EDITING AND IMPUTATION OF THE 1977 TRUCK INVENTORY AND USE SURVEY

Edwin L. Robison and W. Joel Richardson, Bureau of the Census

A. Introduction

The Truck Inventory and Use Survey is one of the surveys included in the 1977 Census of Transportation. The Census of Transportation has been conducted every five years since 1963. The Truck Inventory and Use Survey is a mail survey of truck registrations to obtain data on certain truck characteristics by state. Historical data were not sufficient to formulate editing and imputation procedures for some data items. The Statistical Package for the Social Sciences (SPSS) was used to analyze early survey returns and formulate procedures in a short amount of time.

The resulting bias in published estimates of annual miles and lifetime miles is analyzed in a tabular presentation and the possible bias in estimates generated from the public use tape is discussed. The bias is compared to sampling and nonsampling errors. The relationship between annual miles and lifetime miles is covered in detail.

B. Brief Description of the Survey

The 1977 Truck Inventory and Use Survey is a mail survey of 118,000 truck registrations, designed to obtain reliable data on certain truck characteristics in each state. A separate sample was selected in each state of "small" trucks (pickups, vans, multistops) and "large" trucks. This stratification was used to minimize the variance for some important characteristics. Data are collected concerning body type, vehicle weight, major use, area of operation, annual miles, lifetime miles, fuel type, and several other items. These data are published and are provided in more detail on a public use tape. Users of the data include the Department of Transportation, the Interstate Commerce Commission, state highway administrations, vehicle and parts manufacturers, tire and fuel companies, research and consulting firms, and leasing companies. Good data are sought in all states, with better reliability for states with larger truck populations. There is an overall emphasis upon large (heavy) trucks.

C. General Edit and Imputation Procedures

All responses are subject to a clerical and analytical screening process, and later to a computer edit and imputation. The screening and edit procedures are designed to detect errors and inconsistencies due to the respondent or the keyer. Where possible, analysts correct inconsistencies or errors and provide the correct responses for missing data items, phoning the respondent when necessary. Annual miles, lifetime miles, engine type, and gross vehicle weight are imputed during the computer edit if the data items are still missing.

Editing, imputation, correction and data insertion procedures were based on historical data when possible. Some procedures were very simple and had very low error rates, such as assuming gasoline engines to be in pickup trucks. Other data correction or insertion (imputation) procedures were much more complicated. To impute and correct for cubic inches and horsepower, information derived from the vehicle identification number (VIN) can be used in conjunction with several lists of manufacturers engines and engine characteristics.

For some data items such as annual miles and vehicle size class, historical data were not considered sufficient. The amount and type of imputation on available files was unknown and some important data items were unavailable. For annual miles, some new body types were added and the 1972 data did not identify vehicles used off-the-road. For vehicle size class, the 1972 file did not include the gross vehicle weight (GVW) code provided by the contractor, R. L. Polk and Company. Further, the effects on vehicle operation and purchase due to the recession and increasing petroleum cost were uncertain.

D. Effects of Data Insertion (Imputation) and Error Correction by Analysts

The correction of response and keying errors by analysts significantly reduces the bias and error for several data items. Analysis of the body type, make, and year led to the correction of many responses to the cubic inches and horsepower questions (the quality of response to these items was poor). Analysis of the vehicle identification number (VIN) often allowed the insertion of data for cab types, fuel types, and gross vehicle weights.

The corrections to keying and response errors in the annual miles and lifetime miles items can be of considerable magnitude. Errors for recently manufactured vehicles (model years 1975, 1976, and 1977) are relatively easy to detect and correct. These recently manufactured trucks have special reporting problems, but the bias introduced is near zero. Response and keying errors in older vehicles are much harder to identify positively and, consequently, only major errors are corrected.

E. Annual Miles/Lifetime Miles Nonsampling Errors

Several errors occurred in responses to the annual miles/lifetime miles questions. Among these are:

1. Reporting the same figure for both annual miles and lifetime miles when the truck is not exactly one year old.

2. Switching the answers. This is a problem mainly for 1975, 1976, and 1977 make trucks, especially those less than one year old.

3. "Keyer" type errors such as the introduction or omission of a zero, keying tenths of a mile as miles, or keying 17 instead of 17000. These errors can be caused by the respondent, the data screener, or the keyer.

4. "Respondent" type errors. One major error of this kind is the failure to report odometers turning over 100,000. Other errors occur because the respondent makes a bad approximation of his vehicle's age.

