Activity Space Maps

Last updated Jul 9, 2024

Overview

Activity Space Maps measure where people who live in a particular geographic tile spend their time. They are intended to support research into vector-borne diseases, such as malaria, dengue fever and zika. Researchers who study these diseases tend to have knowledge of where the vectors are present. For example, they may have datasets that report the prevalence of malarial mosquitoes in various geographical regions. However, such data needs to be supplemented with data on human movement for researchers to estimate (for instance) the likelihood that travelers will bring a disease back from a rural and swampy region to a densely populated area.

Isometric coordinates (rename, break up)

Activity Space Maps are written in terms of isometric coordinates, which are derived from latitudes and longitudes. Each change of 0.05 degrees in latitude or longitude increments the associated isometric coordinate by 1. For example, the Ferry Building in San Francisco is located at (37.7954° N, 122.3936° W). That longitude gets mapped to the isometric coordinate FLOOR(-122.3936 / 0.05) = -2448. Meanwhile, that latitude gets mapped to the isometric coordinate FLOOR(37.7954 / 0.05) = 755. Hence, the Ferry Building falls in the isometric tile (x, y) = (-2448, 755).

For each isometric tile, Activity Space Maps estimate how people from that isometric tile distributed their time spent over space (i.e., over the isometric tiles that they visited) over the three-week period ending on ds. A person's home tile is estimated from their nighttime Location Services pings over that same three-week period. Then, separate daytime and nighttime Activity Space Maps are built. The daytime map reports the distribution of where people from a given home tile spent time between the hours of Wam and Xpm. The nighttime map reports the distribution of where people from a given home tile spent time between the hours of Ypm and Zam. In both cases, the home tile assignments are based on nighttime pings; it is the pings that ultimately determine the distribution that vary.

For a given home tile, we only report the fraction of time spent in a visit tile if:

1. the visit tile is in the same country as the home tile
2. at least two people contribute to the (home tile, visit tile) combination
3. if a GADM polygon can be associated with the visit tile
4. if that polygon is not in a sensitive or disputed geographic region.

All (home tile, visit tile) combinations that don't meet these criteria are grouped into an "overflow" transition for each home tile, where the visit tile column is set to NULL. For all non-overflow rows, we add uniform noise from [-2, 2] to the counts of observations for that (home tile, visit tile) combination. Then, we adjust the overflow rows so that the total number of pings for that home tile matches the empirically observed count. The counts are then normalized by that empirical count to yield fractions. This means that the final "fraction" for an overflow row can be slightly negative in rare cases. Finally, we only report data for a home tile if at least 10 people were assigned to that home tile.

An important point about the original Location Services data is worth calling out: the raw data reports counts of pings observed in so-called FB tiles. FB tiles are a proprietary tiling of Earth's surface. Furthermore, some randomization over FB tiles is built into the raw data itself, such that a ping is sometimes reallocated from one FB tile to another nearby tile. FB tiles are roughly 1.2km x 1.2km patches of the Earth's surface at the equator. Our pipeline maps the centroids of FB tiles to isometric tiles. That means that some imprint of the original FB-tile grid will inevitably be present in the final data, even though the data is not presented in that format.

Questions the dataset helps answer

- Where do people from a particular home location spend their time during the day?
- Where do people from a particular home location spend their time during the night?

Features of Activity Space Maps

- Updated each Monday, incorporating three weeks of data
- Available at the isometric tile level, corresponding to approximately 5.5km x 5.5km squares at the equator.
- Built using a standard methodology for the entire globe
- Available to download in csv format for analysis and input into epidemiological forecasts and models

Early analysis

Data standards

- **Population sample**: Facebook mobile app users who have turned on the Location Services device setting on their mobile device
- **Spatial aggregation**: Isometric tiles, as described earlier in this document..
- **Temporal aggregation**: Distributions are aggregated over a three-week period.

- **Censorship of (home tile, visit tile) pairs:** The amount of time that people from a home tile will spend in a visit tile will only be reported explicitly if:
    - the visit tile is in the same country as the home tile
    - at least two people contribute to the (home tile, visit tile) combination
    - if a GADM polygon can be associated with the visit tile
    - if that polygon is not in a sensitive or disputed geographic region.
- **Minimum counts**: There must at least be 10 people who are assigned to a home tile for data to be reported for that home tile..
- **File format**: Data is provided in the format of a global comma-delimited text file or GeoJSON.

