Exploring smartphone data to estimate mobility in India

Robert Steven Banick

Suraj Regmi

Nayantara Sarma

Draft version: December 6, 2021

I. Introduction

Accurate and timely migration data are vital for effective policy on matters of urbanization, structural transformation of labor, disease control and disaster relief, to name a few. However, large scale, representative data on migration are unavailable in many countries. One common challenge relates to the definitions of migration – which can vary in terms of duration, purpose, and distance covered. Correspondingly, data sources are differently equipped to capture specific types of movements and miss others. This note aims to explore the feasibility and limitations of a relatively novel source of mobility information – anonymized smartphone mobility data – in the context of India.

India's national censuses identify migrants as individuals who reside in a location different from their birthplace or previous location. The population census provides only decadal snapshots of movements, presented at high levels of spatial aggregation, such as state or whether the geographic location is rural or urban. Such data are not informative about short-term movements between agricultural seasons or during crises. Although cross-sectional and panel household surveys aiming to measure migration exist, unless they are collected by official sources, such surveys are usually not nationally representative and may have limited generalizability. Using self-reported retrospective information or detailed household rosters, household surveys include information on relevant demographic and labor market covariates that are not available in the census but are pertinent to understanding migration. The timing of cross-sectional surveys is also critical for measuring short-term migration and may miss recording such migrants if conducted during the lean season in rural areas (Imbert and Papp, 2020).[1]

Recent advances in the data landscape have meant that researchers and policymakers can track human mobility patterns with much greater precision. Increased administrative data in the form of digitized vital registrations and population registries provide a means to track internal and international migration for the whole population (United Nations, 2014). However, such databases are available in only a few developed countries. Digital trace data - comprising information created by people's *actions* rather than by self-reported responses in surveys – can also be used to measure migration when they contain locational variables (Chi et al., 2020). The best studied trace data are Call Detail Records (CDR), that is, records of calls, texts, or mobile data usage by cellphones linked to the nearest broadcast tower maintained by a user's mobile network operator (MNO). Such data are rich, representative of all cellphone users, and sufficiently spatially accurate for most analyses. However, these seeming advantages are offset by MNOs' strong historical reluctance to share CDR data (Milusheva et al., 2021).

Another example of trace data is location data collected anonymously by smartphone apps and sold to mobile providers. While such data are globally available, they are not always representative of the whole population, especially in developing countries (Silver and Taylor, 2019). Biases in representation arise from two sources: differential ownership rates of smartphones by different demographic and social groups, and differential smartphone use patterns that trigger recording of the

[1] Kirchberger (2021) provides a comprehensive review of existing sources of migration data and their suitability for different research and policy purposes.

data, such as the use of a location-enabled mobile application (Coston et al., 2021, Milusheva et al., 2021). These representation issues are of particular concern in India, where smartphone usage is less than ubiquitous, gendered, and often shared amongst individuals or family members (Silver and Taylor, 2019; GSMA, 2020). Despite these disadvantages, smartphone mobility data is increasingly used to measure mobility patterns due to its spatial accuracy, the ability to analyze it over any spatial or temporal timescale and the straightforward nature of access relative to CDR data, which must be painstakingly negotiated on a per-country, per-MNO basis.[2]

This note aims to validate the suitability of smartphone locational data to effectively track population movements in India, despite these known biases in representation. Firstly, we collate data from smartphone users to generate mobility measures at various sub-national levels (districts and states). Next, we test summary statistics and mobility trends against measures from existing survey data. Finally, we compare mobility patterns from smartphone data with known patterns of movement triggered by specific events in West Bengal (India) and compare these with similar data from neighboring Bangladesh. We are specifically interested in comparing mobility in scenarios likely to affect the *entire* population to mobility in scenarios affecting specific *subsets* of the population, where representation bias of smartphone datasets is likely to be higher. For instance, we compare religious festivals, natural disasters and elections, which are likely indiscriminate in their impact, to events that disproportionately prompt movement from more vulnerable and low-income populations, like COVID-19 lockdowns and seasonal agricultural migration. By comparing movements during scenarios with different likely migrant population characteristics we attempt to learn about the representativeness of the data and establish a baseline level of confidence in the data's applicability in these contexts.

Our findings suggest the broad applicability of such data to track event-specific migration, but with caveats. While large parts of smartphone datasets are unusable due to unacceptably small numbers of records per user, as well as over-representation of urban samples, we nevertheless find that net gains and losses in population correspond with anticipated patterns of movement at the district level in most scenarios considered. More specifically, we observe strong signals of population mobility around COVID lockdowns, festivals, and disasters, with some minor deviations in the latter case. Seasonal migration largely conforms to anticipated patterns, but more so in some seasons than others; this may be due to overlap between certain seasonal events and other movement triggers (e.g., festivals).

Our paper contributes to two branches of research. The first is the emerging technical body of work on using smartphone data to study mobility, particularly urban-rural migration over short term periods. The second is the long-running discourse on the nature and extent of migration in India.

[2] Another data source is Facebook which collects anonymous mobility data from its users and shares with trusted partners various mobility datasets based on these, aggregated between 0.6 and 2.5 km (usually) over 8-hour intervals, via its Data For Good initiative. However, the questionable representativeness of Facebook users for entire populations, limited information on origins and destinations, and the need for a recognized "disaster-event" to extract data all limit its utility in reliably estimating short-term mobility.

Prior to the pandemic, research using Call Detail Records provided early demonstrations that such datasets could be used to track population movements triggered by natural disasters. Following its use after the 2010 earthquake in Haiti (Bengtsson et al. 2011), analysis of CDR was rapidly deployed after the Nepal earthquake in 2015 to provide spatiotemporally detailed estimates of population displacement. This analysis was used to effectively target and deliver humanitarian relief (Robin et al., 2016). Related analysis in Kenya finds that movement estimates derived from CDR data are robust to known biases in phone ownership amongst geographic and social groups (Wesolowski et al., 2013). Steele et al. (2017) demonstrated the further utility of basic CDR metadata (e.g., user to population ratio, top-up patterns, social network usage, and average mobility) as strong predictors of poverty when preparing intra-census poverty estimates at small spatial scales in Bangladesh. Research using CRD data in developed countries such as Portugal and France is more established, where user densities are shown to closely track actual population distributions at even small spatial scales and anticipated seasonal trends (Deville et al., 2014).

Studies and usage of smartphone-based mobility data have expanded dramatically post- pandemic as policymakers have sought rapid guidance on the efficacy and economic impacts of lockdowns and other public health restrictions. Most notably, Apple's Mobility Trends Reports and Google's Community Mobility Reports have provided general purpose mobility datasets. These were later complemented by more bespoke studies and measures by academics and data scientists – for example, movement patterns extracted from smartphone mobility data were used to create measures of *potential* COVID-19 exposure (Couture et al., 2020), correlate with *actual* COVID-19 incidence and mortality (Badr et al., 2020), and assess the differential impacts of public health restrictions on different socioeconomic groups (Chang et al., 2021). The vast majority of these studies occur in the developed world, where large, representative sample sizes and complementary reference data permit levels of spatial and demographic detail not possible in most developing countries. An exception is a study based in China using real-time travel data from internet service providers to predict spread of COVID-19 and test efficacy of control measures (Kraemer et al. 2020). Travel patterns among smartphone owners are different from basic-feature phone owners, and even less is known about the difference with individuals who do not own a cellphone at all (Milusheva et al., 2021). Within the World Bank, the Global Facility for Disaster Reduction and Recovery (GFDRR) has used mobility data to analyze local economies, commuting, growth patterns, and immediate post-disaster movement in urban centers in Costa Rica, Nepal, and Mexico.

From this literature several useful findings and guidelines emerge, not all of them applicable in the Indian context. Kishore et al. (2020, 2021) note that the data generation process for mobility data (i.e. how often GPS points are collected) and the aggregation methods used shape aggregate metrics are still poorly understood. Moreover, data generation processes may differ by geography due to different transport modalities and trip distances. Costone et al. (2021) find that smartphone-based mobility estimates underestimate elderly and minority groups in the United States and therefore recommend using spatially detailed voter turnout data to audit the representativeness of mobility estimates. In a developing country context, Milusheva et al. (2021) compare cellphone usage data from basic feature phones and smartphones in Uganda and urge caution when relying on mobility estimates solely sourced from smartphones in contexts with low adoption. Comparing CDR data-

based estimates to Google's Mobility reports, Szocksa et al. (2020) recommend using the former for mass population movements and the latter for tracking individuals' or small population groups' compliance with public health measures. The GFDRR addresses the uncertainty around mobile data's representativeness with a simple heuristic that user samples sizes should equal or exceed 0.5% of the actual census population in a study area.

