Developing Vehicle Occupancy Factors and Percent of Non-Single Occupancy Vehicle Travel

Final Report

Publication No. FHWA-PL-18-020

April 2019

U.S. Department of Transportation
Federal Highway Administration

Notice

This document is disseminated under the sponsorship of the United States Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof. This report does not constitute a standard, specification, or regulation.

The United States Government does not endorse products or manufacturers. Trade and manufacturers' names appear in this report only because they are considered essential to the object of the document.

Quality Assurance Statement

The Federal Highway Administration (FHWA) provides high-quality information to serve Government, industry, and the public in a manner that promotes public understanding. Standards and policies are used to ensure and maximize the quality, objectivity, utility, and integrity of its information. FHWA periodically reviews quality issues and adjusts its programs and processes to ensure continuous quality improvement.

Technical Report Documentation Page

1. Report No. FHWA-PL-18-020	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Developing Vehicle Occupancy Factors and Percent of Non-Single Occupancy Vehicle Travel		5. Report Date April 2019 6. Performing Organization Code
7. Author(s) Robert Krile, Andrew Landgraf, Elizabeth Slone (Battelle)		8. Performing Organization Report No.
9. Performing Organization Name and Address Battelle 505 King Ave. Columbus, Ohio 43201		10. Work Unit No. (TRAIS) 11. Contract or Grant No. DTFH61-13-D-00012 / 0213
12. Sponsoring Agency Name and Address Federal Highway Administration Office of Highway Policy Information 1200 New Jersey Avenue SE Washington, DC 20590		13. Type of Report and Period Covered 14. Sponsoring Agency Code

15. Supplementary Notes

The project was managed by Task Managers for the Federal Highway Administration, Wenjing Pu and Daniel Jenkins, who provided technical directions.

16. Abstract

Vehicle occupancy factors (VOF) and percent of non-single occupancy vehicle (NonSOV) travel are important considerations for transportation planners and policy makers. A methodology is proposed to estimate VOF and NonSOV based primarily on police records of occupancy from crashes. These data may be biased due to non-representativeness of occupancy in crashes compared to that of all driving. The bias is proposed to be corrected with post-stratification weighting and an occupancy bias correction from historical data.

VOF and NonSOV were estimated for 10 years for all states and urbanized areas with a population of at least 200,000 using national records from the Fatality Analysis Reporting System (FARS). Crash records from individual states were utilized for estimates of seven pilot states and their urbanized areas. Validation checks were conducted for these estimates. The computer code used to generate the estimates is provided.

The development of credible VOFs and NonSOV from crash records was accomplished on this task. There were certain limitations to the approach that could not be immediately overcome and some potential future limitations. With appropriate documentation of these issues, the general methodology is recommended to be implemented with state-based crash records as the primary source, where available, and the FARS system otherwise.

17. Key Words	18. Distribution Statement
Vehicle occupancy factor, non-single occupancy vehicle travel, FARS, SDS, HSIS	No restrictions. This document is available to the public.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	42	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Table of Contents

List of Figures

List of Tables

Introduction

Vehicle occupancy factors (VOF), also called average vehicle occupancy (AVO), are estimates of the average number of occupants in a single vehicle. They are an important consideration for transportation planners and policy makers. They are used to calculate person-miles traveled, set policies for high-occupancy lanes, and derive traffic delays per person. Non-single occupancy vehicle (NonSOV) travel is a measure of the proportion of person trips as well as trips avoided by telecommuting that are not in a single occupancy vehicle. It is an important multi-modal metric which captures travel behaviors that are more efficient than driving alone. NonSOV travel is estimated, in part, through VOFs.

Unlike vehicle counting, which can be efficiently automated for large-scale coverage of the number of vehicles that travel in certain areas at certain times, occupancy counting has traditionally been done via field collection or through surveys. These methods are more generally resource intensive, which limits the scope of the areas and time periods that can be collected in an efficient manner. Alternatively, methods that use already-available police records of occupancy from crashes have been proposed to estimate vehicle occupancy (Gaulin, 1991; Asante et al., 1996; Gan et al., 2008). These methods hold promise but have to overcome the technical challenge that occupancy at the time of crashes cannot be assumed to be representative of occupancy in general.

This report provides the detailed analytical methodology for estimating VOF and NonSOV from police crash records and other relevant information. Where methodological choices are possible or necessary, options and recommendations are provided. The methodological details are accompanied by a draft of actual estimates and their statistical uncertainties. Validation checks are provided for these estimates. Finally, the computer code used to generate the estimates is provided.

The development of credible VOFs from crash records was accomplished on this task. From the VOFs, NonSOV travel was also successfully estimated. There were certain limitations to the approach that could not be immediately overcome and some potential future limitations, but with appropriate documentation of these issues, the general methodology is recommended to be implemented.

Scope of the Estimates

VOFs and NonSOV travel as detailed in this report are estimated as average values. The reported average estimates are accompanied by estimates of their standard errors. These standard errors provide a measure of the degree of uncertainty introduced in determining the factors through a sample of data rather than being able to determine them from an entire population (i.e., perform a census).

VOF estimates apply for combinations of vehicle type, road type, geographic resolution and time period. These categories and their respective reporting levels are detailed in Table 1. In addition to estimation at a unique category, some estimates are aggregated (e.g., all times of day for an urbanized area). NonSOV travel is estimated just for urbanized area by year. The methodology section details the general steps necessary to make estimates. The implementation section lists the specific choices and assumptions that were made to make the estimates with the data available.

Table 1: List of categories and levels for which vehicle occupancy factors are reported.

Category	Name	Levels
Vehicle Type	Vehicle Class Group	Car (FHWA Vehicle Classes 1, 2, and 3) Bus (FHWA Vehicle Class 4) Truck (FHWA Vehicle Classes 5-13)
Road Type	Highway Type	Interstate Highways Non-Interstate National Highway- System
Geographic Resolution	State	The 50 U.S. States and the District of Columbia
	Urbanized Area	U.S. Census (2010) defined urbanized areas with population of 200,000 or more (n=177)
Time Period	Time of Day	Weekday 6AM-10AM Weekday 10AM-4PM Weekday 4PM-8PM Weekend 6AM-8PM Overnight (Any day of the week 8PM – 6AM)
	Year	Calendar Year

Data Sources

A few different data sources were used to estimate VOF and NonSOV. Broadly, the data sources were used to estimate 1) the occupancy of vehicles, 2) the amount of non-vehicle travel, or 3) the bias adjustments needed since crash data is not necessarily representative of the driving population. Some data sources were used for multiple purposes. Following is a brief description of each data source and its use(s).

State crash records, including State Data Systems (SDS) and Highway Safety Information Systems (HSIS)

Every time a crash is reported to police, the information of the crash is recorded in a police accident report (PAR). The PAR includes information on the characteristics of the crash, vehicles, and people involved. The recorded number of occupants in the vehicles can be used to estimate occupancy. Further, information about the vehicle, driver, and time/location of the crash is used to correct for biases that are observed in the data. Some states, such as TX and MD, make these records freely available to researchers online, and they were accordingly downloaded for this analysis. Further, the National Highway Traffic Safety Administration (NHTSA) collects and maintains these PAR's for 34 states in the SDS, which are available at the acceptance of the corresponding states. The analysis in this report was completed with SDS records from twelve different states. FHWA's Highway Safety Information System (HSIS) has recent crash data on seven states, with Maine and California records being used in this evaluation. The state crash records in general have the advantages of being freely available, regularly updated, and having a large sample of vehicles. The disadvantages include the potential unrepresentativeness of the crashes to the driving population and the fact that each state has different coding schemes and do not necessarily collect the same attributes. Additionally, the data quality for the reports could vary from state to state, by time, or by system.

Fatality Analysis Reporting System (FARS)

NHTSA also collects and maintains FARS, which is a census of all U.S. vehicle crashes from 1975 to 2016 that resulted in a fatal injury within 30 days of the crash. As opposed to the SDS, this data is freely available nationwide, but has a much smaller sample size. Further, because it only contains crashes that resulted in fatal injuries, it may be less representative of the driving population. Like the state crash records, this data source was used in the analysis to estimate occupancy, while information about the vehicle, driver, and time/location of the crash was used to correct for biases that are observed in the data.

National Transit Database (NTD)

The Federal Transit Administration (FTA) collects and maintains the NTD, which contains data on all transit systems that receive benefits from the FTA. Information in this database on passenger trips was used for calculating the percent of non-single occupancy travel and the passenger miles and vehicle miles values were used for vehicle occupancy factors.

National Household Travel Survey (NHTS)

The NHTS is used to control for biases. First, it is used to adjust the estimation of vehicle occupancy for a subset of drivers and vehicles in crashes by their prevalence in the population. This is done by estimating the person trips and vehicles miles for certain subpopulations. NHTS contains information on vehicle trips only for privately-operated vehicles (POV), which need not be owned by anyone in the household. Notably, POV's exclude buses, streetcars, taxis, and school buses. Vehicle trips are trips in which the respondent is the driver and the transportation mode is a POV. Second, NHTS is used in combination with crash records from matching years to estimate how the presence of occupants in a vehicle affects the probability of getting in a crash.

FHWA Traffic Volume Trends (TVT) and Highway Statistics Series (HSS)

While NHTS provides the number of person trips or vehicle miles by many different categories, it was not designed to do so for small subpopulations and may be inaccurate in these cases. NHTS also cannot provide the amount of travel that was on highways versus other roads. The TVT and HSS data were used to provide more information on travel by vehicle type, road type, states, month of year, and by urban/rural designation.

American Community Survey (ACS)

The Census Bureau's ACS samples approximately one percent of the U.S. population each year. The survey contains questions on travel mode to work. These data were used to find the number of telecommuting trips as well as other non-vehicle commuting patterns.

The remainder of this report provides overall methodology behind the calculations, details of how the methodology was implemented for the available data, validation of the resulting estimates, and details regarding the deliverable estimates and corresponding computer code. The actual estimates and computer code are included as attachments to this report.

Methodology

Overview of Calculating VOF and Percent of NonSOV Travel

Vehicle occupancy is defined mathematically so that $\Pr(VO = v)$ is the probability that the vehicle occupancy (VO) equals v (for $v = 1, 2, 3, \ldots$). The VOF is the expected value of VO.

$$VOF = \sum_{v} v \ \Pr(VO = v)$$

The summation in the above equation conceptually could include any non-zero integer value, but it is practically limited under the observation that most passenger vehicles have 4 or fewer occupants.

The overall non-single occupancy vehicle travel (NonSOV) is obtained by combining the proportion of non-single occupancy vehicle travel which takes place in vehicles, $NonSOV_{veh}$, with the probability that a mode of transportation for a trip is a vehicle, $\Pr(Vehicle)$. That is, the NonSOV is the proportion of trips not taking place in a vehicle ($1 - \Pr(Vehicle)$) plus the proportion of trips that are in a vehicle *and* where the vehicle has more than one occupant ($\Pr(Vehicle) * NonSOV_{veh}$).

$$NonSOV = (1 - \Pr(Vehicle)) + \Pr(Vehicle) * NonSOV_{veh}$$

For trips taking place in a vehicle, the proportion of non-single occupancy vehicle travel is the number of passenger trips that are not single occupancy divided by the total number of passenger trips. Equivalently, it is 1 minus the proportion of passenger trips that have only one occupant, which is the proportion of vehicle trips with one occupant divided by the expected number of occupants in the vehicle.

$$NonSOV_{veh} = 1 - \frac{\Pr(VO = 1)}{\sum_v v \Pr(VO = v)} = 1 - \frac{\Pr(VO = 1)}{VOF}$$

This simplifies to

$$NonSOV = \left(1 - \Pr(Vehicle)\frac{\Pr(VO = 1)}{VOF}\right)$$

The calculation of both VOF and $NonSOV_{veh}$ require estimation of the probability distribution of vehicle occupancy.

When using crash records, this evaluation found that only private vehicles and trucks permit reliable estimation of $\Pr(VO = v)$ ($v = 1, 2, 3, \ldots$). Buses are not prevalent enough in crash records, and even when present, the occupancy numbers are often capped. This means crash records can be used to estimate all forms of VOF for cars and trucks, but additional data sources are required to estimate VOF for buses. Estimating overall NonSOV for urbanized areas requires knowledge of non-vehicle traffic to estimate $\Pr(Vehicle)$. This estimate must come from other sources than crash records.

The remainder of the methodology section first discusses how to estimate the vehicle occupancy distribution from crash records and then how to estimate VOF for cars and trucks. The final two subsections discuss the methodology for estimating VOF for buses and the estimation of non-vehicle traffic necessary to determining NonSOV travel.

Estimating Vehicle Occupancy Distribution from Crashes

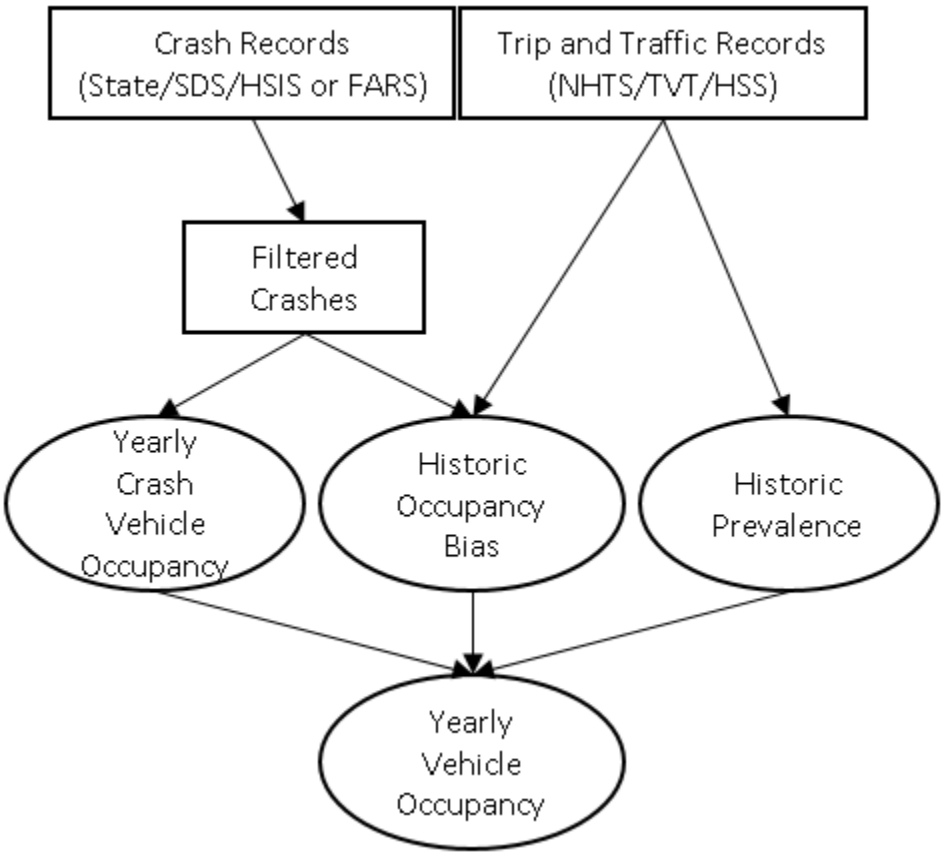

Figure 1: Flowchart of the steps to estimating occupancy in passenger vehicles and trucks from crash records.

