

## CHAPTER 5 . USING THE DATA

### 5-A. TRAVEL CONCEPTS

#### OVERVIEW

The Travel Concepts portion of **Appendix D** is primarily geared toward NPTS 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 D** contains definitions of the following measures of personal travel, when to use each, and how to compute them with the NPTS data:

- Person Trips
- Person Miles of Travel (PMT)
- Vehicle Trips
- Vehicle Miles of Travel (VMT)
- Vehicle Occupancy
- Trip Chains
- Overlap Trips (used when adding Travel Day and Travel Period data)

### 5-B. TABULATING THE DATA

#### SAMPLE TABLES & LOGIC

**Appendix B** contains 12 sample tables, computed at the national level. The sample tables were chosen to illustrate frequently used data tabulations. Tables were chosen to illustrate the national-level estimates which would be tabulated by many data users, such as estimated:

- 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.

The 12 sample tables in **Appendix B** also include 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 SUDAAN survey data analysis software developed by RTI. The computer logic used to prepare the data input to make the tables is also included in **Appendix B**.

**ADDITIONAL  
RESOURCES**

**NPTS Website: <http://www-cta.ornl.gov/npts>**

The NPTS Website offers:

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

**NPTS Training** - FHWA is 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:

NPTS Website: Oak Ridge National Laboratories  
ORNL, (423) 574-5958  
rtg@ornl.gov

User Support (Non-Web) FHWA, (202) 366-5026  
OHIM.gatekeeper@fhwa.dot.gov  
Fax (202) 366-7742

**5-C. CONTROL NUMBERS**

Two kinds of control numbers, control totals and weight sums, are described briefly below.

**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 1995 NPTS weights for:

- (1) the number of U.S. households, and
- (2) the number of persons five years of age and older.

The control categories chosen for the 1995 NPTS and the method used to make the adjustments, also known as a post-stratification

weight adjustment procedure, are described in Section 3-G of this User's Guide. **Appendix A** contains the full complement of Control numbers for the 1995 NPTS data set.

## 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 1995 NPTS total sample sizes and weight sums for the six data files are as follows:

**Exhibit 5.1 - File Sample Sizes and Weight Sums**

Data File	Sample Size	Weight Sum
Household	42,033	98,990,000
Person	95,360	241,675,000
Vehicle	75,217	176,066,658
Travel day trip	409,025	378,930,363,336
Segmented trip	3,779	3,440,664,924
Travel period trip	29,647	1,996,178,135

## 5-D. WEIGHTING THE DATA

### MUST USE THE WEIGHTS

Calculation of survey weighting factors for the 1995 NPTS data was discussed earlier in Section 3-G of this User's Guide. The weights reflect the sample design and selection probabilities, over-sampling of certain strata, and adjustments to compensate for survey non-response and non-coverage.

The weights are multiplicative factors that **must** be applied to the file variables in order to obtain valid estimates of population values. If the weights are not used, the tabulations will give incorrect results. For example, overall unweighted daily sample trips per household are 9.73, whereas overall weighted daily trips per household are 10.49. Sample error can be magnified and lead to serious inaccuracies when weights are not used in tabulating these data.

The estimated weighted totals are obtained by multiplying each data value by the appropriate weight and summing the results. The purpose of weighting the data is to obtain valid estimates of national and regional totals for the U.S. population.

**OVER-SAMPLING**

Large metropolitan areas with subway or elevated rail transit systems were over-sampled in order to increase the number of in-sample transit trips. Also, several geographic areas purchased NPTS add-on contracts, increasing the sample sizes within their planning areas in order to provide small-area data for transportation planning. The target sample size for the national sample was 21,120 useable households. Additional samples of useable households were provided to five add-on areas, as shown in Exhibit 5.2.

**ADD-ON AREAS**

Over-sampling certain strata to increase the sample sizes increases the selection probabilities for each household in the sampling frame for the over-sampled areas. The larger selection probabilities translate into smaller weighting factors for the over-sampled strata, correcting the weighted results for the effect of the over-sampling. Note that Exhibit 5.2 shows that the five add-on areas accounted for 55.2 percent of the final useable households in the 1995 NPTS data set, though they accounted for only 10.8 percent of the initial 1995 NPTS target sample size at the national level, and 10 to 11 percent of U. S. households. It would be especially dangerous to rely on unweighted tabulations made from the 1995 NPTS data files, because of the heavy over-sampling rates applied in the add-on areas. That is, national data tabulations made without weighting the data would look a lot like data for New York and Massachusetts. Weighting the data eliminates this problem and corrects the sample estimates.

