

CHAPTER 6. DESCRIPTION OF DATA FILES

6-A. STRUCTURE OF THE DATA FILES

6-A.1. BASIC STRUCTURE

The 2001 NHTS Public Use Data, January 2004 release (Version 3) is organized into four different data files, which are available to users in SAS, ASCII, or DBF formats. Exhibit 6-1 illustrates the structure of the four files, with a description of which data are included in each file, the applicable questionnaire sections, the record level, and the variables that are needed to uniquely identify a record (ID variables).

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 & length,
- whether the variable was identical to one on the 1995 NPTS dataset,
- 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.

For each of the delivery files, Appendix C provides the SAS Proc Contents, the ASCII file layout and the dBase IV file layout. The Appendix displays the name, label, starting position and length of each variable.

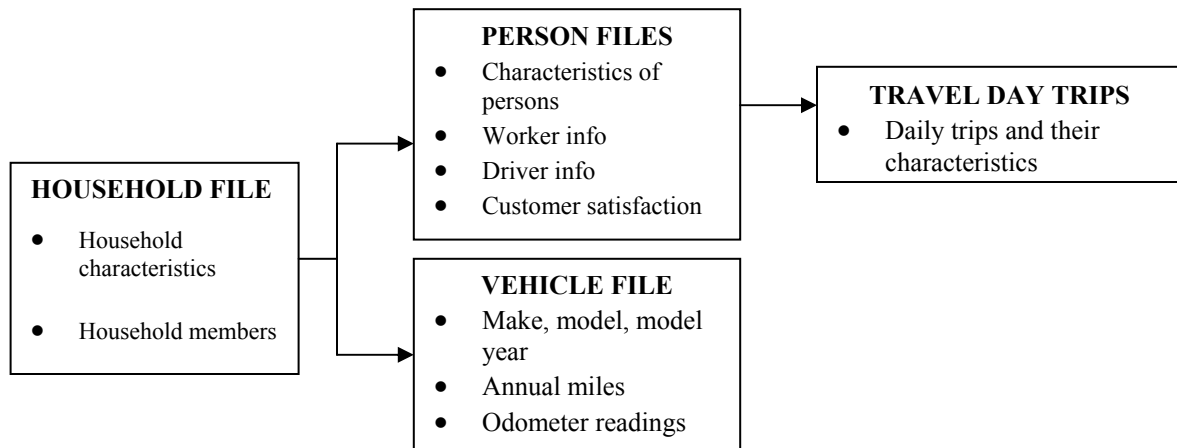
Exhibit 6-1. Structure of 2001 NHTS January 2004 (Version 3) Data Files

Data Files	Information Included	Record Level	ID Variables	Weight Variables¹
House hold file	Data unique to a household, or questions asked once for each sample household. Questions from interview sections: B Number of vehicles C Person Data, Telephone Data, Type of Residence, D Location of Home, M Household Income, Education of Household Respondent.	One record per household	HOUSEID	WTHHFIN WTHHNTL EXPFLHLLH EXPFLHLLH EXPINTHH EXPSCRHH
Person file	Data determined once for each completed person interview. Questions from interview sections: C Age, Race, Driver Status, E Travel to Work, L Miles driven, Customer Satisfaction, M Country of Birth, Education, Person Income, Medical Condition, Internet Use.	One record per person	HOUSEID and PERSONID	WTPERFIN WTPRNTL EXPFLPR EXPFLPRN EXPINTPR
Vehicle file	Data relating to each of the household's vehicles. Questions from interview sections: B Vehicle Data, C Race of Respondent, Type of Residence, L Verified Vehicle Data, Annualized Vehicle Miles, M Education of Respondent, Household Income.	One record per vehicle	VHCASEID	WTHHFIN WTHHNTL EXPFLHLLH EXPFLHLLH
Travel day trip file	Data about each trip the person made on the household's randomly-assigned travel day. Questions from interview sections: C Person Data, G Travel Day Data.	One record per travel day person trip	HOUSEID, PERSONID, and TDTRPNUM	WTTRDFIN WTTRDNTL EXPFLLLTD EXPFLLLTDN

¹ Chapter 7 provides a description of each of the weights.