Our paper is also related to a substantial body of work on measuring the extent and correlates of migration in India and Bangladesh. While we do not attempt a complete literature review here, it is pertinent to note that while early work has relied mainly on Census and National Sample Survey (NSS) data, more recent discussions of migration refer to novel sources like Indian Rail Passenger datasets (Economic Survey of India, 2016-17; Firth et al., forthcoming). Historically, migration in India has been considered low in India (Munshi and Rosenzwieg, 2016) and an 'underinvested technology' in Bangladesh (Bryan et al., 2014). State borders in India are known to act as barriers to migration due to the limited portability of public entitlement schemes (Kone et al., 2018). However, the months following COVID-19 related lockdowns saw mass movements precisely for the same reason, as individuals were not able to access food rations and secure housing to quarantine in place at migration destinations (Jesline et al., 2021).

The rest of this note is organized as follows: the next section describes the smartphone and survey data used for comparison in our analysis. Section III outlines the methodology and Section IV the main results. The last section concludes with some limitations and possible extensions for this work. User guides for reproducing the analysis in this paper, additional considerations and tests related to the use of smartphone data are presented in the Annexes.

II. Data description

Smartphone data

In this study we use mobility data from Unacast and Veraset, which are private-sector aggregators of locational "pings" collected from smartphone users in India, Bangladesh, and countries around the world.[3] Unacast and Veraset obtain their data from application providers who capture users' locations at the time of usage (with users' consent via privacy policies), then clean, anonymize, and aggregate results to sell onwards to data scientists and analysts.[4] The identity of the apps providing data, the demographic profile of their users, and the contexts in which users' locations are captured are trade

[3] The World Bank was granted access to Unacast and Veraset's datasets through late 2020 under the terms of its engagement with the Development Data Partnership (DDP). To help implement the Sustainable Development Goals and operationalize the World Development Report 2016, the World Bank launched the Digital Development Partnership. This partnership makes digital solutions available to developing countries with an emphasis on data, digital access, cybersecurity, etc.

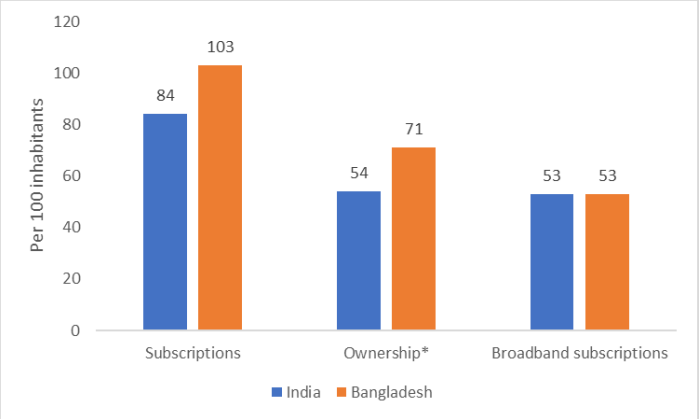
https://www.worldbank.org/en/programs/digital-development-partnership Accessed Dec 5, 2021

[4] This is a rapidly evolving commercial space and other providers such as Outlogic (formerly X-Mode) have emerged in South Asia since this research was conducted; we therefore avoid focusing on any one provider, except to show that results from Unacast generally mirror those from Veraset, enhancing the internal validity of our figures.

secrets not shared with the World Bank. This means that we must infer demographic characteristics of users from secondary data.

Mobile subscriptions per 100 inhabitants are high in both countries at 84 and 103 in India and Bangladesh, respectively. Individuals can have multiple subscriptions as mobile ownership is lower at 54 percent in India and 71 percent in Bangladesh (Figure 1). Only half of the countries' inhabitants have active mobile broadband subscriptions – these inhabitants are presumably richer and more educated than those without subscriptions.

Figure 1: Mobile and mobile internet penetration



Source: ITU, World Telecommunication/ICT Indicators Database (Data collected up to November 2020). *Mobile ownership data for India are from Statista (2020)

Smartphone mobility data offers an enormous collection of timestamped GPS data points, or 'records', but extremely limited metadata – only the device number and type. Each record's location information is a GPS point with spatial "noise" added by providers to protect privacy. The exact amount of noise (usually between 30-100 meters), which compounds existing GPS inaccuracies, depends on the provider, whether the record is located in a rural/urban setting. Temporal information comes from a UTC timestamp accurate to the millisecond.[5] Any other information about the population of users must be inferred from these data. Identifying information comes in the form of a unique device ID, which we treat as a single user, and the iPhone/Android status of the phone.

Most unique device IDs generate very few records even over long periods and as such, complicate mobility analysis by inflating the denominator of statistics. Since little can be deduced from such observations, we employ a minimum records criterion to exclude infrequent users. We anticipate these users to be tourists or users with no/few tracking applications. Additionally, the infrequency of records from most users in the developing world plus the paucity of explanatory information makes it hard to establish whether a record in a changed location represents lasting movement, a fortuitous

[5] No other data is provided through the World Bank's Development Data Partnership (DDP) about individual records or the dataset as a whole and DDP providers are unable to further clarify to protect commercially valuable information. Mobility data cannot be reliably used for real time analytics less than two weeks removed from the present day, as providers don't finish processing most records for seven days and all records for fourteen.

ping during a quick trip, or something in between. This adds an inherent uncertainty to most movement calculations that can only be resolved with stricter filters and movement definitions (which in turn limit sample sizes).

Annex I directs the reader to a repository containing codes and data to reproduce the analysis in this note. Beyond the extraordinary temporal and spatial specificity of smartphone data, the dominant characteristic of such datasets is their large size. We discuss some of these technical considerations for the benefit of other users in Annex II.

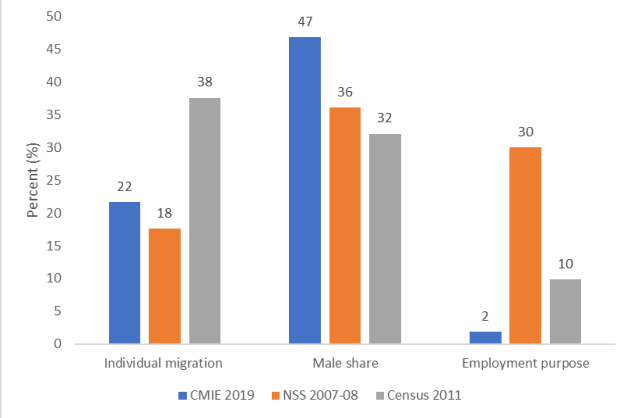
Household survey data

We compare mobility estimates from smartphone data with those from available household surveys in India. Official sources of migration data are dated. Researchers either use the 64th round of the National Sample Survey (NSS) conducted in 2007-08 or the population census from 2011. The Census (2011) measures whether an individual has "come to this village/town from elsewhere". Therefore, short-term migration, which would not require a change in household residence, is unlikely to be captured. The 64th round of NSS contains information on the migration history of each household member, covering both short and longer-term migration.

More recent estimates of migration levels in India are available from the Consumer Pyramids Household Survey (CPHS) collected by a private data collection agency called the Centre for Monitoring the Indian Economy (CMIE). The CPHS is the world's largest ongoing panel survey: around 170,000 households are visited once every 4 months, with visits staggered such that 25 percent of the sample is visited each month. Household members are classified as migrants if they are absent for one or more rounds of the panel survey. CPHS, thus, captures migration of periods greater than 4 months. As this is only a subset of all migrants, we expect CPHS migration data to only weakly converge with net movement rates that are calculated on a month-to-month basis.

The NSS (2007-08) contains different measures of migration. It captures movements occurring any time in the past, and for short durations between 1-6 months. We use the former definition in Figure 2 and show that there are clear differences across migration measures arising from different surveys. This is expected from the varying scope of the underlying definitions and the different years in which the surveys were conducted.

Figure 2: Comparison of migration measures across surveys



Note: Migration definitions vary across surveys. CMIE (2019): Member of household who is currently not residing there. NSS (2007-08): Any former member of the household who migrated out any time in the past. Census (2011): Individual who has come to this village/town from elsewhere.

