

Cities On the Move

How analysis of human movement patterns can inform resilient, inclusive and sustainable cities

A Technical Guidance Note

Nicholas Jones, Takahiro Yabe, Samuel Heroy

Acknowledgements

Introduction

1. Mobility data: types, limitations, availability

Mobility data: three key categories

Data availability: Mobile phone use across countries

Data adequacy: How much is enough?

2. Analyzing Mobility Datasets: A Roadmap

Key concepts and terminology

Methods road-map

- Acquire and pre-process

- Assess data quality

- Filter by user

- Impute home and work locations

- Test representativeness

- Analyze commute patterns

- Point-of-interest analysis

3. Applied use cases

Disaster response: Displacement after the 2019 Puebla Earthquake

Resilient Transportation: Accessibility impacts of Cyclone Nivar in Chennai

Who uses city assets? Inequality and usage patterns of parks

Rapid displacement mapping: Ukrainian refugees in Poland

Covid-19 response: Monitoring adherence to stay-at-home orders

Urban regeneration: Accessibility constraints and local economic growth

4. Home and work locations: comparing spatial form of world cities

5. Selected bibliography

Acknowledgements

This Guidance Note was developed by a core team comprising Nicholas Jones, Takahiro Yabe, Maham Khan, Sam Heroy, Chris Oates and Oliver Lock (Global Facility for Disaster Reduction and Recovery - GFDRR) and Enrico Ubaldi (Mind Earth). Samuel Fraiberger (DEC); Satish Ukkusuri (Purdue University); Alessandra Feliciotti and Emanuele Strano (Mind Earth); and Andrew Schroeder and Shenyue Jia (CrisisReady) contributed content.

The methods and use cases presented in this note were refined in collaboration with World Bank task teams from the Global Practice for Urban, Disaster Risk Management, Resilience and Land (GPURL) and Transport Global Practice. The project team thanks Diana Tello Medina, Oscar Edmundo Diaz and Bernadette Gordyn (Costa Rica), Jonas Ingemann Parby and Jessica Schmidt (Nepal), Tatiana Peralta and Gerald Olliver (Chennai), and Reena Badiani-Magnusson and Tom Bundervoet (Ukraine). The team benefited from technical discussions with Sam Fraiberger, Paolo Avner, Dunstan Matekenya and Sveta Milusheva as well as guidance and support from Nancy Lozano-Gracia (Lead Economist) and Niels Holm-Nielsen (GFDRR Practice Manager).

The team wishes to thank the Development Data Partnership, notably Holly Krambeck and Gabriel Vicente, whose work facilitates secure, efficient and privacy-protected access to datasets discussed in this note. Among private sector partners, the team particularly acknowledges Alex Pompe and Eugenia Giraudy (Data for Good and Core Data Science teams at Meta), Mikel Maron (Mapbox), Tim Solt (Quadrant) and Brennan Lake (Cuebiq).

The Guidance Note will be finalized following peer review to which Jun Rentschler, Megha Mukim, Tatiana Peralta and Ben Stewart kindly agreed to participate. The team gratefully acknowledges funding from GFDRR's USAID Single Donor Trust Fund as well as earlier funding provided by the Spanish Fund for Latin America and the Caribbean under the Disruptive Technologies for Development program.

Introduction

Movement patterns are key for vibrant and resilient cities

Mobility is at the heart of life in cities. Transport planners focus on connecting people with jobs, amenities and essential services smoothly and at low cost. Disaster risk management planners aim to ensure orderly evacuations if a storm is approaching. And urban economists know that mobility is key for growth and poverty reduction. Cities act as motors of economic growth by connecting people together: workers move to cities to find jobs that are a good match for their talents, and firms prosper by connecting with suppliers and customers that demand specialized goods and services.

But despite the importance of urban mobility for urban planning and disaster risk management, the movement of people around cities has historically been difficult to observe. Getting such data has typically required costly and time-consuming surveys in which households are interviewed – face-to-face or by telephone – about their movement patterns. This Guidance Note addresses a new opportunity for urban planners and disaster risk management professionals, namely the rapidly increasing availability of large datasets that describe actual observed mobility patterns of citizens in major cities of the world.

Mobility data analysis: established methods, widening applicability

The first mobile phone was demonstrated in 1973. Since being launched to the public in the 1980s, growth has been exponential, with more than 15 billion mobile devices operating globally in early 2022.

Since the late 1980s, academics pioneered an extensive scientific literature examining human movement patterns based on cellphone data, building on related fields of physics and ecology. In January 2010, the practical value of such data for disaster response became apparent in the tragic context of the Haiti earthquake. Nearly one-quarter of the population of Haiti's capital Port-au-Prince, left the city after a 7.0 magnitude earthquake levelled more than 250,000 buildings. Flowminder, a UK-based research group, used local cellphone information to estimate where people had gone; information that proved invaluable to target humanitarian supplies across towns and villages facing a population influx.

The data used in Haiti earthquake response were Call Detail Records, or CDR, a type of information generated by mobile network operators for billing purposes. Intensive efforts to widen the use of such methods have paid off in a range of countries; yet CDR remain intrinsically difficult for most disaster risk management and urban planning stakeholders to access, and their suitability for fine-grained urban analyses is limited.

More recently, availability of mobile phone location data has entered a new phase due to the growth of smartphones. Large quantities of GPS-accuracy location data are generated every day through Android and iOS phones. These devices are carried by more than 3 billion people globally, a number which continues to increase. Datasets depicting human movement patterns based on smartphone measurements can be accessed by a wider range of stakeholders than before, presenting important opportunities to enhance understanding of how people move within cities in ordinary times and during natural disasters.

Why is this Guidance Note needed?

In the authors' experience, many urban and disaster risk management practitioners appreciate the potential value of large-scale observational data on human movements to provide valuable insights on making cities more resilient, inclusive and sustainable. However, getting started with such datasets can be a struggle:

- **Data barriers.** Lack of guidance about data types and coverage rates make it hard to know which data is promising for a given need – even when such a dataset is readily available.
- **Computational barriers.** Until recently, the lack of open-source tools has led to costly duplication as teams write one-off code to handle the same basic procedures.
- **Methods barriers.** Several challenges come as standard with human mobility datasets – handling large data volumes, testing representativeness, extracting commuting and visit patterns – but lack of familiarity with the steps involved discourages urban specialists from using them.

This Technical Guidance Note is intended to provide readers with a better understanding of smartphone-derived human mobility datasets, the issues associated with using them to inform resilient urban planning, their caveats and limitations, and the potential value they offer¹.

Who is this Guidance Note for?

This note addresses two main audiences. The primary audience is urban planning and disaster risk management professionals who are interested to enrich their work based on insights from human mobility data. The note aims to answer commonly recurring questions: am I likely to find a good-quality mobility dataset for my region of interest; how do I check its robustness; what technical procedures are needed to develop trustworthy conclusions; and what could it show me?

For a second audience, data scientists who expect to work directly with such datasets, the note provides a series of boxes marked ‘Hands on Analysis’. Using short code snippets, these illustrate a typical workflow to clean, inspect, filter and analyze a human mobility dataset. The boxes aim to ‘demystify’ such analysis by demonstrating well-established technical methods as applied to today’s urban challenges. The code snippets are based on Mobilkit, an open-source library developed by the project team.

Looking ahead: from ‘frontier’ to mainstream

Human mobility datasets are one of today’s ‘frontier datasets’. Just ten years ago, satellite imagery represented a new frontier of technical possibilities for urban planners and disaster risk management specialists, complete with a new and unfamiliar set of technical concepts and a scientific literature that was not widely disseminated beyond the research community. Today, urban development specialists increasingly have key concepts of remote sensing at their fingertips and consider satellite imagery a mainstream resource for their work.

Arguably, human mobility data is undergoing a similar journey from technical frontier to mainstream part of urban development and transport planning practice. Indeed, several use cases profiled here show potential for wider adoption. Mapping the density of foot traffic visitors in public plazas (see Spotlight 2 – Place Magnetism) could be used to contrast ‘before’ and ‘after’ following urban regeneration or public transport investments. In the context of natural hazards, rapid post-disaster mapping of displaced populations is increasingly viable (see Case Study 4) offering enhanced potential to target response and recovery resources.

The landscape of mobility data can be expected to change in coming years amid a dynamic regulatory and technological context, but as smartphone penetration grows, data depicting urban movement patterns will become an increasingly salient resource. Urban planning and disaster risk management can benefit from new insights derived from this data.

¹ Readers intending to work with Call Detail Records should refer to guidance material at <https://worldbank.github.io/COVID19-WBG-Mobility-Task-Force/> and www.flowminder.org.

1. Mobility data: types, limitations, availability

Understanding how and where people move is important to help make cities more resilient, inclusive and sustainable. Targeting relief after a humanitarian crisis requires knowing where people have gone. Design of a more equitable city may benefit from knowing whether certain neighborhoods or spaces experience social and economic segregation. Improving transportation systems requires knowledge of commuting patterns and trip demand.

In each case, human mobility data has proven value to help policymakers take informed decisions. It can be particularly valuable in low- and middle-income settings where official datasets are less comprehensive. However, to make use of it, urban planners and disaster risk management practitioners need to navigate several questions and uncertainties. What datasets are available for my region of interest? Are they sufficiently detailed and reliable? How to address privacy and security concerns?

This section discusses the types of mobile phone location data; how data density and representativeness vary across world regions; how such data should be contextualized alongside other urban sources; and key limitations and caveats when analyzing such a dataset.

Mobility data: three categories

Mobile phone location data can be classified into three main categories: mobile phone Call Detail Records (CDR), smartphone GPS location data collected by location intelligence companies, and smartphone GPS data collected and processed by major tech firms. Table 1 presents these three forms of data together with their collection methods, pros and cons, and access considerations.

Table 1. Mobility data types

Data type	Description	Pros and Cons	Access considerations and example providers
Call Detail Records	Primarily collected for billing purposes, this dataset provides the locations of the nearest cellphone tower each time a user places a call or sends a text.	(+) substantial coverage of population including non-smartphone users (-) lower spatial and temporal resolution	Case-by-case legal agreement typically required with Mobile Network Operator. Digicell, Vodafone, Orange
Smartphone GPS location data (location intelligence firms)	This dataset compiles the locations of individual users at GPS accuracy based on movements recorded by third-party smartphone apps.	(+) precise location information of users (-) limited transparency of data generation process; lower population coverage rate than CDR	Available commercially or via data sharing programs such as Development Data Partnership subject to privacy and data ethics requirements. Veraset, Cuebiq, Unacast, Quadrant, Outlogic
Derived data products (major tech firms)	A dataset provided by a technology firm that depicts movement patterns of users in simplified form based on smartphone GPS measurements	(+) available in standardized form across multiple countries and time periods	Available to general public, select researchers or commercial clients (varies by dataset).

	generated through the firm's software.	(-) Outputs restricted to select metrics produced by the tech firms	Example datasets: Facebook Disaster Maps, Google Community Mobility Reports, Mapbox Movement Data
--	--	---	---

(i) Call detail records

Call Detail Records (CDR) are datasets that mobile phone network operators create primarily for billing purposes. They record key information about phone calls and text messages. Information on people's movements can be derived from these datasets because they include the location of the cellphone tower that was used to route each call or text. For each network event – ie. a call or text – the CDR dataset records the time, the location of the nearest cellphone tower, the user's unique ID number, and additional information such as length of phone call and approximate location of its recipient. A key advantage of CDR datasets is their high coverage rate where regular cellphones (as distinct from smartphones) are a high proportion of devices. This makes them uniquely valuable particularly in low-income settings. Compared with GPS-based data from smartphones, the spatial resolution of CDR datasets is lower and varies depending on population density. Locations are derived from the catchment area of the nearest cellphone tower, which may be densely positioned in cities but extend to hundreds of square kilometers in rural areas.

(ii) Smartphone GPS location data

Location datasets derived through GPS devices on smartphones provide highly detailed information on human movement patterns. Smartphones are defined as phones that perform many functions of a computer, typically having a touchscreen and an operating system capable of running apps². Apps utilize the GPS chips on smartphones to offer a range of functions: providing weather forecasts for the user's current location, showing local maps, hailing ride sharing vehicles, and providing locally relevant adverts. Human mobility dataset collected by such apps is aggregated by a range of companies ('location intelligence firms'). Such firms apply a range of anonymization and privacy-protection processing steps ahead of utilizing them for commercial purposes (such as evaluating the impact of advertising campaigns) and making them available to researchers under 'data for good' programs. Typical lines of a GPS location dataset show user ID number, timestamp, and longitude and latitude. Compared to CDRs, GPS logs have higher spatial precision and higher frequency of observations. However, the process through which data are collected and combined from several applications often remain undisclosed, making rigorous analysis of representativeness important.

(iii) Derived data products from tech firms

Similar to the smartphone GPS location data collected by location intelligence firms, major tech firms such as Facebook, Google, and Apple also collect GPS location data from their users. The major difference in the data generative process is that these firms use data collected from their own platform, not by third party services. Often, these data are provided in a pre-processed form, aggregated by both time and space. Meta, through its Data for Good program, provides various types of location information products to researchers, agencies, and non-profits.³ Facebook Disaster Maps provides detailed density maps of the population density and movement

² The two main operating systems for smartphones are Android and iOS. Annual shipments of phones using these operating systems increased from 1.2 billion to 14 billion per year from 2007 to 2021, according to market research firm Gartner.

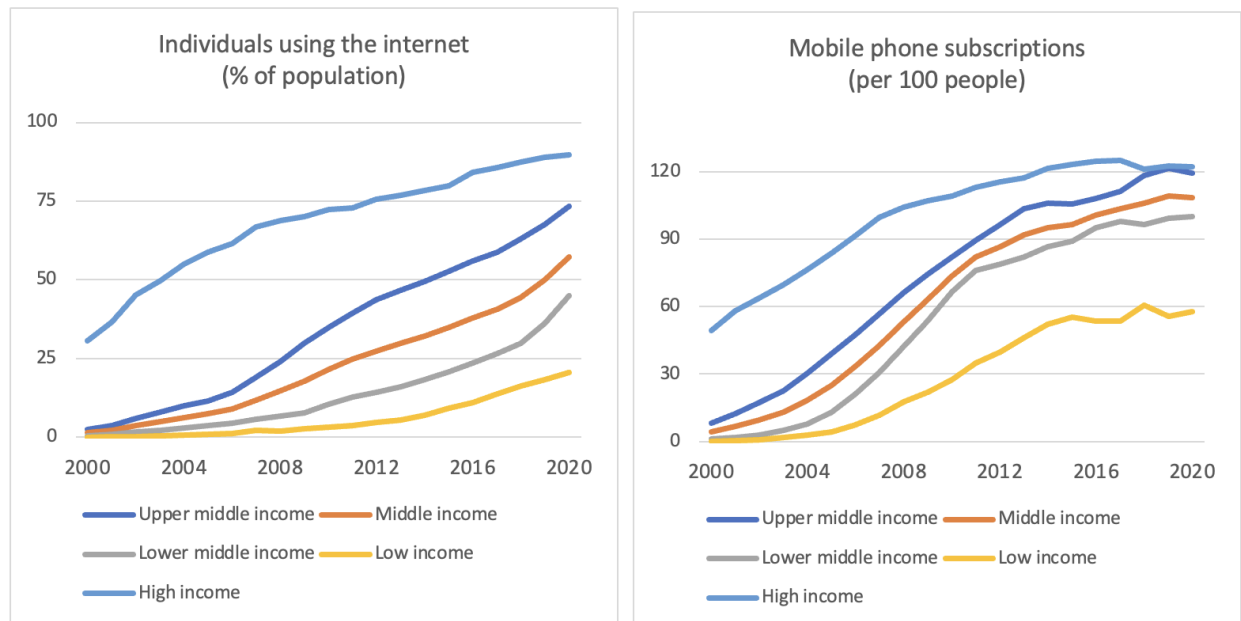
³ <https://dataforgood.meta.com>

patterns before, during, and after disaster events. The data is temporally aggregated (usually every 24 hours), spatially aggregated (usually into 360,000 square meter tiles), and spatially smoothed, to anonymize and protect users' privacy. Other examples of pre-processed mobility data products include Google Community Mobility Reports and Mapbox Movement Data.

Data availability: mobile phone use across countries

Use of mobile phones increased sharply from 2000 to 2020 but remains uneven between rich and poor countries. Mobile phone subscriptions became ubiquitous in upper middle-income countries and the number of mobile phone subscriptions exceeded 90% in lower-middle income countries by 2020. In low-income countries, mobile phone subscriptions grew substantially but remained at 57 per 100 people in 2020 (Figure 1).

