Developing a Best Estimate of Annual Vehicle Mileage for 2009 NHTS Vehicles

Prepared For

Federal Highway Administration
National Household Travel Survey

Prepared By
Oak Ridge National Laboratory
June 2011

Developing a Best Estimate of Annual Vehicle Mileage for 2009 NHTS Vehicles

1. Introduction

In the 2009 NHTS, the number of miles (VMT) driven by an NHTS household vehicle can be estimated in three different ways. First, one can use the single odometer reading to compute an estimate of annual mileage. Second, a designated household member was asked to report the total number of miles driven in each of the household vehicles (hereafter referred to as "self-reported VMT"). Finally, the amount of annual driving can be estimated based on the amount a vehicle is driven during the designated sample day (i.e., the travel day). Ideally, annualizing the odometer readings would probably generate the most reliable VMT estimate compared to estimates based on the other two approaches. Unfortunately, not all vehicles had an odometer reading recorded. Furthermore, of those that had their odometer reading recorded, the quality of some of the odometer readings is less than desirable. As such, ORNL was asked to estimate the number of miles driven by each of the NHTS vehicles based on the best available data. This estimate is hereafter referred to as the BESTMILE. BESTMILEs are computed only for automobiles, pickup trucks, vans, and sport utility vehicles. The value of the BESTMILE for motorcycles, other trucks, and recreational vehicles (RV) equals the value of the self-reported VMT for those vehicles with such information available. The BESTMILE estimates were developed using Version 2 of the 2009 NHTS data.

The preceding description of BESTMILE applies, for the most part, to both the 2001 and 2009 NHTS surveys. The major difference is that the 2001 NHTS collected two odometer readings while in the 2009 NHTS survey, only a single odometer reading was collected. A summary of relevant variables, and any differences found between the 2001 and 2009 surveys, is presented in Table 1 below. Given this data limitation, the challenge of using the single odometer reading was ultimately one of developing an estimate of annual vehicle mileage that effectively used the available information. At the

same time, since one would presume that the use of two odometer readings is better than one, the ideal method would be comparable to that developed in 2001.

Table 1. Comparison of Variables Relevant to Computing the Best Estimate of Annual Vehicle Mileage, 2001 and 2009 NHTS

Name	Description	2001	2009	Notes
Vehicle-related Variables				
OD_READ(1)	First/only odometer reading	✔	✔	
OD_DAY(1)	Day of odometer reading 1	✔	✔	
OD_MON(1)	Month of odometer reading 1	✔	✔	
OD_YEAR(1)	Year of odometer reading 1	✔	✔	
OD_READ2	Second odometer reading	✔	✘	No second reading in 2009
OD_DAY2	Day of odometer reading 2	✔	✘	No second reading in 2009
OD_MON2	Month of odometer reading 2	✔	✘	No second reading in 2009
OD_YEAR2	Year of odometer reading 2	✔	✘	No second reading in 2009
VEHYEAR	Vehicle model year	✔	✔	
VEHTYPE	Type of vehicle	✔	✔	
ANNMILES	Self-reported VMT per vehicle	✔	✔	
Household-related Variables				
MSASIZE	Size of MSA of Household	✔	✔	
CENSUS_D	Census division of Household	✔	✔	
LIF_CYC	Life cycle of Household	✔	✔	
HHSIZE	Number of persons in Household	✔	✔	
HHVEHCNT	Number of vehicles in Household	✔	✔	
Primary Driver-related Variables				
WHOMAIN	Primary driver of vehicle	✔	✔	
EDUC	Level of education of driver	✔	*	Categories of this variable have been collapsed in 2009.
R_AGE	Age of primary driver	✔	✔	
WORKER	Worker status of primary driver	✔	✔	
R_SEX	Gender of primary driver	✔	✔	

Ultimately the process of estimating BESTMILE for the 2009 vehicles followed much of what was done for the 2001 survey. First, an initial overview of data quality was performed (see Section 2). This process involved assessing the number of sample vehicles that had necessary components of BESTMILE estimation, such as an odometer reading, vehicle year, and information on the primary driver. Next, investigation of how to best use the single odometer reading was performed (Section 3). Once that was

accomplished, the actual calculation of BESTMILE was done (Section 4). This calculation step involved iteration – an estimation method was attempted and subjected to a validation step. The validation step, in the absence of a known, reliable estimate of vehicle miles per vehicle, involved simulation of the method using 2001 vehicle data, comparing results of the new method versus BESTMILE estimates produced for the 2001 dataset. In cases where such validation showed inadequate results, alternate methods were attempted until adequate results were obtained. The estimates were then adjusted to fit a precise time frame - April 1, 2008 to March 31, 2009 (Section 5). Finally, the BESMTILE estimates were screened for outliers and flagged or adjusted where appropriate (Section 6).

2. Data Quality

An initial analysis of 2009 NHTS vehicle data quality and availability was performed. Presence of the single odometer reading, combined with data on the vehicle year, was the primary basis for 2009 quality checks. Other items needed for computation of the BESTMILE variable included primary driver of the vehicle, specific vehicle types[1], and vehicle year. Table 2 summarizes the 2009 data.

[1] The out of scope vehicle types included "other trucks," "recreational vehicles," "light electric veh (golf cart)," and vehicles with missing vehicle type information.

Table 2. 2009 NHTS Vehicle Data Quality Checks

Data Quality Checks	Sample Vehicles	%
Total 2009 NHTS Vehicles	**309,163**	**100.0%**
No Odometer Reading	77,469	25.1%
No Vehicle Year	5,412	1.8%
No primary driver associated with the vehicle	21,257	6.9%
Out of Scope Vehicle Types	7,559	2.4%
Vehicles without Data necessary for eventual BESTMILE estimation[2]	12	0.0%
Vehicles with Usable Odometer Data	**197,454**	**63.9%**
Vehicles with Presumed Odometer Rollovers[3]	4,393	1.4%

The percentage of vehicles with BESTMILE based on odometer data (63.9%) was higher than in the 2001 NHTS (47.5%). This reflects the fact the only one odometer reading was taken in the 2009 data, with two required for a vehicle to be considered usable in the 2001 data. Table 3 summarizes the distribution of 2009 NHTS vehicles in terms of the key pieces of data. The structure of this table was the foundation for the differing ways in which BESTMILE was computed for the 2001 NHTS, and shaped computations in the same way for the 2009 NHTS vehicles.