Figure 1 is an outline of the process for estimating occupancy in passenger vehicles and trucks from crash records. The source data on the top row has been identified in the previous section and its detailed treatment is discussed in the implementation section to follow. The next two subsections describe how the crash data is filtered and how the occupancy is estimated. Following that, the processes for estimating the historic occupancy and prevalence biases are described. Finally, the pieces are combined to get a final estimate of vehicle occupancy.

Two strategies are employed to make the estimates of vehicle occupancy from crash records as valid as possible:

1. Filtering the data so that it is as consistently representative as possible
2. Correcting for biases that cannot be filtered out

The filtering process is discussed below, but after it is employed, there is still a likelihood that crash records are not representative of the entire driving population. Two specific biases for this non-representative nature of crash data are considered.

First, as previously observed by Chen et al. (2000), the probability of getting in a crash can be affected by the number of occupants in the vehicle. Specifically, Chen et al. (2000) found that 16 and 17-year-olds are more likely to get into a fatal crash when there are more people in the vehicle. The opposite is true for older drivers. They are more likely to get into a fatal crash when driving alone. The vehicle occupancy

distribution for crash data is therefore generally *conditional on there being a crash* and can be denoted as ($\Pr(VO = v|Crash)$). This distribution is not necessarily the same as the desired unconditional occupancy distribution ($\Pr(VO = v)$) required to calculate the VOF. To relate the conditional and unconditional distributions, the following relationship due to Bayes Theorem can be used,

$$\Pr(VO = v) = \Pr(VO = v|Crash)\frac{\Pr(Crash)}{\Pr(Crash|\, VO = v)}$$

If the probability of getting in a crash changes depending on how many people are in the car, as is the case in the referenced Chen et al (2000) work, then the naïve estimates of occupancy distribution from unadjusted crash data will be incorrect. The ratio $\Pr(Crash|\, VO = v)\ /\ \Pr(Crash)$ is the probability of getting in a crash, conditional on v occupants being in the vehicle, divided by the probability of getting in a crash, regardless of how many occupants are in the vehicle. This quantity, which is the inverse of the final term in the equation, is defined to be the ***occupancy bias***. If it can be satisfactorily estimated, the bias effect it has can be removed from the equation to get the true occupancy distribution.

A second bias in determining the occupancy distribution from crash data stems from the fact that the drivers, vehicles, and time/locations of crashes may not be representative of a random sample of drivers, vehicles, and time/locations of vehicle traffic on the road. To account for the fact that some subpopulations will be over or under represented in the crash data, a post-stratification by subpopulations ($SubPop$) is performed.

$$\boldsymbol{Pr(VO = v) = \sum_{SubPop} \Pr(VO = v|SubPop, Crash)\ \Pr(SubPop)\frac{\Pr(Crash|SubPop)}{\Pr(Crash|SubPop, VO = v)}} \quad (1)$$

In the above equation, $\Pr(SubPop)$ is the ***prevalence*** of the subpopulation, which is the proportion of vehicle miles driven by the subpopulation. Subpopulations consist of combinations of driver, vehicle, and crash characteristics. Variables that are considered are listed in Table 2.

Table 2: List of variables that are used to classify crashes into subpopulations.

Data Type	Variables
Factors required for reporting	Road type, time period, state, urbanized area More detailed vehicle type (Passenger car, light truck, motorcycle, truck)
Crash	Location (urban/rural, Census region/division, state, metro size), Time (season of year)
Driver	Age (grouped), gender
Interactions	Between all the required factors and the other crash and driver factors

One example of prevalence bias can be seen with gender. Table 3 shows the proportion of vehicle miles that were driven by each gender in 2009, as estimated by the NHTS, as well as the proportion of fatal crashes in 2009 and the corresponding estimates of VOF, as reported in FARS. Fatal crashes are much more likely to involve male drivers (72%) than the general proportion of vehicle miles driven (60%). The estimated VOF without reweighting by prevalence (28% * 1.51 + 72% * 1.38 = 1.41) will underweight the VOF of female drivers and overweight the VOF of male drivers. The net result is an estimated VOF, at 1.41, which is biased compared to a correctly weighted VOF estimate (40% * 1.51 + 60% * 1.38 = 1.43).

Table 3: A comparison of the prevalence of male and female drivers on the road versus those involved in fatal crashes.

Driver Gender	Proportion of 2009 Fatal Crashes (FARS)	Proportion of 2009 Vehicle Miles (NHTS)	2009 FARS VOF
Female	28%	40%	1.51
Male	72%	60%	1.38

To estimate the distribution of vehicle occupancy for passenger vehicles and trucks from crashes, three components are necessary for each subpopulation

- The occupancy in crashes ($\Pr(VO = v|SubPop, Crash)$)
- The prevalence ($\Pr(SubPop)$)
- The occupancy bias ($\Pr(Crash|\, SubPop, VO = v)$ / $\Pr(Crash|SubPop)$)

The methodology for estimating each quantity is provided below, after describing how crash data is pre-processed.

Crash data preparation

If measured data have consistent and reproducible biases, these biases can potentially be removed by mathematical adjustment. Some measured data are not well suited to this type of bias adjustment and such records are recommended to be filtered out of the crash data subsequently used. For example, FARS data records imply that at least one person must have died as a result of the crash. Since multiple vehicles can be part of a crash, and a crash can impact those outside a vehicle, this means that each vehicle may or may not have any deaths in it. Further, if there was a death in the vehicle, it could have been the driver, a passenger, or both. Table 4 shows occupancy statistics for passenger vehicles in the FARS data from 1998 to 2015. As a reference, the national average VOF for cars is 1.67 according to FHWA's Transportation Performance Management guidance, which uses the 2017 NHTS (https://www.fhwa.dot.gov/tpm/guidance/avo_factors.pdf). When only a non-driver died, there would have had to have been at least two occupants in the vehicle, which makes the occupancy biased much higher. Consistent with Heidtman et al. (1997), the methodology employed here is to remove records where only a non-driving passenger was a fatality.

Table 4: Passenger vehicles in FARS from 1998-2015.

Casualties	# Vehicles	VOF	% NonSOV
None	277,310	1.52	56.0
Driver (at least)	360,170	1.37	44.8
Non-Driver(s)	87,234	2.96	100

Several other types of vehicles and crashes were removed from both FARS and the state crash records:

- Missing occupancy
- No information on the driver or crash
- Multiple drivers for a vehicle or duplicate vehicle information
- Parked cars (usually have occupancy of 0)
- Pedestrian and bicycle records

Estimating vehicle occupancy distribution in crashes

For both FARS and state crash records, the strategy for estimating the distribution of passenger vehicle and truck occupancy is the same. From Equation (1),

$$\Pr(VO = v|SubPop, Crash)$$

is estimated for $v = 1, 2, \ldots$. As previously detailed, v is limited to taking values $VO = 1, VO = 2, VO = 3$, and $VO \geq 4$.

A simple way to approach this problem is to use empirical estimates of the proportion of crashes for each subpopulation that has v occupants. If the number of subpopulations is small compared to the number of crashes, this is a feasible solution. However, correctly estimating the overall vehicle occupancy requires estimates for this probability for every subpopulation where $\Pr(VO = v|SubPop, Crash)$ changes. For example, if the occupancy is similar for both male and female drivers, then driver gender does not need to be included in the subpopulations. However, if occupancy does differ in a significant way between males and female drivers, then gender should be included. Example subpopulations may include all unique combinations of road type, day of week, time of day, vehicle type, driver age, and driver gender, as outlined in Table 2. A mathematical model of the occupancy as a function of the subpopulation is recommended since it can reduce the variability that would result from a large number of empirical estimates, many of which would come from sparse data.

This is a multinomial regression problem and the probabilities must sum to 1 for every $SubPop$. Say that subpopulations consist of a combination of p variables $(X_1, \ldots, X_p)$ and that $SubPop_i$ has values $X_1 = x_{1i}, X_2 = x_{2i}, \ldots, X_p = x_{pi}$. Continuing the example from above, there would be $p = 6$ variables with X_1 representing road type, X_2 representing day of week, and so on. For $SubPop_i$, x_{1i} would represent the road type for the ith subpopulation, which may be interstate highway, for example. As a technical aside, since the variables are all categorical, x_{ji} is a vector composed of all 0's except for a 1 for the category that the j'th variable belongs to.

Using a binary logistic regression approach, for each v, a model can be fit of the form

$$\Pr(VO = v|SubPop_i, Crash) = f_v(x_{1i}, x_{2i}, \ldots, x_{pi}).$$

One form of $f_v(x_{1i}, x_{2i}, \ldots, x_{pi})$ on the simple end of the spectrum would be standard linear logistic regression (McCullagh and Nelder, 1987):

$$f_v(x_{1i}, x_{2i}, \ldots, x_{pi}) = \sigma\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\right),$$

where σ is the logistic sigmoid function, which constrains the probability estimates to be between 0 and 1. Models including ordered logistic regression, LASSO regularized multinomial regression (Tibshirani, 1996), multi-level logistic regressions, and random forest (Breiman, 2001) were all evaluated. Multi-level logistic regressions performed best generally, as will be shown in the implementation section (Table 7) and had the additional benefit of providing the possibility of drawing approximate posterior samples to help in uncertainty quantification.

The estimation is done separately for each data source and year, including FARS, even though it has a smaller sample size. Validation was done by training on 75% of the crashes and predicting on the remaining 25%. The models were compared based on the log-likelihood of their predictions. All available

variables from Table 2 were included in the model. This form of modeling allows for estimates for all subpopulations that may have an occupancy-related effect, even if it is small. The interactions ensure that the estimates vary for the different levels that are reported in the deliverable, even if the difference is small.

Estimating prevalence

Prevalence for a subpopulation is the proportion of vehicle miles that are traveled by that subpopulation. The NHTS has information on the prevalence of many of the subpopulations of interest. For example, it can be used to estimate the number of vehicle miles driven by a number of driver demographics (gender, age, location of home), vehicle characteristics (type of vehicle, age of vehicle), and time characteristics (day of week, month of year, time of day). More precise traffic data for a subset of characteristics can be found with the Traffic Volume Trends (TVT) and the Highway Statistics Series (HSS). These data sources provide vehicle miles by state, urban/rural area, vehicle type, and route type.

When estimating prevalence for many subpopulations, the NHTS is likely to give highly variable estimates (and often observed values of 0) especially for some smaller subpopulations due to the limited sample size. To get reliable estimates of the prevalence for small subpopulations, raking, also called iterative proportional fitting (IPF), was employed. Raking also allowed combining and refining the prevalence of the NHTS with the TVT and HSS prevalence estimates.

Raking is used when the marginal distribution of a number of variables is known, or even the joint distribution for some crossed variables, but the full joint distribution of all variables is not. For example, assume it is known that males make up 60% of traffic, with females making up 40%, and that 25% of traffic occurs on weekday mornings, with 75% occurring at other times of day, but it is not known what percentage of the weekday morning traffic is male versus female. In this simple example, an estimate of 15% male (25%*60%) and 10% female (25%*40%) would satisfy the requirements, but the solution is more involved with many variables. Raking is a way to mathematically estimate subpopulation prevalence, so it matches, to the extent possible, all the known distribution margins simultaneously.

Raking requires initializing the joint estimation with a seed matrix, which provides the a priori expected proportion of traffic for each subpopulation. A seed matrix of all ones was used for this analysis, which gave equal weight and no a priori information to the estimates. For the marginal distributions, available information from HSS and TVT was used first and then supplemented by added information from NHTS that is not available in the other sources.

- HSS, tables VM-2 and VM-4, provided the vehicle miles travelled (VMT) for combinations of state, route type, urban/rural, and vehicle type.
- TVT provided the VMT for combinations of state and season of the year.
- TVT also provided the VMT for combinations of urban/rural, route type, and season of the year.
- NHTS VMT was used for every other bivariate combination of the variables. For example, combinations of driver age and season of the year were included from NHTS, but vehicle type and urban/rural were not, since that combination was already present in the TVT. The nine US Census divisions were the most detailed geographic area used with NHTS due to small sample sizes for some states. A total of 40 bivariate combinations were included from NHTS.

The raking was implemented to ensure that the marginals of the raked estimates matched the HSS and TVT provided marginals. Prevalence was calculated for the subpopulations listed in Table 2. Since the NHTS is only updated every eight years between 2001 and 2017, this methodology assumed that the NHTS-based prevalence stayed relatively constant from year to year. For example, if males made up

60% of the traffic in 2009, they were expected to continue to make up 60% of the traffic in the near future (2010 to 2016).

Estimating occupancy bias

The occupancy bias for a subpopulation in a vehicle with v occupants can be calculated by dividing the proportion of crashes in the subpopulation that had v occupants by the proportion of the prevalence in the subpopulation that had v occupants, as shown in the following equation, where vehicle miles travelled (VMT) is being used for prevalence in this example.

$$\text{Bias}(SubPop, VO = v) = \frac{\Pr(Crash|SubPop, VO = v)}{\Pr(Crash|SubPop)} = \frac{\#(Crash, SubPop, VO = v)}{\#(Crash, SubPop)} \frac{VMT(SubPop)}{VMT(SubPop, VO = v)}$$

This bias is calculated with the number of crashes for each subpopulation and the number of occupants (from the state crash records or FARS) and the vehicle miles for each subpopulation and number of occupants (from the estimation of the prevalence). For maximum defensibility of the bias estimates, the time periods for the prevalence and crash data should coincide. In the case of the NHTS, though, prevalence estimates are only available every eight years between 2001 and 2017. The approach used for this analysis was to use years that overlap in the crash and prevalence data and then to assume that the estimated biases would still apply into the future. For example, the 2009 NHTS and 2009 state crash records were used to estimate the occupancy biases for each subpopulation (e.g. young male drivers) in 2009, but then were assumed to remain the same for the next several years.