Figure 1: Growth of mobile phone subscribers and internet users by country income group



Source: World Development Indicators

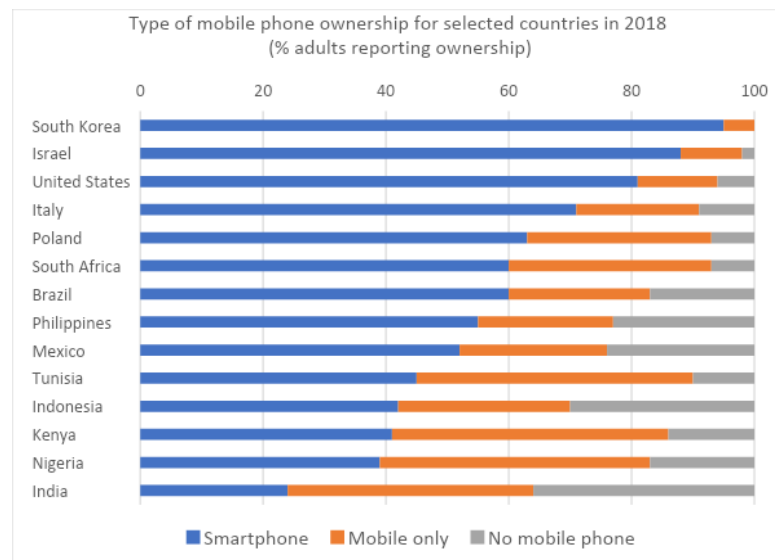
Access to smartphones remains marked by substantial inequality between and within countries. According to ITU statistics, globally 62% of men use the internet compared with 58% of women. However, in Least Developed Countries, the digital divide widens (31% of men compared with 17% of women). Gender differentials are more acute than the world average in Africa and in the Arab States, according to International Telecommunications Union (ITU) figures. Urban–rural divides are another major inequality where internet access is concerned. In Least Developed Countries, an estimated 47% of urban dwellers use the internet compared with 13% of people in rural areas. Lastly, a generational divide affects smartphone use. Globally, 71% of the world's population aged 15-24 is using the internet, compared with 57% of other age groups⁴.

⁴ <https://www.itu.int/en/mediacentre/Pages/PR-2021-11-29-FactsFigures.aspx>

Survey data shows considerable variation in reported ownership rates of smartphones as compared with regular cellphones (Figure 2). A World Bank study comparing smartphone and CDR data in Uganda found that that trips to the capital represents 10% of smartphone user trips, but only 4% of basic user trips, and that smartphone users changed their mobility patterns more than did basic users during Covid-19 lockdowns⁵. In high-income contexts, metrics from human mobility datasets used to inform Covid-19 response have been shown to under-represent elderly and non-white citizens⁶.

In general, the biases inherent in human mobility datasets are more acute in lower-income countries. As such, CDR is particularly advantageous in low-income settings. CDR data is particularly advantageous in low-income settings given this data source encompasses users of basic cell phones as well as smartphones. Smartphone mobility datasets become increasingly suitable as national income (and hence smartphone ownership) rises. In lower-income settings, additional care to quantify and mitigate biases – including gender, generational and urban-rural disparities – is essential.

Figure 2: Mobile phone ownership by type



Source: Pew Research, Global Attitudes Survey 2018

⁵ Sveta Milusheva, Daniel Björkegren, Leonardo Viotti. October 18, 2021 (blog). “Can we trust smartphone mobility estimates in low-income countries?” World Bank Data Blog.

⁶ Amanda Coston, Neel Guha, Derek Ouyang, Lisa Lu, Alexandra Chouldechova, and Daniel E. Ho. 2021. “Leveraging Administrative Data for Bias Audits: Assessing Disparate Coverage with Mobility Data for COVID-19 Policy.” *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FACCT '21)*

Data adequacy: How much is enough?

The use of surveys to understand mobility patterns has a rich history, and important ongoing role, in urban transport planning. Indeed, as state investment in national highways and urban mass transport system increased globally in the 1950s and 1960s, transport demand surveys proved essential tools to understand demand for trips; change in mode share between walking, private automobile use and public transport; and the distribution of trips by purpose, length and population segment⁷.

Comparing location data with transport demand surveys

Transport demand surveys provide an important evidence base to underpin transportation budgets comprising capital and maintenance expenditure for transportation assets. One study of sample size requirements cites 1% as a statistically adequate sample size for transport demand estimation in large urban areas, but notes that 10% of the population may need to be surveyed in small municipal areas⁸. To understand trip behavior disaggregated by purpose, time and mode in complex cities such as Toronto, Calgary or Sydney, the study recommends a sample size of 4%.

Although transportation demand survey data and mobile phone location data both offer valuable insights on urban mobility patterns, they differ fundamentally in their mode of production. Transportation demand surveys typically adopt a random stratified sampling approach to ensure sufficient representation of population cohorts (gender, age, income, vehicle ownership, and occupation). By contrast, mobile phone datasets are ‘convenience samples’: inclusion in the dataset is based on arbitrary factors such as whether an individual downloaded a specific app and kept it on their phone.

Transport demand surveys are costly, locally customized instruments developed to provide transport planners with required insights. Mobile phone movement datasets are produced for commercial reasons entirely independent of their utility for urban planning purposes. A key advantage of local transportation surveys is their inclusion of demographic characteristics alongside information on transportation needs and preferences. Transportation surveys also provide valuable information on user perceptions and satisfaction. By contrast, mobile phone location datasets provide raw information on observed movements with no information about individuals’ demographic characteristics.

However, mobile phone mobility datasets offer important advantages alongside local survey instruments. Trip diaries of household members have been collected since the 1980s as part of travel demand surveys, but despite continuous efforts to boost survey quality, issues such as missing trips and low response rates persist. Biases reported in US travel surveys include under-reported short trips, inaccurate travel times and inaccurate travel distances⁹. By contrast, mobile phone location datasets offer mobility information in unprecedented detail. Such information has a large sample size (often millions of daily observations), at a high frequency (eg. 50 daily data points per user), at high spatial granularity (~10 meters accuracy), for a longitudinal time frame (eg. a user panel observed over 1 year time period).

⁷ John R. Meyer, *The Urban Transportation Problem*, Harvard University Press, 1965

⁸ Khandker Nurul Habib, Wafic El-Assi, Tian Lin. 2016. “How Large is too Large? A Review of the Issues related to Sample Size Requirements of Regional Household Travel Surveys with a Case Study on the Greater Toronto and Hamilton Area (GTHA)”, Presented to the TRB Annual Meeting, Washington DC, <https://arxiv.org/pdf/2005.00563.pdf>.

⁹ Stopher, Peter and Greaces, Stephen. 2007. “Household Travel Surveys: Where are we going?” *Transportation Research Part A*.

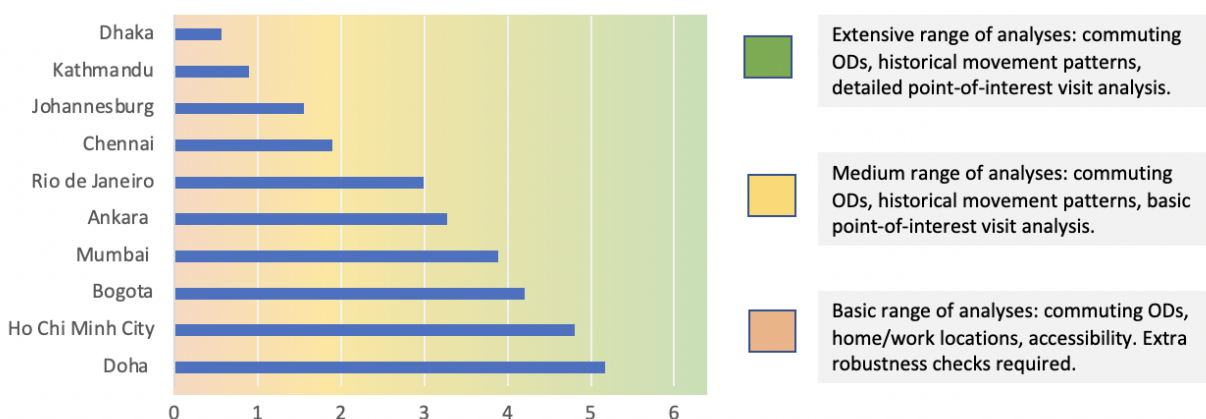
As observational data, reporting biases due to interviewees' limited recall of their trip behavior are not present. In addition, smartphone mobility datasets can provide some insights that a systematic program of transport demand surveys might offer but in low-income or data-poor urban environments where such programs do not currently exist.

Data adequacy: What sample size is required for urban planning studies?

Since smartphone mobility datasets are proprietary, a systematic comparison of sample size across cities is not possible. However, Figure 3 presents reported sample size for selected GPS datasets based on an ad hoc survey of analysts. The figures are aggregated from multiple researchers and presented together with 'rule of thumb' guidance about sample size adequacy for a range of urban planning applications¹⁰.

A key metric is the number of Monthly Average Users (MAUs) in a human mobility dataset. A basic range of analyses is possible even with MAU of between 0.5% to 1.5% of the local population, albeit with a high degree of care required to mitigate against potential biases such as under-representation of low-income people. With higher sample sizes, a wider range of analyses becomes possible and bias concerns may lessen.

Figure 3: Data density for selected GPS datasets and recommended applications suitable for each level of sample size (Monthly Average Users as % local population)



Source: Project team

¹⁰ Reported sample size is based on an ad hoc survey of researchers who worked with GPS datasets during 2021-22. Actual size of human mobility datasets will differ based on data provider and extraneous factors such as operating system changes and trends in app usage. The figures are for illustrative purposes to show order-of-magnitude dataset sizes at a specific point in time.

2. Analyzing Mobility Datasets: A Roadmap

Having confirmed the benefits of analyzing human mobility datasets to inform an urban planning or disaster risk management project, what are the first steps to get started? There is no universal recipe to follow. However, large mobility datasets have basic features in common. They frequently present similar challenges including the need to grapple with large volumes of data and the need to identify and mitigate biases.

This section first presents key concepts and terminology that are widely used in mobile phone data analysis in its raw form. It then presents a practical ‘road map’, comprising seven steps, for conducting an urban planning analysis with human mobility data.

Key concepts and terminology

Mobile phone dataset: stylized description

At their most simple, human mobility datasets comprise data tables showing locations of individual users at points in time. Figure 4 presents a stylized representation of such a dataset, with columns showing time, longitude and latitude, and user ID numbers.

Importantly, mobile phone location datasets are invariably provided with no demographic information whatsoever. User ID numbers facilitate analyzing changes in individuals’ movement patterns over time, but provide no way to link their movements with other personal information. A range of additional anonymization and privacy protection measures are typically provided depending on the data provider.

Figure 4: Stylized example of a human mobility dataset

Timestamp	Longitude	Latitude	User ID
2020-01-23 01:48:49	38.90041	-77.04408	00001
2020-01-23 01:50:55	38.90112	-77.04574	00001
2020-01-23 01:53:06	38.90188	-77.04762	00002
2020-01-23 01:54:09	38.90121	-77.04573	00001

Privacy safeguards and reidentification risk

Mobile phone datasets contain personal information about individuals and handling them thus require the highest level of care. Researchers and analysts should understand the potential harms that can arise from misuse of such datasets and ensure robust procedures to mitigate these risks ahead of time.

Human mobility data is typically provided in anonymized form similar to that depicted in Table 4. Indeed, ahead of acquiring such a dataset, analysts should confirm that the data provider will share only de-identified data. This means that any identifier or attribute that can link movement information with people’s identities has been removed.

However, de-identified human mobility datasets can still result in privacy breaches when misused. The ‘mosaic effect’ describes harms that can arise when datasets are layered with other complementary data sources,

thereby disclosing sensitive information (such as personal identities and place visits) that would otherwise not be revealed. The risk of privacy breaches through combination of datasets cannot be fully mitigated through anonymization techniques. As such, mobility datasets must be considered sensitive and handled accordingly. This implies the need for data to be stored and analyzed in a secure framework limiting access to authorized individuals only.

Data sharing agreements establish conditions for data providers to share information with researchers while ensuring respect for privacy rights. Such agreements should be constantly reviewed in light of potential risk to the privacy and security of individuals and in light of applicable laws and regulations.

Following principles embedded in the EU's General Data Protection Regulation, at a minimum such datasets should be handled in line with the following:

- **Purpose limitation:** Only utilize datasets for study purposes identified ahead of time.
- **Data minimization:** Only acquire the minimum amount of data needed to fulfill the study's objectives.
- **Storage limitation:** Destroy the data after the study is concluded.
- **Integrity and confidentiality:** Undertake best efforts to prevent data breaches.

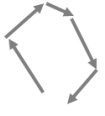
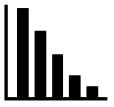
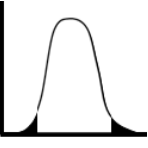
Key concepts for mobility data analysis

Like other technical fields that urban planners now consider mainstream to their work - an example being the use of satellite imagery - mobility data analysis comes with a set of terminology and concepts.

The key concepts can be collected into three main groups:

- **Individual mobility metrics.** Research into human or animal movements has developed a number of standard indicators to describe the characteristics of movement patterns. At the simplest level, Total Travel Distance summarizes how far an individual moves over a given period (eg. one day). Another widely used metric is the Radius of Gyration (RoG), which measures how far an individual typically travels from their home.
- **Collective mobility metrics.** At the level of a city's population, trip patterns can be summarized in the form of an Origin Destination matrix. At its simplest, an OD matrix derived from human movement data might summarize observed movements from users' home to work locations, providing insight into trip demand for commuting between neighborhoods in a city.
- **Algorithms and analysis steps.** Trajectory data analysis makes frequent use of clustering algorithms to identify key locations - including home and workplace - for any given user. The term 'significant locations' refers to places where individuals spend most of their time. To go from raw observations to information about time spent visiting places of interest - such as shops, public amenities or workplaces - stop detection can be performed. Lastly, since human mobility patterns follow strong statistical regularities, anomalies in population density can be identified, helping to assess how populations were affected by natural disasters.

Table 2: Key concepts in human mobility analysis

Icon	Metric / concept	Description
	Total travel distance (TTD)	Total distance traveled along an individual's movement trajectory. Total Travel Distance can be calculated for different time periods (eg. each day for one month; monthly average for the year). Changes in TTD have been used to measure the effectiveness of public health stay-home orders.
	Radius of gyration (RoG)	A measure of the characteristic distance traveled by individuals from the central point of their trajectories. Higher income and occupations linked with high mobility are associated with higher RoG.
	Significant locations	Locations that are regularly visited by the individual; typically includes home location, workplace, and favored leisure destinations. Research highlights strong statistical regularities in time spent at such locations: people spend most time at home, with time spent at the next most favored locations decreasing systematically.
	Home and workplace detection	Use of clustering algorithms to impute home location, defined as where a user spends most time at night, and work location, which may be defined as the location where most time is spent during working hours. Detecting home locations helped match raw movement data with socio-economic information considering the wealth of individuals' home neighborhoods.
	Stop detection and visit detection	Stop detection transforms an individual's raw set of location observations into a list of 'stops', defined as staying within a given radius for a given period of time. For visit detection, a stops database is combined with geographic information about Points of Interest (eg. locations of stores or parks). This makes it possible to estimate the number and duration of visits to such POIs.
	Origin-destination (OD) matrix	A table summarizing movement patterns in the form of number of trips between a set of origins and a set of destinations. For example, a city's commuting patterns may be summarized as an OD matrix representing typical number of weekday trips from census tracts where people live (origins) to the census tracts where they work (destinations).
	Anomaly detection	Use of statistical regularities in human movement patterns to identify unexpectedly large changes. Population density at given locations typically follows the normal distribution (ie. a bell-shaped statistical distribution). Plotting changes in population density against this statistical baseline on a map helps identify areas facing unusually large population inflows or outflows due to natural disasters.

Source: Authors

Introducing Mobilkit

Data scientists may use a wide variety of programming languages and code libraries to perform the type of analysis outlined here; however one available resource is the Mobilkit library, developed by Enrico Ubaldi¹¹. Mobilkit is a Python library that can be easily installed on any computer. It is open source and under continuous development. It allows a user to:

- Load raw mobility data and filter them by spatio-temporal queries;
- Filter users according to per user-ID (UID) statistics, eg. number of daily pings per user or total number of days an user has at least one ping recorded;
- Estimate home and work locations of each user;
- Estimate the land use of cells tessellating a given region of interest (ROI);
- Measure displacement after a major event or shifts in the visiting rates of specific Points Of Interest (POIs).

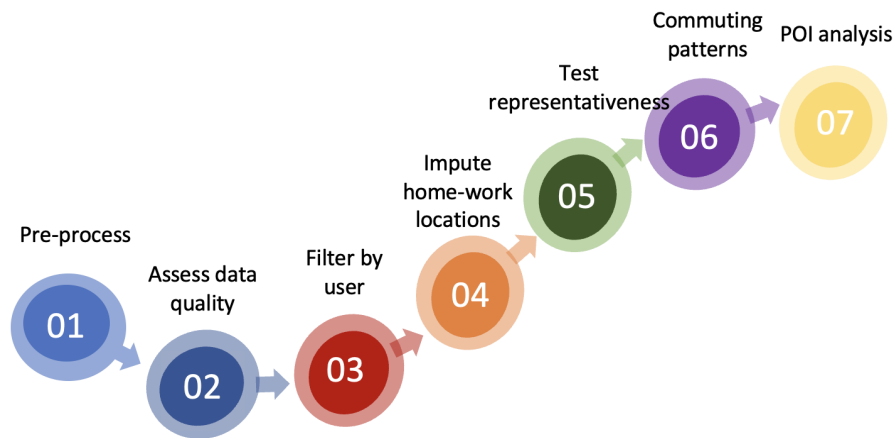
Mobilkit is built on the Dask distributed computing framework, thus enabling analysis of large datasets even on regular personal computers. Its modular structure and open-source licensing allows users to easily expand its functionalities or apply algorithms from other libraries (such as scikit-mobility) to their data.