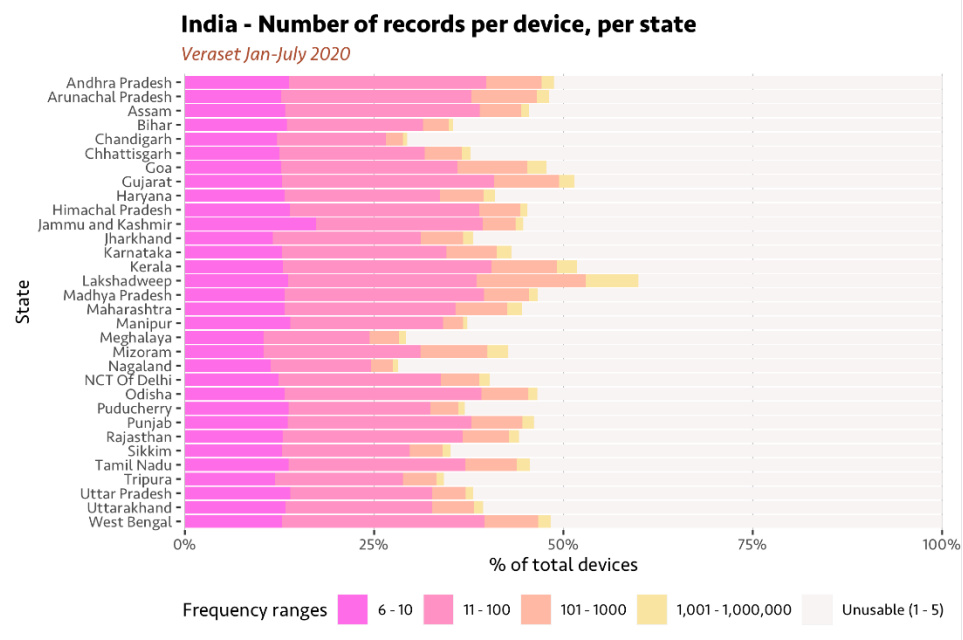
While Figure 2 presents the share of migrants currently in the population, it only captures the "stock" notion of migration and not its "flow". That is, the definitions used across datasets include migrants who moved at *any point in time* in the past. Higher frequency data sources are better equipped to measure the corresponding flows of migration, which occur over a specific period of time. NSS (2007-08) and the CPHS panel data can detect movements occurring in the past year (or longer, in the case of CPHS) and for durations ranging between 1-6 months (longer than 4 months for the CPHS). Keeping these considerations in mind, we use corresponding survey measures only from the CPHS (2019-20) to compare mobility changes from smartphone data using an *overlapping* time period.

III. Methodology

To assess the suitability of smartphone data for mobility estimates, our analysis is conducted in the following steps:

First, we estimate net-movement rates of smartphone users aggregated by different geographic levels at a weekly/monthly frequency. This involves mapping each record of data to the smallest administrative area (mouza in Bangladesh, village/ward in India) which encloses the device's GPS coordinates when it is being used. Boundaries for disaggregated geographical units were sourced from the South Asia Spatial Database for India (Li et al., 2016) and the Bangladesh Bureau of Statistics (BBS). We eliminate inapplicable records which contain too few records or fall outside the target date range or analysis area. Higher record filters trade off higher quality results for less representative sample sizes. The use case determines the appropriate number of records to filter by: for year-long analyses we typically discard users with fewer than 6 records (Figure 3 below), whereas we require only 2 records for analyses of less than 2 weeks. By contrast, some developed world studies report filtering users with less than 10 records per *day*, a standard that would result in us discarding 99.99999% of records.

Figure 3: Visualizing the number of usable devices

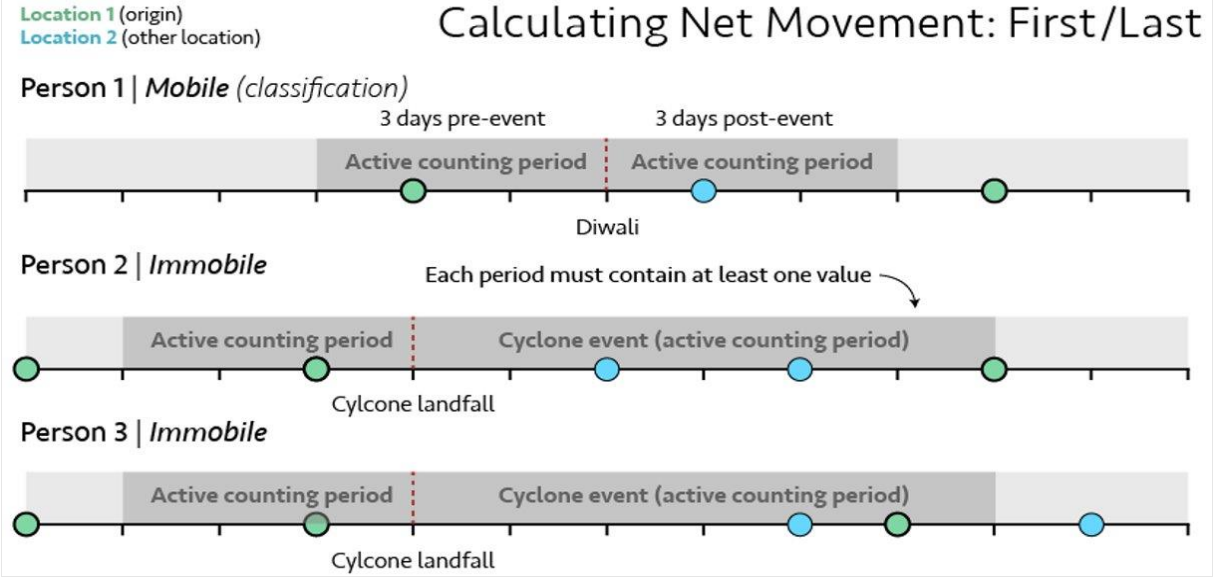


We define mobility as changes in location of a unique device ID, i.e. if the last record in a specific time period is in a different location from the first (illustrated in Figure 4).[6] Person 1 is counted as mobile because she is in a different district at the end of the specific time-period (3 days after Diwali) whereas Person 2 is counted as immobile because her *final* GPS ping is in the same location as her first, despite leaving for Location 2 in between. Person 3 is similarly counted as immobile despite being in Location 2 immediately before and after the final ping of the counting period. In practice, most users' movement is stable -- if they move to Location 2, they tend to stay for a time -- but these examples illustrate the sensitivity of the definition to the counting period lengths chosen.

The sum of all such movements into and out of a district or state is defined as **net movement**, with *positive net movement values indicating net inflows,* or population gains in the district/state. The corresponding **rate of net movement** for a district/state (or other unit of analysis) is net movement divided by the sum of all users in that district/state. It bears mentioning that in any period the vast majority of users are observed to be immobile, so net movement rates are usually small.

[6] Another definition of mobility relies on 'modal movements': A user's location is marked as changed if the last record in the time period is in a different location from the modal values of all the points in the time period. Modal values are frequently employed in other studies with larger samples (Kishore et al., 2020) as they offer a more reliable indicator of the user's primary location in a given time period. The First/Last definition we use is more computationally efficient and was observed to yield results similar to those calculated using the modal values definition in limited sample sizes. Our definition differs from the modal metric recommended by Kishore et al. (2020) because the limited observed sample sizes over small time periods precluded calculating modal values.

Figure 4: Visualizing net movement definition



Source: Authors' illustration

Secondly, we compare trends of mobility against measures from the CPHS data for the same geographic area and time-period. Widespread lockdowns enacted to curb the spread of COVID-19 provide a unique opportunity to study changes in mobility. The impacts of the lockdown should be perceptible in smartphone data, as well as in the CPHS which adapted to phone-based survey protocols during the pandemic. Additionally, since lockdowns were implemented across borders, their impact and the timing of the same should be detected across cities.

Finally, we consider events which may trigger movements of i) the general population and ii) specific subsets of the population based on demographic attributes. For the above analyses, we focus on West Bengal in India and its neighbor, Bangladesh which may serve as logical counterfactuals for each other. While both areas have similar rice-based agricultural economies, their populations have different demographic characteristics and different political calendars. This allows us to test the mobility impacts of specific religious festivals, that would have an impact on one area but not on its 'counterfactual'. On the other hand, due to similar agroclimatic conditions, seasonal mobility in both areas should correspondingly have similar patterns.

