NATIONAL DATASETS; HOW TO CHOOSE THEM, HOW TO USE THEM

Penelope Z. Weinberger
American Association of State Highway and Transportation Officials

1. INTRODUCTION

1.1. The CTPP and the NHTS

The US population is around the 323 million mark today (and grows by 4 a minute). Transportation planners who serve this enormous population are dependent on only two national surveys for the bulk of travel data that informs infrastructure investment, planning, and decision making. The US Census Bureau's ongoing American Community Survey (ACS), from which the Census Transportation Planning Products (CTPP) data is derived, and the United States Department of Transportation (USDOT)'s National Household Travel Survey (NHTS) are the two national, premier, travel behaviour datasets available. Others exist, but these are the best vetted, and most reliable.

This paper is a comparative analysis of the two data sets, it discusses their methodology and statistical nuances, contents, limitations, strengths, and uses. This work is aimed to increase the capacity of practitioners dealing with large and varied data sets and give some insight into how to evaluate, how to assess, and what types of analyses are appropriate for given data.

The US Census Bureau collected commuting data from 15% to 25% of households, decennially from 1960 to 2000. In 2005, the Census Bureau began to collect data continuously via the survey tool that replaced the decennial long form; the ACS. The ACS asks a series of commute based questions that give a statistically robust but partial picture of the trip to work for the nation. Despite the vast coverage, the data are of scant detail, the Journey to Work question (JTW), for example, provides data for one trip for one job, by one mode, in one direction. Other questions on the ACS provide commute related data, such as trip start time, trip length in minutes, work location, and vehicles available per household. The ACS also provides a tremendous amount of socio demographic information for the nation, as it is not only the sole national survey tool designed to supply data to the agencies and administrations that comprise the US Executive branch, but also the tool the US Congress uses to gather information on the nation. The ACS is based on a sample of just under 10 percent of households accumulated over five years. Each year, data are released for both one and five year samples, with the one year sample available only for geographies 65,000 or greater in population. Five year accumulations of sample data are available down to the smallest published geography for sampled data, the "block group" which represents a resident population of 600 to 3000. The data are released as period estimates, data which describe the nation for the reference year or years.

The State Departments of Transportation, through the CTPP, commission a customized tabulation of small area ACS data tailored for transportation planning applications. This data set is used for travel model validation and

calibration; an input to the long range plans required of the 408 Metropolitan Planning Organization (MPO) areas and 50 states as a condition for them to receive Federal aid. Other uses of the CTPP range from generating demographic profiles to corridor planning to Environmental Justice analysis to trend analysis, including national commute trends detailed in a publication called Commuting in America.

The NHTS is a travel diary based survey that has been collected seven times between 1969 and 2009 with an eighth iteration scheduled for completion in 2017. Although smaller and more infrequent, the NHTS collects more detailed travel data, including all trips for all purposes. In recent years the US Department of Transportation has invited regional transportation agencies (State DOTs and MPOs) to purchase additional data for their respective jurisdictions. These “add-ons” are concentrated in a specified region and respondents are asked all the core questions along with up to six extra questions chosen by the purchasing agency. The agencies have some latitude in the additional questions and choice of travel days sampled (one or both weekend days, but not weekday only). In 2001, the NHTS collected all household travel from a national core of 26,000 households, with regional add-ons adding 43,000 more households, constituting about 0.06% of total US households. In 2009, the core sample remained the same, but the add-ons agencies totalled nearly 125,000 households, bringing the sample to about 0.13% of the total household population.

The CTPP is better suited to assess phenomena in small geographies, while the NHTS captures many vehicle characteristics and all trips and purposes, not just commute trips. Both data sets are used to inform transportation policy and assess how previous transportation investments have performed.

