U.S. Department
of Transportation

**Research and
Innovative Technology
Administration**

Bureau of
Transportation Statistics

National Household Travel Survey Pre- and Post-9/11 Data Documentation

TABLE OF CONTENTS

CHAPTER 1. THE NHTS AND 9/11

PRE- AND POST-9/11 DATA FILES

The 2001 National Household Travel Survey (NHTS) was conducted from March 2001 through May 2002. Because the data collection time period includes 9/11/2001, there has been considerable interest in using this data set to assess the effect of the events of 9/11 on travel behavior, especially for long-distance trips. *This document discusses long-distance trip data set files only.*

However, the 2001 NHTS was not designed to assess the effects of 9/11 on long-distance travel and several factors preclude the direct comparison of pre-and post-9/11 travel data. *Travel is influenced by seasonality, economic conditions, and other factors. Therefore, the differences in travel volume and patterns in the pre- and post-9/11 data sets cannot be attributed solely to the impact of the terrorist attacks.*

The pre-9/11 period of 2001 NHTS covers March 2001 to September 2001, a period of over 5 ½ months, and includes the summer season in which a large proportion of long-distance trips are taken. There were approximately 22,000 persons responding about travel prior to 9/11. On the other hand, the post-9/11 period of the survey covers September 2001 to May 2002, a period of roughly 8 months, and includes Thanksgiving and Christmas and other winter holiday travel – a traditionally heavy season for long-distance trips. This survey had responses from approximately 38,000 persons on their long-distance trips after 9/11. The demographic, spatial, and temporal composition of persons who took long-distance trips prior to 9/11 and those who took long-distance trips following 9/11 were not the same.

To facilitate assessment of the effect of 9/11, one approach is to make each of the two groups a nationally representative sample. This is achieved by constructing new weights using statistically sound methods.

The pre-9/11 and post-9/11 data files were created for both person-level and long-distance trip level by dividing the NHTS 2001 public use person data file and long-distance trip data file, respectively, into two parts organized into four different data files: Pre-9/11 Person, Post-9/11 Person, Pre-9/11 Long-Distance Trip, and Post-9/11 Long-Distance Trip. Note that at this point, because of interest, the daily trips level file has not been divided for pre/post- 9/11 comparison. The new full sample weights and replicates for each of these four data files were created by applying statistical adjustment on the original weights to obtain valid annual estimates and to estimate the statistical significance of the estimates. The weight calculation for Pre-9/11 and Post-9/11 data only applied for the national “usable” level, so there is only one level weight for each data set.

How the data, weights, and replicates were created and how to use them are described in detail in Chapter 2 of this documentation.

NHTS 2001--OVERVIEW

The National Household Travel Survey is a U.S. Department of Transportation (DOT) effort sponsored by the Bureau of Transportation Statistics (BTS) and the Federal Highway Administration (FHWA) to collect data on both long-distance and daily travel by the American public. The joint survey gathers trip-related data such as mode of transportation, duration, distance, and purpose of trip. The Survey also collects demographic, geographic, and economic data for analysis purposes. Policy makers, individual state DOTs, metropolitan planning organizations, industry professionals, and academic researchers use the data to gauge the extent and patterns of travel, to plan new investments, and to better understand the implications of travel trends on the nation's transportation infrastructure.

The 2001 NHTS updates information gathered by two earlier series of travel surveys—the Nationwide Personal Transportation Survey (NPTS) conducted in 1969, 1977, 1983, 1990, and 1995, and the American Travel Survey (ATS) conducted in 1977 and 1995. It is a nationally representative survey of the daily and long-distance travel patterns of the nation using travel behavior data collected from all household residents in roughly 26,000 households.

