APPENDIX C: QUALITY OF THE DATA

APPENDIX C

QUALITY OF THE DATA

INTRODUCTION

This section discusses several issues relating to the quality of the National Household Travel Survey (NHTS) data and to the interpretation of conclusions based on these data. In particular, the focus of our discussion is on the quality of specific data items, such as the fuel economy and fuel type, that were imputed to the NHTS via a cold-decking imputation procedure. This imputation procedure used vehicle-level information from the NHTSA Corporate Average Fuel Economy files for model year's 1978 through 2001. It is nearly impossible to quantify directly the quality of this imputation procedure because NHTS does not collect the necessary fuel economy information for comparison. At best, we have indirect evidence on the quality of our imputations, which is addressed in the following sections. Indeed, such an imputation procedure could be vastly improved with the collection of Vehicle Identification Number (VIN), fuel type and retail fuel price for each sample vehicle. However, those collections may represent an unreasonable burden on NHTS respondents.

The quality of the data collection and the processing of the data affect the accuracy of estimates based on survey data. All the statistics published in this appendix, such as total vehicle-miles traveled (VMT), are estimates of population values. These estimates are based on observations from a randomly chosen subset of the entire population of occupied housing units. Consequently, the estimates always differ from the true population values. Because the NHTS is a sample survey, data from the survey are subject to various sources of nonsampling and sampling error.

Nonsampling error is a measure of variability due to the execution and processing of the survey. These errors can include: population undercoverage during sampling; questionnaire wording and format; response bias and variance; interviewer error; coding and/or keypunching error; and nonresponse bias. Nonsampling errors are treated in several sections of this appendix. The main section pertains to the imputation procedures used for "missing" fuel economy, fuel type, and fuel economy adjustments. In the previous sections, fuel economy adjustments were addressed. This section deals mainly with imputing fuel economy or $MPG_{i(EPA\ 55/45)}$ to each appropriate sample vehicle.

NONSAMPLING ERROR

Nonsampling errors are due to the conduct of the survey, and include both random errors and systematic errors or biases. The magnitudes of nonsampling biases cannot be estimated from the sample data. Thus, avoidance of systematic biases is a primary objective of all stages of survey design. Subsequent to conducting a survey, problems of unit nonresponse and item nonresponse need to be addressed.

In surveys with complex questionnaires and procedures, such as the NHTS, the final dataset reflects fundamental approaches taken in the data collection and editing processes. For the 2001 NHTS, two approaches may have had considerable impact on the resulting data.

The first is the reluctance to impute data. If the respondent did not answer an item, its value was generally not imputed, (i.e., determine what the logical response would be given the response to other items). Carefully performed imputation has its place in many statistical surveys, however Westat and U.S. DOT determined that imputation would be limited in the NHTS data. If data were imputed, an imputation/edit flag was set for the variable to indicate the values that were imputed. The treatment in the NHTS of these types of errors is discussed in 3-D.3. APPROACH TO POST INTERVIEW EDITING of the NHTS User's Guide.

Supplemental data, by definition, are 100 percent imputed. Thus, it is important that EIA thoroughly present the approach used to impute energy-related supplemental NHTS data (see Appendix B).

