NextGen NHTS Resources

- Data Portal: https://nhts.ornl.gov/od/
- Methodology Documentation: https://nhts.ornl.gov/od/documentation

Overview of NextGen NHTS Objectives

Q: **Michael Iacono (Minnesota DOT):** How are demographic characteristics imputed, and are there validation efforts to evaluate the accuracy of the imputation methods?

A: Stacey Bricka (MacroSys LLC): The imputation and validation details are discussed in the methodology reports, located here: https://nhts.ornl.gov/od/documentation

Q: **Keith Miller (NJTPA):** Is 2020 the first year that NHTS OD data are available?

A: Stacey Bricka (MacroSys LLC): Yes, 2020 is the first year that the national OD passenger and truck products are available.

A: Michael Pack (Univerity of Maryland CATT Lab): Yes. 2020 is the first year; however, there is the possibility that 2019 may be provided in the future. There's no timeline yet for when/if that might happen.

How It's Made: Behind the Scenes of the NextGen Data Product and Supporting Tools

Q: **Harun Rashid (Northern Virginia Transportation Authority):** What is the stop time threshold to define a passenger trip?

A: Aref Darzi (University of Maryland CATT Lab): In the tour-based trip identification algorithm designed for the NextGen NHTS project, there are several thresholds are used for identifying trip ends. In short, for short-distance trips, a 5-minute threshold is used while for long-distance trips the time threshold is 30 minutes. The details of the tour-based trip identification can be found here: https://nhts.ornl.gov/od/documentation

Q: **John Auble (Timmons):** If 20 unique people have devices reporting while all sharing the same bus ride, do you use all 20 individuals or just one "bus"?

A: Aref Darzi (University of Maryland CATT Lab): NextGen NHTS OD data reports person trips and movements. In the case of 20 people riding the same bus, the OD data will report 20 bus trips.

Q: Keith Miller (NJTPA): How can you say that it's 20 people, not 10 people each with two devices?

A: Aref Darzi (University of Maryland CATT Lab): There is a deduplication algorithm in the framework to ensure that we are not double-counting a single individual. Duplication can happen either due to several data vendors providing information about the same person, or one person having multiple devices. The details of the deduplication algorithm can be found here: https://nhts.ornl.gov/od/documentation

Q: Keith Miller (NJTPA): In congested downtowns, how do you differentiate between signals from people in slow-moving cars, buses, pedestrians, and bicyclists?

A: Aref Darzi (University of Maryland CATT Lab): We developed several machine learning and AI algorithms for imputing the travel mode information and ensuring the desired accuracy of the imputed mode. In addition to the average speed and travel time, the algorithm uses percentile speed information and transportation network features. Details of the mode imputation algorithm are presented here: https://nhts.ornl.gov/od/documentation

Q: **David Nelson (New Mexico DOT):** How do different states and regions vary in the quality of the end data (MOE etc.)?

A: Aref Darzi (University of Maryland CATT Lab): For the end data, the OD data is weighted and expanded to represent all movements across the nation. However, the weighted data expands the sample movements that are being observed in the sample data. Therefore, the sample penetration rate will have an impact on the richness of observations in the end data.

Q: David Nelson (New Mexico DOT): Do you have quantitative information on the MOE by region?

A: Aref Darzi (University of Maryland CATT Lab): The sample penetration rates are provided at the county-, MSA-, and state levels in the methodology document, along with other MOE for the raw datasets. More details can be found here: https://nhts.ornl.gov/od/documentation

Q: **Venu Garikapati (National Renewable Energy Lab):** How do you account for bias in location-based data? Are there any checks you do (or measures you take) to ensure that the trips are representative of various socio-demographic segments in a region or state?

A: Aref Darzi (University of Maryland CATT Lab): In the NextGen NHTS methodology, we designed a multi-level weighting framework to address the biases both for not observing all the population (device-level weighting) and also not capturing all the movement of an observed device (trip level weighting). In the device level weighting, we used the socio-demographic information to make sure that known socio-demographic biases in the data are addressed after weighting.

