

**Ozone Science and Air
Modeling Research
Project Area H-8B:
Modeling and MOBILE6 –
Development of Local Mileage
Accumulation Rates**

FINAL REPORT

Prepared for:

The Houston Advanced Research Center

Prepared by:

Eastern Research Group, Inc.

August 29, 2003

ERG No.: 3381.00.002.001

OZONE SCIENCE AND AIR MODELING RESEARCH

**PROJECT AREA H-8B: MODELING AND MOBILE6 –
DEVELOPMENT OF LOCAL MILEAGE ACCUMULATION RATES**

FINAL REPORT

Prepared for:

The Houston Advanced Research Center
TERC
4800 Research Forest Drive
The Woodlands, TX 77381

Prepared by:

Eastern Research Group, Inc.
5608 Parkcrest Drive, Suite 100
Austin, TX 78731

August 29, 2003

Table of Contents

INTRODUCTION	1
Initial Data Preparation	1
Advanced Data Preparation	2
Final Creation of MARs.....	16

List of Tables

Table 1. Number of Observations for each VIN.....	2
Table 2. MOBILE6 Default & Calculated Houston-Specific MARs	24

List of Figures

Figure 1. Annual Mileage Accumulation Rates by Vehicle Class (LDV).....	4
Figure 2. Annual Mileage Accumulation Rates by Vehicle Class (LDT1)	5
Figure 3. Annual Mileage Accumulation Rates by Vehicle Class (LDT2)	6
Figure 4. Annual Mileage Accumulation Rates by Vehicle Class (LDT3)	7
Figure 5. Annual Mileage Accumulation Rates by Vehicle Class (LDT4)	8
Figure 6. Possible MARs for LDVs Given Different Data Clean Up Scenarios	13
Figure 7. Median MARs Values for LDVs Given Different Data Clean Up Scenarios	14
Figure 8. Number of Records Used for LDVs Given Different Data Clean Up Scenarios	15
Figure 9. LDV Calculated MARs	18
Figure 10. LDT1 Calculated MARs.....	19
Figure 11. LDT2 Calculated MARs.....	20
Figure 12. LDT3 Calculated MARs.....	21
Figure 13. LDT4 Calculated MARs.....	22
Figure 14. Number of Observations Used	23

INTRODUCTION

The objective of this analysis was to create Houston area specific mileage accumulation rates (MARs) for use in MOBILE6 models. The area specific MARs can replace the MOBILE6 defaults and improve the accuracy of the model. The MARs are based on historical odometer readings from the Texas Inspection and Maintenance Vehicle Inspection Database (VID) for Harris County. Although data is not available for the remaining seven counties, Harris County vehicles are responsible for roughly three quarters of all Vehicle Miles Traveled (VMT) in the region. Therefore, development of mileage accumulation rates for Harris County alone will provide an important adjustment to the total on-road inventory.

All of the odometer data in the VID was hand entered by inspectors and, as such, are subject to significant error. The majority of the analysis for creating the MARs focused on removing, or accounting for, these errors as best as possible. This document discusses the data processing steps used to create the MARs and presents the Houston area specific MARs that were produced.

Initial Data Preparation

The VID dataset, as received at ERG, contained readings from Harris County from January 1, 1997 to March 31, 2003 with over 12.5 million observations and over 5 million unique vehicle identification numbers (VINs). All of these records were imported into SAS for data clean up and analysis.

The VIN information contained within the VID data is very useful for data clean up and analysis. Each correct VIN is guaranteed to be unique to an individual vehicle and does not change with change of ownership. Therefore, we are able to use the VIN information to easily track each vehicle's odometer reading through the years. Since we are only interested in the change in the odometer over time, we removed all VINs that only had one odometer entry associated with it. This removed 2.2 million records.

After removal of the singular VINs, we then processed the remaining VINs through ERG's VIN Decoder program. This program decodes each VIN using the manufacturer's information and categorizes each vehicle into its appropriate MOBILE6 vehicle class category. Using the VIN Decoder results, we removed all records associated with vehicles that had VINs that decoded with errors. VIN errors are commonly introduced into the VIN when inspectors make VIN transcription errors. We also retained only gas vehicles that were light duty (LDV, LDT1, LDT2, LDT3, or LDT4). This removed 908,998 observations and 291,280 VINs.

We then examined the remaining VINs to see how many readings we had for each VIN. Through this process we discovered that there was a small number of vehicles that had a large number of repeat readings, including several vehicles that had several hundred readings in the five and a half years of observations. The highest count was 628 readings for one vehicle. The VINs for these vehicles were valid and decoded without errors. It was our assumption that these vehicles were most likely control or certification vehicles and would not appropriately represent the mileage accumulation rates for the general population. We removed all VINs with over 15

measurements. This amounted to 182 VINs, which combined to form 13,689 observations. Table 1 shows the grouping of the remaining VINs by the number of observations for each VIN.

Table 1. Number of Observations for each VIN

Number of Readings (Per VIN)	Number of VINs	Percent of VINs	Cumulative Number of VINs	Cumulative Percent of VINs
2	823,771	31.44	823,771	31.44
3	602,776	23.00	1,426,547	54.44
4	472,825	18.05	1,899,372	72.49
5	362,020	13.82	2,261,392	86.31
6	222,031	8.47	2,483,423	94.78
7	83,753	3.20	2,567,176	97.98
8	30,648	1.17	2,597,824	99.15
9	12,356	0.47	2,610,180	99.62
10	5,394	0.21	2,615,574	99.82
11	2,415	0.09	2,617,989	99.92
12	1,099	0.04	2,619,088	99.96
13	543	0.02	2,619,631	99.98
14	251	0.01	2,619,882	99.99
15	139	0.01	2,620,021	99.99
All VINs with over 15 observations were removed.				

Advanced Data Preparation

Once the initial cleanup was finished, two primary types of data errors remained: rollovers and typos. Rollovers occur when a five-digit odometer accumulates more than 99,999 miles and starts over at zero miles. This causes the odometer in the current test year to read lower than the odometer reading recorded for the previous test year. Most newer model year vehicles have six-digit odometers that must accumulate 999,999 miles before rolling over. While technically possible, it is highly unlikely that any vehicle in this sample set has gone over a million miles. The second common remaining error stems from typos when entering the odometer readings. Typos can cause the odometer reading to be higher or lower than the actual odometer reading. Unfortunately, the existence of typos makes the detection of odometer rollovers more difficult.

The change in odometer (delta odometer) was calculated as the difference between a current odometer reading for a VIN minus the previous odometer reading for this VIN. Similarly, the change in time (delta date) is calculated as the number of days between the current reading and the previous reading for this VIN. The annual mileage accumulation rate is then calculated as shown below.

$$\text{Annual Miles Travelled} = \frac{\Delta \text{Odometer (miles)}}{\Delta \text{Date (days)}} * 365.25 \left(\frac{\text{days}}{\text{year}} \right)$$

Figures 1 through 5 show the observed annual mileage accumulation rate distributions for each of the light duty gas MOBILE6 vehicle classes. The data for these graphs were not corrected for rollovers or typos.

The California Air Resources Board (ARB) has developed a simple method for correcting odometer rollovers. They have established a cutoff of 100,000 miles as the maximum reasonable difference allowed between consecutive test year odometer readings. They then calculated the accumulated miles for a vehicle from Year 1 to Year 2 as:

If YEAR1ODO < YEAR2ODO
Then ACCUM = YEAR2ODO – YEAR1ODO

If YEAR1ODO > YEAR2ODO
and YEAR1ODO < 100,000
and YEAR2ODO < 100,000
Then ACCUM = 100,000 + YEAR2ODO – YEAR1ODO

We applied the above methodology to the Harris County VID data and then analyzed the results. We found that after using this methodology there was a large number of vehicles that had been treated as a rollover and subsequently had over 80,000 miles accumulated in one year. An example of this was a car that had a year 1 reading of 35,000 miles followed by a year 2 reading of 25,000 miles. The above methodology would treat it as a rollover with 90,000 miles accumulated. While this is technically possible, it occurred at a high percentage in the VID dataset and was more likely the result of a typo. We decided that the 100,000 cut off was too high for the majority of the vehicles in this dataset and needed to be refined. Incorrectly assigning rollover odometer readings would introduce a high bias in the MARs.

To develop better rollover cutoff values based on the non-negative values of the Houston VID, we temporarily deleted all records associated with any VIN that had one or more negative annual miles accumulated. This removed 396,140 VINs and a total of 1,428,646 observations, which left 6,911,522 observations. We then removed any delta odometer based on readings less than 6 months apart. A minimum time interval between odometer readings is required to ensure that the annual mileage accumulation is representative of an entire year's driving pattern. In a similar fashion, we also removed any delta odometer based off of readings over 2 years apart. These deltas span multiple vehicle ages and are not useful in distinguishing the changing driving pattern from year to year. This left us with 4,098,979 observations. We then removed any reading over 100,000 annual miles in order to remove the gross positive errors resulting from typos. This left 4,070,127 observations. From this final dataset, we produced temporary annual mileage accumulation rates for each vehicle class and vehicle age to be used in refining the rollover detection process.

