U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**

NHTSA

DOT HS 812 804

September 2019

Crash Investigation Sampling System:

Sample Design and Weighting

DISCLAIMER

Suggested APA Format Citation:

Zhang, F., Noh, E. Y., Subramanian, R., & Chen, C.-L. (2019, September). *Crash Investigation Sampling System: Sample design and weighting* (Report No. DOT HS 812 804). Washington, DC: National Highway Traffic Safety Administration.

Technical Report Documentation Page

1. Report No. DOT HS 812 804	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Crash Investigation Sampling System: Sample Design and Weighting		5. Report Date September 2019	
		6. Performing Organization Code NSA-210	
7. Authors Fan Zhang, Eun Young Noh, Rajesh Subramanian, Chou-Lin Chen		8. Performing Organization Report No.	
9. Performing Organization Name Mathematical Analysis Division, National Center for Statistics and Analysis National Highway Traffic Safety Administration 1200 New Jersey Avenue SE Washington, DC 20590		10. Work Unit No. (TRAIS)	
		11. Contract or Grant No.	
12. Sponsoring Agency Name and Address Mathematical Analysis Division, National Center for Statistics and Analysis National Highway Traffic Safety Administration 1200 New Jersey Avenue SE Washington, DC 20590		13. Type of Report and Period Covered NHTSA Technical Report	
		14. Sponsoring Agency Code	
15. Supplementary Notes The authors would like to thank Phillip Kott for his consultation.			
16. Abstract As part of the effort to modernize NHTSA's crash data collection system, NCSA has designed a new national probability-based crash sampling system – the Crash Investigation Sampling System (CISS) to replace the Crashworthiness Data System (CDS). This document summarizes the sample design and weighting methods of CISS.			
17. Key Words Crash Investigation Sampling System, CISS, CISS Sample Design, CISS Weighting, Crashworthiness Data System, CDS	18. Distribution Statement This document is available to the public from the National Technical Information Service, www.ntis.gov.		
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 58	22. Price

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

TABLE OF CONTENTS

1. Executive Summary

The National Highway Traffic Safety Administration has been collecting motor vehicle crash data through a number of systems including the National Automotive Sampling System (NASS). NASS was established in the 1970s to support vehicle/highway safety research, policy making, and regulation program development.

NASS is comprised of two nested probability sampling systems – the General Estimates System (GES) and the Crashworthiness Data System (CDS). The GES collected general information of the traffic crashes from police crash reports only. The CDS collected detailed information from the crashes involving passenger vehicles to better understand the crashworthiness of vehicles and consequences to occupants in crashes. NHTSA had developed and implemented CDS in the 1980s. CDS is based upon a three-stage, stratified random sample of Primary Sampling Units (PSUs), police jurisdictions (PJs), and police accident reports (PARs). The CDS 24-PSU sample is a subsample of the GES 60-PSU sample. The same PSU and PJ samples have been used for CDS data collection since 1989.

Over the past two decades, however, the general population, vehicles, and highway safety measures have changed dramatically, so that crash characteristics and distributions have changed over the PSU and PJ frame. In addition, the research interest of the transportation community has expanded to topics such as driver performance, crash avoidance, and the effects of new technologies on crash amelioration.

NHTSA recognized the need to undertake a redesign of NASS to better support its own and stakeholders' data needs. Congress authorized NHTSA to undertake a significant effort to re-design and modernize its crash data collection system. NHTSA identified three major areas for improvement – re-designing the survey sample, modernizing the information technology (IT) infrastructure, and revamping its data collection protocols and technology.

The redesign started in January 2012. The majority of the work was in the formation of conceptual research designs, establishment of sampling frames, selection of data collection locations and sources, and documentation of protocol and results for the new surveys. During this process, two new national, probability-based crash sampling systems were designed – the Crash Report Sampling System (CRSS) and Crash Investigation Sampling System (CISS) - to replace GES and CDS. This report summarizes the sample design and weighting methodology used in CISS.