¹¹ www.github.com/mindearth/mobilkit

Methods road map

This section presents a ‘roadmap’ for analyzing human mobility datasets to address urban and disaster risk management questions. It presents seven steps in a typical workflow that data scientists might follow. The first three steps address the crucial need to prepare a dataset for analysis, taking an informed view of the dataset’s quality and removing data that does not meet quality and relevance criteria. The next two steps link mobility traces with socio-economic data and test its representativeness with regard to the city’s population. The last two steps present two widely used analysis methods: constructing origin-destination matrices to represent journeys between home and work, and analyzing patterns of visits to points of interest (POIs)¹².

Figure 5: Key methodology steps



Source: Project team

¹² The examples in this section assume the use of smartphone GPS location data, as described in Table 1 and illustrated in Figure 4. Derived data products – such as those provided by technology firms Meta, Google, Apple and Mapbox – do not require the same processing steps as described here. However, similar processing steps are performed in producing such datasets, so analysts using them would do well to become familiar with the concepts presented.

1. Acquire and pre-process dataset

To acquire human mobility datasets, urban planners and disaster risk management professionals may wish to identify any data access partnerships to which they have access, research 'Data for Good' programs of data providers, or research derived data products available from technology firms (see Table 1 - Mobile Phone Data Types). The essential first step is to ensure a clear set of legal and operational safeguards to ensure all applicable regulations and best practices in data privacy and security are adhered to. Having done so, analysis can commence.

The size of human mobility datasets often poses a challenge. City-scale datasets meet a standard definition of the term 'big data', namely that they are too large to load into the memory of a standard computer. Efficient computing approaches - such as those utilized in the Mobilkit library - are therefore required.

To get started with a mobility dataset, first inspect a small sample of the data and perform basic steps:

- Remove observations with missing or malformed information (eg. missing timestamps or out-of-range coordinates).
- Keep only observations with the required spatial accuracy.
- Define the spatial and temporal boundaries for the research project.

The geographic boundary for the study can be defined using a geographic data format such as a Shape file. Initial pre-processing steps include cropping the dataset in time and space (ie. dropping observations that fall outside the time period and geographic area of interest - see Hands-on Analysis 1).

HANDS-ON ANALYSIS (1):

Mobility datasets are typically divided into individual comma separated value (CSV) files, or equivalent formats, each representing different periods of time.

Using the mobilkit.loader functions, the code snippet below creates an object ('dataframe') representing all of the CSV files in a folder. Using parallel processing, the program will then crop the data by removing all observations outside the period and geography of interest (the latter defined by a Shape file entitled 'city_outline.shp'). The resulting object ('df_filtered') is a cropped mobile phone dataset suitable for further processing.

load_and_filter.py

```
1 # load CSV files from the 'data' folder and crop by date
2 start_date = '2020-05-01'
3 end_date = '2020-11-01'
4
5 dataframe = mobilkit.loader.load_raw_files('data/sample*.csv',
6                                           start_date = start_date,
7                                           end_date = end_date)
8
9 # filter out observations outside area of interest
10 AOI = 'city_outline.shp'
11 df_filtered = mobilkit.spatial.tessellate(dataframe, tessellation_shp = AOI, filterAreas = True)
```

2. Assess data quality

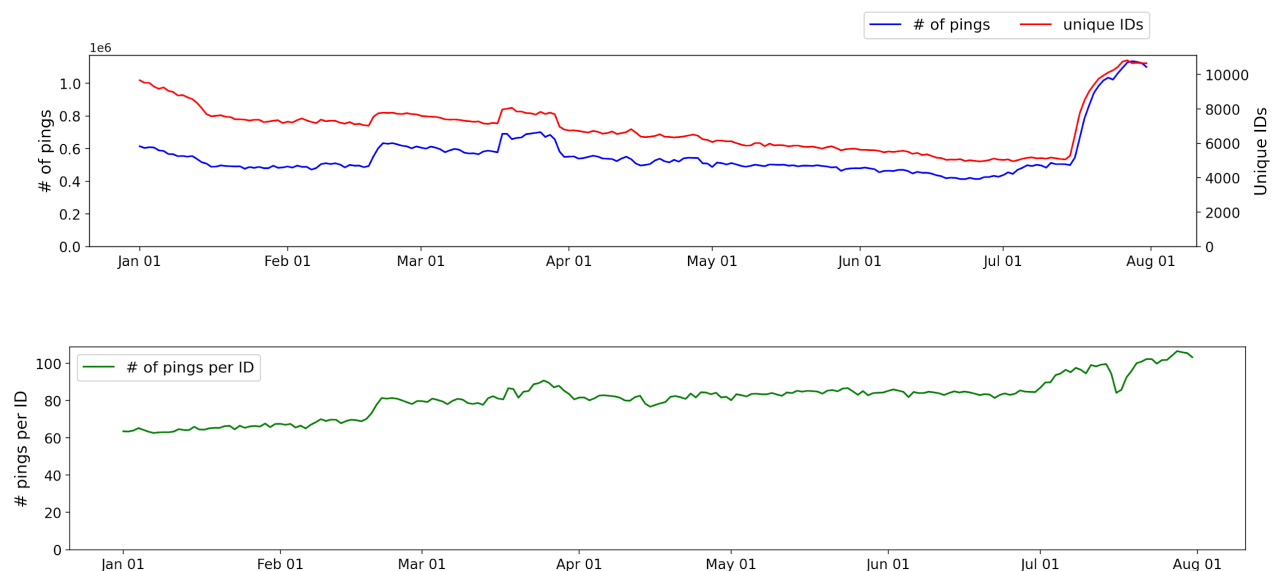
Considering potential data quality issues, it is important to acquire a thorough understanding of the dataset at an early stage. Two key issues are data density and user churn. The amount of data available for a specific city is hard to ascertain without taking a detailed look at the data itself. Data quantity has several dimensions: total number of data points (or ‘pings’), total number of users, number of daily active users (DAUs) or monthly active users (MAUs), and average number of pings per user.

In many cases, location intelligence firms aggregate movement information from many third-party smartphone apps. The set of apps whose data is included may change over time. Such changes can result in sudden increase or decrease in data quantity, as user data is added or dropped. This results in ‘churn’, ie. change over time in the observed panel of users. Assessing the degree of churn is important to establish an effective filtering strategy (see Step 3).

Figure 6 shows the daily number of unique users, and the number of pings, in a GPS dataset from a South Asian city covering January to August 2021. While the number of users fluctuates between 6,000 - 10,000 across the 7 months of observation (from January to July 2021), the number of daily pings ranges between 600,000 and 800,000 during most of the time. Both the signals feature a sudden increase during the last month. Moreover, the ratio between the daily number of pings and users evolves in time, starting from about 60 pings/user/day and increasing up to about 100 pings/user/day at the end of the observation period.

Such time series plots are a helpful step to diagnose user churn and data quantity, whether due to addition and removal of data source apps or user pattern seasonalities. Analysts may consider adjusting the time period of their analysis to benefit from periods where data is most consistent; and adjusting the ambitions of their study considering the number of users and data density per user. Where churn is identified as an issue, extra attention should be paid to filtering users (Step 3) to ensure a consistent observation panel over time.

Figure 6: Diagnosing data quality issues for a GPS dataset of a South Asian city



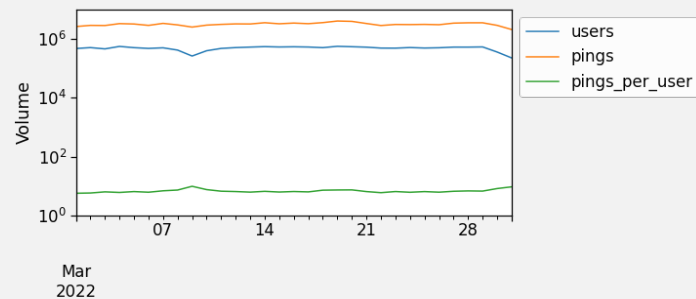
HANDS-ON ANALYSIS (2):

The data density and churn in a mobile phone location dataset can be understood through descriptive statistics. Key metrics include total number of users, number of pings, daily or monthly average users, and daily pings per user.

The program below inspects a dataframe and produces a plot of relevant statistics over time. In the following examples we will assume that we imported mobilkit with `import mobilkit as mk`.

data_quality.py

```
1 # generate a table showing the daily number of users and pings
2 df_volume = mk.computeTemporalProfile(df_filtered, what='both', frequency = '1d')
3
4 # plot a time series chart
5 df_volume.plot()
```



3. Filter dataset by user

Besides fluctuation in the number of users and pings, another source of heterogeneity in smartphone GPS datasets is the difference in amount of data between different users. Highly active smartphone users may have their location recorded many times per day; while infrequent users may show up infrequently, or inconsistently, in the dataset. For some analytical purposes – such as assessing changes in commuting patterns to inform transport investments – it is particularly important to have a stable panel dataset, including only users whose location is consistently reported with an adequate level of detail across the period of interest.

For these reasons, it is important to enforce a minimum set of requirements for a user to be kept in the dataset during the analysis. Filtering the dataset by user statistics achieves this purpose. For the example of a commute patterns analysis, we might wish to retain users that have enough pings per day to reconstruct their home-work journey patterns and are regularly active – ie. are present for a minimum percentage of days. Such filtering can also remove observations deemed irrelevant (eg. tourists who visited the region only briefly¹³.)

Quantity-quality trade-off

The dataset should be filtered to remove users who do not meet data quality thresholds. An inevitable trade-off must be confronted: filtering out more users will increase the data quality and relevance of the remaining observations; but leaving a smaller dataset to work with. The analyst may wish to set thresholds for the following variables:

- Minimum pings: the minimum number of data points ('pings') for a user to be retained.
- Minimum days spanned: the minimum number of days between the first active and last active day for a user to be retained.
- Minimum active days: the minimum number of days on which a user's location was observed for the user to be retained.

As an informed basis to select threshold values, it is helpful to visualize the amount of user data available with differing filtering criteria applied. The figure below shows a histogram in which darker colors represent more users. The ideal users from a data quality perspective are found in the upper-right quadrant; a rich record of pings and a long observation period is available for these users. Many users fall in the lower-left quadrant (few daily pings and short observation period). As these two selection thresholds increase towards high values, data quality increases but the number of selected users falls.

Threshold selection depends on use case

Considering the trade-off between data quality and quantity, the analyst can select a threshold (marked here with a red dotted line) and filter the dataset accordingly.

The decision on threshold selection should be made in view of the planned analysis methods. If interested to measure the density of foot traffic in an urban plaza, an appropriate choice may be to set the *minimum_pings* threshold to 100 and the *days_spanned* threshold to zero (as we do not care to exclude tourists or non-local residents from the analysis). If looking at commuting patterns over the course of a year, the analyst may be interested in users who are observed over a long period but be comfortable with fewer pings per day.

¹³ Scherrer, L., Tomko, M., Ranacher, P. & Weibel, R. Travelers or locals? Identifying meaningful sub-populations from human movement data in the absence of ground truth. EPJ Data Science 7, (2018).

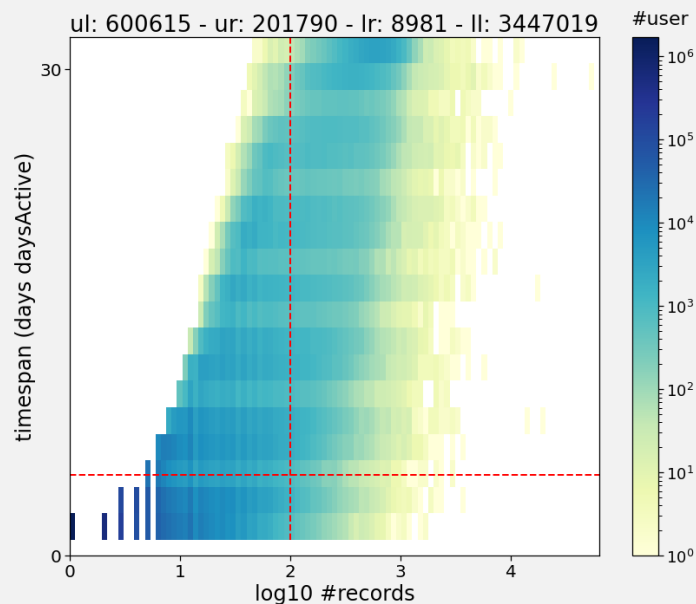
HANDS-ON ANALYSIS (3):

For each user in the sample, the code below calculates total number of location records ('pings') and the number of days on which they show up in the dataset ('active days'). This is plotted on a two-dimensional histogram. The histogram shows the density of users with regard to number of pings and number of active days. Darker areas represent more users and lighter areas represent fewer users. In this case, the densest part of the user distribution (the bottom left quadrant) comprises users with few pings and few active days.

The red lines represent one possible choice of filtering thresholds. Only users with a sufficient number of pings and active days (those in the top right quadrant) will be retained. This results in a smaller sample comprising only users who are observed more frequently and consistently over time.

user_stats.py

```
1 # compute a table of user statistics
2 user_stats_df = mk.stats.userStats(df_filtered).compute()
3
4 # set thresholds of minimum pings and active days
5 min_user_pings = 100
6 min_user_days = 50
7
8 # plot 2d histogram
9 mk.stats.plotUsersHist(user_stats_df, min_pings = min_user_pings, min_days = min_user_days)
10
11 # make a clean dataframe keeping only users that meet requirements
12 valid_users = list(user_stats_df.query("daysActive >= @min_user_days"
13                                     "& pings > @min_user_pings")["uid"])
14 df_clean = mk.stats.filterUsersFromSet(df_filtered, users_set = valid_users)
```



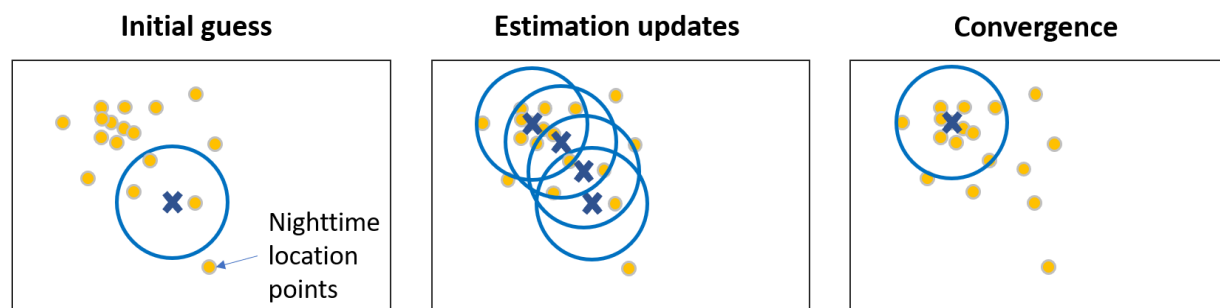
4. Impute home and work locations

Home and workplace locations are among the most important locations of an individual. Human mobility patterns show a high degree of temporal and spatial regularity, where each individual is likely to be observed at a few, highly frequented locations, such as their home location (most likely at nighttime) and workplaces (during daytime on weekdays).¹⁴

Home and workplace detection is essential for mobility data analysis, since it is used in many downstream tasks including representativeness checks, population displacement analysis, commute trip analysis, and POI visit estimation. Moreover, by overlaying such information with census data, estimated home locations can be used to infer a user's socioeconomic status such as an approximate household income.

A user's home location is typically estimated by clustering the users' nighttime location points. The clustering algorithm can be applied to the user's raw pings or to their stops (also called stay points) where users have spent some time and could be further refined by putting more weight on places that were stayed for longer durations.¹⁵ Similar to home location estimation, workplaces are estimated on daytime stop locations. Additional constraints such as "the workplace needs to be visited more than 3 times a week" can be used in the algorithm to improve the estimation accuracy.¹⁶ Such heuristic rules should be tailored to study aims and local cultural and labor market characteristics.

Figure 7: Imputing home location using Mean-Shift Clustering



¹⁴ Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196), 779-782.

¹⁵ Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8), 790-799.

¹⁶ Lenormand, M., Louail, T., Barthelemy, M. & Ramasco, J. J. Is spatial information in ICT data reliable? (2016).

HANDS-ON ANALYSIS (4): Imputing home and work locations

Home and work locations can be imputed from human mobility data - ie. a table showing timestamp, user IDs, and geographic coordinates - using clustering algorithms.

The program below firstly transforms raw location observations to stops. Clusters of points located within a 200-meter radius over a period of at least 10 minutes are classified as a stop. The resulting table specifies the location and the start and end time of each stop.