Figure 1. Annual Mileage Accumulation Rates by Vehicle Class (LDV)

Annual Mileage Accumulation Rates by Vehicle Class
MOBILE6_Vehicle_Class=LDV

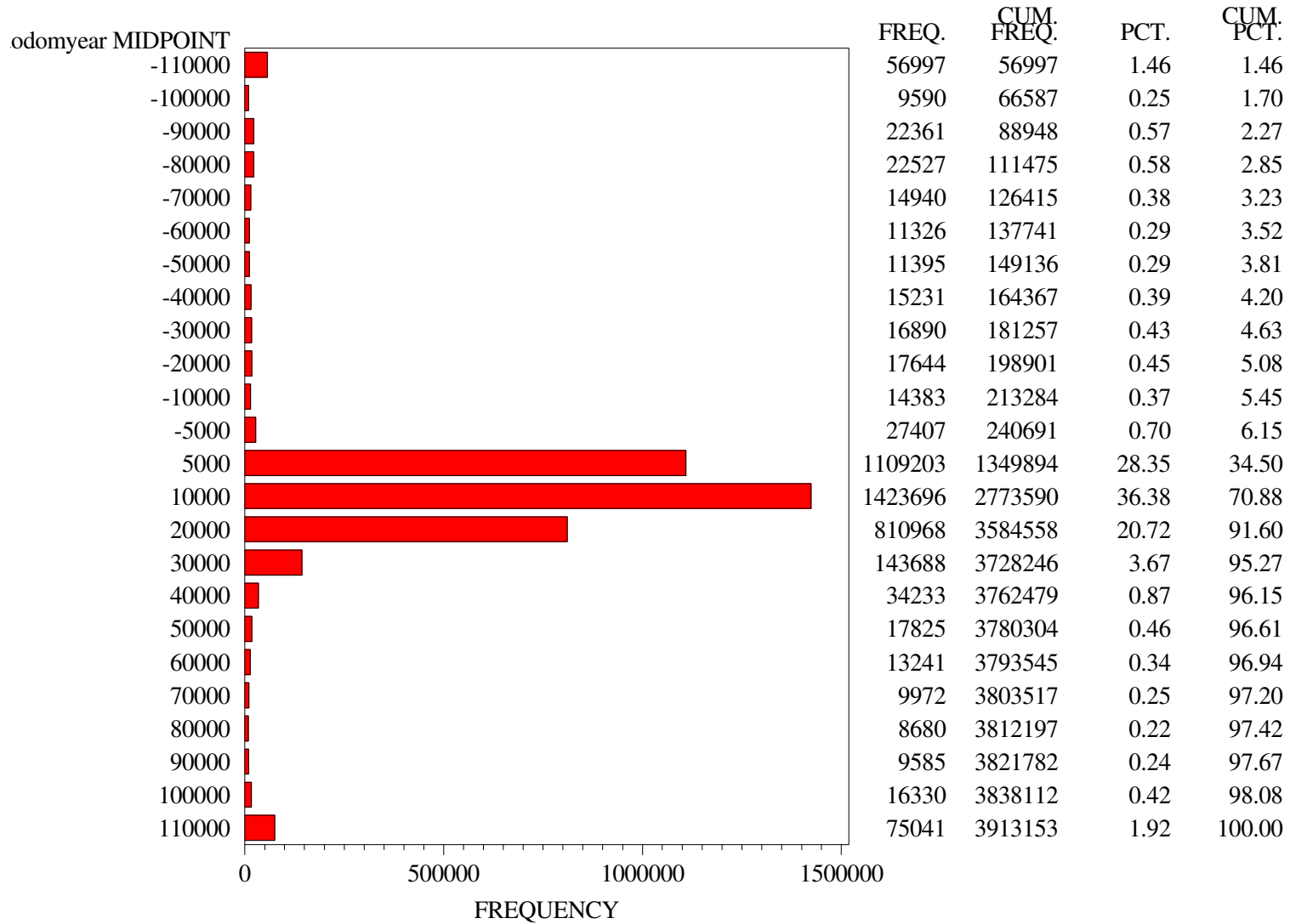
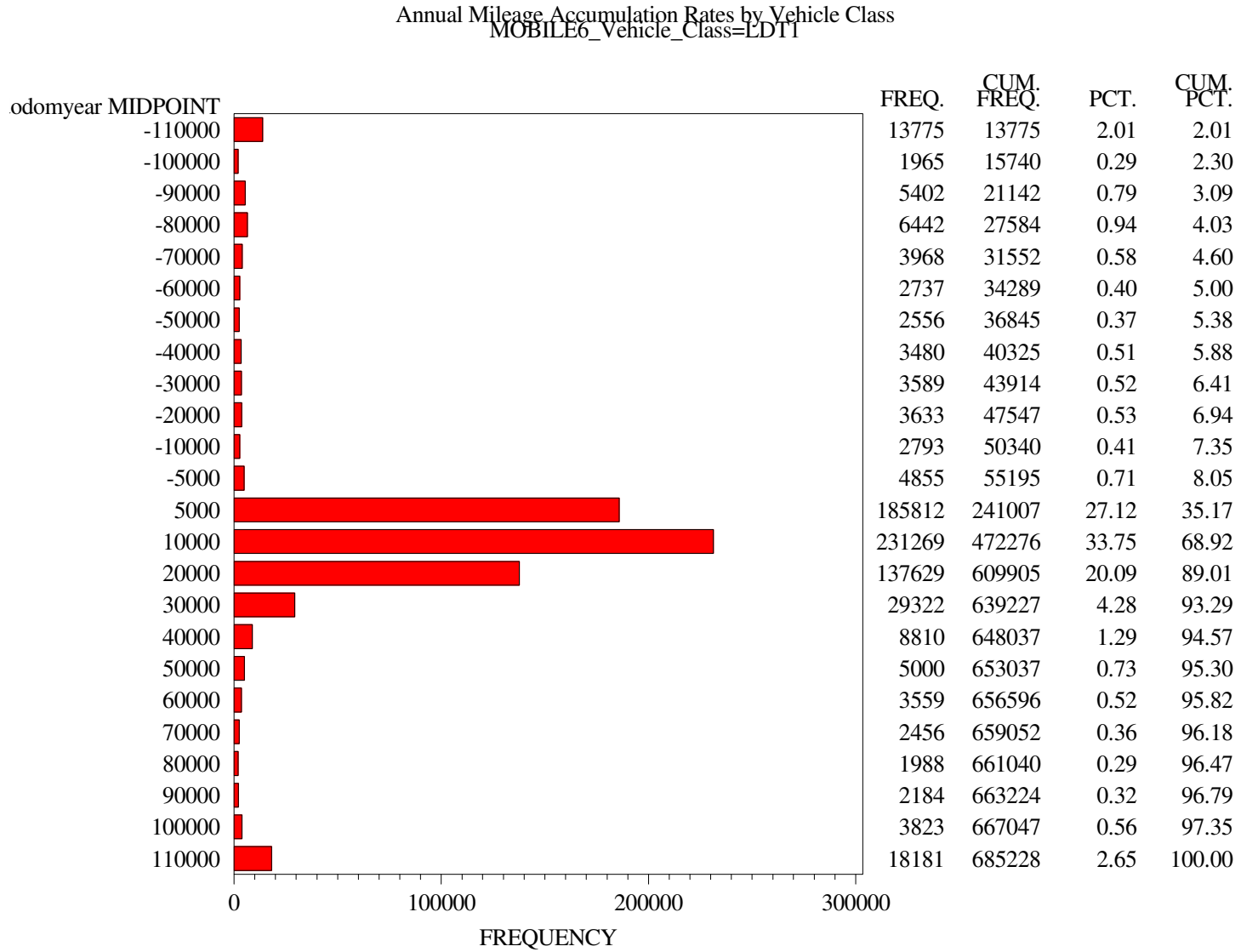


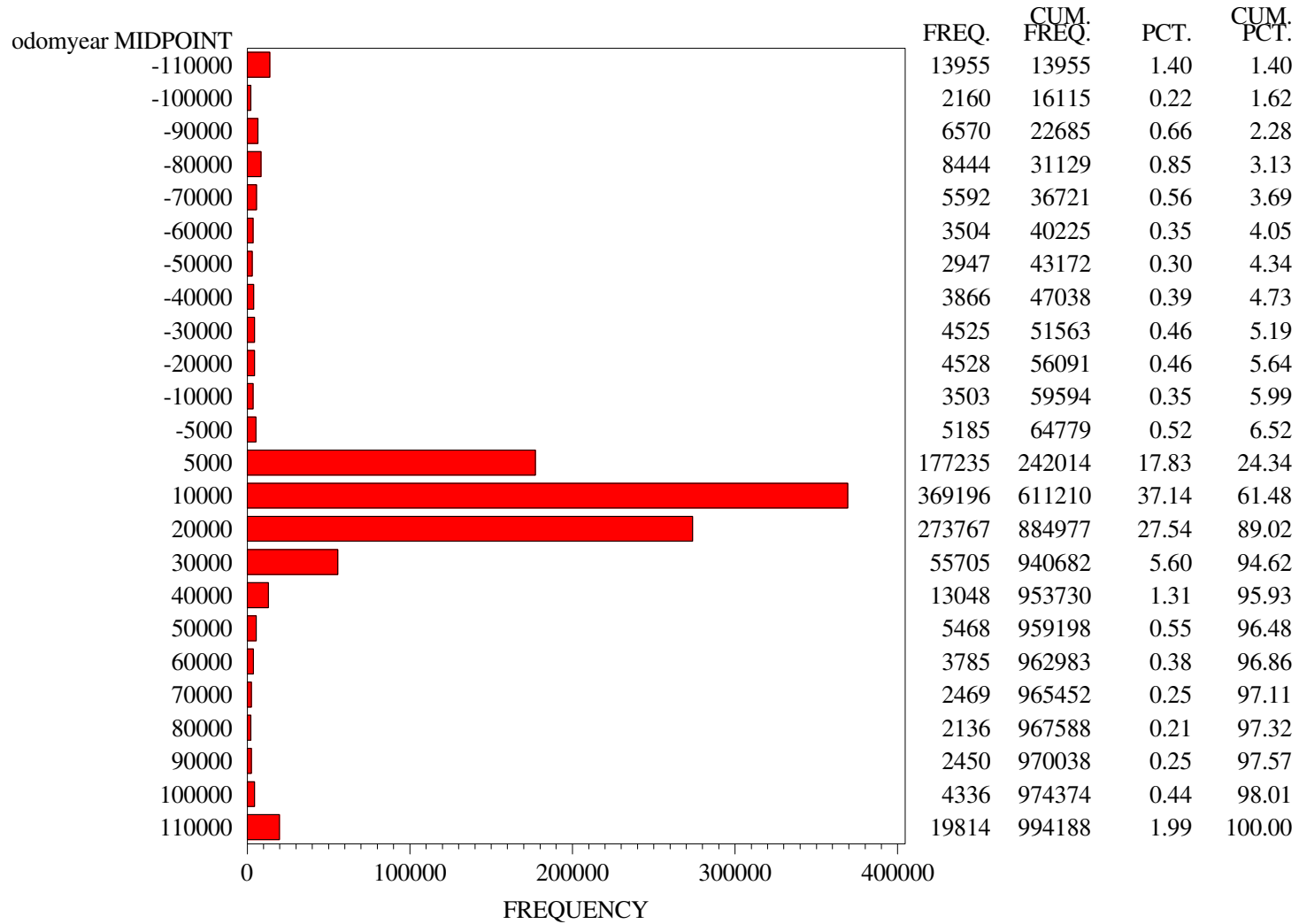
Figure 2. Annual Mileage Accumulation Rates by Vehicle Class (LDT1)