1.2. Other Datasets and Why They’re Not Included

There are other transportation and worker based datasets of national scope that could be employed to address some of the analytical needs answered by the NHTS and the ACS and CTPP. The Census bureau produces a number of other surveys and administrative datasets that each have a piece of the puzzle covered. The most comprehensive is likely the Longitudinal Employer-Household Dynamics (LEHD), whose public facing dataset is the LEHD Origin Destination Employment Statistics (LODES). This is a large, popular, frequently updated, administrative dataset, derived from state employment wage data, privacy protected social security records, State Department of Employment Security reports, other Census datasets, and other sources. The program is voluntary, the data is highly synthesized. The chief reason LODES is not included in this comparison is that the documentation is scant and un-illuminating and the data is incomplete. By way of example here is a description of the data from a US Census Bureau staff person who is chiefly concerned with the dataset: “The LODES data are an extract of the LEHD data infrastructure, which is composed of administrative records, census and survey data focused on the labor market, worker, and firm statistics. State unemployment insurance reporting and account information and federal

worker earnings records provide information on employment location for covered jobs and residential information for workers, which form the basis of the LODES data product. However, these data are not available in all states for all years of the series". Fortunately, LODES is a work in progress and there is high hope that it will be an excellent data source in the near future.

Another new data set is the US DOT's National Performance Management Research Data Set (NPMRDS), a national data set of average travel times and speed on the National Highway System network for use in performance measurement. The data set contains five minute increments of probe based passenger and freight vehicle travel data for the road network. Sample size is unknown and no data are imputed, that is to say, if no probe crosses a link within the increment, no data are reported for that pair. Outliers are included; the data is released in an unprocessed state to public agencies by agreement.

Yet another National dataset compiled from state contributors is the Highway Performance Monitoring System (HPMS). The HPMS is a national level highway information system that includes data on the extent, condition, performance, use, and operating characteristics of the nation's highways. While both of these datasets are inputs to the planning process, they contain no insight into travel behaviour and are not appropriate for comparison in this paper.

2. ABOUT THE DATA

2.1. Methodological and statistical nuance

The ACS is a household based survey which purportedly samples 3.54 million households per year, although, the US Census Bureau reports the unweighted sample count number of households as 2,305,707 in 2015, the most recent year for which data are available. In 2014, the most recent year for a five-year (small area) sample, the unweighted sample count was 10,952,853 which is about 6 million households shy of the Census' Sampling frame. Households are randomly selected from a master address file and should not be sampled more than once within five years. The survey is a mail out, mail back with computer assisted personal interviewing (CAPI) non-response follow up. It is collected continuously from January through December with its reference time being the past calendar year. The JTW reference usual commute patterns for the preceding week.

It is beyond the scope of this paper to discuss the complex weighting methodology used by the census bureau however the survey design and methodology is well documented and publicly available at: http://www.census.gov/programs-surveys/acs/methodology.html Suffice it to say, population and households, race, age, sex, and Hispanic origin are controlled to the Census Bureau's official estimates.

Tabular data are reported at a 90% confidence interval and with margins of error for some data at some areas exceeding the estimate. Anonymized

individual records are available at specified census geographies with a minimum 100,000 population.

The CTPP in particular is subject to rules for special tabulations that include disclosure review, disclosure proofing in the manner of data perturbation, rounding, and suppression. It is produced from the ACS microdata records and re-tabulated according the its own specifications, and thus includes workplace based data at small areas, flows from home to work, and data cross tabulated with transportation planning issues in mind.

The 2009 NHTS dataset contains data for 150,147 completed households in the sample. Publicly available data from the 2009 NHTS include an anonymized microdata record of all survey reponses. Since the nationally dispersed core sample accounts for 26,000 households and the locally specified add-ons account for nearly 125,000 households, weighting factors have been adjusted to account for the oversampling in the add-on areas. “For example,” According to Federal Highway Administration 2009 NHTS Users Guide, “if New York State was oversampled by a ratio of 4 to 1, then the weights for NY samples were reduced to ¼ of their value. This is an oversimplification of the actual weighting process, but the point here is that the data user can be assured that the weighted data from the NHTS is representative of national estimates.” The 2009 NHTS was conducted over a period from March 2008 through May 2009. Travel days were assigned for all seven days of the week, including all holidays. The survey data were weighted to a 12-month period to produce annual estimates of travel. The NHTS was conducted as a telephone survey, using Computer Assisted Telephone Interviewing (CATI) technology. The sample was a list-assisted random digit dialing (RDD) telephone number sample. The 2009 NHTS is raked, weighted and adjusted to 2008 ACS household totals and to population totals which are themselves controlled by the Census Bureau. The next NHTS (currently in the field – 2016) is a household based mail out, mail back.