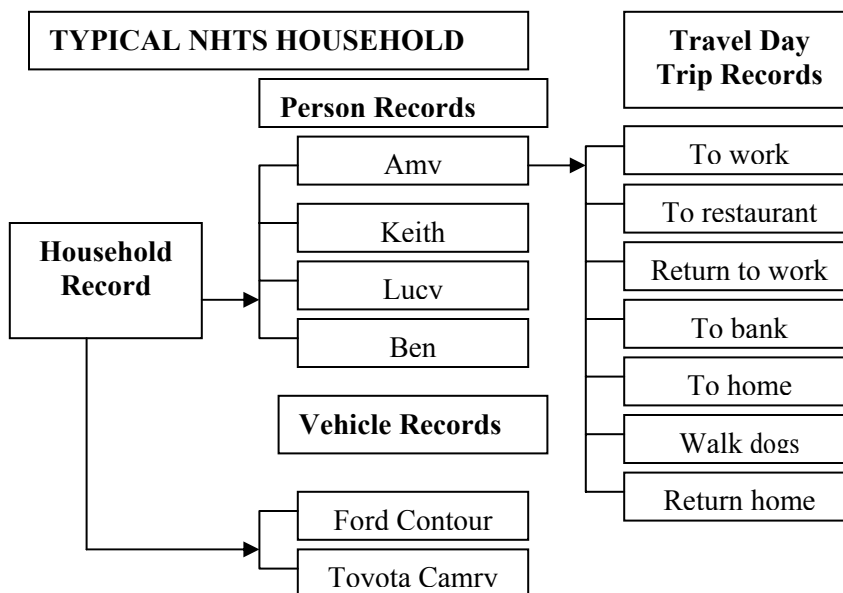
6-B. RELATIONSHIP BETWEEN THE FOUR NHTS DATA FILES

The chart below depicts the four 2001 NHTS Version 3 data files and their relationship.



6-B.1. TYPICAL NHTS HOUSEHOLD

The next chart shows how the records would appear for the data reported by the Typical NHTS Household example introduced in Chapters 1 and 2. Remember that this example household reported only a portion of what would have been reported in an actual NHTS interview.



Note: This follows the Typical NHTS Household material in Chapter 2. In a real household there would probably be trips by each household member; not just by Amy

6-B.2. WHEN IS A RECORD ON THE FILE

The purpose of this subsection is to present information to clarify the NHTS file structure issues that have been confusing to data users in the past.

Household Record - There is one record for each household in the dataset, also called a “useable” household².

Vehicle Record - There is a vehicle record for each vehicle owned, leased or available for regular use by the household members in a useable household. If the household has no vehicles, there are no vehicle records. The number of household vehicles, including zero vehicles, is available on the household record in the variable, HHVEHCNT.

Person Record - There is a person record for each household member who completed a person interview. For example, in our Typical NHTS Household there are four household members. Person interviews were completed for Amy, Lucy and Ben. However, Keith was never available, despite repeated attempts during the six-day travel window. There is a person record for Amy, Lucy and Ben. No person record exists for Keith, but his characteristics, provided by Amy during the household interview, are available to the analyst on the household file (see Section 6-B.3. Household Member Variables).

Travel Day Trip Record - There is a trip record for each trip taken by an interviewed person in a useable household. In Chapter 2, we described the seven trips Amy made in our Typical NHTS Household. Since she made seven trips, there are seven travel day trip records on the file for Amy. If Lucy was ill and stayed home all day there would be no travel day trip records for Lucy, however, there is a person record for her, since she was interviewed. The person file variable, SAMEPLC, i.e. “stayed in the same place all day” would confirm that Lucy was interviewed for travel day and reported no trips. No travel day trip records exist for Keith, since he was not interviewed. Likewise, there is no person file record for Keith.