Home and work locations are then computed based by identifying the stop locations at which each user spent most time during night-time and day-time hours. These are specified respectively as 9.30pm to 7am; and 9.30am to 7pm. Mobilkit allows additional constraints to be placed on the home-work location identification. For instance, an additional parameter can be specified to force the work location to be different from the home one by a minimum distance (see Mobilkit online documentation).

home_work.py

```
1 # compute stops (where user moves less than 'spatial_radius_km' for at least 'minutes_for_stop')
2 stops_df = mk.spatial.findStops(df_clean, tessellation_shp = aoi_grid_file,
3                                 stay_location_kwds = {
4                                     minutes_for_stop:10,
5                                     spatial_radius_km:0.2})
6
7 # compute user locations: groups of stops close enough to be a single venue
8 locations_df = mk.spatial.computeUserLocations(stops_df, method = 'dbscan')
9
10 # create table showing imputed HW locations for each user
11 location_stats_hw = mk.stats.stopsToHomeWorkStats(locations_df,
12                                                    home_hours=(21.5,7),
13                                                    work_hours=(9.5,19))
14 user_HW_table = mk.stats.compressLocsStats2hwTable(location_stats_hw)
15
```

5. Test representativeness

When statistical agencies and research firms conduct surveys, they typically aim to make observations of a limited sample of the population in question but to ensure the resulting dataset can be used to make valid inferences about the general population. Indeed, survey methods provide a range of sampling methodologies to achieve this goal. In complicated populations, stratified sampling ensures enough members of minority groups are included in the overall sample.

Mobile phone location datasets are different. In survey terminology, they can be defined as ‘convenience samples’: they provide observations about those community members who can be most readily observed. This presents challenges since urban development specialists are typically interested to draw inferences about a region’s overall population.

Early in an analysis workflow, it is therefore imperative to test the representativeness of a given dataset vis-a-vis the local population (including socio-economic, ethnic and gender aspects¹⁷). Geographical representativeness can be assessed by checking the spatial variance of the population sampling rate, which can be computed by dividing, for each administrative unit, the number of users in the dataset by the official population. This gives a percentage of the population for each neighborhood that is in the sample.

Home and work locations for the observed mobile phone users are imputed (see Hands-On Analysis 4), aggregated by census tract, and compared with population counts from the most recent national census. A starting point for representativeness analysis is to compute the Pearson’s correlation coefficient. The dataset pictured in Hands-On Analysis 5 exhibits very high correlation of user numbers with census counts ($r^2=0.99$).

Representativeness of mobile phone users may be checked using other sociodemographic and economic variables in addition to population distribution data. For example, plotting the sample rates against the median income ranges allows for assessing disparities in representativeness with respect to wealth levels.¹⁸

As a non-traditional data source, bias in location data can be mitigated but not eliminated. Further filtering and data cleaning should be undertaken to address any identified under-representation of population groups (Step 4). Representativeness tests allow analysts to get a handle on the severity of such biases. Importantly, residual biases should be openly acknowledged and communicated to policymakers alongside analysis findings. Policy makers should be presented with information to judge whether remaining biases – such as under-representation of low-income and elderly people – are mild or severe. Such knowledge is important to inform how much weight should be placed upon the findings when forming policy-decisions and to inform how they will be utilized alongside other information sources such as surveys and official statistics.

¹⁷ Wang, F. & Chen, C. On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C*, 87, 58–74 (2018).

Barbosa, H. et al. Uncovering the socioeconomic facets of human mobility. (2020-12-01).

Gauvin, L. et al. Gender gaps in urban mobility. *Humanities and Social Sciences Communications* 7, (2020).

¹⁸ Zhao, Z. et al. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science* 30, 1738–1762 (2016).

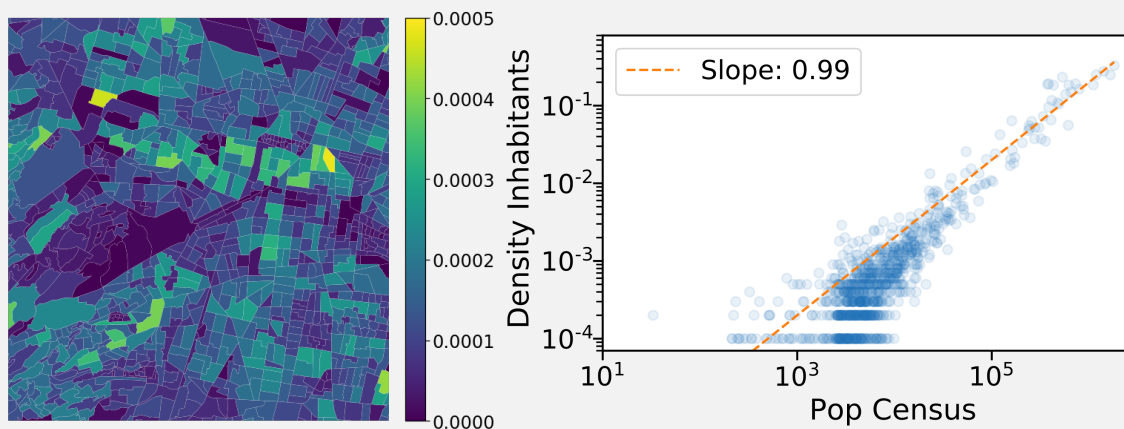
HANDS-ON ANALYSIS (5): Testing representativeness

Once we computed the home and work locations as in HANDS-ON ANALYSIS (4), we can test the representativeness of the data by comparing with census figures.

For instance, one can look at the correlation between the census number of residents and the number of detected homes in each census area (provided the same areas were used to tessellate the mobility data in the first place).

representativeness.py

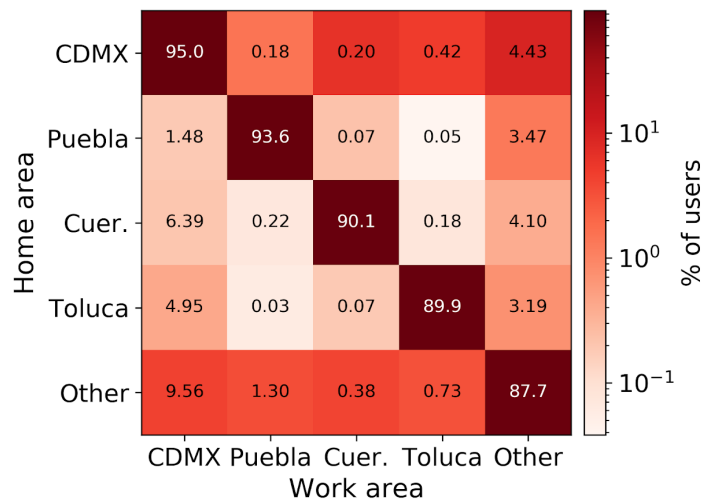
```
1  # summarize number of users per area
2  pop_per_area = location_stats_hw.groupby('home_tile_ID').agg('uid':'nunique',
3                                                                'home_pings':'sum')
4
5  # load census population and merge with cellphone user counts
6  gdf_areas = gpd.read_file('data/population_shapefile.shp')
7  gdf_areas = pd.merge(gdf_areas, pop_per_area, on='tile_ID')
8
9  # compute user density
10 gdf_areas['pop_density'] = gdf_areas['census_pop'] / gdf_areas['census_pop'].sum()
11
12 # plot user density vs. census population
13 gdf_areas.plot()
```



6. Analyze commuting patterns

Movement patterns between pairs of locations can be represented in an Origin Destination matrix ('OD matrix'). For instance, Figure 8 presents journeys between home and work locations for a sample of people observed through mobile phone data in Ciudad de Mexico (CDMX - Mexico's capital region). The diagonal presents the share of people who live and work in the same area. Surrounding municipalities feature a higher proportion of people who commute outside their local area as compared to CDMX (which is the preferred commuting destination for each of the other municipalities).

Figure 8. OD matrix for daily commutes in the Mexico City area



OD matrices are relatively simple to produce, as they only require counting the number of people traveling from one location to another in the time span of interest. Transport demand surveys habitually produce information on trip demand between fine-grained geographic regions of a city disaggregated according to user demographics, time of day and week, and transportation mode. Human mobility data cannot produce such disaggregated data due to its anonymized nature. However, it has other advantages, particularly when analyzed in conjunction with traditional transport planning metrics.

For instance, after creating a baseline OD matrix, it is possible to observe how the population reacts to a disruption or discontinuity - an extreme weather event, the closure of a train line, or the introduction of a new Bus Rapid Transit network - understanding which neighborhoods were most impacted, how trip patterns changed and which destinations were more resilient to adverse disruptions. Another advantage of mobile phone data applied to OD analysis is that it allows researchers to spot the evolution of route choices over long time periods¹⁹. And compared to face-to-face or telephone survey techniques, large passively generated movement datasets may have completeness advantages, since movements are observed directly rather than relying on people to remember and accurately describe their transport choices.

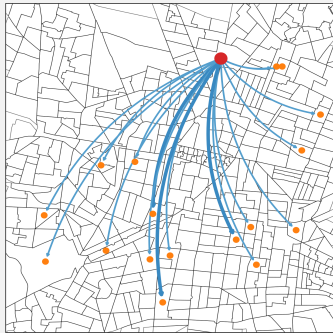
¹⁹ Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, Wei-Ying Ma. Understanding Mobility Based on GPS Data. In Proceedings of ACM conference on Ubiquitous Computing (UbiComp 2008), Seoul, Korea. ACM Press: 312-321.

HANDS-ON ANALYSIS (5): Origin destination matrices of commuting patterns

Another byproduct of the home and work locations obtained in HANDS-ON ANALYSIS (4) is the home-work based commuting matrix. Here we count the number of users having the home and workplace in each couple of possible areas. We then plot the most prominent commuting directions by selecting the top-20 OD edges by user count.

OD_matrix.py

```
1 # group by home and work locations and count users for each combination
2 OD_matrix = location_stats_hw.groupby(['home_tile_ID', 'work_tile_ID']\
3                                     .agg({'uid': 'nunique'})\
4                                     .rename(columns={'uid': 'volume'})
5
6 # visualize top 20 destinations for one selected origin
7 selected_origin = 12
8 slide_od = OD_matrix.loc[selected_origin].sort_values('volume', ascending=False).iloc[:20]
```



7. Point-of-interest analysis

A common type of analysis is to see visiting patterns to POIs: who is visiting, when visits happen, and from which parts of the city visitors come. Catchment areas of businesses or public amenities can be identified and the social mix of visitors can be identified by inferring socio-economic traits from their area of residence.

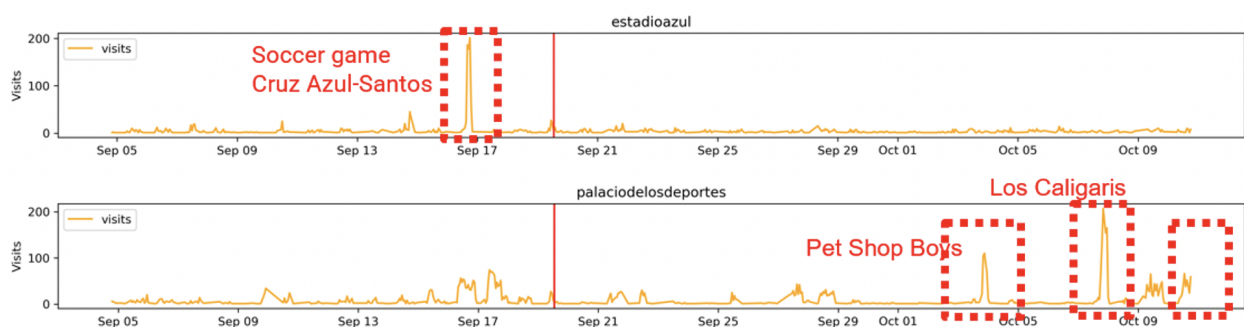
A key requirement for such analysis is the availability of a geographic dataset of places. OpenStreetMap is a freely editable geographic database of the world, established in 2004, that lists places together with their names, coordinates, and category (eg. education, medical, restaurant, store).

There are two key steps to detecting stops and visits to specific places:

- Stop detection.** For many POI visits, such as a 30-minute trip to a supermarket, the individual's location will be recorded multiple times within and around the store. Stop detection is the process of identifying stops (or 'stay points') where a user spent a period of time. Analysts may specify a maximum radius (eg. 100 meters) and a minimum period of time (eg. 20 minutes). Based upon these parameters, a stop would be identified with the coordinates of the supermarket for the period of time between the first and last ping recorded there.
- Visit detection.** Having extracted a list of stops from the raw movement data, the analyst can combine this with a set of POI locations. Where stops correspond with a known POI, a visit is inferred. Aggregating visit trends among the observed sample of city residents allows for a range of analyses: from understanding how trips to metro stations and bus-stops were influenced by Covid-19; to detecting trends in public park usage; to assessing the local economic impact of an urban regeneration development through change in visits to restaurants, coffee shops and specialty stores.

Among growing research and practical applications, visit trends have been used to monitor the disruption and recovery of local businesses after major disaster events²⁰; understand how people from different socio-demographic backgrounds encounter at different types of places in urban environments²¹; and to monitor predict COVID-19 outbreaks to inform public health institutions²².

Figure 9: Mexico City - using rock concerts and football matches to ground-truth visit numbers



²⁰ Yabe, T., Zhang, Y., & Ukkusuri, S. V. 2020. "Quantifying the economic impact of disasters on businesses using human mobility data: a Bayesian causal inference approach." *EPJ Data Science*, 9(1), 36.

²¹ Moro, E., Calacci, D., Dong, X., & Pentland, A. 2021. "Mobility patterns are associated with experienced income segregation in large US cities." *Nature Communications*, 12(1), 1-10.

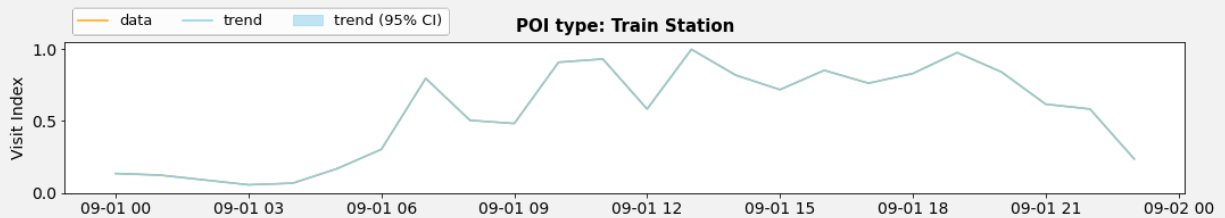
²² Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., & Leskovec, J. 2021. "Mobility network models of COVID-19 explain inequities and inform reopening". *Nature*, 589(7840), 82-87.

HANDS-ON ANALYSIS (7): Point of interest visit analysis

Here we leverage on the stops locations of users as obtained in Hands-on Analysis (3) and the home-work locations computed in Hands-on Analysis (4) to compute the hourly visitation patterns in a set of selected POIs using open-source movement dataset 'GeoLife' in Beijing.

poi_visits.py

```
1 # define Points of Interest
2 df_pois = pd.DataFrame([39.86494076, 116.379170, 'South Station', 0.5, "Train station"],
3                          columns = ['lat', 'lon', 'name', 'radius', 'category'])
4
5 # visits for each POI based on pings more than 200 meters from user's home
6 # pass the HW locations and stops dataframe (from Hands On Analysis 3 and 4)
7 per_POI_visits = mk.spatial.compute_poi_visit(df_pings=stops_df, df_homes=user_hw_table,
8                                               df_pois=df_pois, min_home_dist_km=0.2,
9                                               visit_time_bin='1H')
```



poilat	poilon	Name	radius	type
39.864941	116.379170	South Station	0.5	Train Station
39.942086	116.354219	North Station	0.5	Train Station
39.902534	116.427129	Beijing Station	0.5	Train Station
39.920758	116.433130	Beijing SOHO	0.2	Shopping
39.912633	116.409748	WF Central	0.2	Shopping



3. Applied use cases

The use cases present practical applications of the methods described in Section 3.

- Anomaly detection is used to identify which neighborhoods saw a population decline after the 2019 Puebla earthquake and to pinpoint changes in transport accessibility after major storms in Chennai. Visit trends - derived using Point of Interest (POI) analysis - are used to understand usage patterns of parks in Chennai and San Jose, presenting a new perspective on equity and usage of public amenities.
- Mobility metrics, such as average daily distance traveled by cellphone users, is overlaid with local area poverty maps to understand how communities responded to Covid-19 stay-at-home orders.
- In an applied case of post-disaster population mapping, mobility data is used to identify Polish towns that faced high population inflows during the Ukraine crisis indicating immediate humanitarian response needs as well as medium-term challenges for public service delivery.
- Lastly, the use of mobility data to assess how many jobs and amenities city residents can access - based on actual observations of road travel speed - is presented in the case of San Jose, Costa Rica.

Data sources

The case studies are based on work performed by the GFDRR / World Bank team. The case studies utilized smartphone-derived GPS location datasets describing the movement patterns of an anonymized sample of local residents (see Figure 4 - Stylized example of a mobility dataset)²³.

In addition, two derived datasets were utilized. Case study 4 on mapping displaced populations utilized processed and anonymized data from Meta's Data-For-Good program. Case study 5 utilized the Mapbox Matrix API, a dataset describing road travel speeds based on cellphone observations as well as smartphone GPS data.

²³ Anonymized and de-identified samples of human mobility data based on users who opted into location sharing for defined purposes were provided by Cuebiq, Outlogic and Quadrant, and accessed under the terms and conditions of the Development Data Partnership program (<http://datapartnership.org>).

