normally miss non-work trips. Moreover, such surveys are not available for most low-income countries.

In this paper, we focus on a third type of mobility that has been traditionally difficult to capture: visits. By examining the frequency with which individuals move across space – from rural areas to towns and villages, or between cities – we can assess the salience of some key frictions related to the variable costs of distance. For instance, a world in which rural people travel frequently to distant towns and cities is not one in which narrowly defined costs of mobility (i.e. the cost of a bus ticket) can plausibly explain sectoral or spatial gaps.

We measure mobility using newly available, fine-grained, but anonymized, data on smartphone locations. Unique to our study is the scale at which we can study the phenomenon of short-term population movements. Our 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.

Each observation in our data reflects an instance when a user's phone connects to the internet to use a certain app. For each such use, we observe the GPS location and the precise time. This type of data has been used, for example, to study the length of time that indi-

[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.

viduals spend with their families for Thanksgiving in the US (Chen and Rohla, 2018), to construct a measure of experienced segregation (Athey, Ferguson, Gentzkow, and Schmidt, 2020), to study the effect of chance meetings on knowledge spillovers in the Silicon Valley (Atkin, Chen, and Popov, 2020), or to measure the effectiveness of social distancing (Mongey, Pilossoph, and Weinberg, 2020), social interactions (Couture, Dingel, Green, Handbury, and Williams, 2020) and the consumption value of transport investments (Miyauchi, Nakajima, and Redding, 2021). We add to this literature by focusing on three countries in sub-Saharan Africa and by looking at patterns of mobility.

We use the fact that devices are seen at different locations within countries to characterize movements of users between relatively distant locations. We use the data to map and categorize the movements of people and the connectedness of locations. 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.

The paper makes three main contributions. First, we construct a novel set of metrics for characterizing mobility across space related to frequency, spatial extent, densities and locations visited. Second, we analyze these measures to provide insights into the patterns of human mobility within the three countries where our data originate. Third, we develop a conceptual framework in which individuals decide what locations to visit. This framework delivers testable propositions.

We find striking evidence of a high degree of mobility within these three African countries. Although the population we study is undoubtedly atypical for Kenya, Nigeria, and Tanzania, we find evidence that a substantial fraction of the people in these countries is 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 rural people 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 relatively remote areas are isolated – and therefore ignorant of what goes on in the big cities. The data also cast doubt on the notion that the monetary and non-monetary costs of mobility are simply prohibitive. 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. This too is important for our understanding of spatial frictions. One could imagine that rural people seldom travel beyond the nearest town. But the data show persuasively that people from remote villages and towns do indeed travel to capital cities. These findings can help inform our understanding of spatial frictions in developing countries. In particular, they can discipline models that incorporate spatial and sectoral frictions. 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.

Our analysis requires some serious discussion of the representativeness – or lack thereof – of the population from which our data are drawn. While smartphone users might arguably, in 2021, be representative of the wider population in rich countries, it would be unreasonable to assume that the same holds in low-income countries. We do not have personal information on the users of our devices, and for ethical and privacy reasons, we have not attempted to exploit the data for the purpose of extracting identifying information.[2] The one personal characteristic that we construct for each device is its "home location". We define this as the modal 0.01-degree cell ($\approx 1.1km$ at the equator) at which we observe the user between 7pm and 7am.[3] We then select a subset of “high-confidence” users that we observe at least 10 nights and who spend at least half of these in the home location.

We characterize our users in three steps. First, we compare the distribution of population density at our users’ home locations with that of the overall population. Second, we use nationally representative micro-data from the ICT Access and Usage Surveys to compare basic characteristics such as income, sector of employment, age, and education of individuals by device ownership, examining individual characteristics of those without a phone, a basic mobile phone, a feature phone and a smartphone. Third, we propose a new method to characterize the places our users reside in by linking user’s home locations with widely available micro-survey data. While we do not have characteristics of users such as age, education or gender, this allows us to gauge how representative our users’ home locations are compared to locations where no users reside. We argue that these steps are crucial in order to understand the characteristics of our users, given that our data are generated from a population that does not represent a statistical sample, selected with the benefit of defined sampling frames and protocols. A rough summary of our efforts to characterize the sample is that we find our sample to be, unsurprisingly, more urban than the population as a whole. Smartphone owners tend to be younger and better off. However, our best assessment is that our population is not extraordinarily atypical in any of these dimensions. As smartphones

[2]It might be possible, for instance, to use the location information from individual users to identify or infer their religious observance, gender, and other attributes. In this paper, we do not attempt to use the data for these purposes, in recognition of the obvious privacy concerns.

[3]In practice, we define the modal 2-decimal rounded location at night as the home location, so our 0.01-degree grid cells have 2-decimal rounded coordinates as centroids.

have diffused through sub-Saharan Africa in recent years, smartphone users have come to look more and more like the rest of the population, and our users' home locations seem generally unremarkable. The urban users seem to live in "relatively normal" urban areas (in a sense that we will characterize below), and our rural users live in similarly normal rural areas. This trend is only going to become more pronounced over time. In the sections that follow, we address the nature of our sample in greater detail.

This research builds on a growing literature in economics that seeks to understand the salience of spatial frictions for development and growth. Spatial frictions limit the mobility of goods, information, and people within economies. 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.[4] Spatial frictions also may limit the processes of sectoral reallocation, as in Caselli and Coleman (2001) and Eckert and Peters (2018).

The importance of within-country spatial frictions in the movement of goods has been documented in recent work (e.g., Arkolakis, Costinot, and Rodríguez-Clare (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. In particular, a number of papers (e.g., Aker (2010); Jensen (2007)) have shown the impact of mobile phones on the dispersion of prices across space. Allen (2014) suggests that information frictions can compound spatial frictions.

Both these literatures have used new data sources to understand spatial frictions at a highly localized level. For movement of goods, many frictions occur in the proverbial "last mile", making it important to consider spatially disaggregated data. Similarly, for information frictions, studies have often looked at price dispersion across nearby markets. For movement of people, however, the literature has largely focused on crudely defined measures of migration or broad-brush comparisons of rural-urban gaps in living standards (Young (2013); Hamory, Kleemans, Li, and Miguel (2021a); Bryan, Chowdhury, and Mobarak (2014); Akram, Chowdhury, and Mobarak (2017)). Our paper builds on the work of Blumenstock (2012) and Lu, Wrathall, Sundsøy, Nadiruzzaman, Wetter, Iqbal, Qureshi, Tatem, Canright,

[4] It is not entirely clear whether one should view real frictions – such as transport costs – as a source of inefficiency, in the way that a tariff would be viewed as inefficient. The social planner cannot wish away distance or mountain ranges; real resources would be required to reduce or eliminate these frictions. To avoid this largely semantic issue, we prefer in this paper to use the term "friction" rather than "barrier" or "distortion," and we have sought to avoid language that would imply "inefficiency."

Engø-Monsen et al. (2016) in using more spatially detailed data on human mobility.

Beyond the implications for spatial frictions, our analysis points to a number of interesting features of the data. First, the widespread prevalence of rural people visiting cities suggests that urban areas generate benefits for a much broader set of people than their own residents and nearby commuters. Our data suggests that 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 rural people (and the inhabitants of small towns) to break down the rural-urban binary. Put differently, 'visits' allow people to achieve partial urbanization. In a sense, 'visiting' cities substitutes 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 help 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. The ubiquity of visits suggests that we should be cautious in treating rural and urban areas as entirely distinct; they are instead connected by significant flows of goods, information, and people. Finally, from a policy perspective, our findings highlight that as cities are growing in many parts of the developing world, they will need to accommodate not only a larger number of residents but also visitor populations attracted by the amenities they produce.

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.

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.[5] 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.

[5]In general, users give permission by clicking on an 'I Agree' button; whether this truly represents informed consent is a more complicated ethical question.

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.[6]

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).[7] Second, we consider two additional restrictions: (a) that a user is observed for a minimum of 10 nights; and (b) that the user is at the inferred home location for at least 50% of the total nights when that user is observed anywhere. These two restrictions eliminate cases where the user is seen infrequently at night, or is seen frequently but at multiple locations. Given the central role home location plays in our analysis, we define our core sample – which we call the "high-confidence" sample – as users that satisfy both criteria. Unless specified otherwise, we use our high-confidence users for our analysis.

We then examine the spatial distribution of home locations and identify users with apparent mislabelled locations by tabulating home locations and displaying them visually. This reveals what we call "irregularities" such as data sinks (e.g. a large fraction of users inexplicably assigned by the data-generating software to a spurious location, such as a country centroid) and other apparent errors in the data, e.g. users with equal latitude and longitude coordinates in locations where no people reside (visible on the 45 degree line on a map).[8] Appendix A provides further information on this quantitative extent of misclassification and our procedure to remove observations affected by apparent irregularities.

Table 1 shows the number of users and pings per user for our base sample of users and our

[6]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 Coid-19 pandemic and do not reflect any of the subsequent lockdown restrictions. Note also that we only observe a "ping" when a phone is connected to the internet. We are therefore not able to draw conclusions about areas without internet coverage.

[7]The accuracy of GPS data would theoretically allow us to infer home locations at higher resolutions. We choose to settle for a relatively coarser resolution to reduce computational time and because we deem it preferable for the analyses we conduct throughout the paper. In particular, for our purposes, we would like to consider pings from a few hundred meters apart as belonging to the same home location, rather than defining the home location as a particular house or plot of land.

[8]For example, 372,661 users in Tanzania seem to have been assigned to the identical arbitrary spot in the middle of Dodoma.

high-confidence sample.

Table 1: Sample and pings per user

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

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

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 under 120,000 devices across the three countries, with an average of over 2,000 pings observed per user.[9] 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.[10]

Table 2 summarizes user-level temporal statistics for our high-confidence users considering three different measures.

[9] We also build other subsets based on alternative values for the minimum number of nights observed: (i) the “medium confidence set" includes users with at least 8 nights observed and (ii) the “low confidence set" with users seen at least 5 nights in total. Our results are generally robust to using these alternative subsets.

[10] 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%.

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.

The first statistic that we consider is the duration over which we observe a particular user in the dataset, 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.[11] These statistics are roughly similar for the three countries. We see users on average over a span of about 100 days, on about 40 distinct days, and they have between 40 and 100 pings per day on average.[12] 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. These are issues explored in Bryan et al. (2014), Imbert and Papp (2020), and Lagakos et al. (2018). As smartphones become more common and usage patterns become more similar to those in high income countries, this is likely to change.

Appendix Figure C.1 shows the distribution of users and pings per user over time. The graphs show that there is an upward trend in the number of users over the period in which we observe the sample, likely due to a combination of factors. One is the steady and secular increase in the rate of smartphone ownership and usage. Another possible reason is the introduction of new apps in the sourcing of data during this period. The number

[11] This differs from the pings ratio in Table 1 which simply summed over all pings in the data across all users and divided by the number of users.

[12] 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.

of pings per user shows discontinuities, possibly due to further apps being added (or removed), users switching between apps, and app or operating system upgrades that change the way geolocation is recorded. Some of these changes presumably introduce classical measurement error, but others might not. For example, if users with lower incomes are added to the sample later, this could bias our estimates of mobility if we were to examine seasonal differences. We do not have access to income data for our users, but we can test whether individuals coming into the sample in later months reside in different densities. We find that the R-squared between the date of entering our data and population density at home location ranges from 0.0006 to 0.001 in our three countries. Most of our metrics are aggregated over the entire year, and thus they should not be sensitive to these particular discontinuities.

2.2. 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 (see Table A.2 in the Appendix). In this subset ("work subset"), home and work locations are found within the same 0.01-degree cells for 80% of users (see Table 3) which is in line with high rates of self-employment and short-distance commuting.[13] The lower panel in Table 3 shows that 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.

[13]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.

Table 3: Home and work locations

	Identical home and work locations			Home-to-work median distance (in km)		
	Overall	**Urban**	**Non-urban**	**Overall**	**Urban**	**Non-urban**
	(1)	(2)	(3)	(4)	(5)	(6)
Kenya	80.2%	78.5%	87.7%	4.7	5	2.5
Nigeria	79.8%	79.2%	84.4%	4.5	4.6	4
Tanzania	82.6%	80.6%	87.7%	3.3	3.5	1.6
TOTAL	80.4%	79.3%	86.1%	4.4	4.5	3

Note: Columns (1)-(3) show the fraction of users with identical home and work location. Columns (4)-(6) show the median distance between work and home locations for users with distinct work and home locations.

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 (see Tables A.3 and A.4 in the Appendix).

3. Characteristics of users

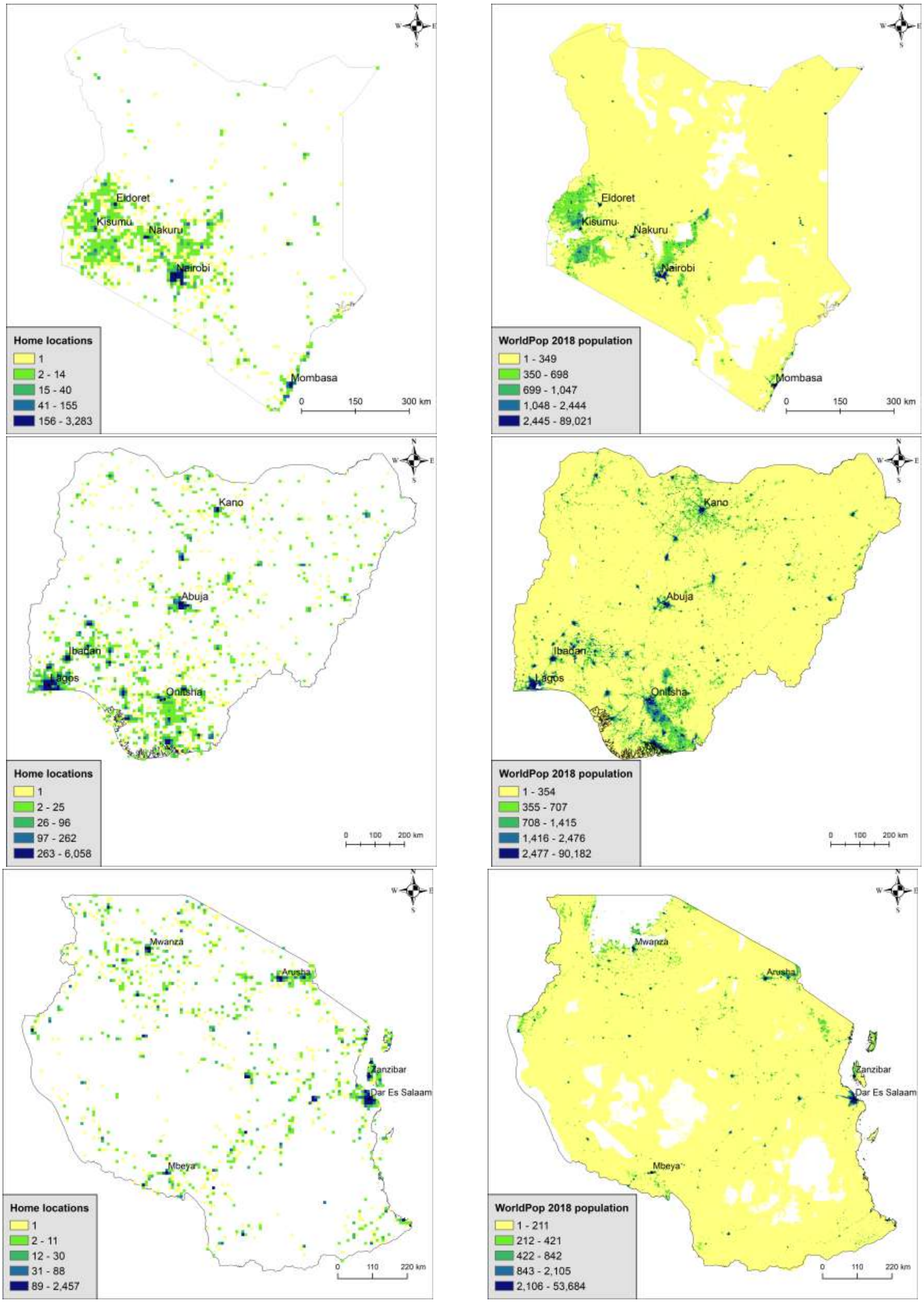
The key selection concern when using smartphone app location data is that we only capture individuals who own a smartphone. A further restriction affecting selection into our sample is that individuals require data credit on their phones, similar to requiring phone credit to make calls or send texts. 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. 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 World-Pop to understand how representative users are for different levels of population densities. Second, we draw on data from other nationally representative surveys – specifically, the ICT Access and Usage Surveys – to examine differences between individuals who own a smartphone and those who do not. To the extent that our population of smartphone app users is typical of all smartphone owners, these survey data will tell us something about how our users compare to the broader national populations of their countries. Third, to measure how representative our users are, in terms of their home locations, we develop a methodology to match home locations with nationally representative micro-data from the

Demographic and Health Surveys (DHS). This allows us to say something about whether the locations where our users live are typical or atypical.

Figure 1 shows the distribution of home locations in the left panel and compares it with the population distribution in the right panel.

Figure 1: Distribution of home locations and population.



Note: This figure shows the distribution of home locations of users (on the left) and the distribution of the population (on the right).

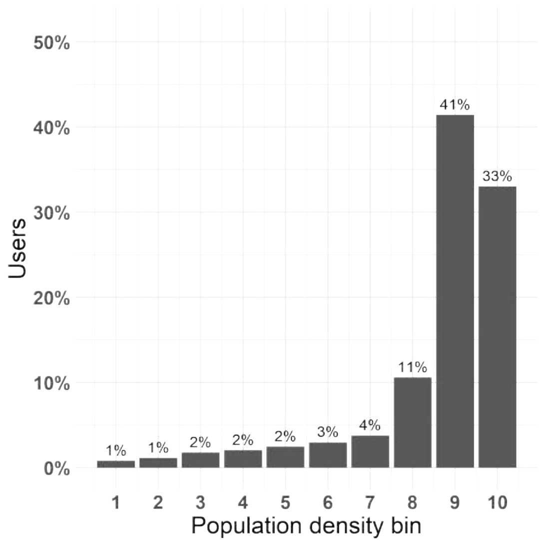
Darker values indicate a higher number of users. Unsurprisingly, we observe a higher number of users in the main cities. However, the figure shows that coverage of users is broadly national, with users residing in fairly distant places as well as in the densest cities. In fact, we have users in all but three of the 115 regional capitals in the three countries we study.[14] Maps in Figures C.2 and C.3 in the Appendix show these comparisons for the three capitals as well as Mombasa, Lagos and Dar es Salaam. Again, we find 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 2).[15] We also used other population density products such as Landscan as illustrated in Appendix Figure C.4. Using these products, our users are more represented in the lower quintiles. We therefore view these results as the most conservative population density distributions.