/roadhog/HARC/MARS/Ewgodom.sas

Figure 3. Annual Mileage Accumulation Rates by Vehicle Class (LDT2)

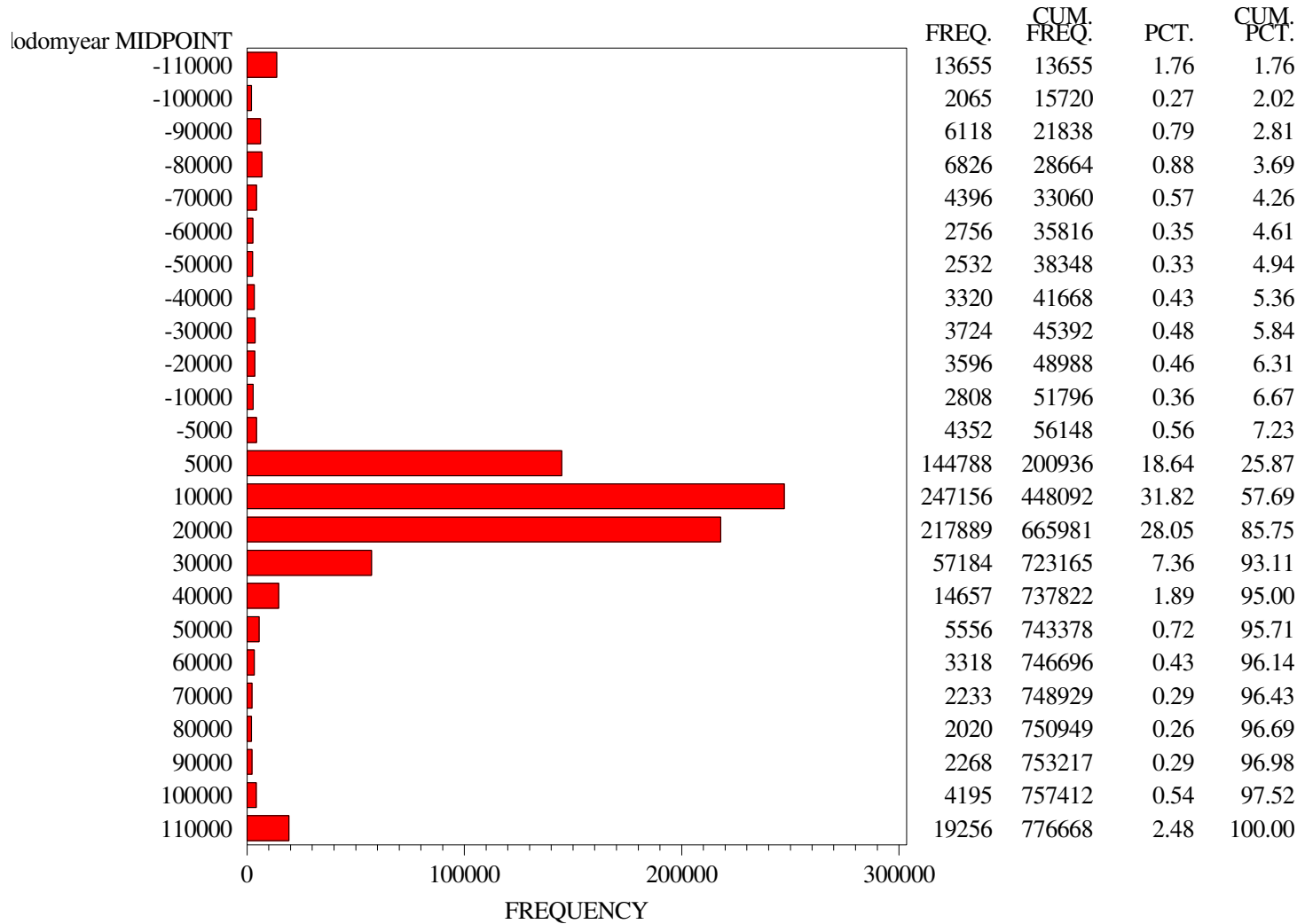
Annual Mileage Accumulation Rates by Vehicle Class
MOBILE6_Vehicle_Class=LDT2



/roadhog/HARC/MARS/Ewgodom.sas

Figure 4. Annual Mileage Accumulation Rates by Vehicle Class (LDT3)

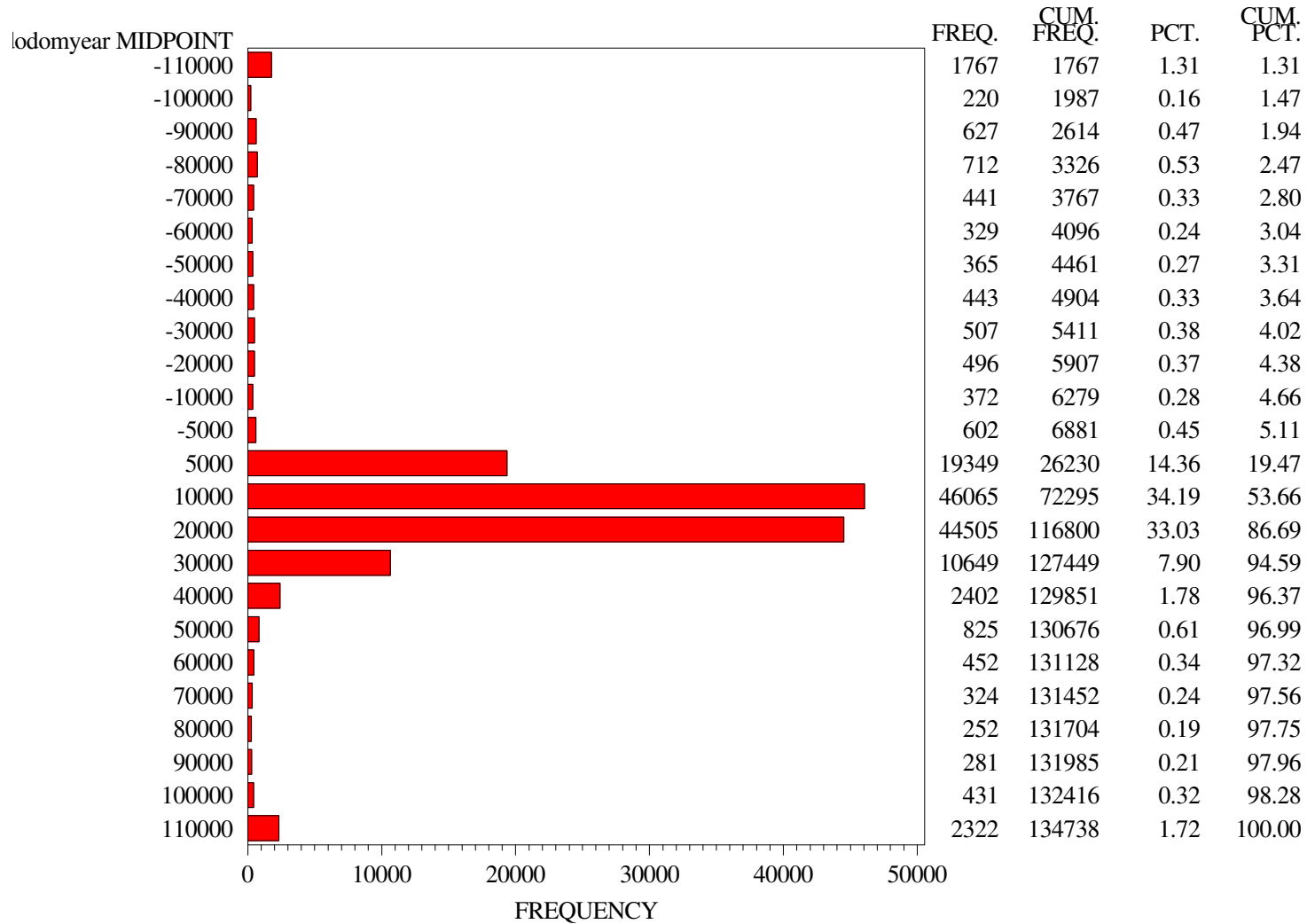
Annual Mileage Accumulation Rates by Vehicle Class
MOBILE6_Vehicle_Class=LDT3



/roadhog/HARC/MARS/Ewgodom.sas

Figure 5. Annual Mileage Accumulation Rates by Vehicle Class (LDT4)

Annual Mileage Accumulation Rates by Vehicle Class
MOBILE6_Vehicle_Class=LDT4



/roadhog/HARC/MARS/Ewgodom.sas

The temporary MARs developed from the above datasets are biased high because we only removed a large number of negative errors and a small number of gross positive errors. There still remain an unknown number of positive errors from typos in the dataset that have not been removed or accounted for. While this will not serve as our final answer, it will provide a decent estimate that can be used to determine rollovers. This temporary MARs dataset contains an annual mileage accumulation rate for each vehicle type for vehicles aged 1 to 30 years. The dataset also contains the standard deviation for each MAR.