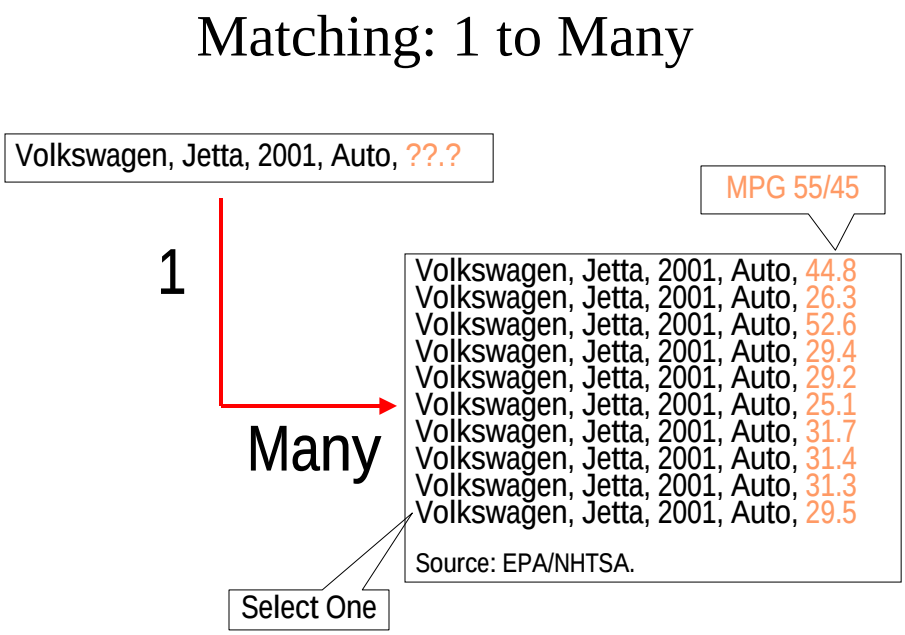
UNIT NONRESPONSE

Unit nonresponse is the type of nonresponse that occurs when no data are available for an entire sampled household. The respondent being unavailable or the respondent's refusal to cooperate causes most unit nonresponse cases. See the NHTS User's Guide, CHAPTER 4. SURVEY RESPONSE RATES, for further details on unit nonresponse.

IMPUTATION PROCEDURES FOR SUPPLEMENTAL DATA

Imputation procedures fill in the gaps of "missing" data. Item nonresponse occurs when the respondents do not know the answer or refuse to answer a question, or when an interviewer does not ask a question or does not record an answer. Or, as in the case of this appendix, item nonresponse occurs when a question was not asked, such that imputation procedures are required to address the need to append supplemental data to a pre-existing file from other external, but related, files. As already mentioned, NHTS took a conservative approach to item nonresponse. For supplemental data, in an effort to facilitate "full-sample" data analyses, imputations were made to provide the most probable responses when responses were "missing." For linking supplemental data, a pseudo cold-decking imputation was employed. Figure C1 depicts the cold-deck approach, using NHTS make, model, model year, and vehicle type information to "match" with eligible donors from the NHTSA CAFE files.

Figure C1. Schematic for Linking or Matching a NHTS Sample Vehicle to Eligible EPA/NHTSA Vehicles



Note: EPA – Environmental Protection Agency, NHTSA – National Highway Transportation Safety Administration.

COLD-DECK PROCEDURE

Because the fuel economy for a sampled vehicle could not be unequivocally determined by its NHTS-collected descriptors, a cold-deck imputation procedure was employed to "match" a NHTSA file record to a sample vehicle. A matching record was chosen from among the several applicable ones, with probability proportional to sales, using the sales figures on the NHTSA files. Once chosen, a record provided (1) EPA Composite MPG, (2) fuel metering, and (3) engine type. Although more attributes were available for selection, EIA limited its "donated" vehicle attributes to those required to assign an appropriate fuel price to a sample vehicle. This matching routine commonly resulted in a 1-to-many record linkage (see Figure C1 for an example).

Cold-deck procedures make use of a fixed set of values, which covers all of the perspective data items. These values can be constructed with the use of historical data, subject-matter expertise, or a combination of both. Such a procedure is an attempt to create a "perfect'" questionnaire in order to fill in the missing data gaps or, in this case, append supplemental data. If these procedures are completed properly and with limited bias, imputation has the ability to derive a complete and accurate record that (1) contains an audit trail for evaluation purposes; and (2) ensures that the imputed records are internally consistent.

Multiple paths were used to "match" recipient NHTS sample vehicles to eligible donor NHTSA file record vehicles. Because matching used a combination of four common linking variables – vehicle manufacturer, vehicle model, vehicle model year, and vehicle type – several

"matching" paths were followed. These paths are denoted (i.e., internally audited) with imputation flags, which are defined for each vehicle as follows:

- 10# denotes a NHTS sample vehicle that had a single model name "matching" to eligible NHTSA file records using four linking variables: vehicle manufacturer, vehicle model, vehicle model year, and vehicle type.