Similar to the occupancy distribution from crashes, the data can be used to directly get empirical estimates of occupancy bias. However, the occupancy bias must be calculated for every subpopulation where it varies, and the empirical bias estimates will be unreliable for small subpopulations. For example, if a subpopulation had no crashes in 2009 for a given number of occupants in the data (even though it is not believed no crashes occurred in truth), the empirical bias estimate will be 0, which will cause the estimate of the VOF to be undefined because it will involve dividing by 0. Further, there may be some subpopulations where the data did not show any crashes in 2009, regardless of the number of occupants, in which case it would be impossible to make the empirical estimate of bias at all.

To avoid having to estimate the occupancy bias directly, a methodology that is commonly used to model disease and other rates was applied. The quantity $\#(Crash, SubPop, VO = v)$ is treated as the response and the other three elements of the bias equation as the known exposures. Since the number of crashes is a count, a Poisson distribution is assumed, and the other three elements are treated as an offset in the Poisson regression. Similar to the occupancy distribution estimation, the occupancy is broken into four groups ($VO = 1, VO = 2, VO = 3$, and $VO \geq 4$) and fit to a model for each group such that $\#(Crash, SubPop_i, VO = v)$ is Poisson distributed with the log of the mean equal to

$$\log\left(\frac{VMT(SubPop_i, VO = v) \times \#(Crash, SubPop_i)}{VMT(SubPop_i)}\right) + f_v(x_{1i}, x_{2i}, \dots, x_{pi}).$$

The estimated bias for subpopulation i is $e^{f_v(x_{1i}, x_{2i}, \dots, x_{pi})}$.

Similar to estimating occupancy, several models were considered, including Poisson regression, LASSO-penalized Poisson regression, and multi-level Poisson regression. Multi-level Poisson regression and standard Poisson regression performed best generally, as will be shown in the implementation section. Standard Poisson regression was ultimately selected, because it was simpler and did not require incorporation of uncertainty into the bias estimates.

The occupancy bias estimation was done separately for each crash data source and NHTS year (2001, 2009, and 2017). Since FARS has smaller sample sizes, multiple years of FARS records were modeled around each NHTS year. Validation was done by training on one year and predicting the bias and the counts on the following NHTS year. The models were compared based on the weighted mean squared error of the predicted bias and the log-likelihood of the predicted counts. All available variables from Table 2 were included in the model, with interactions with the occupancy level.

As an example of the benefit of the modeling approach, Figure 2 compares prediction of FARS occupancy bias in 2009 with both the empirical bias in 2001 and a linear Poisson regression model fit using 2001 data. The model follows the diagonal line much more closely, giving more accurate and less variable predictions of bias.

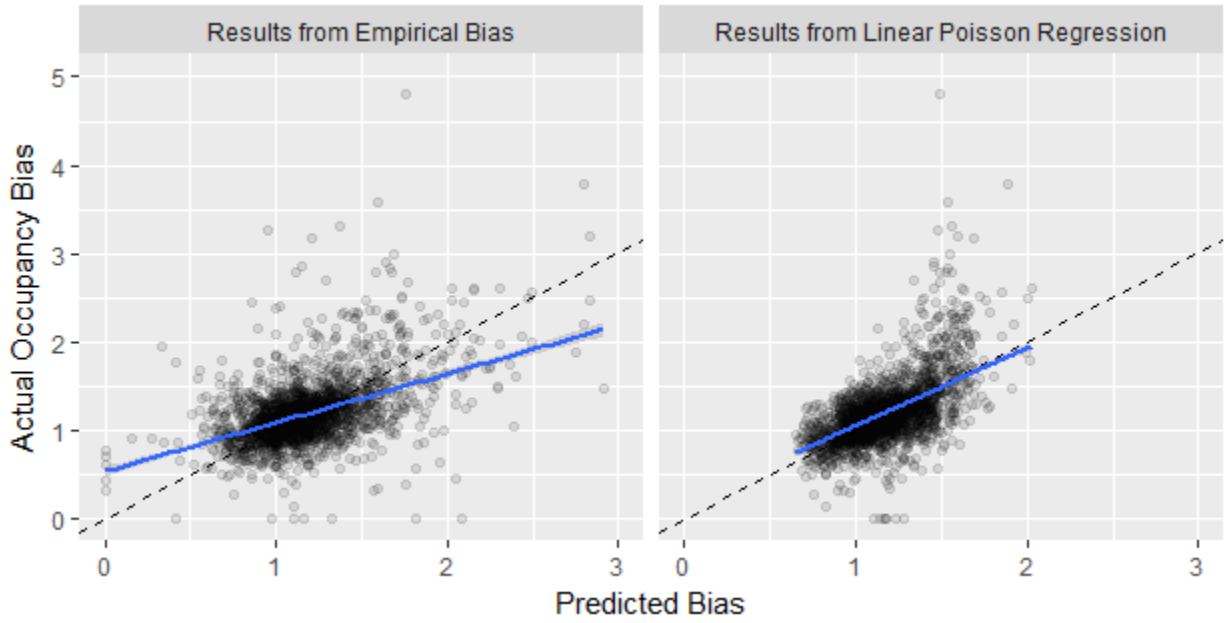

Figure 2: Predicted occupancy bias versus actual occupancy bias in FARS in 2009 for subpopulations with over one billion vehicle miles. The plot on the left uses the empirical bias in 2001 to estimate the bias in 2009. The plot on the right used the proposed linear Poisson regression model fit on 2001 data to predict the bias in 2009. The blue line is the ordinary regression line and the dashed line indicates where predicted equals actual.

Correcting for occupancy bias has the effect of adjusting the crash occupancy estimate to more closely match the occupancy estimate from the source of the prevalence (NHTS). This means that using 2009 crash records and 2009 NHTS data to estimate occupancy bias will produce vehicle occupancy estimates that more closely match NHTS for 2009.

It is possible that NHTS will change what type of data is collected in the future, which would limit the ability to use it as a source of bias correction. The example above has shown that FARS occupancy bias is consistent over an eight-year time span (2001 to 2009). This suggests that the occupancy bias estimates using the 2017 NHTS will be valid for several years before some sort of re-validation would be necessary.

Generating the combined estimate

The final estimation of the vehicle occupancy distribution is accomplished by substituting the estimates of crash occupancy, prevalence, and occupancy bias into equation (1). Additionally, the probabilities are normalized so they sum to 1. That is,

$$\Pr(VO = 1) + \Pr(VO = 2) + \Pr(VO = 3) + \Pr(VO \geq 4) = 1.$$

Vehicle Occupancy Factor for Passenger Vehicles and Trucks

The calculation of VOF involves the estimates of occupancy from crashes, prevalence, and occupancy bias.

$$VOF = \sum_{v \in \{1,2,3,4+\}} E(VO|VO = v) \sum_{SubPop} \frac{\Pr(VO = v|SubPop, Crash)\Pr(SubPop)}{\text{Bias}(SubPop, VO = v)}.$$

This demonstration was completed with FARS data for every state as well as with selected state crash records. FARS and state crash records were separately used to create independent estimates. Each individual VOF estimate may be for a specific condition (e.g., passenger vehicle, morning, NHS, in a state for a year) or for an aggregate of conditions (e.g., all of one state for one year). The estimated VOF includes only the relevant subpopulations for the condition of interest. Additionally, in the above equation, $E(VO|VO = 1) = 1$, $E(VO|VO = 2) = 2$, $E(VO|VO = 3) = 3$, and $E(VO|VO = 4+) = E(VO|VO \geq 4) = 4.5$. The use of 4.5 as the expected value for all records with occupancy 4+ is based on averages observed in NHTS and the crash records.

In addition to the direct VOF estimate, a standard error of the estimate is derived, which provides a measure of the uncertainty in the estimate. To estimate standard errors, the VOF is simulated many times and then a standard deviation is calculated for these simulated estimates. Each simulated VOF is a draw of an approximate Bayesian posterior sample from the crash occupancy distribution, $\Pr^{(l)}(VO = v|SubPop, Crash)$, for $l = 1, \ldots, 50$. This results in 50 simulations from the posterior distribution,

$$VOF^{(l)} = \sum_{v \in \{1,2,3,4+\}} E(VO|VO = v) \sum_{SubPop} \frac{\Pr^{(l)}(VO = v|SubPop, Crash)\Pr(SubPop)}{\text{Bias}(SubPop, VO = v)},$$

for $l = 1, \ldots, 50$. The estimate of the standard error of VOF is the standard deviation of these 50 simulations. This standard error methodology does not incorporate the uncertainty in estimating the prevalence or occupancy bias. Doing so would likely involve the replication weights of the NHTS, which would be computationally intensive. Additionally, the relationship between occupancy bias, prevalence, and crash occupancy would need to be evaluated. This methodological development could represent a future enhancement of the overall VOFs but was beyond the scope of the current demonstration project.

For urbanized areas that overlap with multiple states, the approach was to estimate the VOF of the portion of the urbanized area in each state separately. Then the estimates from all the states were averaged, weighted by the proportion of the urbanized area's population that is in each state. Since the states' records can be considered independent, the combined squared standard error is equal to the sum of the states' squared standard errors, weighted by the squared proportion of the population in each state.

Vehicle Occupancy Factor for Buses

Due to the lack of crashes involving buses, especially in FARS, vehicle occupancy factors for buses are estimated with a different approach. The National Transit Database (NTD) provides passenger miles traveled (PMT) and vehicle revenue miles (VRM), which are divided to get an overall VOF for an area of interest.

Three potential data sources were evaluated:

1. The Annual Database UZA Sums (https://www.transit.dot.gov/ntd/data-product/2016-annual-database-uza-sums) has PMT and VRM for each urbanized area, but no mode to differentiate buses from other methods of transit
2. The Annual Database Service (https://www.transit.dot.gov/ntd/data-product/2016-annual-database-service-0) has PMT, VRM, mode, and time period associated with each agency, but not detailed enough state and city information.
3. Service (https://www.transit.dot.gov/ntd/data-product/2016-service) data contains PMT, VRM, mode, and state and city information associated with each agency, but no time period information.

The third ("Service") data source was selected because it had enough information to make estimates at the state and urbanized area level. However, it does not include enough information to estimate at the time period level. None of the data sources allowed estimation of VOF for different route types.

Records with zero or missing PMT values were excluded from analysis. In addition, only records pertaining to buses, commuter buses, rapid bus transit, and trolley buses were included in the analysis (i.e. records with mode value equal to "CB", "MB", "RB", or "TB").

The VOF for buses in a particular urbanized area or state and year was computed with the following equation from the queried data, which follows the guidance of FHWA's Transportation Performance Management (https://www.fhwa.dot.gov/tpm/guidance/avo_factors.pdf),

$$\mathrm{VOF}_{\mathrm{buses}} = \frac{\sum_{r=1}^{R}[PMT]_r}{\sum_{r=1}^{R}[VRM]_r},$$

where r is a record in the queried data, R is the total number of records in the urbanized area or state, $[PMT]_r$ are the passenger miles traveled in a year for the data record r, and $[VRM]_r$ are the vehicle revenue miles in a year for the data record r.

A major limitation of this methodology is that it only includes transit systems that are in the NTD, which include systems that receive benefits from the FTA. Therefore, the estimates may not generalize to all bus traffic.

Estimating the Proportion of Non-Single Occupancy Vehicle Traffic

The NonSOV travel requires the VOF calculations above, as well as the proportion of passenger trips that are not from vehicles. A preferred source for this latter estimate is the ACS which is an adequately large sample and is updated yearly. The ACS provides the mode of transportation (including telecommuting) that people use for commuting trips. The NonSOV travel must be based on more than just commuting trips, though, so ACS data for percentage of non-vehicle travel in large MSAs was compared to a similar calculation from the NHTS, which takes into account all trips, not just commuting, but has the disadvantage of only being conducted periodically and having a relatively small sample. Figure 3 shows the relationships between the non-vehicle travel on commuting trips (from ACS) to the non-vehicle travel on all trips (from NHTS) at the metropolitan statistical area (MSA) level. In Figure 3, each point is one of the 50 large MSAs in the US that is included in the NHTS in 2009 and 2017. A regression line is also

plotted, assuming a log transformation on the x-axis. The relationship between the two measures is similar for the two years.

Other variables, such as population size and the MSA's density, were evaluated to see if they would improve the fit, but neither did. As a result of this evaluation, the ACS estimate of non-vehicle travel for the urbanized area and year was selected as the appropriate input to the overall NonSOV estimate.

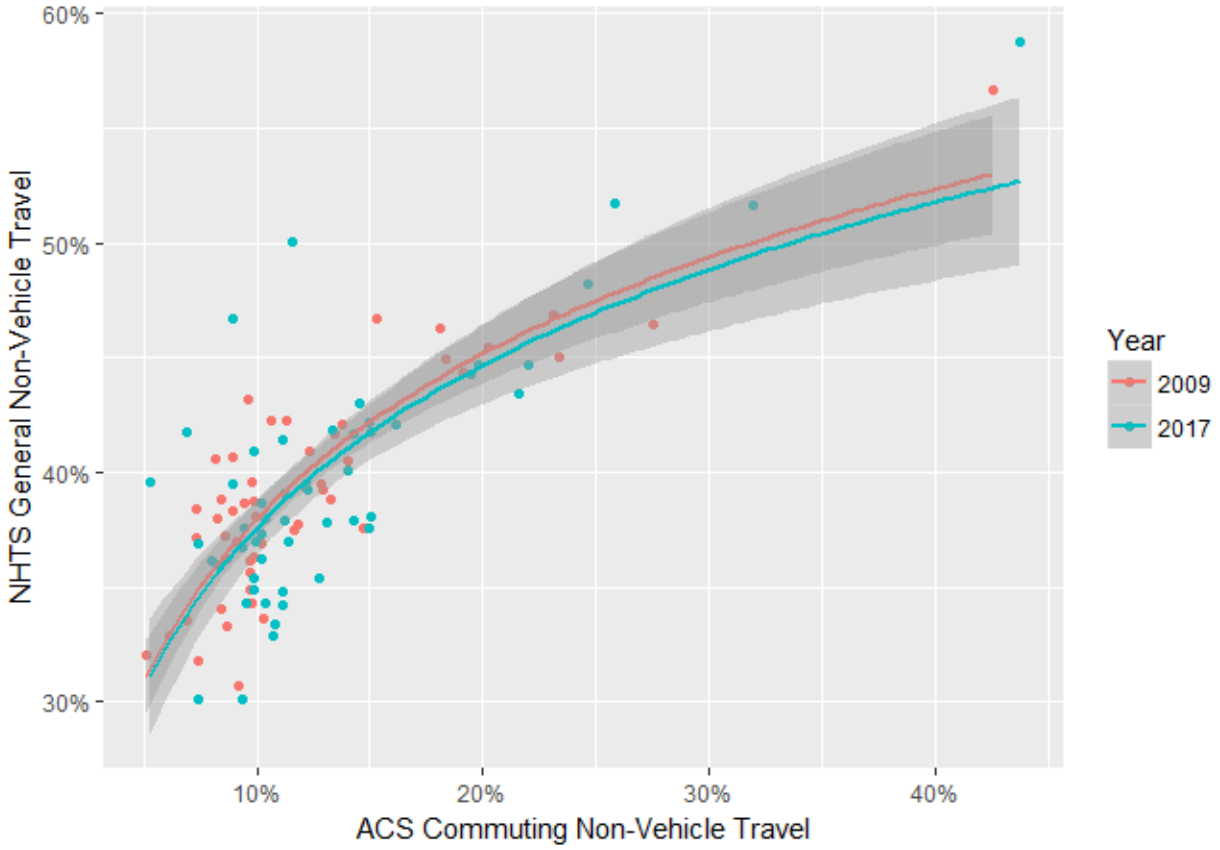

Figure 3: Comparison of the percent of non-vehicle commuting trips from ACS and non-vehicle general trips from NHTS.

