Raw Data Sandbox

What Is the Raw Data Sandbox?

The research team produced a sample of raw location data from mobile devices, called Raw Data Sandbox. The Raw Data Sandbox is prepared in order to allow various data users and the broader researcher and practitioner community to better understand raw mobile device location data and improve their confidence and appreciation of relevant data products (e.g., origin-destination tables and others). The Raw Data Sandbox is a sample of raw data, anonymized and aggregated to small zones to protect privacy. We must note that all raw data flaws, such as duplicate observations and location jumps, are intentionally left as-is in the Raw Data Sandbox, so that the potential users can also have a better understanding over data issues and limitations.

What Can Users Do with the Raw Data Sandbox?

The entire research community can access the sandbox without any data access restrictions. The sandbox includes samples of raw location data for people and vehicles from different data providers. These samples benefit the users by showing them the raw data structure, data frequency, data coverage, and data flaws. The samples can be used to develop and test algorithms for data cleaning, trip-identification, sample weighting, and mode/purpose/socio-demographic imputations. The users can also submit algorithms developed using data sandbox or tested on the data sandbox to be applied on the protected full sample and get the aggregate-level validation back. This process can help make the entire research community involved in advancing the methodologies for utilizing mobile device location data.

What Data Are Included in the Raw Data Sandbox?

The Raw Data Sandbox includes anonymized raw location data for both person location and vehicle trips. The geographical coverage of the data in the Raw Data Sandbox is the Baltimore metropolitan area.

- Person location data: Person location data are generated by the location-based services (LBS) within mobile devices. LBS data are obtained from the interaction between smartphone apps and software development kits that are designed to record the device location. The data sandbox includes data from multiple data providers. An undisclosed share/subsample of all devices from each original data provider is included in the data sandbox, which also protects business-sensitive information for data providers. The temporal coverage is one week in July 2017 (July 23~29, 2017).

- Vehicle trip location data: Vehicle trip location data are generated from the GPS devices inside vehicles. The data include both passenger vehicles and trucks. The GPS devices frequently record the location of the vehicle and produce sightings. GPS location data from one major GPS data provider, separated by passenger cars and trucks, are included in the sandbox. The temporal coverage is one week in July 2018 (July 23~29, 2018).

The data provider partners contributing to the Raw Data Sandbox are AirSage (AirSage website), INRIX (INRIX website), and Cuebiq (Cuebiq website).

How Does the Raw Data Sandbox Protect User Privacy?

After an extensive discussion between MTI, FHWA, and data provider partners, the team decided to aggregate the raw location data to small hexagons to address privacy concerns. The original location data were aggregated to hexagons that cover the entire earth based on the Uber H3 indexing system (Uber's introduction to H3). Each location point was substituted with a hexagon index. Each hexagon index represents a hexagon covering a defined area on the map. The size of the hexagons depends on the H3 resolution, which can lead to small or large hexagons. In the Raw Data Sandbox, H3's resolution 7 was used, which divided the Baltimore metropolitan area into around 6500 zones. At this resolution level, the length of each edge of a hexagon zone is 1.22 km. The hexagons for the Baltimore metropolitan area can be seen in **Figure 1**.