[2] This includes specific variables used in various regression models. For example, a vehicle may have primary driver information, but not have a value for a specific variable, such as EDUC (Education of the driver). Some of this was accounted for in the 2001 models; however, some variables may have specific values in 2009 that are not present in 2001.

[3] If a vehicle was at least 20 years old and the odometer reading was less than 100,000, analysis was performed regarding a possible unrecorded odometer rollover. If adding 100,000 or 200,000 miles to the odometer reading resulted in an average miles per year of less than the 75th percentile of miles per year for vehicles, by age group, for those vehicles at least 20 years old with more than 100,000 miles, then the additional 100,000 or 200,000 miles were added to the odometer reading. The 75th percentile cutoffs were 10,000 miles per year for 20-24 year old vehicles, 7,500 miles for 25-29 year old vehicles, 6,000 miles for 30-39 year old vehicles, and 4,000 miles for vehicles 40 years and older.

Table 3. NHTS Vehicles[4] by Data Required for BESTMILE Estimation

	Usable Data to Estimate Odometer-Based BESTMILE					
	Yes		*No*			
	Usable Self-Reported VMT		**Usable Self-Reported VMT**			
	Yes	*No*	*Yes*		*No*	
	Information on Primary Driver?		**Information on Primary Driver?**		**Information on Primary Driver?**	
	Yes		*Yes*	*No*	*Yes*	*No*
One driver/One vehicle HHs	23,312	651	5,940	62	664	555
Two drivers/two vehicles HHs	71,172	1,915	15,898	9,900	1,089	2,239
Other Drivers=Vehicles HHs	17,275	648	4,949	2,708	448	783
Drivers > Vehicles HHs	10,668	438	2,929	2,823	352	717
Drivers < Vehicles HHs	69,403	1,972	22,984	14,638	1,821	5,456
Subtotal	191,830	5,624	52,700	30,131	4,374	9,750
Subtotal by Usable Data	***197,454***		***96,955***			

[4] There were 309,163 vehicles included in the 2009 NHTS survey. However, 14,754 of these vehicles were out of scope for the BESTMILE estimate. The out of scope vehicle types included "other trucks," "recreational vehicles," "light electric veh (golf cart)," and vehicles with missing vehicle type information. BESTMILE for these vehicles was set to the self-estimated annual miles driven, where available.

3. Initial Determination of An Annualized Odometer Estimate (ODOMMILES)

Investigation into how to use a single odometer reading in place of two odometer readings was conducted. As an initial step, 2001 NHTS national sample vehicle data was examined. Average 2001 self-reported mileage shows a slight decline for each year that a vehicle is owned, indicating that vehicle age should play an important role in the process. A new/used question (such as "Did you purchase this vehicle new or used?") was not asked in either 2001 or 2009; however, for purposes of this analysis a vehicle was considered purchased "used" if it was 2 or more years older (as determined through the vehicle model year) than the amount of time it was owned by the household. Similar analysis was conducted on 2009 NHTS vehicle data.

Given data on self-reported miles driven by new/used status and vehicle age, three regressions (one for new vehicles, one for used, and one for all vehicles – for use on vehicles where new/used status is unknown) were run to determine the relationship between vehicle age and annual miles driven. These three regressions, calculated separately but taking the same form, are summarized by Equation (1)[5]:

$$\text{Self - Reported Annual Miles} = \alpha + \beta_1(\textit{VehicleAge}) + \beta_2(\textit{VehicleAge})^2 \qquad (1)$$

Predicted values for each regression were computed for each vehicle age, which in the 2001 NHTS data ranges from 1 to 40. The predicted values by age are summarized in Figure 1.

[5] Note that regressions for 2001 and 2009, while taking the same form, were computed separately, leading to slightly different parameter estimates between surveys. Admittedly, for both 2001 and 2009, the R-squared values of all models are low (in the .04-.07 range). However, all model terms and the models themselves are statistically significant, and given the large amount of variation among vehicles in both surveys, one would expect R-squared values to be somewhat low.

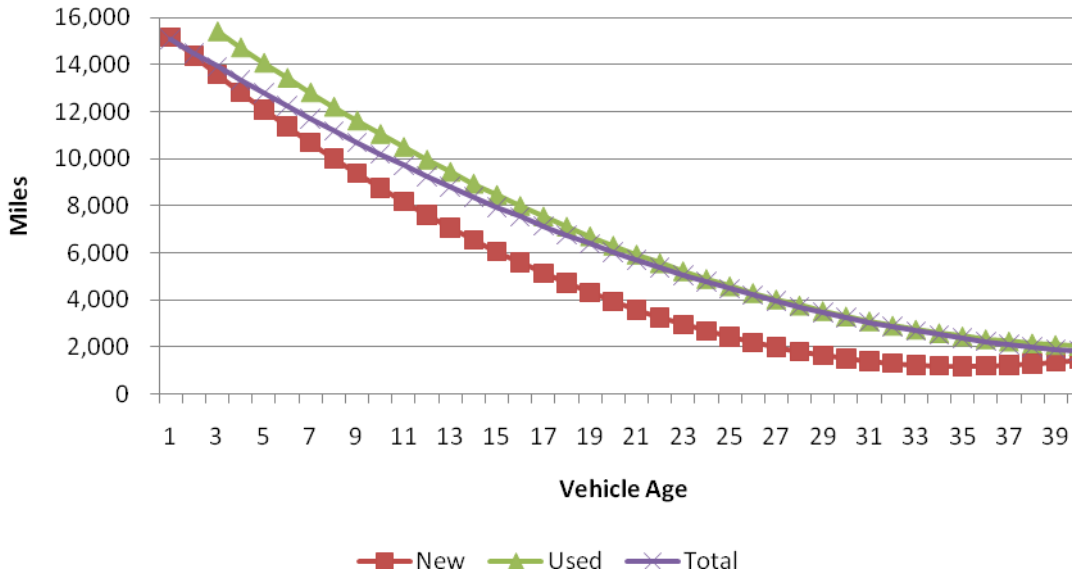

Figure 1. Average Self-Reported Miles (Smoothed via Regression Modeling) by Vehicle Age and New/Used Status, 2001 NHTS National Sample Vehicles

For each vehicle these predicted values were used to determine the percentage of travel that a given vehicle took in the most recent year, given the vehicle age and its subsequent cumulative mileage. Equation 2 summarizes the percentage of the single odometer reading attributed to the current year mileage for new vehicles[6]:

$$\text{New Mileage Percent}_i = \frac{\text{Estimated Self Reported Miles}_t}{\sum_{i=1}^{t} \text{Estimated Self Reported Miles}_i} \text{ x } 100\% \qquad (2)$$

where t is the vehicle age, and the numbers for Estimated Self Reported Miles are estimated using the regression for new vehicles in Equation 1. This percentage is then multiplied by the odometer reading in order to compute the estimated annual mileage (*ODOMMILES*) in the most recent year.