2.2. Contents

The ACS is the survey tool used to collect demographic, socio-economic and housing data for the nation. Data collected of interest to transportation planners includes journey to work by mode, and number of passengers if the trip is by private vehicle (vehicle occupancy). Detailed residence and workplace data are collected. Trip time in minutes, time leaving home, workers per household, auto availability and presence or absence of children for households are all collected, along with myriad other individual and household based data. The dataset includes data on income, industry, occupation, worker class, status of worker, race, sex, age and many other variables that are useful for transportation planning applications.

One may wonder why the transportation community purchases a tabulation when there is so much in the ACS already. The special tabulation goes much further than the ACS. Along with specialized cross tabulations at place of residence, the CTPP includes 63 tables with unique variables and cross tabulations at place of work down to the census tract, including 15 means of

transportation tables crossed with other, or multiple other variables. Compare this to the ACS standard tabs, which contain 42 tables at place of work reported to the place level. Additionally, the CTPP contains home to work flows down to the tract level. While the Census Bureau produces ad hoc flow tables, they are limited to County to County as the most refined geography.

The NHTS collected data on daily trips including: trip purpose (e.g., work, shopping), mode of transportation used, travel time, time of day when the trip took place, travel day, and if a private vehicle trip - number of people in the vehicle (vehicle occupancy), driver characteristics (e.g., age, sex, worker status), and vehicle attributes (e.g., make, model, model year, amount of miles driven in a year). These data are collected for all trips, all modes, all purposes, all trip lengths, and all areas of the country, urban and rural, including household, person, vehicle and daily (travel day) trip level data.

Data common to both datasets include work trips (including mode), commute trip length, departure time, household size, vehicles per household, household income, and age of worker. Identifying the commonalities between datasets, and ensuring that the methodology is such that the variables are comparable is a key to dataset expansion.

It has been suggested that, in addition to reviewing the tables produced or studying lists of available variables, completing the survey used to collect the data is an excellent way to become more familiar with any dataset's contents and methodology.

2.3. Limitations and Caveats

In order to have sufficient data to release at small geographies, the ACS is collected for five years and batched together as a period estimate, which makes comparison with historical point in time data difficult. There is a single mode collected and data collection is subject to reporting errors. Only a single "main" job is collected, there is no trip chain data. The CTPP program is small, and though it is the purchaser of the largest special tabulation of census data it is limited in scope, bandwidth and funding.

The NHTS is not a regularly collected survey; it is ad hoc, occurring once every five to ten years when the budget and political will align. NHTS does not contain specific information on the costs of travel, information about specific travel routes or types of roads used, or how travel of the sampled household changes over time. The NHTS is not a longitudinal survey, which would involve tracking the same sample households over time. Furthermore, information that would identify the exact household or workplace location is collected but not public to protect the confidentiality of respondents.

2.4. Strengths

The CTPP special tabulation has been requested and created for the nation from 1990 forward. A CTPP based on 2006 to 2010 ACS represents the third full national data set in three decades. There were flow and workplace special tabulations requested from the 1980 and 1970 decennial census' but they

were requested by individual agencies and do not represent a national dataset. From 1990 to 2010 the CTPP had grown with more tables, more refined statistical processes, more documentation and overall more utility with each request. In 2018 the special tabulation is scheduled to shrink to one third of the size from 2013 (the release year of the 2006 – 2010 data). Nevertheless the data will still contain more workplace data than the standard ACS products. The data is more current; being continuously collected allows for more frequent tabulation. The CTPP is now more consistent with data sets that come out annually, such as the Bureau of Labor Statistics' Consumer Expenditure Survey. Both the CTPP Program and the Census are proactive about the error terms, publishing them with the estimates and highlighting the caveat in training and presentation. The CTPP is free and available through web interface and ftp download. The program is well covered by technical support, outreach, and training.