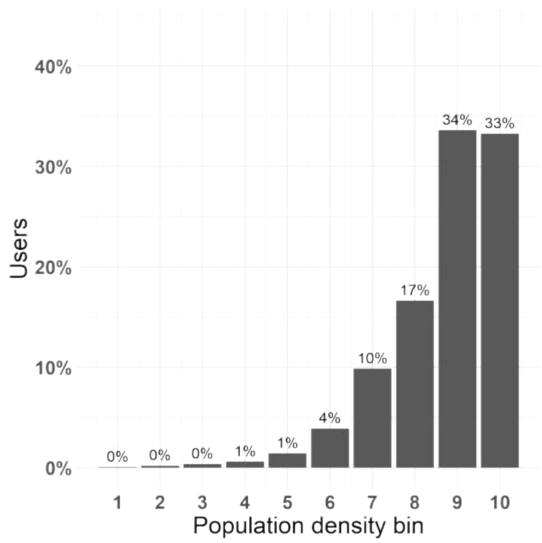
[14]Regional capitals are broadly understood as capital cities for subdivisions of the first administrative level. More specifically, Kenya has 47 counties, Nigeria has 36 states and a Federal Capital Territory and there are 31 regions (or *mikoa*) in Tanzania. Cities' boundaries are defined according to GRUMP 3km-buffered polygons (more details in footnote 18). 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.

[15]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. Note that our users are not evenly allocated across the density decile. As shown in Figure 2, our users are skewed towards the more densely populated bins.

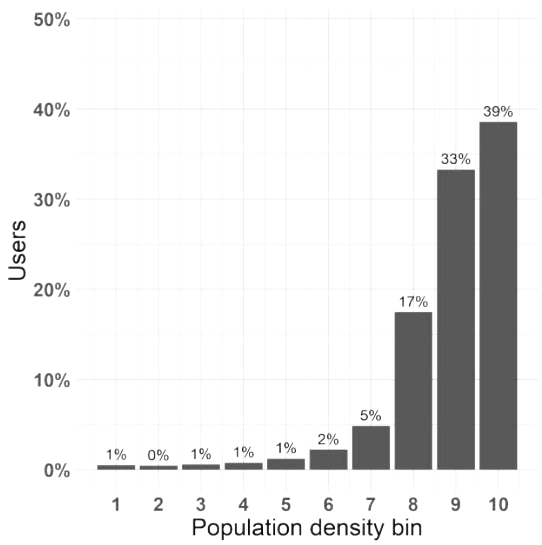
Figure 2: Users by population density decile.



(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).

We compute three further metrics to measure the representativeness of our users across different levels of population density: first, we take all 10-km pixels in a country and regress the number of users in a pixel on population of the corresponding pixel. We find that the R-squared ranges between 0.36 in Kenya to 0.81 in Tanzania, depending on the source of the population density estimates.[16] 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

[16]We find a significantly higher correlation when using Landscan data than other sources.

Tanzania.[17]

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.[18] 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. In all of our analysis, we account for this aspect of the data. The more interesting question is how our urban users compare with other urban dwellers, and how rural users compare with other rural residents. For this, we can turn to other data sources.

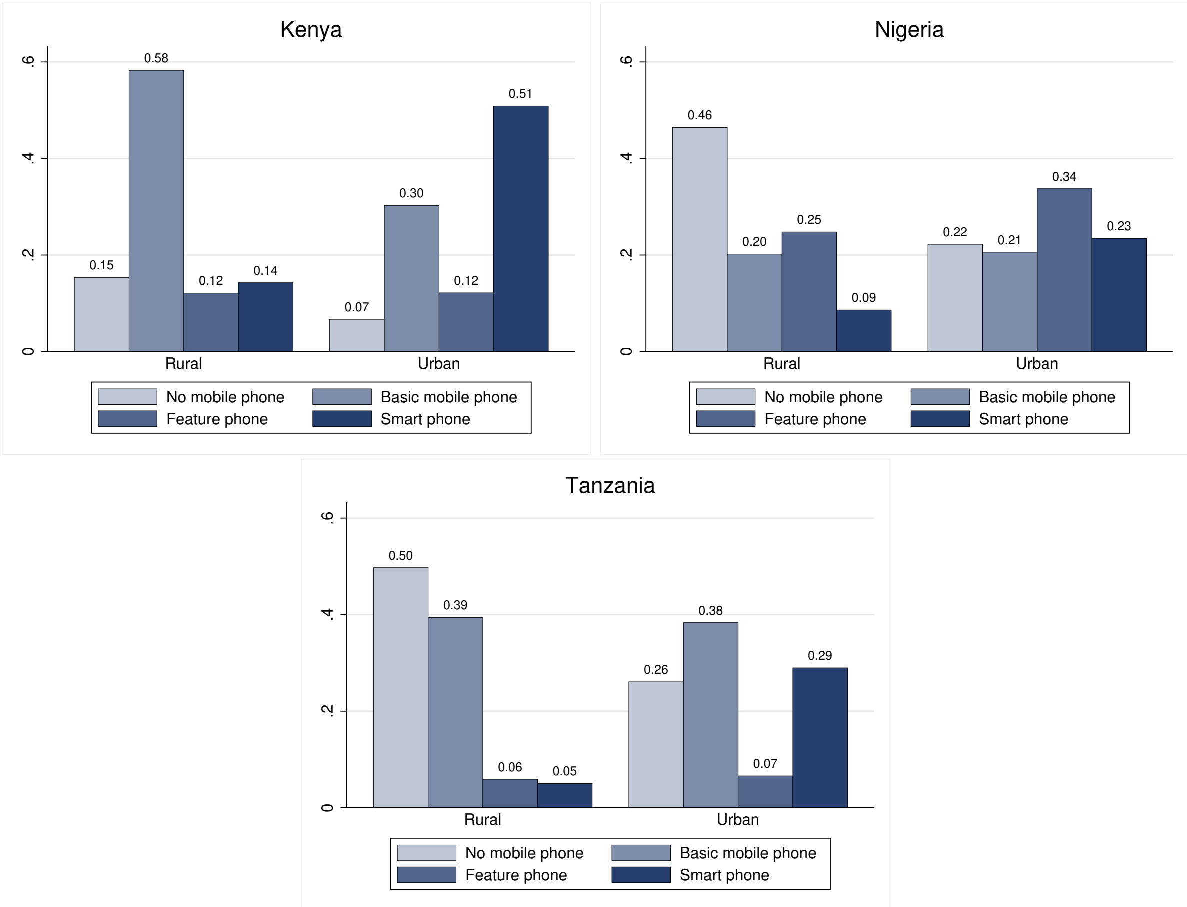
To what extent are smartphone users in the population represented in different locations, and how do they compare to the overall population in terms of income, education and age? To address this question, we use data from the ICT Access and Usage Survey 2017-2018 for Nigeria, Kenya and Tanzania. These 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.[19] Figure 3 shows ownership rates for different types of mobile phones, comparing rural and urban locations.

[17] See Appendix Figure C.5.

[18] 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. Documentation available on the dedicated webpage https://sedac.ciesin.columbia.edu/data/set/grump-v1-urban-ext-polygons-rev02. 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.

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

Figure 3: Device ownership by location.



Note: These figures show device ownership rates for rural and urban respondents. All figures use the sample weights provided.

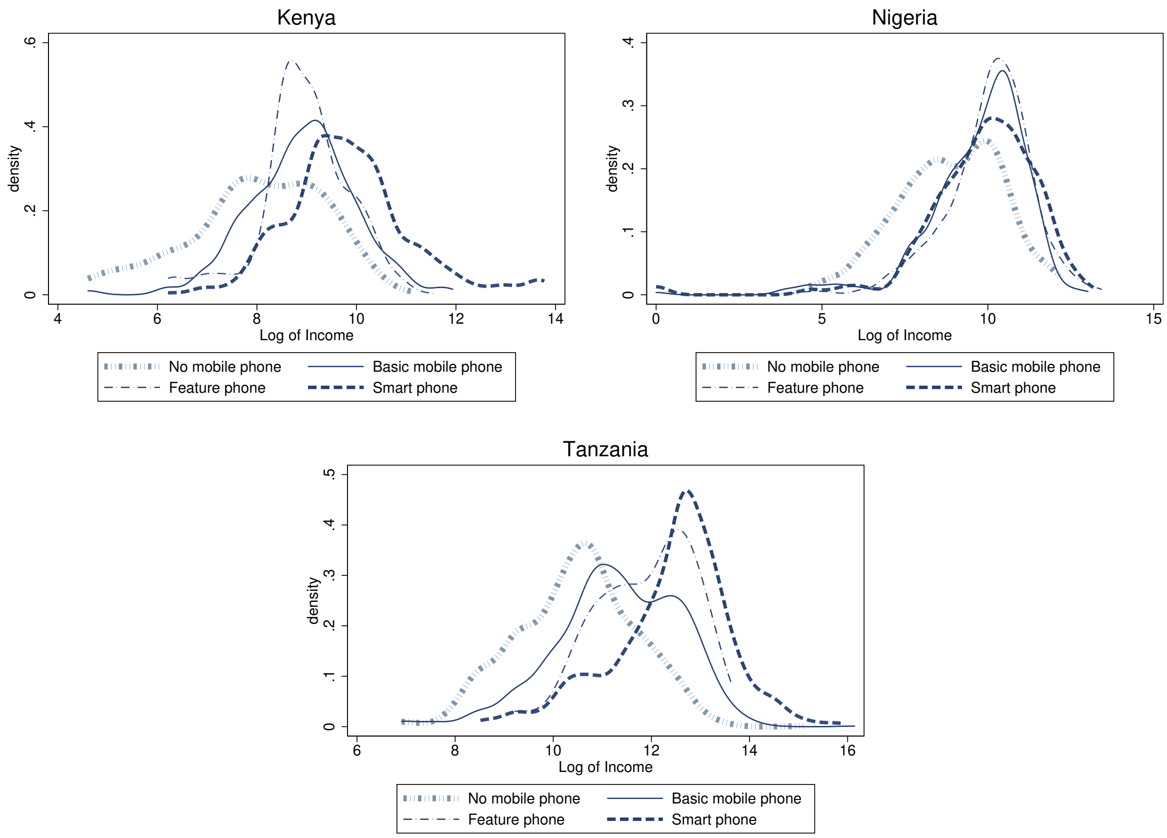
Compared to rural areas, respondents in urban areas are unsurprisingly more likely to own a mobile phone, and the phone is more likely to be a more sophisticated phone. The figure shows that in all countries smartphone ownership is highest in urban areas, with rates between 23 and 51 percent. If we include feature phones, this increases the rate to between 50 and 60 percent. The proportion of individuals with a basic mobile phone ranges between 21 and 38 percent. Across the rural areas of our three countries, smartphone and feature phone ownership is highest in Nigeria, at 31 percent penetration, and lowest in Tanzania, with 11 percent. When asking users for the reasons they do not own a smartphone, the main reasons given are affordability and not needing one, because a feature or basic phone is enough.[20] Figure C.6 in the Appendix examines ownership rates by gender. 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.

To examine how owners of these different devices differ from each other – and perhaps more

[20] This question is available only for a small number of users from Nigeria.

important, from those without a device – we compare users by income, age and education. Figure 4 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. Appendix Figure C.7 compares the number of years of schooling and Appendix Figure C.8 shows the age distributions. Both highlight that these distributions are not distinct.

Figure 4: Income and device ownership.



Note: These figures show the distribution of income by device ownership. All figures use the sample weights provided.

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 RIAS survey does not ask for the sector of employment, but does ask for income from different sources.[21] We use this information to assign a main income source to each respondent.[22]

Table 4 shows the proportion of smartphone owners across different categories in columns

[21]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.

[22]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.

(1), (2) and (3) and we compare this to the sample averages in columns (2), (4) and (6).

Table 4: Smartphone ownership and main source of income.

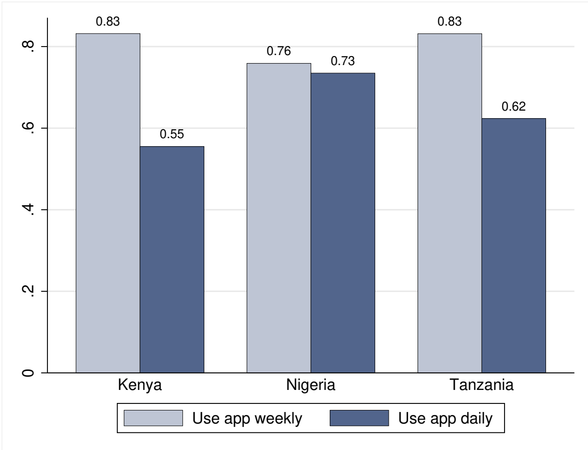
	Kenya		Nigeria		Tanzania	
	(1)	(2)	(3)	(4)	(5)	(6)
Rural						
Salary or wage	54.9	26.7	24.3	8.4	51.6	15.0
Agricultural produce/farming	9.9	34.0	8.7	25.3	18.6	34.2
Vending/trading	3.8	1.8	11.0	15.2		
Work you are doing at home	1.2	5.0	2.4	2.0	0.0	0.5
Income from your business	6.0	8.9	14.4	19.1	20.6	10.7
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 income sources in columns (1), (3) and (5) compared to the sample averages in columns (2), (4) and (6). All figures use the sample weights provided.

A few points are worth highlighting: first, smartphone owners are represented across almost all categories, and the proportions are generally in line for several of these categories. For example, in rural Kenya 6 percent of smartphone users have a business as their main source of income compared to 9 percent of our sample. Second, between one-fourth and one-half of smartphone users earn a salary or wage. We are therefore likely to oversample from this category. Third, there are fewer farmers owning smartphones than their corresponding proportion in the overall sample, so that we are likely undersampling this occupation. Given that these are likely to be among the poorest individuals this is not surprising. Overall, with the caveat in mind that the survey only asks one question about main income, rather than including a full labor market module similar to LSMS surveys, the table suggests that smartphone users are not just from one occupation (e.g., traders) but are represented across different types of economic activities.

The survey also asks respondents about their usage of a range of apps from social networking apps (facebook, Whatsapp, Instagram) to news, weather, trading, business, health and dating apps. Figure 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.

Figure 5: App usage of smartphone users.



Note: These figures show the fraction of smart phone owners using apps weekly or daily. All figures use the sample weights provided.

Finally, 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.[23] 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 B 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.[24]

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

[23] Adding to the challenge is that the published locations for the DHS clusters are randomly displaced (between 0 and 10 kilometers) in an effort to ensure data confidentiality. To be precise, the published geo-referenced locations for the DHS clusters are displaced by selecting a random compass direction and then a random distance. Urban DHS clusters are randomly displaced by 0-2km, and rural clusters are randomly displaced by 0-5km, with 1 percent of clusters randomly selected to be displaced by 10km (Perez-Heydrich, Warren, Burgert, and Emch, 2013).

[24] The country-specific matching rates are 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.

clusters. In other words, if we look at the set of DHS clusters where we find our users, we can ask whether this matched DHS sample looks statistically similar to the overall ("raw") set of DHS clusters. We carry out this analysis by conducting t-tests for equality of means between the raw DHS and matched DHS samples on a range of directly quantifiable household characteristics, such as whether the household has a constructed floor, walls, roof, overcrowding and access to public services such as electricity and tap piped water. Moreover, we produce results for rural and urban sub-samples separately to account for both the prevalence of urban users in our sample and the lower matching rate in low density areas, which together may lead to results being mainly driven by the urban component of the sample. We produce t-tests comparing our two weighted data sets, with bootstrapped standard errors robust to heteroskedasticity. The survey weights are used for the reference DHS sample, while those of the matched DHS sample correspond to the number of users each cluster is paired with. Definitions for all variables used are given in Appendix Table B.1.

Appendix Tables C.1-C.3 show that, perhaps unsurprisingly, 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 is perhaps more striking, 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. 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. In short, users live in locations that are not fully representative of the national population – but at the same time, these locations are not wildly atypical or weirdly distorted. We are not seeing only a small population of people living in gated communities or in rural holiday spots. The locations where our users live look fairly similar to a nationally representative sample of locations where people live.

Our takeaway message from this analysis is that our population of users provides useful information about patterns of mobility in the broader national population. The income distributions of smartphone users are not distinct from the overall sample of mobile phone owners and they are represented across the main economic activities. Our urban users live in places that are similar to the places where other urban residents live; our rural users live in locations that are not especially different from other rural locations. Within urban locations, smartphone users are a relatively large fraction of the population. Speaking very loosely, they represent the most privileged one-third to one-half of the distribution – but not just the top one percent. Even in rural areas, smartphone users account for over one-third of the Kenyan population, and more than 10 percent in Tanzania. It is true that even within the populations of smartphone owners, our users may be atypical. Our high-quality subset consists of people who use their devices relatively frequently, and this may bias us towards users who are more mobile and more sophisticated than the average. But as people in these countries have begun to use their phones for messaging and social media, one suspects that this distinction – to the extent that it ever held true – may not be very pronounced. In sum, we believe that the evidence supports us in using these data to inform a discussion of human mobility within these three countries. We are almost certainly looking at a subset of the population that is more mobile than the average; but equally, we are looking at only a subset of the total mobility. In other words, the trips that our users have taken are a subset of the total trips made; and the connections that we identify between locations are a subset of the total connections.

To conclude this section of the paper, we take up more fully the potential biases that we may have introduced by equating "devices" with "users". We also consider other potential challenges in working with our ping data.

We acknowledge that distinct users may use the same device, and individual users might have multiple devices. We discuss these issues in turn. 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. Unfortunately the survey does not ask which type of mobile phone a respondent used, nor are respondents asked whether this was the respondent's own phone at the time.[25] However, 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. If several people share a device that is used, for example, to travel to the nearby city, this would lead us to capture travel by several household members rather than just one device user. Given that we are interested in the flow of people between locations (and not necessarily the particular person), this would still give us a reasonable measure of human mobility between locations.