For a more concrete example, assume that we want to determine the miles driven for a new vehicle with an age of 5 and an odometer reading of 75,000. The table below shows the first step in the calculation:

[6] This method is also used for vehicles with an unknown new/used status, although the parameter estimates for these vehicles were different from those for new vehicles.

Table 4. Example Computation of Percent Mileage by Vehicle Year for a New Vehicle

Vehicle Year	Annual Miles	Cumulative Miles	Percent of Total
1	15,163	15,163	22.3%
2	14,356	29,520	21.1%
3	13,573	43,093	20.0%
4	12,815	55,908	18.8%
5	12,080	67,987	17.8%

Numbers in the Annual Miles column represent the predicted values from the model computed using Equation (1). Percents for all years are computed using the Cumulative Miles for the last year as a denominator. Since the vehicle is 5 years old, the Year 5 percent of 17.8% is multiplied by 75,000 to obtain the initial estimate for odometer miles (13,326 miles).

Used vehicles present a slightly more complex calculation. The first owner originally purchased the vehicle new, so for the period before the household respondent owned the vehicle, the mileage figures are estimated from the new vehicle regression. At the point at which the current owner (the household respondent) took ownership of the vehicle, the used regression is utilized to generate mileage figures[7]. Equation 3 below summarizes the percentage of the single odometer reading assumed to be the current year mileage for used vehicles:

$$\text{Used Mileage Percent}_i = \frac{\text{Used Vehicle Miles}_t}{\sum_{i=1}^{s-1} \text{New Vehicle Miles}_i + \sum_{i=s}^{t} \text{Used Vehicle Miles}_i} \text{ x } 100\% \quad (3)$$

where s is the vehicle age minus the number of years the household has owned the vehicle (more simply, the vehicle age at which the household obtained the vehicle), t is the vehicle age, New Vehicle Miles numbers are estimated using the regression for new vehicles in Equation 1, and Used Vehicle Miles numbers are estimated using the regression for used vehicles in Equation 1.

[7] Lack of data precludes adjustments for vehicles with more than one owner before the survey respondent. For purposes of this analysis, a single previous owner is assumed for vehicles determined to be "used."

To modify the previous example, assume that a 5 year-old vehicle with an odometer reading of 75,000 miles has been owned by the household for 2 years. To illustrate the mileages used for each year in terms of Figure 1, the figure below shows which estimates are used for each year the vehicle was in use:

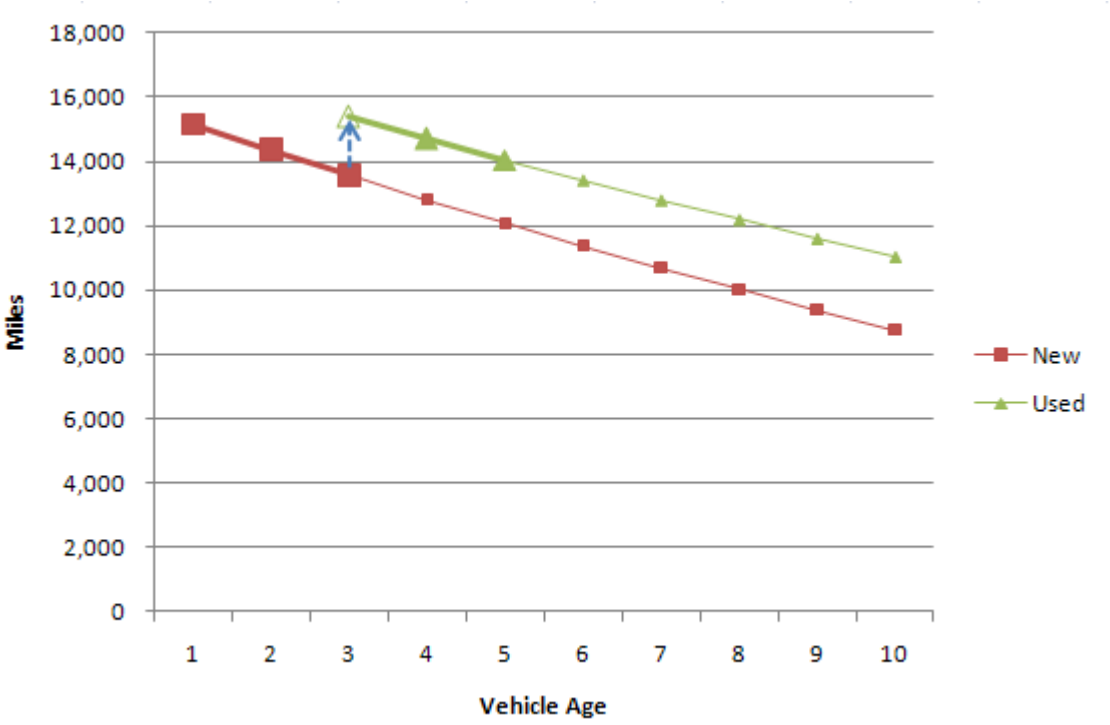

Figure 2. 5 Year-Old Used Car Example of Average Self-Reported Miles (Smoothed via Regression Modeling) by Vehicle Age and New/Used Status, 2001 NHTS National Sample Vehicles

As described in Equation (3), the first three years use the new vehicle mileage, while the next two shift to the used averages. These are then used to calculate the percentage of mileage driven in the most recent year. Table 5 shows the first step in this calculation.