After its assessment of research objectives and operational considerations, NHTSA decided to design the CISS independently from CRSS in order to optimize both CISS and CRSS. Therefore, unlike the current NASS, the formation and selection of the CISS PSUs were independent of the CRSS PSU formation and selection.

CISS has a stratified three-stage sample design similar to CDS: PSU, PJ, and PAR. The CISS PSUs were formed so that a minimum number of severe crashes could be selected from as many PSUs as possible. To keep travel time for data technicians under control, different driving distance constraints were imposed to rural PSUs and urban PSUs. The PSUs are deeply stratified and selected with probability proportional to the expected number of severe crash counts based on previous experience. In addition, the CISS PSU

sample has been designed to be scalable to accommodate future budgetary fluctuation without completely reselecting the PSU sample.

Pareto sampling (Rosén 1997) was used for both PJ and PAR sample selection. Pareto sampling method produces overlapping samples when a new sample is selected. This reduces the changes to the existing sample when a new sample needs to be selected. For PJ sample selection, Pareto sampling produces a PJ sample with selection probabilities approximately proportional to the PJ's crash counts. Pareto sampling makes it easier to handle PJ frame changes such as the creation, closure, or splitting of PJs. For PAR sample selection, Pareto sampling not only allows cases of high interest to be selected with larger selection probability but also allows the PAR sample to be expanded to effectively replace non-responding cases (i.e., crashes with key vehicle information missing) with additionally sampled cases.

An optimization technique was applied to find an approximately optimal sample allocation: the best combination of PSU, PJ, and PAR sample sizes that minimize anticipated variance under a fixed budget. The optimization results indicate when budget is available the most effective way to reduce the standard error of an estimate is to increase the PSU sample size while maintaining the number of PJs per PSU and the number of PARs per PJ at certain levels.

In summary, the CISS has been designed as a stratified multi-stage and multi-phase sampling with unequal selection probabilities. The scalability designed into PSU sample and the Pareto sampling used in PJ and PAR sample selection provide options to adjust for uncertainties such as future budgetary fluctuations, administrative changes in the police jurisdictions or replacing cases that are missing critical information that will enable NHTSA to monitor and react to achieve desired sample allocations.

2. Introduction

NHTSA collects motor vehicle crash data to support its vehicle/highway safety research, policy making, and regulation program development. The NASS, established in the 1970s, has been one of its key crash data systems and an integral part of NHTSA's efforts to fulfill this mission.

NASS was comprised of two nested systems – the GES and the CDS. Both systems were operated by the NCSA and provided national probability samples of crashes.

GES was a survey of police-reported traffic accident reports, PARs. It collected general information of the traffic crashes from PARs only. See Appendix A for an example of a PAR. GES data were used to:

- Provide a general picture of the crash population and trends,
- Identify highway safety problem areas and assess the size of the problem ,
- Provide a basis for regulatory and consumer information initiatives, and
- Form the basis for cost and benefit analyses of vehicle regulations.

See Shelton (1991) for a detailed discussion of GES sampling and weighting procedures.

While the GES captured general information on all types of traffic crashes, CDS focused on collecting more detailed information from severe crashes involving passenger vehicles to better understand the crashworthiness of vehicles and consequences to occupants in crashes. CDS collected more detailed data about the crash, vehicles and occupants through:

- Interviews,
- Medical records,
- Vehicle inspections, and
- Scene inspections.

See Fleming (2010), and Zhang and Chen (2013) for more details on CDS sampling and weighting procedures.

Developed in the 1970s and redesigned in the 1980s, NASS's primary data collection sites, the PSUs, and the secondary data collection sites, the PJs, had not changed for the past 30+ years. During this time, the underlying NASS sampling frame had seen many changes, for example:

- The number and nature of crashes across PSUs,
- Population growth and mobility shift,
- PJ frame (opening, closing, merging, crash distribution changes among PJs),
- Improvements in vehicle and highway safety.

Also, the data needs of the highway safety community have increased and significantly changed over the last three decades. For example, the primary focus of the original NASS design was to enhance crashworthiness knowledge by providing detailed information about vehicle crash profiles, restraint system performance and injury mechanisms. In recent years, the highway safety community has been interested in understanding the factors leading to a crash in order to develop new crash avoidance countermeasures.