The effects of these errors on annual miles data are given in Attachment A for model years 1975, 1976, and 1977 (16% of all trucks in the 4 states analyzed). A summary is given here. The "high" column represents the amount that uncorrected errors on the high side would overestimate the true total annual miles for the model year. Similarly, the "low" column represents the amount that uncorrected data would underestimate the true total annual miles for that model year. Errors not affecting annual miles data (or affecting it very little) were counted as "not in error."

Model Year	Percent Trucks in Error	Percent High Errors	Percent Low Errors	Total Annual Miles (000)
1977	35	3	18	3,000
1976	25	3	8	14,000
1975	10	7	4	13,000

Nearly all errors in annual miles for these years were detected and corrected. Before model year 1975, most errors in annual miles cannot be identified reliably without referring to the original form or contacting the respondent. The gross errors of introducing an additional zero or keying tenths of a mile as miles would be consistently detected. A study at a future date is planned.

Lifetime miles is seriously underreported, especially for vehicles manufactured prior to 1971 where about 10% of the cases are underreported by 100,000 miles or more. Estimates of lifetime miles made from the public use tape would be understated by approximately 5%. The chief reason for this underreporting is that odometers generally turn over at 100,000 miles. Errors of this sort are usually not corrected unless the respondent is contacted. Some low lifetime miles figures are bona fide and would be improperly "corrected" in a computer process. Imputations of annual miles based on underreported lifetime miles would on the average be too low, but of the 7% of vehicles imputed this appeared to be a problem in less than 0.5% of the cases. This type of nonsampling error is of the same magnitude as relative standard errors for state totals on lifetime miles.

F. Preliminary Statistical Analysis for the Imputation of Miles

Early survey returns were analyzed to formulate procedures for the imputation of annual miles and lifetime miles. Approximately 16,000 trucks were available, many of which were discarded because of important missing data or absurd data relationships. Only simple tabulations and cross tabulations were used. Regressions were originally attempted but were dropped because of the brief time available and the presence of multinomial data items with no structured order (the major uses are coded 01 through 13 with no particular relationship to annual miles).

The ratio $\text{RAT} = \frac{\text{lifetime miles}}{\text{annual miles x age}}$ proved to be a stable quantity, with age of the vehicle (1978 - year of make) being the dominating variable. Many variables were tested individually and jointly for the effects on this ratio. These were body type, year of make (age), major use, fuel type, area of operation, and vehicle type. (The vehicle types are: 1 - Straight truck, 2 axle; 2 - Straight truck, 3 axle; 3 - Straight truck, other; 4 - Tractor truck, 2 axle; 5 - Tractor truck, 3 axle; 6 - Tractor truck, other; 7 - Pickup, van or multistop.)

Age of the vehicle was the dominant factor affecting the ratio. The denominator, annual miles x age, is an estimate"of sorts"of lifetime miles. As vehicles age they are driven fewer miles. Thus annual miles x age falls far short of lifetime miles. For recent vehicles the ratio is less than 1.0 because of part year ownership during the model year of the vehicle (example: using 1978 as the base year, most 1976 trucks are less than 2 years old). For these vehicles the month and year of acquisition should perhaps be used.

Most of the effects apparently caused by the other factors were almost entirely due to the ages of the vehicles in the cells examined. A portion of older vehicles used on the farm appeared to have lower ratios than other vehicles. Presumably this could be due to low yearly usage and extended useful life. These vehicles cannot be identified except on a case by case basis when both annual miles and lifetime miles are available. The same can be said for a portion of the off-the-road vehicles. In neither case did the data give clear justification for distinguishing these vehicles from the others.

G. Ratios for Use in Imputing Annual Miles and Lifetime Miles

If lifetime miles is available, the annual miles is imputed from the ratios given below. Similarly, if annual miles is available, lifetime miles may be imputed using the ratio. The imputation formula is $\text{RAT} = \frac{\text{lifetime miles}}{\text{annual miles x age}}$. Approximately 7% of the early returns had responses to only one of the items.

Model Year	Ratio	Model Year	Ratio
< 40	6.0	66	1.65
40 - 44	3.4	67	1.6
45 - 55	3.2	68	1.55
56	3.0	69	1.45
57	3.0	70	1.35
58	2.95	71	1.2
59	2.7	72	1.05
60	2.6	73	.95
61	2.25	74	.9
62	2.1	75	.8
63	2.0	76	.7
64	1.85	77	.7
65	1.8	78	.5
		Other	1.0

H. Imputation of Annual Miles Without the Ratio

Approximately 1% of the forms had no response to either annual miles or lifetime miles. For these vehicles, an annual miles entry is imputed from the ratio. There is no imputation for total nonrespondents. Each truck is assigned a base annual miles and is then subjected to four adjustments. For each body type there are two base annual mileages. One is for tractor trucks not including gasoline engine tractor trucks; the other is for all other vehicles of that body type. The adjustments are for: (1) vehicle type and fuel type (2) year model (3) major use, and (4) area of operation.