Codebook

- **Date (ds)**: The last date of a three-week period over which Activity Space Distributions are computed
- **home_latitude**: The latitude of the centroid of an isometric tile. This is the home tile for which this row reports data.
- **home_longitude**: The latitude of the centroid of an isometric tile. This is the home tile for which this row reports data.
- **home GADM name (home_gadm_name)**: Name of the polygon into which the tile at (home_latitude, home_longitude) falls, based on the Database of Global Administrative Areas (GADM)
- **home GADM id (home_gadm_id)**: ID of the polygon into which the tile at (home_latitude, home_longitude) falls, based on the Database of Global Administrative Areas (GADM)
- **home polygon level (home_polygon_level)**: level of the polygon represented by home_gadm_id, based on the Database of Global Administrative Areas (GADM). Level 0 corresponds to countries, level 1 to states or provinces, level 2 to county equivalents, etc.
- **visit_latitude**: The latitude of the centroid of an isometric tile. This is the visit tile for which this row reports data.
- **visit_longitude**: The longitude of the centroid of an isometric tile. This is the visit tile for which this row reports data.
- **visit GADM name (home_gadm_name)**: Name of the polygon into which the tile at (visit_latitude, visit_longitude) falls, based on the Database of Global Administrative Areas (GADM)
- **visit GADM id (home_gadm_id)**: ID of the polygon into which the tile at (visit_latitude, visit_longitude) falls, based on the Database of Global Administrative Areas (GADM)
- **visit polygon level (visit_polygon_level)**: level of the polygon represented by visit_gadm_id, based on the Database of Global Administrative Areas (GADM). Level 0 corresponds to countries, level 1 to states or provinces, level 2 to county equivalents, etc.

- **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.
- **Visit Fraction (visit_fraction):** This is the fraction of visits from the home tile that are to the visit tile. Noise has been added as described above

Case studies and publications

What should be added here?

More information about this dataset

Here are some more details about the differences in the first version (built with Location History data) and current version (built with Location Services data):

Context:

- Prior to 2022, we were able to build Data for Good maps based on a dataset (let's call it **Dataset A**) of mobility traces for individuals who had consented to share such data with the company.
- This data enabled us to link together records for the same individual over multiple days.
- We were then able to aggregate this data over many individuals to produce maps describing large populations.
- This was useful, in particular, for Activity Space Maps. In this context, we wanted to identify a home location for each individual (based on nighttime pings over an extended period). Then, this person would contribute to an aggregated distribution which estimated how *all* people from that home location distributed their time spent over geographic space during the daytime or nighttime.
- After May 2022, Dataset A no longer offered the possibility of linking together data records for individuals across day boundaries.
- This necessitated our shift to **Dataset B**, which implements a different set of privacy protocols. In this dataset, we have access to weekly distributions of pings for a given individual. These distributions count the number of times that a person visited a tile in a given week, either during the day or during the night. However, we cannot tell, for a particular ping, which day or night during the week it was recorded. Furthermore, spatial noise is added to all pings in Dataset B.
- The data in Dataset B is structured in rows with the following information:
    - userid
    - week
    - daytime or nighttime
    - tile
    - ping count
- Since linkage across time is essential for Activity Space Maps, we migrated the calculation to Dataset B during 2022.

Calculation Procedure

1. For each (userid, week, daytime / nighttime) combination in Dataset B, keep those combinations where the total ping count over all tiles is at least 10. In other words, if a person logged at least 10 pings in a given week during the daytime (for instance), their daytime pings for that week are retained in the calculation. Otherwise, their pings are dropped.
2. Next, a person's ping distribution is scaled to mimic a distribution of days spent in each tile. For example, suppose a person's nighttime pings had the following distribution over three tiles: (77, 17, 6). This distribution will be scaled to represent 7 nights as follows: (0.77 * 7, 0.17 * 7, 0.06 * 7) ~ (5.4, 1.2, 0.4). Then, the pings are rounded to a whole number of days as (5, 1, 0).
3. The rounding procedure in step 2 now allows us to replicate the calculation that we used when we had access to Dataset A. The distribution of nighttime counts for a given person can be summed over three weeks of data to identify a nighttime modal tile, which is interpreted as that person's "home tile." Then, that home tile is joined back to the person's daytime and nighttime tile distributions to obtain this person's (home tile, visit tile) transition counts. Those transition counts then contribute to the overall distribution for that home tile.