Furthermore, the scope of traffic safety studies has also been expanding with emerging traffic safety issues. Because of the limited CDS sample size, it has not provided enough sampled cases to support detailed domain analysis. While substantial reductions in passenger vehicle fatalities have been realized, data on emerging traffic safety areas were not collected in the CDS and need to be identified and analyzed.

Recognizing the importance as well as the limitations of the current NASS system, NHTSA is undertaking a modernization effort to upgrade its data systems by improving the information technology infrastructure, updating the data collected, and reexamining the NASS sample sites and size.

The United States Congress supported the effort to examine the deficiencies in NASS and to plan for a modernized and comprehensive data system. In MAP-21, Congress instructed:

> "(*a) IN GENERAL.—Not later than 1 year after the date of enactment of this Act, the Secretary shall submit a report to the Committee on Commerce, Science, and Transportation of the Senate and the Committee on Energy and Commerce of the House of Representatives regarding the quality of data collected through the National Automotive Sampling System, including the Special Crash Investigations Program.*
>
> *(b) REVIEW.—The Administrator of the National Highway Traffic Safety Administration (referred to in this section as the ''Administration'') shall conduct a comprehensive review of the data elements collected from each crash to determine if additional data should be collected. The review under this subsection shall include input from interested parties, including suppliers, automakers, safety advocates, the medical community, and research organizations.*
>
> *(c) CONTENTS.—the report issued under this section shall include—*
>
> *(1) The analysis and conclusions the Administration can reach from the amount of motor vehicle crash data collected in a given year;*
>
> *(2) The additional analysis and conclusions the Administration could reach if more crash investigations were conducted each year;*
>
> *(3) The number of investigations per year that would allow for optimal data analysis and crash information;*
>
> *(4) The results of the comprehensive review conducted pursuant to subsection (b);*
>
> *(5) The incremental costs of collecting and analyzing additional data, as well as data from additional crashes;*
>
> *(6) The potential for obtaining private funding for all or a portion of the costs under paragraph (5); H. R. 4348—367*
>
> *(7) The potential for recovering any additional costs from high volume users of the data, while continuing to make the data available to the general public free of charge;*
>
> *(8) The advantages or disadvantages of expanding collection of non-crash data instead of crash data;*
>
> *(9) Recommendations for improvements to the Administration's data collection program; and*

(10) The resources needed by the Administration to implement such recommendations."

As part of the effort to modernize NHTSA's data collection system, NCSA has designed two new national probability-based crash sampling systems – the Crash Report Sampling System and Crash Investigation Sampling System replacing GES and CDS. This document summaries the sample design and estimation methodology of CISS. A companion report summarizes the sample design and weighting methodology for CRSS.

The following sections discuss how data needs define the scope of CISS, an overview of the sample design, how crashes are selected from multi sampling stages, and how weights are created for both parameter estimation and variance estimation.

3. The Scope of CISS

Crash data needs and the focus of traffic safety research have significantly changed since the establishment of NASS in the 1970s. It is critical to identify current data-user needs to properly define the scope of CISS. This not only involves identifying data elements critical to the identification of safety issues, monitoring of trends and evaluation of the effectiveness of countermeasures, but also includes information that is no longer or less relevant to the traffic safety research community. To this end, NHTSA conducted two studies to evaluate internal and public data needs.

3.1 NHTSA's Data Needs

In August 2009 NHTSA assembled a project team to conduct a review of the crash databases and an assessment of current and projected data needs. Sixty NHTSA employees, representing all offices across the agency and with a broad range of expertise and perspectives, were interviewed. The team supplemented the interview data with documented rulemaking and research plans.

Through this review, NHTSA identified a number of broad based goals for a modernized NASS system. These included adding new data elements that would support the development of safety countermeasures, especially those related to crash avoidance and behavioral safety; expanding data collection on crashes involving motorcycles, commercial vehicles, pedestrians, bicyclists, school buses, and low speed vehicles, collecting more data on injuries and on the performance of advanced vehicle technologies, enhancing analysis through more complete case information and greater data accessibility, and modifying the research design to better reflect current crash populations.

