STREETLIGHT InSight

Larger and More Representative Samples:

How Big Data Can Support Equitable Transportation Analytics & Decisions

Version 1.0

October 2020

Table of Contents

Introduction

In recent years, Big Data resources and practices have provided significant opportunities to improve transportation planning and policy outcomes by broadening the scope of possible analysis. However, some researchers and practitioners have voiced concerns that Big Data may have biased samples (and thus, potentially biased or invalid metrics may be derived from Big Data). We agree that it is important to understand whether a data set used to make transportation infrastructure and policy decisions actually reflects the same underlying distribution of the population.

Fundamentally, we believe (and research backs this up) that more equitable data and sampling will lead to more equitable transportation policies and infrastructure planning. These policies and plans have generational impacts due to the life cycles of these investments.

As a very simple example, if no one from a certain neighborhood is included in a survey, then the planners may never know that it takes twice as long for that community to get to work, compared to the average, and will not take actions to correct this inequitable distribution of transportation and accessibility. If lower-income block groups are more often the ones not included, then a systemic bias in allocation of transportation planning can result.

The U.S. Department of Transportation includes a core principle of equitable transportation planning as "full and fair participation by all potentially affected communities in the transportation decision-making process." If data describing the experience of a group of people is not collected, they simply cannot fully participate.

Using representative data requires two steps:

1. Collecting a sample that has broad representation across all groups of the relevant population (minimize sample bias).
2. Using robust statistical and data science techniques to normalize that sample to scale up and represent the whole. All samples have some bias and must be normalized. The less bias you start with, the better the normalization can be.

In this white paper, we attempt to address these concerns related to the first point (minimizing sample bias) and show that the Location-Based Services (LBS) data that drives the StreetLight InSight® platform, sometimes known as Big Data, has a sufficient and representative sample across the country. While we summarize our normalization techniques here, much more information is available in our deeper methodology documents.

Sample Penetration Methodology Calculation

One of the many ways StreetLight considers our data set is to look at what percentage of the population our LBS devices represent. To determine device sample rates for LBS data, we estimate the number of useful devices in the sample that "live" in a particular area. Next, we compare that number to the region's total population per the most recent U.S. Census. The LBS device count used for the analysis in this paper was from November 2019. Population data was

taken from the 2010 Census population by county extrapolated to 2019[1] levels. Additional information about how closely StreetLight Metrics represent the population at large is available in our webinar about using LBS data for Environmental Justice Studies.

A device's home location is based on the device's location during nighttime hours, when people tend to be near their residences. Based on how many nighttime hours a device spends at a location, we assign a probability that the device is affiliated with a particular Census block. Because the home location is probability-based, there can be more than one home location for a device in a month. For example, a Census block may get a 50% probability of being a home location, another block 30%, and a third block 20%. Consistent with StreetLight's strong privacy guidelines, we do not have any personally identifiable data (just points in space and time) about device owners.

Next, we add up all the devices in our sample that are most likely affiliated with each Census block. If we assign 15 devices to a given Census block that 100 people live in, that means our device sample share for that Census block is 15%. Because published StreetLight Metrics include information only about groups of people, we aggregated all of the Census blocks in this study into tracts. For context, about 30 people live in the average Census block, and about 6,000 people live in the average Census tract.[2]

Overview of the Different Ways to Assess Coverage

The following sections give details on the different angles that are important to consider when assessing how robust a sample is.

Sample Penetration

To understand sample representativeness, we look at the penetration rates and how they vary for different populations. StreetLight's average LBS device penetration rate for the United States is 10.6%. Diving into regional penetration rates, as you can see in Figure 1 below, the distribution of counties by penetration rate is tightly centered around the average. In addition, Figure 1 shows that 82% of the counties have a penetration rate in the 7%-16% range, and 70% of the counties in the 8%-15% range. The median penetration rate is 11.6%, which is close to the average penetration rate, showing that there is no concentration of counties with extreme penetration rate values.

[1] "Population, Population Change, and Estimated Components of Population Change: April 1, 2010 to July 1, 2019 (CO-EST2019-alldata)" https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html#par_textimage
https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/totals/co-est2019-alldata.csv
[2] https://www.streetlightdata.com/big-data-supports-environmental-justice-in-transportation/

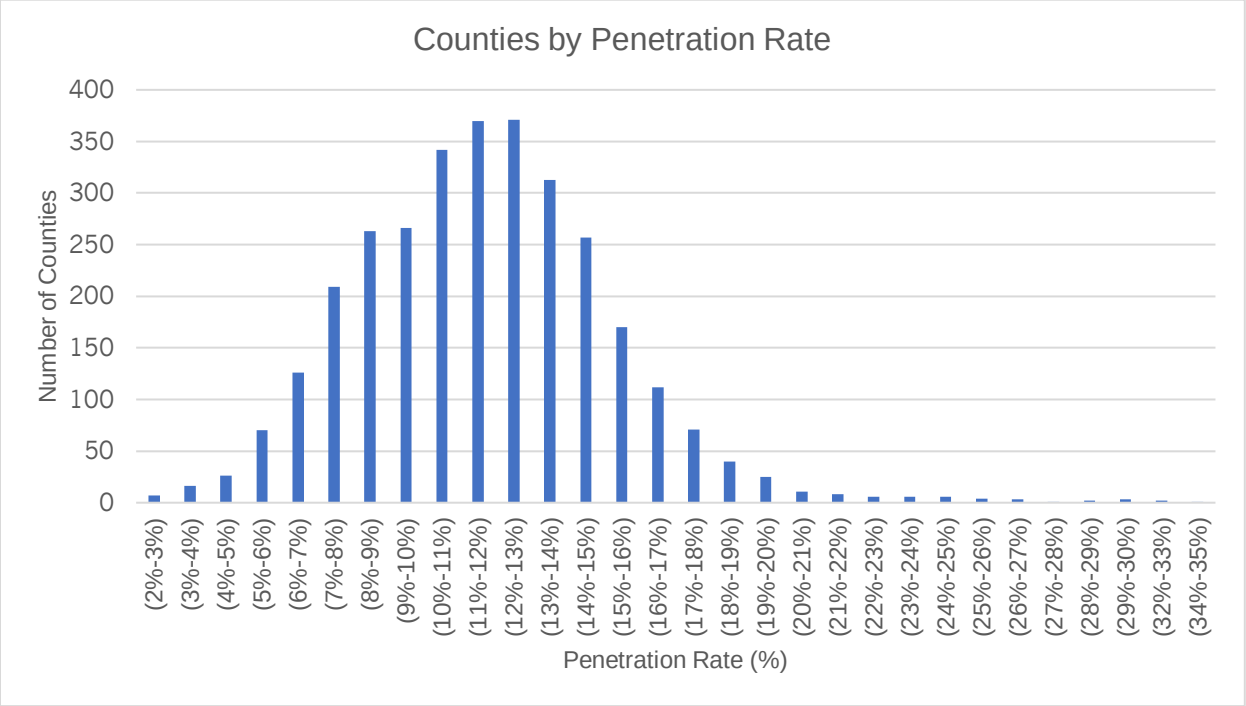

Figure 1: Comparison of number of counties by penetration rate.

