Frequently Asked Questions (FAQs)

Cambridge Systematics, Inc. (CS) develops trip tables by employing thoughtful algorithms to transform raw location-based services (LBS) data into carefully validated travel patterns for different states and regions around the nation. This document serves as a reference for questions related to the data description, data processing, and expansion techniques.

Learn more about LOCUS on www.camsys.com/locus

CAMBRIDGE SYSTEMATICS

Data Description

Q1. What are location-based services (LBS) data?

Location-based services data are timestamped geolocation data generated by smartphone applications where users have explicitly granted permissions for the application to track geolocations. This data can be generated in the foreground (when the application is turned on) or the background (location information shared even when application is not actively used). LBS data are GPS-driven and therefore spatially very accurate. Our processed data are delivered within our LBS product called LOCUS.

Q2. If smart-phone GPS or app-based, are individual users tracked over time? If so, for how long? Are user IDs rotated to maintain privacy? If yes, what is the rotational schedule?

Location-based services data is purchased from a third-party vendor and then used to identify potential travelers through data science methods that process the data before providing the data through LOCUS. The third-party vendor resells data or tracks individuals over time using what is called a Marketing and Advertising ID (MAID), which is unique to the device, regardless of the app being used for data collection. There is no set amount of time when the ID gets reset. These IDs can be reset for a variety of reasons including the following: manual reset by the user, reset if the operating system is updated, and if the smartphone is disconnected.

Some devices have very little data in our dataset, which could be the result of limited app usage patterns, an ID that is reset very quickly, or a variety of other reasons. That is why our algorithms and filters that we apply are so critical to ensure that only the most robust and usable data are retained. Typically, among devices we find to be usable, the device IDs are active between 2 and 6 months.

It is important to note that the data provided through LOCUS is aggregated (spatially and temporally) and contains no personally identifiable information.

Q3. How does CS ensure data privacy in the development of LOCUS trip tables?

The data obtained by CS cannot be linked to a cell phone number or an individual. We use data from apps where people have explicitly opted in for location tracking, and we work only with vendors who employ rigorous standards to safeguard personally identifiable information. Further, any data distributed through our licenses are aggregated across spatial and temporal dimensions to add an additional layer of privacy protection before we share the data with our partners and clients. Privacy is our priority - whenever there is a choice to be made between privacy and accuracy, we have always and will continue to choose privacy.

Q4. What is the standard geographic scale/resolution?

While the raw data is collected using GPS technology and is spatially very accurate, our standard products include Census Block Group level detail for origin-destination flows. Based on the needs of a study, we allow and can provide custom zone selection, but we require that zone sizes be large enough to protect privacy and also to avoid small sample sizes.

Whenever examining the data at fine spatial resolutions, we always look at sample sizes to determine the level of detail that the data can support. We will not provide fine-grained spatial data if the data are too limited.

Q5. Does the data screen out minors? For example, the data from devices registered to minors is not supposed to be tracked.

Our vendor follows all applicable laws and standards relevant to data tracking, aggregation, privacy and anonymization.

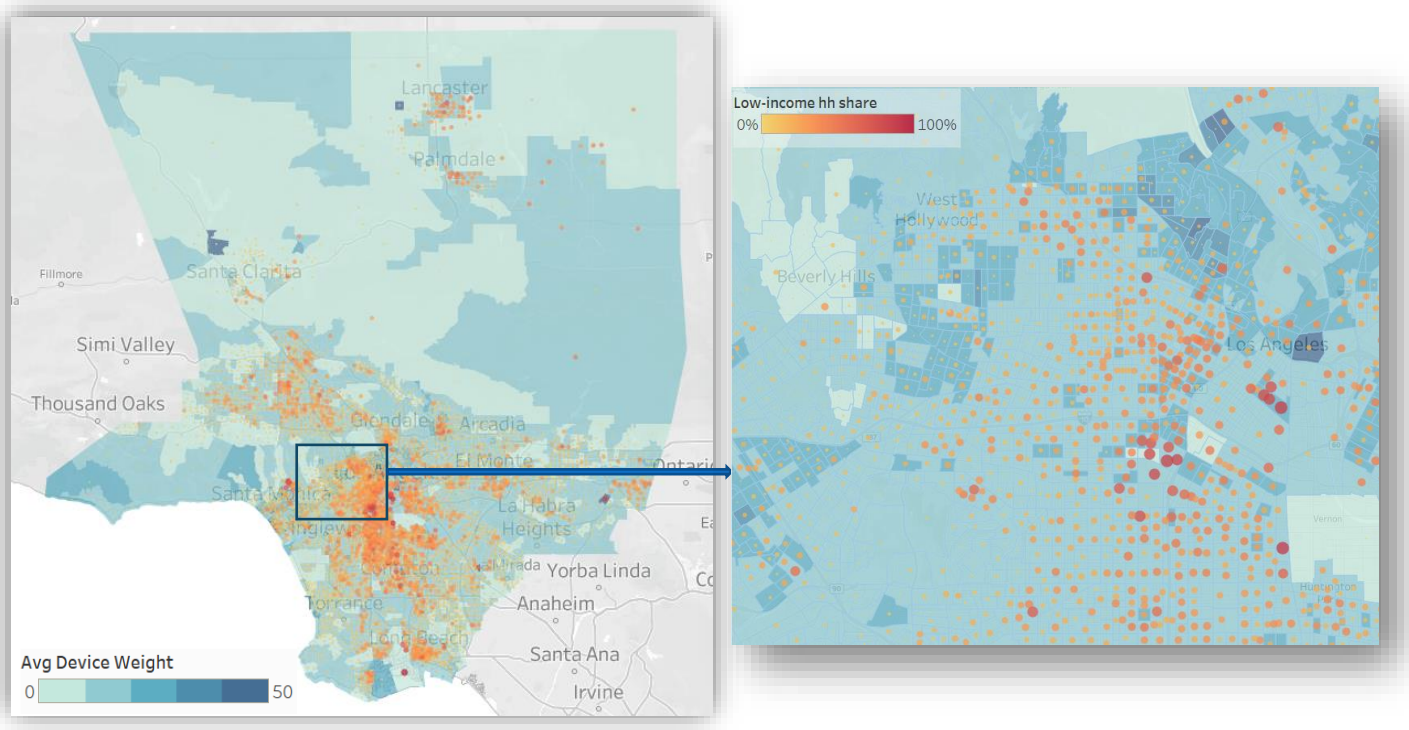
Q6. What is the sample size of the LBS data?