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. This is an advantage of our data relative to the CDR data widely used in many development applications; although ownership of multiple active SIM cards is relatively rare in Kenya (more than 80 percent of individuals have only one active SIM

[25] It is possible, for instance, that the respondent lost his/her phone or had it stolen during that time period, or perhaps even that the phone died or was sold.

card), it is fairly common in Nigeria and Tanzania where less than 57 percent of individuals have only one active SIM card.

However, 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 – for example, because different mobile providers may offer better coverage in certain geographies. 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. 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 then disaggregate both the origin and destination locations into densities and summarize our data in the form of a spatial transition matrix to examine the connections between remote and dense areas.

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.[26] 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 $p_i(x) = 0$

[26]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.

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 $\bar{d} = 10$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 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:

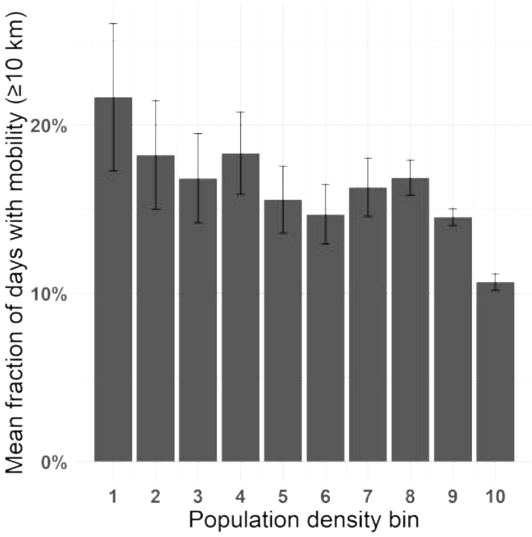
$$F^j = \frac{1}{n_j} \sum_{i \in I_j} F_i. \tag{2}$$

Figure 6 shows this frequency for all three countries, broken down by density bin. The pattern is consistent across countries: on roughly 10-20 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 (see Figure C.12

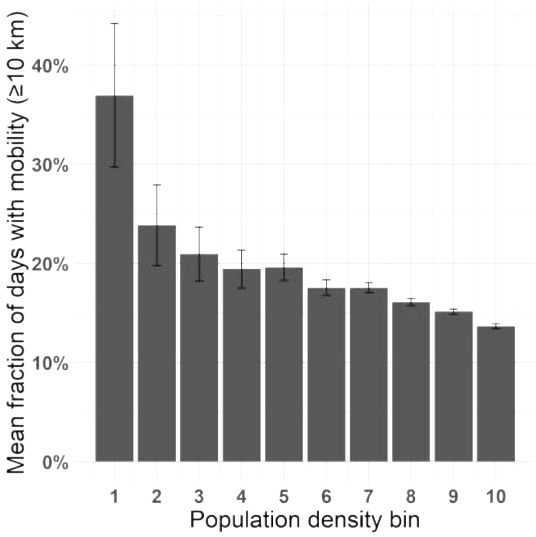
in the Appendix). One plausible interpretation is that those who live in relatively remote areas are likely to travel more frequently than those who live in towns and central cities. We cannot, of course, distinguish between the frequency of trips and the frequency with which users turn to their phones 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.[27]

[27] As a robustness check, we reproduce Figure 6 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. Results are provided in Figure C.13 in the Appendix. We observe small decreases in the average fraction of days away in all density bins but no change in the overall pattern of decreasing frequency with population density.

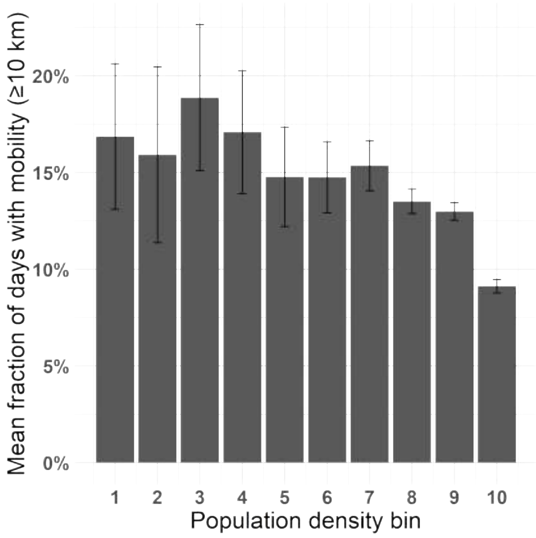
Figure 6: Fraction of days with mobility beyond 10km by density bin.



(a) Kenya



(b) Nigeria



(c) Tanzania

Note: These figures show 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:

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

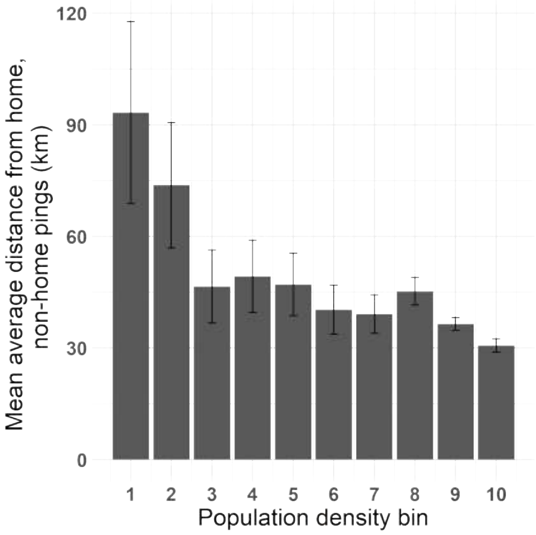
We find that the average distance of non-home pings ranges from 31.8 km in Tanzania to 45.5 km in Nigeria. In extrapolating this measure to a group of people, we can once again take an average. For example, we can measure the average of our spatial extent measure for the individuals belonging to a population density bin b_j by simply averaging the individual values of S_i. Thus:

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

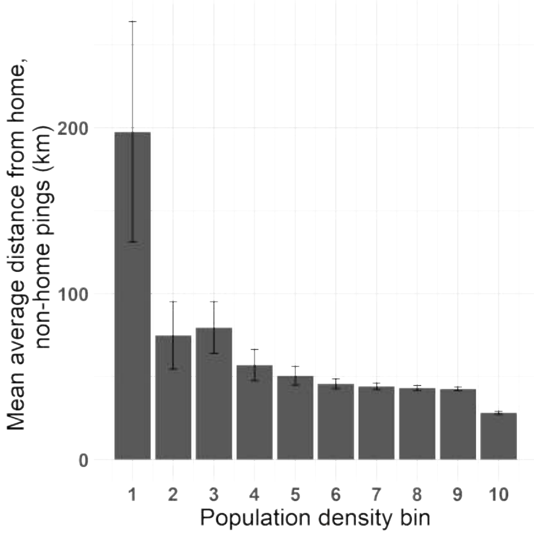
Figure 7 shows that non-home pings are not all highly local. In fact, the average distance – across countries and density bins – ranges from 30-100 km. As in Figure 6, we see a pattern across density bins suggesting that those in relatively sparsely populated areas seem to travel the farthest – in the 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. As for the frequency metric, we reproduce Figure 7, discarding values in the top 5 percentile to eliminate the effect of outliers; results are shown in Figure C.14 in the Appendix. Once again, the mean distance away from home falls in absolute terms for all density categories, but the overall pattern still holds for all three countries.

Taken together, Figures 6 and 7 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.

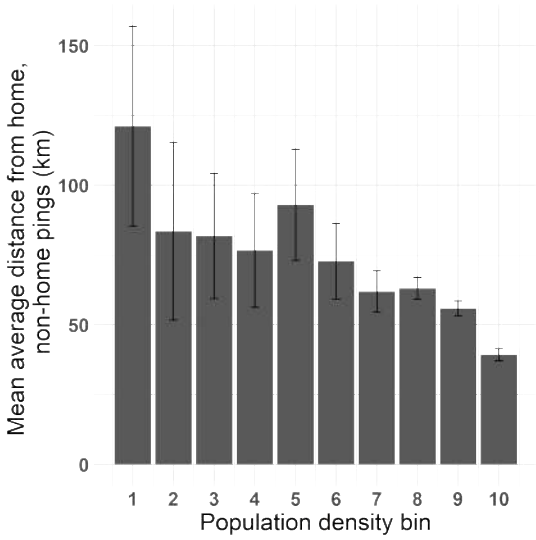
Figure 7: Mean distance away from home by density bin.



(a) Kenya



(b) Nigeria



(c) Tanzania

Note: These figures show the average distance from users' home locations of non-home pings by density decile over the period of a year.

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 visits (i.e., 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:

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

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

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

Table 5 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.[28] This table gives the fractions of users residing in a given density bin who are seen over the course of the observation span on at least one occasion in a non-home location within each of the ten density bins. For instance, this tells us that 6.7% of those Kenyans living in the most densely populated locations in the country were observed on at least one occasion during the year in a cell that falls within the *least* densely populated parts of the country. At the other end of the distribution, 32.1% of the users whose home locations are in the most sparsely populated areas of the country were observed at least once during the year in the most densely populated parts of the country.

To examine the role of potential biases caused by users being observed while travelling, i.e. at locations they might not actually visit, we rely on the procedure described in Section A.4 to purge the data from transit pings and calculate V_{od} and Δ_{od} on these purged subsets of pings. We find only minor differences (see Appendix Tables C.5 and C.6).

[28] Results for V_{od} (the average distribution of non-home pings across density bins) are shown in Appendix Table C.4 for our three countries.

Table 5: 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
Visited density	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%
	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
Visited density	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%
	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
Visited density	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%
	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.

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 Table 6 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 [29].

Table 6: Origin of visitors in top 5 cities.

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

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

[29] See footnote 18 for the definition of city boundaries. The reference year for city-level population counts is 2018.

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. 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").

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 Table 7.

Table 7: Top 5 destinations of residents from top 5 cities, Kenya.

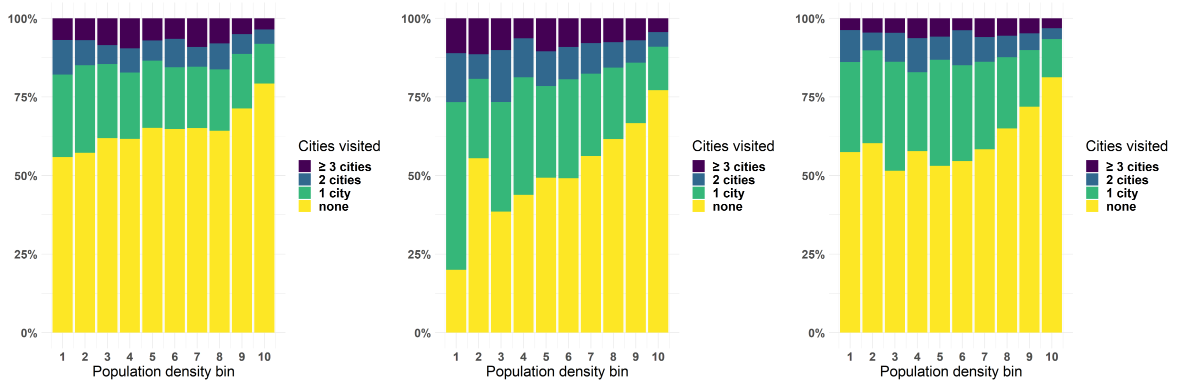
Kenya									
Nairobi (11,290 residents)		Mombasa (1,683 residents)		Nakuru (413 residents)		Eldoret (340 residents)		Kisumu (258 residents)	
Destination	*Residents*	*Destination*	*Residents*	*Destination*	*Residents*	*Destination*	*Residents*	*Destination*	*Residents*
Mombasa	5.8%	Nairobi	20.4%	Nairobi	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.

A striking feature of these tables is that the largest city is the leading destination for those living in almost all other cities – regardless of distance. Curiously, urban dwellers are also relatively likely to have been seen in non-urban areas. This is suggestive of the possibility that secondary cities are relatively substitutable for one another, but the largest cities (and perhaps also non-urban areas) offer benefits that are somehow distinct. This may reflect a lack of specialization and differentiation between secondary cities – an issue that has been raised previously in sub-Saharan Africa (see, for example, Henderson and Kriticos (2018)).

As a final step in our characterization of mobility, we examine in more detail the number of distinct visits individuals make as well as what type of amenities the data suggest people consume when making these visits. Appendix A.5 provides the details on how we define visits. Figure 8 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 a larger number of cities.

Figure 8: Distribution of users according to the number of cities visited, by population density bin.

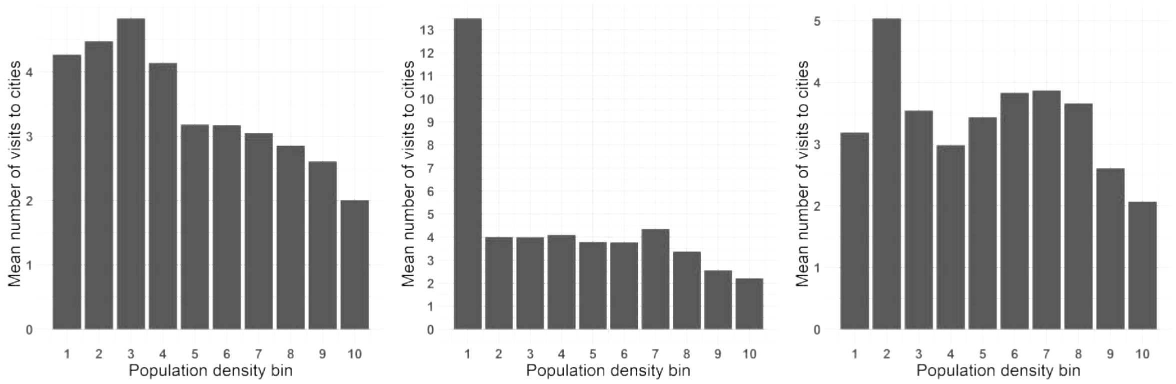


(a) Kenya (b) Nigeria (c) Tanzania

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.

To what extent are visits to cities events that occur as an exception rather than journeys individuals embark on with some regularity? Figure 9 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 9: Average number of visits to cities, by population density bin.

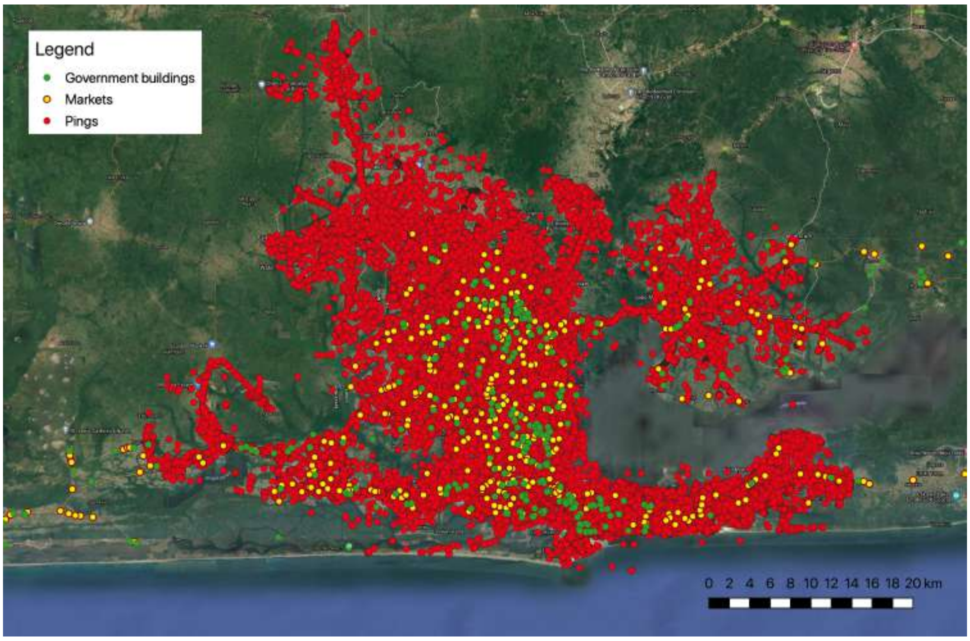


(a) Kenya (b) Nigeria (c) Tanzania

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

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, we can inspect the locations that visitors to cities are seen at. We present here the data for one such case study that we conduct in Lagos. Figure 10 shows pings by visitors alongside markets and government buildings. The figure shows that pings by visitors are spread out throughout the city rather than simply capturing visits to the main markets or government buildings, for example.