3.2 The Data Needs of the Public

In order to solicit inputs from the broadest possible group of stakeholders, NHTSA published a notice in the Federal Register announcing the survey modernization effort on June 21, 2012 (see NHTSA-2012-0084 at www.regulations.gov) and conducted a listening session to hear additional comment on July 18, 2013. This notice reflected NHTSA's intent to upgrade the information technology, research design, data elements, and data collection methods to meet the needs of government agencies, industry and academia in the United States and abroad. NHTSA also sent the Federal Register Notice to more than 500 interested parties by letters and e-mail. These public stakeholders include:

- Automotive manufacturers,
- Government agencies,
- Universities and other research organizations, and
- Advocacy groups

More than 20 organizations and individuals submitted over 300 comments to NHTSA. The comments and suggestions received from data users outside of the NHTSA reflected similar needs to users within NHTSA. Comments regarding the importance and relevance of the various data systems were universally

positive. However, data users wanted to see NASS updated so it remains relevant. The comments covered a wide range of topics including the following.

- Data elements
- Data availability
- Sampling Plan
- Quality control
- Contracting
- Training
- Data collection

In addition to continuous interest in crashworthiness data, both internal and external comments indicated the motor vehicle safety initiatives are now and will continue to be largely focused on crash avoidance technologies, behavioral safety, and vehicle systems that can enhance human performance and vehicle control.

Some thought that the scope of the CISS should be broadened to include crashes involving motorcycles, commercial vehicles, pedestrians, bicycles, and other road users such as low speed vehicles and ATVs. Alternatively, it was suggested that the new CDS narrows its scope to collect data only on severe crashes, data of most interest to users, especially under constrained funding scenarios.

3.3 CISS Analytic Objectives

Based on the assessment of internal and public data needs, NHTSA determined that the purpose of the CISS is to gather accurate, detailed information about a nationally representative probability sample of passenger vehicle[1] crashes.

Crashes involving motorcycles, commercial vehicles, pedestrians, bicycles, and other road users such as low speed vehicles and ATVs are relatively rare crash populations. Capturing these crashes needs either a very large sample size or a sample design tailored for and targeted towards a particular type of crash. Motorcycle crashes, for example, are most likely to happen in warmer states and are concentrated in a fewer geographic areas and roadways. A sampling system for general passenger vehicle crashes with a relatively small sample size such as CISS will not be able to sample many motorcycle crashes. The most efficient way to study a rare population is to design a special study that solely targets that particular rare population. Therefore, NHTSA decided to capture motorcycle, pedestrian, bicycle and large truck crashes through CRSS because CRSS was planned to have a much larger sample size than CISS. If more information about these rare crash populations is needed, a special study will be designed using an appropriate sample. This approach will allow both CISS and the special study to be efficient for its own respective purpose. See Zhang, Noh, Subramanian and Chen (2018) for more information on CRSS.

[1] CISS passenger vehicles are in-transport automobiles, automobile derivatives, SUVs, van-based light trucks, light conventional trucks (pickup-style cab) and other light trucks with GVWR less than or equal to 4,536 kilograms or 10,000 lbs.

The data provided by the CISS may then be used for a variety of purposes including:

- Identifying emerging issues in vehicle safety.
- Examining detailed data on the crash performance of passenger cars, light trucks, vans, and utility vehicles.
- Evaluating vehicle safety systems and designs.
- Increasing knowledge of crash related injuries, including injury mechanisms.
- Assessing of the effectiveness of motor vehicle and traffic safety program standards.
- Designing future crash avoidance and crash mitigation technologies.

NHTSA determined that non-severe crash PAR strata would be necessary to estimate both crashworthiness and crash-avoidance measures of relative risk. Excluding the non-severe crash PAR strata would greatly jeopardize these types of analyses. Therefore, NHTSA decided *not* to narrow CISS's scope to only severe crashes.