In the previous rollover routine, we used a constant cut off of 100,000 miles as the maximum reasonable difference allowed between consecutive test year odometer readings. Our revised rollover methodology uses the MARs calculated in the temporary dataset plus 2.6 times the standard deviation as the maximum cut off. Please see the example below for clarification. 2.6 was chosen as the 99% confidence interval for a 2 tail distribution.

EXAMPLE:

From the temporary MARs dataset, we can look up the MARs and standard deviation for a 3-year-old, LDT1 vehicle (note: values given below are for example only and do not reflect actual data):

MARs = 30,000 miles
STD = 2,000 miles

Vehicle 1: 3-year-old LDT1 vehicle

Year 1 odometer reading = 85,000 miles
Year 2 odometer reading = 16,500 miles

Since Year 2 odometer is less than Year 1 odometer and both are below 100,000 miles, we consider this to be a potential rollover. We must check to see if this is true.

Delta odometer = $(100,000 + 16,500) - 85,000 = \underline{31,500 \text{ miles}}$
Cut off = $30,000 + 2.6 * 2,000 = \underline{35,200 \text{ miles}}$

Since Delta odometer (31,500 miles) is less than Cut off (35,200 miles) we consider this to be a true rollover.

Vehicle 2: 3-year-old LDT1 vehicle

Year 1 odometer reading = 75,000 miles
Year 2 odometer reading = 20,000 miles

Since Year 2 odometer is less than Year 1 odometer and both are below 100,000 miles, we consider this to be a potential rollover. We must check to see if this is true.

Delta odometer = $(100,000 + 20,000) - 75,000 = \underline{45,000 \text{ miles}}$
Cut off = $30,000 + 2.6 * 2,000 = \underline{35,200 \text{ miles}}$

Since the delta odometer (45,000 miles) is greater than the cut off (35,200 miles), we consider this not to be a rollover and we treat it as a typo.

By using the above logic and the temporary MARs that we developed, we designated fewer delta odometers as rollovers and instead considered more to be typos. This left only the typos remaining to be accounted for. We thought of several different possible solutions of how to handle the typos in the data. Five of the scenarios that we created are as follows:

Scenario A. Simple

1. Simple Roll over detection. (static cut off of 100,000 miles).
2. Delete any negative delta odometer.
3. Delete any annual mileage accumulation over 100,000 miles.

Scenario B. Ultra Conservative

1. Do not perform any rollover corrections.
2. Delete any vehicle and all of its records if it has one negative delta odometer.
3. Delete any annual mileage accumulation over 100,000 miles.

Scenario C. Ultra Loose

1. Do not perform any rollover corrections.
2. Keep all records regardless of negative delta odometers.
3. Delete any annual mileage accumulation over 100,000 or under -100,000

Scenario D. Improved Roll Over

1. Use the results from Scenario B to refine the rollover calculation
2. Delete all records associated with a VIN with one or more negative delta odometers.
3. Delete any annual mileage accumulation over 100,000 miles.

Scenario E. Improved Roll Over plus Ultra Loose

1. Use the results from Scenario B to refine the rollover calculation
2. Delete all records associated with a VIN that has an annual mileage accumulation over 100,000 or under -100,000 miles.

The average annual mileage accumulation rates from the above five scenarios are shown in Figure 6. The median annual mileage accumulation rates for each scenario is given in Figure 7 and the number of vehicles used in each scenario is shown in Figure 8. By examining these figures we can start to see several trends. First, there is close agreement among all five scenarios for the younger vehicles and all of them show a significant increase in annual mileage accumulation when compared to the MOBILE6 defaults. This increase was expected due to the large driving distances present in Houston.

As the vehicle age increases, the five different scenarios begin to spread out greatly. This shows that the results for the older vehicles are very dependent on the data clean up scenario that is used to correct the typos and rollovers. Part of this dependence stems from the reduced number of records present for the older vehicles (Figure 8). As the number of observations drops substantially, the process used to handle the typos and rollovers becomes increasingly important and greatly affects the final results. The older vehicles are also more likely to have 5 digit odometers and higher mileage, making them much more dependent on the methodology used to correct for rollovers.

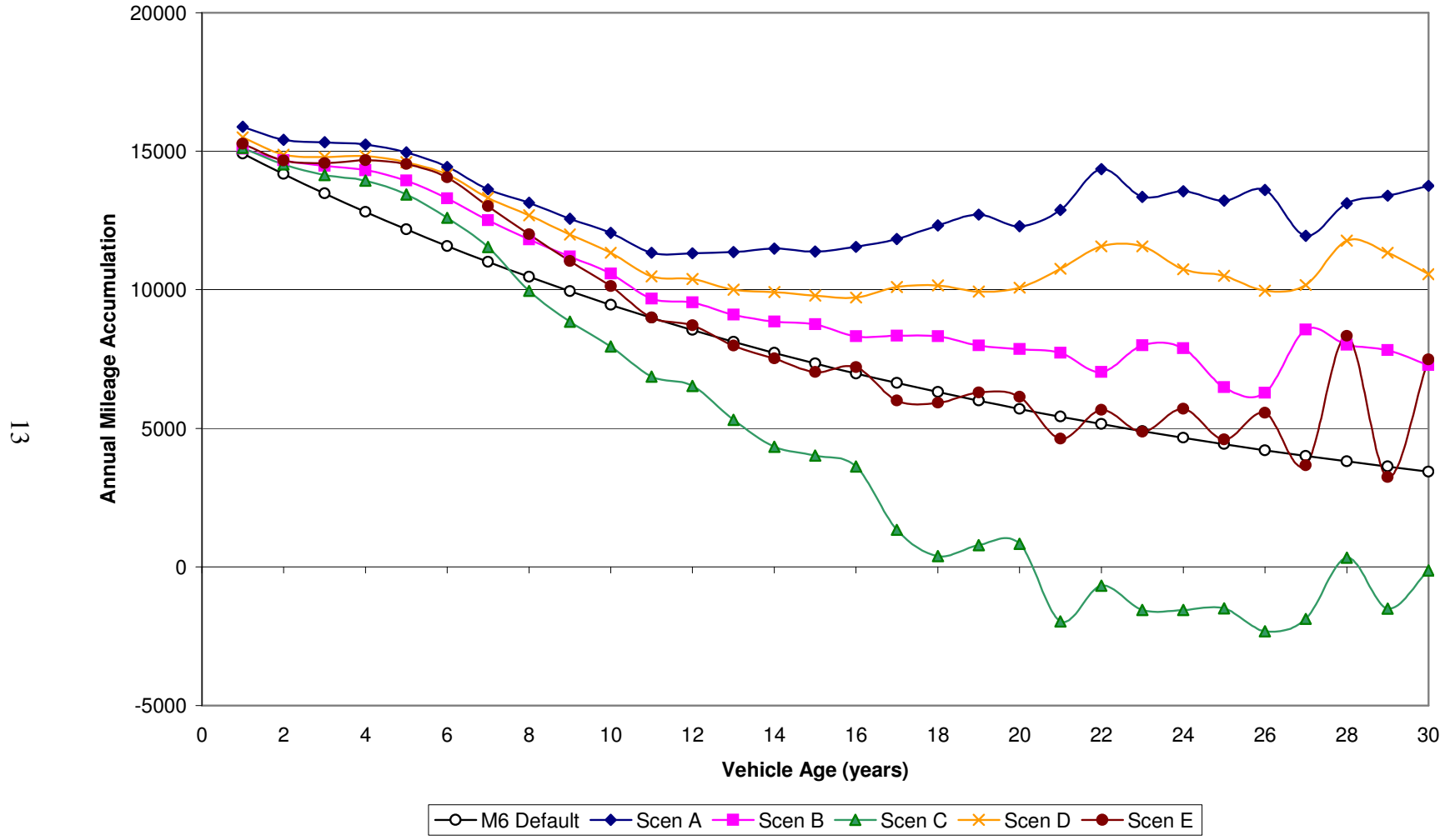
Some of the scenarios presented have a known bias in them, which is introduced by the data cleanup routines. The high bias in Scenario A is first introduced through using 100,000 miles for the rollover cutoff. This methodology misidentifies a large number of typos as rollovers that results in large delta odometers. Secondly, Scenario A also only deletes negative delta odometers. As discussed earlier, a typo in an odometer can cause a positive direction error, a negative direction error, or both. By deleting only negative delta odometers, all of the positive directional errors remain. These two positive biases combine to set Scenario A as the upper end of our possible mileage accumulation rates.

In the other direction, Scenario C serves as the lower end of the spectrum with a strong negative bias. In this scenario we did not correct for any rollovers and we only removed annual mileage if they were above 100,000 miles or below negative 100,000 miles. By not correcting for rollovers, we greatly underestimated the older vehicle mileage accumulation. This bias, combined with negative typos, actually drove the annual mileage accumulation rates below zero for older vehicles.

The median values shown in Figure 7 show a much closer agreement among all five scenarios. This is as expected because the median, by definition, is less susceptible to extreme values. The median values will always be below the mean values shown in Figure 7 because the delta odometers, by their nature, are a positively skewed distribution. Because of this, the median values cannot be used as the basis for the MARs. They can, however, be used to give the trend in annual miles accumulated which can then be used to check the final answer.