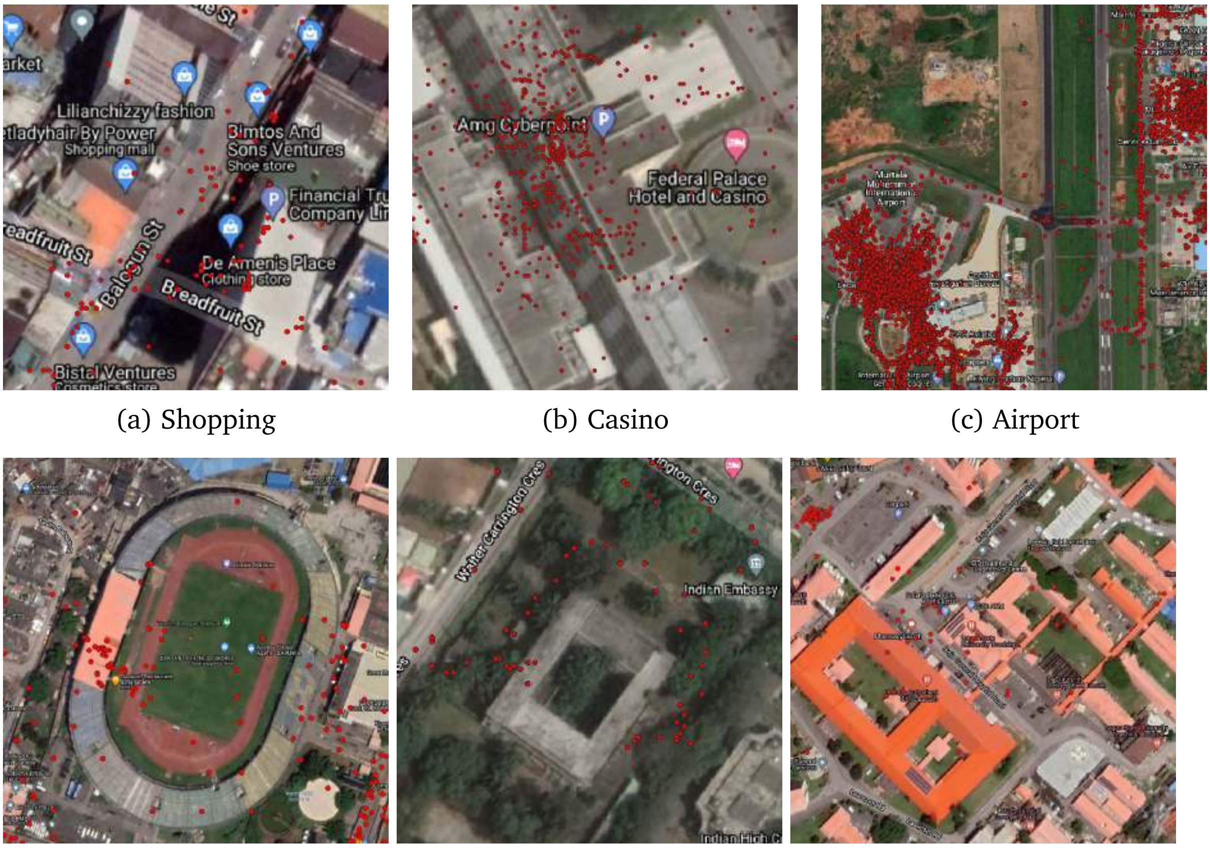
Figure 10: Destinations visited within cities: Lagos case study.



Note: This figure shows the distribution of pings of visitors to Lagos. Data on the location of markets and government buildings come from Federal Republic of Nigeria (2021).

Figure 11 gives a number of examples where visitors are seen from shopping areas to entertainment locations (e.g., casino, stadium), at the airport, public offices or health facilities. This is also in line with one of the key assumptions of the model, namely that cities provide a range of amenities increasing in city size and that amenities can be defined at the level of the city.

Figure 11: Destinations visited within cities: Lagos case study examples.



(a) Shopping (b) Casino (c) Airport

(d) Stadium (e) Embassy (f) Hospital

Note: This figure shows the distribution of pings of visitors to Lagos at specific selected locations.

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 more than 50 to 100 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.

The data point to a world in which mobility frictions are insufficient to choke off human mobility. At least some subset of the population in our three countries is highly mobile and provides a network of information flows across locations. If individuals are moving, there must also be comparable movement of goods and of information. Even if our data comes from a selected subset of the population, their movements may be sufficient to generate flows of knowledge and information about spatial gaps in living conditions or opportunities. The data suggest that we should question frameworks in which spatial gaps arise from a complete failure of information to flow across locations. Rather, they suggest that high-frequency travel is a way that allows individuals to consume amenities in distant locations without having to entirely relocate there.

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. (2021) 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 j on visit v, then:

$$\theta_{ix} = \sum_{v=0}^{V_{ix}} \theta_{ivx}$$

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:

$$\pi_a(h)a_i + \pi_c(h)c_i + \sum_x \left[V_{ix}(\gamma D_{ix} + \lambda) + \sum_{v=0}^{V_{ix}} \tau_x \theta_{ivx}^{\alpha} \right] \leq w(h).$$

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 $N_1, N_2, \ldots N_{\bar{K}}$ denote the populations of the different settlements; furthermore, without loss of generality, we can order the indexing such that $N_1 < N_2 < \ldots < N_{\bar{K}}$. 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).[30]

5.2.1. Location amenities

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

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

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 $w_c = \pi_c^*$.

[30] 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.

5.4. Equilibrium

We consider first 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, they will be higher in larger cities than in smaller cities, since larger cities are more productive in supplying these 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:[31]

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

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.*

5.5. Dynamic extensions and migration

Note that in this framework, we do not allow for permanent migration. This reflects the limitation of our data; we cannot observe with confidence any migration choices in our data,

[31]This equation is similar in flavor to a gravity equation coming out of quantitative spatial models developed by Ahlfeldt, Redding, Sturm, and Wolf (2015) and Kreindler and Miyauchi (2021).

since our observations cover only one year. However, we note that this framework would in principle allow for a long-run equilibrium in which individuals can choose to migrate; i.e., to change their home location. Migration would entail some further costs, reflecting the disruptions that migration imposes on work and family life. Given a longer data series, it might be useful to extend this model further.

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.[32] 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 8 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. Formally, the t-tests show that we can reject the null in favor of the alternative hypothesis that H_a which is that $V_k(1,2)/N_{k_1} > V_k(2,1)/N_{k_2}$, lending support to Proposition 1. Given that some of the location pairs might have small differences in populations, we also explore whether the distribution of visits become more distinct when we vary the difference between the origin and destination populations. Appendix Figures C.15-C.17 show that this is indeed the case.

[32]As for the rest of the paper, city boundaries are based on 3km-buffered GRUMP polygons (see footnote 18). Here we consider the subset of cities above 50,000 inhabitants – based on 2018 WorldPop population map. We exclude visits that originate in non-urban locations.

Table 8: 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.

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

Table 9: 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.

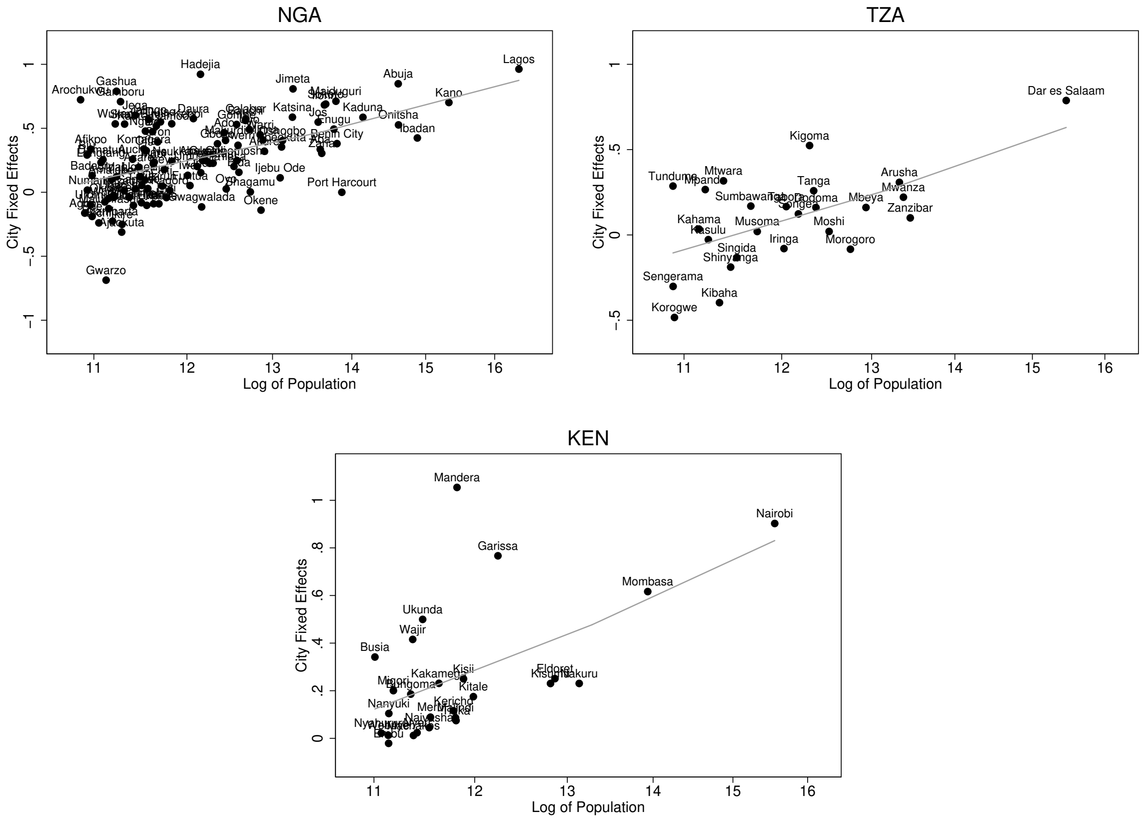
at the origin. Table 9 shows the results from estimating this relationship using all visits in our dataset, where we exclude visits that originate from rural areas. The table shows a clear negative relationship between distance and the fraction of days users spend visiting a city, after controlling for origin and destination fixed effects. The results are very similar when we use travel time instead of distance.[33] 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.

[33]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 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 12 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

Figure 12: Destination fixed effects and city size.



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

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

Table 10 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

Table 10: 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.

significantly more likely to spend a higher fraction of days in larger settlements. This underlines the magnetic forces large cities play. This also highlights that as as cities in Africa are expected to grow, their infrastructure will have to not only accommodate their resident populations, but also the large numbers of people from all over the country who are attracted by the amenities these locations offer.

7. Conclusion

We return finally to the questions that motivated this research. Previous literature has found evidence that spatial and sectoral gaps in many developing countries are large (e.g., Gollin et al. (2014, 2021)). If gaps are large, one might expect to see rapid migration of people across locations – particularly from rural to urban areas, with corresponding shifts from agriculture to other sectors. However, micro data do not seem to show rural-urban migration at the rates that might be expected from measured differences in wages, productivity, or living standards. One possible explanation is that gaps are overstated, as argued in Hamory, Kleemans, Li, and Miguel (2021b).

This paper points to another possible explanation. We find evidence of substantial movements of people across locations within countries, suggesting that people may be able to break down sectoral gaps without permanent migration, or even seasonal migration. Trips between rural and urban locations (or between smaller cities and larger ones) allow people to benefit from the amenities of large cities without moving to them. 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 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.

In this paper, we use a new data source – pings from smartphone apps – that provides objectively measured observations on the frequency with which people move around within three African countries, and on the locations that they visit. The data reveal a strikingly high degree of mobility – albeit for a non-representative subset of the population. Our smartphone users travel frequently and relatively far. Travel is not limited to peri-urban commuting, nor to city-to-city interchanges. On the contrary, we see substantial flows of people between rural areas and big cities. The data point to several important findings.

One is that cities in these countries are providing benefits to many people beyond their own residents. The largest cities, in particular, provide benefits that extend far beyond their geographic boundaries, acting as magnets for visits from the entire country. Although our data do not allow us to specify the nature of these benefits, the evidence suggests that they are complex and multidimensional. People do not visit these cities only to visit government offices or the central markets; instead, we observe them at points scattered throughout the destination cities.

A second key finding from our research is that 'visiting' locations appears to offer an affordable and desirable alternative to migration, for many people. The frequency with which people leave home to visit other locations, often far away, suggests that visiting cities provides travellers with substantial benefits, without the fixed costs and dislocations involved in permanent migration (or even temporary migration).

A third implication is that, at least in these countries, within-country migration seems not to be greatly constrained by the direct costs of travel – i.e., the distance-related costs of moving from one location to another. We observe many people travelling far from home. People make multiple trips over the course of the year when we observe them. Interestingly, some people travel to the same destination multiple times during the period of observation. This is difficult to square with any model in which spatial and sectoral gaps are sustained by the direct costs of travel.

Moreover, the frequency and ubiquity of travel poses a challenge to explanations in which gaps are sustained by information frictions. It is of course possible that those individuals who travel between locations do not report accurate information to acquaintances in their home locations. But with sufficiently large and frequent flows of people, it becomes harder to believe that information frictions can play a quantitatively important role in sustaining spatial gaps.

Our research thus argues against the idea 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, with the largest cities serving as particular magnets of economic activity. Using data on observed trips we estimate high returns to visiting cities.

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. But our paper can be viewed as an illustration of the potential for using such data to address deeper questions about a range of questions 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. We cannot entirely overcome the selection issues that make our sample unrepresentative. Although we filter out many 'transit pings,' we cannot fully determine which places people visit deliberately.[34] But we benefit from the large number of observations and the large number of users. And our efforts to characterize smartphone users shows persuasively that these individuals are not outliers in their own countries, even if they are in some degree unrepresentative.

In spite of these limitations, our analysis offers insights that can help guide future work on spatial frictions and their relevance for development. The findings of this paper encourage us to think critically and carefully before invoking models in which human mobility costs are prohibitive (i.e., models in which people are unable to move across locations and restricted to consume in their home location). The data also suggest that we need to be careful in using models in which information does not flow across locations; even though we observe only a fraction of the total mobility in our three countries, the movements of people seem sufficient to spread information across space. We need to look elsewhere, perhaps, for frictions that seem more capable of explaining persistent sectoral and spatial gaps. For instance, movements from rural to urban areas may involve the loss of social connection or informal insurance, or the loss of claims to land and other resources in rural areas. There may be barriers for rural people – particularly those who are older – in learning new kinds of work or new ways of life. Certainly there is no shortage of potential frictions to consider.

We should also bear in mind that, for the poorest people in our three countries, the story may be very different. Our data shed little light on the mobility of the poor within these countries, and there is much work still to be done to understand the costs of mobility frictions for the poor. Even small monetary costs of mobility can be highly salient for poor people, and information may not flow freely across barriers of ethnicity, social class, age, and other dividing lines. We also have no clear ability to extrapolate 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 –

[34]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.

especially in areas where geography and social strictures make mobility much more challenging.

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.

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.

AKRAM, A. A., S. CHOWDHURY, AND A. M. MOBARAK (2017): "Effects of Emigration on Rural Labor Markets," Working Paper 23929, National Bureau of Economic Research.

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. A. FERGUSON, M. GENTZKOW, AND T. SCHMIDT (2020): "Experienced Segregation," Working Paper 27572, National Bureau of Economic Research.

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. (2012): "Inferring Patterns of Internal Migration from Mobile Phone Call Records: Evidence from Rwanda," *Information Technology for Development*, 18, 107–125.

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.

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.

Federal Republic of Nigeria (2021): https://grid3.gov.ng, accessed September 2021.

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.

Hamory, J., M. Kleemans, N. Y. Li, and E. Miguel (2021a): "Reevaluating Agricultural Productivity Gaps with Longitudinal Microdata," *Journal of the European Economic Association*, 19, 1522–1555.

——— (2021b): "Reevaluating Agricultural Productivity Gaps with Longitudinal Microdata," *Journal of the European Economic Association*, 19, 1522–1555.

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

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.

Lu, X., D. J. Wrathall, P. R. Sundsøy, M. Nadiruzzaman, E. Wetter, A. Iqbal, T. Qureshi, A. Tatem, G. Canright, K. Engø-Monsen, et al. (2016): "Unveiling Hidden Migration and Mobility Patterns in Climate Stressed Regions: A Longitudinal Study of Six Million Anonymous Mobile Phone Users in Bangladesh," *Global Environmental Change*, 38, 1–7.

Miyauchi, Y., K. Nakajima, and S. Redding (2021): "Consumption Access and Agglomeration: Evidence from Smartphone Data," Unpublished Manuscript.

Mongey, S., L. Pilossoph, and A. Weinberg (2020): "Which Workers Bear the Burden of Social Distancing Policies?" Working Paper 27085, National Bureau of Economic Research.

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.

Young, A. (2013): "Inequality, the Urban-Rural Gap, and Migration," *Quarterly Journal of Economics*, 128, 1727–1785.

Appendix (For Online Publication)

A. Details on data smartphone app data

A.1. Construction of the base sample

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[35]. We entirely remove users with home location coordinates at the data sink from the sample.

A.2. Inferred 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. Unsurprisingly, the sample size decreases with the minimum number of observed nights imposed and nearly doubles between the high- and low-confidence subsets.

[35] See footnote 18 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.

However, differences in the distributions of users across density deciles between the subsets is only minimal, as shown in Figures C.9 to C.11 , which supports the idea that the minimum number of nights criterion used to build our core sample does not imply a selection of atypical users.

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

A.3. Inferred work locations

Table A.2: Work subset sample size.

	High-conf.		**weekdays>0**		**Work subset**	
	users	%	users	%	users	%
Kenya	18,545	100	18,359	99	13,053	70.4
Nigeria	78,750	100	78,211	99.3	58,452	74.2
Tanzania	22,994	100	22,808	99.2	16,415	71.4
TOTAL	120,289	100	119,378	99.2	87,920	73.1

Note: This table shows the base sample of high-confidence users in the first two columns. The next two columns show the fraction of users seen at least on one weekday. The final two columns show the fraction of high-confidence users that meet the criteria to identify work locations confidently.

Table A.3: Home and work locations.

	Identical home and work locations			Home-to-work median distance (in km)		
	Overall	**Urban**	**Non-urban**	**Overall**	**Urban**	**Non-urban**
Kenya	79.5%	78%	87%	4.7	5	2.5
Nigeria	79.7%	79.1%	84.9%	4.6	4.6	4
Tanzania	82.7%	80.9%	87.7%	3.3	4	1.6
TOTAL	80.2%	79.2%	86.2%	4.4	4.6	3

Note: Home and work locations are defined at a 2-decimal degree resolution. The subset is restricted to users seen on at least 10 distinct week days.The top panel shows the fraction of users with identical home and work location. The bottom panel shows that median distance between work and home locations for users with distinct work and home locations.

Table A.4: Home and work locations.

	Identical home and work locations			Home-to-work median distance (in km)		
	Overall	**Urban**	**Non-urban**	**Overall**	**Urban**	**Non-urban**
Kenya	80.7%	79.4%	87.5%	3.6	3.8	1.3
Nigeria	80%	79.4%	85%	3.1	3.2	1.8
Tanzania	81.8%	80.6%	85.4%	2	2.4	0.8
TOTAL	80.5%	79.6%	85.6%	3	3.2	1.4

Note: Home and work locations are defined at a 3-decimal degree resolution. The subset is restricted to users seen on at least 10 distinct week days. The top panel shows the fraction of users with identical home and work location. The bottom panel shows that median distance between work and home locations for users with distinct work and home locations.

A.4. Identifying 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) 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} [36]. 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

[36] 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.

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.[37] 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 [38] and (ii) at least 25% of speed values are greater than 20km/h.[39] Visits that are neither stays nor transits are classified as undefined.

Table A.5 applies this algorithm to the three countries.

Table A.5: Total and mean (user-level) fractions of transit, stay, and undefined pings.