² A useable household is one where at least 50 percent of the adult household members have completed a person interview.

In earlier NHTSs, before “stayed in same place all day” was asked, data users assumed that the lack of a travel day trip record for Lucy meant that she was not interviewed for her travel day travel. This was not true for the 1990, 1995, and 2001 surveys. If there is a person record for that person, they were interviewed for the details of their travel day. Note that about 12 percent of the 160,758 persons in useable households in the 2001 NHTS reported no travel day trips. Not surprisingly, more of the stay at home days fall on a weekend. Of all persons who did not make a trip on their travel day, 19.9 percent had Saturday as their travel day and 16.8 percent had Sunday. While some of these non-travelling people may be “soft refusals” who did not want to bother reporting their trips, many of them are legitimate non-travelers. Remember that the NHTS travel days encompass all 365 days of the year, including holidays and weekends.

6-B.3. HOUSEHOLD MEMBER VARIABLES

For the 2001 NHTS, the characteristics of all household members, whether interviewed or not, are available on the Household File. These characteristics were included to allow the user to address a number of travel behavior and survey method research issues. They provide the full profile of the household and allow users to know the characteristics of those household members who completed the person interview and those who did not. The characteristics are contained in variable names that end with P1 through P14 which is the maximum number of household members in a 2001 NHTS household. The information provided for each household member includes:

- Age (AGE_P1 through AGE_P14),
- Driver status (DRV_P1 through DRV_P14),
- Relationship to Household Respondent (REL_P1 through REL_P14),
- Sex (SEX_P1 through SEX_P14),
- Person interview response status, i.e., whether a person interview was completed, etc. (STAT_P1 through STAT_P14), and
- Worker status (WKR_P1 through WKR_P14).

6-C. CODEBOOK

6-C.1. CODEBOOK FORMAT

Appendix B contains the codebook with sections for each of the data files. The codebook contains nine items of information about each variable in each of the files. Exhibit 6-2 lists the items that are arranged in the codebook as columns, along with a brief description of the contents of each column. Appendix D, Derived Variables provides additional detail on how each of the derived variables in the codebook was calculated.

6-C.2. CODEBOOK EXAMPLE

As an example, the third column of Exhibit 6-2 shows the codebook information for the variable named VEHOWNMO.

- It is a numeric variable of length eight including the decimal point. The decimal point position is not fixed. The format for the variable in SAS is 6.2 (up to three digits before the decimal point and two after). Formats for each variable are provided in Appendix A, Data Dictionary;
- This is a derived variable (denoted by an * to the right of the question number). The variable it was derived from was not asked the same way in the 1995 NPTS;
- The variable contains the length of time the vehicle was owned converted to months. It is derived from question L8. The reported length of time the vehicle was owned (days, weeks, months or years) was converted to months based on the questionnaire variable OWNUNIT (question L8) (which is not included in the dataset).
- The value range and the frequencies show that the file contains 21,005 reports ranging from 0 to 11.7 months; that there were no subjects who did not know the distance, and none refused to answer the question. It also shows that the question was legitimately skipped for 118,366 subjects. Another 11 subjects had a value of not ascertained in the field; and
- Further details regarding this variable are found in Appendix D.