The primary difficulty in handling the typos is the detection of typos that cause positive errors in the change in odometer. It is impossible to say whether a large sudden increase in a vehicle odometer is due to a typo or caused by some change in the vehicle driving pattern, such as a long road trip. We can assume that a typo in an odometer reading is just as likely to create a positive error as it is to create a negative error. Due to the randomness of the typos, we assume that the errors caused by the typos are evenly distributed around the mean value. As such, the best way to handle the typos to minimize the bias in the MARs will be to leave the negative errors caused by the typos in the dataset so they balance out the positive type errors when we average delta odometers to calculate the mean MARs. In this manner, the typos will merely cause an increase in the standard deviation, but should not bias the mean mileage accumulation rate for each vehicle type in the fleet. When we applied this methodology to Scenario E and compare the mean versus the median (Figure 6 and 7), we see that the trends of mean and median follow closely to each other as expected.

Figure 6. Possible MARs for LDVs Given Different Data Clean Up Scenarios



13

Figure 7. Median MARs Values for LDVs Given Different Data Clean Up Scenarios

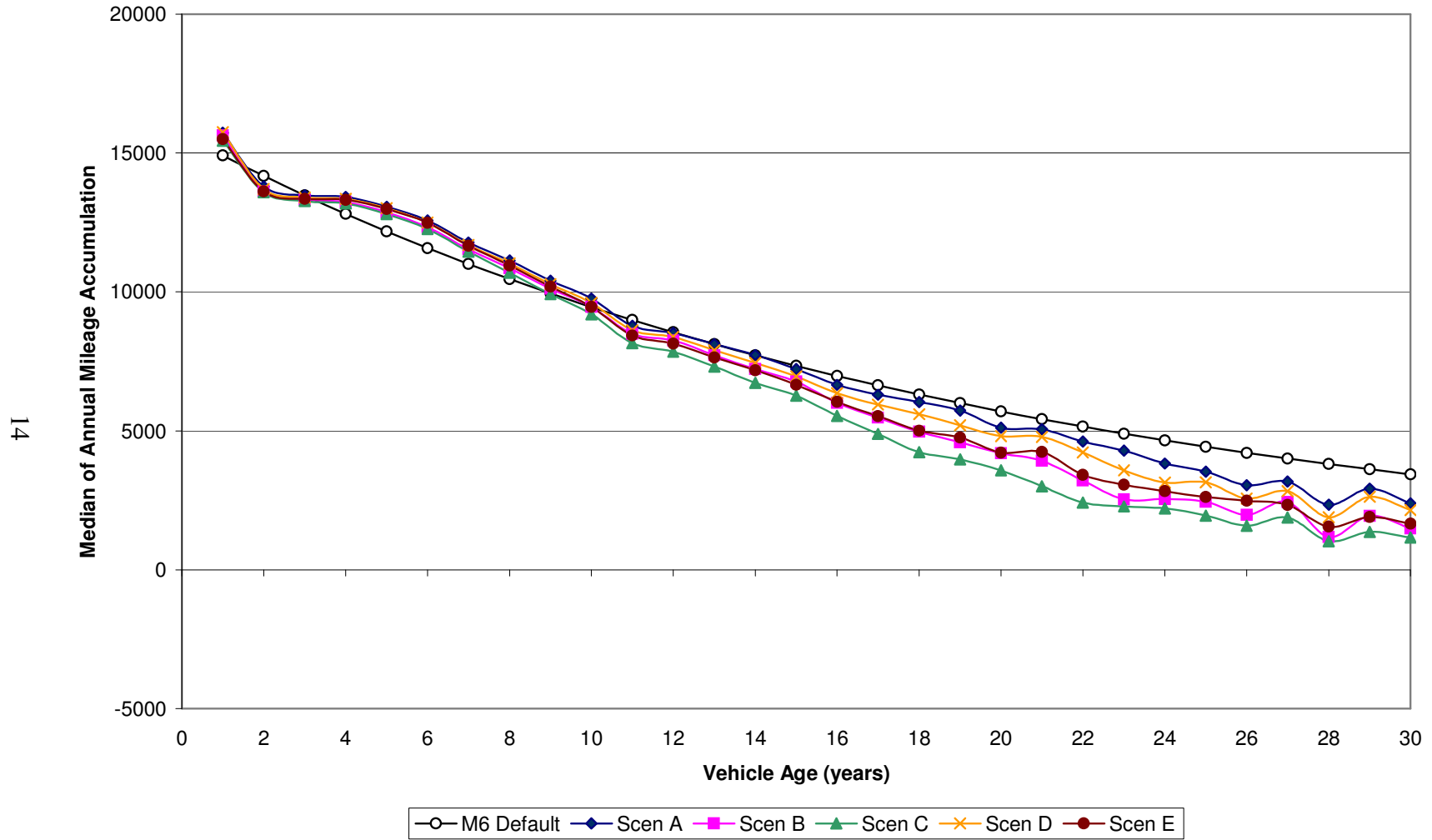
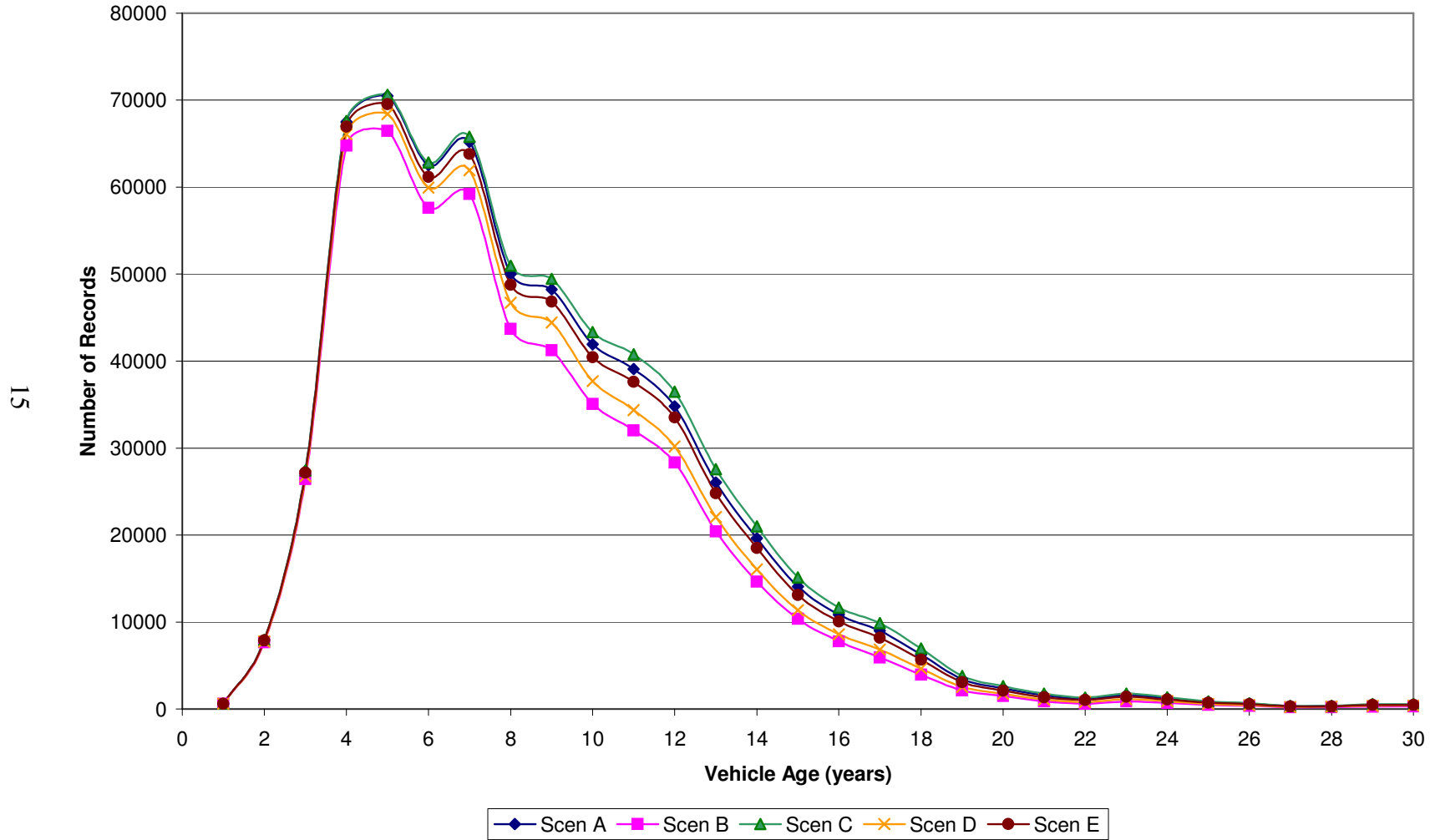


Figure 8. Number of Records Used for LDVs Given Different Data Clean Up Scenarios



Additionally, it was discovered that there were some typos that created gross errors of either over 100,000 miles in a year or under –100,000 miles in a year, such as one delta odometer showing the vehicle traveled six million miles in one year. Due to the extreme size and infrequency of these gross errors, they do not average out. We therefore decided to remove a VIN and all of its associated records if it had under –100,000 miles accumulated in a year. This removed 99,936 VINs with a total of 369,405 observations. We then removed all VINs and all of their associated records that had over 100,000 miles accumulated in a year. This removed 93,879 VINs with a total of 309,173 observations. We then removed any delta odometer based on readings less than 6 months apart or over 2 years as discussed earlier. This left 4.6 million observations spread across 5 different recording years and 5 vehicle classes.