- 20# denotes a NHTS sample vehicle that had multiple model names "matching" to eligible NHTSA file records using four linking variables: vehicle manufacturer, vehicle model, vehicle model year, and vehicle type.

- 30# denotes a NHTS sample vehicle that had a single model name "matching" to eligible NHTSA file records using three linking variables: vehicle manufacturer, vehicle model, and vehicle model year.

- 40# denotes a NHTS sample vehicle that had multiple model names "matching" to eligible NHTSA file records using three linking variables: vehicle manufacturer, vehicle model, and vehicle model year.

- 50# denotes a NHTS sample vehicle that had a single model name "matching" to eligible NHTSA file records using three linking variables: vehicle manufacturer, vehicle type, and vehicle model year.

- 60# denotes a "match" based on EIA expert analysis using subject matter experience, in conjunction with past RTECS. Additionally, this imputation flag value represents recreational vehicles (VEHTYPE = "06"), where $MPG_{(EPA\ 55/45)}$ has been fixed at a yearly estimate based on the U.S. Department of Transportation, *National Transportation Statistics 2000*.[76]

- 001 denotes a NHTS sample vehicle that was internally hot-decked to match with its median Composite fuel economy value as defined by one or more vehicle characteristics, such as make, model, model year, and vehicle type. These flagged values become more meaningful with pre-1978 model year vehicles since NHTSA's CAFE exclude pre-1978 model years. EIA, therefore, recommends that users take extreme caution when making inferences concerning pre-1978 model year vehicles from this report.

- 999 denotes an imputation flag where no eligible NHTSA file records were found to "match" a NHTS sample vehicle.

In the above listing, # is a number between 0 and 5. This number, #, represents a year increment. Due to the errors in respondents reporting accurate model year or, to a lesser extent, due to deficiencies in the NHTSA files, it was necessary to incrementally increase or decrease (not simultaneously increase and decrease) the model year for "matching" to successively larger range of years. If, for example, an eligible match was not found for a NHTS sample vehicle

[76] Table 4-11. Passenger Car and Motorcycle Fuel Consumption and Travel, *National Transportation Statistics 2000* Bureau of Transportation Statistics, U.S. Department of Transportation. (Washington, DC).

having the following attributes: Volkswagen, Scirocco, 1990, Automobile. Toggling of model years, by a single year increase followed by a single year decrease of the reported model year, resulted in a match with a Volkswagen, Scirocco, 1988, Automobile. In this example, the Volkswagen, Scirocco, 1990, Automobile, while seemingly a respondent reporting error, would receive an imputation flag of "102" due to the "match" with the NHTSA file record corresponding to a Volkswagen, Scirocco, 1988, Automobile.

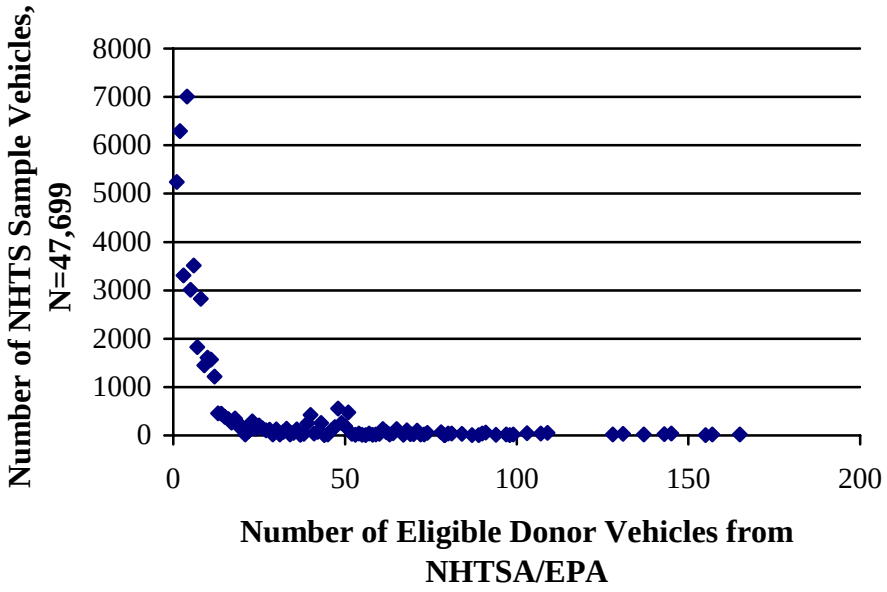
Table C1. Distribution of NHTS Sample Vehicles by Fuel Economy Imputation Flag, 2001