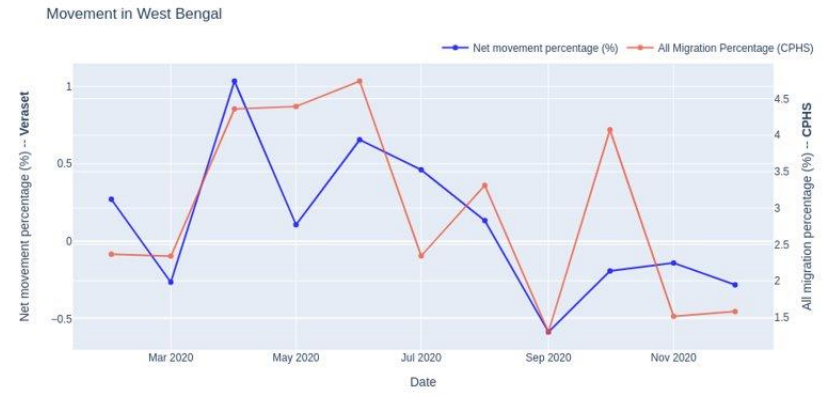
The correspondence of smartphone mobility trends with measures from household surveys, along with event-based tests using a geographic counterfactual, should indicate the reliability and limitations of smartphone data.

IV. Results

Comparing smartphone mobility measures with household survey data post-COVID-19

We begin our analysis by comparing rates of net movement in West Bengal derived from smartphone data to that from survey data (CPHS) during post-pandemic months. Despite differences in the measurement methodology and data types we find a strong correlation coefficient of 0.65 between monthly net movement rates from smartphone data and migration rates in CPHS. In Figure 5 below, the CPHS trend closely tracks net movement over most months (on different Y axes). Particularly notable is the spike in April 2020 when nation-wide lockdowns came into effect in India, accompanied by extensive journalistic coverage of a 'migrant crisis'.[7] While the magnitudes of the two measures differ, their synchronicity suggests some validity to the measures from both data sources.

Figure 5: Comparing net movement rates from smartphone data and migration from CPHS

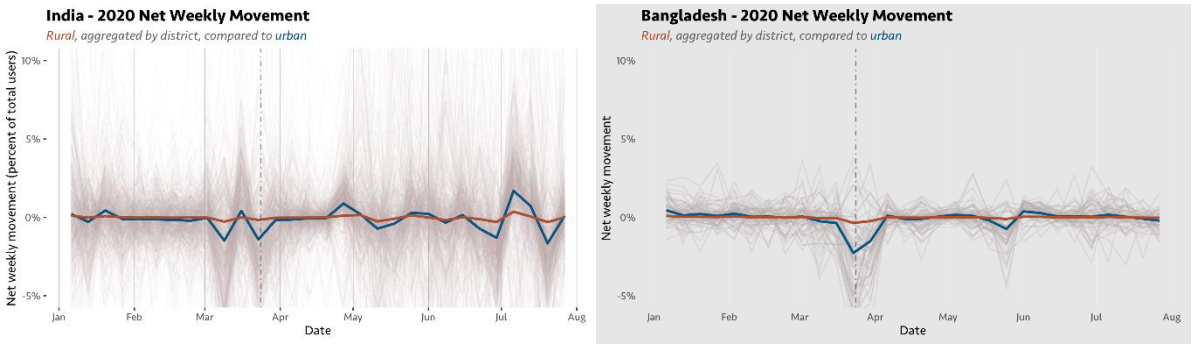


Source: CPHS (2020) and Veraset
Note: Migration in CPHS is defined as the share of individuals who have returned to their households from a different location (within/outside of the state) since the last time the household was surveyed. Net movement rate from Veraset measures the share of users who are in a different district at the end of 1 month (the active counting period) from their first location recorded in the same month. Districts may or may not be within the same state. The denominator is all individuals surveyed, and all smartphone users recorded in each month in the state for CPHS and Veraset, respectively.

The above patterns are consistent at the national level and follow similar patterns for both India and Bangladesh (Figure 6). On average, there is a spike in movement *away from urban areas, i.e. outflows* shown by the blue lines after the lockdown announcements in March 2020 as migrants returned to their native villages, followed by several months of limited movement as people stayed put (observable in the reduced noise of individual grey district lines). This contrasts with a relatively flatter line for the average net movement for rural districts, where there was lower net movement or a larger base population.[8] Subsequent positive spikes in India's net movement for urban areas around July 2020 possibly indicate *inflows* into cities as lockdowns were relaxed and workers returned. Similar patterns are observed in Bangladesh.

[7] https://www.bbc.com/news/av/world-asia-india-55434594 Accessed November 29, 2021.
[8] The majority of migration in India is from rural areas to other rural areas, so it is likely that we are not able to detect a positive or negative average trend as flows cancel each other.

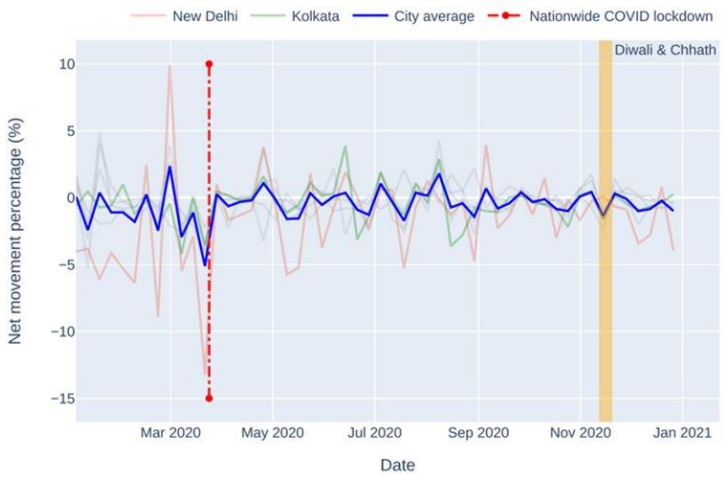
Figure 6: Net weekly movements after COVID-19 related lockdowns in India (left) and Bangladesh (right)



Source: Veraset
Note: The graph above disaggregates net movement in urban (in blue) and rural (in red) districts. Districts are classified based on the share of population living in urban/rural areas within the district in Census 2011.

Limiting our analysis to two of India's major urban centers (Kolkata and Delhi) in Figure 7 shows that the overall outflows, or net population loss, observed in Figure 6 (left) comes from such urban centers.[9] This is consistent with extensive news reporting and preliminary studies of low-income and transient populations in India and Bangladesh leaving cities for their home districts when the lockdowns were initially implemented.[10] Accordingly, these figures can be interpreted as indicative of smartphone data's ability to track mobility trends for low-income and transient populations at large scales, despite lower rates of smartphone ownership amongst them in developing countries.

Figure 7: Net movements triggered by COVID-19 in India's major cities



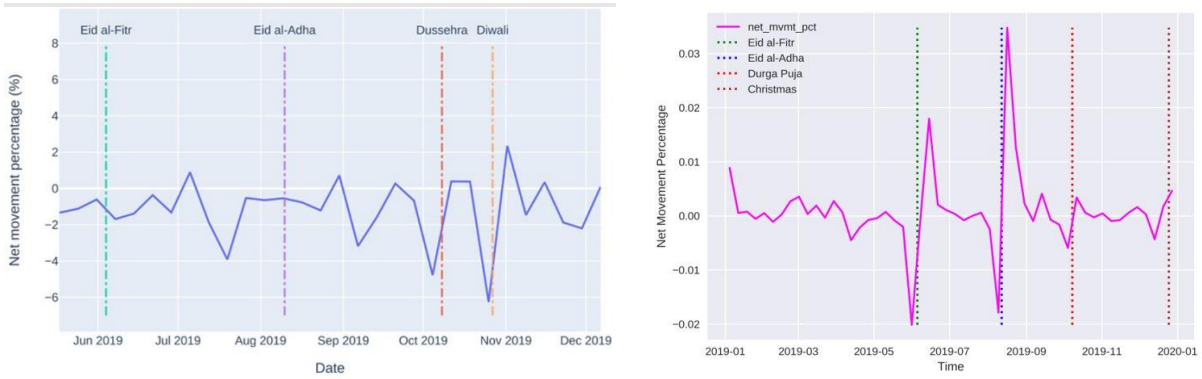
Source: Veraset

[9] Kolkata and Delhi are among India's densest cities (Census 2011).
[10] https://www.theguardian.com/world/2020/mar/30/india-wracked-by-greatest-exodus-since-partition-due-to-coronavirus Accessed November 29, 2021

Event-based tests of smartphone mobility measures: festivals, cyclones, and elections

We next compare net movement rates in Kolkata (West Bengal, India) to Dhaka (Bangladesh), with attention to local religious festivals. In Figure 8 (left), Kolkata has *negative net movement*, or outflows of population, before major Hindu festivals like Dussehra and Diwali, as city-based migrants visit their rural districts, but less so before Muslim festivals. On the right (Figure 8), the same trend is observed before major Muslim festivals in Dhaka, but not for Hindu festivals.