The survey was conducted from March 2001 through May 2002. The first telephone call to recruit a household was made on March 19, 2001, and the last telephone call was made on May 4, 2002. Each household was sent a survey form and asked to report all travel by household members on a randomly assigned "travel day"Interviewers followed up with a phone call and asked respondents about their travel on the travel day and the preceding 27 days. This four-week period is called the "travel period." Travel days for daily-travel trip reporting were assigned for all seven days of the week, including all holidays. The first travel day assigned was March 29, 2001, and the last travel day assigned was May 4, 2002. The last travel period assigned was April 7 through May 4, 2002. The respondents were asked to report on trips 50 miles or more, referred to as a"long-distance trips" or "travel-period trips" taken by household members during the travel period, including the travel day.

Public-use national data from the 2001 NHTS is organized into five different data files:

1. households,
2. persons,
3. vehicles,
4. daily travel, and
5. long-distance travel.

Weights and replicates were provided for each of these five data files. This information is used to extrapolate the entire population's annual travel patterns from this representative survey of resident's travel patterns. Weights and replicates match the studied variables of the sample size to the same information of the population at large.

The data files contain two kinds of weights: one is from "usable" households in which person interviews were completed with at least 50 percent of adults in the household (26,038 households in the sample), and another is from"100 percent" households in which person interviews were completed with all adults in the household (22,178 households in the sample).

For more information about weights see Person-Level Weight, below.

For more information about long-distance and daily trip data in NHTS 2001, please refer to the National Household Travel Survey (NHTS)—National Data and Analysis Tool CD.

CHAPTER 2. WEIGHT CONSTRUCTION

Weights are needed to produce valid population-level estimates so that the results of a survey of the population are representative of the population as a whole. Adjustment and poststratification are performed on collected data to reduce bias of estimates. Poststratification reweights the data so that the characteristics of the respondents are the same as the characteristics of the population.

INITIAL PRE-9/11 AND POST-9/11 PERSON DATA SETS

The 2001 NHTS person-level data file was separated into two data sets: the Pre-9/11 group of respondents and the Post-9/11 group of respondents, based on the assigned day of September 25, 2001. A person interviewed before and on 9/25/01 would be in the Pre-9/11 group, while a person interviewed after 9/25/01 would be in the Post-9/11 group. The rationale behind selecting September 25, 2001 as the cutoff is explained below. The separation of the 2001 NHTS person-level data file is only in the sense of time frame. All information or variables including the weights were kept intact.

PERSON-LEVEL WEIGHT

Data were weighted at the person level so that the survey respondents and their demographics and characteristics would accurately represent the characteristics and demographics of the population at large.

For example, if in the survey 47% of the respondents were male (not accurate, an example), and in the U.S. population 49% of the population was known to be male, then the male survey respondents would be weighted stronger so that their data and travel information would count for 49% of the population's travel patterns. Likewise the female respondents' data would be weighted less than the males, so that their 53% of the survey results would accurately represent the 51 % of the population that are female.

The number of respondents in the Pre-9/11 group was 22,204, while the Post-9/11 group had 38,078 respondents. In addition, the composition of the two groups was different. To see whether the Pre- and Post-9/11 groups were representative samples from the national population we constructed population control totals for key characteristics that were related to key survey variables. The control totals were independently obtained by adjusting the Census 2000 numbers for growth between 2000 and September 2001, which was the midpoint of the survey. The control totals used for reweighting the Pre-9/11 data set are given in Table 1 of Appendix B while the control totals for the Post-9/11 data set are given in Table 2 of Appendix B. We saw differences in estimates of the number of Hispanics in the United States., the number of Blacks in the United States, and other variables, from population control totals. To make each of the Pre-9/11 group and the Post-9/11 group a nationally representative sample, we post-stratified each group to population control totals.

REWEIGHTING—QUESTIONS AND ANSWERS

Q: Why do we need to reweight the new data sets?

A: The original survey was weighted by the different variables over the entire 14-month period to correspond with the population at large. By splitting the data up, the original weights cannot be used in either subsample to accurately represent the population at large. (For example, in the first half of the survey 48.39% of the interviewed were males which is less than the 48.80% over the entire study of the survey respondents, but in the second half 49.11% of respondents were male. For this reason our initial weight, which calculates to 48.80% for males, would not be an accurate weight for either subsample.) This same principle applies to the respondents and their other characteristics before and after 9/11. The weights used over the entire study were representative of the characteristics of respondents over the entire study, but not of the Pre-9/11 respondents' characteristics or of the Post 9/11 respondents' characteristics. The whole is equal to the sum of the parts, but in this case, the parts were not equivalent. The respondents characteristics, before and after 9/11 were not equivalent and therefore the responses needed to be reweighted.