	Total fraction			Mean fraction		
	Transit (1)	**Stay** (2)	**Undef.** (3)	**Transit** (4)	**Stay** (5)	**Undef.** (6)
Kenya	17%	71%	12%	0.02	0.46	0.52
Nigeria	9%	67%	24%	0.02	0.47	0.51
Tanzania	8%	76%	16%	0.02	0.5	0.48
TOTAL	11%	70%	19%	0.02	0.48	0.5

Note: This table shows the proportion of pings that are transit pings, stay pings or undefined pings averaged across all pings in columns (1)-(3) or all users in columns (4)-(6).

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 (columns (4)-(6) of Table A.5).[40] 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

[37]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}.

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

[39]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.

[40]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.

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. Distinct visits to cities

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

First, we detect sequences of consecutive pings at a visited city. In this first step, we use a rather conservative criterion and, for any given user i, we allow for a maximum inter-ping time ϵ^i_{visit} that corresponds to a return trip in straight line between the considered ping and the home location at a constant speed of 40 km/h. We introduce "home flags" that indicate when a user was observed back to her home location between two consecutive sequences of pings at a visited 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 a same visited city according to a set of well-defined rules: (i) consecutive sequences of pings at a same visited city within a single day are grouped to form a unique visit[42], (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

[41] See footnote 18. 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.

[42] 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

less than 12 hours apart[43], (iii) if the travel time between the visited city centroid and the home location is strictly beyond 2 hours, we group together sequences of pings that are less than 36 hours apart. With criterion (i), we allow for the possibility of commuters being observed early in the morning and late in the afternoon in their destination city. This is also relevant for visits to the closest cities where $\epsilon^i_{visit,c}$ is small and potentially leads to separate sequences of pings to a visited city on a given day when those are most likely part of the same visit. Criterion (ii) basically allows for users to spend a night in a nearby city and therefore be unobserved for that period of time. For instance, a sequence of pings in Nairobi ending at 9pm one night followed by another starting at 7am the day after from a user residing in Thika (approximately a 1h drive) will be considered as a single visit to Nairobi. Similarly, criterion (iii) allows for two nights away to more distant cities without being observed, i.e. it is sufficient to see the user at the visited city on one night and in the morning two days after to consider that we are observing the same visit.

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

[43] 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.

B. Pairing users with DHS information

In Section 3, we link users' home locations with data from the most recently available Demographic and Health Survey (DHS) data to characterize areas where our users live: the 2014 standard DHS in Kenya, the 2018 standard DHS in Nigeria and the 2015-2016 standard DHS in Tanzania.[44] DHS data are geo-referenced at the cluster level and cluster coordinates are randomly displaced to maintain respondents' confidentiality. 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.

We first classify our users within urban and rural categories based on the overlay of users' home location with city polygons[45]. 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 (10km for urban users and 5km for rural users). This yields a set of DHS clusters that are comparable, in some sense, to the home location of our user. The number of these comparison clusters will be either zero or a strictly positive number of clusters. Not all these nearby clusters will offer valid comparisons, however. For example, a user at the outskirts of Dar Es Salaam might be associated with a nearby rural cluster as well as a number of urban clusters. To ensure that we do not falsely assign an urban cluster as a comparison location for a rural user (or vice-versa), we add the second criterion that the cluster's average population density (calculated over a 5km buffer) must be within 25% of the average population density that we have computed for the user's home location. If this does not hold, we drop the DHS comparison cluster.

Following that methodology, we pair 70% of our users in the high-confidence sample with at least one DHS cluster (90% in Kenya, 66% in Nigeria, 72% in Tanzania). We call the subset of respondents within paired clusters the "matched DHS" sample.[46] 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 including housing quality (floor, walls, roof, over-crowding) and access to public services (electricity, water). Moreover, we produce results for rural and urban sub-samples separately to account for both the prevalence of urban

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

[45] See footnote 18 for details on the definition of city boundaries.

[46] 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 begin equal to the number of users the corresponding cluster is matched to.

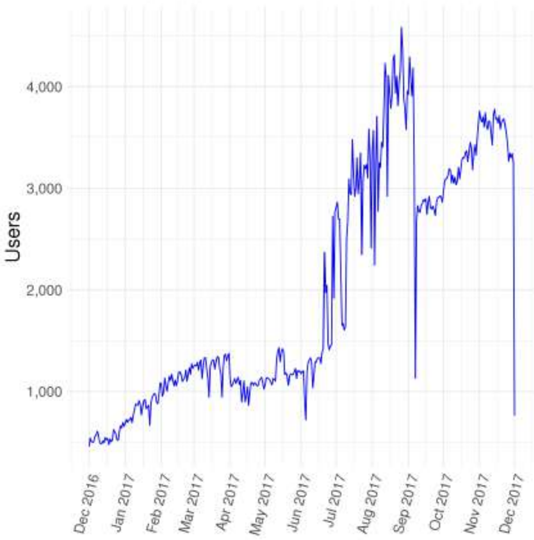
users in our sample and the lower matching rate in low density areas, which together may lead to results being mainly driven by the urban component of the sample. We produce t-tests comparing our two weighted data streams, with bootstrapped standard errors robust to heteroskedasticity. The survey weights 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. Definitions for all variables used are given in Table B.1.

Table B.1: Definitions of variables used in mean testing between DHS and matched DHS samples.

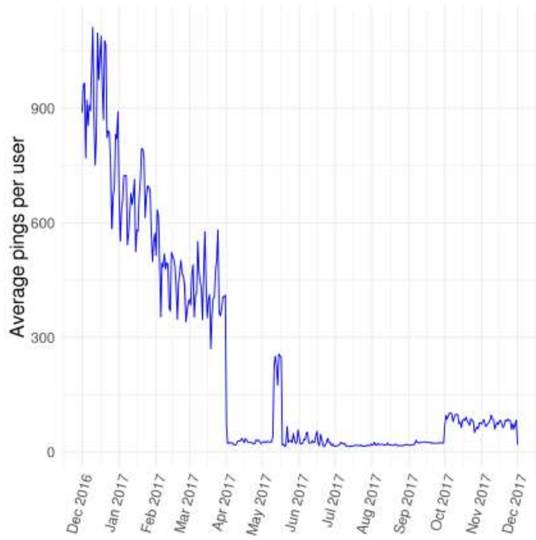
Variable	Definition	Notes
Household size	Total number of (*de jure*) household members	
Age of HH head	Age of the household head	
Education of HH head	Level of education of the household head in single years	
Access to electricity	Dummy equal to 1 if the household has access to electricity	
Radio	Dummy equal to 1 if the household has a radio	
Television	Dummy equal to 1 if the household has a television	
Rooms per adult	Number of rooms used for sleeping divided by the number of household members older than 6	
Access to piped water	Dummy equal to 1 if the household has an access to piped water	Calculated based on the main source of drinking water for household members. Our dummy is equal to 1 if the source of water is one of "piped water", "piped into dwelling", "piped to yard/plot", "public tap/standpipe" or "piped to neighbor".
Constructed floor	Dummy equal to 1 if main material of the floor is not one of "natural", "earth/sand", "dung", "rudimentary", "wood planks", "palm/bamboo"	
Constructed walls	Dummy equal to 1 if main material of the walls is not one of "natural", "no walls", "cane/palm/trunks", "cane/palm/trunks/bamboo", "poles with mud", "dirt", "rudimentary", "dung/mud/sod", "grass", "bamboo with mud"	
Constructed roof	Dummy equal to 1 if main material of the roof is not one of "natural", "grass/thatch/palm leaf", "no roof", "thatch/grass/makuti", "mud", "rudimentary", "rustic mat", "dung/mud/sod", "palm/bamboo", "wood planks", "cardboard"	

C. Additional tables and figures

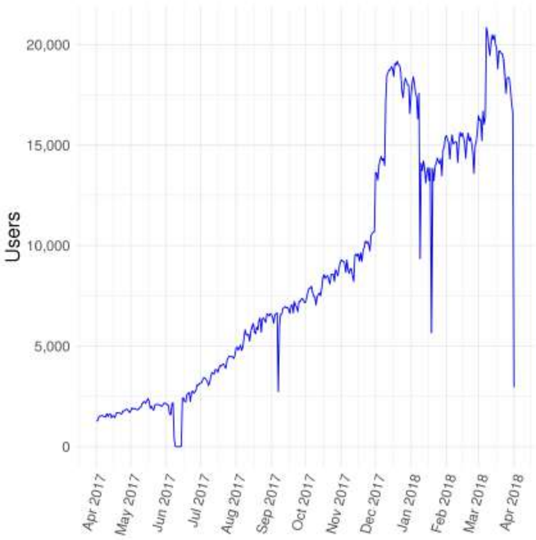
Figure C.1: Users and pings per user over time.



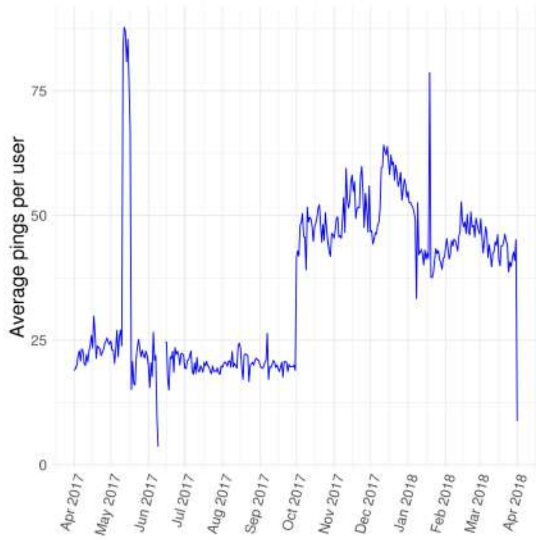
(a) Users - Kenya



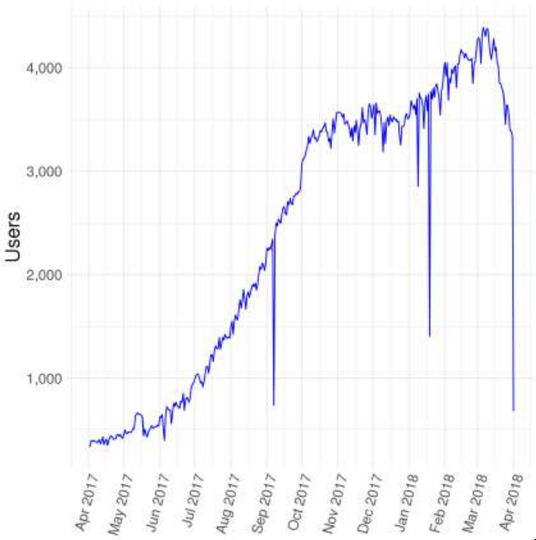
(b) Pings per users - Kenya



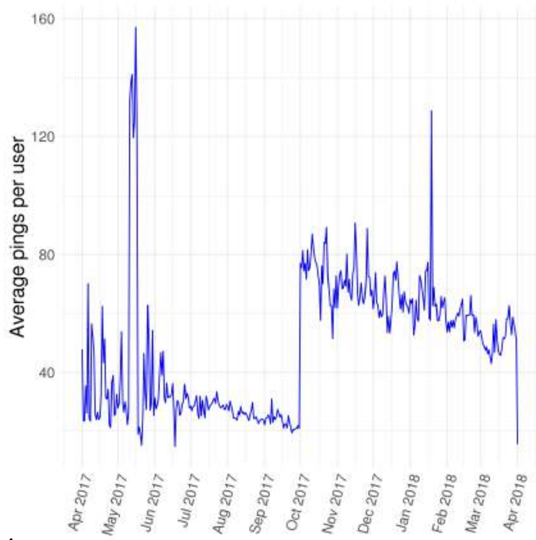
(c) Users - Nigeria



(d) Pings per users - Nigeria

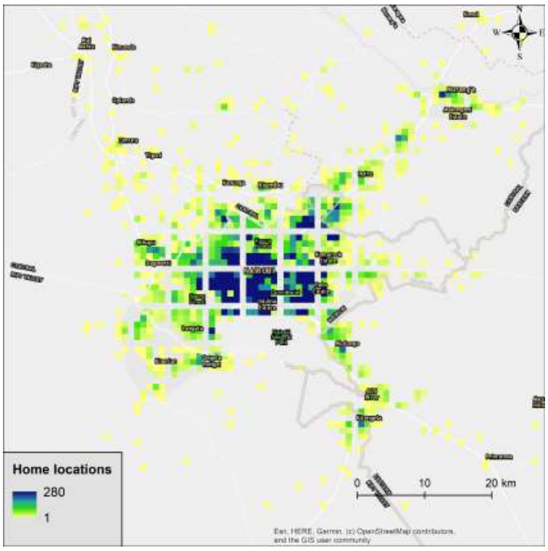


(e) Users - Tanzania

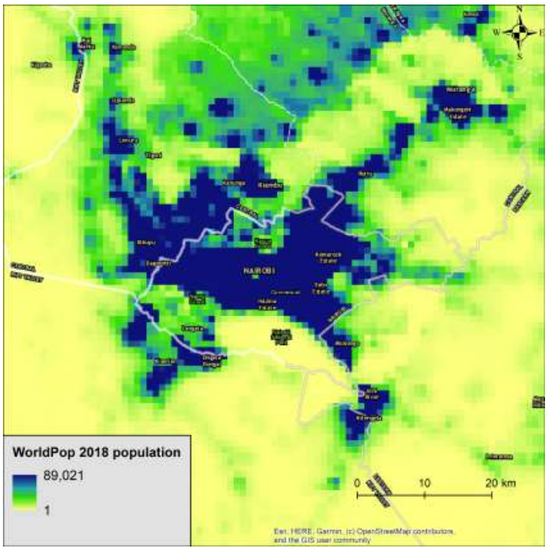


(f) Pings per users - Tanzania

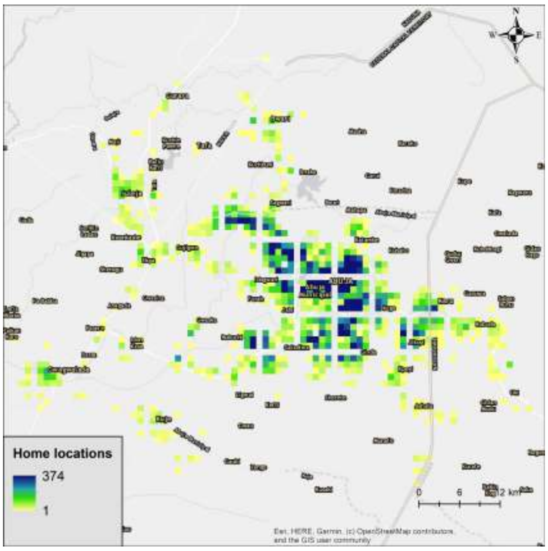
Figure C.2: Distribution of home locations in capital cities, high-confidence sample.



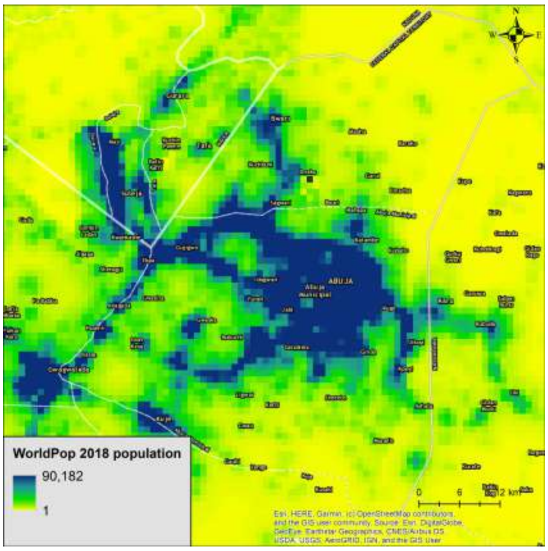
(a) Nairobi



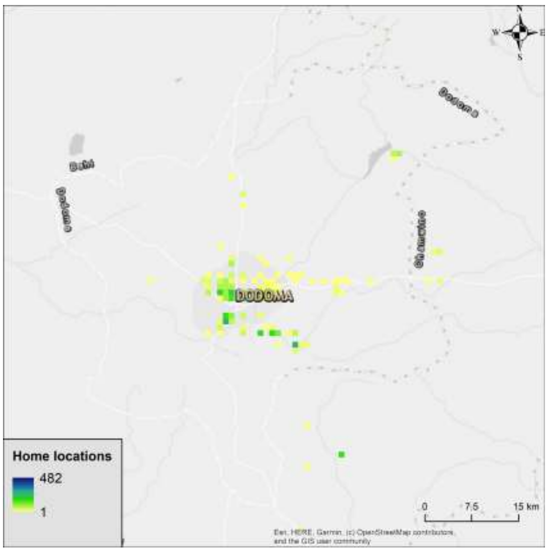
(b) Nairobi



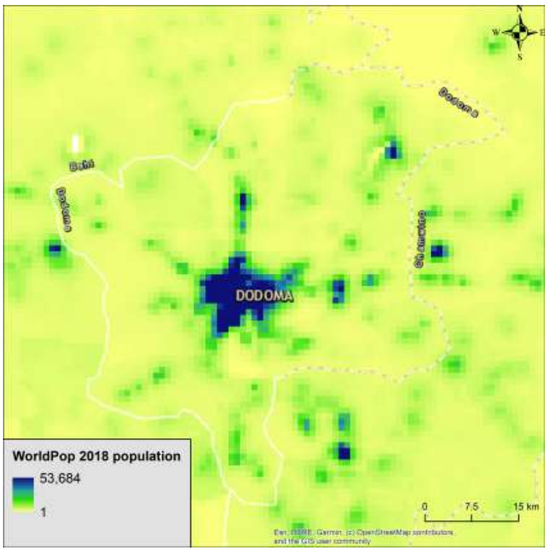
(c) Abuja



(d) Abuja

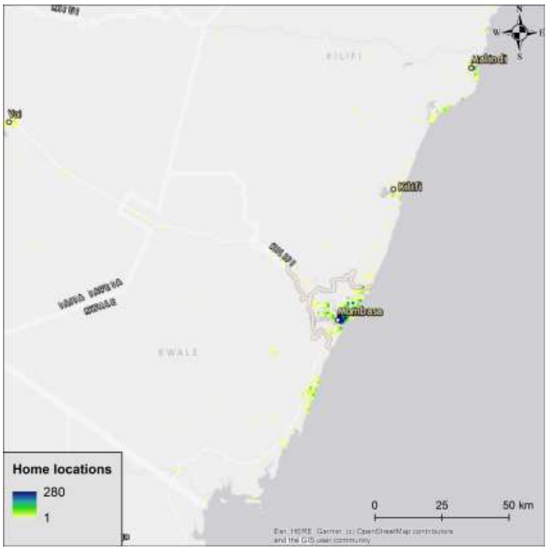


(e) Dodoma

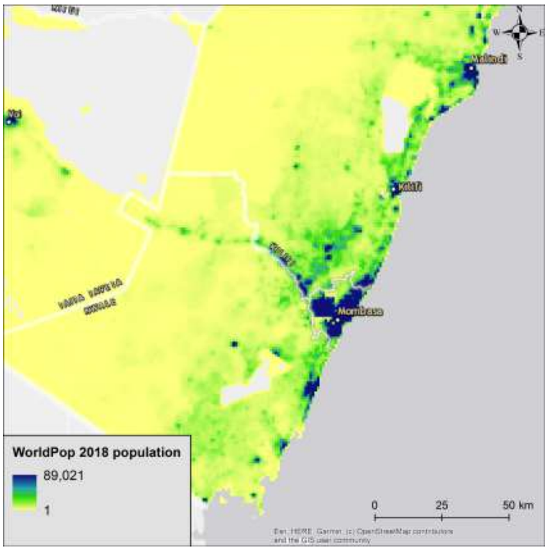


(f) Dodoma

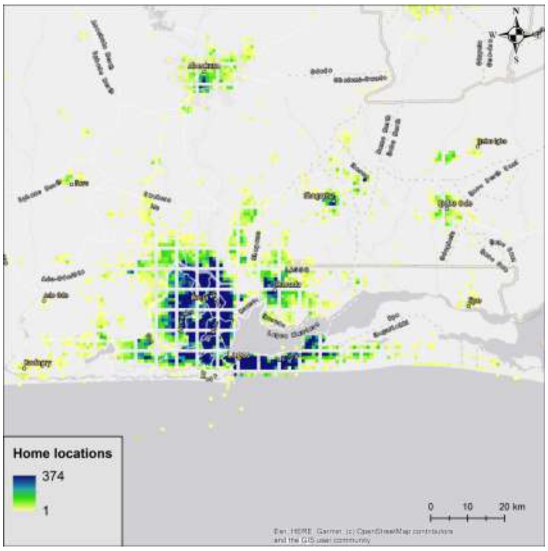
Figure C.3: Distribution of home locations in major cities, high-confidence sample.