Table 5. Example Computation of Percent Mileage by Vehicle Year for a Used Vehicle

Owner	Vehicle Year	Annual Miles	Cumulative Miles	Percent of Total
1 (presumably non-NHTS)	1	15,163	15,163	21.0%
	2	14,356	29,520	20.0%
	3	13,573	43,093	18.9%
2 (NHTS respondent)	4	14,719	57,812	20.5%
	5	14,062	71,874	19.6%

Numbers in the Annual Miles column for Owner 1 are predicted values from the New Car model computed using Equation (1), and from the Used Car model for Owner 2. Again, since the vehicle is 5 years old, the Year 5 percent of 19.6% is multiplied by 75,000 to obtain the initial estimate for odometer miles (14,674 miles). According to this calculation, the annual miles increase when ownership of the car is transferred and the used car, given the same mileage, was driven more in the most recent year. Intuitively this makes sense. If a person sells a car, that car may be more likely to be either in disrepair or underutilized. A person purchasing a used car, however, will tend to treat that car as if it were new, which it is from their usage perspective.

In 2001 a key component of calculating BESTMILE was the use of a crude daily estimated odometer mileage, taking the difference in the two odometer readings and dividing that by the difference in the dates of when those readings were taken. The calculation of ODOMMILES should be seen as an approximation of this crude method. The ODOMMILES calculation is subject to assumptions in driving patterns – mainly that driving of a given vehicle declines over time - that may lead to bias in the estimates. Thus, ODOMMILES is merely used as a piece in the BESTMILE estimation process, and not an end in itself.

4. Calculation of BESTMILE for Vehicles in the 2009 NHTS

As with the 2001 BESTMILE, estimation of 2009 BESTMILE utilized six different approaches, depending on which data was available for each vehicle. A seventh approach involved merely assigning self-estimated miles to vehicles of out-of-scope types, where no other information was present. Odometer readings are a key part of

Approaches 1 and 4 (detailed later in this section), and the estimate from the previous section (ODOMMILES) was integrated into the BESTMILE methodology for 2009.

Ideally, similar methodology to that used in creating BESTMILE for the 2001 NHTS vehicles would be used for the 2009 BESTMILE estimates in order to ensure comparability of estimates. In order to measure just how compatible 2009 estimates using this new methodology would be, the method was first simulated using 2001 NHTS vehicles. New 2001 BESTMILE estimates were then compared with the original 2001 estimates as a validation step. In cases where the new methods produced results that differed greatly, other alternatives were investigated (and detailed for each approach later in this section).

Approach 1. *For vehicles with a usable odometer reading, self-reported VMT, and information on the primary driver.*

Estimation

There were 191,830 vehicles in this category (Table 3). This approach assumes that the daily driving of a vehicle is a function of:

- the daily driving based on self-reported VMT,
- characteristics of the primary drivers, and
- other household characteristics and geographical attributes.

In the 2001 computation[8], the annualized estimate was computed using Equation (4):

$$Y = X\beta + R, \qquad (4)$$

[8] More fully described in the *2001 NHTS User's Guide, Appendix J.*

where *Y* was the difference in the two odometer readings divided by the difference in the dates of those readings (essentially a crude daily estimated mileage), *X* is a vector of independent variables, β is the matrix of model parameter estimates, and *R* is the vector of residuals containing the differences between the observed crude daily mileage and the estimates daily mileage. The vector of independent variables, *X*, included annual self-reported VMT (*ANNMILES*), education level (*EDUC*), age class of the primary driver (*R_AGEC*), vehicle age class (*VEHAGEC*), vehicle type (*VEHTYPE*), area size (*MSASIZE*), Census division (*CENSUS_D*), life cycle of the household (*LIF_CYC*), worker status and gender of the primary driver (*WORKER* and *R_SEX*, respectively), and size of the household (*HHSIZE*). The model for the case with an unequal number of drivers and vehicles also used a categorical variable for the driver to vehicle ratio (*DRVEH*).

In order to approximate the data available in 2009, this model substituted *ODOMMILES* (as computed in Section 3) as the dependent variable *Y* in Equation (4). This differs slightly from the 2001 method in that the dependent variable for 2001 was daily rather than annual miles. However, such an adjustment would merely affect parameter estimates but have no effect on predicted values for each vehicle; thus, ODOMMILES was left in annual terms and not divided by 365. In addition, the independent variable *EDUC* was modified to match those levels provided in 2009. If one odometer reading is truly enough to provide an adequate estimate of annual mileage, one would expect similarities in the results when compared to actual 2001 BESTMILE estimates. In addition to demonstrating the similarities of the approaches, such consistency would be desirable for comparison purposes by data users.

Two methods of comparison between the model estimates and the 2001 *BESTMILE* were devised. First, the standard error of the 2001 estimate, available in the *ANULZDSE* variable, was compared to the difference between the new model estimate and the *BESTMILE* value. This difference was classified in terms of the number of standard errors that the new estimate was different from *BESTMILE*. The second method involved the distribution of the percentage difference between the two estimates for each

vehicle, such that one can see, for example, that 25%-50%-75% of new model estimates are no more than a respective percentage away from the *BESTMILE* estimate.

The comparison of results is described in Tables 6a and 6b. Ultimately, using *ODOMMILES* as the dependent variable was extremely poor. Less than 30% of vehicles had estimates within two standard errors of the original *BESTMILE*, with nearly 40% of vehicles having estimates that differed by a factor of 5 or more standard errors. In other terms, with a median difference of 21%, half of all vehicles had differences with the original *BESTMILE* of 21% *or more***.**

Table 6a. Distribution of 2001 NHTS Vehicles by Differences between Estimate using *ODOMMILES* as the Dependent Variable and BESTMILE Value, in terms of standard errors of BESTMILE estimates, Approach 1

	% of Vehicles
0-1 StdErrs	15.00%
1-2 StdErrs	13.57%
2-3 StdErrs	12.19%
3-4 StdErrs	11.17%
4-5 StdErrs	9.63%
5+ StdErrs	38.44%

Table 6b. Distribution of the Percentage Differences between Estimate using *ODOMMILES* as the Dependent Variable and BESTMILE Value, 2001 NHTS, Approach 1

Percentile of Vehicles	% Difference from Original BESTMILE
100% Max	194977%
99%	1256%
95%	189%
90%	98%
75% Q3	44%
50% Median	**21%**
25% Q1	9%
10%	3%
5%	2%
1%	0%
0% Min	0%

Alternate models were tried, using a single model instead of multiple models based on the driver-vehicle relationship, for instance. The inclusion of new independent

variables, such as whether a vehicle was new or used, or purchased in the last year, was also investigated. Models using the self-reported VMT for each vehicle (ANNMILES) as the dependent variable were also tried. None of these attempts produced anything other than marginally different results from those in Tables 6a and 6b.