Q: Venu Garikapati (National Renewable Energy Lab): Which data source is used to develop these weights? ACS-PUMS or something else?

A: Aref Darzi (University of Maryland CATT Lab): The ACS data has been used for matching the population information.

Q: **Stan Young (The Eastern Transportation Coalition):** We've been exposed to data supply chains due to changes in policy either at Google or Apple or changes in state regulations. In your opinion how vulnerable is this NHTS moving forward as far as consistency and availability?

A: Michael Pack (University of Maryland CATT Lab): Data comes and goes all the time. You'll see press releases about data being misused and vendors going out of business because of privacy concerns. There's even impending legislation that might disrupt some

of the access to these data sets. What we try to do is have a large panel of data providers so we're not dependent on any one vendor. If we lose one, we've still got enough data from others that we can be statistically confident in the products that we're putting out. But there are concerns about (a large chunk of) this data potentially disappearing if certain legislation is passed. Although we can't do anything about that, we do have plenty of other methods to help prevent dramatic implications from the loss of just one provider.

Q: **Alan M Warde (New York State DOT):** Can through trips that are not stopping in an area be calculated?

A: Aref Darzi (University of Maryland CATT Lab): Theoretically, they can be calculated since the sighting data provides travel trajectory information as well. However, as the NextGen NHTS product focuses on OD movements, cut-through trips are not considered in the OD data.

Q: **Allen Greenberg (Federal Highway Administration):** How much do we know about the individuals from whom we're getting the data? The "old" NHTS combined individual demographic and other survey-acquired data with travel data. Does the new NHTS include individualized data and, if so, how? Thank you.

A: Michael Pack (University of Maryland CATT Lab): For the next-gen OD, demographics are imputed based on the home location of the device. Census data is used to impute demographics which is used in our weighting algorithms to help account for biases.

A: Stacey Bricka (MacroSys LLC): The NextGen NHTS program includes both the core data collected through a household travel survey as well as the complementary OD products derived from passive data products. The core data continues to provide individual demographic and survey-acquired data. Is your question about the biennial core data/household travel survey component or the OD component?

A: Aref Darzi (University of Maryland CATT Lab): There is no demographic information in the raw sighting data due to privacy protection. However, UMD designed an algorithm to impute the socio-demographic attributes of devices based on several features including the home location and travel behavior of devices. The details of the socio-demographic imputation algorithm can be found here: https://nhts.ornl.gov/od/documentation. The national OD does not have any socio-demographic information, but for the pooled fund partners age, gender, and household income information are provided.

Q: Allen Greenberg (Federal Highway Administration): For my clarity, though, are you trying to tie it to a specific individual (e.g., a person who is part of a 5-person family with income X) or instead to the common characteristics of individuals/families in the area where the individual lives? Much appreciated.

A: Patrick Zhang (Federal Highway Administration): The data is not linked to individuals; income is imputed using other data sources.

A: Aref Darzi (University of Maryland CATT Lab): The socio-demographic information is imputed at the individual level. As Patrick mentioned all the socio-demographic attributes are imputed.

Demonstration of Origin-Destination Data Visualization Tool

Q: **Brian Gillis (Miami-Dale County):** What is NHTS an acronym for?

A: Ross Wang (Oak Ridge National Laboratory): National Household Travel Survey.

Q: **Mike Bruff (Capital Area MPO):** What is the link to the OD data portal?

A: Stacey Bricka (MacroSys LLC): The OD data portal is located here: https://nhts.ornl.gov/od/

Q: **Kyle Titlow (USDOT - Bureau of Transportation Statistics):** Are the truck datasets also LBS-based, or are they from a different source (CV, telematics)?

A: Michael Pack (University of Maryland CATT Lab): Truck data are from in-vehicle telematics.

Q: **Keith Miller (NJTPA):** What is the pricing structure for the add-on data? What order of magnitude, and what is the price based on (e.g., the population of the region, number of zones, etc.)?