Final Creation of MARs

The final creation of the MARs for the Houston area incorporated both the modified rollover analysis as well as the averaging of the positive and negative typos. Figures 9 through 13 present the final MARs that were developed for this project. Each plot shows the calculated MARs with error bars as well as the MOBILE6 defaults. The error bars show the 95% confidence interval about the mean. Figure 13 shows the number of mileage accumulation observations used for each MAR calculated. As seen before, the number of observations greatly decreases as the age of the vehicles increase. This is also reflected in the increase in the size of the error bars and the oscillations of the averages for the calculated MARs as age increases. These error bars show only the error that is present in the scenario used to create this set of MARs. They do not show the uncertainty that arises from the choice of the scenario as Figure 6 illustrates. Inclusion of this error would make only a small difference for the younger vehicles but greatly increase the error bars for the older vehicles.

After analyzing Figures 9 through 14, it becomes apparent that the calculated MARs match the MOBILE6 defaults quite well. This matching serves as independent verification of our methodology. On the whole there does seem to be a small but systematic increase in MARs across vehicle types, typically between years three and ten, relative to MOBILE6 defaults. However, on the whole older vehicle MARs are not significantly different than the MOBILE6 defaults. This is shown by the MOBILE6 defaults being encompassed by the error bars for half of the older age vehicles. Given the 95% confidence intervals and the scenario-to-scenario uncertainty, there is insufficient evidence to support that LDV MARs for Harris County are significantly different than the MOBILE6 defaults for essentially all vehicles over 11 years old. A similar pattern exists for other vehicle types, although decreases in sample sizes, especially for the LDT3 and LDT4 categories, yield erratic calculated trends.

In addition, the relative uncertainty associated with the first two years of MAR data is relatively high, especially for the truck categories, due to unusually low sample sizes for these years. (Also, the first two years of calculated MARs for the LDT1, 3, and 4 categories also show an anomalous peak at year 3, adding to our suspicion about their credibility.) The low sample sizes are an artifact of the I/M program itself, which does not require vehicle testing for a full two years after new vehicle purchase. In fact, one would expect no data for this time period at all, if all vehicles received their first I/M test at the correct time. For these reasons we believe the first two years of calculated MAR data should be replaced. A conservative solution would be to set year one and two values equal to year three levels.

Given these observations, we propose to use the Harris County calculated MARs for the first 11 years of vehicle age, with year one and two levels set to equal year 3, and MOBILE6 defaults for the remainder of the vehicle age. We believe that this captures the majority of the area specific driving patterns as possible with the existing dataset.

Table 2 shows the final recommended vehicle MARs for use with MOBILE6.