Even more granular, as shown in Figure 2, the distribution of tracts by penetration rate is tightly centered around the average penetration rate of 10.6%. Approximately three-fourths of the tracts have a penetration rate in the 5%-14% range, 70% of the counties are in the 8%-15% range, and 38% are in the 7%-11% range. The median penetration rate is 9.8%, which is close to the average penetration rate, suggesting that there is no concentration of tracts with extreme penetration rate values. There are a few tracts with extremely high penetration rates – these generally occur in tracts with very low populations.

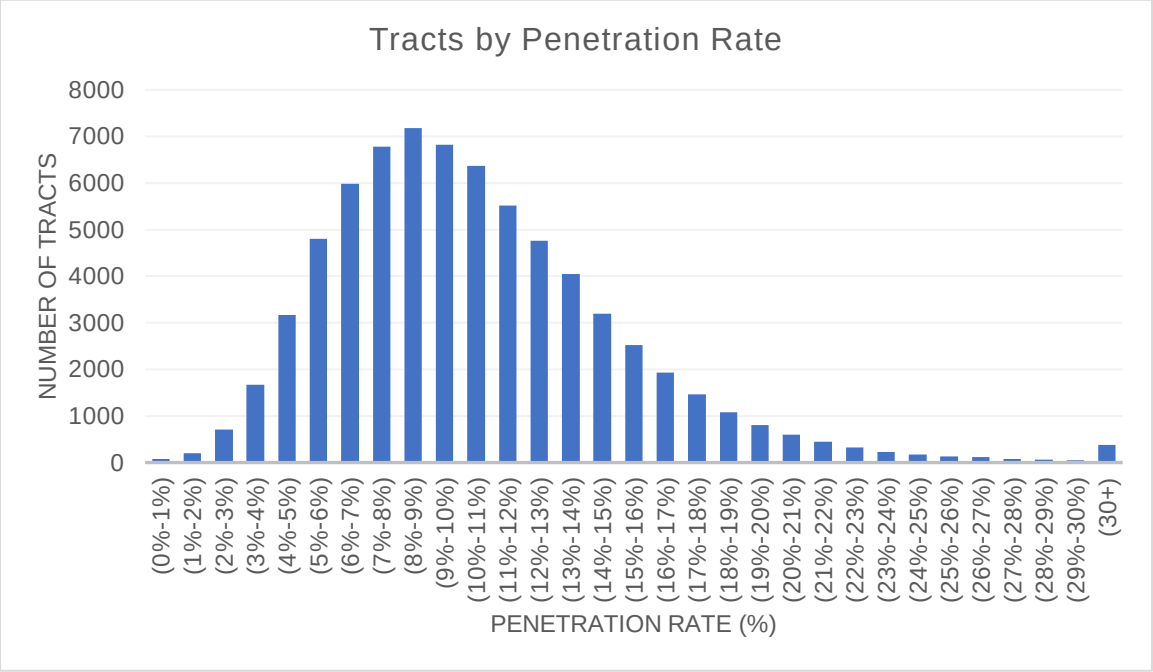

Figure 2: Comparison of number of Census tracts by penetration rate.

Urban vs. Rural Areas

There is a common misconception that Big Data sourced from LBS devices has less coverage in rural areas, compared to urban areas. (Learn more about how LBS data is widely available and useful for rural areas.) Due to the technologies used by LBS devices to provide location data (anonymized, of course), the StreetLight LBS penetration rate in rural Census tracts is slightly higher than that in urban tracts. The average device penetration rate for urban tracts is 10.2%, while that for rural tracts is 12.2%.

Note: A Census tract is defined as urban if more than 10% of it lies within the urban areas based on the cartographic boundaries outlined by the U.S. Census. The 10% threshold is used with an 80-20 urban-rural population split, because urban areas now account for ~80% of the U.S. population (learn more about the population split here).

To look at this from a granular regional perspective as well, we assessed the distribution of states and counties in terms of penetration rate in urban and rural tracts within each of the counties, as shown in the figures below.

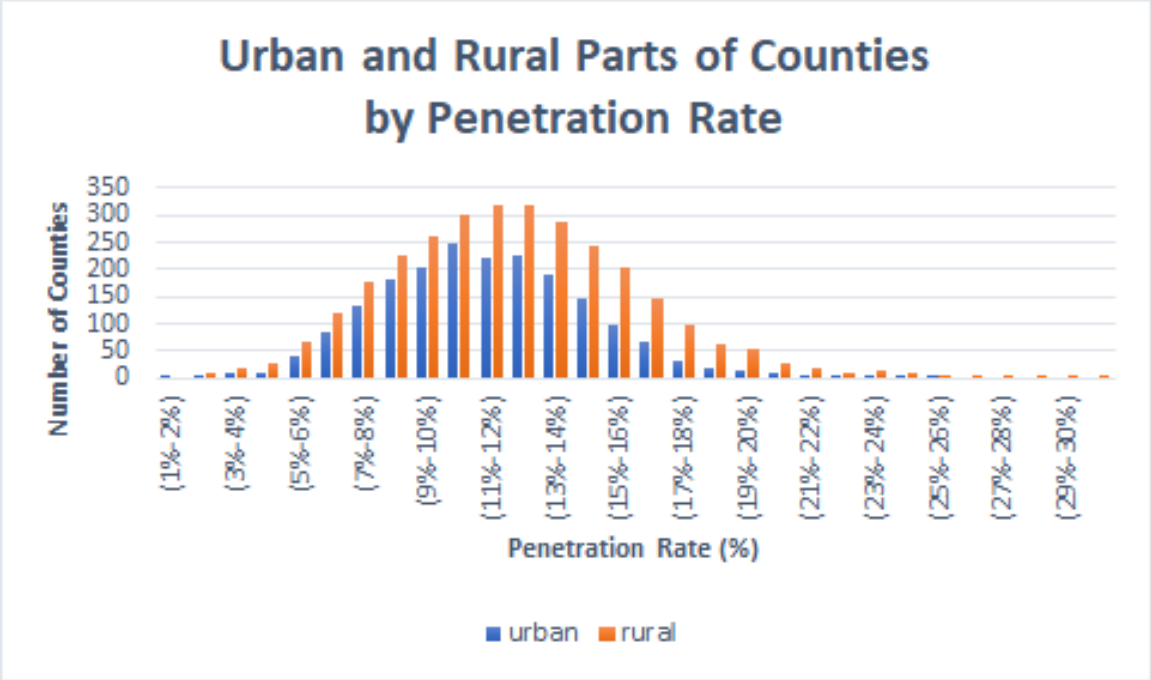

Figure 3: Comparison of number of urban and rural parts of counties by penetration rate.

Figures 4 and 5 below demonstrate penetration rate by geography, looking at states and counties.