Exhibit 6-2. Contents of the 2001 NHTS Codebook

Column Heading	Description of Contents	Example Variable (From Person File)
Target Variable	The variable name	VEHOWNMO
Variable Type	C = Character; N = Numeric	N
Variable Length	Maximum variable length	8
1995 Variable Comparison	Y Identical to 1995 NR Response categories are different NQ Question is different NQR No match with 1995 SD Some difference in the derived variable X Derived variable did not exist in 1995	NQR
Variable Label	Short description of the variable	How long vehicle owned - months
Question Number	Source item(s) in the questionnaire section	L8*
Value Range & Codes	Either lists all possible values of the variable, a range of the values, or a combination of the two	0 – 11.7 -1 = Legitimate skip -7 = Refused -8 = Don't know -9 = Not ascertained
Unweighted Frequencies	Shows the number of records in the file for each listed value	0 – 11.7 = 21,005 -1 = 118,366 -7 = 0 -8 = 0 -9 = 11

Exhibit 6-2. Contents of the 2001 NHTS Codebook (continued)

Column Heading	Description of Contents	Example Variable (From Person File)
Weighted Frequencies	Shows the corresponding weight for each listed value for the variable	0 – 11.7 = 30,542,988 -1 = 172,023,762 -7 = 0 -8 = 0 -9 = 19,450
Footnote	Refers the user to other sections of the User's Guide for more information	Refer to Appendix D for more detail on derived variables

6-C.3. COMPARABILITY WITH 1995 NHTS

Emphasis was placed on making the 2001 NHTS data files comparable to the 1995 NPTS data files. We compared each of the questions in the 1995 NPTS with those in the 2001 NHTS. The fourth column in the codebook, 1995 Variable Comparison, provides a code that compares the questions and the response categories under each question in the 1995 and 2001 surveys. When comparing data values in variables across surveys, we recommend data users pay attention to the codes in this column irrespective of whether the variable names are identical in the two surveys. If the code indicates that the questions were identical, then the values can be compared with no adjustments. However, if questions were different, adjustments are recommended when comparing results across survey years. The codes in the column are:

- Y (Identical to 1995) Indicates that the 1995 and 2001 questions were identical. That is, both the wording of the question and the response categories were identical,
- NR (Response Categories are Different) Indicates that the wording of the questions were identical to 1995 but the response categories were different,
- NQ (Question is Different) Indicates that the response categories were identical to 1995 but the wording of the question was different,

NQR (No match with 1995) Indicates both the wording and response categories were different,

SD (Some difference in the Derived Variable) Indicates the variable was derived and that the derived variable is not identical to the one in 1995, and

X (Derived Variable did not exist in 1995) Indicates the variable was derived and did not exist in 1995.

6-D. VARIABLES REPEATED

In addition to the information specific to its file, each of the four files includes variables from other files to be used along with its own variables (e.g., the travel day file contains data on the individual travel day trips). This is done for the convenience of the data user to minimize the need to merge data from multiple files. Although this format is less desirable from a data storage standpoint, it significantly simplifies subsequent data manipulation.

On the following page we list the commonly used variables that have been included in all four data files:

Variable Name	Label
CDIVMSAR	HHs by Census div., MSA size, rail
CENSUS_D	Household Census Division
CENSUS_R	Household Census Region
DRVRCNT	Count of drivers in HH
HBHRESDN	Housing units per sq mile - Block group
HBHTNRNT	Percent renter-occupied - Block group
HBHUR	Urban / Rural indicator - Block group
HBPPOPDN	Population per sq mile - Block group
HHC_MSA	MSA / CMSA code for HH
HHFAMINC	Total HH income last 12 months
HHINCTTL	Total income all HH members
HHR_HISP	Hispanic status of HH respondent
HHR_RACE	Race of HH respondent
HHSIZE	Count of HH members
HHSTATE	State-household location
HHSTFIPS	FIPS state code for HH
HHVEHCNT	Count of HH vehicles
HOMEOWN	Housing unit owned or rented
HOMETYPE	Type of housing unit
HOUSEID	HH Identification Number
HTEEMPDN	Jobs per sq mile - Tract level
HTHRESDN	Housing units per sq mile - Tract level
HTHTNRNT	Percent renter-occupied - Tract level
HTHUR	Urban / Rural indicator - Tract level
HTPPOPDN	Population per sq mile - Tract level
LANG	Language interview was conducted in
LIF_CYC	HH life cycle
MSACAT	MSA category
MSASIZE	Population size of HH MSA
NUMADLT	Number of adults in HH
RAIL	Rail (subway) category
SMPLAREA	Add-on area where HH resides
SMPLFIRM	Firm collecting the data
SMPLSRCE	Sample where the case originated
TDAYDAT2	Travel day date (YYYYMMDD)
TDAYDATE	Travel day date (YYYYMM)
TDBOA911	Travel Day Before or On/After 9/11
TRAVDAY	Travel day - day of week
URBAN	Household in urbanized area
URBRUR	Household in urban/rural area
WRKCOUNT	Count of HH members with jobs