NonSOV travel was estimated for a given urbanized area by the previously given equation, combining the occupancy distribution and the estimated proportion of vehicle traffic,

$$NonSOV = \left(1 - \Pr(Vehicle)\frac{\Pr(VO = 1)}{VOF}\right)$$

where

$$\Pr(VO = 1) = \sum_{SubPop} \frac{\Pr(VO = 1|SubPop, Crash)\Pr(SubPop)}{\text{Bias}(SubPop, VO = 1)}.$$

Uncertainty estimates for $\frac{\Pr(VO=1)}{VOF}$ can be obtained by the same method discussed as used for VOF, but the addition of the non-vehicle estimation makes the overall estimation of NonSOV uncertainty more challenging. NonSOV uncertainty estimates were not produced in this evaluation. Estimates from state

crash records of urbanized areas that overlap with multiple states used the same method as used for VOF, weighting the vehicle NonSOV estimates by the proportion of the urbanized area's population that lives in each state.

Implementation Details

The implementation of the proposed methodology on specific data required additional considerations which are detailed in this section.

Data Sources

The data sources and all variables considered in the modeling process are shown in Table 5 and Table 6. Not every desired variable was available for each data source. In addition, some variables were only available for a subset of the years. The tables further show that more variables were collected and normalized than those listed in Table 1 and Table 2. Many variables were consistent from data source-to-data source, but others, such as car type and road type, required a manual process of matching the source data to the predetermined categories. These tables also list some variables that were ultimately not included in the analyses. Driver race and ethnicity were excluded because they were unavailable in most data sources. Severity of the crash and the weather were examined for the possibility of using them to filter the data but were not used due to the data source-to-data source variability, which would necessitate subjectivity in the filtering process.

Table 5: Variable Availability among NHTS, FARS, Online, and HSIS Datasets.

Variable	NHTS	FARS	MD Online	TX Online	CA HSIS	ME HSIS
Years	'01,'09,'17	'98-'16	'15-'16	'10-'16	'10-'14	'06-'10
Urban/Rural	●	●		●	●	
Road Type	● [a]	●	●	●	●	●
Time /Date	●	●	●	●	●	●
Weather		●	●	●	●	●
Overall Crash Severity		●		●	●	●
Number Vehicles		●	●	●	●	●
Vehicle Type	●	●	●	●	●	●
Vehicle Year (Age)	●	●	●	●	●	
Number Occupants	●	●	●	●	●	●
Number Fatalities		●	●	●	●	●
Driver Fault		●	●	●	●	●
Vehicle Severity		●	●			
Driver Age	●	●	●	●	●	●
Driver Gender	●	●	●	●	●	●
Driver Race	●	● [b]		●		
Driver Ethnicity	● [a]	● [b]				
Driver Injury /Severity		●	●	●	●	●
Latitude/Longitude		● [b]	●	●	● [c]	
City				●		
County			●	●		●

a NHTS: Road type 2009 only, ethnicity 2009 and 2017 only, Urban Area provided in urban size of household
b FARS: Latitude/Longitude missing 1998-2000, Race and Ethnicity provided only in fatalities 2000-2016
c HSIS CA: Latitude/Longitude only available 2010-2011

Table 6: Variable Availability among SDS Datasets by State.

Variable	CA	FL	IA	MD	MT	DE	IL	NE	NJ	NM	PA	VA
Years	'06-'10	'06-'14	'06-'14	'01-'15	'01-'08	'01-'14	'01-'14	'01-'13	'01-'14	'01-'13	'06-'13	'01-'15
Urban/Rural	●	●					●	●		●	● [g]	● [h]
Road Type	●	●	●	● [c]	●	●	●	●	●	● [a]	●	●
Time /Date	●	●	●	●	●	●	●	●	●	●	●	●
Weather	●	●	●	●	●	●	●	●	●	●	●	●
Overall Crash Severity	●	● [b]	●	●	●	●	●	●	●	●	● [g]	● [h]
Number Vehicles	●	●	●	●	●	●	●	●	●	● [f]	● [g]	●
Vehicle Type	●	●	●	●	●	●	●	●	●	●	●	●
Vehicle Year (Age)	●	●	●	●	●	● [d]	● [e]	●	●	●	●	●
Number Occupants	●	●	●	●	●	● [d]	●	●	●	●	● [g]	● [h]
Number Fatalities	●	●	●	●	●	●	●	●	●	●	●	●
Driver Fault	●	● [b]	●	●	●	● [d]	●	●	●	●	● [g]	●
Vehicle Severity		●	●	● [c]	●	●		●		●	●	●
Driver Age	●	●	●	●	●	●	●	●	●	●	●	●
Driver Gender	●	●	●	●	●	●	●	●	●	● [f]	●	●
Race	● [a]	● [b]				● [d]				● [f]		
Ethnicity						● [d]						
Driver Injury /Severity	●	●	●	●	●	●	●	●	●	●	●	●
Latitude/Longitude												
City	●	●	●	●	●		●	●	●	●	●	● [h]
County	●	●	●	●	●	●	●	●	●	●	●	● [h]

a CA: Race available 2009 and 2010 only
b FL: Crash severity, driver fault, and Race 2006-2010 only
c MD: Vehicle Severity missing 2009-2014
d DE: 2001-2004, 2007-2014; Vehicle Age, Occupancy, and Driver Fault 2007-2014; Race and Ethnicity 2010-2014
e IL: Vehicle Age 2001-2003, 2007-2014
f NM: Road class 2001-2011; Number vehicles, Race, and Ethnicity 2012-2013
g PA: 2006-2012; Urban vs. Rural 2006-2012; Overall crash Severity missing 2006-2007
h VA: Urban vs. Rural 2001-2007; Overall Severity missing 2008-2009, 2015; Occupancy 2001-2009, 2013-2015; City 2001-2004, 2008-2013; County 2001-2004, 2008-2011, 2013

Urbanized areas were not included in any of the data sources. For data with latitude and longitude of the crash, the location of the crash was used to assign the urbanized area. For all other sources the city name of the crash was mapped to the urbanized area, using relation files from the 2010 U.S. Census. If a city name was not available, then the county was mapped to the urbanized area also using relation files.

The census relation files only provide the proportion of the land area, population, or housing units in each urbanized area, not the amount of traffic. The proportion of the population was used as a proxy for traffic, with the following adjustment. According to HSS, 70% of traffic takes place on urban roads, however 81% of people live in urban areas, according to ACS. This means the odds of driving in an urban area (70/30) are 1.83 times lower than the odds of living in an urban area (81/19). For each city and county, the odds of living in the urbanized area was reduced by a factor of 1.83 to estimate the adjusted proportion of the traffic that takes place in the urbanized area. Crashes are weighted by their probability of being in the urbanized area for estimation purposes.

Another variable that was required for the VOF estimates is whether the crash took place on a National Highway System (NHS) highway. Only FARS included this as a stand-alone variable, but it was possible to reasonably derive it for crashes with longitude and latitude so long as the coordinates of the crash's location was within 150 feet of any NHS highway.

Finally, all of the crash data sets had at least some missing values. Single imputation was performed using the method of Stekhoven and Buehlmann (2012). This provided the benefit of being able to estimate crash occupancy and occupancy bias using as many data records as possible.

Occupancy from Car and Truck Crashes

The methodology section identified different modeling options that were considered for the occupancy estimation. The selection of the best model was based in part on the validation accuracy of the predicted distribution for 25% held out crashes. Table 7 shows a representative example of the model comparisons using FARS for 2015. With lowest negative log likelihood as the metric and smaller values preferred, the multi-level regression with interactions is the preferred model in this example. A large enough number of similar results led to selection of this model as the basis for all occupancy distribution estimation.

Table 7: Validation of crash occupancy distribution models for FARS.

Model	Negative Log Likelihood
Multi-level Regression w/Interactions	0.1746
Multi-level Regression	0.1747
LASSO	0.1756
LASSO w/Interactions	0.1768
Random Forest	0.1785
Ordered Logistic Regression	0.1795

Geographic attribution of crashes to urbanized areas represented a challenge since the urbanized area was not a directly coded value in the databases. When the geographic definition of the crash location made urbanized area assignment unclear, a randomization approach was employed. For each geographic division (e.g., city or county), the relative proportion of its traffic associated with a particular urbanized area was estimated. For a crash in that geographic division, it was either assigned to the urbanized area or excluded from the urbanized area using a random probability compared against the relative proportion.

In all crash data sources, there are very few occurrences where a motorcycle has 3 or 4 occupants. This very rare occurrence caused some technical issues with the estimation of multi-level models and increased the simulated variance. To deal with this, motorcycles were assigned 0 probability of having 3 or more occupants.

Prevalence

Prevalence estimates were made for the years of the three most recent NHTS surveys: 2001, 2009, and 2017. HSS and TVT data were used from the same years, with two exceptions. The most recent complete HSS data was for 2015, so that was used with the 2017 NHTS data. The oldest TVT year was 2003, so that was used with the 2001 NHTS data. The 2017 NHTS did not have any information on truck traffic, and hence the 2009 NHTS marginals related to trucks were used for information not available from HSS or TVT.

When estimating occupancy, 2001 prevalence was used for years 2005 and prior, 2009 prevalence for years 2006 through 2012, and 2017 prevalence for years 2013 and later.

Occupancy Bias

Occupancy bias is only estimated for years with prevalence estimates: 2001, 2009, and 2017. For FARS-based estimation, several years of crash records were used surrounding each prevalence year. Crashes from 2001 to 2003 were used for 2001 occupancy bias, crashes from 2007 to 2011 were used for 2009 occupancy bias, and crashes from 2014 to 2016 were used for 2017 occupancy bias. For estimation with state data records, only the year of crash records that was closest to each prevalence year was used, as long as it was within two years. If no crash records were available within two years, the bias was not estimated. Occupancy estimation was done with the most recent bias estimate that was made with prior crash records. For example, occupancy estimation for 2011 with FARS data used the 2001 bias estimates, since the 2009 estimates would have included 2011 FARS data.

Model selection for occupancy estimation was based in part on the validation accuracy of the predicted bias and counts for the next bias period. For FARS, crash and prevalence data from 2001 were used to predict the bias from 2009. The results from FARS are in Table 8. The weighted MSE is the mean squared error of the predicted bias, weighted by the number of VMT. The negative log likelihood is for a Poisson distribution, predicting the number of crashes with a certain number of occupants given the offset term, which includes the total number of crashes, and the proportion of VMT with that number of occupants. The models “by occupant” assume the relation between the variables and bias varies by occupancy. Smaller values of the weighted MSE and negative log likelihood correspond to better model fits.

The results provided for FARS are representative of the results for all crash data sources. Poisson Regression by Occupant and multi-level by occupant typically are the top two, with nearly identical performance. Poisson Regression by Occupant was used for all occupancy bias estimation.

Table 8: Validation of model for FARS' occupancy bias.

Model	Weighted MSE	Negative Log Likelihood
Poisson Regression by Occupant	1.865	0.432
Multilevel by Occupant	1.865	0.431
Poisson Regression	1.870	0.443
Multilevel	1.870	0.443
Lasso by Occupant	1.937	0.529
Lasso with Interactions	1.983	0.501
Naive Estimate	2.808	

Validation

A critical component of the evaluation was to determine whether the estimates were accurate. Where available, estimates were compared between measurement systems. In the absence of this option, several internal consistency checks were identified that could provide confidence in the quality of the estimates.

Internal Consistency

One method of validation that Heidtman et al. (1997) recommended was to confirm that the estimates follow known patterns of vehicle occupancy. This validates the relative values of the estimates, but not the absolute values. Specifically, they recommend confirming that:

- Weekday AM estimates are lower than weekday PM estimates,
- Weekend estimates are larger than the weekday estimates,
- Off peak estimates are larger than on peak estimates, and
- Winter estimates are smaller than summer estimates.

A natural additional test is to confirm that:

- Car estimates are larger than truck estimates.

The estimates produced in this evaluation are determined for subpopulations that include highway type, time of day, and vehicle class within each year and geographic division. A sampling of these estimates was selected to evaluate the recommended validation checks. They include:

1. Weekday AM peak to weekday PM peak for cars on non-interstate NHS highways,
2. Weekend day time to weekday midday for cars on non-interstate NHS highways,
3. Weekday AM to weekday midday for cars on non-interstate NHS highways, and
4. Cars to trucks for weekday midday on non-interstate NHS highways.

Comparison of winter to summer was not possible since all estimates produced were for full years. To complete the comparison, a statistic was calculated for whether the comparison provided the expected outcome with regard to which VOF was larger. As a control, the statistical comparison was repeated with VOF estimated by a more naïve methodology, where the average of the number of occupants for all crashes fitting the given criteria was used without application of the bias and prevalence estimates.

The results of the four validation tests are in Table 9, using both FARS and the state crash records as the data sources. Pass and fail are counts of the number of unique state (or urbanized area) and year combinations where the VOFs determined by the two methods generated the expected result (Pass) or the opposite (Fail). The table shows that the bias and prevalence adjusted methodology passed every test for all geography/year combinations using both data sources and for both geography levels. The results for the naïve estimates are generally not 100 percent consistent with the expected trends. Additionally, many of these naïve estimates could not even be developed because there were no crashes in the database for a given year/geography and other filtering criteria.