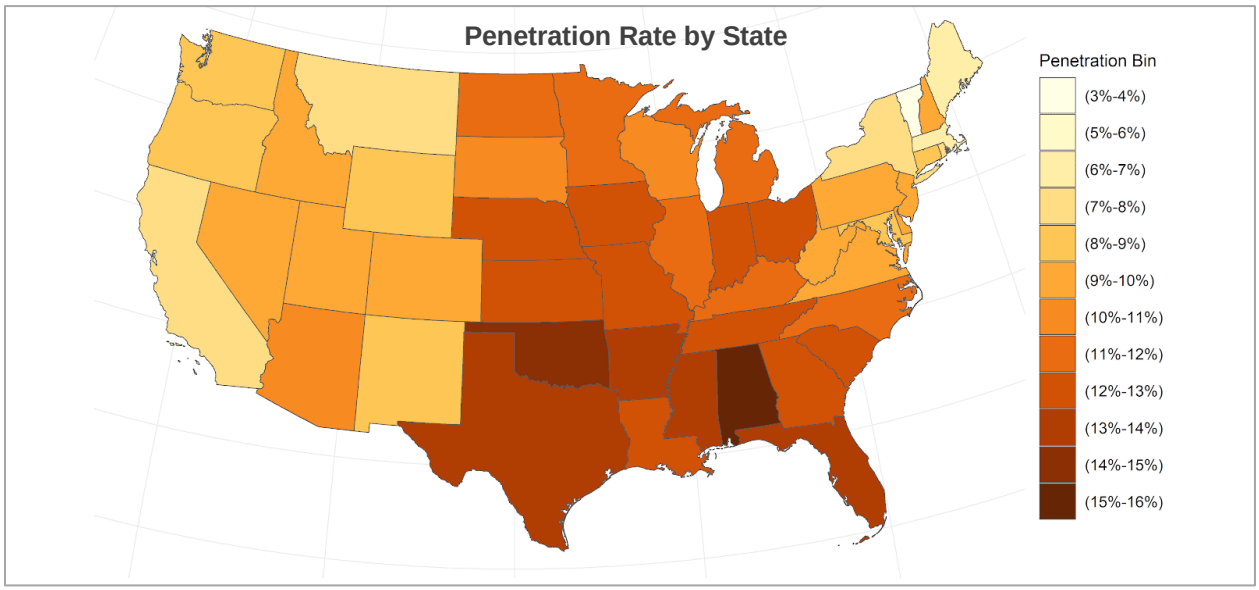

Figure 4: Heatmap of penetration rate by state.

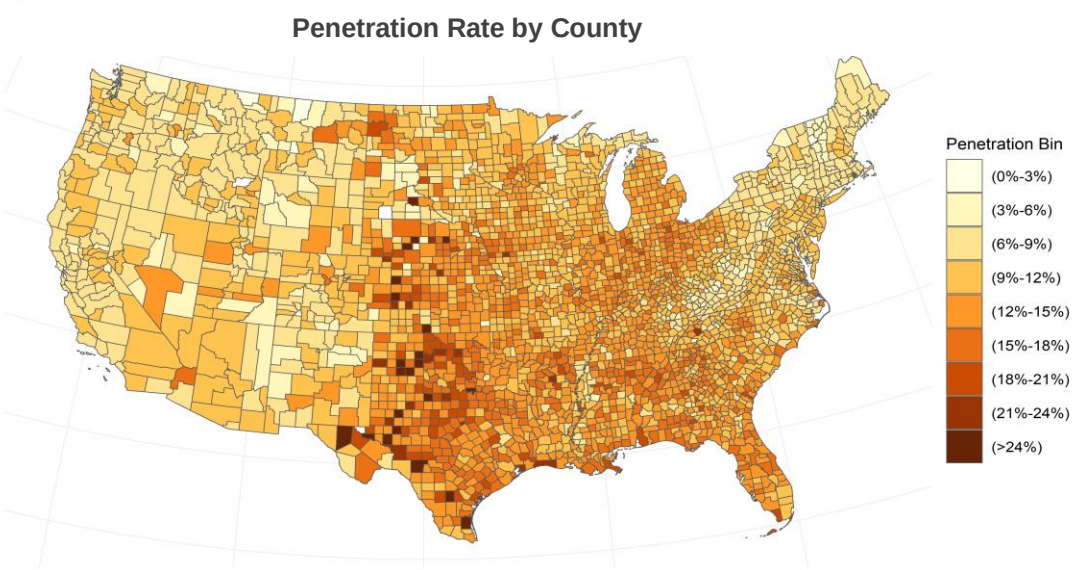

Figure 5: Heatmap of penetration rate by county (the white patches are counties with missing data).

Sample Representativeness Across Demographic Categories

Travel surveys face great challenges in ensuring that the sample is representative. The difficulty in selecting a sampling frame that is representative is made more difficult by the fact that surveys can cost-effectively capture only a very small percentage of a region's population. There are other limitations, such as surveying people only on a particular day(s), which means that they capture only those that travel on the days of the survey, and recall bias (when participants tend to remember certain types of trips more than others). In addition, surveys depend on an expected number of responses, and significant differences in response rates by geography, demographic groups, and mode-preference groups usually impact the quality of sample.

StreetLight's LBS sample addresses many of the above limitations of travel survey sampling by capturing trip details from more people every day of the year, as opposed to some people a few days of the year. Because the sample of trips is many times larger than surveys, it is inherently more likely to capture different groups of people traveling in different ways. However, we still ask the question, how representative is the StreetLight LBS data sample?

PENETRATION RATE BY HOUSEHOLD INCOME

We assessed how the LBS device penetration rate varies across tracts as the average household income varies. Figure 6 below shows the variation of average penetration rate for tracts with change in household income. The data labels show the number of census tracts that fall into each bin. The horizontal line shows the average penetration rate for all tracts in the U.S. As the chart shows, about two-thirds of the tracts have an average household income in the $50K-$125K range, and within these tracts, the penetration rate varies from 11% to 11.4%, which means that there is not much variation.