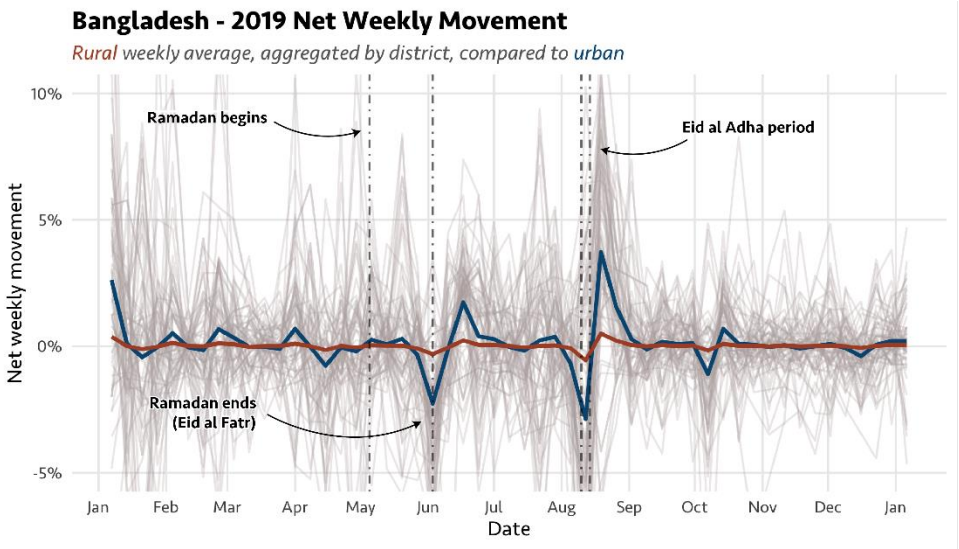
Figure 8: Net movements in Kolkata (left) and Dhaka (right) around festivals



Source: Unacast

When viewed at a national scale for Bangladesh in Figure 9, these trends are repeated, as populations leave cities for other parts of the country. The weekly average urban trend of net-movement dips just before Eid celebrations as net movement is negative and populations flow out of cities (in blue). The trend peaks sharply immediately afterwards indicating positive net migration, or *inflows*, when migrants return to their work locations. Again, we witness a relatively flatter trend for rural areas (in red).

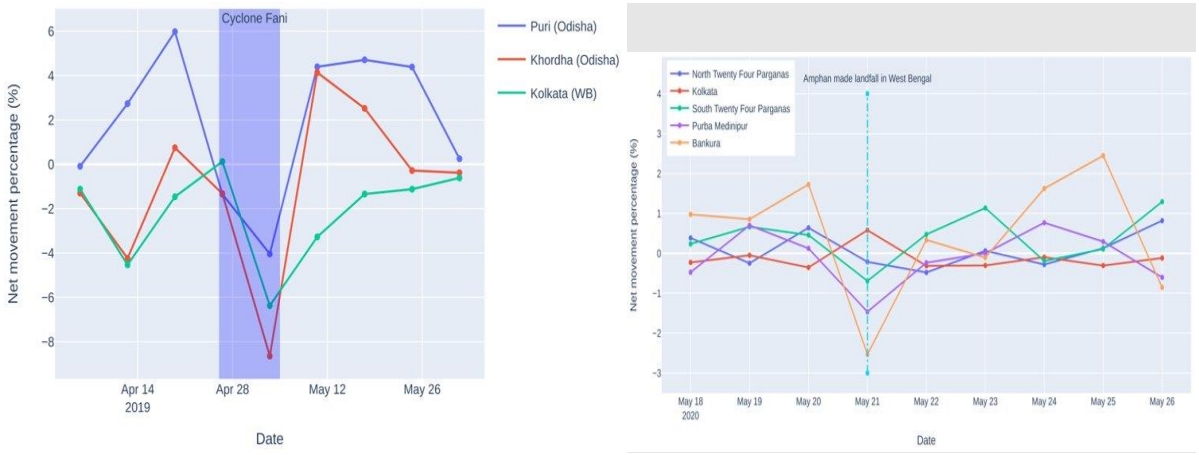
Figure 9: Weekly net movements in Bangladesh, by district



Source: Unacast

Tropical cyclones are frequent and potentially devastating phenomena in coastal regions along the Bay of Bengal. Cyclones Fani (2019) and Amphan (2020) strongly impacted the coast of Bangladesh and Eastern India and prompted general evacuations. We aim to detect the impacts of these cyclones in net movement rates from smartphone data. In India, such movement is visible for cyclone Fani in Figure 10 (left) and for Amphan in Figure 10 (right). Cyclone Fani made landfall in Odisha on May 3, 2019 and we detect high outflows, or *negative net movements, from* key districts, as well as in Kolkata, suggesting that people were evacuated to minimize damage. Similarly, when Cyclone Amphan hit coastal districts of West Bengal, net movements dropped to -3 percent as populations moved out of them. Notably, Kolkata is the only district significantly gaining during this period, possibly suggesting that evacuated populations may have temporarily decamped to the city, in order to shelter in the houses of relatives.

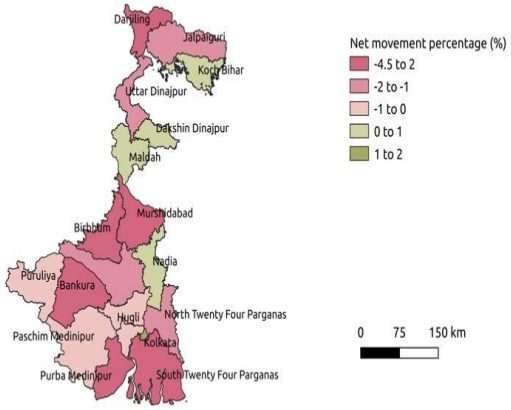
Figure 10: Net movement triggered by Cyclone Fani (left) and Cyclone Amphan (right)



Source: Veraset

A spatial distribution of net movements during Cyclone Amphan is presented in Figure 11. The districts in red identify those which saw up to 4.5 percent of users change location *away* from them, while district in green witnessed inflows (e.g., Kolkata, Maldah). However, it is more difficult to explain the outflows in the far north of the state (Darjeeling, Jalpaiguri, and Uttar Dinajpur), where cyclone rains perhaps triggered flooding in these mostly hilly terrains.

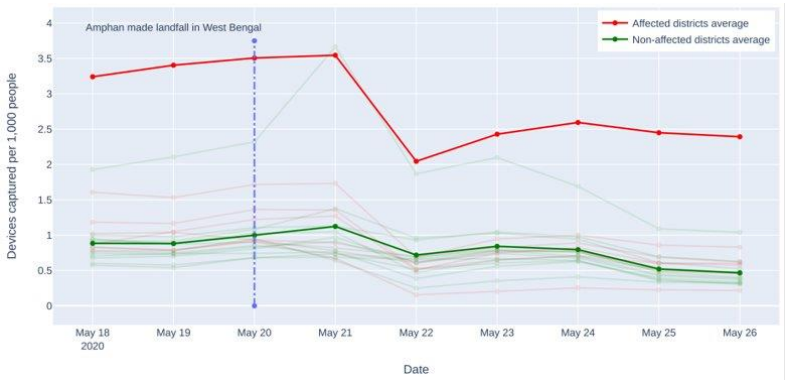
Figure 11: Map of net movements in districts of West Bengal during Cyclone Amphan



Source: Veraset

In addition to locational information contained in smartphone data, the absolute number of users available during natural disasters is itself informative. Smartphones denied a recharge for one or more days due to power cuts and/or evacuation to sites without power will eventually shut off and hence drop out of the sample, possibly complicating our analysis. Sharp reductions in unique users without corresponding changes in net movements can, however, help identify geographic areas in need of humanitarian relief. After Cyclone Amphan we, indeed, see a sharp drop in sample rates for affected districts in West Bengal (Figure 12), although movement was still visible from users who remained charged or moved before their phone died (see Figure 10, right). It is possible that loss of power may have distorted observed movement rates, but this cannot be tested or quantified with available data and remains a matter for future research.