Table 9: Results of the validation tests for VOF.

			Full Adjustment			Naïve		
Test	Data Source	Geography	Pass	Fail	% Pass	Pass	Fail	% Pass
1	FARS	State	510	0	100%	355	106	77%
1	FARS	UZA	1770	0	100%	396	114	78%
1	State Crash Records	State	53	0	100%	53	0	100%
1	State Crash Records	UZA	383	0	100%	339	19	95%
2	FARS	State	510	0	100%	354	126	74%
2	FARS	UZA	1770	0	100%	468	216	68%
2	State Crash Records	State	53	0	100%	53	0	100%
2	State Crash Records	UZA	383	0	100%	350	10	97%
3	FARS	State	510	0	100%	340	128	73%
3	FARS	UZA	1770	0	100%	417	136	75%
3	State Crash Records	State	53	0	100%	53	0	100%
3	State Crash Records	UZA	383	0	100%	332	28	92%
4	FARS	State	510	0	100%	390	34	92%
4	FARS	UZA	1770	0	100%	248	31	89%
4	State Crash Records	State	53	0	100%	50	3	94%
4	State Crash Records	UZA	383	0	100%	305	27	92%

Year-Over-Year Consistency of Estimates

Another measure of validity would be consistency of the estimates from year-to-year. While some change in VOF is certainly possible, it seems unlikely that large scale geographies would see large year-over-year changes. Figure 4 shows the yearly overall state-wide VOF estimates using the bias and prevalence-adjusted methodology compared to the naïve average based solely on FARS crash occupancy. The more robust methodology leads to more consistent year-over-year results (plot on left) compared to the more random appearing pattern on the right. While more consistent, the fully adjusted VOFs are still susceptible to movement as evident by the separation of the states into two groupings at 2012.

After review, this is caused by a change in the occupancy bias estimation at 2012 that drove most states results higher, but seemed to depress the results for the states in the Mountain Census division (CO, MT, ID, WY, UT, NM, AZ, NV). Figure 5 shows the VOF by Census division[1] as estimated by the NHTS for 2001, 2009, and 2017. Only the Mountain division had a significant decrease in VOF from 2001 to 2009. Since the occupancy bias estimates used for 2012 and on were derived from the 2009 NHTS, this caused those estimates to likewise decrease.

[1] Map available at https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

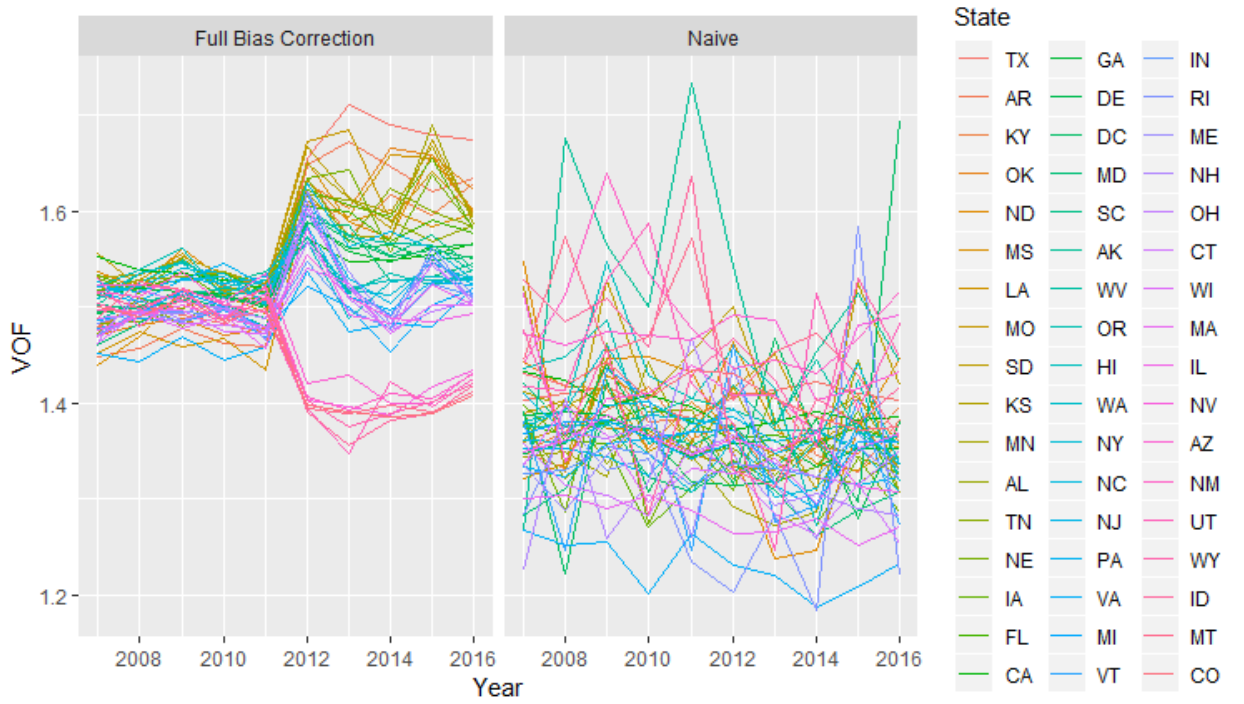

Figure 4: State-wide car and truck VOF estimates using the fully adjusted estimation methodology and the naïve average using FARS data.

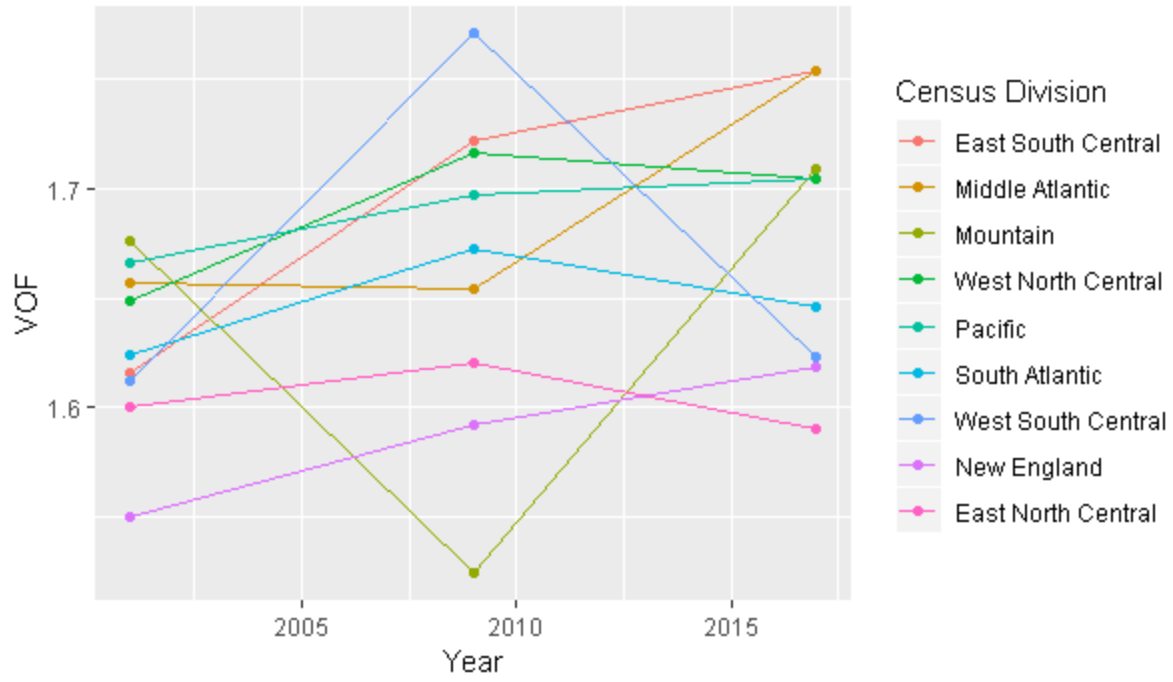

Figure 5: NHTS VOF estimates by Census division.

A similar analysis for year-over-year consistency of VOFs was conducted using the state crash records, as shown in Figure 6. This plot includes error bars indicating one standard error and adds a post-stratified estimate without the occupancy bias correction in addition to the naïve estimate and the fully adjusted estimates. In contrast to the results of the consistency analysis for FARS, the naïve estimates from the state crash data appear just as consistent from year-to-year as the adjusted estimates. This is most likely due to the much larger amount of data on crashes in the state data. When comparing the absolute values of the estimates, though, the fully adjusted estimates appear to be systematically larger and closer to the national NHTS estimate of 1.67. The one exception is for CA HSIS data, where the naïve estimates are much higher than the fully adjusted estimates. Investigation of this result suggests there may be a bias in the way the CA HSIS data are recorded, where the number of passengers were less likely to be recorded if there was only the driver in the car.

The inclusion of the intermediate estimate with post-stratification was intentional and it provided an important observation in that the results of this method were more like the naïve estimates than to the fully adjusted ones. This indicates that the population of drivers who get in some sort of crash is not too dissimilar from the general population of drivers. The larger adjustment from the occupancy bias indicates that the methodology finds that the number of occupants in the car effects the likelihood of getting in a crash.

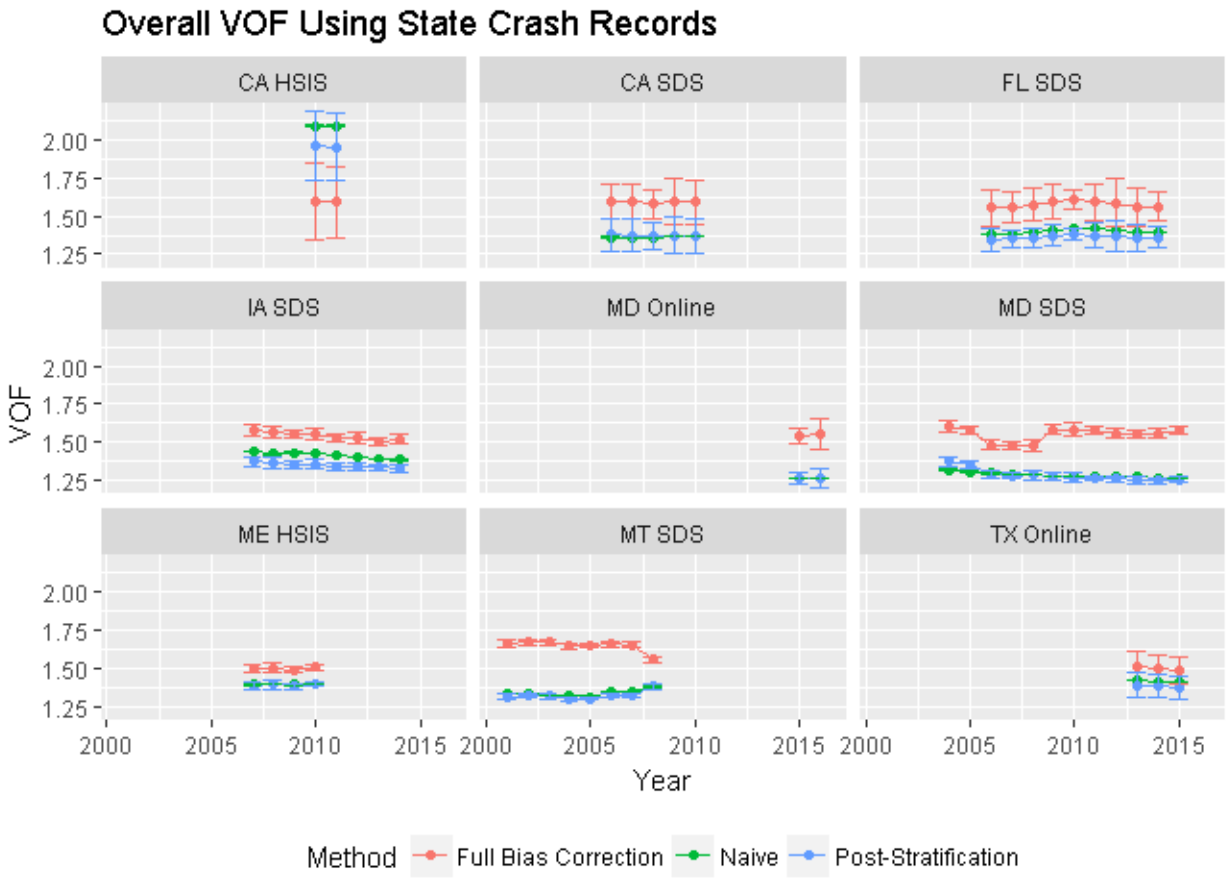

Figure 6: State-wide car and truck VOF estimates with standard errors using the fully adjusted estimation methodology, post-stratification, and the naïve average using state crash records.

Comparison of Different VOF Estimates

The use of different source data in FARS and the various state-based systems, as well as the overall NHTS VOF provide another way to validate the proposed methodology. Figure 7 shows the comparison of the state-wide car and truck estimates from the seven pilot states using the different data sources: SDS, HSIS, and Online for state crash records, FARS for deadly crashes, and NHTS for a survey of the driving population. ME and MT do not have NHTS estimates for 2001 because their sample sizes were too small to be included in the survey. Standard errors are provided for all estimates. The estimates from the crash records use the methodology outlined in this report, while the NHTS estimates use replicate weights.

In general, the estimates from the different crash data sources (FARS, SDS, HSIS, Online) are consistent, with the standard errors often overlapping. The estimates from NHTS are either in line with the crash-based estimates or a little higher. One reason the NHTS estimates are higher could be due to a difference in the relative proportion of truck traffic. In the NHTS, the percent of VMT driven by trucks were 2.7%, 1.6%, and 0% in the three surveys. In contrast, according to the 2015 Highway Statistics Series, 9.0% of VMT were driven by "single-unit 2-axle 6-tire or more and combination trucks". Since the fully adjusted method uses the HSS data to post-stratify and truck occupancy is generally much lower than car occupancy, it is not unexpected that crash-based estimates are lower than NHTS.