Disaster response: Displacement after the 2019 Puebla Earthquake

The 2017 Puebla Earthquake affected a large swathe of central and southern Mexico. Using a smartphone mobility dataset, the extent of population displacement was mapped and 'recovery curves' constructed. Citizens of lower-income districts were found to have a lower rate of population displacement compared with residents of equally-affected wealthier areas.

On September 19, 2017 just after 1pm, citizens in Mexico City had just concluded an orderly earthquake drill when a magnitude 7.1 quake struck with an epicenter 55 kilometers away, near the city of Puebla.

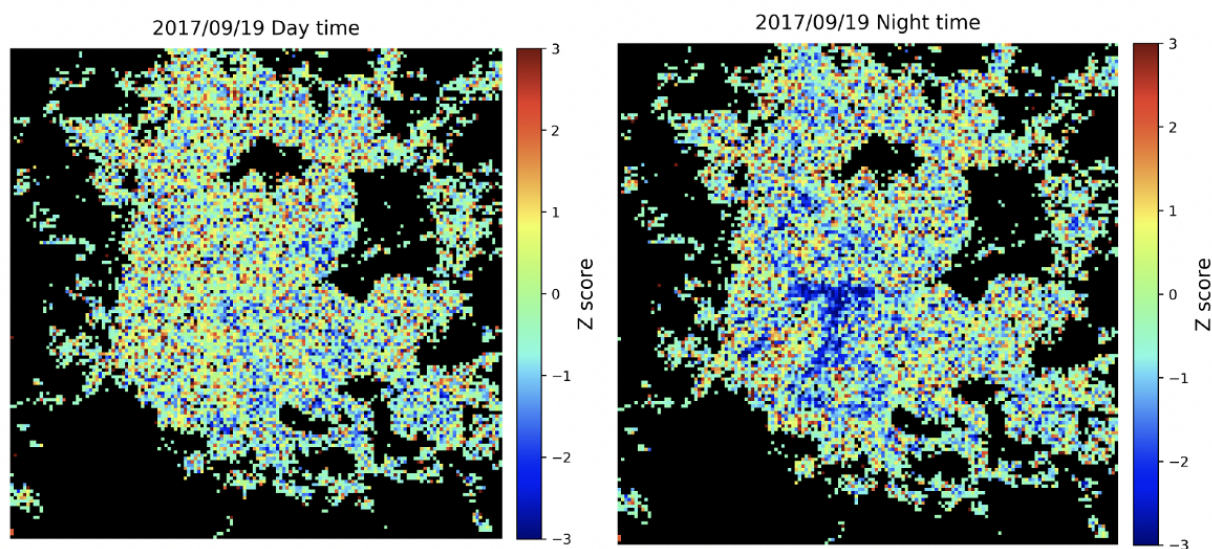
Mobility data offers a way to understand the movement of a population in greater detail and with less lead time. Since people take their phones with them when fleeing a disaster, such devices produce information even as other systems go silent. This study utilized a dataset of roughly 200,000 smartphone users in Mexico who opted into sharing their GPS locations for defined purposes. The dataset was provided by technology firm Cuebiq under its Data-for-Good program.

Mapping displacement intensity

In a first step, the home and work locations of users in the study area were imputed based on their most frequent locations at night and during working hours. The night-time population of users prior to the earthquake was plotted in order to provide a baseline for subsequent analysis.

Using the Z-score anomaly detection method, the mean population in the 30 days before the earthquake was compared with the night after the earthquake struck. As shown in Figure 19, on the night of August 19, 2019, the Z-score highlights a population drop in central and southern districts of Mexico City (corresponding to areas in which building collapses took place).

Figure 9: Anomaly detection maps highlight post-earthquake displacement from Mexico City districts



Poorer people stayed closer to home

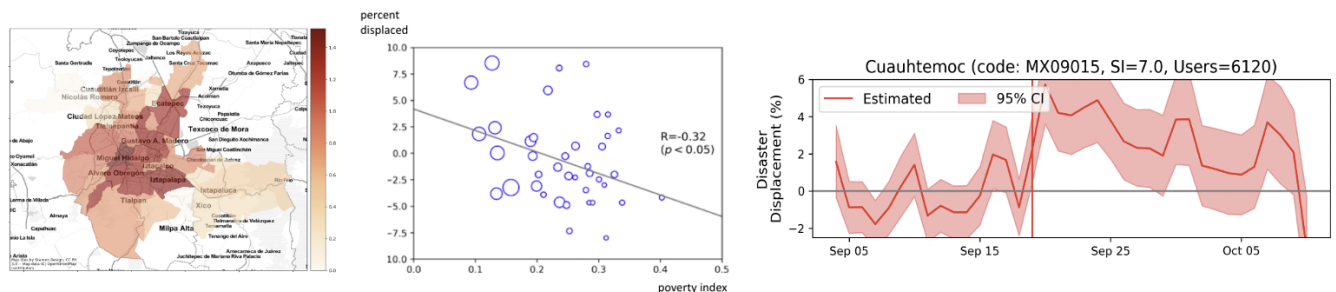
Having estimated the home locations of individuals in the mobility dataset, the study estimated two additional variables for any user whose night-time location changed after the earthquake:

- **Length of displacement:** Defined as the number of nights spent away from home.
- **Displacement destination:** The main location where the individual relocated (estimated using the same mean-shift clustering approach used to estimate home locations).

In addition, a geographic wealth index was created based on household survey data and overlaid with the estimated home locations of users. As a proxy for individual wealth, users were allocated the wealth level of the census tract where they live (scaled from 0 for most poor to 1 for most wealthy).

After the Puebla earthquake, mapping displacement destinations highlighted that the majority of displaced people moved to nearby districts within Mexico City – but displacement was detected across the country including to provinces and cities hundreds of miles away. Overlaying wealth and displacement patterns revealed that citizens of poorer areas were more likely to ‘remain in place’ compared with residents of equally affected wealthy areas. As observed in diverse disaster cases (eg. floods in New Orleans and Haiti), poorer households who may lack savings, social capital, access to transport and secure property rights are less likely to effectively safeguard their lives and properties by evacuating at-risk areas, suggesting the need for additional disaster preparedness support²⁴.

Figure 10: Analysis of displacement data, Puebla Earthquake. Left to right: destinations of displaced households within Mexico City; relationship between displacement rate and wealth of home neighborhood; disaster displacement curve for the heavily-affected neighborhood of Cuauhtemoc.



Post-disaster recovery curves

For urban areas, post-disaster recovery involves multiple actions unfolding over short, medium and long terms, including search and rescue operations, return of displaced populations, and restoration of urban services. The right panel in Figure 10 presents an example ‘recovery curve’ indicating one aspect of recovery, namely the return of displaced populations. For the heavily affected district of Cuauhtemoc, more than 4% of the local population was displaced from their homes according to the mobility dataset, with more than three weeks taken until the population returned to its pre-crisis level.

²⁴ Jimena Lopez et al. “Using Behavioral Insights To Improve Disaster Preparedness, Early Warning And Response Mechanisms In Haiti.” World Bank. <https://documents1.worldbank.org/curated/en/465051578683565433/pdf/Using-Behavioral-Insights-to-Improve-Disaster-Preparedness-Early-Warning-and-Response-Mechanisms-in-Haiti.pdf>

Resilient Transportation: Accessibility impacts of Cyclone Nivar

Chennai, India, is a city of 7 million people that regularly experiences coastal storms. In November 2020, Tropical Cyclone Nivar caused flooding that affected an estimated 40,000 homes and wind gusts that reached 70 kilometers per hour. To inform efforts to make the transportation system more resilient, changes in transport accessibility outcomes were evaluated. Data showed that trips to work and essential amenities were impacted by the storm, with remote and low-income neighborhoods most severely affected.

After Cyclone Nivar developed as a Category-1 tropical storm, heavy rainfall began to affect Chennai on November 23, 2020, with the greatest impact coming on the morning of November 25.

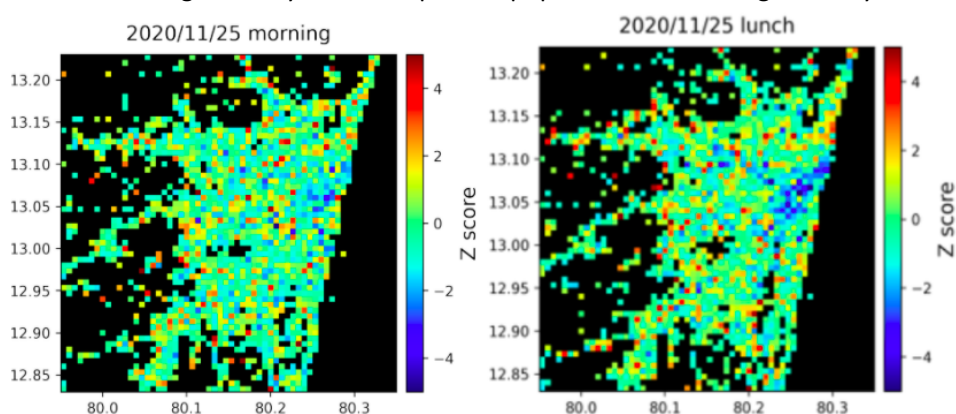
Identifying transport disruptions

By using a Z-score analysis of the population across different areas of Chennai, a remarkable drop in trips was observed on November 25 compared with a pre-disaster baseline period (Figure 11). In the central business district of the city, clusters of blue emerge with growing intensity during the morning and mid-day periods. This is a clear indication that many people who work in this district and buoy its daytime population did not travel to work leaving a 'cold spot' in the corresponding area.

Over the course of the week of the cyclone, the active population of the central business district was significantly lower during the day, with November 25 and 26 seeing the steepest drop. Interestingly, the shift in commuting patterns continued in the days following the worst of the cyclone. Typically, the central areas of big cities see a population increase during morning and evening rush hours, reflecting the movement of those who travel daily in and out of the central areas on their way from and to work.

Figure 11: Displacement during Cyclone compared to normal day

Subtitle: Blue is significantly below expected population. Red is significantly above expected population

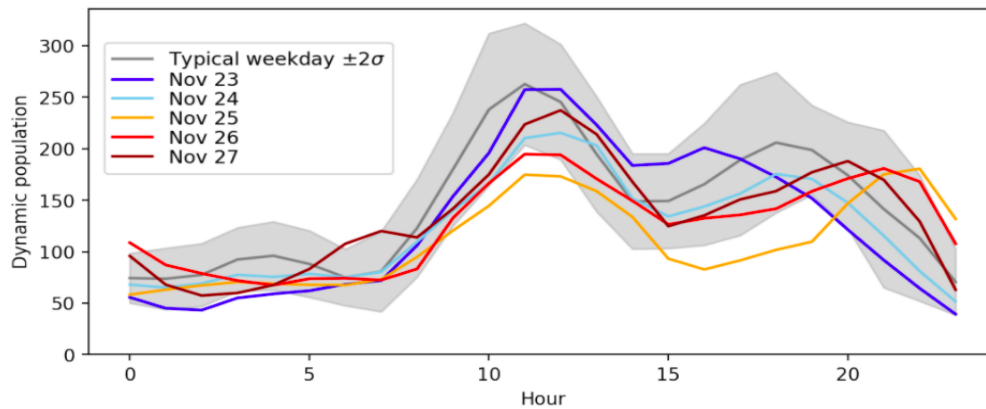


Change in trip length during the storm

As can be seen from Figure 12, between November 25 and 27, evening rush hours occurred later compared to a typical workday. This could reflect people attempting to wait out the rush during the day of the cyclone. If so, this suggests that an extreme weather event of this kind might trigger a 'tragedy of the commons' situation, in which people attempting to avoid rush hour merely recreate it at a later time at which the transportation system is not designed to handle the inflated volume of people.

Figure 12: After Cyclone Nivar, rush hour shifted later

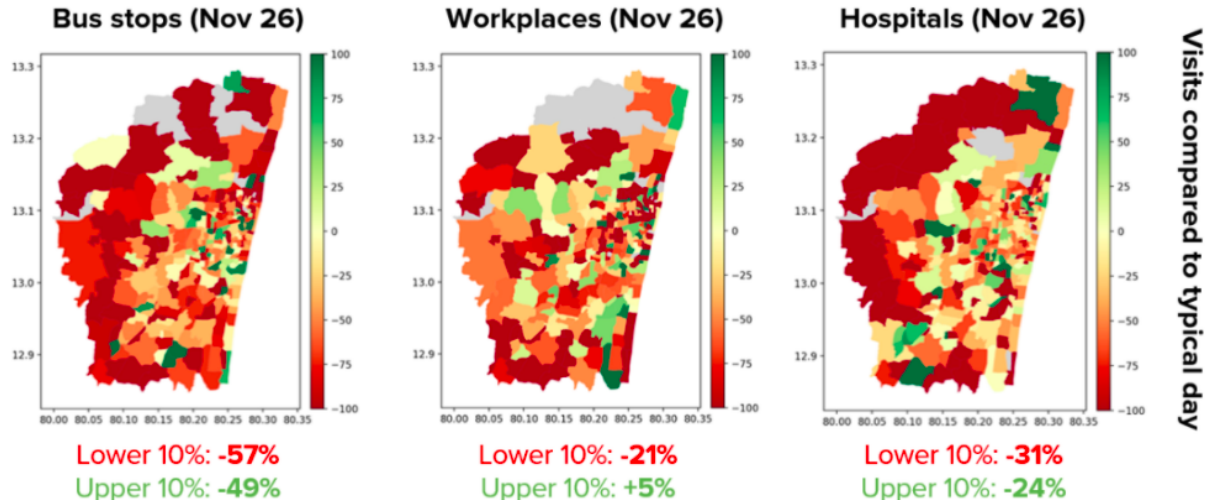
Subtitle: Population in central commercial areas throughout the day during the week of cyclone



Visits to Points of Interest

After matching mobility data with POI information from OpenStreetMap, we are able to estimate visits to different categories of places. On the day after the heaviest rainfall, there was a significant drop in the number of people going to bus stops and hospitals. This followed a roughly similar pattern among the poorest and richest people, although affecting the poorest decile comparatively more. In turn, the lowest income group had experienced a significant reduction in attendance at workplaces, whereas the highest income group even had a slight increase from the baseline.

Figure 13: Change in visits to locations by income



People were also less likely to travel to their local stores during the cyclone. Travel to supermarkets and local stores both fell during the worst of the cyclone. But trips to supermarkets quickly recovered, whereas trips to local stores remained depressed. This could be due to the assumption that supermarkets would be better stocked compared to local stores, reflecting a belief that deliveries to smaller neighborhood stores might be more heavily affected by a disrupted transportation network.

Who uses city assets? Inequality and usage patterns of parks

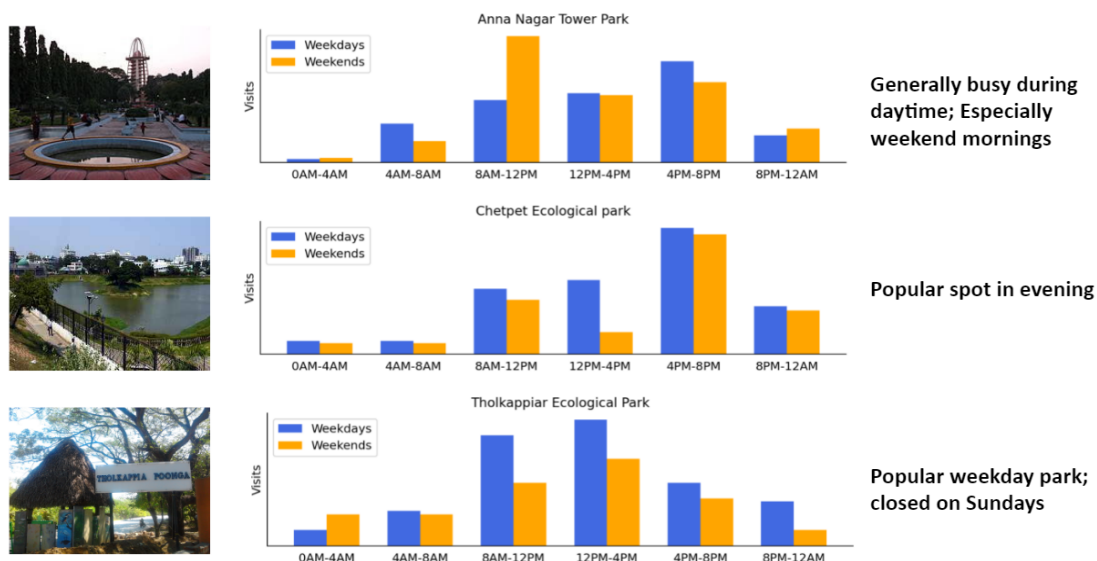
Parks are a critical part of urban life, with extensive research showing the value of open space on quality of life. Parks also serve as public places that allow for people from all socioeconomic backgrounds to meet and encounter each other.²⁵ Human mobility data enables us to track the movements of individuals at high spatial and temporal granularity, which allows us to see to what extent people in a city visit parks and from what neighborhoods they arrive.²⁶ Rapid utilization estimates for parks and other public amenities can help build a stronger evidence base to support public investment programs.