Figure 12: Using sharp changes in number of records as indicative of mobility/access to power

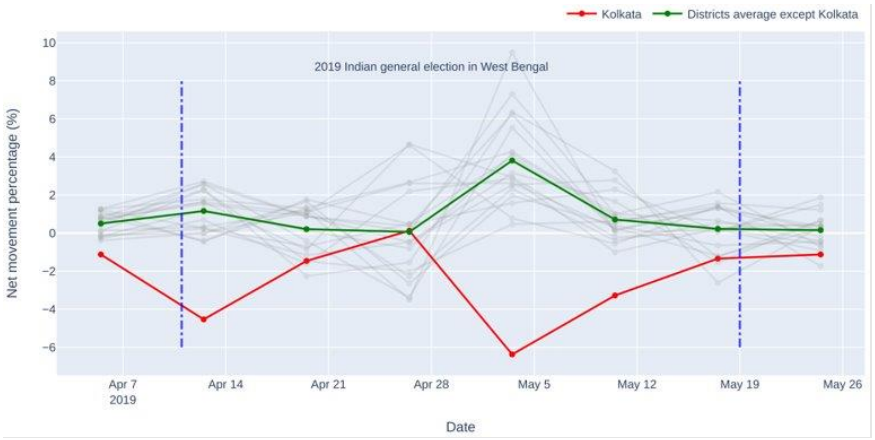


Source: Veraset

Another 'movement-triggering event' we consider is India's 2019 general election. National and regional elections provide a useful test case of smartphone mobility data's ability to detect movement amongst transient populations. Migrants temporarily settled in cities for short- and medium-term

work would be expected to leave cities during election periods to cast votes in their home districts. This precise pattern can be observed in Figure 13 which visualizes net movements of people in Kolkata and the rest of West Bengal during the 2019 Indian general election period. The negative net migration – i.e., away from Kolkata – is mirrored by positive inflows into other districts.

Figure 13: Net movements in West Bengal during general elections, comparing Kolkata and other districts



Source: Veraset

Impact of agricultural seasons in West Bengal and Bangladesh

Finally, we look at seasonal migration occurring at pace with local agricultural calendars. Seasonal migration provides a well-researched and important test case of migration unique to populations of interest to policymakers. Both West Bengal and Bangladesh's agricultural economy is heavily linked to rice production seasons, called Aman, Boro, and Aus (visualized in Figure 14). These seasons contribute 39, 54, and 7 percent of rice production in Bangladesh, respectively (Tisdell et al., 2019). Each season is characterized by intensive periods of work during plantation and harvest followed by periods of limited food and work in rural areas known as *Monga,* or the lean season (Khandker, 2012). Temporary labor migration to cities during interim lean periods significantly boosts development outcomes amongst low-income farming households (Bryan et al., 2014). Yet tracking such short-term movement at scale is at present very difficult: censuses and other expensive traditional instruments can only capture infrequent snapshots. Bryan, et al. (2014) manage to do so only within the framework of a randomized control trial.[11]

[11] The CPHS only captures movement greater than 4 months and thus misses movements if they fall entirely within the panel interregnum.

Figure 14: Illustration of rice-agricultural seasons in West Bengal and Bangladesh

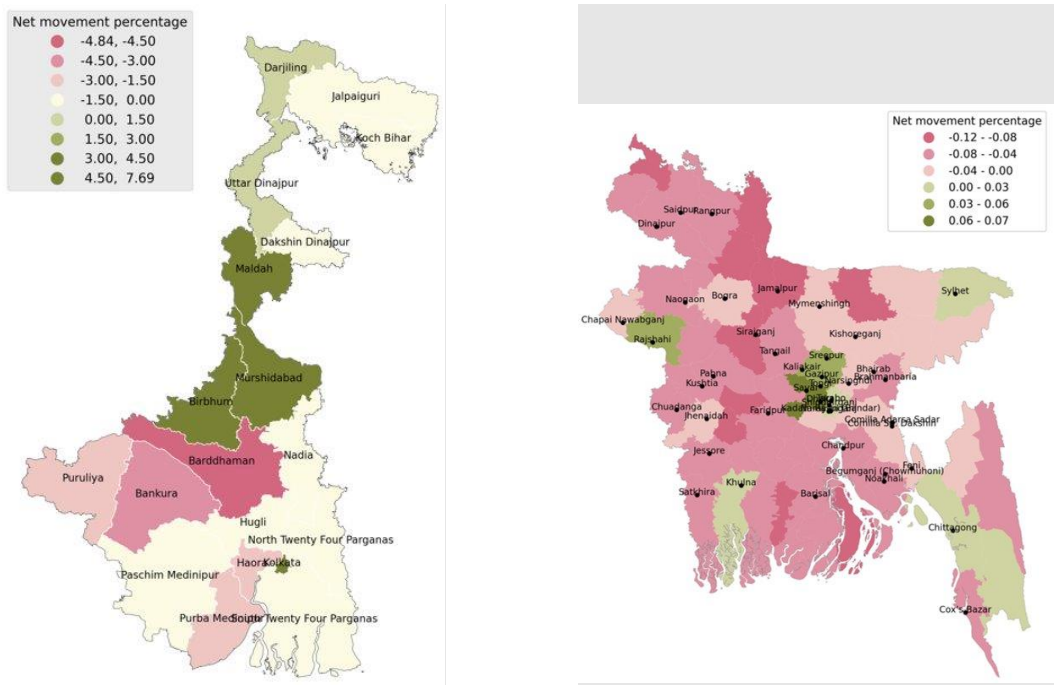
Source: Authors' illustration using time periods for the predominantly used high-yield rice variety

We focus our analysis on the Aman season given its significance to households' agricultural production and its dependence on the main Southwest summer monsoon in the region. As expected, the maps below show higher net movements into rural areas before planting (shown in green in Figure 15). These patterns of net movement are consistent for both West Bengal and Bangladesh. Kolkata and Dhaka lose between 7 to 1 percent of their total userbase, respectively, while rural areas gain them with varying magnitudes just before the Aman planting season.

Spatial patterns of movement after the Aman planting, when the first *Monga* or lean season starts, indicate movements in the reverse direction, especially strong in Bangladesh. We see movements *into cities* at the beginning of Monga or lean periods when there is less agricultural work available in the villages (shown in green in Figure 16). Inflows are observed into Dhaka and a few other major cities in Bangladesh, with corresponding negative movements out of the more rural districts. In West Bengal, Kolkata gains population as anticipated, but so do the mainly rural northern districts of Birbhum, Murshidabad and Maldah. The reasons are unclear and merit further study: it may be that the agricultural calendar is behind in these areas, or that other unobserved variables are driving migration inflows towards these districts. An additional factor is that the end of the shorter *Aus* harvet season extends slightly beyond the end of the Aman plantation and may delay return movement to cities. However, why this overlap of seasons differently impact movements in West Bengal and Bangladesh is unclear.

Figure 15: Net movement before Aman plantation season in West Bengal (left) and Bangladesh (right)

Figure 16: Net movement in the post-Aman 'Monga' period in West Bengal (left) and Bangladesh (right)



Source: Unacast

Annex III discusses some limitations of extending our analysis to administrative areas below the district level and tests correlations of net movement rates to other migration-related outcomes, such as wages.

V. Discussion and limitations

This note sets out to examine the utility of smartphone mobility data in the Indian context. We do so by first comparing mobility estimates derived from this relatively novel source with measures from existing household survey data. Examining trends during COVID-19 related lockdown periods in India and Bangladesh suggests the consistency of smartphone mobility measures at the state and district levels. Although magnitudes of mobility measures vary across sources, their directions match that of widely reported anecdotal and journalistic evidence despite valid concerns about the under-representation of poorer, migrant populations in smartphone mobility datasets due to lower ownership rates. These findings therefore suggest that smartphone mobility data is a potential means of tracking the short-term mobility.