Ultimately, the solution that would best ensure compatibility between results used the existing BESTMILE as the dependent variable, while the right hand side of the equation was populated with independent variables that would be available in the 2009 data, including the ODOMMILES measure computed in Section 3. Using BESTMILE as the dependent variable on first glance may raise some concerns; however, one should note that such a model is not used to make inferences on the statistical validity of relationships between dependent and independent variables. The model is merely being used as an algorithm that relates the 2001 BESTMILE to a set of variables in the 2001 NHTS.

The other obvious problem with using BESTMILE as a dependent variable is that no BESTMILE estimates exist for 2009 data. Thus, the models using 2001 data were "transferred" to the 2009 data in order to create such estimates. In other words, these models were developed using 2001 data, then applied to the 2009 data to produce estimates.

Similar to what was done in the 2001 computations, models were estimated separately for three different types of households, as classified by the driver to vehicle relationship. These types consist of (1) households with one vehicle and one driver, (2) multi-driver households with an equal number of vehicles and drivers, and (3) households with unequal numbers of vehicles and drivers. The models are represented in Equation (4) shown earlier, where Y is the vector of *BESTMILE* estimates from 2001, X is the vector of independent variables, β is the matrix of model parameter estimates, and R is the vector of residuals. The vector of independent variables, X, includes the initial annualized odometer estimate based on the first odometer reading as described in Section

3 (*ODOMMILES*)[9], as well as the other independent variables detailed in the model with ODOMMILES as the *dependent* variable.

Use of a term in all models to account for year-to-year variation was investigated. *Highway Statistics* shows a 0.4% overall decrease in annual miles driven per passenger car between 2001 and 2008 (the latest year for which data is available), with slightly larger decreases in other types of vehicles (Table 7). Overall, annual miles for passenger cars and other 2-axle, 4-tire passenger vehicles fell 1.4% between 2001 and 2008. The self-reported annual miles estimates also dropped between the 2001 and 2009 NHTS surveys, for a much larger overall decline of 8.9%, with drops between 6% and 15% depending on vehicle type (Table 7). Since the declining self-reported mileage is a component in the modeling process, such information will ultimately influence the final estimates in a downward fashion, thus eliminating any need for a year-to-year term.

Table 7. Comparison of 2001 and 2009 Average Miles per Vehicle, *Highway Statistics* and NHTS

	2001	2009*	% diff
Highway Statistics			
Passenger Cars	11,831	11,788	-0.4%
Other 2-Axle, 4-Tire Vehicles	11,204	10,951	-2.3%
Passenger Cars & Other 2-Axle, 4-Tire Vehicles	11,593	11,432	-1.4%
NHTS ANNMILES (Self-Reported Mileage)			
Automobile/car/station wagon	10,695	10,054	-6.0%
Van (mini, cargo, passenger)	12,717	11,030	-13.3%
Sports utility vehicle	12,722	11,584	-8.9%
Pickup truck	11,729	9,891	-15.7%
All	11,078	10,088	-8.9%

* The most recent data for Highway Statistics is for the year 2008. Data can be found at http://www.fhwa.dot.gov/policy/ohpi/hss/hsspubs.cfm.

Comparison of 2001 Method vs. 2009 Method Using 2001 Vehicles

[9] Models with the single unadjusted odometer reading, as well as population density data and data from the travel day were also investigated, but were found to be less adequate in estimating 2001 BESTMILE. Note that use of ODOMMILES differs between what was used to calculate BESTMILE for the 2001 dataset. However, ODOMMILES was found to be the best bridge between the 2001 approach and the data available in 2009.

The model approach with BESTMILE as the dependent variable produced a close estimate of the 2001 *BESTMILE*. In order to compare just how close, the 2001 data where Approach 1 was used in the original *BESTMILE* computation was split into two equal groups. The model for the 2009 method was fitted to the data in the first group, and then estimates were computed for the vehicles in the second group. These estimates were then compared to the *BESTMILE* variable (computed using the 2001 method) in the 2001 NHTS dataset. This process was then reversed, with the model fitted to the second group of vehicles, and estimates using that model produced for the first group.

For the first group of data, 86.9% of all vehicles had estimates based on the new models that were within 2 standard errors of the *BESTMILE* estimate found in the 2001 NHTS dataset, with 96.2% within three standard errors (Table 8a). For the second group of data, these numbers were 86.9% within two standard errors and 96.4% within three standard errors (Table 8a). In terms of percentage differences, 50% of vehicles had new estimates that had a difference of 4.5% or less when compared to the *BESTMILE* estimate for both groups of data (Table 8b). 75% of all vehicles had differences of 10% or less for the both groups of vehicles (Table 8b).

Table 8a. Distribution of 2001 NHTS Vehicles by Differences between Estimate using One Odometer Reading and BESTMILE Value, in terms of standard errors of BESTMILE estimates, Approach 1

	Group 1	Group 2
0-1 StdErrs	57.16%	57.25%
1-2 StdErrs	29.73%	29.66%
2-3 StdErrs	9.31%	9.48%
3-4 StdErrs	2.16%	2.39%
4-5 StdErrs	0.70%	0.64%
5+ StdErrs	0.94%	0.58%
% 0-2 SEs	**86.89%**	**86.91%**
% 0-3 SEs	**96.20%**	**96.39%**

Table 8b. Distribution of the Percentage Differences between Estimate using One Odometer Reading and BESTMILE Value, 2001 NHTS, Approach 1

	Group 1	Group 2
100% Max	207891%	14558%
99%	403%	408%
95%	47%	47%
90%	24%	24%
75% Q3	10%	10%
50% Median	**4.5%**	**4.6%**
25% Q1	2%	2%
10%	1%	1%
5%	0%	0%
1%	0%	0%
0% Min	0%	0%

Given that the estimates from the new modeling scheme so closely match those of the *BESTMILE* variable on the 2001 dataset, this modeling approach, using both groups of data above, was used to compute 2009 *BESTMILE* estimates. Note that the initial annualized odometer estimate (*ODOMMILES*) for 2009 vehicles was computed based on Equations (2) and (3), which were re-calculated using 2009 data.

Residuals

In estimating 2001 *BESTMILE*, the residual from Equation (4) was retained since the goal was to create annualized estimates, as opposed to predictions completely free from random noise. Based on the assumption that the residuals from these new models based on 2001 data would be similar in distribution to residuals for 2009 data (assuming 2009 data could be used to create such as model), the residuals for vehicles from these new models were randomly assigned to the 2009 NHTS vehicles (referred to hereafter as "pseudo-residuals")[10].