To build a picture of public amenity usage in Chennai, India, a smartphone location dataset featuring anonymized trajectories of 320,000 local residents was utilized. To get a proxy for the wealth level of each individual, home locations were imputed based on most frequent night-time locations, and cross-referenced with geographic information on asset ownership provided by the Relative Wealth Index dataset.

Constructing visit profiles of public amenities

The usage profile of three parks in Chennai is depicted in Figure 14, each park having a distinct visitation profile during week and weekend days (blue and orange respectively). The affluent Anna Nagar Tower Park is generally frequented during the daytime of both weekdays and weekends, but is especially busy on weekend mornings. Chetpet Ecological Park is relatively more popular in the evenings. Tholkappiar Ecological Park has a distinct visitation profile from the previous two, with more visits on weekdays than weekends. Visitation patterns measured via mobility data can be used to know whether certain parks are over-crowded or underutilized at particular times, which might prompt further investigations on potential barriers to use such as fear of crime, poor condition, or substandard lighting.

Figure 14: Visitation patterns of parks by time-of-day and day-of-week



²⁵ Douglas S. Massey and Nancy A. Denton, The Dimensions of Residential Segregation, Social Forces Vol. 67, No. 2, pp. 281-315

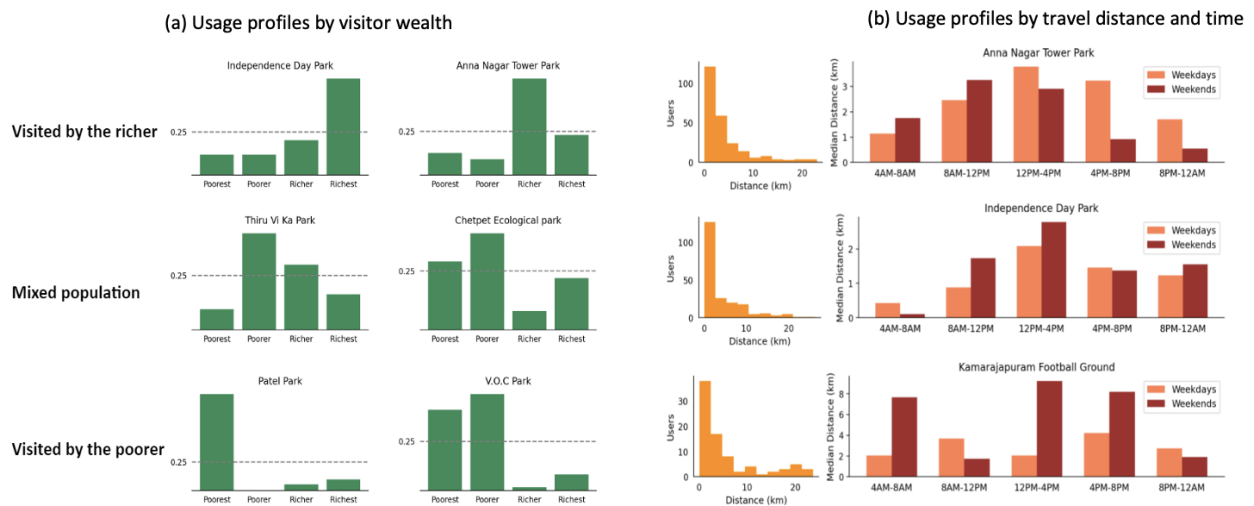
²⁶ Esteban Moro, Dan Calacci, Xiaowen Dong & Alex Pentland, Mobility patterns are associated with experienced income segregation in large US cities, Nature Communications, Volume 12, Article number: 4633 (2021)

Quantifying inequality in park usage

Figure 15 (right) shows the quartile distributions of wealth of the visitors for nine representative parks in Chennai. In each plot, each bar from left to right shows the proportion of visitors in each wealth quartile from poorest to richest. In a perfectly “mixed” situation, where public parks are visited by people belonging to all socioeconomic backgrounds, each of the four bars will be aligned to the 0.25 mark (that is $\frac{1}{4}$ of the total number of visitors), which is shown as a horizontal dotted line.

In Chennai, the socioeconomic profiles of the people who visit each park significantly vary. The three parks shown at the top of the image are visited mainly by members of the richest quartile including Anna Nagar Tower Park, which covers an area of 15.35 acres and features a 135-foot-tall, 12-storied tower located right at its center. The park is equipped with various amenities, including an amphitheater, a bird-watching deck, badminton courts, play area for kids, a skating rink, a lake, and a range of convenience facilities for the visitors. The middle row of the image shows examples of parks visited by a mixed population while the bottom row gives examples of parks where most visitors are from low-income neighborhoods.

Figure 15: Mobility data reveals how parks in Chennai are not equally used by all population groups



Building the investment case for public assets

By combining visitation pattern data and estimated home locations individuals extracted from mobility data, we are able to estimate the typical distance traveled to reach each park in Chennai. For instance Independence Day Park has a geographically small catchment, with more than 50% of its visitors traveling less than 1km, while Kamarajapuram Football Field attracts a high number of visitors coming from a greater distance (10km - 20km). As the plots in the right column show, in both these parks, and especially in Kamarajapuram Football Field, travel distances are significantly longer on weekend afternoon and evening periods. Identifying locations where people need to travel long distances to access public facilities, can be tied in with transport planning strategies to improve the accessibility and inclusiveness of public facilities. Moreover, socioeconomic footprints of parks can be used to assess whether or not park investments equally benefit different groups of society.

Rapid displacement mapping: Ukrainian refugees in Poland

More than 6.5 million people fled Ukraine after the Russian invasion in the largest population movement in Europe for 50 years. Data from Meta's Data-for-Good program was used to map intensity of population movement to neighboring countries. The detailed information on which towns experienced largest inflows relative to population proved valuable to inform policy responses on service delivery, labor markets and humanitarian support, as well as showing feasibility of gender-disaggregated insights.

With Ukraine's neighbors experiencing arrival of displaced people on an unprecedented scale, smartphone data derived from the Facebook app proved a valuable complement to official refugee statistics. Approximately 59% of Ukrainians use Facebook making the sample size large.

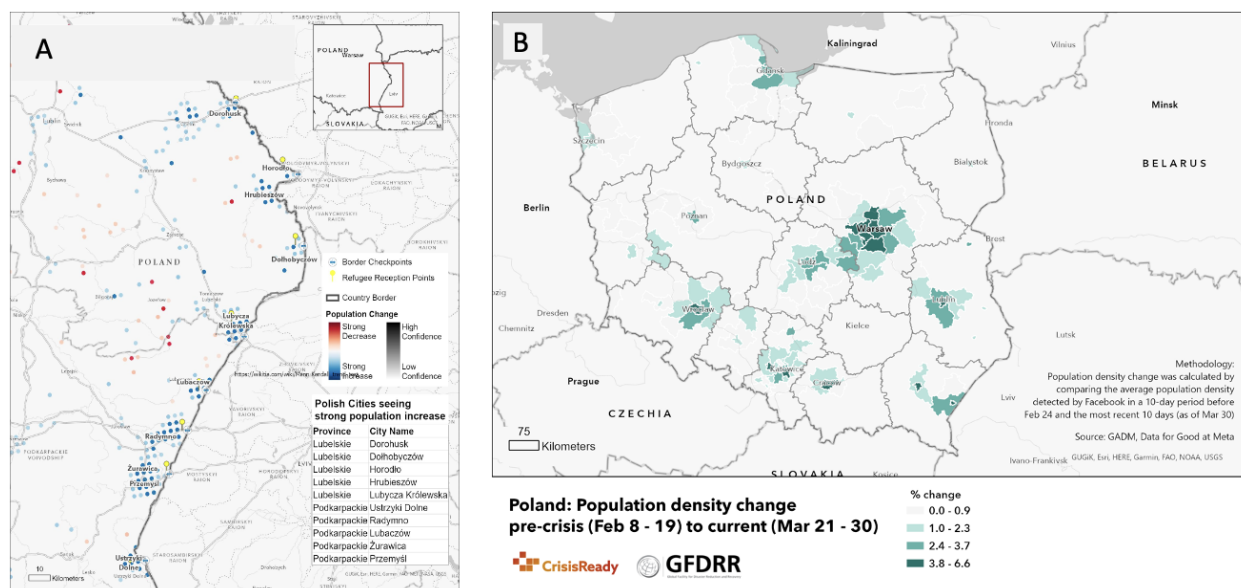
Mobile phone data complements official statistics

Official refugee figures and datasets provided by Meta each have pros and cons. Official refugee statistics use information collected at national border crossings: these may be incomplete, but they are subject to established protocols for verification and triangulation. Meta's displacement data is based on large-scale observation of people's locations, but methods for estimating population displacement with it are recent. The datasets can be examined together to draw reliable conclusions.

Displacement mapping highlights support needs

Early in the crisis, the maps clearly highlighted border towns in Poland, Romania and Moldova that experienced highest short-term population influx (Figure 16A). The maps confirmed a diffuse pattern of arrivals to neighboring countries with displaced people arriving to secondary towns and cities, not just national capitals. Smaller towns such as Zielona Gora and Slupsk had high population inflows of 15% or more relative to their usual population (Figure 16B) highlighting needs for support.

Figure 16. Displacement of Ukrainians in Poland. (A) Population influx to Polish border regions during first days after Russian invasion; (B) Regions with highest relative population influx in late March.



Displacement timeline: arrivals, returns, onward movements

Cellphone-derived data on population movements is updated frequently, making it possible to distinguish areas experiencing a short-lived surge in population - such as zones through which refugees transited en route to other destinations - from areas that are dealing with sustained population increases. In Poland, the border town of In Figure 17, the border town of Hrubieszow saw a major population uptick in the first week of the war, followed by a return to baseline levels; this contrasts with the sustained population increase experienced in the national capital Warsaw.

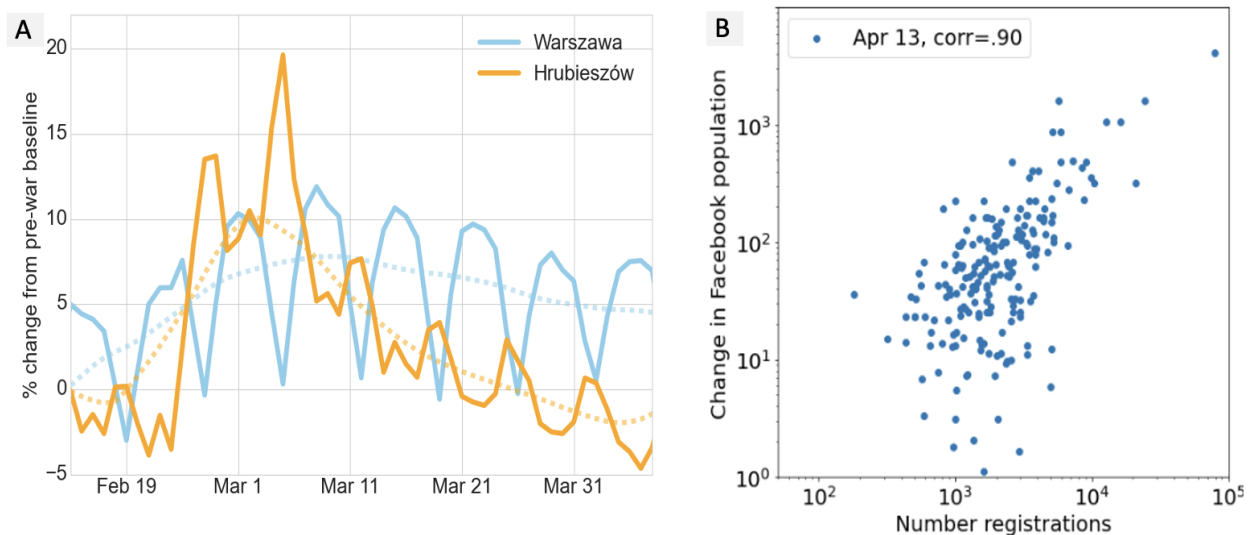
Meta's data also highlighted that significant numbers of Ukrainians moved beyond the first country they arrived in. An estimated 669,000 of the 1.9 million people who entered Poland moved onwards to a third country (Germany being the top destination). Complementing official refugee statistics, the openly available datasets from Meta allowed a detailed understanding of onward movement to third countries and the pace of return travel to Ukraine.

Gender dimension and validation

Unlike other mobile phone datasets, data products made available by Meta are disaggregated by gender. The data confirmed a strong gender divide, with around 70% of observed displaced persons being women.

To validate use of the Facebook-derived data for refugee mapping, the data was compared with official refugee registrations (Figure 17A). The two datasets correlate very strongly early in the conflict (correlation coefficient of 0.9 on April 13). The correlation becomes less strong later in the conflict; a finding that is expected since official refugee counts may not capture onward and return movement.

Figure 17. Displacement timelines and validation. (A) Mobility data shows a sustained population increase in Warsaw compared to a shorter-lived population surge in border town Hrubieszow; (B) Strong correlation between cellphone and official refugee statistics.



Spotlight 1: Can the destinations of post-disaster displacement be predicted?

As of May 2022, more than 6 million people had been displaced from Ukraine due to the military invasion. In the short term, movements of forcibly displaced people were conditioned by the security situation in their home regions, accessibility of border crossings, and access to transport and funds. Surveys underscored that, in many cases, displaced people intended to continue onwards to other destinations within Europe. While population movements cannot be predicted ahead of time, extensive research has established that density of social ties between pairs of regions correlates closely with migration flows.

Density of social ties predicts evacuation behavior

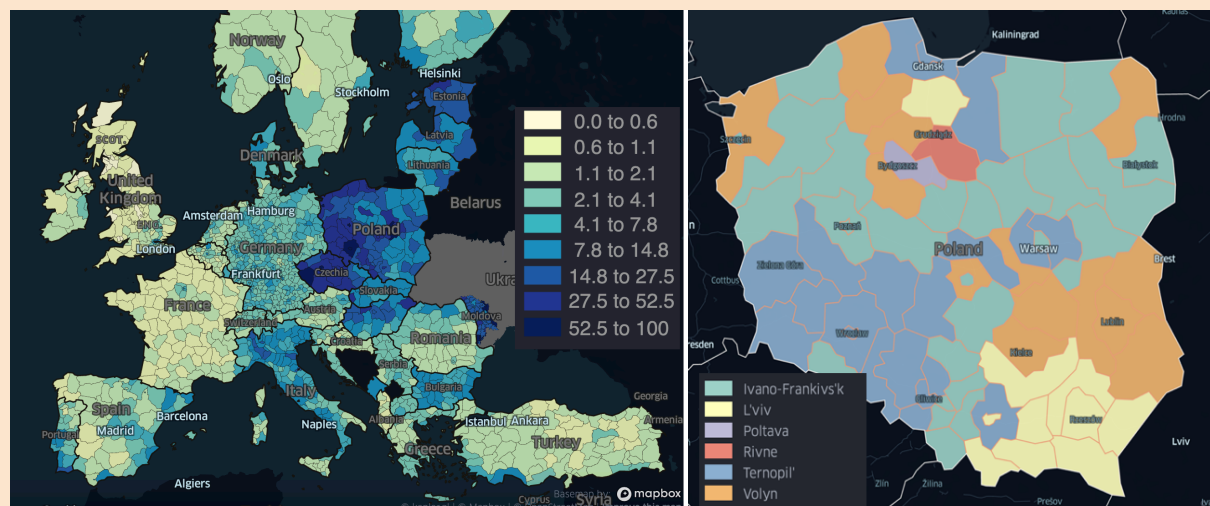
Figure 19 presents a mapping of the density of social ties between Ukraine and other regions of Europe using the Social Connectedness Index. The Index presents the relative probability of two people from each geographic unit being friends on Facebook. Past studies have confirmed that high Social Connectedness Index values predicts higher trade volumes, tourism flows and disease transmission between region pairs²⁷.

In a study of Cyclone Yaas, which hit eastern India in May 2021, cell phone data was used to estimate displacement from affected regions to other districts of India. Distance, wealth of destination region, and strength of social ties were evaluated as explanatory factors. The study found that the Social Connectedness Index was the strongest predictive factor, reflecting the tendency of displaced people to head for places where they have friends and family members (Yabe et al, forthcoming).

Anticipating population inflows

In the case of forced displacement from Ukraine, mapping the density of social ties in early stages of the crisis proved informative as a leading indicator of population displacement several weeks later. Detailed social ties mapping provides an empirical basis to identify areas that may see population inflows. Moreover, mapping out connectedness between regions in Ukraine to regions in Europe can elucidate how conflict may influence different destinations. For example, while Polish cities are very connected to all of Ukraine, they tend to be most connected to Ivano-Frankivsk, Lviv, or other western Ukrainian regions. Bulgarian cities are most connected to Odesa and southern Ukraine, while Slovakia and Romania (except its Baltic coast) are most connected to Chernivitsi and Zakarpattia (southwest Ukraine).

Fig 19: Density of social ties between Ukraine and European regions (left), most connected oblast for Polish region (right)



Source: GFDRR analysis. Left figure: darker colors indicate higher connectedness to Ukraine (metric: Social Connectedness Index).