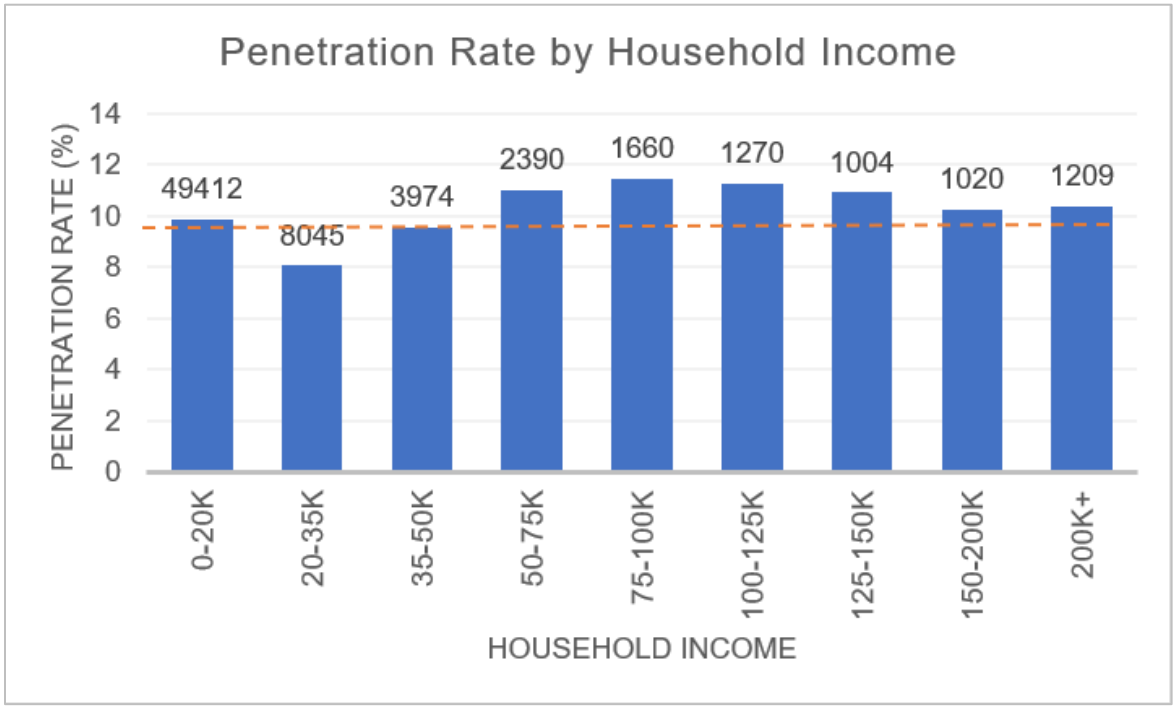

Figure 6: Comparison of penetration rates by household income, with data labels highlighting the number of Census tracts that fall into each bin.

There is a bit more variation in the $0K-50K ranges, but this is vastly less variation than seen in many surveys – for example, the recent Southeast Florida Household Travel Survey sample, shown below in Figure 7. More importantly, this relatively minor variation can be corrected with our normalization process.

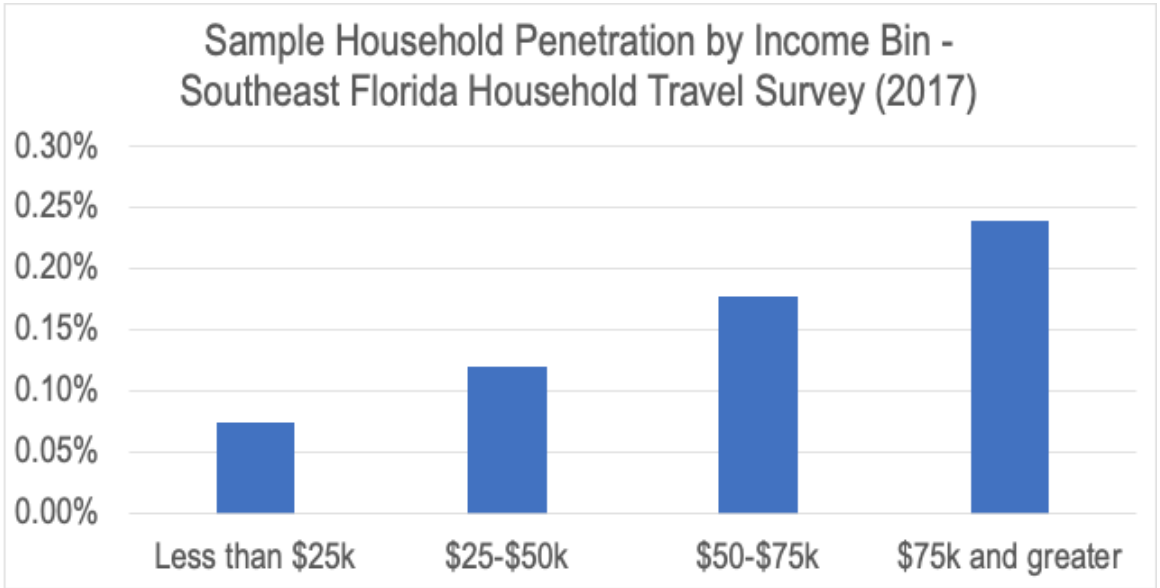

Figure 7: Southeast Florida Household Travel Survey sampled higher-income households at 3.2x the rate of lower-income households, unlike StreetLight, which has a very similar rate for high and low income. Also, StreetLight's sample covered ~100x more devices for many more days.

Another way of looking at this comparison is understanding if there is a correlation between the penetration rate by tract and the average household income. As shown in Figure 8, a simple scatterplot shows that there is no meaningful correlation. However, despite the lack of correlation, the scatterplot further colors the tracts by urban/rural classification.

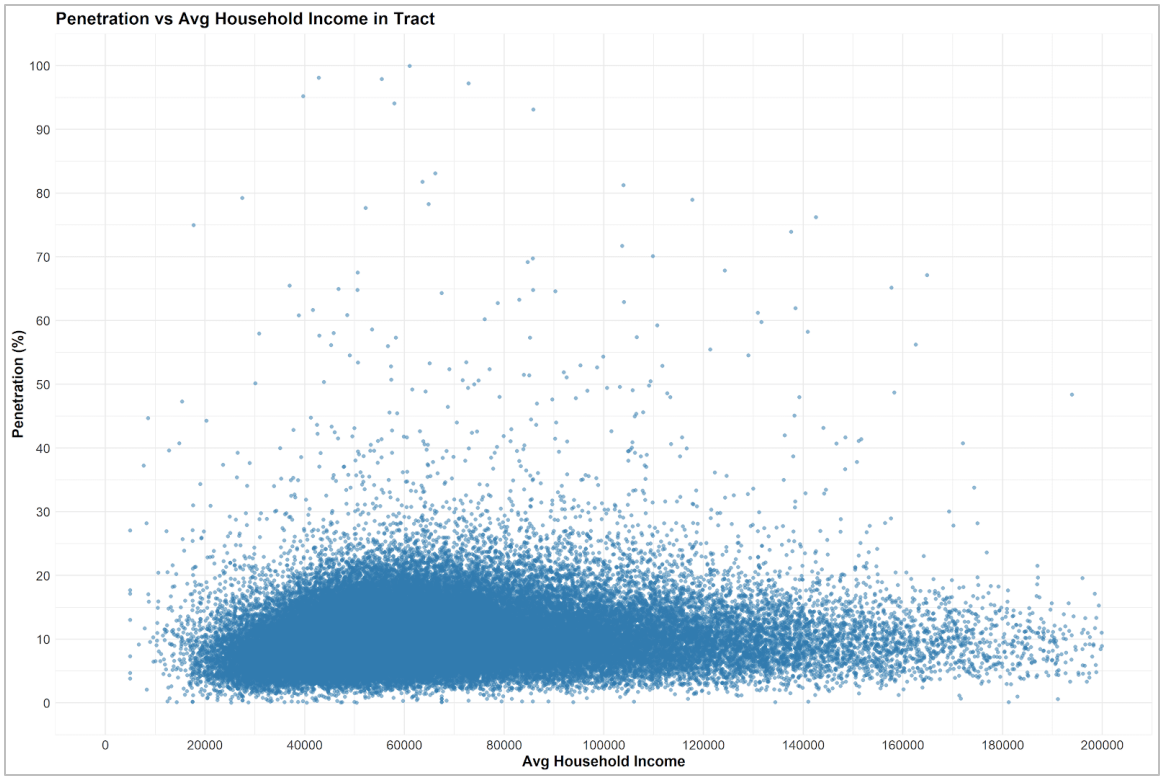

Figure 8: Scatterplot highlighting if the penetration rate by tract is correlated with the average household income.

PENETRATION RATE BY RACE

In addition to household income, we assessed how the LBS device penetration rate varies across tracts as the percentage share of population that belongs to a few races varies. The race categories we looked at were White, Black, Asian, and Hispanic.

