

## **CHAPTER 7. USING THE DATA**

### **7-A. TRAVEL CONCEPTS**

#### **7-A.1. OVERVIEW**

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 the transportation planning professional because the use of certain travel terms and concepts often vary by individual survey. Appendix E contains definitions of the following measures of personal travel, when to use each, and how to compute them with the NHTS data:

- Person Trips
- Person Miles of Travel (PMT)
- Vehicle Trips
- Vehicle Miles of Travel (VMT)
- Vehicle Occupancy

### **7-B. TABULATING THE DATA**

#### **7-B.1. SAMPLE TABLES AND LOGIC**

Appendix G contains 12 sample tables that provide tabulations of some of the most commonly used variables. Tables were chosen to illustrate the national-level estimates that would be tabulated by many data users, such as:

- total households by income and vehicle ownership patterns,
- total persons by age, race and gender,
- total numbers of workers, drivers, person trips, person miles, vehicle trips, and vehicle miles, and

- vehicle occupancy and commute time tabulations.

Each cell of each of the tables contains the:

- sample size
- weighted estimate, and
- sampling error of each weighted estimate.

These tables were prepared using the WesVar survey data analysis software developed by Westat.

## **7-B.2. ADDITIONAL RESOURCES**

**NHTS Website** - <http://nhts.ornl.gov/2001>

The NHTS Website offers:

- analysis capability which will include production of user-defined tables,
- a component for exploratory analysis of the data,
- a number of standard NHTS tables,
- a conference portion to allow the data user to communicate with others, share code, etc., and
- papers and articles analyzing the NHTS data.

NHTS Training - FHWA and BTS are developing an interactive CD-ROM as a stand-alone training tool. This will allow individuals to obtain training that fits with their needs.

Contact information for user support:

NHTS Website:	Oak Ridge National Laboratory ORNL, (865) 946-1257 <a href="mailto:rtg@ornl.gov">rtg@ornl.gov</a>
Other User Support:	NHTS Team FHWA, (202) 366-0160 BTS, (202) 366-2546
Statistical Issues	Lee Giesbrecht, BTS (202) 366-2546

## **7-C. CONTROL NUMBERS**

Control totals and weight sums, which are the two most useful control numbers are described briefly below.

### **7-C.1. CONTROL TOTALS**

Control totals are known values, external to the survey itself, which are used to adjust the survey weights for non-response and non-coverage. Control totals were used to adjust the 2001 NHTS weights for:

- the number of U.S. households, and
- the number of persons in these households.

The control categories chosen for the 2001 NHTS and the weighting procedure are described in Chapter 5 of this User's Guide. Appendix F contains the full complement of Control numbers for the 2001 NHTS data set. The variables used to define nonresponse adjustment cells are in Appendix H.

## 7-C.2. WEIGHT SUMS

Weight sums are simply the calculated sums of the survey weights. These values are helpful to users in verifying the correctness of data tabulations. The 2001 NHTS total sample sizes and weight sums for the four data files are in Table 7-1. A full set of sample sizes and weight sums that can be used for checking output are contained in Appendix F, Table 3.

**Table 7-1. File Sample Sizes and Weight Sums**

<b>Data File</b>	<b>Sample Size</b>	<b>Weight Sum</b>
Household	69,817	107,365,346
Person	160,758	277,203,235
Vehicle	139,382	202,586,200
Travel day person trips	642,292	407,262,485,207

## 7-D. WEIGHTING THE DATA

### 7-D.1. WHY USE WEIGHTS

Chapter 5 described how the weights were calculated for the 2001 survey. 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, 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.

## 7-D.2. WHICH WEIGHT TO USE

There are several different weights, and it is important that the appropriate weight is used for a particular estimate. There are sets of weights for the full sample and for the national-only sample. For each set, there are household weights, person weights, travel day and travel period weights. Travel Period data have not been included in the January 2004 release of the data, but will be released later in the year by BTS.