Q: Why was September 25, 2001 chosen as the cutoff for the Pre- and Post-9/11 data set?

A: There is no way to divide the long-distance trip data by the exact date of September 11, 2001, and then adjust the corresponding weights: a. Because it is impossible to adjust the weights directly at the long-distance trip level, we have to adjust the weights of the Pre-911 and Post-911 data files at the person level first and then incorporate the person-level weights into the long-distance trip data files; b. The NHTS did not collect long-distance trip data from all household members for the entire data collection period (March 2001 – May 2002). Instead, people were asked to report their long-distance trips for a 4-week period prior to and including their randomlyassigned travel day. Because a cross-sectional sample of people was interviewed throughout the data collection period, these 4-week "travel period" reference periods are spread out across the data collection period. This means that most survey respondents were interviewed only about their trips either before or after September 11, 2001 (with a small number of people whose reference periods spanned September 11). This made it necessary for us to choose a date to divide those whose reference period was before September 11 and those whose reference period was after September 11 (and evenly divide the group whose reference period includes September 11). The assigned travel day of 9/25/2001 was chosen as the cutoff point to achieve this: The respondents who interviewed on the assigned travel day of 9/25 were asked about all long-distance trips taken on that day and the preceding 28 days. So respondents would give information about all trips they took between and including August 29, 2001, and September 25, 2001, and 9/11 is the exact midpoint of the time interval. This means that some long-distance trips taken after 9/11 are in the Pre-9/11 long-distance trip data file and also some long-distance trips taken before 9/11 are in the Post-9/11 long-distance trip data file. For assessing the effect of 9/11 one can delete trips taken after 9/11 by persons in the Pre-9/11 group by using the variable TPBOA911 (travel period before on or after 9/11). Similarly, one can delete trips taken on or before 9/11 by Post-9/11 persons easily.

Chapter 5 in this documentation will explain how to deal with these cases in more detail.

ADJUSTMENT FOR DATA SPLITTING

After the data set was divided, the sum of the weights left both in the Pre-9/11 and Post-9/11 person data file were not equal to the population total. To adjust for this, the weights are multiplied by factors for each of the two data files.

Factor = Population total/ Sum of the original weights, where
Population total = 277,208,169

RAKING

The next step in recalculating the useable person weight was to match survey estimates to independent controls for various demographic categories, in a process called raking. In this study controls are used to match the survey respondents measured characteristics to what is known about the occurence of the characteristics in the population at large, so that respondents' weights will represent accurately the travel patterns of the entire population.

There are eight dimensions used in the raking process. The dimensions are: race, ethnicity, race by month, ethnicity by month, sex by age, census region, MSA status, and month by day of the week. These dimensions were chosen because the Pre- and Post-9/11 groups were different from the population in terms of these dimensions and we had available control totals for these variables from Census 2000. These control totals were constructed separately for the Pre- and Post-9/11 groups by adjusting Census estimates for growth between 2000 and 2001, when the majority of data collection on the NHTS was done, by using estimates from the Census Bureau's Current Population Survey.

Weights were first adjusted to assure agreement on the first raking dimension, then weights were adjusted for the second raking dimension, then for the third, etc. This process was repeated, again and again, assuring agreement with each of the raking dimensions. The process continued, with iterative controlling for each variable, until simultaneously close agreement for each of the variables was obtained. The raking process was done separately for both the Pre-9/11 and Post-9/11 data files.

The variables and the control totals are provided in appendix D, along with the average adjustment factor for each category.