The chart below shows the variation of average penetration rate for tracts when the percentage population in the tract that is White varies. The data labels show the number of census tracts that fall in each bin. The horizontal line shows the average penetration rate for all tracts in the U.S. As Figure 9 shows, our sample size increases with the share of the tract that is White. However, the range is between 7% and 12%, which is well within normalize-able range, and we do not believe it is manifested as biased or invalid final metrics. We will evaluate future data suppliers to mitigate this bias in sampling.

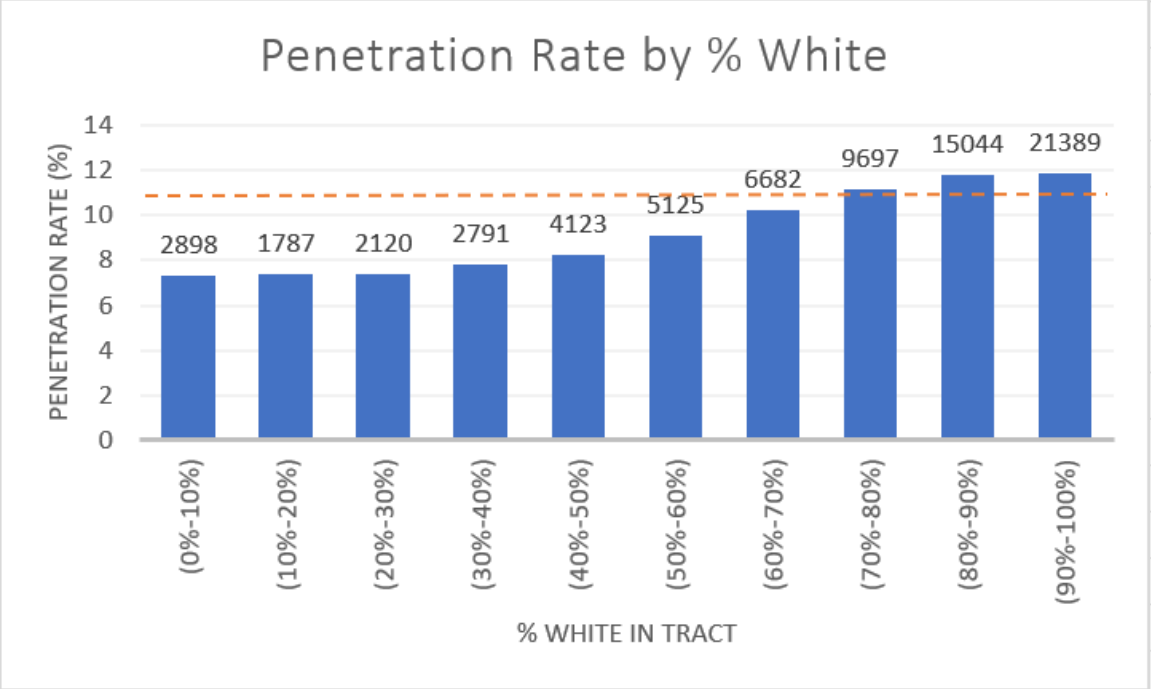

Figure 9: Comparison of penetration rates by percentage of White population in the tract, with data labels highlighting the number of Census tracts that fall into each bin.

The chart below shows the variation of average penetration rate for tracts when the percentage population in the tract that is Black varies. Figure 10 shows a similar, but inverted, trend as Figure 8. As the share of a tract that is Black goes up, the penetration rate goes down slightly. Again, the swing (7%-11%) is well able to be normalized.

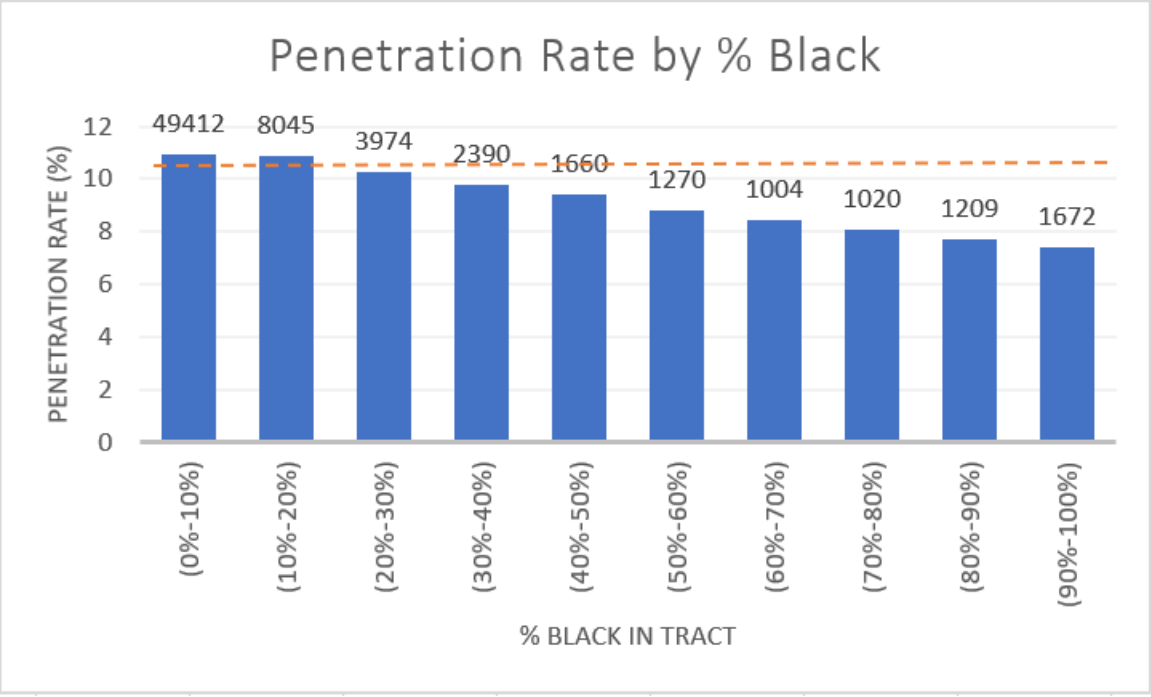

Figure 10: Comparison of penetration rates by percentage of Black population in the tract, with data labels highlighting the number of Census tracts that fall into each bin.

The chart below shows the variation of average penetration rate for tracts when the percentage population in the tract that is Asian varies. As Figure 11 shows, about two-thirds of the tracts have less than 2% of the population that is Asian. The penetration rate varies from 11% to 9% as the percentage of Asians in the tracts increases from 0% to 20%.