- Compared to national-only weights, full sample weights have the advantage of being based on a larger sample size and therefore produce estimates with lower sampling errors. Since the additional sample is largely concentrated in some small population geographic areas, the sampling errors are not reduced very much for most national estimates. For sub-national estimates specifically for an add-on area or an area that is only a little larger than an add-on area, the sampling errors will be much smaller for the full sample weights and therefore should be used.
- Response rates were significantly higher for the national sample than for most of the add-on sample areas. Thus, there is potentially higher bias in estimates based on the full sample than on the national-only sample. Estimates for small subgroups tend to have large relative sampling errors and thus any bias due to nonresponse is likely to be small compared to the sampling error. For such estimates, it is preferable to use full sample weights. For most estimates, however, bias may be large compared to sampling error, and thus it may be preferable to use the national sample.
- Household weights are used whenever one is tabulating an estimate at the household level as opposed to the person level, such as number of households by household vehicle ownership and distribution of households by number of household drivers.
- Travel day weights are used for estimates involving numbers of trips or miles of travel, for example, number of vehicle trips by trip purpose. Only trips in personally owned vehicles that are reported by the driver should be counted in estimating personal vehicle trips. (For example, if a person reports being a passenger in a vehicle driven by another member of the household, that trip would not be counted.)
- Travel period weights are used for estimates involving numbers of trips or miles of travel for trips of more than 50 miles as obtained for the 28-day travel period. The travel period household weight is used for estimates of household trips, and the travel period person weight is used for estimates of person trips.

- Person weights are used for all other estimates (i.e., for non-household and non-travel day items of interest).

Note that for some estimates requiring ratios, different weights should be used for the numerators than for the denominators. For example, for estimates of daily trips per household, travel day weights are used for the numerator (since the numerator involves person trips) and household weights are used for the denominator (since the denominator is the weighted number of households). As a second example, for estimates of average time spent driving by all drivers, travel day weights are used for the numerator and person weights are used for the denominator (since drivers are a subset of persons).

Table 7-2 gives the variable names for full sample weights, and Table 7-3 gives the variable names for national weights.

### **7-D.3. WHICH HOUSEHOLD WEIGHT TO USE**

There are two different household weights as shown in Table 7-2 and in Table 7-3. If one wishes to use those households for which there were completed interviews for at least half of the adults, the useable household weight should be used. If one wishes to use only those households for which there were completed interviews with all adults in the household, the 100 percent reported weight should be used. Finally, there are two different travel period (household) weights that differ from the household weights only in that they have a multiplier of 365/28.

### **7-D.4. WHICH PERSON AND TRAVEL DAY WEIGHT TO USE**

There are also two different person weights, one for persons interviewed in useable households and one for persons interviewed in 100 percent reported households. Table 7-2 and Table 7-3 provide variable names for both weights. There are two different travel day weights that differ from the person weights only in that they have a multiplier of 365. Finally, there are two different travel period person weights that differ from the person weights only in that they have a multiplier of 365/28.

Tables 7-2 and 7-3 provide the variable names for the weights and the replicate weights. Section 7-E, Source of Errors discusses how they may be used to estimate sampling errors.

**Table 7-2. Description of the Different Full Sample Weights on the 2001 NHTS**

		Household	Person	Travel day person
Useable Households	Weight	WTHHFIN	WTPERFIN	WTTRDFIN
	Replicates	WTHFIN1-99	WTPFIN1-99	WTDFIN1-99
100% Reported Households	Weight	EXPFLHH	EXPFLPR	EXPFLTD
	Replicates	EXPFH1-99	EXPFR1-99	EXPFTD1-99

**Table 7-3. Description of the Different National Weights on the 2001 NHTS**

		Household	Person	Travel day person	Travel period household	Travel period person
Useable Households	Weight	WTHHNTL	WTPRNTL	WTTRDNTL	WTTRPNTL	WTPTPFIN
	Replicates	FHHWT01-99	FPERWT01-99	FTRDWT01-99	FHTPWT01-99	FPTPWT01-99
100% Reported Households	Weight	EXPFLHHN	EXPFLPRN	EXPFLTDN	EXPFLTPN	EXPFLPTP
	Replicates	EHHWT01-99	EPERWT01-99	ETRDWT01-99	EHTPWT01-99	EPTPWT01-99

## 7-E. SOURCE OF ERRORS

### 7-E.1. SAMPLING ERRORS

Since every person and household in the U.S. were not included in this survey, the sample estimate may differ from the result that would have been obtained if a census were conducted under the exact same circumstances. Calculating sampling errors provides the basis for measurement of the variability in the estimated statistics,

and allows analysts to make probability statements about how large the difference may be between an estimated sample statistic and what would have been obtained for that statistic had a census been conducted.