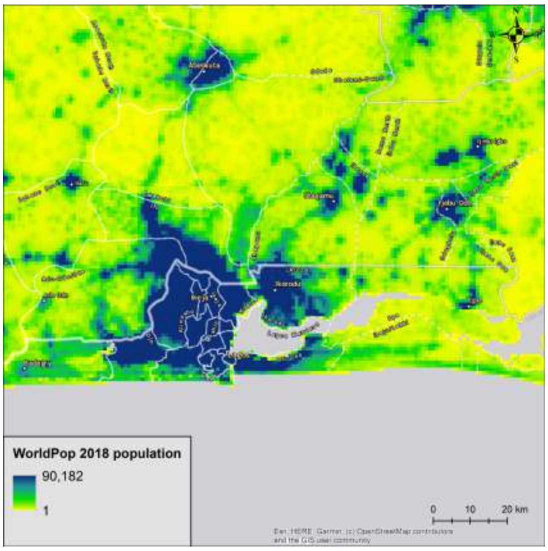
(a) Mombasa



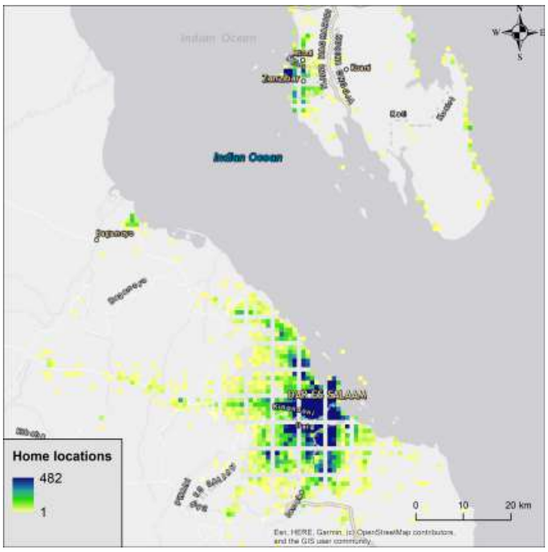
(b) Mombasa



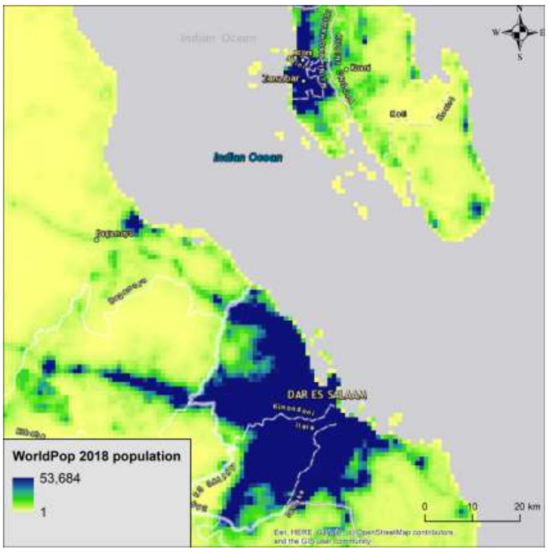
(c) Lagos



(d) Lagos

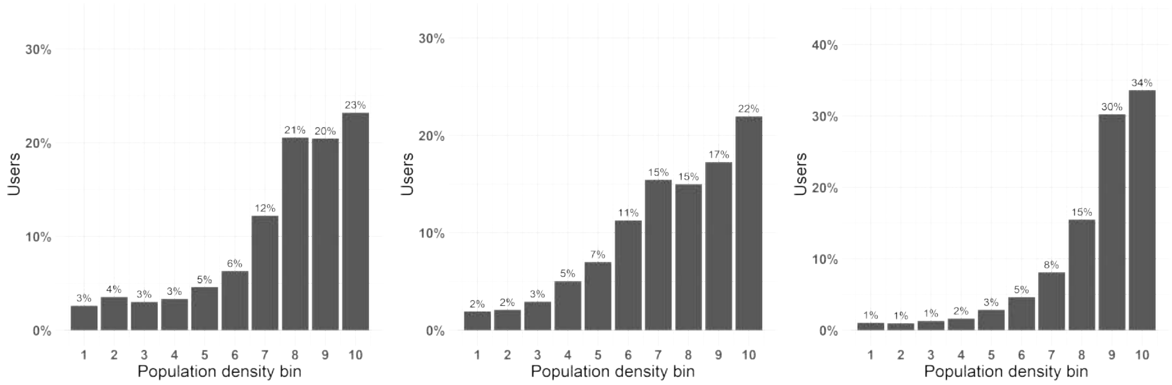


(e) Dar Es Salaam



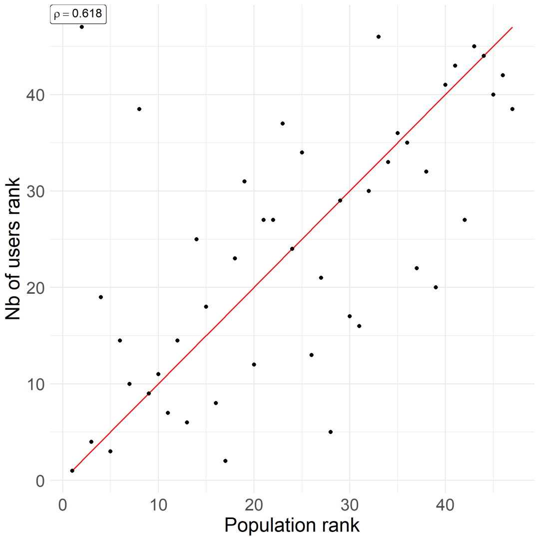
(f) Dar Es Salaam

Figure C.4: Users by population density decile, Landscan.

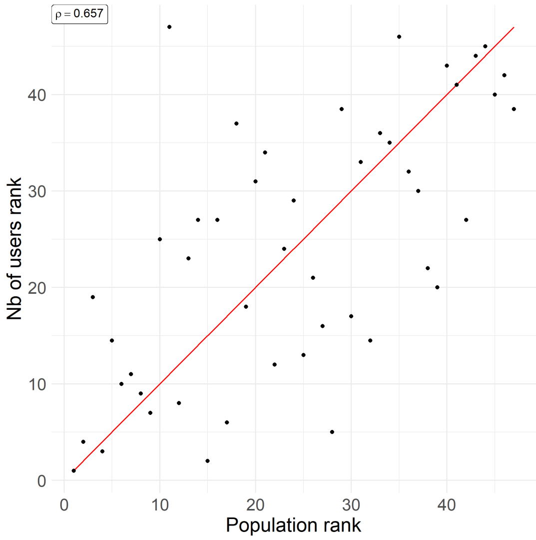


(a) Kenya (b) Nigeria (c) Tanzania

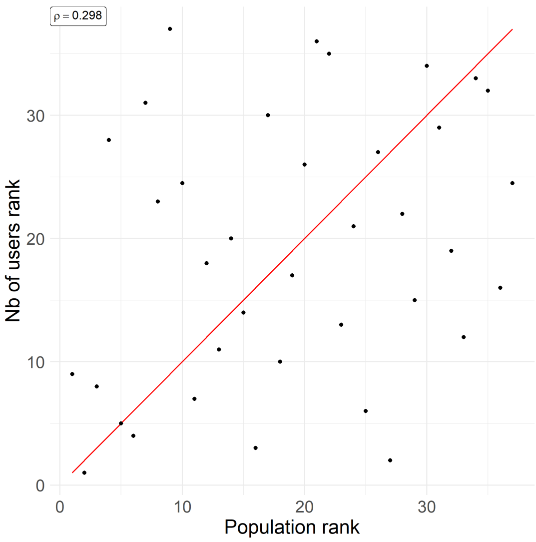
Figure C.5: Comparing user and population ranks at the first administrative level.



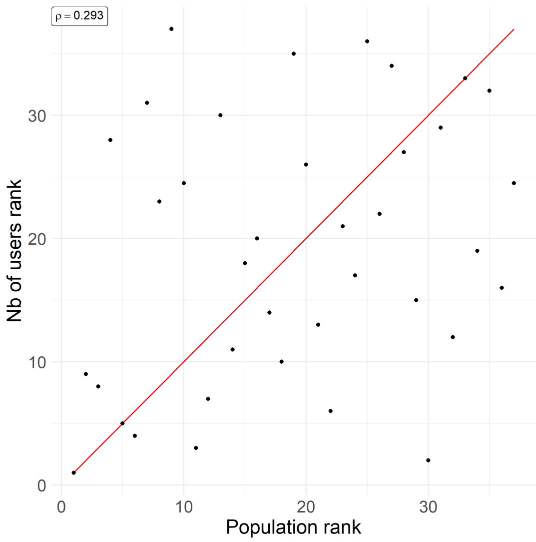
(a) World Pop, Kenya



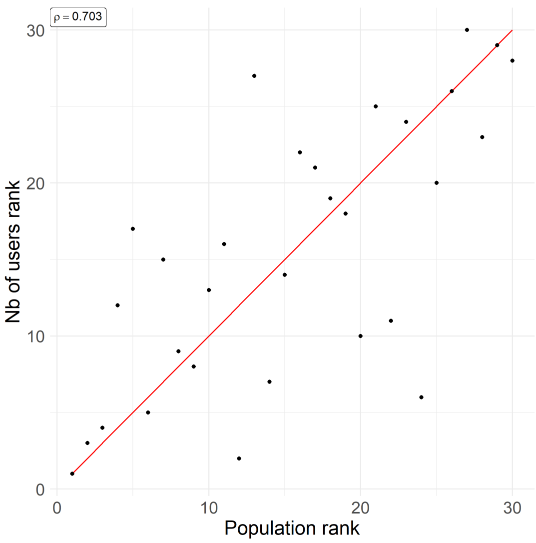
(b) Lanscan, Kenya



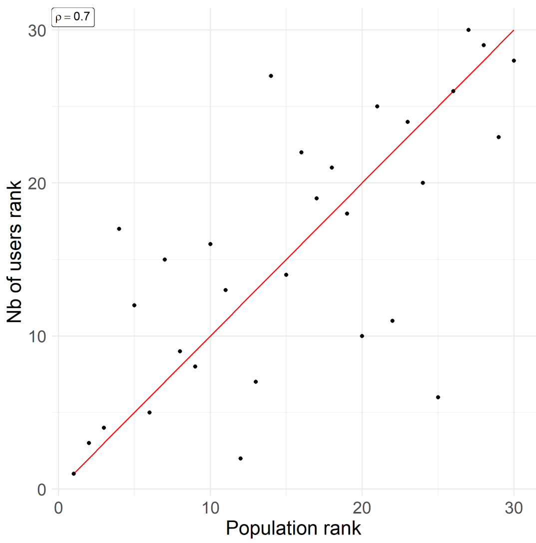
(c) World Pop, Nigeria



(d) Lanscan, Nigeria



(e) World Pop, Tanzania



(f) Lanscan, Tanzania

Table C.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***

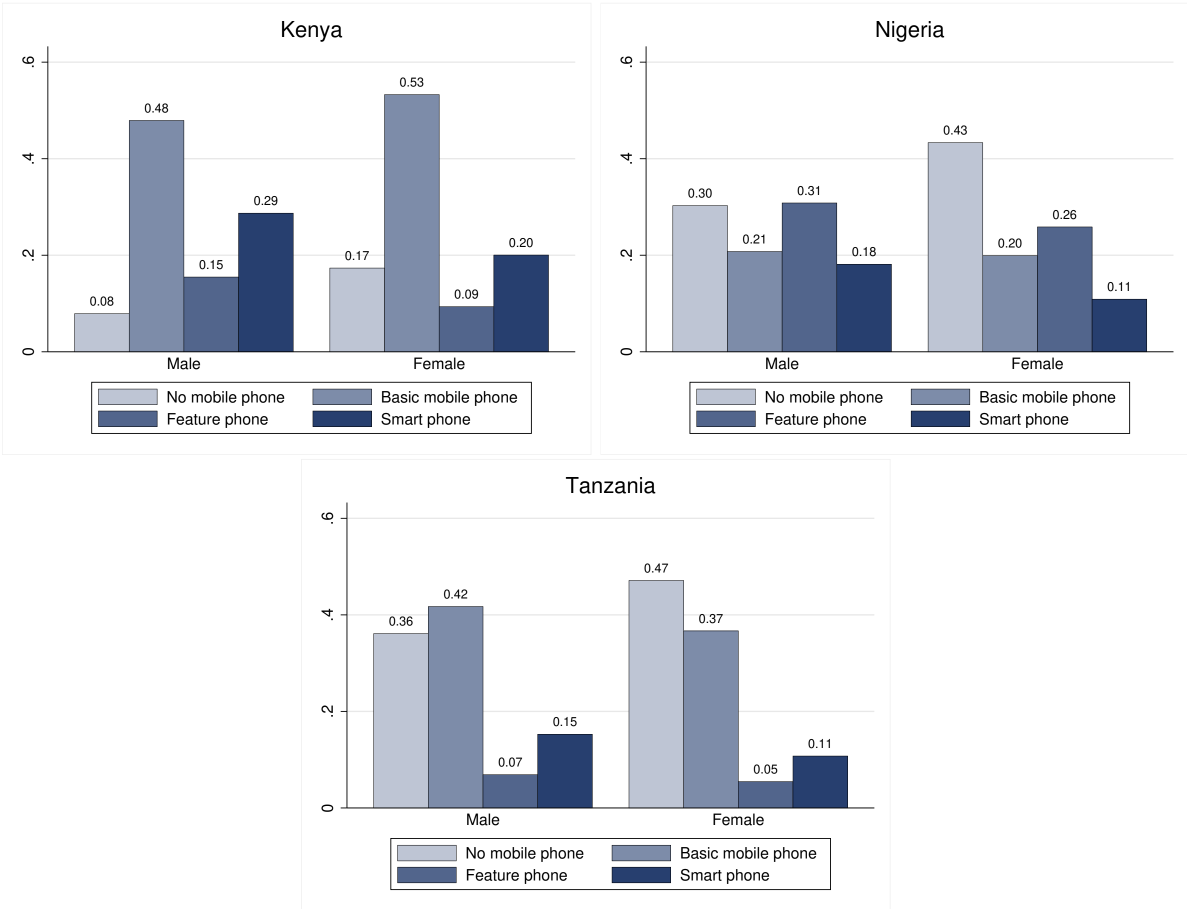
Table C.2: T-tests for equality of means between DHS and matched DHS samples, Nigeria.