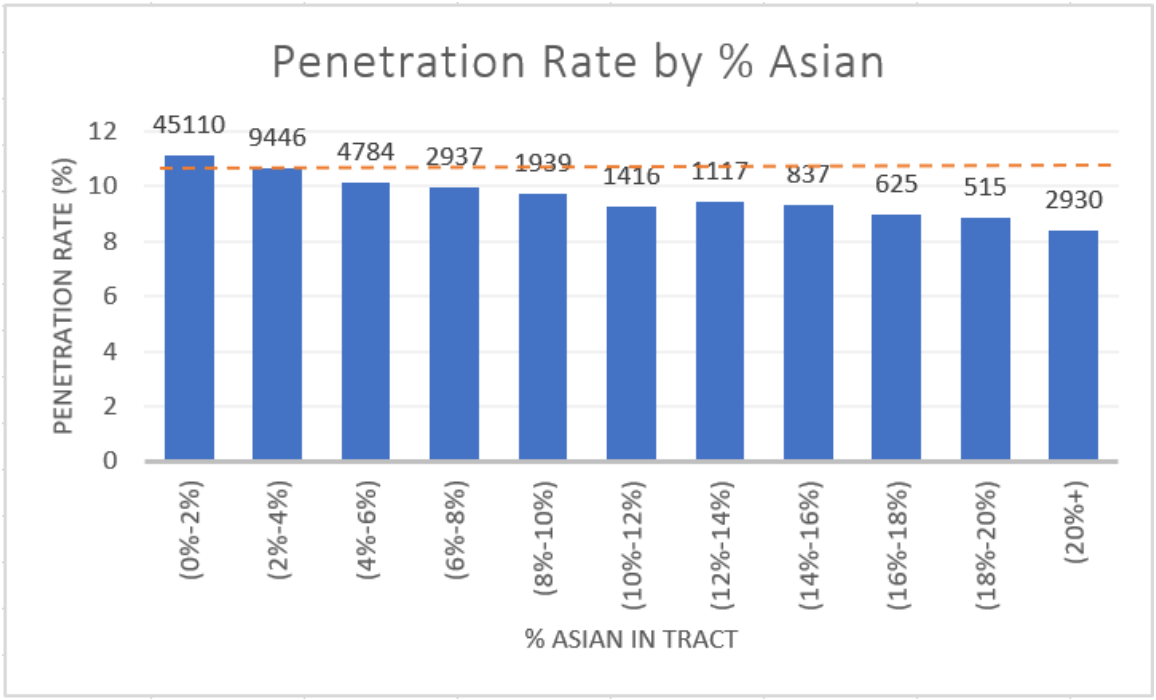

Figure 11: Comparison of penetration rates by percentage of Asian population in the tract, with data labels highlighting the number of Census tracts that fall into each bin.

Figure 12 shows the variation of average penetration rate for tracts when the percentage population in the tract that is Hispanic varies. As the chart shows, 80% of the tracts have less than 25% of the population that is Hispanic. The penetration rate varies from 11.4% to 10.2% as the percentage of Hispanics in the tracts increases from 0% to 25%. At the other end of the spectrum, 9% of the tracts have more than 50% of the population that is Hispanic, and the average LBS device penetration for these tracts is 7.8%.

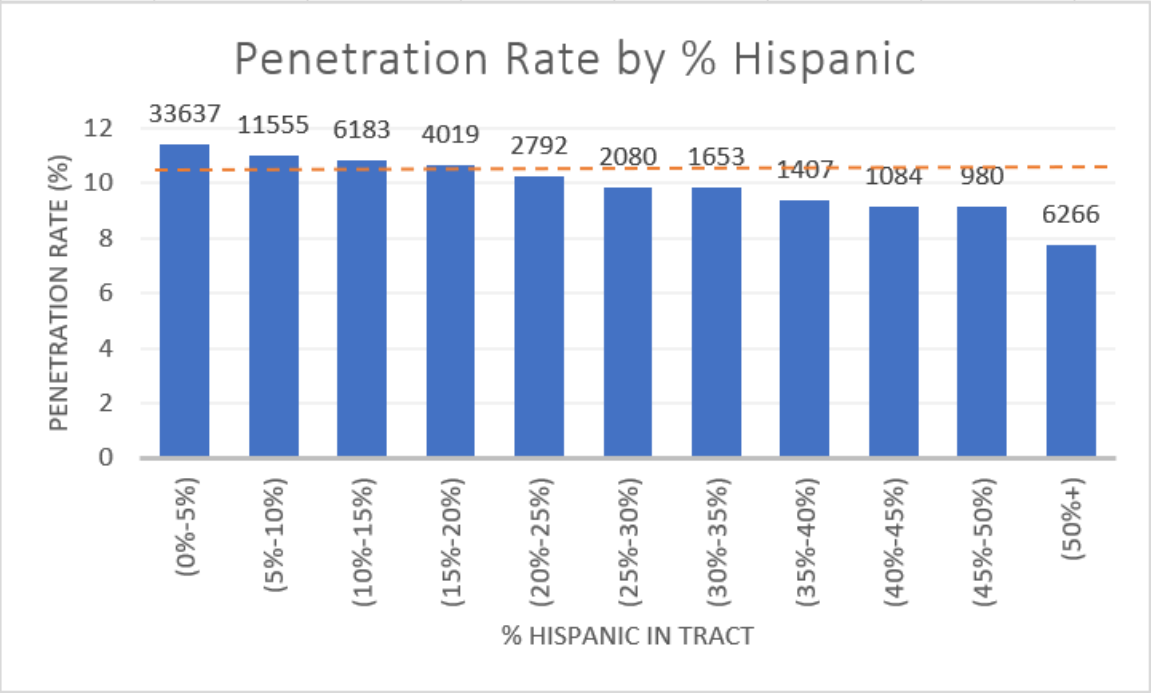

Figure 12: Comparison of penetration rates by percentage of Hispanic population in the tract, with data labels highlighting the number of Census tracts that fall into each bin.

We note that some of the nationwide variance we show in race may in fact be "restating" our higher penetration rates in rural counties in the middle of the country, which tend to have a higher share of White residents. Another way of looking at this is understanding if the penetration rate by tract is correlated with the share of population that is of a particular race. The scatterplots in Figure 13 show that there is no meaningful correlation. They also highlight the higher concentration of heavily White tracts in rural areas, heavily Black tracts in urban areas, etc.

Again, while this bias in the nationwide data does exist, it is much, much smaller than that in most surveys, and it is corrected in our normalization process.