Our validation exercise tries to examine the limitations of representability more closely. We next use a variety of test cases relying on well-known 'mobility-triggering events', such as festivals, natural disasters and elections to detect anticipated movements among smartphone users. The results indicate confirmatory signals at subnational units of sufficient granularity (districts) to inform policy making. However, we observe some seasonal mobility results in rural areas which do *not* match the anticipated pattern of changes between agricultural seasons. This is particularly true for spatial disaggregation of mobility estimates below the district level as the smartphone user sample shrinks. Additional validatory work is needed to uncover links between mobility and events that trigger it, explain departures from anticipated movements, and unpack any generalizable implications for the use of mobility data.

Our findings do not eliminate the possibility that these data sources insufficiently represent poorer migrant households. A serious limitation in this regard is the lack of recent, official, and nationally representative migration data to establish conclusive comparisons with smartphone mobility data sources. Moreover, the relative novelty of mobility data means that there are few guidelines on appropriate usage or known shortfalls. What guidance is available is primarily developed in the developed world, usually in urban settings, and is thus of questionable applicability to rural areas of the developing world.

This initial exercise serves to inform future research agendas on the reliability and appropriate uses of mobility data in rural areas of developing countries. An important foundational piece of follow-up research would be to quantify the precision of mobility movement signals like net movement in rural areas of the developing world and certify their continued validity at small subnational scales where sample sizes shrink. This is a non-trivial challenge as it would require carefully selected representative test cases and rich comparison datasets at small subnational and temporal scales. Cross-disciplinary qualitative work establishing patterns of smartphone usage within families and between genders could further nuance conclusions – there is likely a rich anthropological literature in this regard.

Regardless of these caveats, the findings from our work validate smartphone mobility data's potential in understanding short- and medium-term migration dynamics in South Asia. With smartphone ownership and computing technologies only expected to grow, the limitations of mobility data's analysis will only wane in the coming years. Looking forward, incorporating additional mobility metrics, such as origin destination matrices, distance to or popularity of key local services like markets, hospitals, or clinics, can address or advance the state of knowledge for many policy-relevant migration and mobility questions.

References

Aranda-Jan, Clara, et al. "The Mobile Disability Gap Report 2020. London: GSMA." *The Mobile Disability Gap Report* 4 (2020): 4.

Badr, Hamada S., and Lauren M. Gardner. "Limitations of using mobile phone data to model COVID-19 transmission in the USA." *The Lancet Infectious Diseases* 21.5 (2021): e113.

Bengtsson, Linus, et al. "Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti." *PLoS medicine* 8.8 (2011): e1001083.

Bryan, Gharad, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak. "Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh." *Econometrica* 82.5 (2014): 1671-1748.

Chang, Serina, et al. "Mobility network models of COVID-19 explain inequities and inform reopening." *Nature* 589.7840 (2021): 82-87.

Chi, Guanghua, et al. "A general approach to detecting migration events in digital trace data." *PloS one* 15.10 (2020): e0239408.

Coston, Amanda, Neel Guha, Derek Ouyang, Lisa Lu, Alexandra Chouldechova, and Daniel E. Ho. "Leveraging Administrative Data for Bias Audits: Assessing Disparate Coverage with Mobility Data for COVID-19 Policy." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 173–84. Virtual Event Canada: ACM, 2021. https://doi.org/10.1145/3442188.3445881.

Couture, Victor, et al. "JUE Insight: Measuring movement and social contact with smartphone data: a real-time application to COVID-19." *Journal of Urban Economics* (2021): 103328.

Deville, Pierre, et al. "Dynamic population mapping using mobile phone data." *Proceedings of the National Academy of Sciences* 111.45 (2014): 15888-15893.

Firth, John, Felix Forster, and Clement Imbert. "Internal migration in India: New evidence from rail passenger travel data."." (2017).

Government of India (2010) Migration in India 2007-2008, NSS Report No. 533 (64/10.2/2), National Sample Survey Office Ministry of Statistics & Programme Implementation Government of India June 2010

Government of India (2017a) Economic Survey 2016-17, January 2017, Ministry of Finance Department of Economic Affairs Economic Division

Imbert, C., & Papp, J. (2020). Short-term migration, rural public works, and urban labor markets: Evidence from india. *Journal of the European Economic Association*, *18*(2), 927-963.

Jesline, Joshy, et al. "The plight of migrants during COVID-19 and the impact of circular migration in India: a systematic review." *Humanities and Social Sciences Communications* 8.1 (2021): 1-12.

Khandker, Shahidur R., MA Baqui Khalily, and Hussain A. Samad. "Seasonal migration to mitigate income seasonality: evidence from Bangladesh." *Journal of Development Studies* 48.8 (2012): 1063-1083.

Kirchberger, Martina. "Measuring internal migration." Regional Science and Urban Economics 91 (2021): 103714.

Kishore, Nishant, Mathew V Kiang, Kenth Engø-Monsen, Navin Vembar, Andrew Schroeder, Satchit Balsari, and Caroline O Buckee. “Measuring Mobility to Monitor Travel and Physical Distancing Interventions: A Common Framework for Mobile Phone Data Analysis.” *The Lancet Digital Health* 2, no. 11 (November 2020): e622–28. https://doi.org/10.1016/S2589-7500(20)30193-X.

Kishore, Nishant, Aimee R Taylor, Pierre E Jacob, Navin Vembar, Ted Cohen, Caroline O Buckee, and Nicolas A Menzies. “Evaluating the Reliability of Mobility Metrics from Aggregated Mobile Phone Data as Proxies for SARS-CoV-2 Transmission in the USA: A Population-Based Study.” *The Lancet Digital Health*, November 2021, S2589750021002144. https://doi.org/10.1016/S2589-7500(21)00214-4.

Kone, Zovanga L., et al. "Internal borders and migration in India." *Journal of Economic Geography* 18.4 (2018): 729-759.

Kraemer, Moritz UG, et al. "The effect of human mobility and control measures on the COVID-19 epidemic in China." *Science* 368.6490 (2020): 493-497.

Li, Y., Rama M., Galdo V, & Pinto, M. F. (2016) A Spatial Database for South Asia. World Bank, Washington, DC.

Milusheva, S., Bjorkegren, D. & Viotti, L., 2021, June. Assessing Bias in Smartphone Mobility Estimates in Low Income Countries. In *ACM SIGCAS Conference on Computing and Sustainable Societies* (pp. 364-378).

Milusheva, Sveta, Anat Lewin, Tania Begazo Gomez, Dunstan Matekenya, and Kyla Reid. “Challenges and Opportunities in Accessing Mobile Phone Data for COVID-19 Response in Developing Countries.” *Data & Policy* 3 (2021): e20. https://doi.org/10.1017/dap.2021.10.

Munshi, Kaivan, and Mark Rosenzweig. "Networks and misallocation: Insurance, migration, and the rural-urban wage gap." *American Economic Review* 106.1 (2016): 46-98.

Silver, Laura, and Kyle Taylor. “Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally.” Pew Research Center, February 5, 2019. https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/.

Steele, Jessica E., et al. "Mapping poverty using mobile phone and satellite data." *Journal of The Royal Society Interface* 14.127 (2017): 20160690.

Szocska, Miklos, Peter Pollner, Istvan Schiszler, Tamas Joo, Tamas Palicz, Martin McKee, Aron Asztalos, et al. “Countrywide Population Movement Monitoring Using Mobile Devices Generated (Big) Data during the COVID-19 Crisis.” *Scientific Reports* 11, no. 1 (December 2021): 5943. https://doi.org/10.1038/s41598-021-81873-6.

Tisdell, Clement, et al. "Agricultural diversity and sustainability: general features and Bangladeshi illustrations." *Sustainability* 11.21 (2019): 6004.

United Nations. Statistical Office. *Principles and recommendations for a vital statistics system*. No. 19. United Nations Publications, 2014.

Wesolowski, Amy, et al. "The impact of biases in mobile phone ownership on estimates of human mobility." *Journal of the Royal Society Interface* 10.81 (2013): 20120986.