3.4 CISS Target Population and Analysis Domains

From the assessment of the CISS analytic objectives, NHTSA has determined the target population for CISS shall be all police-reported motor vehicle crashes on a traffic way involving a passenger vehicle in which a passenger vehicle is towed from the scene for any reason. This is slightly different from the CDS target population, which required that the vehicle be towed due to damage. This change was made because sometimes it is difficult to determine why a vehicle was towed. This change will reduce the number of misclassified PARs and speed up the PAR listing process in the field.

CISS-applicable vehicles are the same as CDS-applicable vehicles: passenger cars, light trucks, vans, and sport utility vehicles with GVWR less than 10,000 lbs.

The police-reported motor vehicle crashes refer to motor vehicle crashes that result in police crash reports - the PARs. PARs are part of the national sampling frame for CISS.

The research questions and analytic objectives mentioned in the previous section also suggest specific important domains of analysis[2] for CISS. Table 1 lists these domains and their target percent of the total sample allocation. Two variables are used to identify these domains: the vehicle age and the injury severity. Unlike CDS, whether the injured person is transported is no longer considered. This should speed up the PAR listing process and reduce the number of misclassified PARs.

In Table 1 the "Target Percent of Sample Allocation" column specifies the desired distribution of the sampled cases – for example, "5%" in domain 1 means that 5 percent of the sampled cases would be selected from domain 1. The "Estimated Population" column is the expected population count for each analysis domain estimated from 2011 CDS. The "Population Percent" column is the population distribution of each analysis domain estimated from current NASS. If the "Population Percent" is lower

[2] Analysis domains: sub-populations of research interest.

than “Target Percent of Sampling Allocation,” then the corresponding analysis domain is oversampled relative to its incidence.

3.5 The Relationship Between the CRSS and the CISS Samples

In NASS the 24 CDS PSU sample was a subsample of the 60 GES PSU sample. In other words CDS was nested within GES. The main advantage of this nested design is cost savings from sharing resources between the two surveys. It may also allow the use of auxiliary information from the larger sample for estimation in the smaller sample.

The main disadvantage of a nested design is that it forces compromise in both survey designs since the set of PSUs selected must meet the needs of both surveys. For example, PSU formation and PSU sample selection must be the same for both surveys rather than tailored to the data needs and operational concerns of the specific survey.

NHTSA evaluated the possibility of nesting CISS within CRSS. It was determined that the cost savings that result from nesting CISS are mainly a reduction in the cost of driving from one police jurisdiction to another. This cost can be attenuated by reducing the number of visits per year.

On the other hand, there are major differences between CISS and CRSS that suggest that separate designs might be more efficient. These differences include:

- CISS and CRSS have different target populations: CISS targets towed passenger vehicle crashes while CRSS targets all police-reported crashes.
- CISS and CRSS have different operational requirements: CISS data collection requires vehicle inspection, driver interview, hospital visit and scene investigation which requires a lot travel. Therefore CISS PSUs must not exceed a certain geographic size in order to limit the travel time.

Because of the differences between CISS and CRSS, independently tailored PSU formation, stratification, PSU measure of size definitions, and sample selection can produce more efficient samples for both systems. To optimize both CISS and CRSS, NHTSA decided to design CISS independently from CRSS.

Table 1: CISS Analysis Domains, Crash Sample Size Allocation, and Population Sizes

CISS Analysis Domains	Description	Target Percent of Sample Allocation	Estimated Population	Population Percent
1	At least one occupant of towed passenger vehicle is killed	5%	9,576	0.51%
2	Crashes not in Stratum 1 involving: • A recent model year passenger vehicle in which at least one occupant is incapacitated	10%	17,304	0.93%
3	Crashes not in Stratum 1 or 2 involving: • A recent model year passenger vehicle in which at least one occupant is non-incapacitated, possibly injured or injured but severity is unknown.	20%	162,037	8.71%
4	Crashes not in Stratum 1-3 involving: • A recent model year passenger vehicle in which all occupants are not injured	15%	325,332	17.48%
5	Crashes not in Stratum 1-4 involving: • A mid-model year passenger vehicle in which at least one occupant is incapacitated	6%	23,739	1.28%
6	Crashes not in Stratum 1-5 involving: • A mid-model year passenger vehicle in which at least one occupant is non-incapacitated, possibly injured or injured but severity is unknown	12%	210,407	11.31%
7	Crashes not in Stratum 1-6 involving: • A mid-model year passenger vehicle in which all occupants are not injured	10%	418,702	22.51%
8	Crashes not in Stratum 1-7 involving: • An older model year passenger vehicle in which at least one occupant is incapacitated	6%	28,690	1.54%
9	Crashes not in Stratum 1-8 involving: • An older model year passenger vehicle in which at least one occupant is non-incapacitated, possibly injured or injured but severity is unknown.	10%	220,815	11.87%
10	Crashes not in Stratum 1-9 involving: • An older model year passenger vehicle in which all occupants are not injured	6%	443,151	23.83%
Total		100%	1,859,752	100%