The replicate weights that use the full sample variable names given in Table 7-2 as prefixes may be used to calculate standard errors. The idea in replicate variance estimation is that sample estimates are made for 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.

The replicate weights were calculated using the delete-one Jackknife method<sup>9</sup>. 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:

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

where  $x$  is the full sample estimate (calculated by using the full sample weights) and  $REP(i)$  is the estimate calculated by using the replicate weights and the summation over the index  $i$  is from 1 to 99. For example, suppose one is interested in an estimate of persons for Option 1 using the full sample. The weight  $WTHHFIN$  is used to calculate the overall estimate  $x$ . The weight  $WTHHFIN1$  is used to calculate the estimate  $REP(1)$ , the weight  $WTHHFIN2$  is used to calculate the estimate  $REP(2)$ , etc. Replicate weights are provided only for households and persons. For vehicles, use the household replicate weights. For travel day trips, use the person weight times 365. For travel period trips, use the person weight times 365/28.

As an example of the use of standard errors, the weighted survey estimate of household vehicles is 202,586,200 with an estimated standard error of 672,072. This standard error estimate allows one to conclude with 95% confidence probability that the interval 201,242,056 to 203,930,344 contains the estimated number of household

---

<sup>9</sup> Wolter, KM. (1985) *Introduction to Variance Estimation*. New York: Springer-Verlag

vehicles that would have been obtained if a census were conducted using the same procedures.

## **7-E.2. NONSAMPLING ERRORS**

There are many sources of error in addition to error occurring because only a sample was selected. Some examples of nonsampling include:

- A respondent misunderstands a question and answers it incorrectly,
- A respondent does not recall a trip or remembers details of the trip incorrectly,
- An interviewer does not correctly record what the respondent says,
- A person or household is a nonrespondent, and
- A person does not answer a specific question.

Undercoverage may also be a source of error. Undercoverage occurs for several reasons, including that a household has no telephone, a person states incorrectly that the telephone number we have dialed is not residential, and the household respondent either accidentally or purposely does not report all the people living in the household.

Note that nonsampling error can sometimes be much larger than sampling error. Furthermore, for this survey good estimates of sampling error are possible but, as in most surveys, it is impossible to estimate nonsampling error.

## **7-F. FINDING THE VARIABLES YOU WANT**

### **VARIABLE LISTS**

The 2001 NHTS datasets are large and complex, containing numerous survey and external (derived) variables. In addition to the codebook for each of the four NHTS data files, the following variable lists are available to assist users in locating NHTS variables:

1. SAS Proc Contents - Appendix C contains SAS proc contents lists for each of the four NHTS data files. The survey variables are listed in alphabetic order on each of these four listings.
2. ASCII File Variable Lists - Appendix C also contains the list of each ASCII variable, with its position and length on each of the four files. The ASCII variables for each NHTS file are ordered as follows:
  - First: ID and weight variables,
  - Second: questionnaire variables in order by question number, and
  - Last: all stratification variables, computed or derived variables and external variables.
3. Data Dictionary Listing - This list shows all of the variables that are contained in all four 2001 NHTS data files in a single alphabetic listing. Since many variables are in more than one file, the data dictionary list has four columns indicating which data files contain each of the variables. The data dictionary is Appendix A.

## **7-G. USING THE DATA FROM MULTIPLE FILES**

### **7-G.1. MERGING FILES**

Despite the effort to include often used variables on multiple files (see Section 6-D), there still comes a time when it is necessary to use information from separate files for an analysis. For example, to study the daily trip patterns of different types of privately-owned vehicles (POVs), one needs to use the variable VEHTYPE (vehicle type) from the Vehicle file and link it to trip characteristics maintained in the Travel Day file. In these types of circumstances, one needs to merge together two or more of the four files.

File merging can be complicated and confusing, and a mistake can lead to invalid analysis results. However, an understanding of how the four files are structured and relate to each other can significantly help clarify the process.

## 7-G.2. ID NUMBERS

Each unit (e.g. households, persons) in the survey has its unique identification number (ID). Specifically, each household is identified by a unique nine digit household ID (HOUSEID). Within each household, household members are identified by a two digit person number (PERSONID) and, similarly, household vehicles are identified by a two digit vehicle number (VEHID). Again, trips taken by an individual are numbered by a trip number (TDTRPNUM for a travel day trip and TPTRPNUM for a travel period trip).