If the vehicle is not in use, the annual miles is changed to zero after lifetime miles have been imputed.

I. Bias Due to the Imputation of Annual Miles

The annual miles imputation procedure is examined in two ways: (1) Response data is compared to completely fabricated data (100% imputed data); and (2) The effect upon published estimates is given.

The imputation procedure outlined above underimputes individual trucks by an average of 15%. The impact on published data is a negative bias of about 1%, since only 7% of trucks have annual miles data imputed. A standard method of estimation, as discussed below, has positive bias of the same order of magnitude. The bias due to the imputation procedure is fairly stable from one data cell to another, as opposed to the standard method which is relatively volatile. The downward bias of 1% compares favorably to a three percent relative standard error (RSE), typically the smallest RSE in a state. For most data cells, the bias due to the imputation method can be assumed to be negligible. The underimputation is in the process of being remedied. Attachment B, table B1 compares aggregate reported data to imputed data. A 15% underimputation was the average (16% for small trucks, 14% for large trucks). The data file used was partially corrected. The four states on the file were Illinois, Minnesota, South Carolina, and New Mexico. Illinois, one of the first states processed, had a high error rate and had most of its trucks on the file originally used for formulating the imputation process. Minnesota and South Carolina had about half their trucks on the original file, and New Mexico had very few trucks on the original file.

The underimputation was stable among the body types and the major uses of the vehicles. The only sizeable exceptions were trucks from the large vehicle strata with a major use in agriculture, the under imputation for these being 2%.

The downward bias of 15% per imputed truck resulted mostly from the exclusion of "absurd" data relationships from the file originally used to develop the imputation procedure, and from the failure to exclude many of the errors on the file for new trucks (model years 1975, 1976, and 1977). In particular, some older vehicles recently acquired by new owners were excluded for having unusually high annual miles compared to lifetime miles. The "absurd" data was valid since the new owners had different truck use habits.

The imputation of annual miles for new trucks (model years 1973 and above) was less precise than the imputation for older trucks. The main reason is that 1978 was chosen as the base year for computing age, a 1976 model year truck being taken as 2 years old (the actual average was one year). The lack of precision can be remedied by calculating an age in months for vehicles which were bought new by the present owner.

Tables B2 and B3 compare an "ideal" estimate to the estimate using imputed annual miles data and another estimate using a standard item nonresponse adjustment technique. The "ideal" estimates are generated assuming that the item nonrespondents would have annual miles imputed 15% on the low side, and compensates for this assumption. The standard technique in effect imputes a cell average for missing data. The downward bias of the imputing technique for small trucks is 0.9% (0.8% for large trucks) and the upward bias of the standard technique for small trucks is 0.5% (1.4% for large trucks). The standard technique is biased since it is based on the assumption that nonrespondents in a cell resemble the respondents, which is not usually the case. Table B4 displays relative standard errors (RSE) for some estimates in Minnesota (the other states are very similar). The downward bias of 1% is dominated by the relative standard errors, even by the 3% RSE for the entire state. Only some of the estimates for the United States would have standard errors under 1%.

J. Sampling Frame Uncertainties

The sampling frame itself is the greatest cause of uncertainty in some states. Registration procedures of the states differ and each has its unique coverage problems. Some out-of-scope vehicles are sampled as trucks, and the responses are received. If these vehicles cannot be identified as being out-of-scope, they are tabulated as if they were in-scope. Some states have staggered registrations making the sampling frame a composite of time instead of a fixed point in time. This coverage problem can easily cause uncertainty exceeding the sampling errors. In some states the coverage problems are even greater due to inefficiencies of the state registration procedures. In such states this source of error and uncertainty completely dominates all others.

Preliminary counts of 2 years of registrations in Louisiana are given below. The 160,000 decline in the number of small trucks available for sampling shows massive problems of registration collection in the state. The total number of 1976 registrations is believed to be nearly ten percent short.

Louisiana Registrations Available for Sampling

	1976	1977
Small trucks	398,000 (88%)	238,000 (80%)
Large trucks	54,200 (12%)	58,000 (20%)
Total	452,000	296,000

K. Conclusion

The present annual miles imputation procedure causes a 1% negative bias in estimates. However, estimates in small data cells have less bias than the usual procedure of imputing the cell average. For most estimates, the bias is dominated by the relative standard errors.