	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***

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

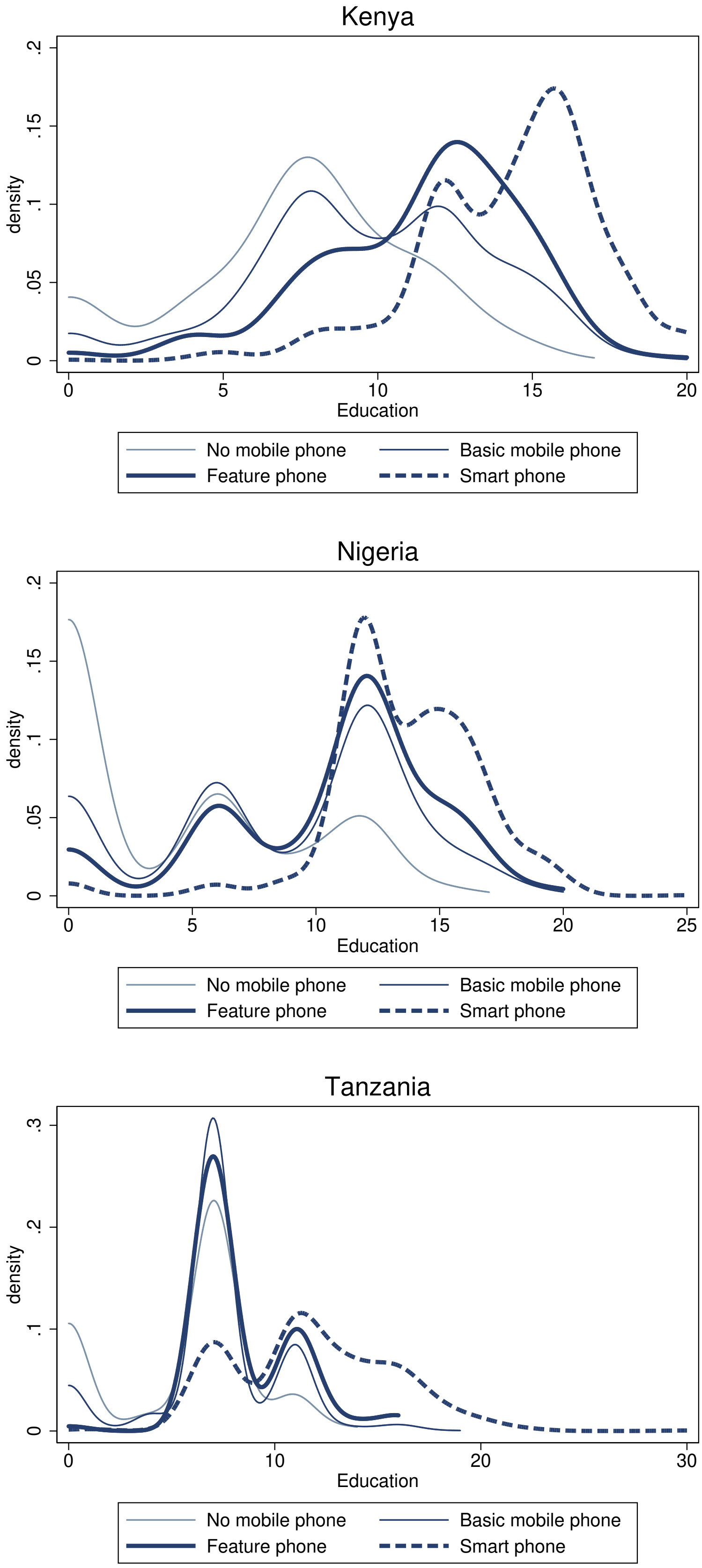
	Variable	DHS	Matched DHS	Difference	SE	p-value
All	Household size	5.03	4.33	-0.70	0.04	0.000***
	Age of HH head	45.43	41.66	-3.77	0.22	0.000***
	Education of HH head	5.90	8.33	2.42	0.05	0.000***
	Access to electricity	0.23	0.78	0.55	0.02	0.000***
	Radio	0.52	0.66	0.14	0.01	0.000***
	Television	0.21	0.65	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***

Figure C.6: Device ownership by gender.



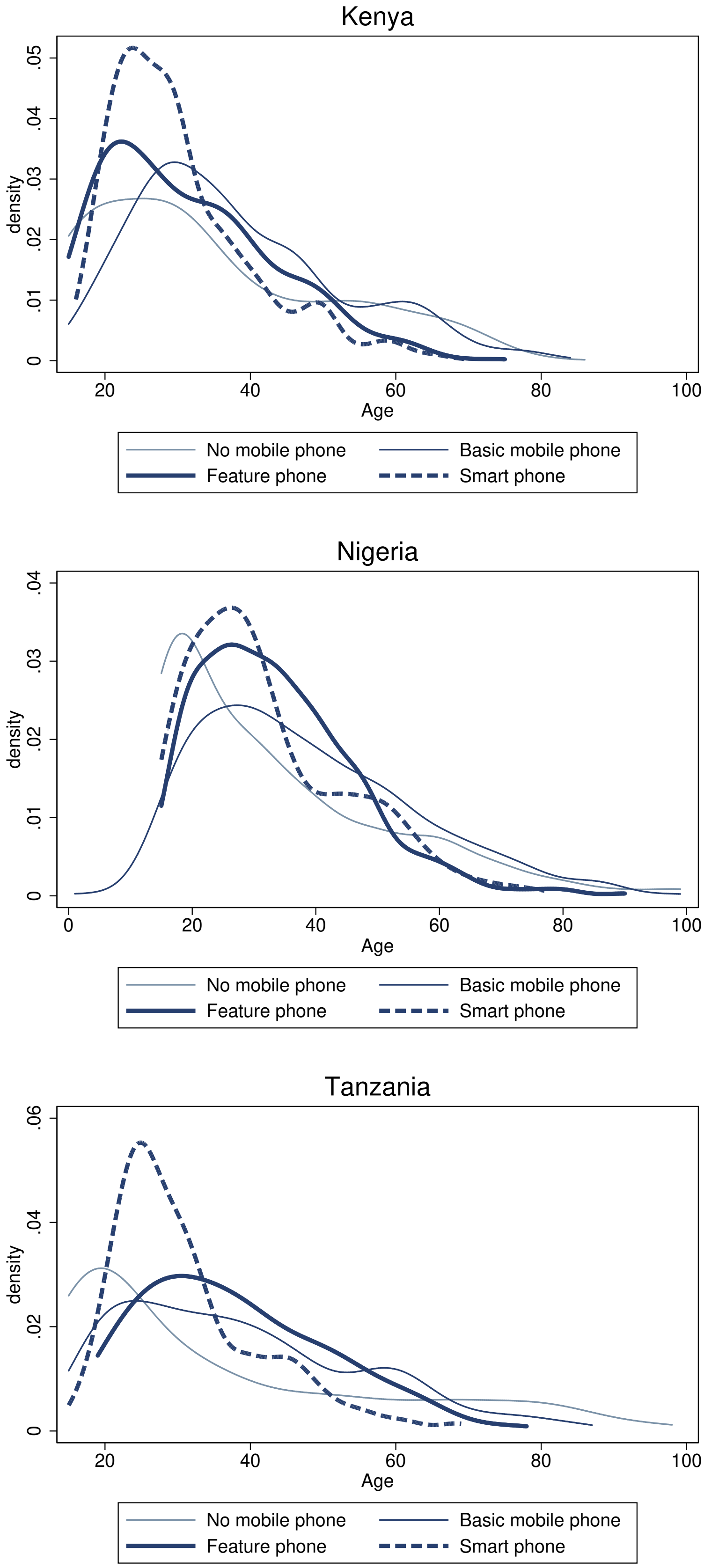
Note: These figures show device ownership rates for female and male respondents. All figures use the sample weights provided.

Figure C.7: Education and device ownership.



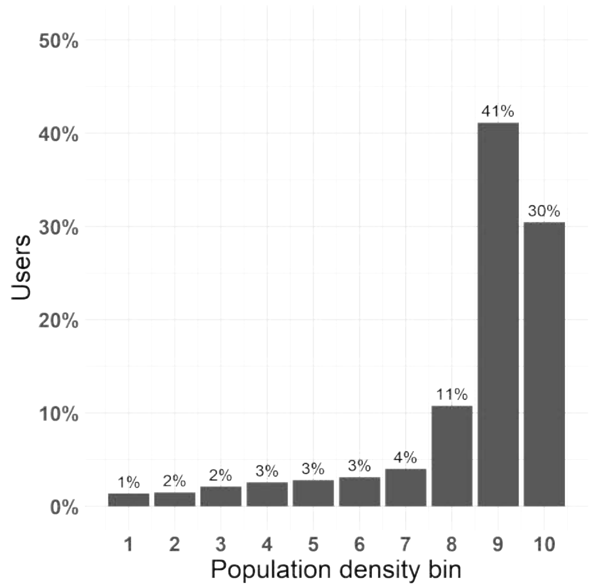
Note: These figures show the distribution of education by device ownership. All figures use the sample weights provided.

Figure C.8: Age and device ownership.

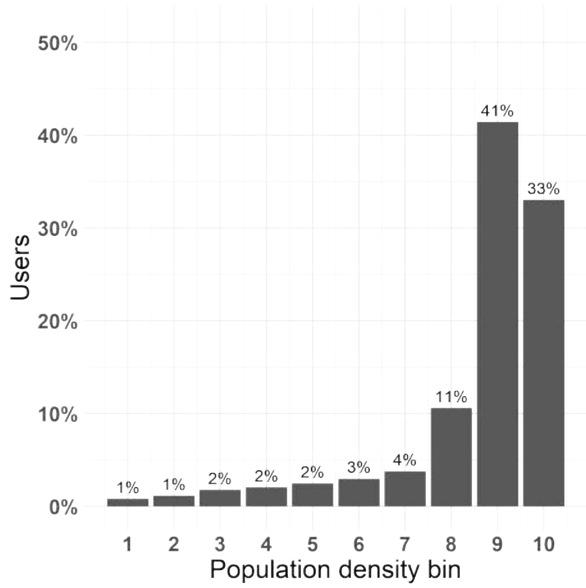


Note: These figures show the distribution of age by device ownership. All figures use the sample weights provided.

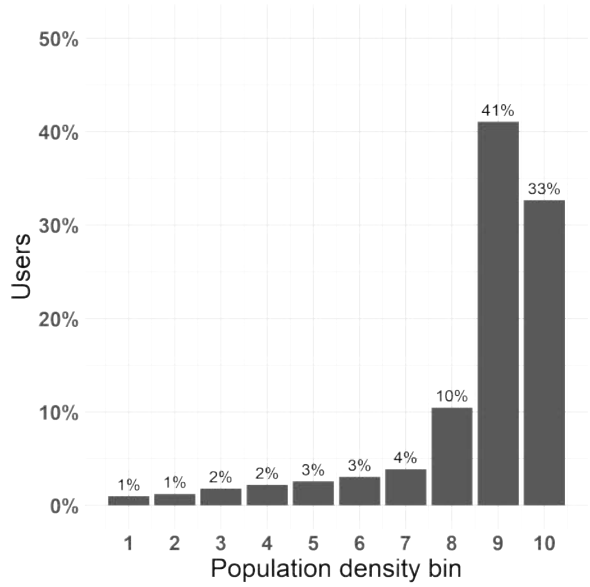
Figure C.9: Fraction of users by population density deciles in Kenya.



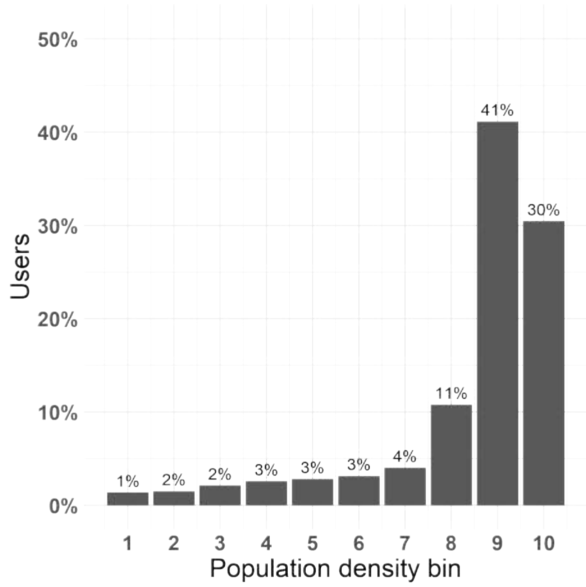
(a) Base sample



(b) High-confidence subset

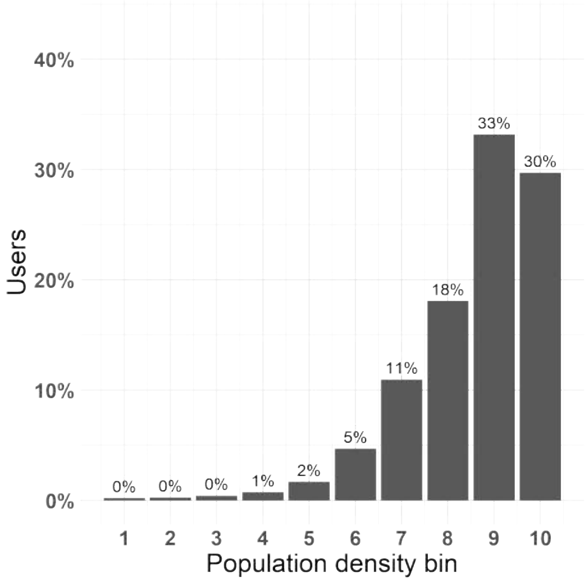


(c) Medium-confidence subset

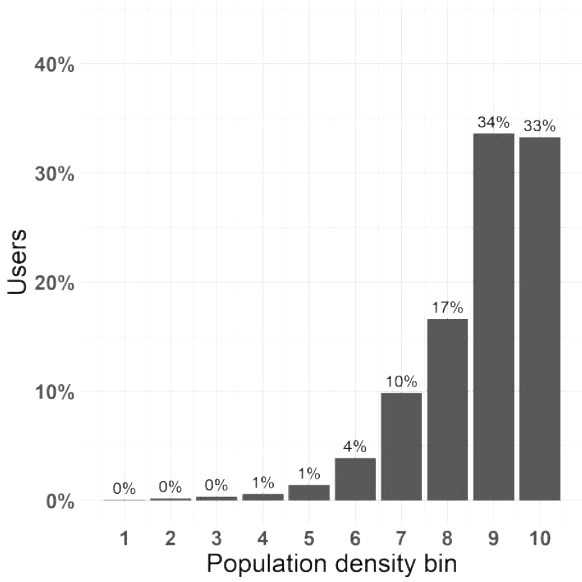


(d) Low-confidence subset

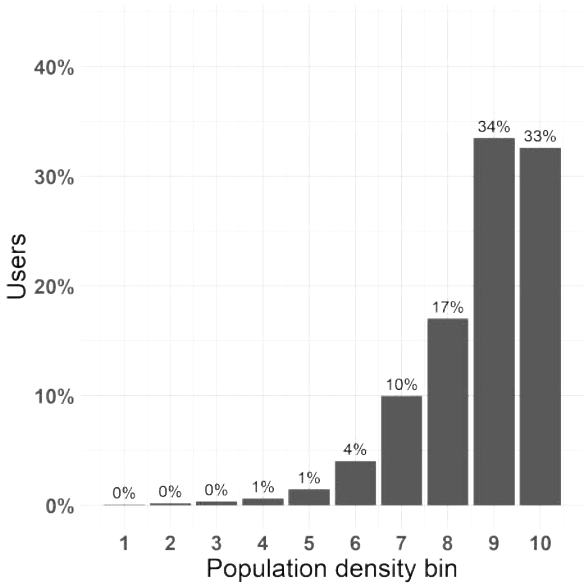
Figure C.10: Fraction of users by population density deciles in Nigeria.



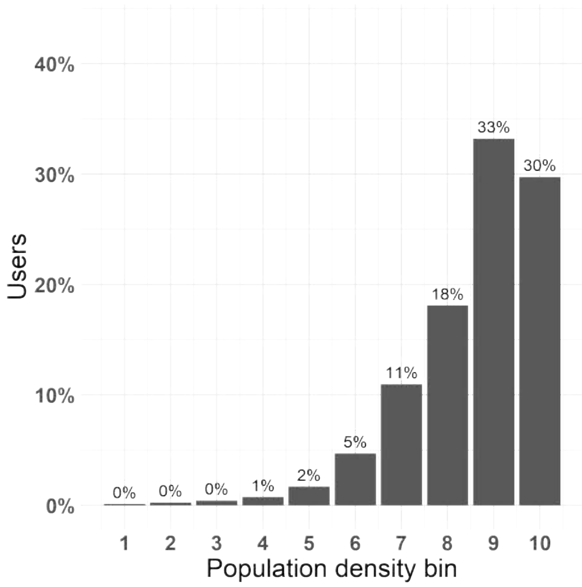
(a) Base sample



(b) High-confidence subset

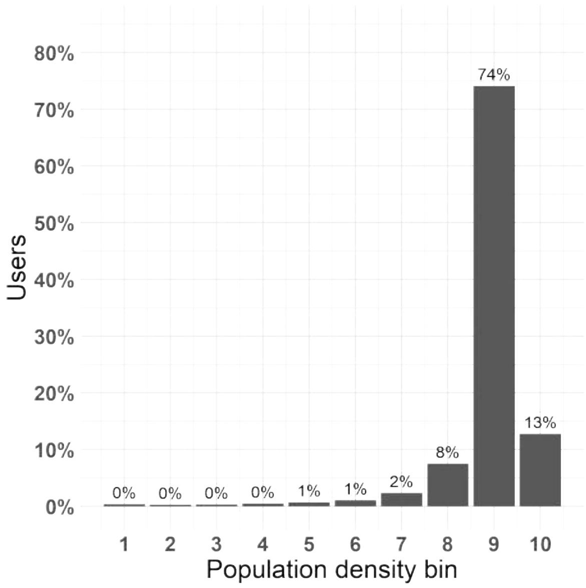


(c) Medium-confidence subset

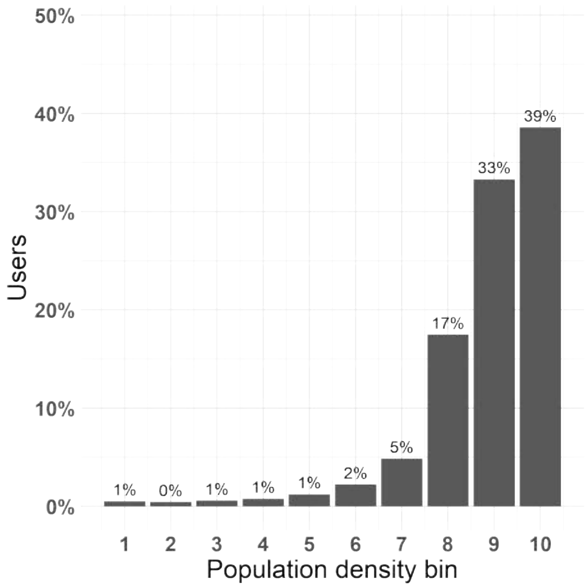


(d) Low-confidence subset

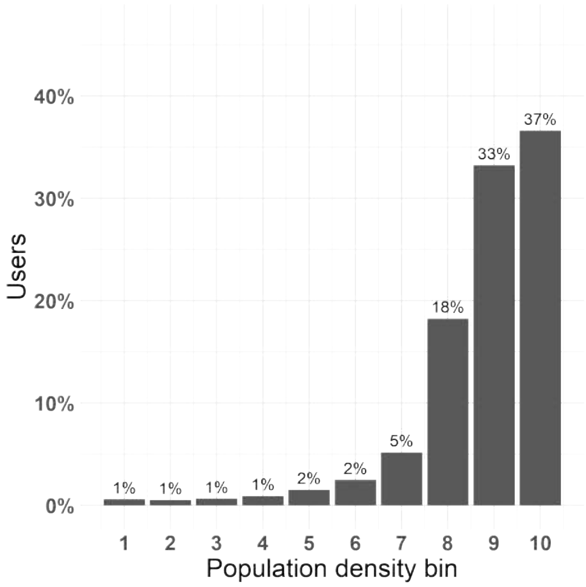
Figure C.11: Fraction of users by population density deciles in Tanzania.