## Exhibit 5.2 - Target and Final Sample Sizes, at the National and Add-on Levels

Geographic Area	National Sample	Add-on Sample	Total Target	Final Actual
New York	1,683	9,189	10,872	11,004
Massachusetts	490	7,500	7,990	7,801
Central Oklahoma	68	2,944	3,012	2,956
Tulsa, Oklahoma	51	962	1,013	976
Puget Sound	-	300	300	326
Remainder of United States	18,828	-	18,828	18,970
Totals	21,120	20,895	42,015	42,033

### 5-E. SAMPLING ERRORS

#### EXAMPLE

Sample surveys are conducted when time or resources are not available to enumerate every household or person. Because every person was not included, the sample has an error associated with the results. Calculating sampling errors allows the measurement of the variability in the estimated statistics, and allows analysts to make probability statements about how large the difference may be between a sample statistic and its population value.

For example, the 1995 NPTS estimated number of household vehicles in the United States is 176,067,000 with an estimated standard error of 828,000 (see Table 2 in **Appendix B**). This standard error estimate allows one to make the following probability statement

"We are 95 percent confident that the number of household vehicles in the United States in 1995 was between 174,411,000 and 177,723,000."

That is, statistical theory tells us that estimated statistics will be within two standard errors of the census value in 95 percent of the possible samples that we may select. Here the census value is the value that would have resulted had the 1995 NPTS survey

been conducted in all United States households, rather than in a sample of households.

## **USE THE WEIGHTS**

When calculating sampling error estimates, it is absolutely necessary to use the survey weights and formulas which properly account for the sample design used for the survey. The 1995 NPTS survey data set is based on a complex sampling design that includes stratification, unequal weighting and clustering of persons, vehicle, and trips. Sampling errors are typically decreased by stratification and increased by sample clustering and unequal weighting, with clustering normally being the dominant factor. **Many standard statistical packages, including SAS, do not calculate sampling errors properly using data from the NPTS or other complex samples.** See **Appendix G** for additional information about properly computing NPTS sampling errors.

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

### **VARIABLE LISTS**

The 1995 NPTS data sets are large and complex, containing numerous survey and external variables. In addition to the code books for each of the six NPTS data files, the following variable lists are available to assist users in locating NPTS variables:

1. SAS Proc Contents - **Appendix I** contains SAS proc contents lists for each of the six NPTS data files. The survey variables are listed in alphabetic order on each of these six listings.
2. ASCII File Variable Lists - **Appendix I** also contains the list of each ASCII variable, with its position and length on each of the six files. The ASCII variables for each NPTS file are ordered as follows:
  - first, ID and weight variables
  - second, questionnaire variables in order by questionnaire section and item 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 six 1995 NPTS data files in a single alphabetic listing. Since many variables are in

more than one file, the data dictionary list has six columns indicating which data files contain each of the variables. The data dictionary is **Appendix H**.

## 5-G. USING THE DATA FROM MULTIPLE FILES

### MERGING FILES

Despite the effort to include as many "common" variables as possible (see Section 4-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 six files.

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

### ID NUMBERS

Each unit (e.g. households, persons) in the survey has its unique identification number (ID). For example, each household is identified by a unique household ID (HOUSEID). Within each household, household members are numbered by a person number (PERSONID) and, similarly, household vehicles are numbered by a vehicle number (VEHID). Again, trips taken by an individual are numbered by a trip number (TRPNUM for a travel day trip or TRIPNUM 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 12345678, 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 12345678**01** and the second member 12345678**02**.

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 trips during the sample day. Thus, the number TRPNUM for the first trip is 01, the second trip 02 and the third trip 03. An ID of 1234567801**01** will uniquely identify the first trip taken by the first household member of Household 12345678. Likewise, an ID of 1234567801**02** and an ID of 1234567801**03** will uniquely identify the second and the third trips taken by the same person, respectively. The last trip ID is represented as:

HOUSEID;PERSONID;TRPNUM = {12345678}{01}{03}

Exhibit 5.3 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 TRIPNUM would not be used because it is only on the Travel Day file, not on the Person file.

**Exhibit - Examples of Link Variables Between the Six 1995 NPTS 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
Household file	Travel period file	HOUSEID
Person file	Vehicle file	HOUSEID
Person file	Travel day trip file	HOUSEID and PERSONID
Person file	Travel period file	HOUSEID and PERSONID
Vehicle file	Travel day trip file	HOUSEID
Travel day trip file	Segmented trip file	HOUSEID, PERSONID, and TRPNUM
Travel day trip file	Travel period file	HOUSEID and PERSONID

## **ID VARIABLES NOT ALWAYS SEQUENTIAL**

The ID variables within a file are not always sequential. There are a number of reasons for this, including the following:

- 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
- When a person took more than 15 travel day trips, the additional trips were numbered starting with 21 in numbering the person's trips (TRPNUM) and starting with 101 in numbering the household's trips (HHTRIPID) .