²⁷ Michael Bailey, Abhinav Gupta, Sebastian Hillenbrand, Theresa Kuchler, Robert Richmond, Johannes Stroebel, International trade and social connectedness, Journal of International Economics, Volume 129, 2021
 Bailey, Michael and Kuchler, Theresa and Russel, Dominic and State, Bogdan and Stroebel, Johannes, The Determinants and Effects of Social Connectedness in Europe (2020). CESifo Working Paper No. 8310

Covid-19 response: Monitoring adherence to stay-at-home orders

The Covid-19 pandemic led governments to impose unprecedented restrictions on people's activity to stop the spread of the virus. However, the novelty of the restrictions meant that there were limited tools to measure their effectiveness or the extent to which people were following them. Mobility data makes it possible to assess degrees of compliance with lockdown orders, and also identify how adherence to such health-promoting measures differed among socio-economic groups - highlighting the need to support poorer communities that face tough trade-offs between health and livelihood considerations.

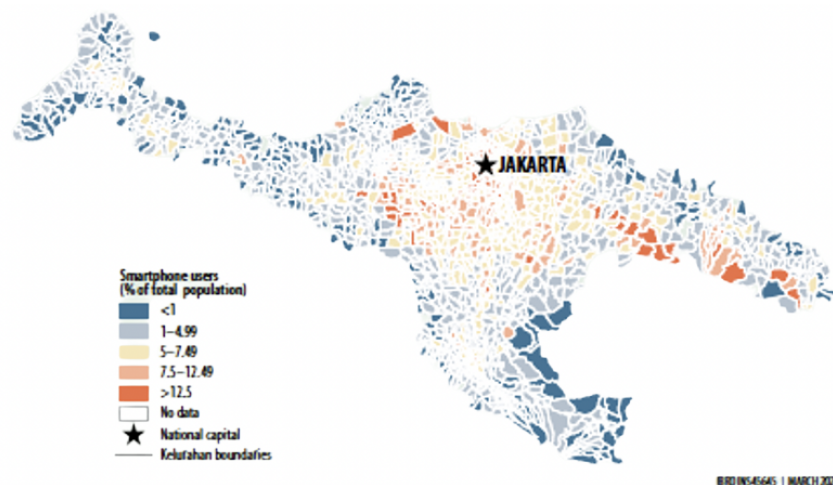
The onset of Covid-19 led to policymakers scrambling to respond to a virus that spread fast with high potential fatality levels. One of the tools adopted around the world were stay-at-home and social distancing orders. By keeping people in their homes and away from workplaces, restaurants, movie theaters, or other places where coronavirus could spread, governments hoped to slow the virus.

Given the high economic costs of lockdown orders and the public health consequences if they failed, governments understandably wanted to know to what extent such policies were being adhered to. They also wanted to know where they were being followed more than others, what groups might be unable to do so, and where the next hotspots might emerge from places where the orders were not being followed.

Mobility data sample

Mobility data offered one way for governments to understand the movement of their populations. Using data from technology firm Cuebiq drawn from more than 275,000 users in Jakarta, there was a sample size large enough to draw conclusions about movements. Jakarta is a city with sufficient smartphone users to provide this data, although it is sparser in outlying areas. We can assume that this would be even sparser in rural areas of Indonesia, showing how mobility data for now is best used in urban environments.

Figure 20: Smartphone users in Jakarta metropolitan area



Socio-economic discrepancies in Covid response

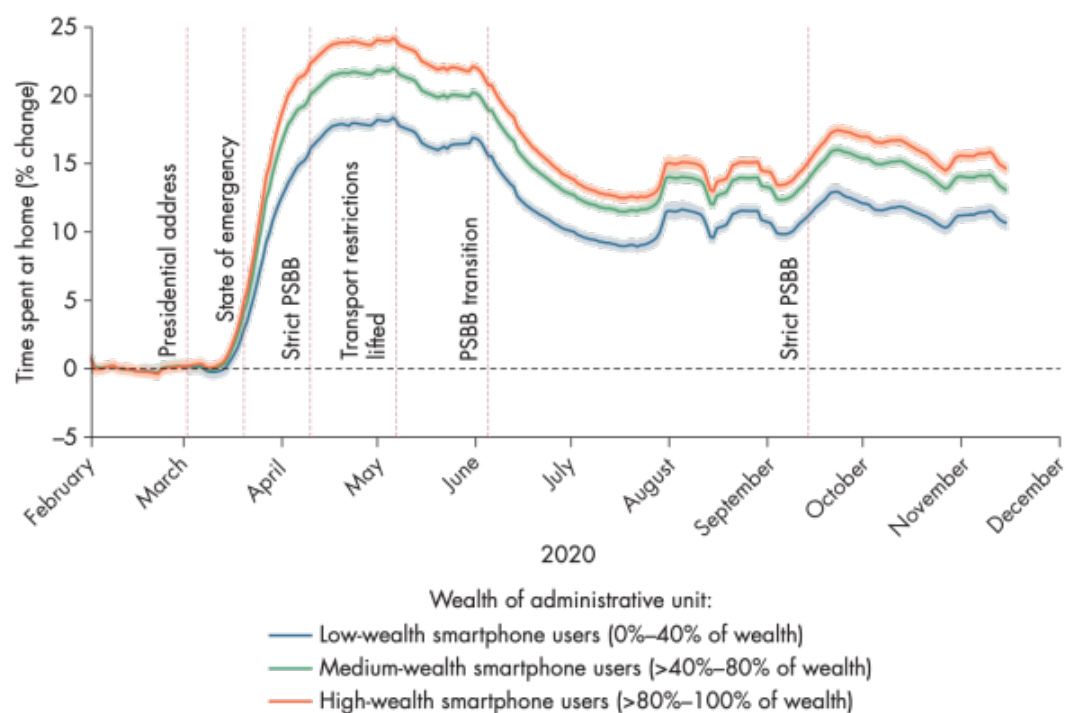
The home neighborhoods of users were identified and then matched to economic data. Users from the wealthiest 10% of neighborhoods and the poorest 10% of neighborhoods were used to compare the difference in how stay-at-home orders were followed across the economic spectrum. A clear differential is evident, with residents of wealthy neighborhoods reducing their daily mobility more than their neighbors from less affluent districts did.²⁸

Across the dataset, the average user increased their time at home by an additional 2.5 hours per day. However, those in the bottom decile stayed home only an additional 11% more than usual, while the wealthiest users stayed home for 20% longer than before the pandemic.

This data follows patterns seen elsewhere in the world from the first wave of the pandemic, where wealthier residents stayed home more easily than lower income workers, as well as evidence as the pandemic continued. Office workers and knowledge workers have been much slower to return to the office than retail, manufacturing, and other types of workers who cannot telework.

This type of information can be used by policymakers seeking to either adjust new orders impacting movement, and to assess the need for supplemental policies, such as social welfare payments, that may be required for lower income citizens to opt not to commute to work and risk losing income.

Figure 21: Impact of COVID-19 restrictions by income



Source: Fraiberger et al, 2020

²⁸ Samuel Paul Fraiberger, Nick Jones, Nancy Lozano-Gracia. 'Poor people respond differently to stay-at-home orders. Here's what the data says.' World Bank Sustainable Cities blog, June 23, 2020

Urban regeneration: accessibility constraints and local economic growth

Urban planners often place a high importance on accessibility when developing master plans or prioritizing local investments. Indeed, the concept of a ‘15 minute city’ in which jobs and amenities are available locally has gained traction. Mobility data can enrich our understanding of urban accessibility, as illustrated in San Jose, Costa Rica. Using observations of traffic speed and commute patterns, transportation constraints emerge as a key challenge for urban regeneration in the economically lagging Arco Sur districts.

With 2 million residents, the Gran Área Metropolitana (GAM) is Costa Rica’s vibrant capital region, centered on the historic city of San Jose. Recent decades have seen significant population growth in the GAM, in tandem with the region’s growth emergence as a diversified services economy that helped drive national-level economic growth. But as in most metropolitan areas, not every district of the GAM has shared equally in the increased prosperity. The region’s southern corridor (known as the Arco Sur) has lagged in terms of growth and job creation. Examination of mobility patterns in the GAM provides a window into urban dynamics and opportunities for policy initiatives to energize growth in lagging districts.

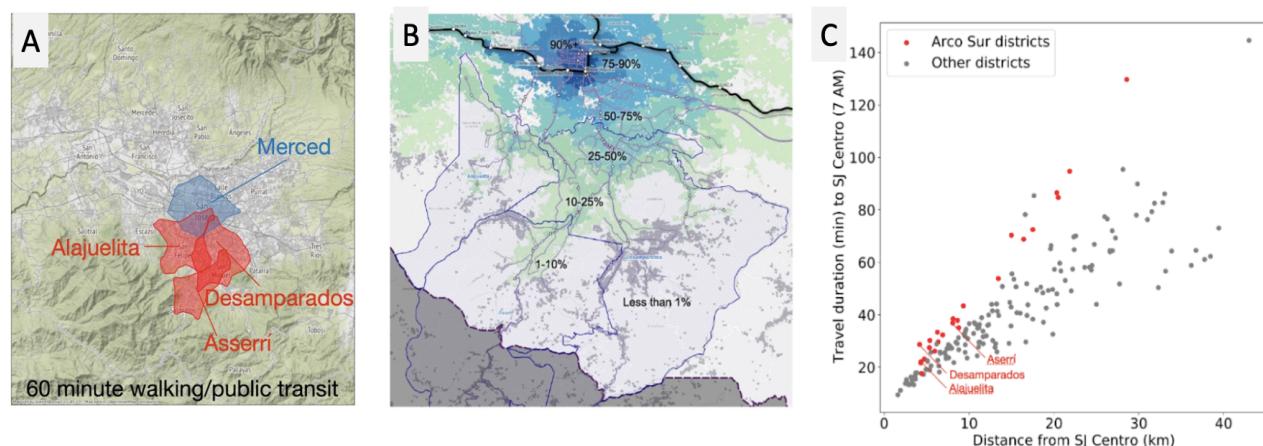
Cities as labor markets: how many jobs are accessible?

About 60% of Costa Rica’s population live in the GAM, a metro area that has developed from four historically distinct cities - San Jose, Cartago, Alajuela and Heredia. Having expanded its territorial footprint rapidly, the city is now tending towards densification. Public transport provision includes a train system running on an east-to-west axis as well as a bus network. For the Arco Sur districts (Alajuelita, Desamparados and Aserrí), travel to job locations and amenities is inherently more challenging due to hilly terrain.

Traditional urban geography methods highlight the destinations that can be reached in a 60-minute walk or public transit ride from the Arco Sur (Panel A). But using travel speed data and bus timetables - together with census data on job locations - a more nuanced picture of accessibility emerges. Residents of the most central regions can access more than 75% of jobs using public transport, but this proportion drops sharply in the Arco Sur. In Desamparados, for instance, residents can reach about 38,000 jobs compared to 85,000 for residents of Merced (a city center district).

Trips by car also take longer. Using travel speeds along San Jose’s roads measured by smartphone apps, we observe that trips from Arco Sur to San Jose city center are about 35% slower per km traveled (Figure 22C).

Figure 22: Inequalities in public transport and driving accessibility



Spotlight 2: Place Magnetism - Mapping Foot Traffic in a Redeveloped Public Space

La Sabana Metropolitan Park, known as “the lungs of San Jose”, is a park public park in the center of Costa Rica’s capital. The site was previously Costa Rica’s national airport. In 1977, a major redevelopment project transformed the space into an amenity-rich destination that enhances quality of life in the city.

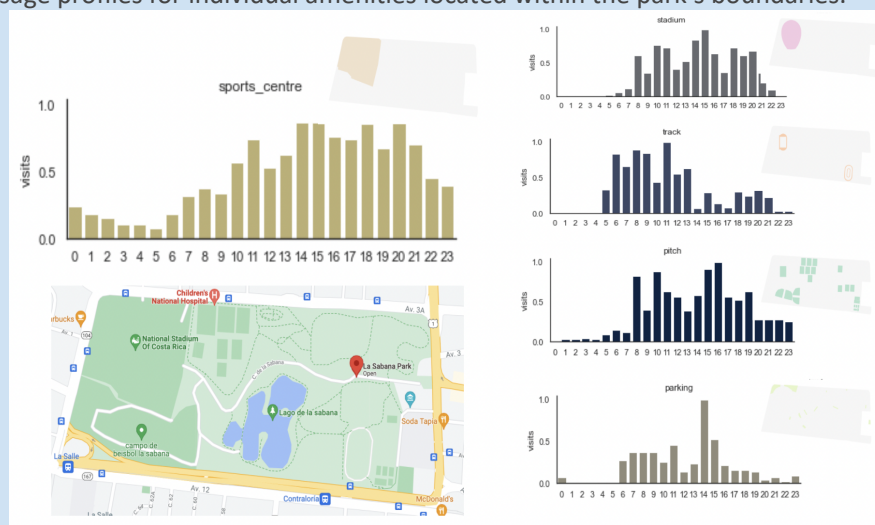
The site contains several distinct elements. Costa Rica’s national stadium is located on the upper-west quadrant; a swathe of soccer and baseball pitches occupy central areas; water features create a scenic environment for walks and picnics; and an art museum occupies the site’s eastern extremity.

Using a smartphone GPS movement dataset, the density of foot traffic across the park was visualized across a one-year period. The park as a whole is a well-utilized public space, with foot-traffic hotspots at prominent landmarks and key facilities.

Part of La Sabana’s character is its multi-functional nature encompassing nature, sports and public events such as concerts. Time-usage profiles show that the sports fields are utilized most extensively in afternoon and early evening hours. Running tracks attract early morning joggers. The national stadium, meanwhile, attracts many visitors in evening hours and at times of major events. Inspecting how visit profiles vary over time-of-day and day-of-week, a clear portrait emerges of the diversified nature of this public asset.



Figure 23: Foot traffic in La Sabana Park. Above: Foot traffic density of La Sabana Metropolitan Park. Below: Time usage profiles for individual amenities located within the park’s boundaries.



4. Home and work locations: comparing spatial form across world cities

People's decisions on where to live and work are key forces shaping cities. Demand for housing drives property markets, while the number of jobs that workers can access is crucial for economic specialization. Urban planners and economists seek to promote vibrant cities whose spatial form supports a thriving society and economy. Household surveys can be used to map commuting patterns and understand where jobs are concentrated. However, surveys are costly and are most widely employed in higher-income settings. Mobility data offers the possibility to map job locations and understand urban form in a wider range of cities. This section presents a set of standardized maps, derived from GPS datasets, that depict the spatial distribution of homes and workplaces for ten world cities.

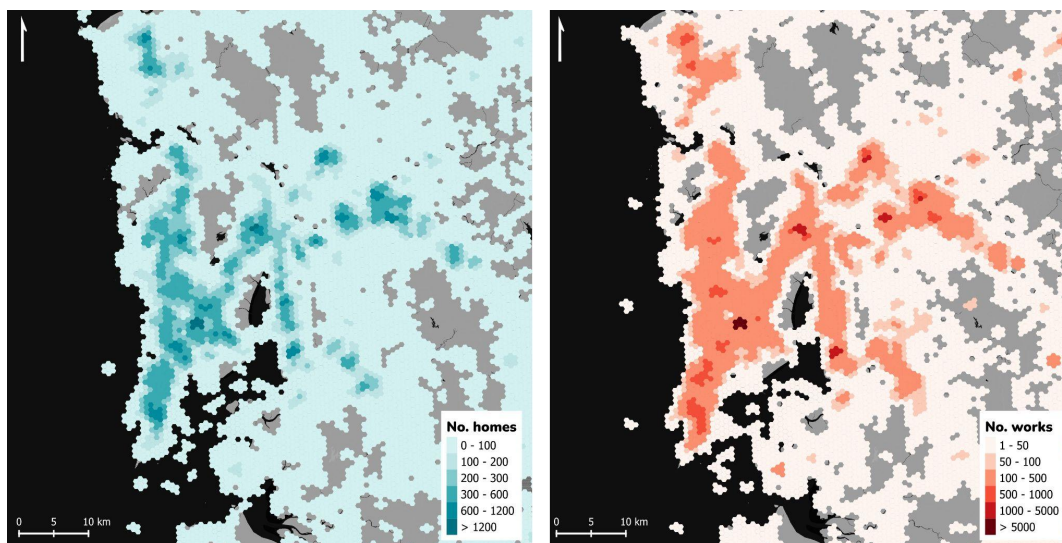
Urban mobility: classic models

In his book *Order Without Design: How Markets Shape Cities*, urban planner Alain Bertaud presents several classic models representing how people move across cities. In classic urban economics models, cities are assumed to be monocentric, with a daily movement pattern from the periphery (where most residences are located) to the center where most jobs are. Alternatively, cities may follow a polycentric model with several centers of employment attracting daily inward trips. Lastly, urban planners have at times sought to advance an 'urban village' model with multiple employment centers each surrounded by closely linked residential areas²⁹.

Identifying home and work locations

Using smartphone GPS location data for Mumbai, Figure 24 presents estimated home and work locations for a sample of local residents. The dataset was first filtered for data quality; home and work locations were then imputed based on the most frequently visited locations during night-time and day-time hours.

Figure 24: Number of estimated home and work locations for Mumbai.