If, after adding the pseudo-residual, the estimated $\hat{y}$ was less than 0 or greater than 200,000 miles per year[11], then a second randomly assigned residual was used. In this process for the 2001 *BESTMILE* computation, a third randomly assigned residual was used if the second residual also resulted in a $\hat{y}$ less than 0 or greater than 200,000 miles

[10] All sampling was done with replacement.

[11] Cutting off mileage at 200,000 miles per year has been standard in the NHTS/NPTS series. This amounts to approximately 550 miles per day, which is a practical maximum for a single driver.

per year[12]. However, after this point, if ŷ was still outside this range, then *BESTMILE* was set at 0 or 200,000. The percentage of total values in 2001 set to 0 or 200,000 after pseudo-residual assignment was approximately 0.2-0.5% depending on the modeling approach used. A comparable percentage in the 2009 ŷ estimates was obtained only when using an additional fourth residual, when needed. Thus, for Approach 1 and all other approaches in 2009, a fourth pseudo-residual was used in cases where necessary.

Approach 2. *For vehicles with self-reported VMT, and information on the primary driver, but without a usable odometer reading.*

Estimation

In the 2001 calculation of *BESTMILE*, the equivalent to Equation (4) was used to estimate vehicles with self-reported VMT and information on the primary driver but without usable odometer readings. In terms of estimation of 2009 *BESTMILE*, this subset of vehicles can be calculated using Equation (4), excluding the annualized single odometer reading term (*ODOMMILES*). The same setup was used as in Approach 1, with an initial model fitted using 2001 NHTS vehicles in two groups. As with Approach 1, pseudo-residuals were assigned, with the process repeated if the resulting ŷ was below 0 or above 200,000 annual miles per vehicle.

Comparison of 2001 Method vs. 2009 Method Using 2001 Vehicles

The results of calculations for these models for 2001 vehicles using the 2009 approach are in Table 9a and 9b, and are comparable to, and even slightly better than, the results in Approach 1.

Table 9a. Distribution of Vehicles by Differences between Estimate using One Odometer Reading and BESTMILE Value, in terms of standard errors of BESTMILE estimates, Approach 2

	Group 1	Group 2
0-1 StdErrs	65.07%	63.63%
1-2 StdErrs	28.70%	29.04%

[12] Note that if the sole purpose was to find a residual that led to an estimate within 0 to 200,000, a more efficient method could have been chosen. However, the main point was to assure that assignment of residuals was random in nature.

2-3 StdErrs	5.56%	6.28%
3-4 StdErrs	0.63%	0.97%
4-5 StdErrs	0.03%	0.08%
5+ StdErrs	0.01%	
% 0-2 SEs	**93.77%**	**92.67%**
% 0-3 SEs	**99.33%**	**98.95%**

Table 9b. Distribution of the Percentage Differences between Estimate using One Odometer Reading and BESTMILE Value, Approach 2

	Group 1	Group 2
100% Max	158033%	34617%
99%	410%	414%
95%	40%	42%
90%	19%	21%
75% Q3	8%	9%
50% Median	**3.8%**	**4.0%**
25% Q1	2%	2%
10%	1%	1%
5%	0%	0%
1%	0%	0%
0% Min	0%	0%

Approach 3. *For vehicles with self-reported VMT, but <u>without</u> a usable odometer reading and information on the primary driver.*

<u>Estimation</u>

There were 30,131 vehicles in this category (Table 3). Although the single odometer reading was missing for these vehicles, the strong relationship between self-reported VMT and odometer readings (and thus, the *BESTMILE* estimate from 2001) suggested the following estimation approach:

$$BESTMILE_i = \hat{\alpha} + \hat{\beta} * ANNMILES_i + R_i \qquad (5)$$

where $\hat{\alpha}$ is the intercept and $\hat{\beta}$ is the estimated coefficient for *ANNMILES*. The pseudo-residuals were assigned in similar fashion to Approaches 1 and 2.

<u>Comparison of 2001 Method vs. 2009 Method Using 2001 Vehicles</u>

This method contains no changes in terms of variable availability from 2001 to 2009. Thus, the 2009 method is identical to the 2001 method, and produced the same results when applied to 2001 vehicles.

Approach 4. *For vehicles with a usable odometer reading and information on the primary driver, but <u>without</u> self-reported VMT.*

<u>Estimation</u>

There were 5,624 vehicles in this category (Table 3). The estimation model was similar to Equation (4), except for the omission of the self-reported VMT term. In order to remain consistent with the approach used in creating the 2001 *BESTMILE*, the DRVEH variable was included in the model in lieu of estimating separate models for households with different ratios of vehicles to drivers.

Modeling similar to that in Approach 1 was performed. First, using ODOMMILES as a dependent variable was attempted, with results similar to those in Tables 6a and 6b. Thus, modeling with BESTMILE as a dependent variable was attempted with reasonable success, with the added benefit of being consistent with modeling in Approach 1.

<u>Comparison of 2001 Method vs. 2009 Method Using 2001 Vehicles and Pseudo-Residuals</u>

The standard error results of calculations for these models for 2001 vehicles using the 2009 approach in Table 10a are nearly identical to the results in Approach 1. The distribution of the percent differences (Table 10b), however, indicate that this approach is not as transferable, with 50% of all vehicles having values more than 12 to 14% different from 2001 values.. This may indicate that the self-reported estimate heavily influences prior approaches, and the lack of this variable hurts comparisons with 2001 estimates. However, given that less than 2% of all 2009 NHTS vehicles fall into this group, such an approach may be acceptable.