Source: Estimated from 2011 CDS data.

Note: This table uses the following definitions.

- Recent Model Year (or Late Model Year) Vehicles: vehicles that are <= 4 years old.
- Mid-Model Year Vehicles: 5-9 year old vehicles
- Older Model Year Vehicles: vehicles that are 10 years old or older

4. An Overview of CISS Sampling Design

The target population of CISS is all police reported motor vehicle crashes on a traffic way that involve at least one passenger vehicle towed from the scene. A direct annual one-stage selection of a national crash probability sample is infeasible because it would require access to more than 5 million PARs in the nation. In some jurisdictions PARs can only be accessed by viewing paper copies at local police stations. Therefore, the CISS uses a three-stage sampling method to select a nationally representative probability sample from the target population. The PSU is a county or a group of adjacent counties. The secondary sampling unit (SSU) is a PJ or a group of police jurisdictions. The tertiary sampling unit (TSU) is a PAR.

NHTSA's data need studies identified important analysis domains. To meet the data needs, rare crashes need to be oversampled in order to provide enough cases for analysis. This oversampling introduces unequal selection probabilities to the CISS design.

Multi-stage and unequal selection probability sampling often inflates variances. To reduce the variance, stratification is implemented at every stage of the CISS sample selection.

Sample allocation and sample size determination are in part driven by the budget level which is currently unknown for future years and is likely to fluctuate. A fixed sample size allocation may not be suitable for variable budget scenarios. Reselecting the sample, either the PSU sample or PJ sample, may require the renewal of the data collection sites. Renewing PSUs is inefficient because of the high cost of setting up PSUs, establishing cooperation from PJs, and recruiting and training on-site crash investigation technicians. A major challenge to CISS has been to select a scalable sample to avoid reselecting the sample in the future when the budget changes. To this end, a multi-phase sampling method was designed for CISS sample selection. This multi-phase sampling method allows for the selection of a deeply stratified and scalable sample with minimum change to the existing sample if changes arise in the future.

In summary, the CISS uses a stratified, multi-stage, and multiphase sampling system with unequal selection probabilities and scalable sample sizes. In the following chapters the sampling frame, sample selection method, and sample allocation are discussed.

5. PSU Sample Selection

5.1 PSU Sampling Frame

The sampling frame refers to a device through which the units of the target population can be sampled and accessed. The CISS target population is all police-reported motor vehicle crashes on a traffic way involving a passenger vehicle and in which a passenger vehicle is towed from the scene. Also, the responding police officer must have filed a PAR for the crash. A one-stage direct selection of a national probability sample of crashes for time sensitive data collection would require timely access to all crashes or PARs in the nation, which is not feasible.

Instead, the country is partitioned into smaller areas called PSUs - a county or a group of adjacent counties for CISS. Then a probability sample of PSUs is selected and technicians are employed to collect time-sensitive information on selected crashes within the selected PSUs. This design is equivalent to grouping the crashes in the country into clusters (PSUs) and then selecting a probability sample of clusters. Accessing PARs and collecting data locally is much easier and more operationally efficient than a national one-stage crash sample. The drawback is that it introduces the clustering effect that often inflates the variances of the resulting estimates.