Very large – we often start with a sample that is 60-80% of the state or region's population each quarter. But all devices are not equal, in terms of quality and quantity of data. After applying rigorous analytics that retain only the most robust devices, we end up with an enriched sample of about 6-10% of the population. Importantly, each device has multiple days, weeks, and months of data allowing us to capture many trips made in the state.

Q7. How are the LBS data representative of LA County population and therefore, regional travel patterns?

Although we do not have information on individual device owners (due to the privacy protection protocols), we work with our vendors to ensure that the data are derived from both Android and Apple operating systems and from a broad spectrum of apps - multilingual, lifestyle, travel, dating, weather, gaming, and news (of all political leanings). We also perform data penetration and data quality checks at different Census geographies. More specifically, here are answers to questions that our clients have asked us:

- ***Does the data represent low-income population****? We believe so, because in most cases, low-income earners use smartphones as their primary computing device. A study by Pew Research (https://www.pewresearch.org/internet/fact-sheet/mobile/) indicates that the smartphone penetration for low-income population (less than $30,000) is 76%, which is comparable to overall market penetration of smartphones (85%). The following maps show the 2019 Q3 and Q4 device penetration and the distribution of low-income households (HH) based on the Census in Los Angeles county. As can be seen below, LBS data penetration rates appear consistent both among low-income and high-income.



- ***Does the data represent non-English speakers?*** *Yes, because the data are derived from apps that feature content in multiple languages (English and Spanish, amongst others).*
- ***Does the data represent respondents of all ages?*** Not always, groups that have lower smartphone penetration rates (61%) such as the elderly (65+) are likely not captured in this sample.
- ***Does the data represent people with disabilities?*** We are not sure, since we have no information about the demographics of individual device owners. To the extent that individuals with disabilities are able to operate and use smartphones, we believe that our data does capture their travel patterns.

Data Processing, Expansion and Validation

Q8. What are the known biases in the data samples? For example, does the data skew towards higher incomes?

We have examined our expansion weights by census tract and correspondingly looked at census tract demographics to determine if there is any correlation between demographics and penetration rates in our sample. We have not found any in most locations. However, this is something we monitor for each new client to ensure data quality. It should be noted that because we do not capture any user demographic information, our checks do not necessarily mean that biases do not exist; however, based on our expansion rate calculations, we believe that any biases that exist are not systemic.

Q9. How are activity stays and trips extracted?

The raw LBS samples/trajectory events are first processed and filtered to extract visit events, which represent a device in stationary state (as opposed to trajectory state, where the device is in motion). These visit events are then clustered to transform consecutive visits into longer "activity stay" observations, which represent a device's engagement in an activity (such as work, school, shopping, etc.). These records are cleaned further by checking for unreasonable speeds between stays. Stays were filtered if they did not meet a minimum stay duration criteria (10 mins) – however, this condition is relaxed under certain conditions to allow for shorter stops (such as coffee stop on the way to work). Once activity stays were identified, trips are inferred from the data by connecting consecutive activity stays completed by the same device.

Q10. What are the data sources used to perform data expansion?

The main control data sources used in the expansion process are ACS population and national estimates for employment at a Tract Level.

Q11. Is the sample size increasing over time? If so, how does this influence the data outputs?

When we process the data, we have found that the data sample size has remained steady, but the data quality of devices in the sample has improved over time. Because we expand the data based upon Census benchmarks, this has no real influence on the expanded data outputs, like numbers of trips or trip rates – however, we may be capturing more variability in travel (i.e., one-off vacation or long-distance travel).

Q12. How are the data expanded to the population?

Our processed data are expanded at the device level to replicate the travel patterns of the national population, including that of LA County residents and visitors to the region. We also scale/normalize trips to represent average daily (Weekday, Friday, Saturday, and Sunday) person travel.

Q13. How are the LBS trip patterns validated?

We validated our expanded travel pattern data against established, standardized external data sources such as the National Household Travel Survey (NHTS) as well as regional travel demand model if/when available. Validation metrics include trip rates, trip purpose, time of day, trip lengths, in addition to other travel patterns. This process ensures that no errors were made in processing the LBS data and provides assurance that the processed data make sense in comparison to more traditional datasets.