The NHTS core data are comprehensive and consistent. The travel diary yields trip rates that are comparable to urban travel surveys. Household rostering of trips helps with respondent burden and coherence of the data. The add-on program allows serving the needs of states and MPOs while enhancing the sample. The web page, http://nhts.ornl.gov, designed and operated by Oak Ridge National Laboratory, provides comprehensive tools that allow access to the data and has greatly broadened the user community. Appended data provide significant items, such as miles per gallon, that respondents often do not know. The transferability statistics files that were created by the Bureau of Transportation Statistics allow all regions to perform travel analysis for their respective regions by using files that are representative of various census tract areas.

3. USES

This section is limited to uses of the CTPP; a comprehensive list of uses of ACS data is too broad and not necessarily transportation related. In a recent survey of CTPP users half of analyses performed employed flow data for a variety of purposes, the top two being modelling and demographic profiles. Modeling purposes include trip generation, trip distribution, mode choice and population synthesis as well as travel model validation. Planning analyses include determining day-time population, Environmental Justice (social equity), supporting planning regulations, and corridor analysis. The CTPP includes sets of ready-made, place based demographic profiles that include trend analysis and statistical significance testing. Profiles include data going back to the 1990 CTPP. Subjects covered include basic worker and population growth, changes in means of transportation, and other commute characteristics, comparisons of commute patterns of Millenials to Baby Boomers, and others. Users also create their own profiles based on custom geographies, or demography of interest to their individual purpose.

The NHTS data are used primarily for gaining a better understanding of travel behavior. The data enable national, state and regional transportation planners and the transportation research community to assess program initiatives, review programs and policies, study current mobility issues, and plan for the

future. The NHTS is a tool in the urban transportation planning process; it provides national data on personal travel behavior, trends in travel over time, and trip generation rates. It is used as a benchmark in reviewing local data, and data for various other planning and modeling applications. The NHTS program maintains a Compendium of Uses assigned into 11 broad categories:

- Energy Consumption
- Trend Analysis and Market Segmentation
- Bicycle and Pedestrian Studies
- Policy and Mobility
- Travel Behavior
- Survey, Data Synthesis, and Other Applications
- Special Population Groups
- Environment
- Traffic Safety
- Transit Planning
- Demographic Trends

within these 11 categories, in 2015 the NHTS recorded 377 individual cases for the data being used in research or articles. Use of the NHTS is seeing steady growth. An annual compendium was collected each year from 2012 forward documenting the number and characteristics of research papers and articles citing the NHTS as a data source. From 2006 to 2011 there are 106, in 2012; 210, in 2013; 280, in 2014; 322. The compendia are not only an excellent source of use documentation but also a rich guide to the possibilities for research using this dataset.

4. CONCLUSIONS

There are many local household travel surveys, state based data sets, national travel surveys and myriad tools that generate what planners need to do their work. Learning of the peculiarities in any data set is critical to proper use. Ensuring that definitions are consistent across data sets or even iterations of data collections is imperative. Understanding weighting schema, error terms, and universe definitions is vital. Read the survey tool, read the documentation, look at past research, and lend full credence to local knowledge.

BIBLIOGRAPHY

2009 NATIONAL HOUSEHOLD TRAVEL SURVEY USER'S GUIDE (2011), USDOT, Washington

Graham, M. R., Kutzbach, M. J. and McKenzie, B. (2014) Design Comparison of LODES and ACS Commuting Data Products, CES, Suitland