TRIMMING

A final step was to "trim" very large weights, which were a byproduct of the raking process. Inordinately large weights tend to substantially increase sampling variance. By keeping weights small, sampling variance is reduced, although there is some loss in bias reduction, which was due to the adjustment and raking process. Trimming is only used to reduce large weights, not for editing data in any way.

Trimming was performed so that the maximum weight a response could have was four times the mean weight of all the respondents. The weights which were more than four times the mean of weights were trimmed to equal a maximum of four times the mean weight. After trimming large weights, the raking process was then repeated so that survey estimates would still agree with the control total. This trimming process was performed twice, separately for both the pre-9/11 and post-9/11 data files.

LONG-DISTANCE TRIP DATA SETS

The 2001 NHTS long-distance trip data file was basically divided into two parts. Any trip taken by persons whose assigned travel day was before or on the cutoff point of September 25th, belongs to the pre-9/11 long-distance trip data file. Otherwise, the trips fall into the post-9/11 long-distance trip data files. This means that some long-distance trips taken after 9-11 are in the pre-9/11 long-distance trip data file, and also some long-distance trips taken before 9/11 are in the post-9/11 long-distance trip data file. Chapter 5 in this documentation will explain a way to deal with these cases.

LONG-DISTANCE TRIP WEIGHTS

The person-level weights were incorporated into long-distance trip data files by merging with the long-distance trip data by house ID and person ID. Long-distance trip weights are simple functions of the person weights described above , modified only for the purpose of producing annual estimates of the number of person trips. The long-distance trips were recorded in a 28-day period. The long-distance trip weight is simply equal to the final person-level weight multiplied by 365/28.

CHAPTER 3. REPLICATE WEIGHT CONSTRUCTION

To calculate correct standard errors for comparing estimates, it is crucial to account for the complex survey design. Relying on methods that assume a simple random sample will typically underestimate the true sampling error associated with estimates from the NHTS. Replicate weights are provided to allow you to correctly compute estimates of standard errors using appropriate software such as WesVar or SUDAAN.

PERSON-LEVEL REPLICATES

Replicate Construction

To create replicate weights, the data were sorted by geographical characteristics, MSA status, and census division. After sorting, 99 replicates were created using the delete-one jackknife method (JK1). The *i*th replicate was constructed by setting the *i*th item of every 100 records of the full sample weight equal to zero. This procedure was done separately for both the Pre-9/11 and Post-9/11 data files.

Adjustment ror Delete-One Jackknife Method

After "deleting one," the totals of each of 99 replicates are not equal to the control totals. To adjust for this, the replicates are multiplied by factors of each of the 99 replicates for both the Pre-9/11 and Post-9/11 data files, separately.

Factor = Population Total / Sum of the *i*th replicate, where
Population total = 277,208,169

Raking

Again, the raking processes were performed on each of the 99 replicates for both the Pre-9/11 and Post-9/11 person data files, applying the same poststratification process for each of the replicates as for the full sample weight. These raking processes are done in exactly the same manner as described above, using the same eight dimensions and the same control totals.

LONG-DISTANCE TRIP REPLICATES

Long-distance trip replicate weights are simple functions of the person-level replicate weights described in section 3.1, modified for the purpose of producing annual estimates of the number of person trips taken by the respondents. The long-distance trip replicate weights are equal to the final person-level replicates' weights multiplied by 365/28.

CHAPTER 4. DESCRIPTION OF DATA FILES

STRUCTURE OF 2001 NHTS PRE-9/11 AND POST-9/11 DATA FILES

The public use 2001 NHTS pre-9/11 and post-9/11 data are organized into four different data files, which are available to users in ASCII and SAS format. These four data files contain all information including weights, replicates and all other variables.