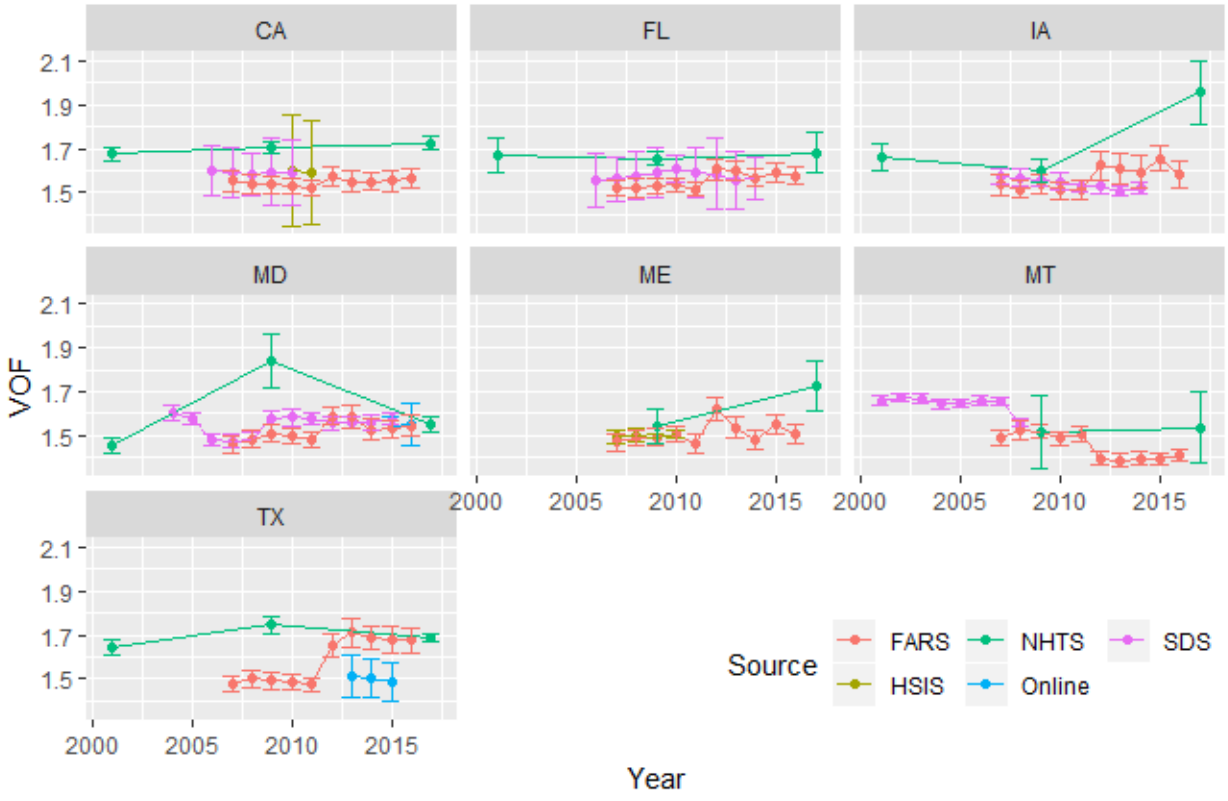

Figure 7: Comparison of state-wide car and truck VOF estimates from different data sources.

Local municipal planning organizations and state DOT's sometimes conduct their own VOF surveys. Among the seven pilot state DOTs, ME provided their VOF estimates based on crash data. FL DOT commissioned a report to estimate VOF with crash records. An estimate of Washington, D.C.'s VOF was obtained from Washington Council of Governments' 2007/08 Household Travel Survey.

The comparisons of VOFs in this evaluation to the targeted ones listed above is shown in Figure 8. The FL external estimate was transcribed from a graphic. Its values unfortunately pre-date the earliest estimates in this evaluation. The ME external estimate does not have an associated year, so it was simply plotted for the entire 2007 to 2016 time period. The DC external estimate is within the standard error bars of the fully adjusted FARS-based estimate. The FL and ME external estimates are lower than the fully adjusted estimates. This may be due to the fact that they are based on crash data and as was demonstrated in the previous section, the proposed occupancy adjustment in this evaluation typically increases estimates. ME makes no adjustments to the crash data, while FL does some post-stratification by driver age and gender.

Finally, in 2013, the Federal Motor Carrier Safety Administration conducted a nationally-representative survey of commercial motor vehicles to estimate seatbelt usage (https://www.trucking.org/ATA%20Docs/What%20We%20Do/Trucking%20Issues/Documents/Safety/SBUCMVD%202013%20Final%20Report%20020414.pdf). They observed a VOF of 1.06 in the front seat. Since they only observed the front seat and there may be occlusion issues, this estimate is a lower bound of the true VOF. The 2013 truck VOF morning estimates using the proposed methodology with FARS data varied from 1.07 to 1.16, depending on the road type and state. For all road types and states, 41% of the estimates have a 95% confidence interval that contains the survey estimate (1.06). This provides weak evidence that the proposed methodology overestimates truck VOF, possibly due to pooling the estimates with cars. The difference could also be caused by random variation and so additional verification is needed to make any conclusions.

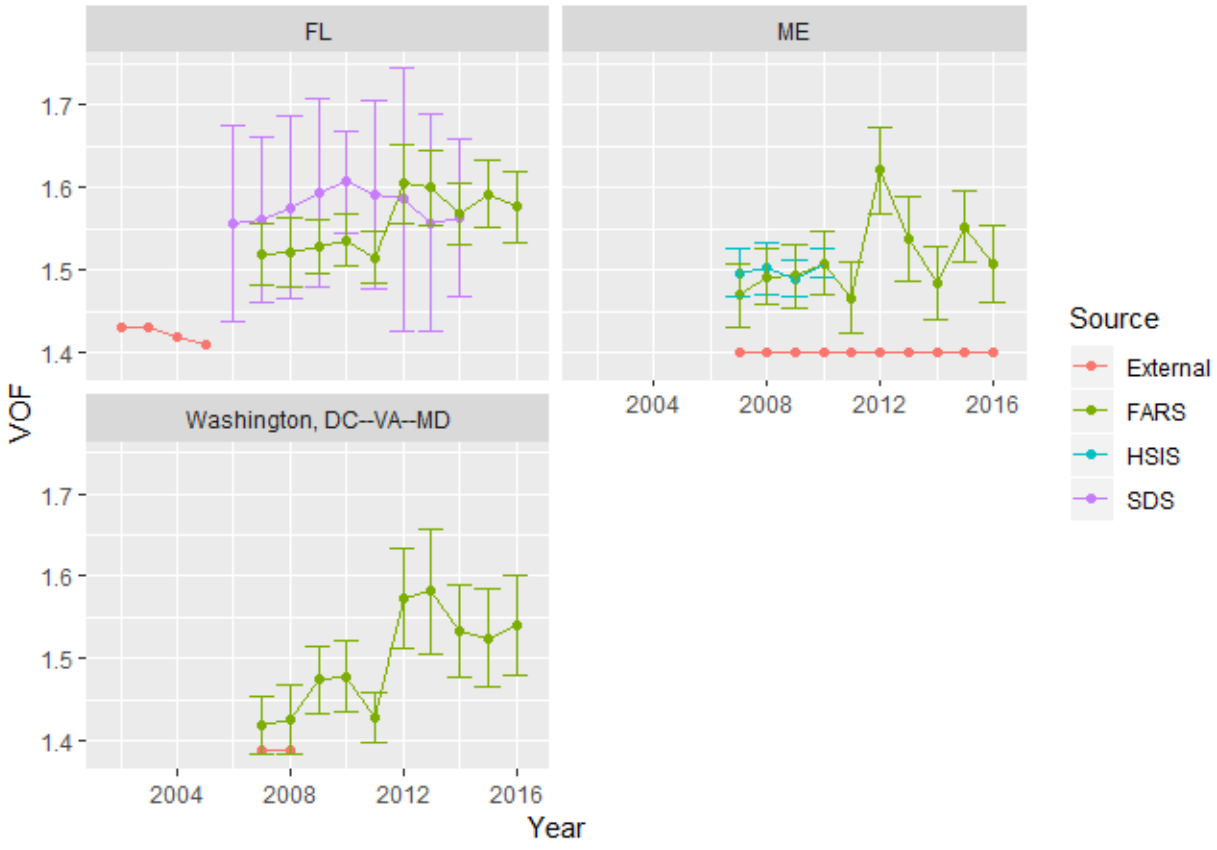

Figure 8: External comparison for select geographies.

NonSOV Comparison

To better understand whether the estimated NonSOV are of the correct magnitude, the distribution of the estimates from the urbanized areas is compared to similar estimates for metropolitan statistical areas from both ACS and NHTS. This comparison of NonSOV estimates is based on FARS data since all urbanized areas are estimated every year.

The ACS provides NonSOV estimates, but it is only for commuting trips and so is much lower than general NonSOV. Figure 9 compares commuting-based non-SOV from the ACS to general NonSOV from the NHTS for metropolitan statistical areas in 2009 and 2017. The ACS commuting NonSOV is typically between 15% and 35%, while the NHTS general NonSOV is typically much higher, between 50% and 70%.

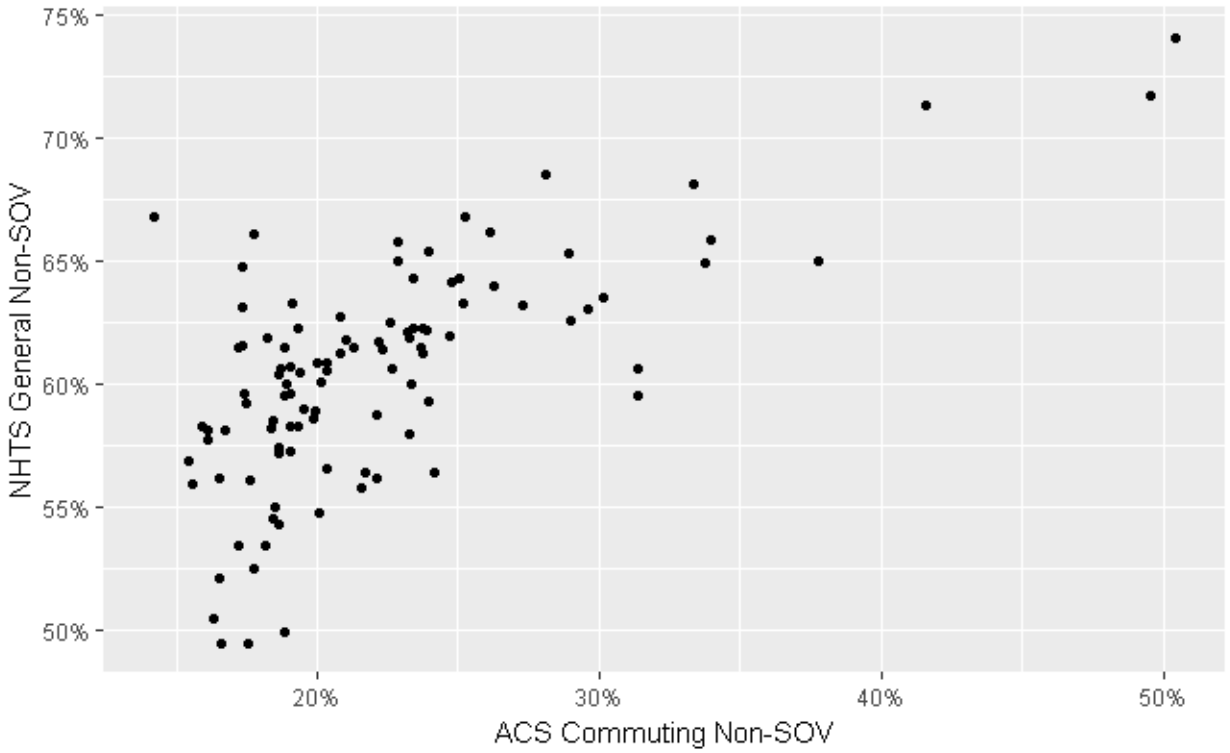

Figure 9: Comparison of ACS commuting NonSOV and NHTS general non-SOV.

The distribution of the NonSOV estimates made using FARS data (2007-2016) are next compared to NonSOV estimates made with NHTS data (2001, 2009, 2017) in Figure 10. Each line is a kernel density estimate for one year. The FARS-based estimates are at the urbanized area level, while the NHTS estimates are at the metropolitan statistical area level. Two methods of calculating NonSOV with NHTS data are presented. The first is trip-based, where the proportion of trips, regardless of driver, are used to estimate NonSOV. This is the recommended method and gives estimates lower than the proposed estimates. To explore why the proposed estimates are higher, NonSOV is also calculated with a vehicle-based method, that more closely mirrors the proposed methodology. For the vehicle-based method, the proportion vehicle SOV is calculated by dividing the proportion of vehicle trips with only one occupant by the VOF. Since crashes occur at the vehicle level, estimates made using crash data must use this methodology. When doing so, the NHTS estimates become even higher. Therefore, it is not possible to definitively state whether the proposed estimates are too high or too low.

Further, there may be a small effect of differing city sizes on NonSOV. Since metropolitan statistical areas are typically larger than their corresponding urbanized areas, there may be less sprawl and more vehicle traffic in MSAs.

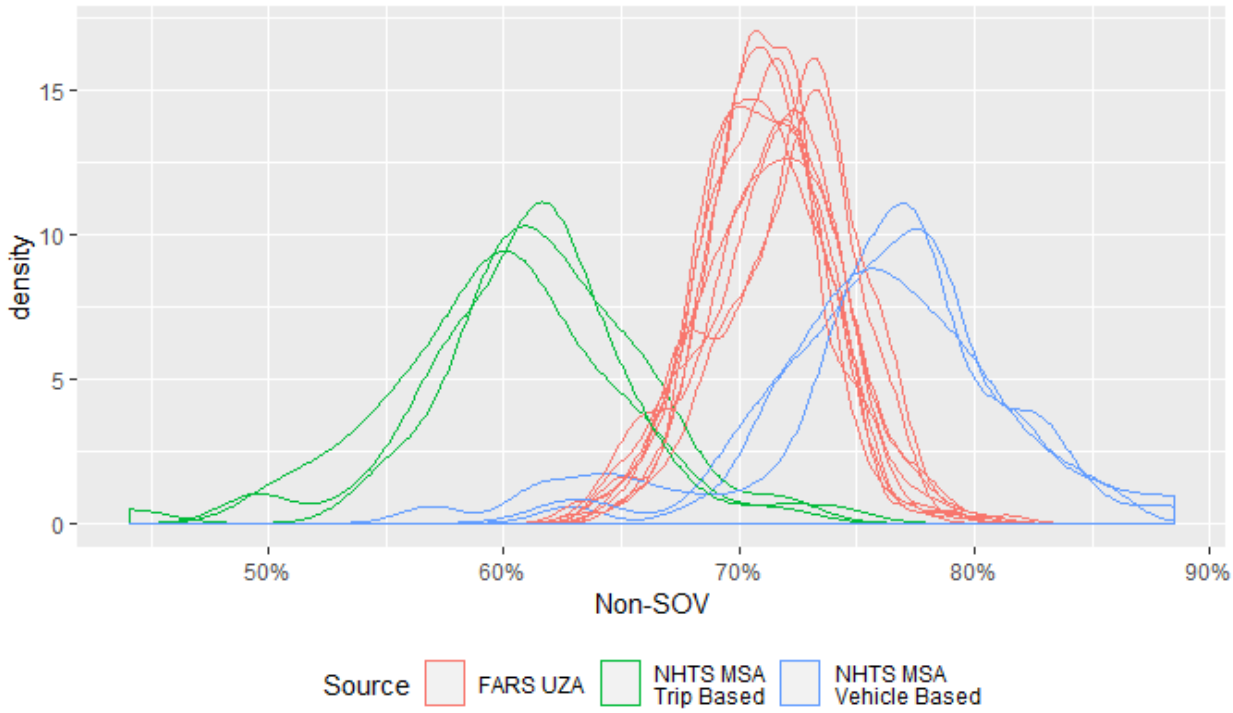

Figure 10: Comparison of the proposed non-SOV urbanized area estimates based on FARS and two methods of calculating non-SOV with NHTS data for metropolitan statistical areas. Each density curve visualizes the distribution for one year.