6-E. DERIVED VARIABLES

Over 239 derived variables were created during the development of the four public use and DOT research files released in January 2004 for the 2001 survey. These exclude variables created for the travel period and most recent research files for DOT. These variables are included in Appendix D, Derived Variables. The Appendix provides documentation on how each of the variables was derived. These variables are considered derived as they do not appear in the questionnaires included in Appendix M and therefore no data was stored in these variables during data collection. The variables were derived by:

- renaming a questionnaire variable to match names used during the 1995 survey or new names provided by DOT,
- calculating the variable from one or more variables in the questionnaires to provide summary variables to aid data users,
- obtaining the variable from external sources to provide additional descriptors, or
- creating flag variables to identify data records that had been edited or imputed.

Among the derived variables are eight variables that concern the estimation of annualized mileage for each household vehicle. These variables were derived by Oak Ridge National Laboratories using NHTS survey data and are described in detail in Appendix J – Methods to Estimate Annual Miles Driven per Vehicle.

The U.S. Energy Information Administration provided ten other derived variables that estimate the vehicle fuel economy, vehicle fuel consumption, and vehicle fuel expenditures. These were derived from data from the 2001 National Household Travel Survey (NHTS); the U.S. Energy Information Administration (EIA) 1985, 1988, and 1991 Residential Transportation Energy Consumption Survey (RTECS); the U.S. Environmental Protection Agency (EPA) fuel economy test results; and the EIA's retail pump price series for 2001 and 2002. The details of these variables are presented in Appendix K, Estimation Methodologies for Fuel Economy and Fuel Cost.

Nine additional derived variables were added to describe the characteristics of the areas where the NHTS survey respondents live. These variables were derived from 2000 Census data and estimated forward to 2002-2002 by Claritas, Inc. Details of these variables are presented in Appendix Q, Tract and Block Group Variables.

6-E.1. 1990 TRAVEL DAY TRIP PURPOSES

The travel day trip purpose definitions for the 2001 NHTS differed from those used in the 1990 and 1995 NPTS. The recoded 1990 trip purposes will be particularly useful for analyses comparing the 1990, 1995 and 2001 data by purpose. For each travel day trip, the data set includes both variables that provide the 2001 trip purpose and the derived 1990 trip purpose. The 1990 trip purpose was calculated by recoding the 2001 trip purpose to match the way trip purposes were collected during the 1990 NPTS. This recoded 1990 trip purpose is stored in the variable WHYTRP90.

The 2001 trip purposes use a “from-to” format, while the 1990 purposes were based on coding a “main reason” for the trip. As a result, the trip purpose codes used in 2001 differed from the 1990 trip purposes in the following ways:

- A trip "to home" after completing an activity is categorized as "return home" in 2001 purposes but was not a 1990 trip purpose. In 1990, the trip purpose was assigned to the activity that was the main reason the subject was away from home,
- In 1990, if one of the purposes was work, the return trip home was assigned a work purpose, even if there were incidental trips made on the way home,
- In 1990, if there were multiple purposes for being away from home and work was not one of them, the respondent was asked the main reason for the trips. Because this “main reason” format was not used in the 2001 survey, when the 2001 purposes were recoded to the 1990 scheme, the activity the person spent the most time at while away from home was assigned as the main purpose for the return trip home. The variable, DWELTIME, was used to determine this.