Structure of 2001 NHTS Pre-9/11 and Post-9/11 Data Files

Data Files	Information Included	Record Level	ID Variables	Weight	Replicates
Pre-9/11Person	These files include education level, worker status, occupation and employer information, driver status, annual miles driven, country of birth, income, and other characteristics of the person.	One record per person (person whose assigned travel date is before or on 9/25/2001)	HOUSEID and PERSONID	WTPRPRE	PERPRE1-PERPRE99
Post-9/11Person		One record per person (person whose assigned travel date is after 9/25/2001)	HOUSEID and PERSONID	WTPRPST	PERPST1-PERPST99
Pre-9/11Long-Distance Trip	Information on these files includes trip purpose, trip distance, and mode of transportation used to and from destination. Household and person characteristics are included with long-distance trip characteristics.	One record per long-distance trip (trip taken by person who is in the pre-9/11 Person-level data file)	HOUSEID and PERSONID	WTLDPRE	LDPRE1-LDPRE99
Post-9/11Long-Distance Trip		One record per long-distance trip (trip taken by person who is in the post-9/11 person-level data file)	HOUSEID and PERSONID	WTLDPST	LDPST1-LDPST99

DERIVED DATA SET

Beside these four data files, we derived two extra SAS data sets for each of these four data files. One only contains ID variables, weights, and replicates. Another contains weights and all other variables. The purpose of creating the extra data sets is user's convenience in downloading manageable files – users can download just the data files needed for their analysis.

The table below shows the data name, data file level, data file time frame, and the variables included for all of the 12 SAS data sets.

Data File Name	Data File Level	Data File Time Frame	Variables Included
PRE_PER_ALL	Person-level	Pre-9/11	Weight, replicates and all other variables
PRE_PER_WGT	Person-level	Pre-9/11	Weight and all other variables, no replicates
PRE_PER_RPL	Person-level	Pre-9/11	Weight, replicates and ID variables, no other variables
PST_PER_ALL	Person-level	Post-9/11	Weight, replicates and all other variables
PST_PER_WGT	Person-level	Post-9/11	Weight and all other variables, no replicates
PST_PER_RPL	Person-level	Post-9/11	Weight, replicates and ID variables, no other variables
PRE_LD_ALL	Long-Distance Trip	Pre-9/11	Weight, replicates and all other variables
PRE_LD_WGT	Long-Distance Trip	Pre-9/11	Weight and all other variables, no replicates
PRE_LD_RPL	Long-Distance Trip	Pre-9/11	Weight, replicates and ID variables, no other variables
PST_LD_ALL	Long-Distance Trip	Post-9/11	Weight, replicates and all other variables
PST_LD_WGT	Long-Distance Trip	Post-9/11	Weight and all other variables, no replicates
PST_LD_RPL	Long-Distance Trip	Post-9/11	Weight, replicates and ID variables, no other variables

DATA DICTIONARY

The file variables are identified by the variable name in the SAS versions. For each file variable, the codebook (Appendix B) contains:

* the variable type and length;
* whether the variable was identical to the one on the 1995 NPTS data set;
* the label, which is a brief description of the variable content;
* the section and item number of the questionnaire or other source of the data;
* value ranges and special codes;
* the unweighted frequency of responses for each value or code shown; and
* the weighted frequency of responses for each value or code shown.

CHAPTER 5. USING THE DATA

TRAVEL CONCEPTS

Appendix E provides abbreviations used in this report, key travel concepts, and a glossary of terms used in the 2001 NHTS. The Travel Concepts portion of Appendix E is primarily geared toward data users who are not familiar with household travel survey data. However, it may also be useful to transportation planning professionals because the exact usage of certain travel terms and concepts often vary by individual survey.

WEIGHTING THE DATA

Chapter 2 described how the weights were constructed for the 2001 NHTS Pre-9/11 and Post-9/11 data sets. The weights reflect the selection probabilities and adjustments to account for nonresponse, undercoverage, and multiple telephones in a household. To obtain estimates that are minimally biased, these weights must be used. Tabulations without weights may be significantly different than weighted estimates and may be subject to large bias. Estimates of the totals are obtained by multiplying each data value by the appropriate weight and summing the results.