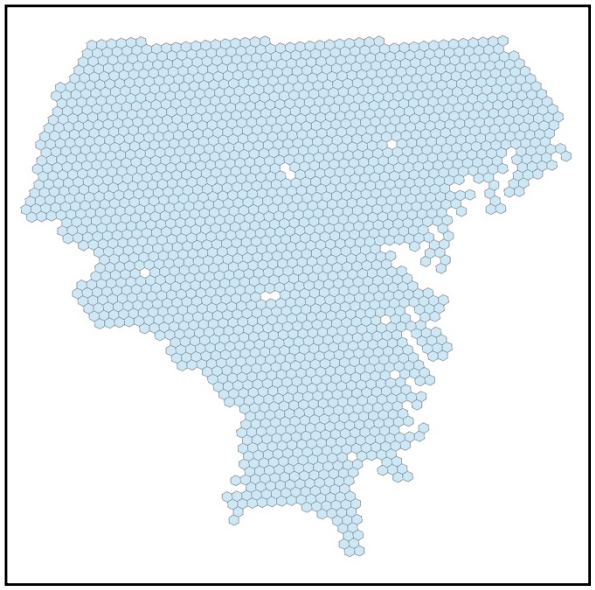

Figure 1[1]. H3 resolution 7 hexagon zones for the Baltimore metropolitan area

After this aggregation process, if the original dataset includes fewer than 10 devices inside a hexagon, all sightings inside the hexagon were removed from the Raw Data Sandbox to further protect the privacy.

How does the Raw Data Sandbox Look Like?

The Raw Data Sandbox has two components: the person location sandbox and the vehicle trip sandbox. The structure of each component is described below:

- *Person Location Sandbox:* This component is a CSV file containing anonymized location data from mobile devices. Each row of the file represents one observation of one mobile

[1] Section 508 Compliance: The figure is a hexagon-base zone system for the Baltimore metropolitan area using the Uber H3 indexing system. This figure shows the geographic distribution of all the hexagons.

device. Hashed device ID, time stamp, and hexagon ID are available for each location data observation. Hexagon ID's are based on Uber H3 zone system (Resolution 7). **Table 1** shows a sample row of the person location sandbox.

Table 1[2]. Cellphone data sandbox format

Device_ID	Time_stamp	Hexagon_ID
46f046aaceca8fec2770ce	2017-07-24 12:35:06	87f042129ffffff

- *Vehicle Trip Sandbox:* The raw location data from in-vehicle GPS devices form this sandbox component. The in-vehicle GPS data are different from the person location data in that these locations come in a trip trajectory format. For each trip, the origin, the destination, and the waypoints are available in the original dataset. The vehicle trip sandbox itself has two components: passenger cars and trucks. For each component, the sandbox includes a trip CSV file which has information on trip origin (aggregated to Uber H3 resolution 7 hexagons), trip destination (aggregated to Uber H3 resolution 7 hexagons), hashed device ID, hashed trip ID, and time stamp. Each component's sandbox also includes a waypoint CSV file which has information on waypoints (latitude, longitude, time stamp, trip ID, device ID). The waypoints falling into the origin and destination hexagons are removed from the sandbox for protecting privacy. The waypoint file can be linked to the trip file using the trip ID. **Table 2** and **Table 3** show sample rows of the trip CSV file and the waypoint CSV file respectively.

Table 2[3]. GPS data sandbox trip file format

Trip_ID	Device_ID	Start_Time	End_Time	Origin_Hexagon	Destination_Hexagon
22cf4ff57	0763a7	2018-07-23T16:30:07	2018-07-23T17:04:07	87f04216f	87f042a93ffffff

Table 3[4]. GPS data sandbox waypoint file format

Trip_ID	Device_ID	Time	Latitude	Longitude
22cf4ff57	0763a7	2018-07-23T16:38:07	39.27155	-76.7228

The data sandbox directory includes two folders, one for person locations and one for vehicle trips. The vehicle trips folder is further divided into two folders, one for passenger cars and one for truck. Each folder includes the required sandbox CSV files in addition to a readme file. The shapefile for the study area and the Uber H3 hexagons is also available in the sandbox directory.

Readers are encouraged to read the project final report to better understand the datasets and methodologies used for this product.

[2] Section 508 Compliance: The table shows Cellphone data sandbox format. It has three columns, Device_ID, Time_stamp, and Hexagon_ID.

[3] Section 508 Compliance: The table shows PS data sandbox trip file format. It has six columns, Trip_ID, Device_ID, Start_Time, End_Time, Origin_Hexagon, and Destination_Hexagon.

[4] Section 508 Compliance: The table shows PS data sandbox waypoint file format. It has five columns, Trip_ID, Device_ID, Time, Latitude, and Longitude.