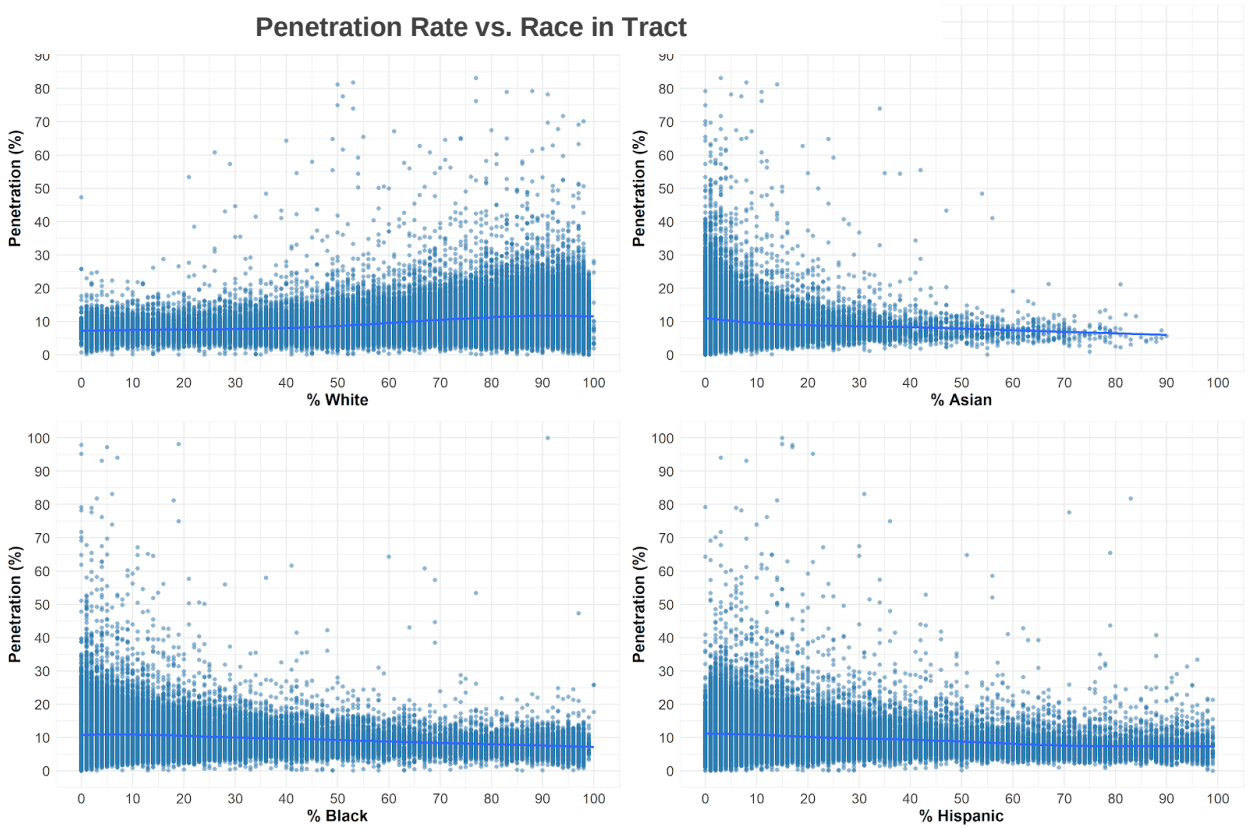

Figure 13: Scatterplots highlighting that the penetration rate by tract is correlated with the share of population that is White, Asian, Black, and Hispanic.

PENETRATION RATE BY AGE

Another common misconception is that smartphone devices are not widely used by the older population and hence this age group is missing from the LBS data sample. To demonstrate the diverse age groups that the LBS data sample covers, the chart below shows that there is no meaningful correlation between penetration rate and average age in a tract or the share of tract population that is 70 years or older.

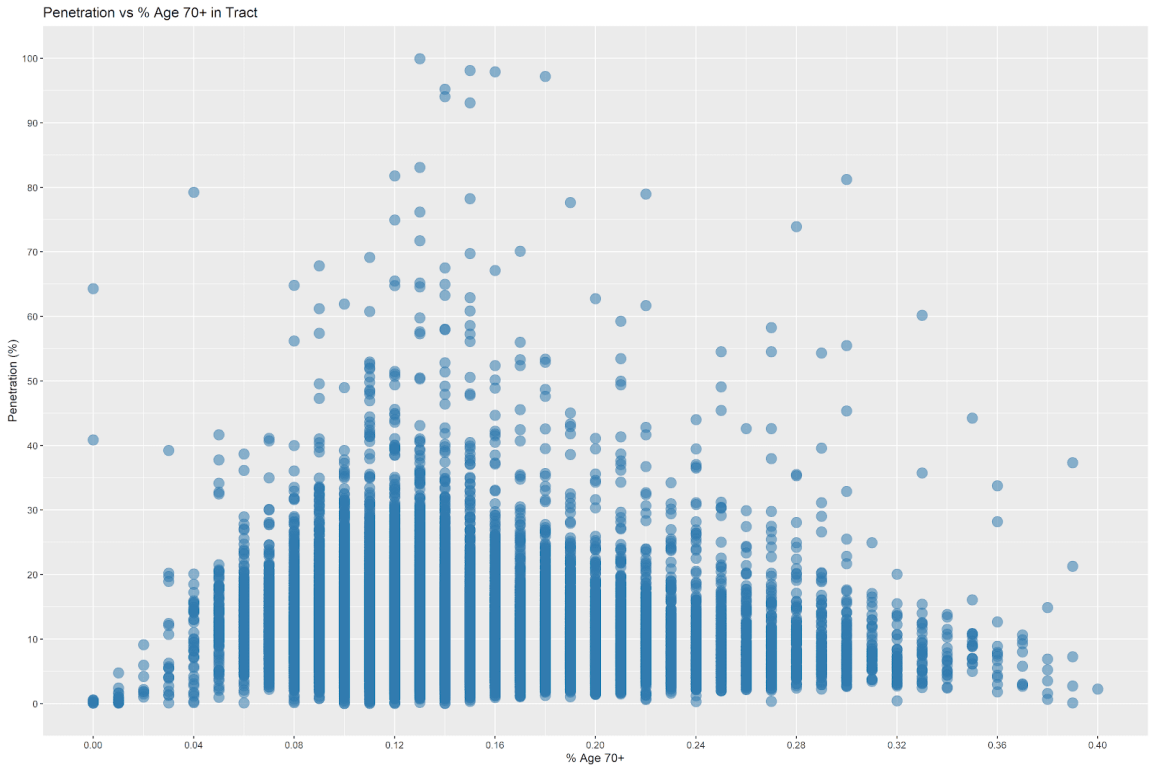

Figure 14: Scatterplot highlighting if the penetration rate is correlated with the share of tract population that is 70+.

Comparison of StreetLight's Sample Penetration to the NHTS Sample Penetration

As mentioned above, travel surveys are the most widely used traditional source of information on travel behavior and travel patterns. Travel surveys are designed and implemented to capture the travel representative of a target population. However, as outlined above, there are various limitations, such as the survey sample being not adequately representative of the population, either in terms of demographics (minorities and hard-to-reach populations) or behavior (transit, pedestrians).

To put the StreetLight sample penetration rates and characteristics into perspective, we looked at the sample data from the 2017 National Household Travel Survey (NHTS), conducted by the Federal Highway Administration. The NHTS is the largest nationwide effort in the United States to collect comprehensive data about all persons and their transportation behaviors.

Surveying the entire population is cost prohibitive and time intensive, so only a small sample of the full population participates in the survey. Details about the survey sample rates compared to StreetLight's sample rates are presented below.

Total Penetration Rate

The 2017 NHTS sample size was 264,234 persons, which has an implied penetration rate of 0.09% of the total population. For comparison, as shown in Table 1, StreetLight's device penetration rate for the U.S. is 10.6%.

Source	Nationwide Penetration Rate
StreetLight Data	10.6%
2017 NHTS Sample	0.09%

Table 1: Comparison of NHTS Sample nationwide penetration rate to StreetLight's nationwide penetration rate.

Income

The NHTS provides socioeconomic data on the survey respondents, such as household income.

While the distribution of the sample income is important, it is also crucial to understand the penetration rate by income bin. A comparison to StreetLight's data is presented in Table 2 below.