Table 10a. Distribution of Vehicles by Differences between Estimate using One Odometer Reading and BESTMILE Value, in terms of standard errors of BESTMILE estimates, Approach 4

	Group 1	Group 2
0-1 StdErrs	55.02%	59.91%
1-2 StdErrs	31.03%	29.07%
2-3 StdErrs	10.27%	7.80%
3-4 StdErrs	2.45%	2.07%
4-5 StdErrs	0.69%	0.38%
5+ StdErrs	0.54%	0.77%
% 0-2 SEs	**86.05%**	**88.98%**
% 0-3 SEs	**96.32%**	**96.78%**

Table 10b. Distribution of the Percentage Differences between Estimate using One Odometer Reading and BESTMILE Value, Approach 4

	Group 1	Group 2
100% Max	408290%	36289%
99%	1847%	1767%
95%	257%	230%
90%	113%	106%
75% Q3	39%	34%
50% Median	**17.9%**	**15.3%**
25% Q1	7%	6%
10%	3%	2%
5%	1%	1%
1%	0%	0%
0% Min	0%	0%

Approach 5. *For vehicles with usable information on the primary driver, but <u>without</u> odometer readings and self-reported VMT.*

<u>Estimation</u>

There were 4,374 vehicles in this group (Table 3). The estimation model again was similar to Equation (4), except for the exclusion of both self-reported VMT and the annualized single odometer term (*ODOMMILES*). As with all approaches, pseudo-residuals were assigned to develop the final BESTMILE estimate.

<u>Comparison of 2001 Method vs. 2009 Method Using 2001 Vehicles</u>

The only change between the 2001 and 2009 methods for this approach is the differing levels of the EDUC variable, which resulted in a negligible amount of difference between 2001 and 2009 methods.

Approach 6. *For vehicles with no driving information except that collected on the travel day.*

Estimation

The 9,750 remaining vehicles of usable vehicle types had no usable odometer readings, self-reported VMT, or information on the primary driver. Of these, 2,811 were used on the travel day. Thus, for these 2,811 vehicles, the total miles driven on the travel day were adjusted by simple annualization and probability factors. Equation (6) shows how the *BESTMILE* estimate for these vehicles was computed:

$$\begin{aligned} \textit{BESTMILE} = 365 \times & (\text{Miles driven on the travel day}) \\ & \times \textit{Prob}\ (\text{vehicle was driven on weekday}) \\ & \times [\textit{Mean}\ (\text{miles driven in a day})]/[\textit{Mean}\ (\text{miles driven on a weekday})] \end{aligned} \tag{6}$$

where *Prob* (vehicle was driven on weekday) is the weighted proportion of vehicles driven on a *weekday* travel day to all vehicles (essentially, the *probability* that a vehicle was driven on a weekday); and [*Mean* (miles driven in a day)]/[*Mean* (miles driven on a weekday)] is a factor to adjust the average of miles per vehicle for vehicles driven on a *weekday* travel day to average miles for any day of the week. A similar approach was used for vehicles that were driven on a travel day that was on a *weekend*. This is the same computation as was done for the 2001 *BESTMILE* variable.

Comparison of 2001 Method vs. 2009 Method Using 2001 Vehicles

This method contains no changes in terms of variable availability from 2001 to 2009. Thus, the 2001 method is identical to the 2009 method, with both methods producing exactly the same results when applied to 2001 vehicles.

Approach 7. *For vehicles not assigned a BESTMILE estimate using the other approaches, or for out of scope vehicle types*

All remaining vehicles with a self-reported mileage estimate (ANNMILES) were simply assigned values of BESTMILE equal to ANNMILES. This includes out of scope vehicles as well, and accounts for 13,961 vehicles.

5. Adjustment to a Fixed Time Frame

In the 2001 BESTMILE computations, the estimates were adjusted in the modeling stage such that they represented annual travel from May 1, 2001 to April 30, 2002. This time frame was selected because it contained the largest proportion of odometer readings compared to all other possible time spans beginning on the first day of a given month. For the 2009 estimates, the time frame of April 1, 2008 to March 31, 2009 used for the 2009 BESTMILE estimates was chosen since the majority of the survey (and thus the majority of odometer readings) was conducted during this time.

Given that the time frame adjustment in 2001 relied on the two odometer readings and their dates, and that the 2009 survey lacked this information, the adjustment was performed differently in 2009. An adjustment factor was computed for each vehicle based upon the date of the household's travel day. This adjustment factor was then applied to the final BESTMILE estimate – not in the modeling stage – and before any screening was performed. Information from *Traffic Volume Trends* (Table 11) compiled by FHWA was used as the basis for this adjustment. The numbers highlighted in green represent those in the chosen time frame.

Table 11. Monthly VMT Estimates (in millions) from Traffic Volume Trends[13]

Month	2007	2008	2009
Jan		233,276	226,296
Feb		221,006	219,145
Mar	259,343	252,297	249,159
Apr	252,398	252,220	252,853
May	267,240	261,345	
Jun	265,336	255,894	
Jul	267,019	261,785	
Aug	271,474	261,095	
Sep	246,265	238,790	
Oct	261,623	256,368	
Nov	245,955	236,902	
Dec	240,776	242,493	

Since the purpose of the adjustment factor was to adapt a BESTMILE estimate so that it reflects the April 2008 to March 2009 time period, this time period's total VMT (2,961,492 million miles) was used as a fixed numerator in the adjustment for all vehicles. The denominator was computed separately for each vehicle using VMT from Table 11 which reflected the year ending with each vehicle's travel day. The adjustment can be summarized by Equation 7 below:

$$BESTMILE_{adjusted} = BESTMILE_{original} * \frac{\text{TVT VMT from Apr. 1, 2008 to Mar. 31, 2009}}{\text{TVT VMT from X to Y}}, \quad (7)$$

where X is the date a year prior to the travel day plus one, and Y is the travel day date. Thus, the adjustment factor will always have one year's worth of VMT in both the denominator and the numerator, and the adjustment factor will be exactly 1 for vehicles where the travel day is March 31, 2009.

As an example on how travel days that were not the last day of the month were handled, say a household's travel day falls on September 13, 2008. The denominator of the adjustment factor would be computed using 13/30 of September 2008's TVT VMT

[13] http://www.fhwa.dot.gov/policyinformation/travel/tvt/history/, accessed Nov. 23, 2010.

according to Table 11, 17/30 of September 2007's TVT VMT, and the entire amount of VMT from October 2007 to August 2008. Table 12 illustrates this example.