Imputation Flag for $MPG_{(EPA\ 55/45)}$	Number of Vehicles in NHTS Sample
001	2,347
100	3,122
101	103
102	39
103	28
104	23
105	44
200	31,407
201	1,428
202	397
203	272
204	172
205	63
300	33
301	4
302	1
303	4
400	582
401	35
402	54
403	28
404	10
405	5
500	1,907
501	38
502	33
503	20
504	8
505	19
600	510
Total	**42,736**

Source: U.S. Department of Transportation, Federal Highway Administration, *National Household Travel Survey 2001*, augmented release by the Energy Information Administration, (Washington, DC).

While the distribution of imputation flags is helpful, further evidence is needed to quantify the quality of this procedure. Figure C2 expands our coverage to include all national sample vehicles (of which "100-percent-reporting household" is just a subset) and charts the 1-to-many "matching" relationship for the 47,669 "matched" sample vehicles, or 53,278 less 3,696 (for 999 flag) less 1,856 (for 600 flag).[77]

Figure C2. Distribution of NHTS Sample Vehicles "Matched" with Vehicles "Donated" by NHTSA File Records



Source: U.S. Department of Transportation, Federal Highway Administration, *National Household Travel Survey 2001*, preliminary National and Add-on release, January 2004. (Washington, DC).

To make the "match" distribution display more revealing, values from the above figure are tabulated to present range categories of donor vehicles in Table C2.

Table C2. Distribution of All NHTS Sample Vehicles "Matched" by Range of Donor Vehicles

Range of Eligible Donor Vehicles	Number of Vehicles in NHTS Sample
1	5,238
2 to 5	19,617
6 to 10	11,225
11 to 20	5,399
21 to 30	1,273
31 to 40	1,148

[77] Estimates were drawn from the public-use file released by FHWA in January 2003. In the January 2004 public-use file, three vehicles were deleted, yielding 53,275 vehicles in the complete national vehicle sample, including 100-percent and 50-percent reporting households, as defined by NHTS.

Table C2. Distribution of All NHTS Sample Vehicles "Matched" by Range of Donor Vehicles

Range of Eligible Donor Vehicles	Number of Vehicles in NHTS Sample
41 to 50	1,672
51 or more	2,127
Total	47,699

Source: U.S. Department of Transportation, National Household Travel Survey 2001, preliminary National and Add-on release, Federal Highway Administration, January 2004. (Washington, DC).

QUALITY OF SPECIFIC SUPPLEMENTAL DATA ITEMS

COLD-DECK PROCEDURE: SENSITIVITY ANALYSIS

Although the accuracy and robustness of the cold-deck procedure and subsequent fuel economy adjustments are not quantifiable because we lack both fuel purchase and mileage diaries for calculating a vehicle's actual on-road, in-use fuel economy, we can assess the sensitivity of the cold-deck procedure in an effort to measure its robustness.

Because we use a single value imputation approach, multiple imputations is one approach available for investigating the uncertainty of our imputed values. Indeed, imputing a single value may result in estimating measures of precision (e.g., standard errors) that are too small because a single value ignores the uncertainty found in selecting from a listing of donated values. Rather than perform a series of multiple imputations, we have assumed that each sample vehicle's list of eligible donors represents a complete set of values for its "missing" fuel economy variable. Therefore, the uncertainty associated with the imputation procedure may be assessed by imputing a pre-determined subset of values; that is, ones that represent the extremes and average of eligible donors. P5 and P95 – the 5th and 95th percentiles of sales-weighted fuel economy, respectively – represent our extreme distribution values, while the average value corresponded to the sales-weighted average of the eligible donor vehicles. Using Figure 2 as an example, we calculate: P5 = 25.1, P95 = 44.8 and a sales-weighted average of 30.8 miles per gallon.