As a first datapoint, the relevant information to calculate vehicle-based NonSOV ($NonSOV_{veh}$) was available from a 2010/2011 survey by the New York Metropolitan Transportation Council and the North Jersey Transportation Planning Authority, which covered an area slightly larger than the New York--Newark, NY--NJ--CT urbanized area (NYMTC and NJTPA, 2014). They found that for all weekday travel, 68.2% of trips were single occupant, 21.7% had 2 occupants, 6.6% had 3 occupants, and 3.5% had 4 or more occupants (Table 4-53 of NYMTC and NJTPA, 2014). Assuming the conditional mean occupancy is 4.5 when there are 4 or more occupants, this implies the $NonSOV_{veh}$ was 53.7%. The proposed methodology using FARS estimates $NonSOV_{veh}$ was 56.0% in 2010 and 52.7% in 2011, with an average value of 56.7% between 2007 and 2016. This is in line with other urbanized areas, as the average $NonSOV_{veh}$ for all urbanized areas over all years estimated is 54.0% using FARS. This data point provides validation that our $NonSOV_{veh}$ estimates using crash data are in line with this external survey. The other input to NonSOV estimation, $\Pr(Vehicle)$, is not estimated using crash records and gives estimates in line with the NHTS (see Figure 9).

A more systematic survey was done to determine if the larger values of NonSOV are reasonable and possibly provide validation of the estimates. Metropolitan planning organization (MPOs) of all urban cities/counties across the nation with population more than 200,000 were contacted. A total of 177 MPOs were contacted to provide the non-SOV or VOF values they use for planning purposes. A generic email was sent to lead individuals in the planning division of the MPOs or their equivalent. Of the 177 MPOs, 20 replied, and 12 provided relevant information. A few agencies mentioned that they do not collect travel

data to estimate the regional non-SOV and VOF values, rather they adopt values that neighboring or other agencies use.

Table 10 shows the details of the information obtained from the MPOs that responded, as well as the 2016 5-year estimates from ACS, which only includes commuting trips, and the estimates proposed in this report. In addition, New York, Philadelphia, Atlanta, Detroit, Denver, and Cincinnati provided estimates, but were not included because they only included either work trips or auto-based trips.

In general, the estimates from the agencies are much higher than the ACS commuting estimates, and closer to the proposed methodology, although the estimates from the proposed methodology are typically larger. This gives further credibility the general NonSOV should be much higher than the ACS estimates, which only account for commuting trips. The one outlier is St. Louis, which has NonSOV values closer to the ACS than the methodology proposed in this report. They did not explicitly state the source of their estimate, and so may not be comparable.

Table 10. Summary of NonSOV values used by major MPOs.

Urbanized Area	Agency Name	Study Year	NonSOV from Agency	NonSOV from 2016 ACS 5-Year	NonSOV Estimate using FARS
Chicago, IL--IN	Chicago Metropolitan Agency for Planning	2015	57%	31%	75%
Miami, FL	Miami-Dade Transportation Planning Organization	2014-2017	64%	22%	74% (2014-2016)
Minneapolis--St. Paul, MN--WI	Minneapolis-St Paul Twin Cities Metropolitan Council	2010	52%	23%	74%
Tampa--St. Petersburg, FL	District Seven Planning & Environmental Management Office	2009	60%	20%	73%
St. Louis, MO--IL	East West Gateway Gov Association	Not Given	28%	18%	74% (for 2016)
San Antonio, TX	Alamo Area MPO	2006	66%	20%	69% (for 2007)
Orlando, FL	Metro Plan Orlando	2015	Home-based work 48% Non-home-based work 50%	20%	75%
Salt Lake City--West Valley City, UT	Wasatch Front Regional Council	2012	55%	25%	67%
El Paso, TX--NM	El Paso MPO	2012	59%	20%	74%
Reno, NV--CA	Washoe County Regional Commission	Not Given	52%	22%	65% (for 2016)
Stockton, CA	San Joaquin Council of Governments	2015	60-65%	23%	71%
Visalia, CA	Tulare County Association of Governments	2015	63%	18%	69%

Deliverables

VOF and NonSOV estimates

Table 11 shows the car and truck VOFs, and NonSOV estimates generated using FARS data and the different state crash data options. This includes NonSOV estimates for every urbanized area in Table 12. Further, car and truck VOF estimates are provided for the same urbanized areas as well as for the seven pilot states (California, Florida, Iowa, Maine, Maryland, Montana, and Texas). The car and truck VOF estimates include subgroup estimates by time of day and highway type within each vehicle class, year, and geographic region. All estimates are reported with mean values and standard errors

Table 11: Estimates provided by geography and data source for car and truck VOF and NonSOV.

Geography	FARS	SDS	HSIS	Online	Notes
CA State and UZAs	2007-2016	2006-2010	2010-2011		Missing Reno, NV--CA using SDS and HSIS
FL State and UZAs	2007-2016	2006-2014			
IA State	2007-2016	2007-2014			
• Omaha, NE—IA	2007-2016	2007-2013			
• Des Moines, IA	2007-2016	2007-2014			
• Davenport, IA—IL	2007-2016	2007-2014			
MD State	2007-2016	2004-2015		2015-2016	
• Philadelphia, PA—NJ—DE—MD	2007-2016	2010-2012			
• Washington, DC—VA—MD	2007-2016				
• Baltimore, MD	2007-2016	2004-2015*		2015-2016	
• Aberdeen--Bel Air South--Bel Air North, MD	2007-2016	2004-2015*		2015-2016	
ME State and UZA	2007-2016		2007-2010		
MT State	2007-2016	2001-2008			
TX State and UZAs	2007-2016			2013-2015	Missing El Paso, TX--NM using Online
All Other States and UZAs	2007-2016				

* NonSOV estimates are limited to 2006 to present, due to ACS availability

Urbanized areas that overlap more than one state presented some challenges to the analysis.

For the Washington, DC--VA--MD urbanized area, state crash data from VA and MD was adequate, but not from DC. The data from DC was missing vehicle and person identifiers necessary to count the number of occupants in the vehicles. Therefore, estimates were only possible through FARS.

The Reno, NV--CA urbanized area was only estimated through FARS. Virtually all of the population of this urbanized area is in Nevada and no state crash records were available in this analysis from Nevada.

The El Paso, TX--NM urbanized area was only estimated through FARS. State crash records were available for TX from 2013 to 2015 and for NM from 2001 to 2011, but this did not provide any single year where both sets of data were available.

The Pensacola, FL--AL urbanized area was estimated through FARS and also through Florida state crash records. Alabama state crash records were not available in this analysis, so the state-based analysis of the urbanized area utilized only Florida crash records. This decision was made as less than 2% of the urbanized area population is estimated to live in AL.

Estimates of bus VOF were generated for 2015 and 2016. Due to limited data availability, these estimates could only be made at the state and urbanized area level, and not for the time of day and highway type categories that were possible for passenger vehicles and trucks. Additionally, of the 59 urbanized areas in the seven pilot states, adequate data were only available to estimate the bus VOF for 45 urbanized areas.

Computer Code

The code to reproduce these estimates is also provided. Since there is a fair amount of variability between the crash records from different states and nationally, code is provided on how the data was processed, combined, and filtered for each state separately. For each crash data source, code to preprocess, filter, and normalize of the data was written in SAS. After the data has been pre-processed, the following procedure was used for each crash data source, all written in R:

1. If latitude and longitude are available, geographic information system techniques were used to identify the urbanized area, metropolitan statistical area, and whether it was on a national highway system highway
2. Additional processing, filtering, normalization, and variable reduction to retain the same variables for all sources
3. Data codebook creation and comparison of variable distributions to FARS
4. Data codebook creation and comparison of variable distributions to NHTS
5. Model comparison and report for estimating crash occupancy distribution
6. Model comparison and report for estimating occupancy bias
7. Estimate and save occupancy bias
8. Estimate and save model of crash occupancy distribution, and
9. Combine estimated prevalence, bias, and crash occupancy to estimate and save VOF and vehicle NonSOV.

The above nine files are used for each data source but can generally be applied to new data by updating the source and destination of the files, as well as the appropriate years. A template version of the 9 files is provided as well.

In addition to the code for estimating occupancy from crash records, code is included to:

- Process and normalize the NHTS
- Calculate prevalence, including data from NHTS, HSS, and TVT
- Estimate the proportion of non-vehicle traffic using ACS data
- Combine the non-vehicle estimates with the vehicle NonSOV, and
- Estimate the bus VOF using NTD data.

Selected Results

Selected graphical summaries are provided below to show the relative range of results obtained using this methodology. Figures 11 and 12 show the range of vehicle occupancy factors and NonSOV travel, respectively, across the 177 urbanized areas in the continental US based on FARS data for the 2016 calendar year. From Figure 11, six of the lowest nine VOFs for 2016 were UZAs in Colorado and Utah,

while 13 of the top 18 were in Texas. For the NonSOV data in Figure 12, results ranged from a low in Fort Wayne (0.643) to a high in New York City—Newark (0.803).

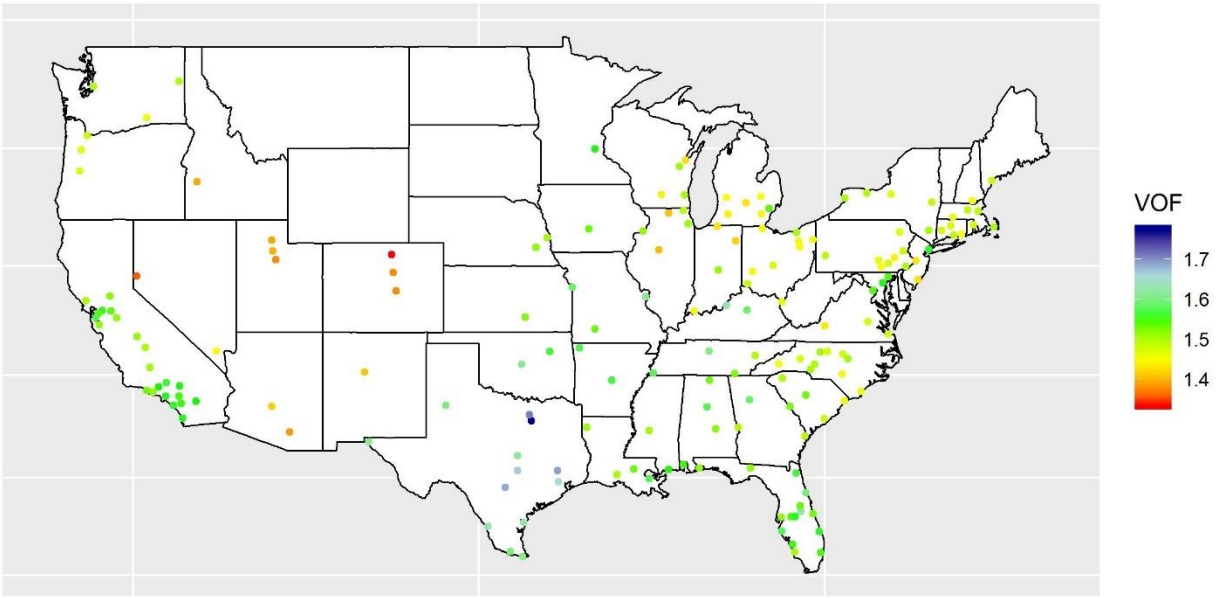

Figure 11: FARS based Vehicle Occupancy Factors by census urbanized area for 2016.

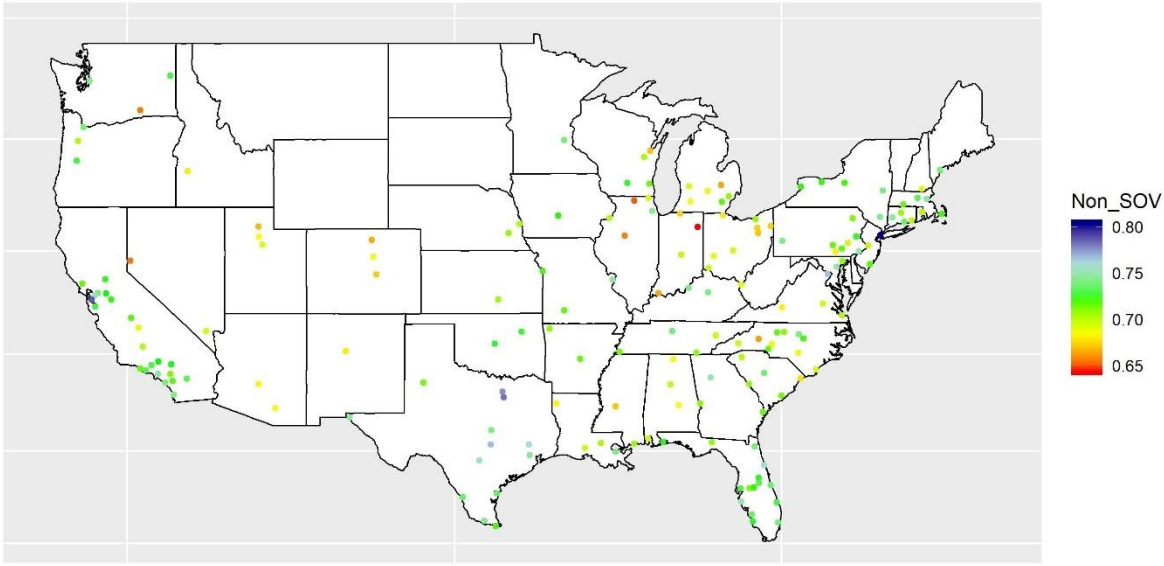

Figure 12: FARS based NonSOV by census urbanized area for 2016.