5.2 PSU Formation

The CISS is a follow-on and potentially on-scene data collection survey. That means technicians have to drive to the crash scene, tow yards, or wherever the case vehicles are located, as well as interview the drivers to collect various data. It becomes operationally inefficient when PSU's geographic area is too big. To better ensure efficient data collection, CISS PSUs were formed to be geographically contiguous and to meet a specified maximum end-to-end distance of a PSU: 65 miles for urban PSUs and 130 miles for rural PSUs.

Census region and urbanicity were identified as effective PSU stratification variables (see section 5.4 for more details on PSU stratification). The Office of Management and Budget defines the metropolitan statistical area (MSA) as one or more adjacent counties or county equivalents that have at least one urban core area of at least 50,000 population. In the CISS, an MSA is considered as urban area and CISS PSU formation respects this urbanicity definition. A CISS PSU was considered urban if there is an MSA in it and all other PSUs were considered rural. PSUs were also formed to respect region for effective PSU stratification. However, PSUs were allowed to cross state lines.

As shown in Table 1, the analysis domain 1 has the lowest population percent (i.e., 0.5% of all eligible crashes in the population). This makes the domain 1 PARs the rarest cases. For domain estimation, it is desirable that the rare cases be selected from as many PSUs as possible. The target sample allocation (desired portion of all sampled crashes) for domain 1 is 5 percent. Therefore, PSUs were formed so that 5 percent of the CISS sample would ideally consist of crashes having at least one fatality in a towed passenger vehicle. Each PSU will employ at least one technician, and one technician collects about 100 cases per year. Therefore, PSUs were formed with an estimated 90 percent probability of yielding at least 5 fatal crashes involving a passenger vehicle in a year. In order to do this, we assumed that the number of

fatal crashes involving a passenger vehicle in a PSU follows a Poisson distribution with mean λ (i.e., $X \sim \text{Poisson}(\lambda)$). Then, the probability of having at least 5 fatal crashes involving a passenger vehicle is

$$P(X \geq 5) = 1 - \sum_{x=0}^{4} e^{-\lambda} \frac{\lambda^x}{x!}$$

In the above equation, $P(X \geq 5) = 0.9$ when the expected value $\lambda = 8$. This means in order to have at least 5 CISS fatal crashes in 90% of the years, a PSU should have annual average of 8 CISS fatal crashes or more.

The CISS PSU frame was formed using Westat, Inc's proprietary PSU formation software WesPSU with consideration of the above factors. Starting with NHTSA's desirable distance constraint: 65 miles for urban and 130 miles for rural PSUs, and an initial λ value of 8, after several iterations of WesPSU, a total of 1,784 PSUs were formed from 3,117 counties in the country. During this process, the distance constraints and the values of λ were modified as necessary for the region and urbanicity where severe crashes are rare as shown in Table 2.

Table 2: Distance Constraints and λ Value Used for the CISS PSU Formation

Region - Urbanicity	**Distance (miles)**	$\boldsymbol{\lambda}$
Northeast - Urban	65	8
Northeast - Rural	130	8
Midwest - Urban	125	8
Midwest - Rural	150	3.9
South - Urban	80	6.7
South - Rural	150	3.9
West - Urban	250	8
West - Rural	250	3.9

The outlying counties that do not contain a city in Alaska and Hawaii were excluded because they are remote and have few crashes. See Appendix C for a complete list of excluded counties.

In summary, CISS PSUs were formed according to the following criteria.

- PSUs were formed as groups of adjacent counties
- PSUs were formed with driving distance constraints described in Table 2
- With 90 percent chance, PSUs have at least (with few exceptions) five CISS fatal crashes in a given year
- Only counties from the same region and urbanicity could be combined to form a PSU, but counties from different states could be combined
- Outlying areas of AK and HI were excluded (see Appendix C)

5.3 PSU Measure of Size

The measure of size (MOS) is a quantity used to assign the selection probability to the frame units for the unequal selection probability sampling. One of the main CISS analysis interest areas focuses on recent model year passenger vehicles and severe crashes. Since these crashes are rare in the population it is necessary to oversample them. Therefore, PSUs with more high interest crashes should be given a larger MOS so these PSUs are more likely to be selected and more high interest crashes can be selected.