With this numbering system, the number that identifies a unit within a household (e.g., the household's vehicles and household members) needs to be used in conjunction with the household ID to uniquely identify that unit. For example, if a household has a HOUSEID of 123456789, its first member has a PERSONID of 01, and its second member has a PERSONID of 02, then the first household member is uniquely identified by an ID of 12345678901 and the second member 12345678902.

Similarly, the number that identifies a trip taken by an individual needs to be used in conjunction with the person's unique ID (i.e., HOUSEID and PERSONID) to uniquely identify that trip.

Continuing the above example, assume that the first household member took three travel day trips on the assigned travel day. Thus, TDTRPNUM for the first trip is 01, the second trip 02 and the third trip 03. An ID of 1234567890101 will uniquely identify the first trip taken by the first household member of Household 123456789. Likewise, an ID of 1234567890102 and an ID of 1234567890103 will uniquely identify the second and the third trips taken by the same person, respectively. The third trip ID is represented as:

$$\text{HOUSEID} + \text{PERSONID} + \text{TDTRPNUM} = \{123456789\}\{01\}\{03\}$$

Table 7-4 shows which ID variables to use in the most common data linking of any two data files. Note that the linking ID must be common to both the "from" and "to" files. For example, in linking Person file data with Travel Day trip data, the variable TDTRPNUM would not be used because it is only on the Travel Day file, not on the Person file.

**Table 7-4. Examples of Link Variables Between the Four 2001 NHTS Data Files**

<b>From File 1</b>	<b>To File 2</b>	<b>Linking ID Variables</b>
Household file	Person file	HOUSEID
Household file	Vehicle file	HOUSEID
Household file	Travel day trip file	HOUSEID
Person file	Vehicle file	HOUSEID
Person file	Travel day file	HOUSEID and PERSONID
Vehicle file	Travel day file	HOUSEID

### **7-G.3. ID VARIABLES NOT ALWAYS SEQUENTIAL**

The ID variables within a file are not always sequential. There are a number of reasons for this. Examples explaining these reasons were provided in Section 3-D, Data Editing. Some of the reasons why the numbers are not sequential are:

- Some persons and vehicles reported by the household respondent were later found not to belong with the household and were deleted from the data set,
- Some trip segments reported as separate trips were combined during editing, and
- Some trip segments reported as a single trip were split into two.

### **7-G.4. MERGING DATA FILES**

Depending on the nature of the analysis, merging files is typically based on a variable common to the files. The file-merging approach is illustrated here using an example. In this example, the user wants to analyze the impact, if any, of occasional telecommuting on the number of daily trips. The trip-making data are contained in the Travel Day file while the variable indicating occasional telecommuting is in the Person file (WKFMHM2M). That is, the Travel-day file needs to be merged with the Person file.

The variables HOUSEID and PERSONID combined enable one to use the Person file to identify those who occasionally telecommute and those who do not. Using the combined identification number for HOUSEID and PERSONID, one can identify trips taken by that person in the Travel Day file. In this case, HOUSEID and PERSONID combined is the common identification needed to merge the Travel Day and Person files.

In layman's language, the computer is first instructed to "grab" the variable WKFMHM2M, which holds the data on whether the respondent occasionally telecommutes, along with the associated HOUSEID and PERSONID variables from the Person file. Next, the computer is instructed to identify from the Travel Day file all trips that are taken by that person. That is, having the same combined HOUSEID and PERSONID identification number.

Finally, the computer is told to "match" information on occasional telecommuting to the travel day trips based on the combined HOUSEID and PERSONID identification number.

After the files are successfully merged, the next question in using the merged file is which weighting factor to use. Section 7-D provides details on the weights to use.

## **7-H. SPECIAL USER NOTES**

### **7-H.1. DATA FILE CONVENTIONS**

There are a number of conventions 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
- Calendar Dates - multiple variables contain these dates, usually the year and month are shown as follows:

- YYYYMM = year followed by the month
- Times - all reported time variables are in military time as:  
0000 to 2359
- 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 - Variables in the dataset that were derived from one or more other variables are listed in Appendix G.
  - 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 is 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
- Survey weights - there are two weight variable on each file. Section 7-D provides guidance on which weight to use.