By separately totaling the consumption of transportation fuel for each of these 3 outcomes and, then, comparing them to our single-value total, it is not surprising that we find that

- applying sales-weighted fuel economy values yields a energy consumption total 2 percent less than the single-value total;
- applying 5th percentile values yields an energy consumption total 7 percent more than the single-value total; and,
- applying 95th percentile values yields an energy consumption total 9 percent less than the single-value total.

Clearly, applying extreme distribution values – P5 and P95 – to each and every eligible sample vehicle results in biased energy-related estimates. While these extreme values are not acceptable to a researcher, such biased estimates illustrate the upper and lower uncertainty bounds associated with cold-deck estimates. Given these bounds, along with survey sampling and non-sampling errors, the use and usefulness of an enhanced 2001 National Household Travel Survey should be evaluated against a researcher's project requirements.

VEHICLE FUEL PRICE AND EXPENDITURES

In the 2001 NHTS, fuel price data were not collected via fuel purchase diaries, compared to previous EIA studies (e.g., RTECS). Instead, fuel prices were determined from EIA price series. Unfortunately, there is no way to validate the price methodology used to assign a monthly price paid for transportation fuel because EIA lacks the necessary fuel purchase diaries from NHTS repondents.

The Bureau of Labor Statistics (BLS) *Retail Pump Average Gasoline Prices* and the Lundberg Survey, Inc. offer alternate price series. However, there was a general consistency with using a price series from one statistical agency.

GASOLINE EQUIVALENT GALLON

The following table provides the gasoline equivalent gallon conversion used in this appendix. All conversion values, to the extent possible, have been made to mirror the conversion values used in deriving equivalent-gallon fuel economy estimates found in the NHTSA CAFE files.

Table C3. Gasoline Equivalent Gallon Conversion Values

Transportation Fuel	Gasoline Equivalent Gallon
Diesel	1 diesel gallon = 1 gasoline equivalent gallon
Electricity	33,705 Watt-hours = 1 gasoline equivalent gallon
Compressed Natural Gas	121.5 cubic feet = 1 gasoline equivalent gallon

Sources: 40 CFR Parts 80, 85, 86, 88, and 600 and 10 CFR Part 474.

GREET MODEL

Of course, there are other conversion factors available, depending on the various fuel-specific factors. For the Greenhouse Gases, Regulated Emissions, and Energy Use in Transportation (GREET) model, the U.S. Department of Energy, Argonne National Laboratory uses the following:

Table C4. Lower and Higher Heating Values for Select Transportation Fuels Based on the GREET Model

Transportation Fuel	LHV (net) Btu per gallon	HHV (gross) Btu per gallon	Density Grams per gallon	Carbon Content (% by wt)	Sulfur Content (ppm by wt)
Conv.Gasoline	115,500	125,000	2,791	85.5%	200
Ref. Gasoline	112,265	121,456	2,795	82.9%	30
Diesel	128,500	138,700	3,240	87.0%	250
Methanol	57,000	65,000	2,996	37.5%	0
Ethanol	76,000	84,500	2,996	52.2%	0
LPG	84,000	91,300	2,000	82.0%	0
Natural gas	928	1,031	21	74.0%	7

Table C4. Lower and Higher Heating Values for Select Transportation Fuels Based on the GREET Model

Transportation Fuel	LHV (net) Btu per gallon	HHV (gross) Btu per gallon	Density Grams per gallon	Carbon Content (% by wt)	Sulfur Content (ppm by wt)
Electricity	3,412	Btu/kWh			

Source: M. Wang, GREET 1.5 -- *Transportation Fuel-Cycle Model*, Volume 1: Methodologies, Development, Use, and Results, Center for Transportation Research, Argonne National Laboratory, ANL/ESD-39, Vol.1, Aug. 1999. M. Wang, GREET 1.5 -- Transportation Fuel-Cycle Model, Volume 2: Appendices of Data and Results, Center for Transportation Research, Argonne National Laboratory, ANL/ESD-39, Vol.2, Aug. 1999. Notes: 1) Gasoline results are for the mix of 70% conventional gasoline and 30% reformulated gasoline. 2) LPG results are for the mix of 40% LPG produced from crude and 60% from natural gas. 3) Electricity results are for the current U.S. average electricity generation mix.