Income Bin	StreetLight Data	2017 NHTS Sample	StreetLight Data Bin compared to < $20K Bin	NHTS Data Bin compared to < $20K Bin
Less than $20,000	9.80%	0.07%	1.00	1.00
$20-$35K	8.10%	0.08%	0.83	1.14
$35-$50K	9.60%	0.09%	0.98	1.29
$50-$75K	11.00%	0.09%	1.12	1.29
$75-$100K	11.40%	0.09%	1.16	1.29
$100-$125K	11.20%	0.09%	1.14	1.29
$125-$150K	10.90%	0.09%	1.11	1.29
$150-$200K	10.20%	0.10%	1.04	1.43
$200K+	10.40%	0.09%	1.06	1.29

Table 2: The yellow columns show the comparison of the NHTS Sample's penetration rate by income to StreetLight's penetration rate by income. The blue columns show the sample for each bin indexed to the lowest income bin.

StreetLight's sample for the highest income bins are only 4%-6% higher than the lowest; the NHTS highest income bins were 29%-43% higher than the lowest income bin.

Urban and Rural Areas

Figure 15 below shows the distribution of the NHTS sample respondents by urban rural area classifications. As you can see, a high percentage of respondents is from a rural area.

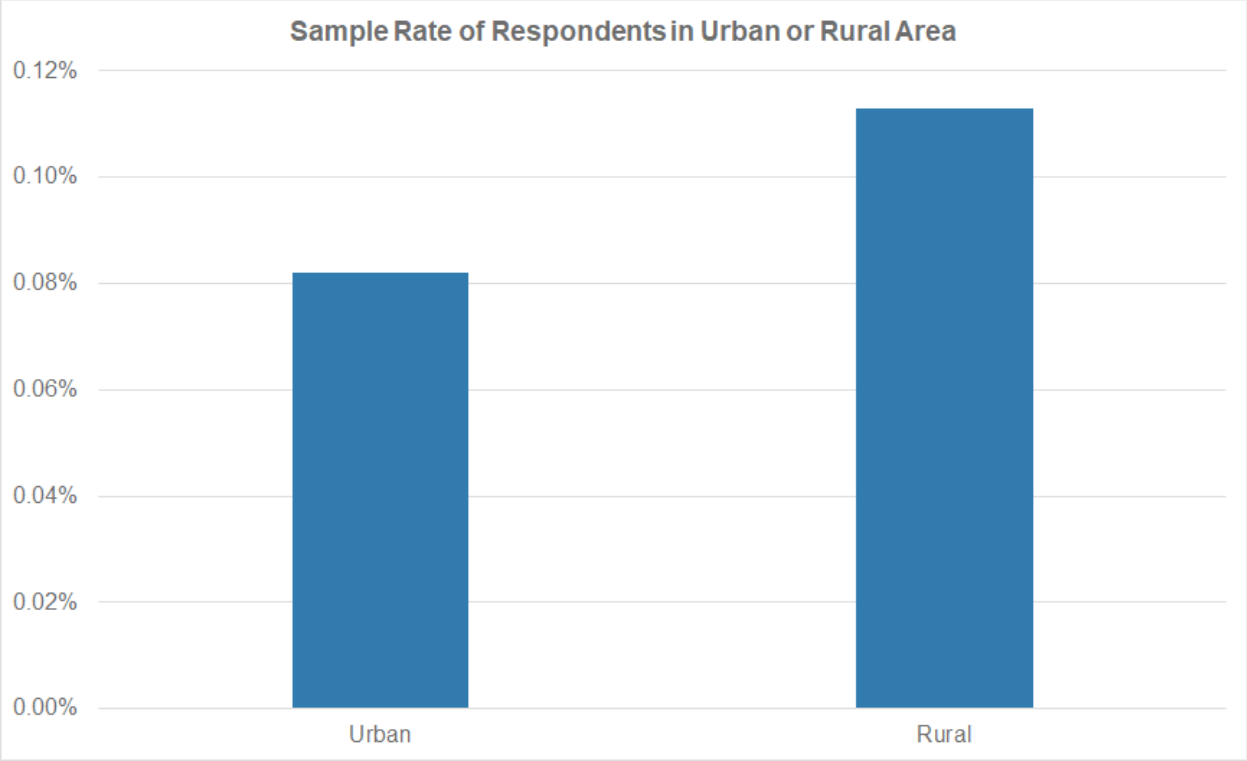

Figure 15: Distribution of urban vs. rural by sample rate of respondents.

A comparison of the penetration rate by geographic areas is presented in Table 3 below.

Density	StreetLight Data	2017 NHTS Sample
Urban	10.2%	0.08%
Rural	12.2%	0.11%

Table 3: Comparison of the NHTS Sample's penetration rate by density to StreetLight's penetration rate by density.

Demographics

NHTS also provides demographic information about their respondents. Table 4 below shows the distribution of respondents by demographic characteristics, as well as what the representative population is. We have included the StreetLight Device Weighted Sample distribution for comparison. The share of each race under the StreetLight Device Weighted Sample is the share of the race in each tract averaged over all the tracts, with the number of devices in a tract as the weight. In a sense, the distribution of race under the StreetLight Device Weighted Sample represents the sampling of devices in StreetLight sample with respect to racial composition. The racial distribution in the StreetLight sample is more similar to the ACS Data than the 2017 NHTS Sample. For example, in the StreetLight LBS sample, if a tract with 50 LBS devices has 4% of its population as Black and another tract with 100 LBS devices has 1% of its population as Black, then the percentage of Black population in the StreetLight Device Weighted Sample is (50*.02+100*.01(/(50+100) = 2%.

	StreetLight Device Weighted Sample	**2017 NHTS Sample**	**ACS Data – Population Weight**
Percentage of Population – Black	11.2%	7.1%	12.3%
Percentage of Population – White	76.1%	82.4%	73.5%
Percentage of Population – Asian	4.1%	4.3%	5.1%
Percentage of Population – Hispanic	14.3%	9.1%	17.2%

Table 4: Comparison of the NHTS Sample's penetration rate by demographic characteristics to StreetLight's Device Weighed Sample by demographic characteristics. (Note: Totals do not sum to 100%. Questions about Hispanic origin are separate from questions about race. Also, there are racial group data included in the NHTS sample that are not collected by StreetLight Data.)

Conclusion

LBS data is a sufficient and representative sample of the population, more so than large surveys such as the NHTS sample. The biases found are correctable. Looking at both the sample penetration rate and the sample representativeness, it is clear that Big Data covers an even distribution of demographic characteristics such as age, household income, and race, and what we see is that our sample represents the population at large. When we talk about how Big Data can help support social equity and environmental analyses, it is important to call out that equitable transportation policies, planning, and infrastructure depend on input from equitable data that is based on a large and truly representative sample.

About StreetLight Data

StreetLight Data, Inc. pioneered the use of Big Data analytics to help transportation professionals solve their biggest problems. Applying proprietary machine-learning algorithms to over 100 billion location data points every month, StreetLight measures multimodal travel patterns and makes them available on demand via the world's first SaaS platform for mobility, StreetLight InSight®. From identifying sources of congestion to optimizing new infrastructure to planning for autonomous vehicles, StreetLight powers more than 6,000 projects every month. For more information, please visit: www.streetlightdata.com.

STREETLIGHT DATA