Figure 13 shows boxplots of the range of VOF estimates of the aggregate of cars and trucks (weighted by VMT) across U.S. States (and the District of Columbia) using FARS data. From this Figure, the overall shift in median state VOF from approximately 1.5 in years 2007 through 2011 can be plainly seen. The

impact of the use of the 2009 NHTS starting in 2012 coincided with an almost 0.1 increase in VOF from the previous year. Following 2012, the median VOF returned to levels of 1.55 and below, but the state to state variability that was discussed above continued through 2016.

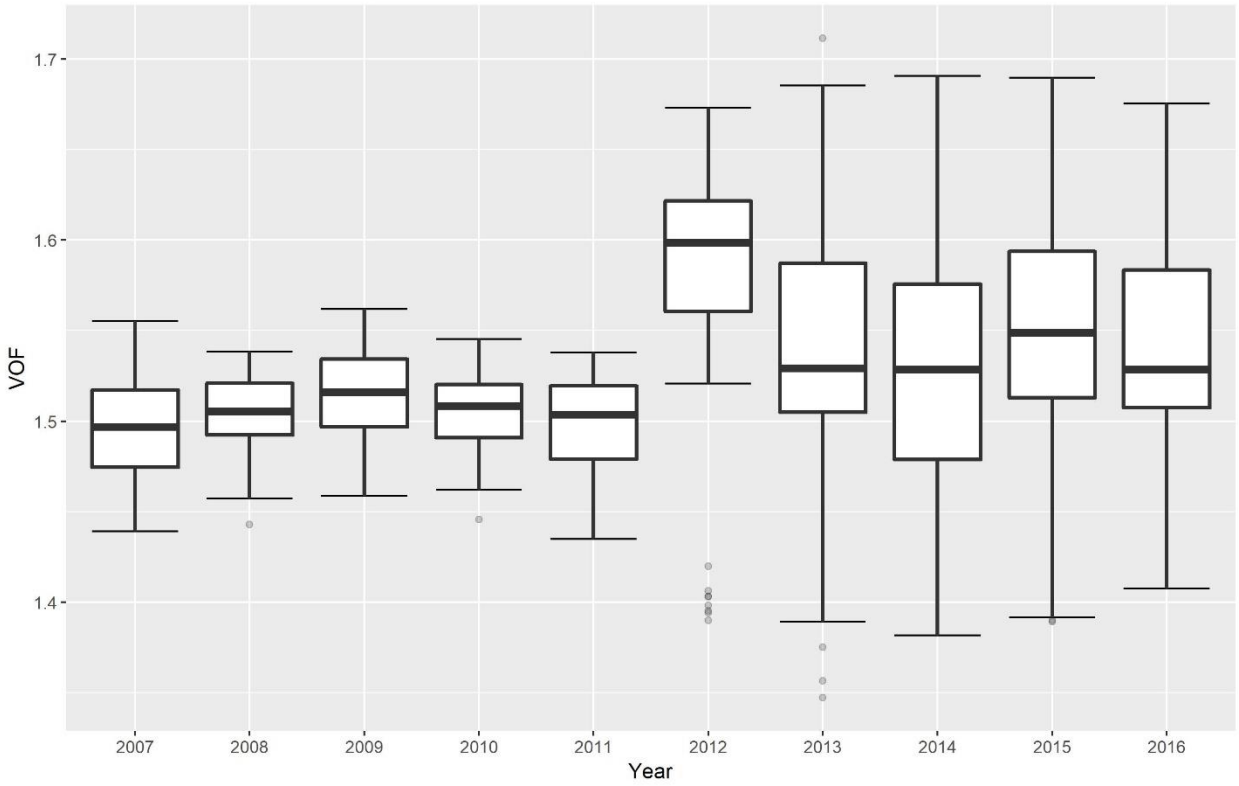

Figure 13: FARS based Vehicle Occupancy Factors of Cars and Trucks across states by Year.

Figure 14 provides some indication of how VOF varied by the covariates of time of day and whether travel was on the Interstate highway system. Whether for cars or trucks, the non-interstate NHS VOFs tended to be a little higher than for interstate, though the differences were small. Time of day was associated with much more impactful VOF measurement differences, with weekday AM rush (6-10 AM) consistently the lowest VOF, and weekday PM rush (4-8 PM) running second. The weekend daytime VOFs showed the largest estimates of any time period. The substantially lower VOFs for trucks are also apparent in Figure 14.

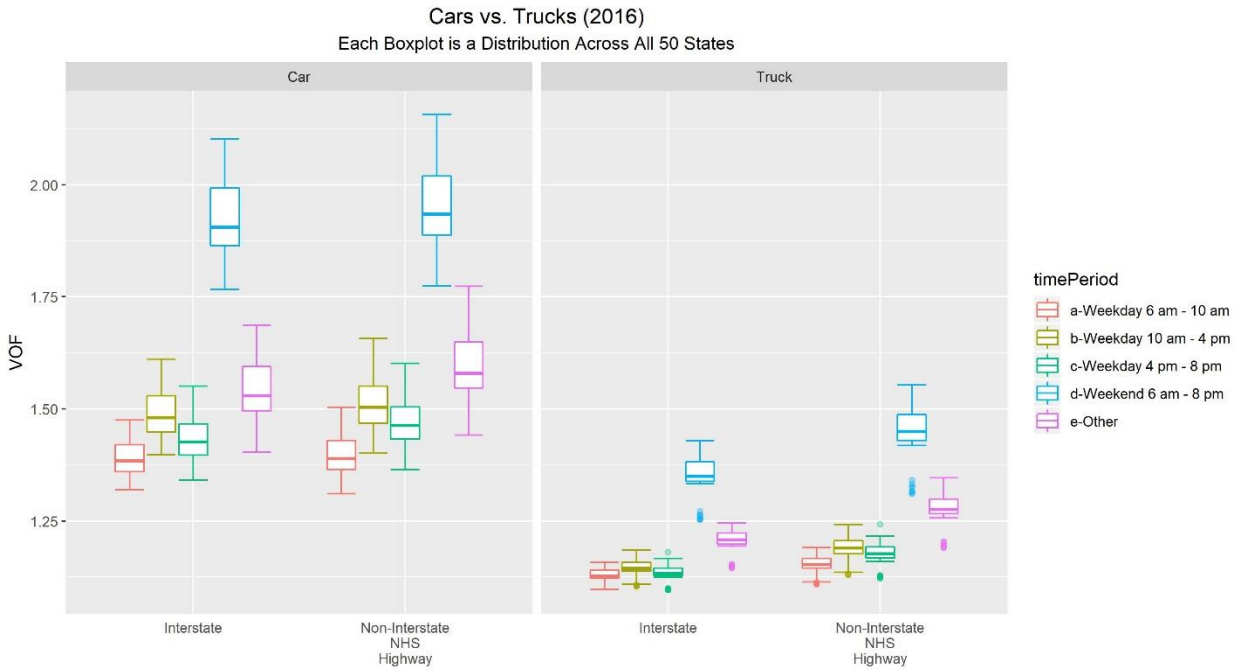

Figure 14: FARS based Vehicle Occupancy Factors by State in 2016 Between Cars and Trucks at Different levels of the interstate and time of day covariates.

Conclusions and Recommendations

The following conclusions and recommendations are provided:

1) For passenger cars and trucks, the method as proposed appears to work to produce credible VOF and NonSOV. The estimates produced were compared in both internal and external validation checks and found to be within the ranges that would be expected.
    a. Estimates were also demonstrated for buses, but these estimates were not so easily validated, and they were only able to be calculated for higher level geographic divisions.
2) Future use of the methodology is always dependent on quantity and quality of crash data as well as some source of occupancy bias since this did appear to be an important factor in the methodology. The uncertainty of the future form of NHTS is a concern since it forms one central part of the bias adjustments. The use of the 2017 NHTS for occupancy bias in passenger vehicles may be acceptable for a few years before drift in this bias over time would have to be investigated.
3) Individual state records are more numerous and can produce a more representative estimate than FARS but have the disadvantage on not necessarily being uniform in their content and data quality, both of which are issues for a sustainable system. However, using FARS or the state system, or even an average of the two, is arguably superior to the current system of only a single national NHTS value, or the great cost of implementing a statewide system.
4) The code used to generate a large subset of VOF and NonSOV estimates has been provided. The system of obtaining and preparing the input data records for this task is non-trivial and would best be completed by someone trained and knowledgeable in data management, data science, and statistics. The code does not anticipate all future issues that could generate questionable results and estimates generated from the code should still be evaluated for realism before being published.

If new users wish to adopt this methodology, this code can be used, with some understanding and modifications, to generate data from other crash records, especially in future time periods. The contents of this evaluation were originally meant to be a demonstration of a proposed methodology and were required for only a subset of U.S. geographic and temporal divisions. The estimates delivered herein greatly exceed those original requirements, in many cases achieving a true national scope. In some areas, though, the estimates could be further enhanced, especially with state level crash data. Additionally, the methodology as implemented assumes the continued availability of data of the type and fidelity currently available. This assumption is discussed where there are concerns that this condition will not persist.

References

Asante, S., Adams, L., Shufon, J., & McClean, J. (1996). Estimating average automobile occupancy from accident data in New York State. *Transportation Research Record: Journal of the Transportation Research Board*, (1553), 115-123.

Breiman, Leo (2001). Random forests. *Machine learning* 45.1: 5-32.

Chang, Li-Yen, and Fred Mannering (1998). Predicting vehicle occupancies from accident data: An accident severity approach. *Transportation Research Record: Journal of the Transportation Research Board* 1635: 93-104.

Chen, Li-Hui, et al (2000). Carrying passengers as a risk factor for crashes fatal to 16-and 17-year-old drivers. *JAMA* 283.12: 1578-1582.

Gan, Albert, K. Y. Liu, and Rax Jung (2007). Vehicle occupancy data collection methods (Phase II). *BD015-14, Transportation Statistics Office, Florida Dept. of Transportation, Tallahassee, FL*.

Heidtman, K., B. Skarpness, and C. Tornow (1997). *Improved vehicle occupancy data collection methods*. No. DTFH61-93-C-00055. Office of Highway Information Management, Federal Highway Administration.

Gaulin, R. (1991). A Procedure to Calculate Vehicle Occupancy Rates from Traffic Accident Data. Bureau of Planning, Connecticut Department of Transportation, Newington.

McCullagh, P., Nelder, J. (1987). *Generalized Linear Models*, Second Edition. CRC Press, New York.

NYMTC and NJTPA (2014). 2010/2011 Regional Household Travel Survey Final Report. https://www.nymtc.org/portals/0/pdf/RHTS/RHTS_FinalReport%2010.6.2014.pdf.

Stekhoven, D.J. and Buehlmann, P. (2012). MissForest - nonparametric missing value imputation for mixed-type data, *Bioinformatics, 28*(1), 112-118, doi: 10.1093/bioinformatics/btr597

Tibshirani, Robert (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*: 267-288.

Wang, Wei, et al (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting* 31.3: 980-991.

Appendix

Table 12: Urbanized area with a population of at least 200,000 in the seven[2] designated states. The population was recorded in the 2010 Census.

State	Urbanized Area	Population	Other States Needed
California	Los Angeles--Long Beach--Anaheim, CA	12,150,996	
	San Francisco--Oakland, CA	3,281,212	
	San Diego, CA	2,956,746	
	Riverside--San Bernardino, CA	1,932,666	
	Sacramento, CA	1,723,634	
	San Jose, CA	1,664,496	
	Fresno, CA	654,628	
	Concord, CA	615,968	
	Mission Viejo--Lake Forest--San Clemente, CA	583,681	
	Bakersfield, CA	523,994	
	Murrieta--Temecula--Menifee, CA	441,546	
	Reno, NV—CA	392,141	NV
	Stockton, CA	370,583	
	Oxnard, CA	367,260	
	Modesto, CA	358,172	
	Indio--Cathedral City, CA	345,580	
	Lancaster--Palmdale, CA	341,219	
	Victorville--Hesperia, CA	328,454	
	Santa Rosa, CA	308,231	
	Antioch, CA	277,634	
	Santa Clarita, CA	258,653	
	Visalia, CA	219,454	
	Thousand Oaks, CA	214,811	
Florida	Miami, FL	5,502,379	
	Tampa--St. Petersburg, FL	2,441,770	
	Orlando, FL	1,510,516	
	Jacksonville, FL	1,065,219	
	Sarasota--Bradenton, FL	643,260	
	Cape Coral, FL	530,290	
	Palm Bay--Melbourne, FL	452,791	
	Port St. Lucie, FL	376,047	
	Palm Coast--Daytona Beach--Port Orange, FL	349,064	
	Pensacola, FL—AL	340,067	AL

[2] Note that Montana is one of the seven states to be evaluated in this study, but it is excluded from this table due to no qualifying urbanized area.

State	Urbanized Area	Population	Other States Needed
	Kissimmee, FL	314,071	
	Bonita Springs, FL	310,298	
	Lakeland, FL	262,596	
	Tallahassee, FL	240,223	
	Winter Haven, FL	201,289	
Iowa	Omaha, NE—IA	725,008	NE
	Des Moines, IA	450,070	
	Davenport, IA—IL	280,051	IL
Maine	Portland, ME	203,914	
Maryland	Philadelphia, PA—NJ—DE—MD	5,441,567	PA, NJ, DE
	Washington, DC—VA—MD	4,586,770	DC, VA
	Baltimore, MD	2,203,663	
	Aberdeen--Bel Air South--Bel Air North, MD	213,751	
Texas	Dallas--Fort Worth--Arlington, TX	5,121,892	
	Houston, TX	4,944,332	
	San Antonio, TX	1,758,210	
	Austin, TX	1,362,416	
	El Paso, TX—NM	803,086	NM
	McAllen, TX	728,825	
	Denton--Lewisville, TX	366,174	
	Corpus Christi, TX	320,069	
	Conroe--The Woodlands, TX	239,938	
	Lubbock, TX	237,356	
	Laredo, TX	235,730	
	Killeen, TX	217,630	
	Brownsville, TX	217,585	