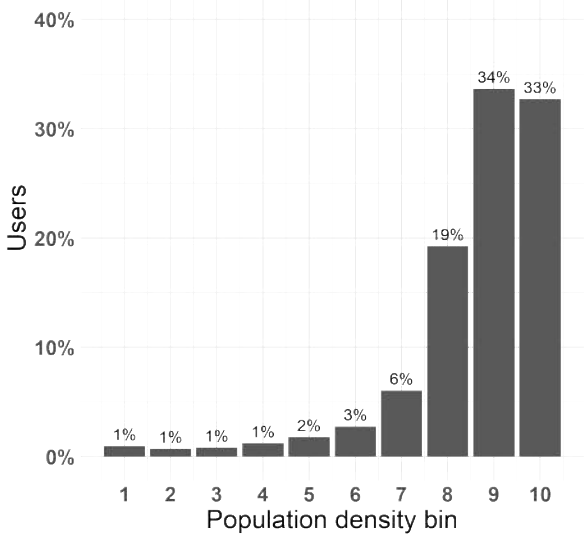
(a) Base sample



(b) High-confidence subset

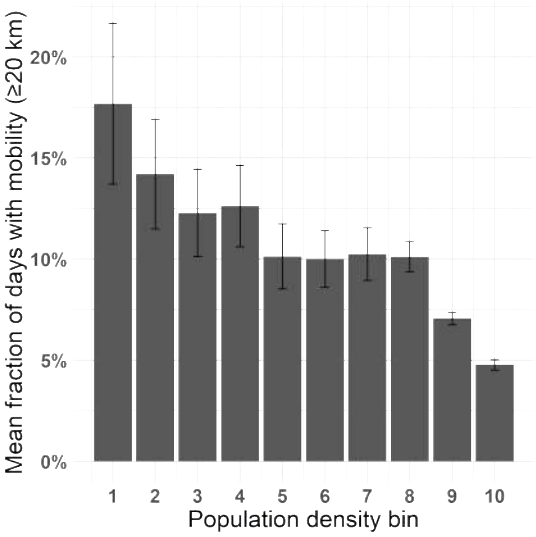


(c) Medium-confidence subset

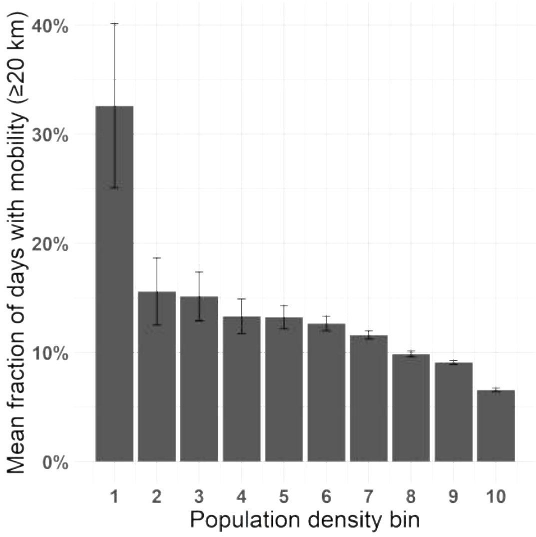


(d) Low-confidence subset

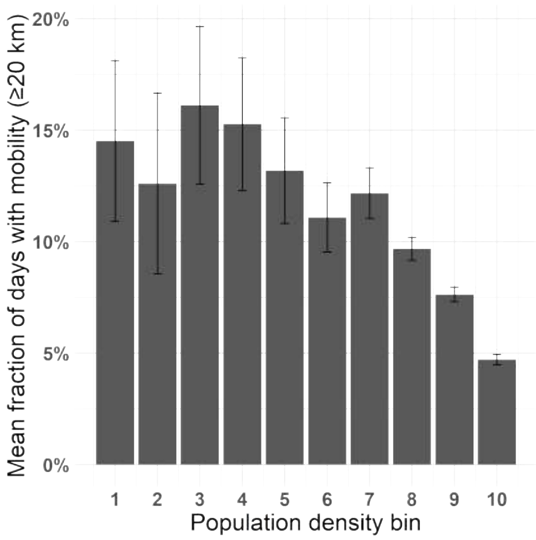
Figure C.12: Fraction of days with mobility beyond 20km by density bin.



(a) Kenya



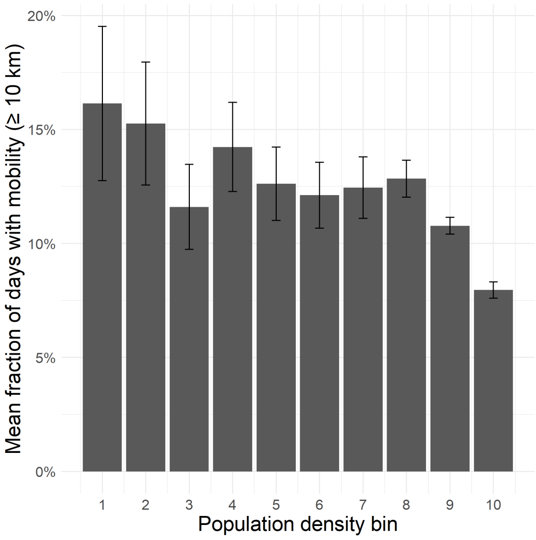
(b) Nigeria



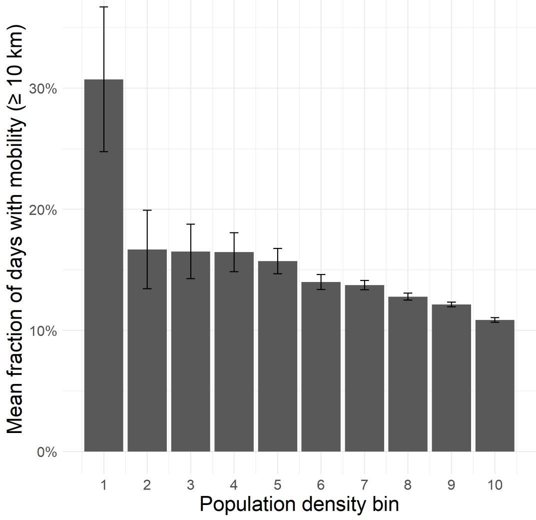
(c) Tanzania

Note: These figures show the fraction of days on which a user is seen more than 20km away from their home location by density bin.

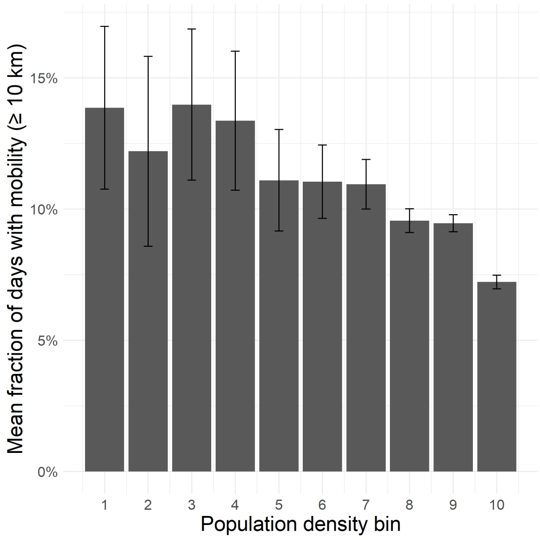
Figure C.13: Fraction of days with mobility beyond 10km by density bin, top 5 percentile trimmed.



(a) Kenya



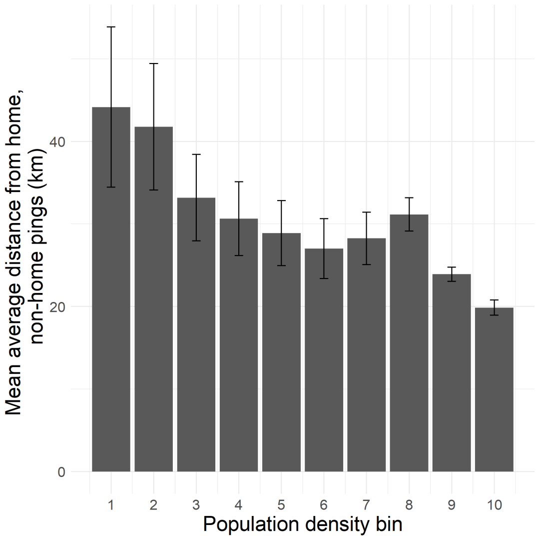
(b) Nigeria



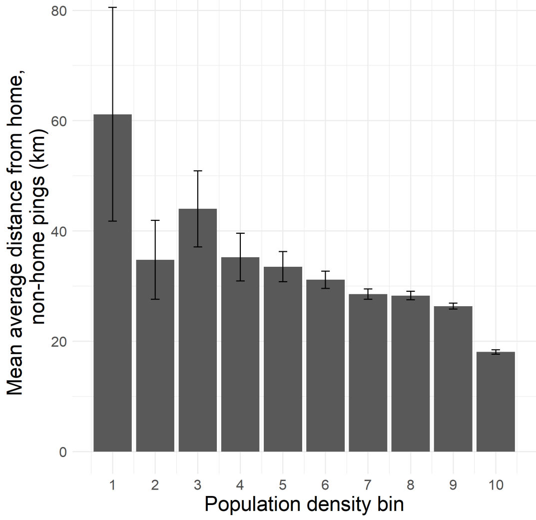
(c) Tanzania

Note: These figures show the fraction of days on which a user is seen more than 10km away from their home location by density bin. For each density bin, values in the top 5 percentile are removed.

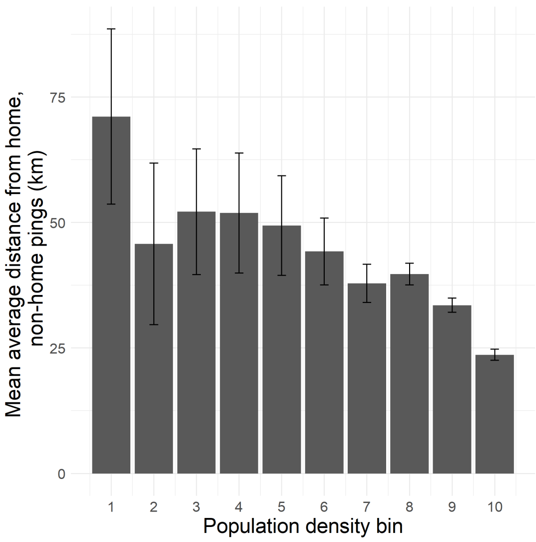
Figure C.14: Mean distance away from home by density bin, top 5 percentile trimmed.



(a) Kenya



(b) Nigeria



(c) Tanzania

Note: These figures show the average across users of the mean distance between home location and non-home pings by density bin. For each density bin, values in the top 5 percentile are removed.

Table C.4: 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

Table C.5: 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
Visited density	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%
	5	6%	4.5%	8.5%	11.2%	12.3%	10.2%	4.9%	4.1%	1.7%	0.9%
	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
Visited density	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%
	5	2.8%	5.1%	8.1%	9.9%	10.8%	5.6%	2.6%	1.4%	1%	0.5%
	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
Visited density	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%
	5	1.7%	6.3%	9.6%	9.4%	9.6%	8.6%	4.2%	1.6%	0.7%	0.3%
	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.

Table C.6: Share of users by home bin-visited bin pair, transit pings excluded.

		Home density bin									
		1	2	3	4	5	6	7	8	9	10
	1	71.2%	32.9%	14%	11.1%	11.4%	13.2%	12.1%	13.9%	8.7%	5.6%
	2	43.2%	60.8%	37.7%	24.9%	18.9%	17.1%	17.2%	19.3%	13.3%	9.6%
	3	25.2%	45.6%	55.1%	41.1%	34.9%	28.4%	26.5%	24.1%	17.6%	13.2%
	4	34.2%	32.9%	51.3%	56.6%	46.2%	37.7%	33.9%	27.5%	22.1%	16.5%
Visited density	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
	1	35.7%	18.8%	18.8%	6.3%	5.7%	3.5%	2.9%	2.7%	2.2%	1.3%
	2	23.8%	31.9%	35%	12.2%	12.1%	8.3%	6.2%	5.5%	4.6%	2.8%
	3	26.2%	29%	40.6%	31.6%	18%	12.5%	9.6%	8%	6.5%	4.2%
	4	31%	26.1%	44.9%	35%	31.7%	21.7%	14.1%	11.3%	10.4%	6.4%
Visited density	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%

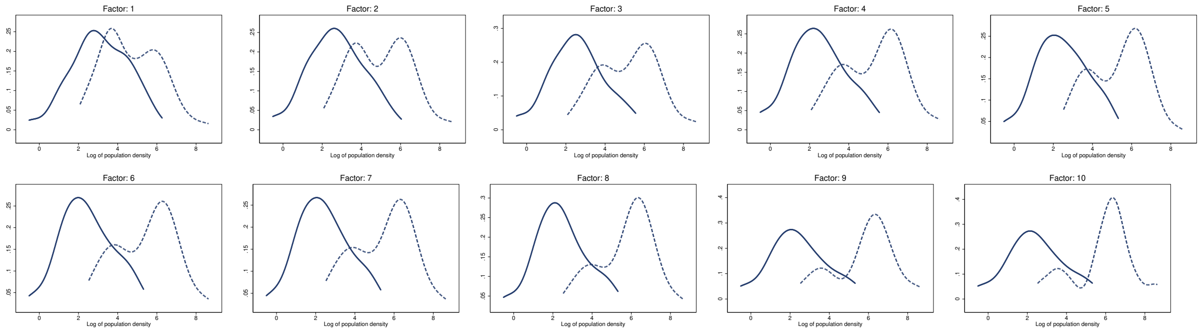
(b) Nigeria

		Home density bin									
		1	2	3	4	5	6	7	8	9	10
	1	73.6%	33.8%	18.2%	14.3%	13.4%	8.5%	10.2%	7.9%	6.5%	3.6%
	2	18.7%	50%	39.1%	27.9%	21.2%	13.1%	13%	9.7%	7.1%	4.1%
	3	11%	38.2%	43.6%	38.8%	29%	18.3%	13.9%	11%	8.6%	5.1%
	4	13.2%	35.3%	40%	40.1%	37.8%	22.4%	19.3%	13.1%	10%	6.4%
Visited density	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

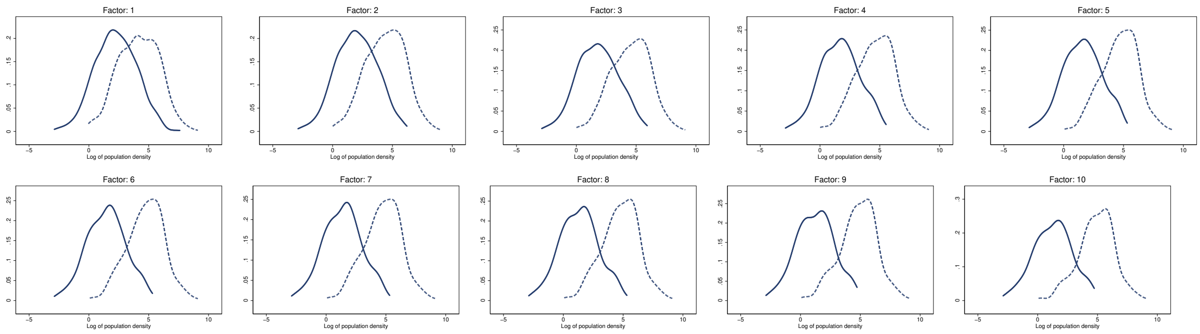
Note: These matrices show the proportion of users residing in home density bin i that are seen at least once in visited density bin j over the period of a year.

Figure C.15: Differences in flows between locations in Kenya



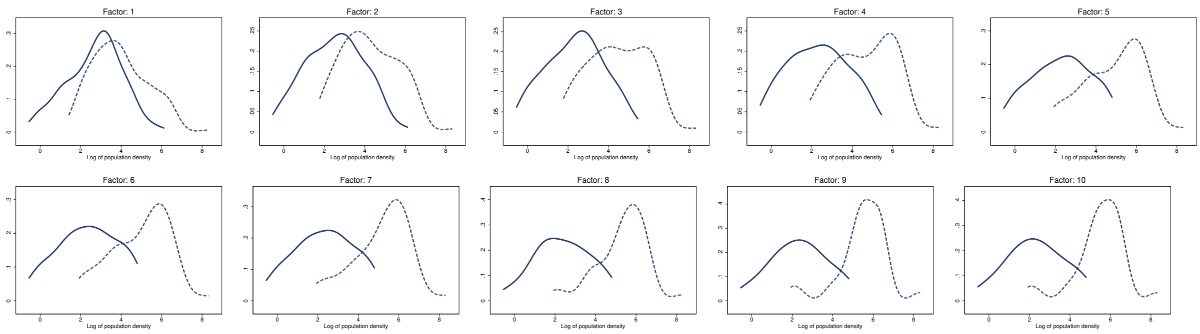
Note: This figure shows how the distributions of $\ln(V_k(1,2)/N_{k_1} * 1000)$ (dashed line) and $\ln(V_k(2,1)/N_{k_2} * 1000)$ (solid line) vary as we change the ratio of populations at origin and destination. We multiply the number of visits per resident by 1000 and take logs for expositional purposes.

Figure C.16: Differences in flows between locations in Nigeria



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.

Figure C.17: Differences in flows between locations in Tanzania



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.