The procedure will be improved before publication. An age computed on date of acquisition for recent vehicles will be used. The downward bias caused by filtering unedited data will be remedied by computing new ratios from an edited file.

Some nonsampling errors are sizeable. About 10% of all vehicles manufactured prior to 1971 have lifetime miles underreported by 100,000 miles or more, causing negative biases of about 5%. The coverage problems in several states create uncertainty exceeding any bias or sampling errors.

ACKNOWLEDGEMENT

Thomas L. Mesenbourg of the Bureau of the Census provided the necessary programming support for this paper. The Truck Inventory and Use Survey is conducted by the Transportation Branch, Economic Surveys Division, Bureau of the Census.

ATTACHMENT A

Nonsampling Errors for Model Years 1975, 1976, and 1977

A1 Some Nonsampling Errors - Model Year 1977

Error Type	Errors in Uncorrected File as Percent of Total (3,010)		Number of Trucks	Corrected Annual Miles (000)
	Understatements	Overstatements		
1) Same Figure	8%	--	24	
2) Switched	4%	--	12	
3) Keyer Type	4%	3%	7	
4) Respondent Type	2%	--	6	
All Errors	18%	3%	49	1,150
Not in Error			87	1,860
Total			136	3,010

A2 Some Nonsampling Errors - Model Year 1976

Error Type	Errors in Uncorrected File as Percent of Total (13,790)		Number of Trucks	Corrected Annual Miles (000)
	Understatements	Overstatements		
1) Same Figure	1%	1%	37	
2) Switched	2%	1%	25	
3) Keyer Type	4%	1%	67	
4) Respondent Type	1%	--	14	
All Errors	8%	3%	143	2,690
Not in Error			368	11,100
Total			511	13,790

A3 Some Nonsampling Errors - Model Year 1975

Error Type	Errors in Uncorrected File as Percent of Total (12,950)		Number of Trucks	Corrected Annual Miles (000)
	Understatements	Overstatements		
1) Same Figure	--	1%	6	
2) Switched	--	--	2	
3) Keyer Type	2%	1%	12	
4) Respondent Type	2%	5%	22	
All Errors	4%	7%	42	950
Not in Error			536	12,000
Total			578	12,950

ATTACHMENT B

Tables Comparing Reported and Imputed Data

B1 Reported Data Versus Fabricated (100% Imputed) Data

	Small Trucks			Large Trucks		
State	Reported	Imputed	Under Imputation	Reported	Imputed	Under Imputation
Illinois	10,453	8,551	22%	14,895	12,425	20%
Minnesota	11,116	9,695	15%	10,709	9,850	9%
South Carolina	10,377	9,221	13%	18,278	16,437	11%
New Mexico	10,275	9,052	14%	11,819	10,415	13%
All	10,573	9,141	16%	14,097	12,363	14%

B2 Comparison of Estimate Types - Small Trucks

	Partially Imputed			Standard Method		
State	Downward Bias	Estimate	"Ideal" Estimate	Estimate	Upward Bias	Trucks Imputed
Illinois	.7%	10,331	10,400	10,453	.5%	5.3%
Minnesota	.6%	11,021	11,085	11,116	.3%	4.7%
South Carolina	1.3%	10,127	10,256	10,377	1.1%	10.7%
New Mexico	1.1%	10,152	10,264	10,275	.1%	8.5%
All	.9%	10,419	10,511	10,573	.5%	7.3%

B3 Comparison of Estimate Types - Large Trucks

	Partially Imputed			Standard Method		
State	Downward Bias	Estimate	"Ideal" Estimate	Estimate	Upward Bias	Trucks Imputed
Illinois	.6%	14,657	14,742	14,895	1.0%	5.4%
Minnesota	.9%	10,603	10,695	10,709	.1%	6.7%
South Carolina	1.0%	17,756	17,941	18,278	1.9%	9.6%
New Mexico	.7%	11,293	11,377	11,819	3.9%	9.2%
All	.8%	13,790	13,900	14,097	1.4%	7.4%

B4 Some Estimates and Standard Errors in Minnesota

	Number of Sampled Trucks in Cell		Average Annual Miles	Relative
	Small	Big	(Rounded)	Standard Error
	83	758	6,600	10%
	3	27	12,200	35%
	32	98	15,100	15%
	3	4	10,400	60%
	90	20	11,800	10%
	2	312	8,600	20%
	0	11	13,000	55%
	0	56	13,600	20%
Whole State	534	1,380	10,400	3%