A: Stacey Bricka (MacroSys LLC): There are several factors considered in pricing. For more details, please email Patrick Zhang (patrick.zhang@dot.gov).

Q: **David Souleyrette (Kentucky Transportation Cabinet):** I am not sure if I understand the difference between aggregating at the state level and not aggregating at the state level. Is there an advantage of one over the other?

A: Ross Wang (Oak Ridge National Laboratory): Aggregating to the state level will only give you the state total, however, one might prefer to have zone-level numbers instead of just one value for the state-level info. It depends on what question you're trying to answer.

Q: **Venu Garikapati (National Renewable Energy Lab):** What is the lowest level of geographical resolution at which the aggregated (add-on) OD data is available? Also, is it possible to obtain OD data with socio-demographic information for the whole nation?

A: Michael Pack (University of Maryland CATT Lab): The national product does not provide socio-demographic information. We are only providing socio-demographic data (age, gender, and income) for the add-on passenger data product. The lowest level we go to is the census block group level.

Purchase and Use of Origin-Destination Data: Add-On Product

Q: **Keith Miller (NJTPA):** Habte/Guy, how much did the NHTS OD add-on data cost for Georgia?

A: Patrick Zhang (Federal Highway Administration): $250K for Georgia DOT.

Q: **Zoe Neaderland (Vermont AOT):** Stan or Michael-Would you compare the NextGen data with O-D from the TDM?

A: Stanley Young (TETC): The products from NHTS Next-Gen are similar in that they provide insight into the mobility behavior of the population in terms of trip-making. The technical differences between NHTS Next-Gen and the O-D in the TDM would include the base data incorporated into each of the products, and the types of products (vehicle trips, ped trips, micro-mobility, long-distance trips). Some products (such as NHTS NextGEN) report on inferred trips of the entire population. Some TDM vendors do the same, while others directly report counts of their observations. As far as quality and value – the Coalition has yet to perform any testing.

Q: **Zhongze Wang (Chatham County - Savannah MPC):** What is halo county?

A: Aref Darzi (University of Maryland CATT Lab): The halo area is defined as the immediately adjacent counties to the core study area (in the case of GDOT/ARC product the state of Georgia).

Q: **Robert Diogo (NJTPA):** Does hybrid work scheduling impact the imputation of the work trip purpose since someone might only go into the office a couple of days a week (and they might be different days)? Also, does hybrid scheduling define a typical weekday?

A: Aref Darzi (University of Maryland CATT Lab): The fixed workplace location identification algorithm is a behavior-based method that relies on observing the device at the work location. The algorithm is trained based on the American Time Use Survey. It is more likely to misidentify the work location in cases where working behavior is more unique.

Q: **Keith Miller (NJTPA):** Am I correct to conclude that there are no data connections between the "core" and "OD" NHTS products?

A: Aref Darzi (University of Maryland CATT Lab): Correct, the two datasets are compiled in completely different ways and data subjects in the two programs are not linked with each other.

Q: **Matt Glasser (Arcadis):** Did you say NHTS was also in RITIS TRIP analytics?

A: Greg Jordan (University of Maryland CATT Lab): Correct. NHTS data has been integrated into the CAT Lab's Trips Analytics Tool. Agencies who fund the Lab's Trips Analytics Tools will be able to see the NHTS data in addition to other trips/O-D data sets that they may have already purchased from other sources (INRIX, Wejo, etc.). This is limited to the MSA-to-MSA trips from Next-GEN.

Q: **Harun Rashid (Northern Virginia Transportation Authority**): Guy, on trip length distribution, how does the data look by trip purposes? for example, HBW (Home-Based Work) trip length on average is higher than HBO (Home-Based Other) trips.

A: Guy Rousseau (Atlanta Regional Commission): TLFD (Trip Length Frequency Distribution) by trip purpose is our next step in the analysis, so more to come on this excellent question

Q: **Jeevanjot Singh (New Jersey DOT):** What are the concerns with the imputed data, especially for non-motorized modes? It is a concern for us in developing the Vulnerable Road User Assessment.