## **EXAMPLE OF A MERGE**

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, one 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 i.e., 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.

**WHICH  
WEIGHT TO  
USE**

After the files are successfully merged, the next question in using the merged file is which weighting factor to use. In our example, there is a weighting factor in the Person file and one in the Travel-day file. Chapter 3-G describes the calculations of the different weights in the NPTS. In essence, a weighting factor expands the sample data to a population from which the sample is selected. Thus, a household weight indicates the number of households with similar characteristics in the overall population that are represented by the sampled household.

For example, a household with a weight of 100 means that it represents itself and 99 other households of similar characteristics that were not sampled for the survey. This implies that these 99 households have travel patterns that are similar to those of the sampled household. One purpose of a sample design is to ensure that such similarity is maximized.

The rule in deciding which weight to use depends on the unit (e.g., households, persons, vehicles, or trips) on which the analysis is performed. For example, if an analysis is to be performed on a collection of trips, then the trip is the unit and the trip weight should be used. On the other hand, if an analysis is to be performed on a set of vehicles, then the vehicle is the unit and the vehicle weight should be used. In the above example, number of daily trips by telecommuting status, the main interest is on the trips, the individual trip is the unit and thus the trip weight is the appropriate factor.

Another way to explain this, using our example is:

**Distribution of Persons by Telecommuting Status and Number of Daily Trips - Hypothetical Data**

Tele-commute Status	0-4 daily trips	5-9 daily trips	10 or more daily trips	All
Sometime	45.9 %	38.9%	15.6%	100.0%
Never	56.7	33.7	9.6	100.0
Total	54.9	34.2	10.9	100.0

In this example, the row data on telecommuting frequency is from the Person file, and the column data, number of daily trips, is computed from the travel day file. The determining factor in which weight to apply is always "where does the cell data come from?". For this example, the cell data is percent of persons, which is from the person file, and the person weight, WTPERFIN, is the correct weight to apply.

## 5-H. SPECIAL USER NOTES

### DATA FILE CONVEN- TIONS

There are a number of conventions followed throughout the NPTS data files. These are also listed in **Appendix J**, Documentation Notes, and they include:

Yes/No questions - coded as 01 = yes and 02 = no.

Calendar Dates - separate variables were constructed for the month, day and year of reported dates.

Times - all reported time variables are in military time from 0000 to 2359.

Legitimate skip codes - questions intentionally skipped in the instrument were generally denoted by a field filled with 9's with a 4 in the last digit.

Don't know - responses of don't know or not ascertained were generally denoted by a field filled with 9's with an 8 in the last digit.

Refused - responses of refused were generally denoted by a field completely filled with 9's

Survey weights - there is one only one weight variable on each file. It is the weight that is appropriate for use in preparing tabulations of data from that file.

### ADDING TRAVEL DAY AND TRAVEL PERIOD DATA

Special procedures must be followed for adding the data from Travel Day and Travel Period. See Section 4-B for a description of the relationship between these two files.

If the respondent took a trip of 75 miles or more and returned

home on Travel Day, that trip will be collected in both the travel day and the travel period sections of the questionnaire. Note that, for travel period trips, it does not matter when the outgoing portion of the trip took place, the return trip must be made during the 14-day travel period. And the trip will be collected twice only for the travel that took place on the travel day.

Because of the difference in the definition of travel day and travel period trips, it is likely that the long-distance travel will be one trip on the travel period file, but will be counted as several trips on the travel day file. The variable, OVERLAP, will identify which travel day trips are part of the long trip reported in the travel period file.

To run a combined estimate, run the travel day file omitting the OVERLAP trips, and combine that result with all trips from the travel period file.

## **ESTIMATES OF VMT FROM THE 1995 NPTS**

There are multiple ways of computing vehicle miles of travel (VMT) from the 1995 NPTS. Which one is used for a specific analysis should depend on the nature of that analysis. For many data inquiries, more than one way would be appropriate. The intent of this subsection is to make the data users aware of the various ways VMT estimates can be made, which are:

- travel day
- travel day plus travel period
- travel day plus travel period plus commercial driving
- annual estimate of driver miles
- annual estimate of vehicle miles
- annualized estimate of odometer readings

FHWA will be conducting analysis of the differences in the estimates derived from each of these sources.