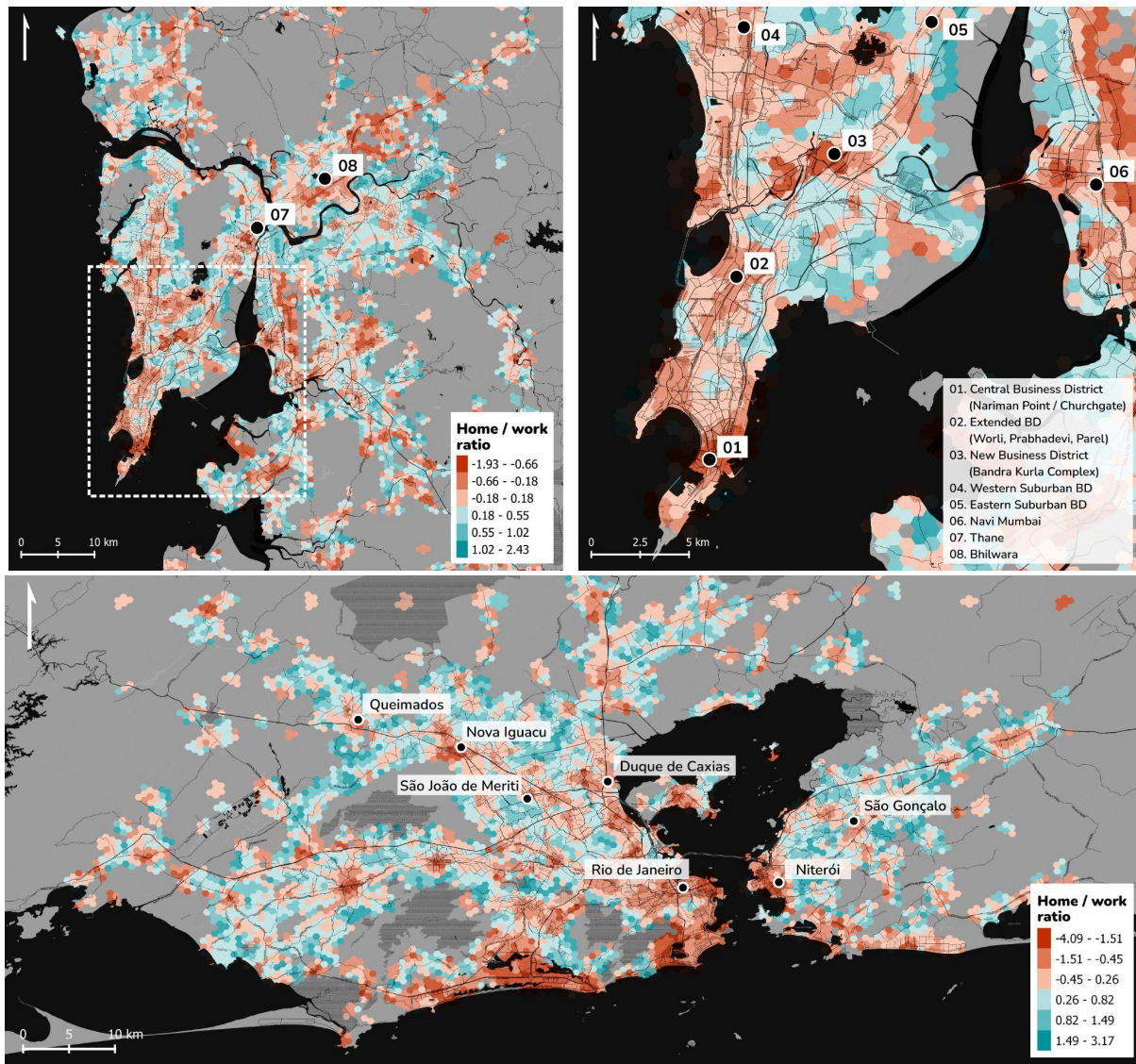
²⁹ See Figure 27 below.

Distinguishing monocentric vs polycentric cities

The ratio of home to work locations provides a convenient way to identify which locations serve as a city's economic hubs. Areas with more jobs compared to residents attract daily inward traffic. Figures 26 and 27 present the ratio of home to work locations for Mumbai, Rio de Janeiro, Doha and Riyadh. Dark red indicates areas with a high ratio of workplaces to homes; dark green shows more homes than workplaces.

Mumbai stands out as a markedly polycentric city. Major workplace centers include the historic administrative and financial core of the city on the Mumbai Peninsula (including Nariman Point and Churchgate), but also more recently developed commercial hubs such as the Bandra Kurla Complex. Workplace hubs can also be observed in Navi Mumbai, Thane and Bhilwandi, reflecting employment growth in formerly peripheral regions. Rio de Janeiro also shows a highly polycentric form: a strong workplace hub is in the Zona Central (the historic administrative and financial center) with significant hubs in Nova Iguaçu, Duque de Caxias, and Niterói.

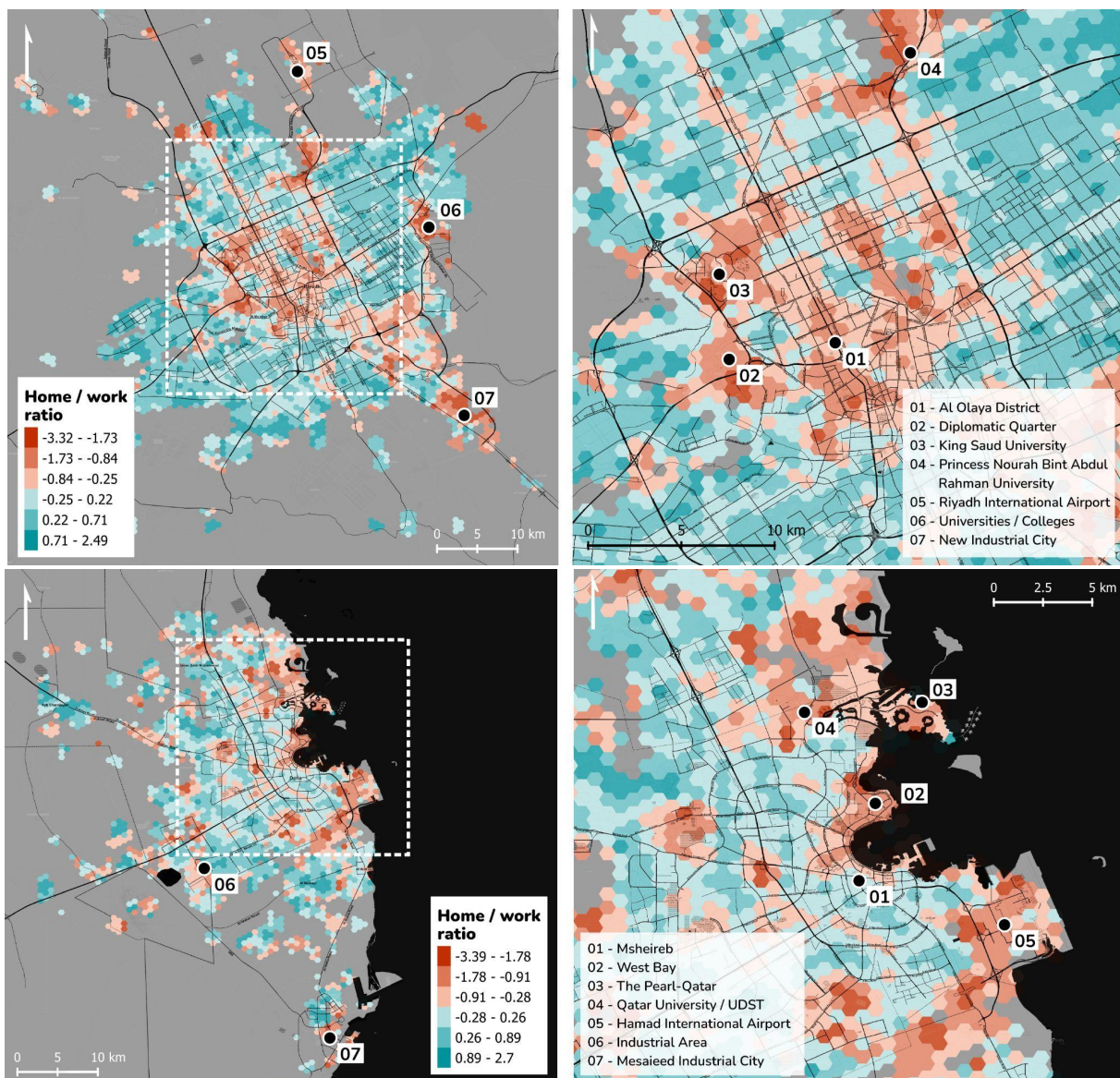
Figure 25. Home and work ratios in polycentric cities: Mumbai (top) and Rio de Janeiro (bottom).



Monocentric cases: Doha and Riyadh

By contrast, Doha and Riyadh display a more monocentric structure. In Riyadh, work locations are concentrated in the Al Olaya district and Diplomatic Quarter - and along the central portion of King Abdullah Road – with minor hubs in outer sections of the city corresponding with industrial estates and higher education campuses. Residential locations (blue) are arranged all around the core, spreading outwards into the periphery. In Doha, the workplace locations are concentrated along the coastline in the historical Msheireb District and the prime commercial locations of West Bay and Pearl-Qatar, with home locations mostly found in peripheral areas.

Figure 26. Home and work ratios in two monocentric cities: Riyadh (top) and Doha (bottom)



Land use regulations: theory vs practice

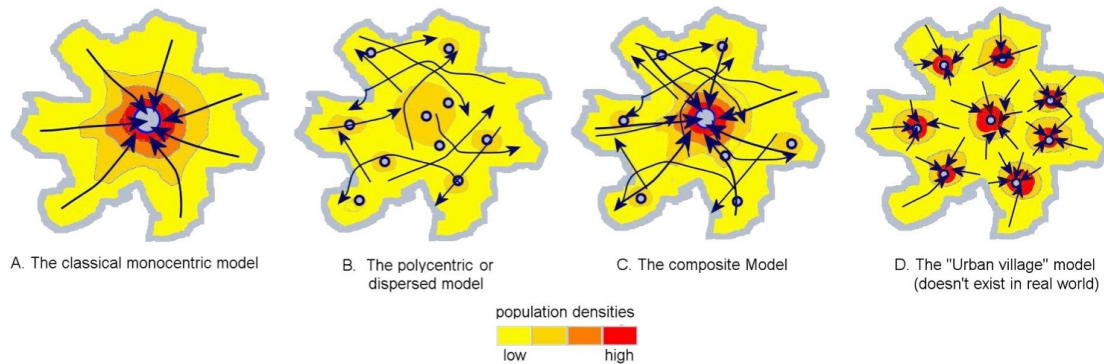
Governments intervene in urban land and property markets in a variety of ways. These range from site-specific investments – such as transport infrastructure and urban regeneration schemes – to regulations, such as zoning schemes that designate plots of land for residential, commercial or industrial use. Master plans seek to shape real estate investments by signaling commitment to a vision of the city's spatial development.

Such interventions often assume a model of how people move – or should move – across the city (Figure 27). Human mobility data presents an opportunity to 'reality check' how far actual movements correspond with planners assumptions, for example:

- Are 'new cities' intended as commercial hubs actually receiving inward trips during the work day?
- Does the traditional city center remain the primary attractor of work trips, rather than newer 'edge cities'?

This section has presented preliminary outputs from analyzing home and work locations across cities. Increased accessibility of mobility data promises to make such questions increasingly answerable.

Figure 27: Mobility within cities - four conceptual schemes



Source: Order without Design: How Markets Shape Cities. Alain Bertaud, 2018.

5. Select bibliography

Human mobility: key concepts

Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. “Understanding individual human mobility patterns”. *Nature* 453, 7196 (2008), 779.

Luis M. A. Bettencourt. 2021. “Introduction to Urban Science: Evidence and Theory of Cities as Complex Systems”, *MIT Press*.

Marc Barthelemy. 2017. “The Structure and Dynamics of Cities: Urban Data Analysis and Theoretical Modeling.” Cambridge, UK: *Cambridge University Press*.

Julian Candia et al. 2008. “Uncovering individual and collective human dynamics from mobile phone records”. *Journal of Physics A: Mathematical and Theoretical* 41, 224015

Analysis methods

Yu Zheng. 2015. “Trajectory data mining: an overview”. *ACM Transactions on Intelligent Systems and Technology*, Volume 6, Issue 3.

Md Shahadat Iqbal, Charisma F Choudhury, Pu Wang, and Marta C González. 2014. “Development of origin–destination matrices using mobile phone call data”. *Transportation Research Part C: Emerging Technologies* 40.

Pappalardo, L., Ferres, L., Sacasa, M., Cattuto, C., & Bravo, L. 2021. “Evaluation of home detection algorithms on mobile phone data using individual-level ground truth”. *EPJ data science*, 10(1), 29.

Alexander, L., Jiang, S., Murga, M. & González, M. C. “Origin–destination trips by purpose and time of day inferred from mobile phone data”. 2015. *Transportation Research Part C: Emerging Technologies* 58, 240–250.

Scherrer, L., Tomko, M., Ranacher, P. & Weibel, R. 2018. “Travelers or locals? Identifying meaningful sub-populations from human movement data in the absence of ground truth.” *EPJ Data Science* 7.

Quantifying and mitigating biases

Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W. & Buckee, C. O. 2013. The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of The Royal Society Interface* 10.

Zhao, Z. et al. 2016. “Understanding the bias of call detail records in human mobility research”. *International Journal of Geographical Information Science* 30, 1738–1762.

Sveta Mulusheva, Daniel Björkegren, Leonardo Viotti. October 18, 2021. “Can we trust smartphone mobility estimates in low-income countries?”. *World Bank Data Blog*. <https://blogs.worldbank.org/opendata/can-we-trust-smartphone-mobility-estimates-low-income-countries>

Open source software and datasets

Luca Pappalardo, Filippo Simini, Gianni Barlacchi and Roberto Pellungrini. 2019. “scikit-mobility: a Python library for the analysis, generation and risk assessment of mobility data.” <https://arxiv.org/abs/1907.07062>

Enrico Ubaldi, Takahiro Yabe, Nicholas K. W. Jones, Maham Faisal Khan, Satish V. Ukkusuri, Riccardo Di Clemente and Emanuele Strano. 2021. “Mobilkit: A Python Toolkit for Urban Resilience and Disaster Risk

Management Analytics using High Frequency Human Mobility Data.” *KDD Humanitarian Mapping Workshop*, <https://arxiv.org/abs/2107.14297>

Paige Maas, et al. 2019. "Facebook Disaster Maps: Aggregate Insights for Crisis Response & Recovery." *KDD*. Vol. 19.

Eugenia Giraudy, et al. December 22, 2021 (blog). "Making Our Displacement Maps More Representative." *Meta Research blog*, <https://bit.ly/3t2V3L1>.

Mobility data and natural disasters

Xin Lu, Linus Bengtsson, and Petter Holme. 2012. "Predictability of population displacement after the 2010 Haiti earthquake." *Proceedings of the National Academy of Sciences* 109, 29, 11576–11581

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

Takahiro Yabe, Nicholas K. W. Jones, P. Suresh C. Rao, Marta C. Gonzalez, and Satish V. Ukkusuri. 2021. "Mobile Phone Location Data for Disasters: A Review". *Computers, Environment and Urban Systems*, Volume 94

Shenyua Jia, et al. 2020. "Patterns of population displacement during mega-fires in California detected using Facebook Disaster Maps." *Environmental Research Letters* 15.7.

Case studies

Princeton University and the International Organization on Migration. 2021. "Assessing the Use of Call Detail Records (CDR) for Monitoring Mobility and Displacement".

Batran, Mohamed, et al. 2018. "Inference of human spatiotemporal mobility in greater maputo by mobile phone big data mining." *ATT@ IJCAI*.

Franco Basso, et al. 2020. "Accessibility to opportunities based on public transport gps-monitored data: The case of Santiago, Chile". *Travel Behaviour and Society*, Volume 21, <https://doi.org/10.1016/j.tbs.2020.06.004>.

Tracey Li, Rachel Bowers, Omar Seidu, Gloria Akoto-Bamfo, David Bessah, Victor Owusu and Laurent Smeets. 2021. "Analysis of call detail records to inform the COVID-19 response in Ghana—opportunities and challenges." *Data and Policy*, Volume 3

A. Akhavan, N. E. Phillips, J. Du, J. Chen, B. Sadeghinassr and Q. Wang. 2019. "Accessibility Inequality in Houston". *IEEE Sensors Letters*, vol. 3, no. 1.

Governance and data ethics

The Open Data Institute. 2021. "The Use of Mobility Data for Responding to the COVID-19 Pandemic DATA4COVID19 Deep Dive", http://theodi.org/wp-content/uploads/2021/03/Data4COVID19_0318.pdf

Stefaan G. Verhulst, Andrew Young, Andrew J. Zahuranec, Susan Ariel Aaronson, Ania Calderon, and Matt Gee. 2020. "The Emergence of a Third Wave of Open Data How To Accelerate the Re-Use of Data for Public Interest Purposes While Ensuring Data Rights and Community Flourishing". *Open Data Policy Lab*, <https://opendatapolicylab.org/images/odpl/third-wave-of-opendata.pdf>

Oliver, N. et al. 2020. "Mobile phone data for informing public health actions across the covid-19 pandemic life cycle." *Sci. Adv.* 6.