TRANSPORTATION ENERGY DATA BOOK: EDITION 22 — 2002

Likewise, the Energy Information Administration, U.S. Department of Energy (according to the latest *Transportation Energy Data Book*) applies the following approximate heat content for various fuels:

Table C5. Lower and Higher Heating Values for Various Transportation Fuels

Transportation Fuel	HHV (gross) equivalent to LHV (net)
Automotive gasoline	125,000 Btu/gal (gross) = 115,400 Btu/ gal(net)
Diesel motor fuel	138,700 Btu/gal (gross) = 128,700 Btu/gal (net)
Biodiesel	126,206 Btu/gal (gross) = 117,093 Btu/gal (net)
Methanol	64,600 Btu/gal (gross) = 56,560 Btu/gal (net)
Ethanol	84,600 Btu/gal (gross) = 75,670 Btu/gal (net)
Gasohol	120,900 Btu/gal (gross) = 112,417 Btu/gal (net)
Aviation gasoline	120,200 Btu/gal (gross) = 112,000 Btu/gal (net)
Propane	91,300 Btu/gal (gross) = 83,500 Btu/gal (net)
Butane	103,000 Btu/gal (gross) = 93,000 Btu/gal (net)
Jet fuel (naphtha)	127,500 Btu/gal (gross) = 118,700 Btu/gal (net)
Jet fuel (kerosene)	135,000 Btu/gal (gross) = 128,100 Btu/gal (net)
Lubricants	144,400 Btu/gal (gross) = 130,900 Btu/gal (net)
Waxes	131,800 Btu/gal (gross) = 120,200 Btu/gal (net)
Natural Gas	
Wet	1,109 Btu/ft^3
Dry	1,027 Btu/ft^3
Compressed	20,551 Btu/pound
	960 Btu/ft^3
Liquid	90,800 Btu/gal (gross) = 87,600 Btu/gal (net)

Table C5. Lower and Higher Heating Values for Various Transportation Fuels

Transportation Fuel	HHV (gross) equivalent to LHV (net)
Fuel Oils	
Residual	149,700 Btu/gal (gross) = 138,400 Btu/gal (net)
Distillate	138,700 Btu/gal (gross) = 131,800 Btu/gal (net)

Source: U.S. Department of Energy, Oakridge National Laboratory, Center for Transportation Analysis, *Transportation Energy Data Book Edition 22*, Washington, DC, 2002, Table B.1, ORNL-6967.

According to ORNL's latest *Data Book,*

> *The heat content of a fuel is the quantity of energy released by burning a unit amount of that fuel. However, this value is not absolute and can vary according to several factors. For example, empirical formulae for determining the heating value of liquid fuels depend on the fuels' American Petroleum Institute (API) gravity. The API gravity varies depending on the percent by weight of the chemical constituents and impurities in the fuel, both of which are affected by the combination of raw materials used to produce the fuel and by the type of manufacturing process. Temperature and climatic conditions are also factors.*
>
> *Because of these variations, the heating values in above table may differ from values in other publications. The figures in the Edition 22 report are representative or average values, not absolute ones. The gross heating values used here agree with those used by the Energy Information Administration.*
>
> *Heating values fall into two categories, gross and net. If the products of fuel combustion are cooled back to the initial fuel-air or fuel-oxidizer mixture temperature and the water formed during combustion is condensed, the energy released by the process is the higher (gross) heating value. If the products of combustion are cooled to the initial fuel-air temperature, but the water is considered to remain as a vapor, the energy released by the process is lower (net) heating value. Usually the difference between the gross and net heating values for fuels used in transportation is around 5 to 8 percent; however, it is important to be consistent in their use.*

EIA strongly encourages a consistent use of heating values.