Wilson, Robin, et al. "Rapid and near real-time assessments of population displacement using mobile phone data following disasters: The 2015 Nepal earthquake." *PLoS currents* 8 (2016).

Acknowledgements

Pablo Tillan and Joaquin Endera assisted with basic data preparation from HIES, Census, and Economic Census in Bangladesh. The entire Bangladesh Poverty team provided crucial review and feedback on earlier versions of this analysis.

Annex I: Reproducing this work

All of the code used to develop this study is currently shared on Github at https://github.com/Suraj1127/west-bengal/. This includes code used to aggregate and measure net movement as well as code to prepare charts and analysis thereof. This draft repository has been reviewed by the Development Data Partnership for any potential data privacy issues or misrepresentations of providers' data / methods) and cleared. Some code used for chart production is still outstanding and will be shared before the public release of the final codebase on the World Bank's own Github organization (https://github.com/worldbank). The repository's README includes a longer discussion of the skills, technology, and approaches used to create this data and recommendations for those looking to replicate this work. This extends the discussion in Annex II below.

In addition to our code, we strongly recommend that users look into the mobility4resilience toolkit developed by the World Bank's Global Facility for Disaster Risk and Reduction (GFDRR) (https://github.com/GFDRR/mobility_analysis) concurrently to this research. Their tools are built on relatively easy-to-adopt stack of Python + Dask and come with many more prepared analysis functions beyond net movement. They are thus likely to be more compatible with many data scientists' skills, budgets, and use cases.

Annex II: Big data and its implications

Beyond its extraordinary temporal and spatial specificity, the dominant characteristic of smartphone mobility data as a dataset is its size: records for a single Indian state for a single day can measure in the tens of millions and processed annual data can approach a terabyte in size. The processing requirements for such datasets can well exceed the available memory and storage of any desktop computer or server using standard processing environments like STATA, R, or Python, or spatial equivalents. Properly resourcing even an exploratory analysis is therefore complicated; to ease analysis by other teams we have described at a high level the considerations involved.

Because of their size, analyzing smartphone mobility datasets requires the use of distributed processing tools and computing clusters. Distributed computing is a computer science technique which uses multiple computers in tandem to solve a computational problem divided up between them. Distributed computing requires the use of specialized frameworks and coding techniques to efficiently divide up tasks and organize them amongst many computers; the difference in processing time between well- and poorly-organized distributed computing tasks is often exponential. Common frameworks include Spark, Hadoop, and Dask, common cloud-based computing "clusters" include Databricks or Coiled, and common storage facilities include Microsoft Azure and Amazon Web Services. Actual analysis code may be written in Scala, PySpark, Spark SQL, Python, R, or other specialized languages, with recourse to Java for troubleshooting errors.

Although distributed computing has become increasingly central to many companies' and governments' analysis capabilities as the availability of "big data" has grown, many distributed spatial analysis capabilities remain underdeveloped and/or poorly documented. The answers to even moderately difficult questions may only be findable through tedious viewing of conference presentations or contact with the actual maintainers of libraries. In some cases, they may not yet be implemented in a library at all. This is problematic because spatial operations are by their nature computationally "expensive" and efficiency is

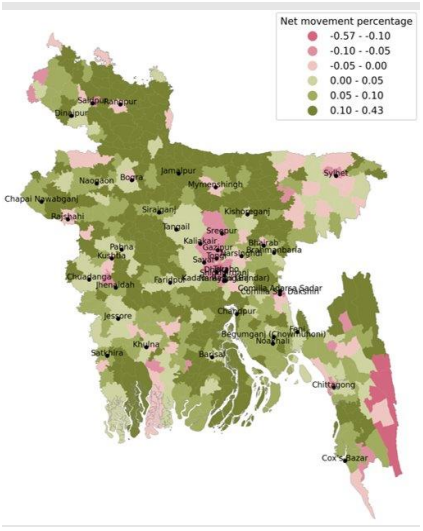
therefore of paramount importance when implementing them at large scales. To give a representative example, table joining 1 million rows of data in STATA or R might take a few minutes while spatial joining 1 million rows of data will crash most desktop computers.

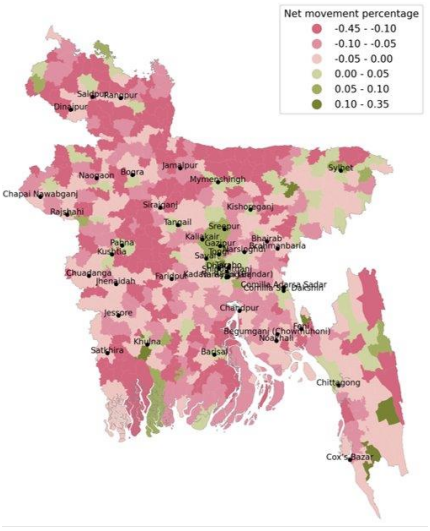
As a result of the above considerations, distributed computing imposes a much greater outlay in terms of human, hardware, and financial resources than standard analysis of survey and geospatial data. In terms of human resources, standard analysts may struggle to implement distributed processing code, even if they are familiar working in a code-based environment, as a decent understanding of computer science fundamentals and more specialized constructs is required. Multiple powerful computers must be accessed and set up to implement distributed computing; in most cases, the more economical solution will be to rent (by the computer-hour) pre-arranged computing clusters from a provider like Databricks or Coiled. Storage space (a "data lake") must be rented to house and efficiently serve intermediate and finalized datasets.

The above aspects of data processing are expensive in terms of staff and server time: for our analysis, around 3.5 months in staff time and $9,000 in server costs were expended to process and analyze approximately two years of data for India and Bangladesh. However, the fixed costs of mastering distributed computing basics, setting up processing code and a data engineering pipeline, and puzzling through the appropriate analysis methodology were our principal outlays. With these fixed costs accounted for, and the requisite skills and codebases developed, we anticipate much lower costs for future analyses. Furthermore, we anticipate that coding equivalent analysis in Python using Dask with Coiled or implementing more efficient spatial analysis routines based on Uber's H3 would speed up computing and reduce server rental costs.

Annex III: Additional tests and possible extensions

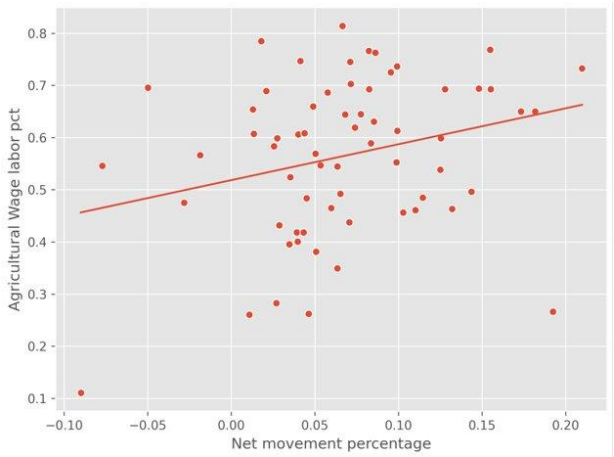
To further examine use-cases of smartphone mobility data in the region, we conducted additional tests on the spatial granularity of estimates, and their correlation with other related labor market indicators from secondary data. When explored at a smaller spatial level in Bangladesh, these patterns hold, but inconsistently, suggesting the instability of mobility estimates in small rural subnational units due to small sample sizes.





Agricultural wage rates and mechanization

We next look at whether correlations of migration, measured by net movement rates, moves with other secondary variables in predictable directions. Specifically, we look at the share of agricultural wage labor and anticipated agricultural productivity in districts of Bangladesh. A modest positive correlation was observed between agricultural wage shares (Bangladesh Household Income and Expenditure Survey, 2016-17) and net movement in Bangladesh. However, positive relationship notwithstanding, the large variance around the mean value suggests caution in interpreting this value too strongly.



Aman "exposure"

We compared net movement rates pre-Aman plantation to a Likert Scale measure of each upazila's generic anticipated productivity during Aman, Boro, and Aus seasons prepared by the Bangladesh Agricultural Research Council (BARC), with values of 1 indicating high productivity and 4 low. Pre-Aman plantation rates were chosen as they provided the strongest signal in the above maps. Despite this, a comparison of mean net movement rates to Aman exposure showed almost no different in mean net movement between strongly and lightly exposed upazilas. This suggests caution in linking movement to seasonal migration too strongly without further investigation of alternative push/pull factors.