The long-distance trip data in this file cannot be used in a simple manner to produce realistic distributions of individual households or persons by number of annual trips. The survey provides the number of trips taken in a 28-day period. Thus, for example, if a person reports taking two long-distance trips in the 28-day travel period, we have no direct knowledge of how many trips the person takes in a year. A simple estimate of number of annual trips is 26 (2*365/28), but of course it is quite likely that the person will have taken fewer trips than this in a year. Similarly, if a person reports taking zero long-distance trips in the 28-day travel period, a simple estimate of number of annual trips is also zero, but of course it is quite possible that the person will have taken a few trips during the year.

EVALUATING THE IMPACT OF 9/11 ON LONG-DISTANCE TRIP TRAVEL PATTERN

The Pre-9/11 and Post-9/11 long-distance trip data files may be used to estimate the annual long-distance trips at the national level for Pre-9/11 and Post-9/11 time frames, respectively. Some long-distance trips taken after 9/11 are in the Pre-9/11 long-distance trip data file and some long-distance trips taken before 9/11 are in the Post-9/11 long-distance trip data file. An analyst may want to delete such trips from the two files. The variable that can be used for this purpose is TPBOA911 (travel period before/on or after 9/11.)

The person-level data file can be merged with corresponding long-distance trip data file, by house ID and person ID, to combine the personal information with long-distance trip information.

CALCULATING THE STANDARD ERROR

The replicate weights may be used to calculate standard errors. Replicate variance estimation is useful because sample estimates are made by using a number of subsamples of the fully conducted survey. One then looks at the difference between each replicate sample estimate and the full sample estimate and squares the difference. Finally, one sums up the squared differences across all the replicates, with an appropriate multiplicative factor. Replicate weights in the NHTS Pre-9/11 and Post-9/11 were constructed using the delete-one Jackknife method (Wolter, K.M. 1985. Introduction to Variance Estimation. New York: Springer-Verlag). These weights can be used to calculate standard error estimates using WesVar or SUDAAN. Standard error estimates can also be easily calculated using the following formula:

$$\sqrt{\frac{98}{99}\sum_{i=1}^{99}\left[REP\left(i\right)-x\right]^{2}}$$

where x is the full sample estimate (calculated by using the full sample weights), *REP(i)* is the estimate calculated by using the replicate weights, and the summation over the index i is from 1 to 99. For a brief introduction to delete-one jackknife variance estimation and the above formula see Raj, D. and Chandhok, P. K. 1998. *Sample Survey Theory*. London, U.K.: Narosa Publishing House.

DATA FILE CONVENTIONS

A number of conventions are followed throughout the NHTS data files. Some of these are also listed in Appendix B, Codebook, and they include:

- Yes/no questions - coded as 1 = yes; 2 = no or 1=yes; 0=no.
- Calendar dates - Multiple variables contain these dates, and usually the year and month are shown as YYYYMM (year followed by the month).
- Times - All reported time variables are in military time as 0000 to 2359.
- Reserve codes - On the ASCII file, the reserve codes of –1, -8, -7 and –9 were used to indicate legitimate skips, don’t knows, refused, and non-ascertained values.
    - Legitimate skip codes - Questions intentionally skipped in the instrument were generally denoted by a -1 in the field.
    - Don’t know - When the respondent indicated that they did not know the response to a question, it was denoted by an -8 in the field.
    - Refused - When a respondent refused to provide a response to a question, it was denoted by a -7 in the field.
    - Not ascertained - When a question should have been asked of the respondent but was not (the question was not a legitimate skip (code -1) for that respondent) or the response provided did not seem correct because it failed an edit check and could not be corrected, the response was set to not ascertained. A not ascertained is denoted by a -9 in the field.

- Missing information for derived variables
    - If a derived variable was derived from just one primary variable, the missing values for the derived variable are identical to the primary variable and could be -1, -7, -8 or -9.
    - If the derived variable was derived from multiple variables, the missing values for the derived variable are -1 or -9. That is, responses of -7, or -8 were set to -9.
    - If the derived variable was not derived from a CATI variable, for example, the weight variables, then missing values are coded as follows:

        - . = missing value for a numeric derived variable
        - Blank = missing value for a character derived variable