Figure 9. LDV Calculated MARs

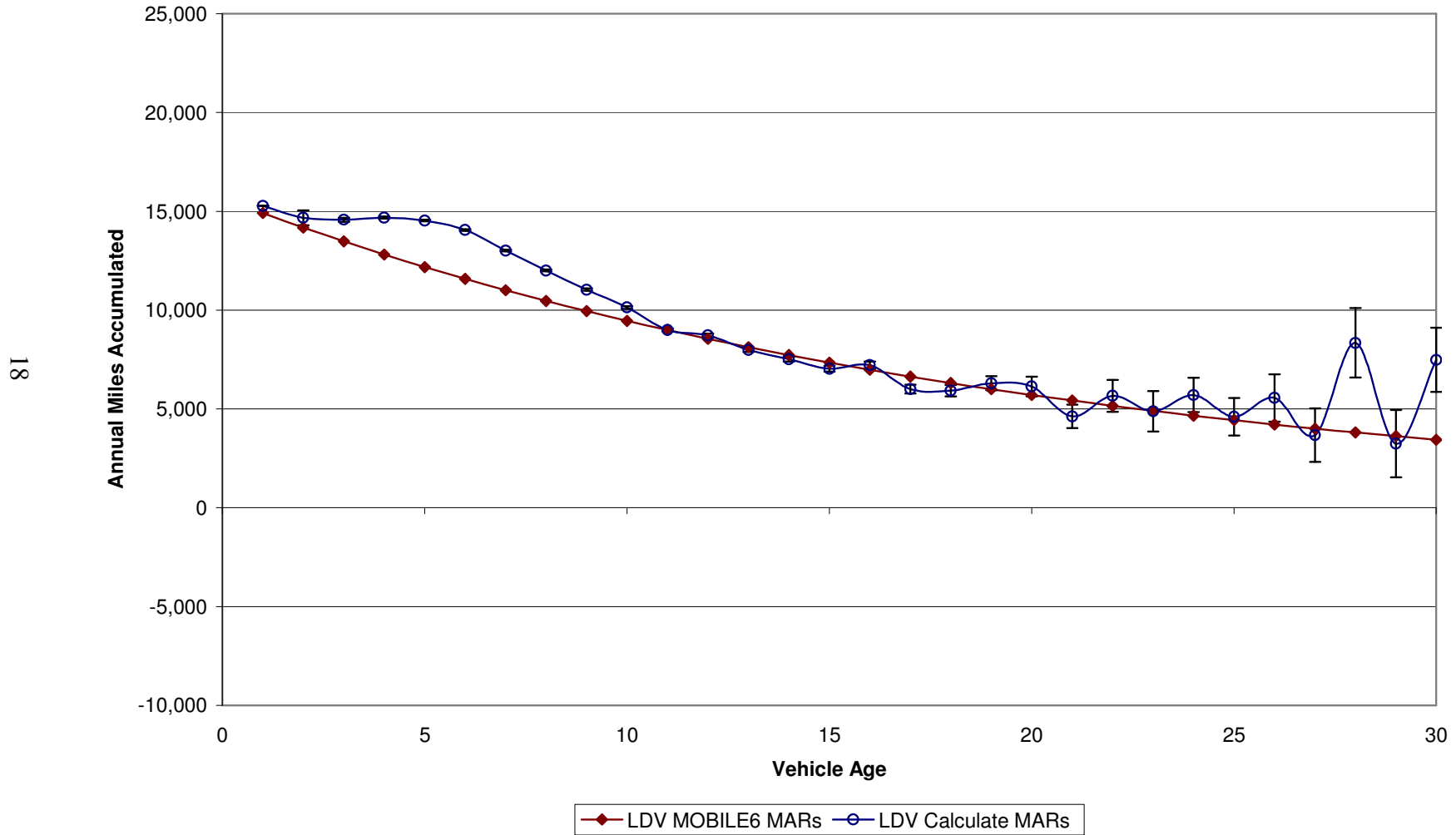


Figure 10. LDT1 Calculated MARs

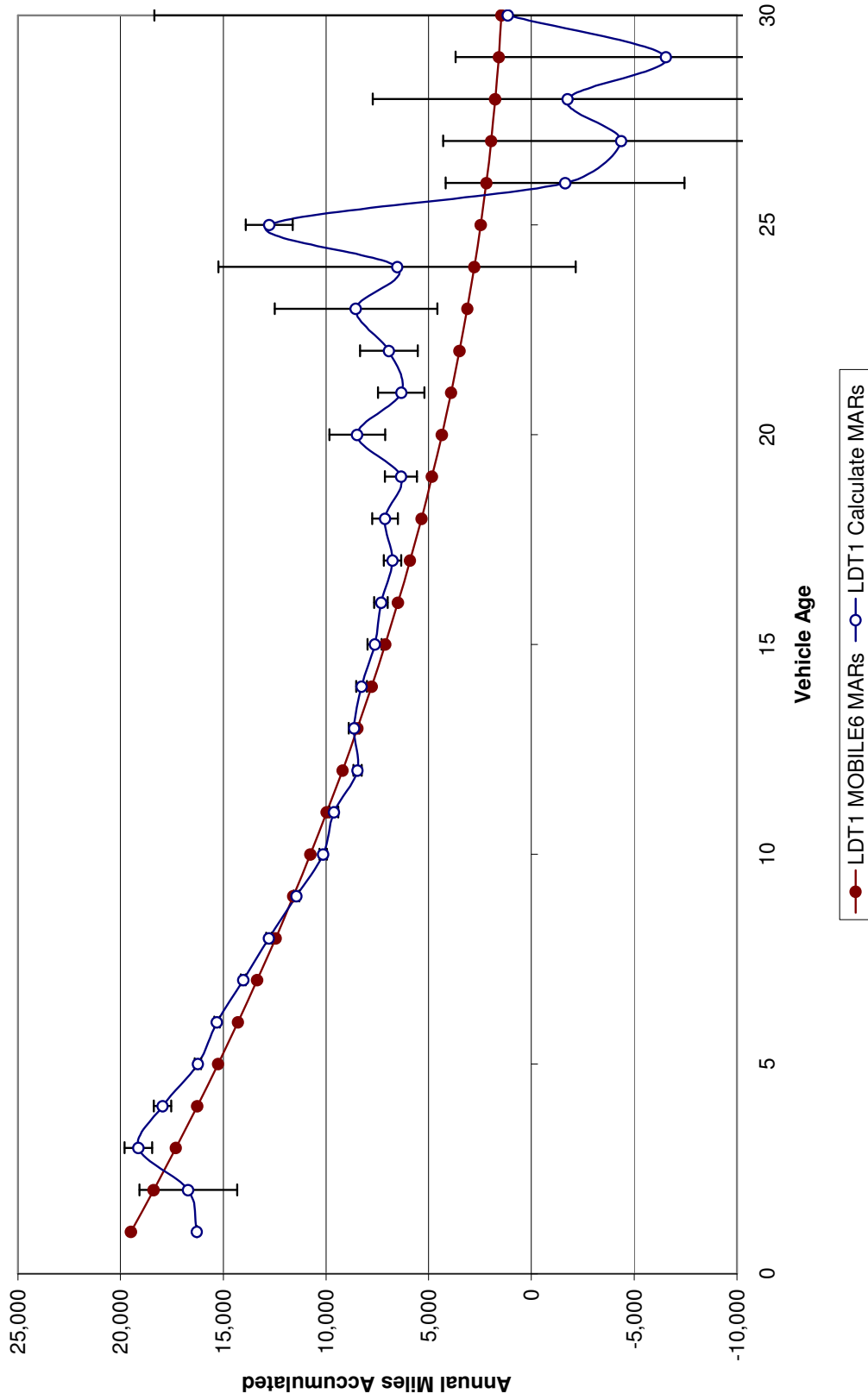


Figure 11. LDT2 Calculated MARs

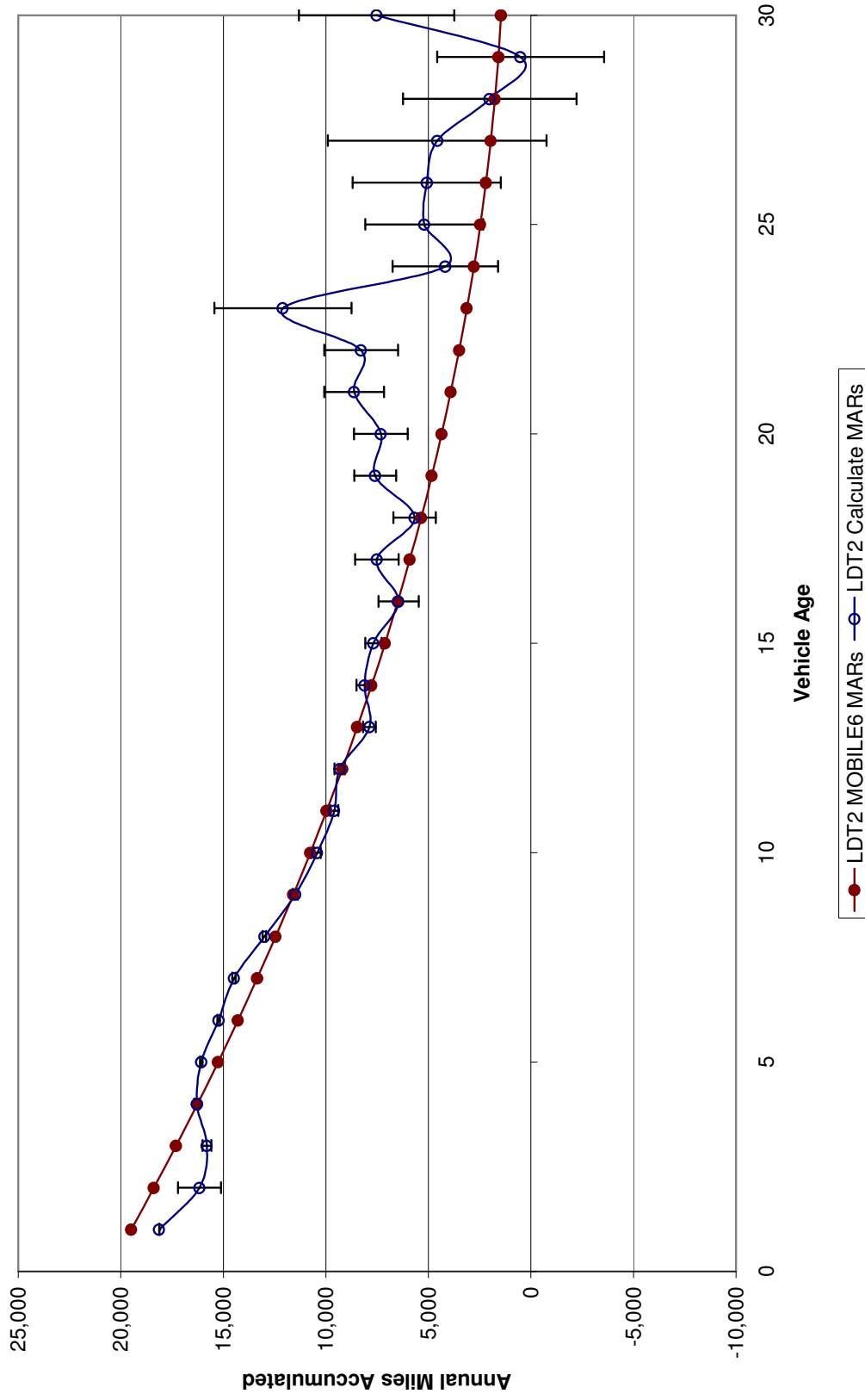


Figure 12. LDT3 Calculated MARs

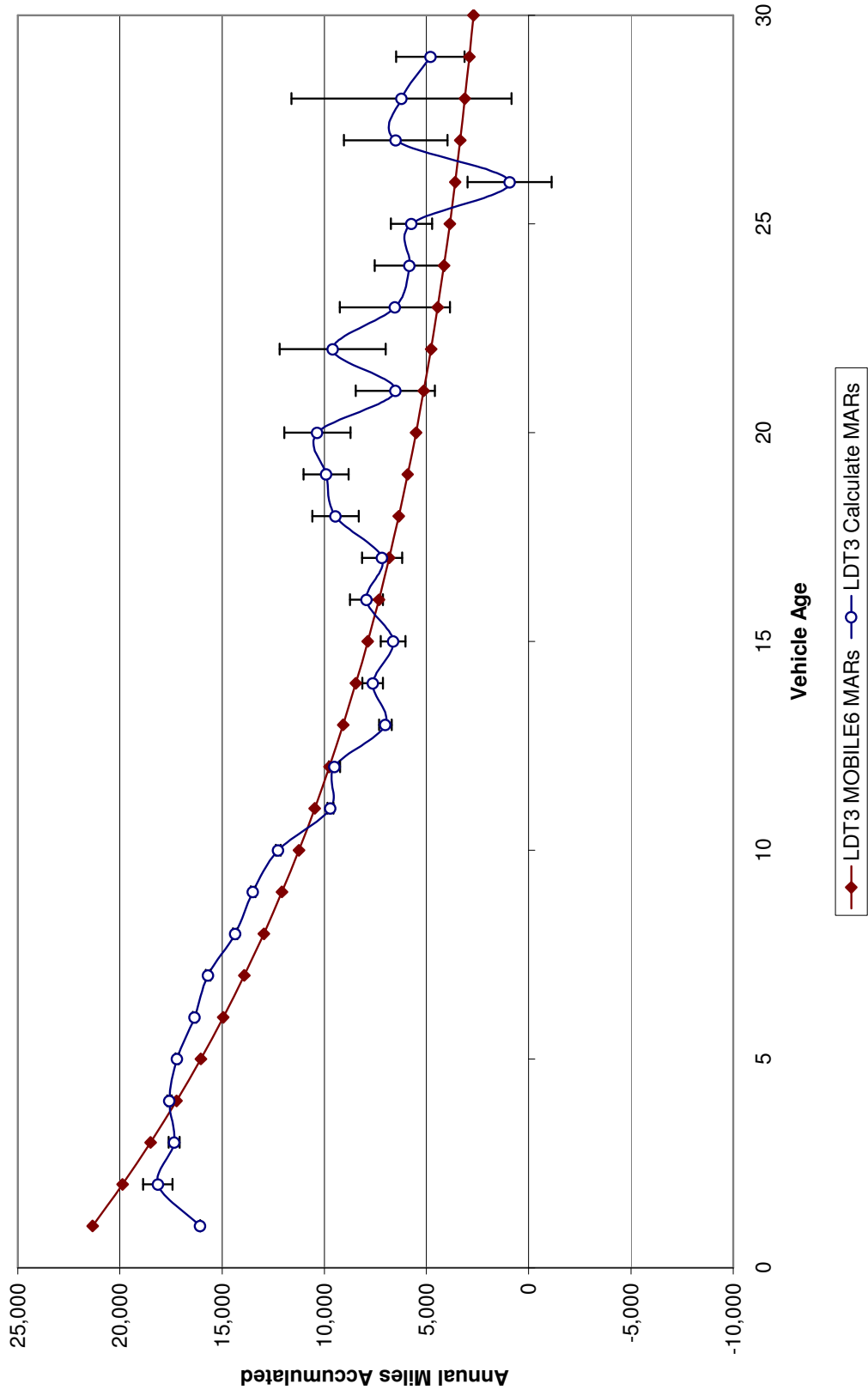


Figure 13. LDT4 Calculated MARs

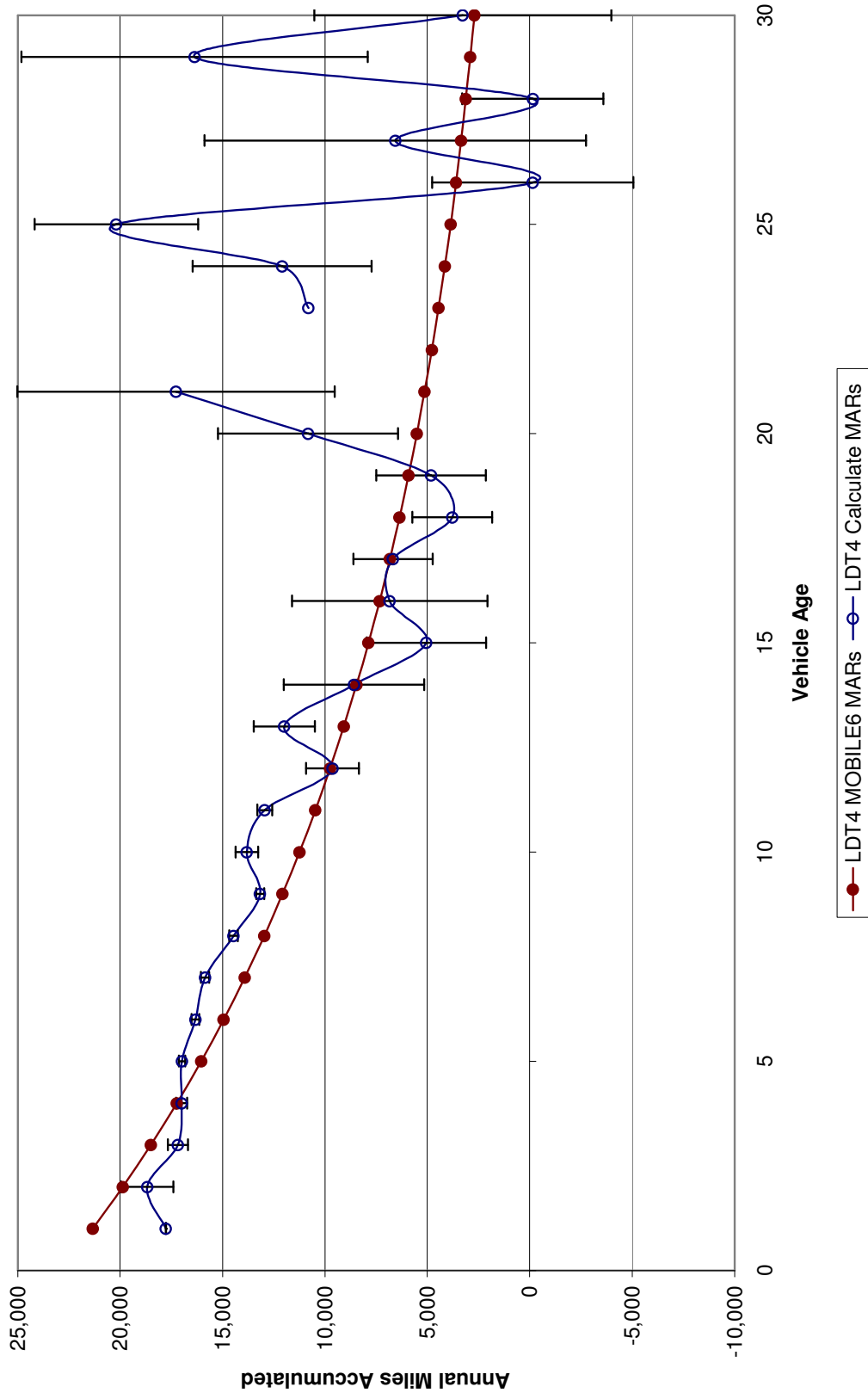


Figure 14. Number of Observations Used

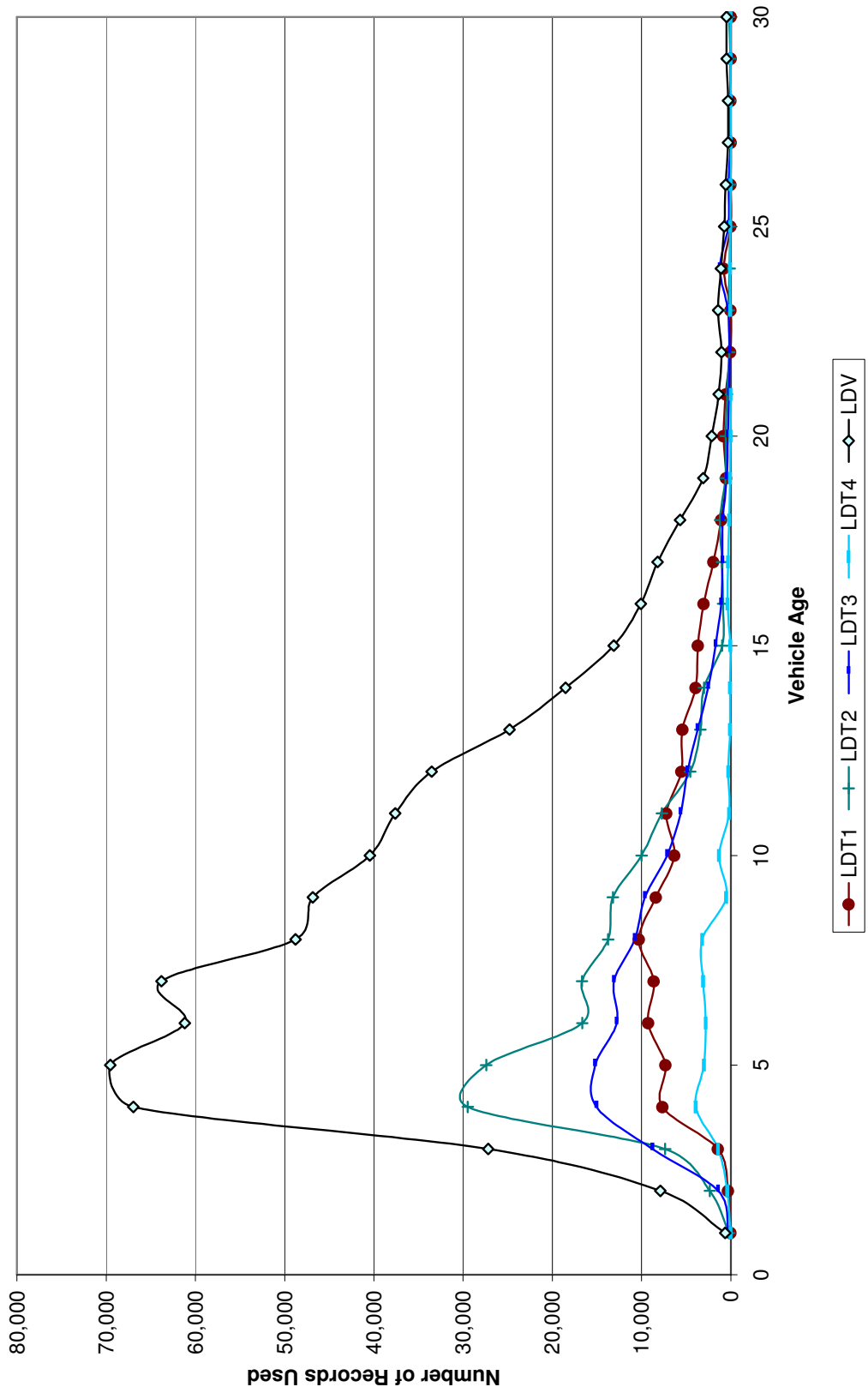


Table 2. MOBILE6 Default & Calculated Houston-Specific MARs

Age	LDV		LDT1		LDT2		LDT3		LDT4	
	Calculated	MOBILE6	Calculated	MOBILE6	Calculated	MOBILE6	Calculated	MOBILE6	Calculated	MOBILE6
1	15,264	14,910	16,293	19,496	18,130	19,496	16,081	21,331	17,766	21,331
2	14,665	14,174	16,706	18,384	16,158	18,384	18,143	19,865	18,679	19,865
3	14,569	13,475	19,135	17,308	15,804	17,308	17,349	18,500	17,177	18,500
4	14,676	12,810	17,952	16,267	16,293	16,267	17,576	17,228	16,996	17,228
5	14,529	12,178	16,243	15,260	16,075	15,260	17,202	16,044	16,979	16,044
6	14,049	11,577	15,304	14,289	15,229	14,289	16,346	14,942	16,328	14,942
7	13,012	11,006	14,025	13,352	14,477	13,352	15,690	13,915	15,853	13,915
8	12,004	10,463	12,781	12,451	12,992	12,451	14,359	12,959	14,458	12,959
9	11,035	9,947	11,434	11,584	11,486	11,584	13,486	12,068	13,162	12,068