A: Aref Darzi (University of Maryland CATT Lab): The details of the imputation algorithms are described in the methodology document: https://nhts.ornl.gov/od/documentation. Additional journal papers and reports are also available for specific imputation algorithms. Also, I want to mention that, after completing the imputations, the weighting/expansion algorithm is also designed to address the potential biases in the data.

Q: **Zhongze Wang (Chatham County - Savannah MPC):** Do the downloads require a login?

A: Ross Wang (Oak Ridge National Laboratory): No Zhongze. Data files are available for anyone to download.

Final Thoughts

Q: **Stan Young (The Eastern Transportation Coalition):** What is your primary concern about data quality or accuracy?

A: Guy Rousseau (Atlanta Regional Commission): One of my primary concerns is protecting the privacy of individuals and the population. That's why we have these filters that allowed us to remove the possibility of fabricating micro-data sets. Especially when you work with the public or in an MPO or you work with personally identifiable information, you want to make sure that there's no way to fabricate microdata.

A: Habte Kassa (Georgia DOT): I don't have any concerns.

A: Patrick Zhang (Federal Highway Administration): One concern is data coverage. Specifically, the data coverage does not have the same density in all areas, as shown in Michael's presentation. Some areas have more coverage than others. In a larger country, data density changes dramatically.

A: Michael Pack (the University of Maryland CATT Lab): There are a lot of things we do to try to account for changes in penetration rates which then impact device and trip level biases and weightings. That's one of the harder things that we must deal with and there's never a way to be perfect. I think we've got a decent methodology.

Q: **Stan Young (The Eastern Transportation Coalition):** In addition to an LBS-based NHTS, we also have a biennial survey. What is the status of that?

A: Danny Jenkins (Federal Highway Administration): We have just completed the data collection phase of the 2022 household survey which encompasses 7,500 households all around the country. We've started the process of processing and waiting for the data. we hope to have public data sets available in late spring or early summer of this year.

Q: **Stan Young (The Eastern Transportation Coalition):** What are your final thoughts?

A: Patrick Zhang (Federal Highway Administration): Our goal is to deliver the best quality data on time. There are some challenges. I think we've resolved all the issues it's going to come out soon. In the next year, we will have this back on a very good track, and we'll get the data on time and with good quality that people can trust. We are planning to release the 2021 data. We are getting to the final stages now.

A: Michael Pack (University of Maryland CATT Lab): The data is only useful if people are using it. Please try it and give us your feedback.

A: Ross Wang (Oak Ridge National Laboratory): The tools are only useful if you're using them. I encourage people to use the tool and play with it. We have a newer version of the OD data that will be coming out soon. That version of the data will be integrated with the tool and 2021 data.

A: Habte Kassa (Georgia DOT): GDOT became an NHTS add-on state back in 2009 and purchased additional surveys. The long and short-distance trips among other things helped us significantly to develop our statewide travel demand model. We joined NHTS again in 2017. This time we focused on the 14 MPO (regional) models that GDOT develops to assist MPOs. The trip rate for each region helped us to get good results. The previous survey was done more than 20 years ago in Augusta, Georgia and the same trip rate was applied elsewhere in the state. It took us a long time to develop and calibrate the model. The 2017 surveys helped us significantly. GDOT was preparing to get NextGen NHTS 2020 core survey. Because of the pandemic, the 2020 survey was on hold. GDOT is currently using the 2019 OD data and looking forward to the 2024 core survey data.

A: Guy Rousseau (Atlanta Regional Commission): I invite everyone to take a deep dive into their respective data sets with this tool; especially with the local origin-destination data add-on. In my opinion, you get the best of all data vendors and get one aggregated data set. It uses data aggregation methods from the CATT Lab. I want to thank everybody that's working with the trip analytics component to ingest this very large data set and make it more user-friendly for all of us. It's nice to use our in-house tools and GIS but I like this additional tool here. I look forward to more collaborative efforts with the group.