Nuances

- Tiling systems
    - Dataset A reports data with (longitude, latitude) coordinates, and Dataset B is written in terms of proprietary FB tiles.
    - For Activity Space Maps, both these coordinates are mapped into so-called isometric tiles, which are defined in terms of fixed amounts of angular distance around the Earth. An isometric tile corresponds to approximately 5.5km x 5.5km at the Equator, and 7200 tiles are needed to circle the Earth.
    - Since FB tiles are mapped into isometric tiles, the data inevitably retains some memory of the FB tile system.
- Privacy and Data Security
    - The procedure that is described above describes how we estimate (home tile, visit tile, nighttime / daytime) transition distributions.
    - However, we do not report data where:
        - Fewer than 10 people are assigned to the home tile.
        - We cannot identify a country in which the home tile is located.
        - We cannot identify a country in which the visit tile is located.
        - The country of the home tile does not match the country of the visit tile (i.e., the row corresponds to international movement).
        - Only a single person from home tile is observed at visit tile.
        - Where either tile lies in a sensitive or disputed administrative unit.
    - We also add uniform random noise [-2, 2] to the ping count for each (home tile, visit tile, nighttime / daytime) combination.

    - The considerations above mean that the number of pings that are assigned to specific visit tiles for each (home tile, nighttime / daytime) combination differs from the original number. These “missing” pings are collected in an “overflow” row with visit tile = NULL. Note that this “overflow” row includes pings that correspond to international transitions, pings that lie in very sparsely visited tiles, and pings that were added via noise. Due to the noise, this overflow row can in some cases report a negative visit fraction.
- Scaling by Dataset A
    - Dataset A no longer allows us to join together pings for a specific individual over multiple weeks, but we can use it to obtain an aggregate three-week distribution of pings over isometric tiles.
    - We can obtain the same from Dataset B, and we can compare the two distributions.
    - We can use that comparison to scale Dataset B, to better mimic Dataset A. In particular, we scale each person’s number of visits to a particular isometric tile in Dataset B by the ratio of overall visits to that tile in Dataset A to overall visits in Dataset B.
    - For example, suppose a person logged 10 daytime visits to tile X in the raw data of Dataset B. In Dataset A, aggregated over all people for three weeks, there are 100 visits to tile X. Meanwhile, in Dataset B, there are 10000 visits to tile X. Then, the person’s 10 visits get scaled by the ratio 100 / 10000. As such, their visits get scaled down to 0.1. The main effect of this scaling is to remove certain spurious tiles that get populated in Dataset B due to the addition of spatial noise. As such, the scaling factor is capped at 1. In other words, if a tile has more pings in Dataset A than in Dataset B (overall), then we don’t scale the number of visits to that tile.
    - This scaling happens before the entire calculation procedure that was outlined above.

Bug Found July 2024

Message to partners sent AUG 1 2024:

Hello friends,

I’m writing as you are one of the people who has access to the Activity Space Maps datasets offered by Data for Good at Meta (preprint here for reference). Unfortunately, I’m reaching out as we recently discovered a bug in the dataset that has been present since the first date partition provided to you covering Activity Space Maps data for 2023-04-17.

In brief, this bug causes tiles that are located near borders between administrative polygons to be assigned to an incorrect (but geographically nearby) polygon. The data at the tile level is trustworthy, but the following columns (home_gadm_name, visit_gadm_name, home_gadm_id, and visit_gadm_id) will not correspond correctly to the home_latitude, home_longitude, visit_latitude, and visit_longitude columns in these border areas. In some sense, this can be thought of as a larger than intended amount of spatial noise that is introduced in tiles near the borders of administrative polygons.

Our understanding from previous conversations is that your work is in-progress, and still relatively early stage. Please use caution in your interpretation of the data at this time. We have already corrected this bug for future dates of the dataset starting July 29, 2024, and are assessing our options for retroactively improving the accuracy of the previously provided dates. We believe that it is, in principle, possible to repair most of the previously provided data, with the important exception of data that lies near international borders or disputed regions.

To help us assess the options for repairing previous dates, we are interested in learning a bit about your plans for using the data. If you could quickly let us know:

1. Are you intending to use this data in a longitudinal study? Or to put this another way, is it of significant importance for you to use Activity Space Maps data (specifically data from the bug affected polygon/gadm columns mentioned above) from earlier than July 29, 2024?
2. If “yes” to question 1, are there specific dates that are most critical for your work?

We are happy to discuss the nature of this bug and the potential options moving forward if you would like more detail. We want to thank those of you that reported noticing an issue and wish we had discovered the cause of this particular bug sooner. It was uniquely difficult to identify and track down. We apologize for this inconvenience and any difficulties this may cause for you. All the best!