Movement Distribution

Last updated December 13, 2022

Overview

Movement Distribution Maps provide an overview of how far people in different regions travel from home on a given day.

Questions the dataset helps answer

- How far do people travel from home on average?
- How does this level of mobility vary with changes in disease prevalence, public-health messaging and travel policy?

Features of Movement Distribution Maps

- Updated daily
- Available at administrative level 2 (equivalent to counties in the US) wherever there are enough users to ensure privacy, or at level 1 (equivalent to states in the US) if level 2 is unavailable
- Built using a standard methodology for the entire globe
- Available to download in csv format for analysis and input into epidemiological forecasts and models

How movement distributions are calculated

Concretely, these maps begin by identifying a home Bing tile (based on location updates during nighttime hours, 8 PM - 6 AM for each person using the Facebook app who has turned on the Location Services device setting on their mobile device. The maps then choose a random location update for each person, which can have been logged at any time during the day. This is referred to as the visit tile. For each person, the maps compute the distance between the home and visit tiles.
Each home tile is mapped to a county-level Database of Global Administrative Areas (GADM) polygon, and then the percentage of users falling into 4 ranges of home-to-visit distances is reported to form a distance distribution:

1. 0km
2. Greater than 0km but less than 10km
3. At least 10km but less than 100km
4. 100km or more

Noise is added to the distribution to implement a differential privacy framework.

Movement Distribution Maps vs. Movement Range Maps

Movement Distribution Maps are a successor to Movement Range Maps, which we developed in 2020 to support response to the COVID-19 pandemic. The original Movement Range Maps were based on Location History data and provided partners with 2 metrics:

1. **Users staying put:** Fraction of people from a home tile who stayed there the whole day
2. **Change in mobility:** Percentage of people who stayed in a single location

The new dataset is based on Location Services data. With this data, there are fewer pings for each person, so it is easier for it to appear as if someone spent the whole day at home or visited fewer tiles. These considerations have, in part, motivated the switch to a new set of metrics, which in turn inspired a new name for this dataset: Movement Distribution. Its 0km value is comparable to the **users staying put** metric in Movement Range. There is no direct analogue in the new dataset of Movement Range's **change in mobility** metric, but the other categories of distribution (except for 0km) are intended to give a sense of how much people from a given home polygon are moving around.

Data standards

- **Population sample**: Facebook mobile app users who have turned on the Location Services device setting on their mobile device
- **Spatial aggregation**: US county-equivalent administrative boundaries. We use the territorial boundaries and names provided by the GADM project.
- **Temporal aggregation**: Distributions are aggregated over a single calendar day defined by the Pacific Time zone. The time windows are fixed and do not change based on local time zones.
- **Differential privacy:** Since we sample a single ping per person, the sensitivity of each count (i.e., the number of pings observed at a certain distance from home) is 1. This means that we can implement differential privacy with epsilon = 1 by adding Laplace noise with strength 1 to these counts, as well as to the sum of all the counts for each home polygon. As a technical point, differential privacy is implemented after the mapping of people to home polygons is fixed and is not incorporated into the estimation of people's homes.
- **Minimum counts**: We add noise to the counts of pings that are observed at different distances from home, as well as to the total number of pings that are observed for each home polygon. Because we sample one ping per person, prior to the addition of noise,

the latter quantity is equivalent to the number of people who are assigned to the polygon. We impose a threshold on the noisy version of that quantity, so that the person count with noise added must exceed 10 for the polygon to be included in the final data.
- **File format**: Data is provided in the format of a global comma-delimited text file.

Codebook

- **GADM ID (gadm_id)**: Unique identifier for the polygon from the Database of Global Administrative Areas (GADM)
- **GADM name (polygon_1_name)**: Name of the polygon based on the Database of Global Administrative Areas (GADM)
- **Country (country**): The 2-letter abbreviation (ISO alpha-2 code) for this row. The country value is assigned according to the Database of Global Administrative Areas (GADM) defining country boundaries.
- **Polygon level (polygon_level)**: Administrative level of this **country (gadm_country)** as defined in the Database of Global Administrative Areas (GADM).
  Using the United States as an example:
  GADM0=country (United States)
  GADM1=state (Florida)
  GADM2=county (Dade County)
  Whenever possible, we use level 2. For countries that don't have level 2 polygons, we fall back to level 1.
- **Home-to-ping distance category (home_to_ping_distance_category):** Indicates which of the 4 distance categories this row corresponds to. Distances are in kilometers and represent ranges. The 4 categories are:
  - 0
  - (0, 10)
  - (10, 100)
  - 100+
- **Ping fraction in distance category (distance_category_ping_fraction):** The fraction of pings from the polygon represented by **GADM ID (gadm_id)** that fall into the distance category indicated by **home-to-ping distance category (home_to_ping_distance_category)**. This fraction incorporates differential privacy noise. In rare cases, because of this noise, the fraction can be negative. The chance that the fraction becomes negative increases with the distance category: In the first 2 distance categories, the rate at which fractions are negative is much less than 0.1%, but for the largest distance category (100km or more), around 10% of fractions can be negative.
- **Date (ds)**: The date in Pacific Time for which the movement distributions are computed. For example, 8/8/2022 means the dataset aggregates over the period from 12 AM on 8/8/2022 to 12 AM on 8/9/2022 in Pacific Time.

Case studies and publications

These are based on the previous methodology.

- Mobility data to aid assessment of human responses to extreme environmental conditions
- COVID-19 and Unemployment Risk: Lessons for the Vaccination Campaign
- Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile
- Many more available on the Data for Good website

More information about this dataset

These are based on the previous methodology.

- Movement Range Maps overview
- Tutorial for identifying at-risk populations using open-source tools