Table 12. Computation of the Denominator of the Adjustment Factor for a Vehicle with a September 13, 2008 Travel Day

Month	Fraction	TVT VMT (millions)	Denominator VMT (millions)
Sep-07	17/30	246,265	139,550
Oct-07	1	261,623	261,623
Nov-07	1	245,955	245,955
Dec-07	1	240,776	240,776
Jan-08	1	233,276	233,276
Feb-08	1	221,006	221,006
Mar-08	1	252,297	252,297
Apr-08	1	252,220	252,220
May-08	1	261,345	261,345
Jun-08	1	255,894	255,894
Jul-08	1	261,785	261,785
Aug-08	1	261,095	261,095
Sep-08	13/30	238,790	103,476
TOTAL			**2,990,298**

So if a vehicle with a Sep. 13, 2008 travel day had a BESTMILE value of 12,000, the adjustment factor would be 2,961,492/2,990,298, or 0.990, and the adjusted BESTMILE would then be 12,000*0.990, or 11,884 miles.

The adjustment factors ranged from 0.97 to 1.0. At first glance this appears odd – one would expect that adjustment factors would range from a number below 1 to a number above 1. However, the time period from April 2008 to March 2009 was an unusual one in historical terms, with VMT actually declining year over year. Table 13 below shows the differences in monthly TVT VMT using the chosen time frame as a base (meaning April 2007 numbers are compared with April 2008 numbers, for example). As one can see, the April 2008 to March 2009 numbers are always lower than the same month's VMT for a previous or subsequent year, with the sole exception being December 2007 vs. 2008. However, since the adjustment factor uses a yearly total, this December

difference was always negated by the larger drops in VMT in other months before and after.

Table 13. Differences in Monthly VMT Estimates (in millions) from Traffic Volume Trends, Using April 2008 to March 2009 as a Basis of Comparison

Month	2007	2008	2009
Jan		(6,980)	0
Feb		(1,861)	0
Mar	(10,184)	(3,138)	0
Apr	(178)	0	(633)
May	(5,895)	0	
Jun	(9,442)	0	
Jul	(5,234)	0	
Aug	(10,379)	0	
Sep	(7,475)	0	
Oct	(5,255)	0	
Nov	(9,053)	0	
Dec	1,717	0	

Once the adjustments were made, screening of the results was completed.

6. Screening of BESTMILE Estimates

Table 14 below shows a comparison similar to Table 7, adding the eventual results of BESTMILE computations for 2001 and 2009 datasets. As a whole, the estimates are much closer to *Highway Statistics* estimates, and reflect the trends of *Highway Statistics* more closely than the self-reported mileage. This suggests that the BESTMILE does indeed improve upon available data to provide better estimates for a given vehicle, particularly when considering that the self-reported mileage numbers were not universally adjusted downward to account for the time frame, as explained in Section 5.

Table 14. Comparison of 2001 and 2009 Average Miles per Vehicle, *Highway Statistics* and NHTS Self-Reported (ANNMILES) and Best Available (BESTMILE) Estimates

	2001	2009*	% diff
Highway Statistics			
Passenger Cars	11,831	11,788	-0.4%
Other 2-Axle, 4-Tire Vehicles	11,204	10,951	-2.3%
Passenger Cars & Other 2-Axles, 4-Tire Vehicles	11,593	11,432	-1.4%
NHTS ANNMILES (Self-Reported Mileage)			
Automobile/car/station wagon	10,695	10,054	-6.0%
Van (mini, cargo, passenger)	12,717	11,030	-13.3%
Sports utility vehicle	12,722	11,584	-8.9%
Pickup truck	11,729	9,891	-15.7%
All	11,078	10,088	-8.9%
NHTS BESTMILE			
Automobile/car/station wagon	11,609	11,118	-4.2%
Van (mini, cargo, passenger)	13,400	12,255	-8.5%
Sports utility vehicle	13,905	12,590	-9.5%
Pickup truck	12,473	11,240	-9.9%
All	11,979	11,176	-6.7%

* The most recent data for Highway Statistics is for the year 2008. Data can be found at http://www.fhwa.dot.gov/policy/ohpi/hss/hsspubs.cfm.

Once calculation of the best estimates was completed, the estimates were checked for reasonableness at the individual vehicle level. Once again, the lack of two odometer readings prohibits most of the adjustments done in the 2001 NHTS from being done to the 2009 data. Negative best estimates were set to zero, and estimates over 200,000 miles were capped at 200,000. An additional new check comparing the single odometer reading to the best estimate was also performed. If the annualized best estimate was greater than the odometer reading, and the vehicle age was greater than 1, the best estimate was set to the initiate annual estimate (*ODOMMILES*) computed in Section 3. These adjustments are summarized in Table 15.

In order to identify outliers, each *BESTMILE* estimate was compared to the initial annual estimate (*ODOMMILES*) as well as the self-reported estimate (*ANNMILES*). Outlier codes were then assigned based on subjective criteria. If *BESTMILE* was different from either *ODOMMILES* or *ANNMILES* by a factor of 4, with an absolute difference of more than 10,000 miles, an outlier code was assigned. These codes are found in Table 16.

Table 15. Adjustments to *BESTMILE*

Adjustment Code	Frequency	Percent	Criteria	Adjustment
No Code	303,000	98.01%	No adjustment	
1	5,330	1.72%	*BESTMILE* > Odometer Reading, *BESTMILE* > Self-Reported VMT, and Vehicle Age > 1	*BESTMILE* set to *ODOMMILES* value
2	350	0.11%	*BESTMILE* > Odometer Reading and Vehicle Age > 1 (for vehicles without Self-Reported VMT)	*BESTMILE* set to *ODOMMILES* value
3	405	0.13%	*BESTMILE* < 0	*BESTMILE* = 0
4	7	0.00%	*BESTMILE* > 200,000	*BESTMILE* = 200,000
5	71	0.02%	*BESTMILE* > 200,000 after Adjustment #1 or #2	*BESTMILE* = 200,000
Total	**309,163**	**100.00%**		

Table 16. Outlier Codes for *BESTMILE*

BEST_OUT	Frequency	Percent	Criteria
No Code	287,805	93.09%	
1	6,392	2.07%	$BESTMILE < \frac{ODOMMILES}{4} and \mid BESTMILE - ODOMMILES \mid > 10{,}000 miles$
2	1,321	0.43%	$BESTMILE < \frac{ANNMILES}{4} and \mid BESTMILE - ANNMILES \mid > 10{,}000 miles$
3	3,033	0.98%	$BESTMILE < ODOMMILES * 4$ $and \mid BESTMILE - ODOMMILES \mid > 10{,}000 miles$
4	10,612	3.43%	$BESTMILE < ANNMILES * 4$ $and \mid BESTMILE - ANNMILES \mid > 10{,}000 miles$
Total	**309,163**	**100.00%**	