10	10,132	9,456	10,135	10,752	10,416	10,752	12,254	11,239	13,814	11,239
11	8,993	8,989	9,616	9,955	9,591	9,955	9,702	10,466	12,939	10,466
12	8,719	8,546	8,464	9,194	9,332	9,194	9,501	9,747	9,631	9,747
13	7,983	8,124	8,632	8,467	7,872	8,467	7,019	9,077	11,987	9,077
14	7,516	7,723	8,277	7,775	8,104	7,775	7,626	8,453	8,579	8,453
15	7,029	7,342	7,625	7,118	7,687	7,118	6,635	7,872	5,051	7,872
16	7,211	6,980	7,318	6,496	6,459	6,496	7,933	7,331	6,836	7,331
17	5,998	6,636	6,762	5,909	7,514	5,909	7,172	6,827	6,679	6,827
18	5,923	6,308	7,123	5,356	5,672	5,356	9,446	6,358	3,783	6,358
19	6,294	5,997	6,348	4,839	7,598	4,839	9,917	5,921	4,815	5,921
20	6,134	5,701	8,479	4,357	7,316	4,357	10,341	5,514	10,824	5,514
21	4,621	5,420	6,336	3,909	8,620	3,909	6,523	5,135	17,273	5,135
22	5,664	5,152	6,939	3,497	8,280	3,497	9,587	4,782	N/A	4,782
23	4,875	4,898	8,542	3,120	12,098	3,120	6,549	4,454	10,806	4,454
24	5,703	4,656	6,534	2,777	4,176	2,777	5,833	4,148	12,090	4,148
25	4,600	4,427	12,757	2,470	5,204	2,470	5,740	3,863	20,184	3,863
26	5,557	4,208	-1,644	2,197	5,082	2,197	930	3,597	-147	3,597
27	3,669	4,001	-4,367	1,959	4,569	1,959	6,508	3,350	6,558	3,350
28	8,339	3,803	-1,760	1,756	2,016	1,756	6,218	3,120	-150	3,120
29	3,235	3,616	-6,546	1,589	509	1,589	4,809	2,905	16,364	2,905
30	7,478	3,437	1,148	1,456	7,522	1,456	N/A	2,706	3,269	2,706