Based on the internal and public data user's need, NHTSA identified the analysis domains 1, 2, 3, 4, 5, 6, and 8 in Table 1 as high interest domains and used the estimated population counts of these domains to calculate the composite MOS of PSU. Table 3 lists the high interest domains and their rescaled relative sample allocations along with the variables used to estimate domain population counts. The descriptions and source of these variables are listed in Table 4.

For each PSU i in the frame, the composite MOS defined as

$$MOS_i = \sum_{s=(1,2,3,4,5,6,8)} p_s * \frac{N_{i+s}}{N_{++s}}$$

The composite MOS is computed over the high interest domains listed in Table 3. In this formula, p_s represents the rescaled target sample allocation (column 3) in Table 3. N_{++s} is the estimated population counts in analysis domain s as defined in Table 3 and 4. N_{i+s} is the estimated population counts in analysis domain s and PSU i as defined in Table 3 and 4. The PSUs with the largest composite MOS are those that have unusually many crashes in one or more high interest domains, particularly in those domains with the highest sample allocations.

Table 3: High Interest Domains and Variables Used to Estimate Domain Population Counts

High Interest Domains	Description	Target Sample Allocation (Scaled to Seven Domains)	Variables Used to Estimate Domain Population Counts (See Table 4 Below)
1	Fatal	6.76%	FATAL_AVG_PV
2	Incapacitated Recent MY	13.51%	A07_08 × PROPNEW
3	Non-Incap injury Recent MY	27.03%	BC07_08 × PROPNEW
4	No Injury Recent MY	20.27%	ACSPop × PROPNEW
5	Incapacitated Mid MY	8.11%	A07_08 × PROPOLD
6	Non-Incap Injury Mid MY	16.22%	BC07_08 × PROPOLD
8	Incapacitated Older MY	8.11%	A07_08 × PROPOLD

Table 4: Descriptions and Sources of the Variables Used to Estimate High Interest Domain Population Counts

Variable Name	Description	Source [1)]
ACSPop	2010 ACS 1-Year Estimates of Population	ACS
PROPNEW	Proportion of passenger vehicles registered in 2011 that are model year 2007 or newer	POLK
PROPOLD	Proportion of passenger vehicles registered in 2011 that are model year 2006 or older	POLK
FATAL_AVG_PV	Average number of fatal crashes, 2007-2011	FARS
A07_08	Number of incapacitating injury crashes, 2007-08	SDS
BC07_08	Number of non-incapacitating and possible injury crashes, 2007-08	SDS

[1)] ACS - The U.S. Census Bureau's American Community Survey.
POLK - R. L. Polk & Company.[3]
FARS - NHTSA's Fatality Analysis Reporting System.
SDS - NHTSA's State Data System.[4]

5.4 PSU Frame Stratification

Stratification refers to the partitioning of the sampling frame into non-overlapping sub-populations in order to allow independent sample selection from each sub-population. A careful selection of stratification variables can produce a more balanced sample and reduce the variance of estimates of population parameters. Stratification also allows better control of the sample size for sub-population estimation. An efficient stratification variable forms homogeneous sub-populations, that is, it minimizes the within sub-population variance and maximizes the between sub-populations variances for variables of interest.

Census regions were used as a PSU stratification variable resulting in a more geographically balanced and representative PSU sample. The Census regions are Northeast, West, South, and Midwest.

Urbanicity was also used as a PSU stratification variable resulting in a more demographically balanced and representative PSU sample. Urbanicity also produces more homogeneous sub-populations. In CISS, urbanicity included two categories:

- Urban – the PSU includes at least one Metropolitan Statistical Area, and
- Rural – otherwise.

Census region and urbanicity formed eight (4×2) primary CISS PSU strata. Within each primary CISS PSU stratum, Westat, Inc.'s proprietary software WesStrat was used to further stratify the CISS PSUs within each primary PSU stratum.

WesStrat stratified PSUs into equal and homogeneous nested strata. Within each primary PSU stratum, PSUs with similar characteristics were grouped into nested strata with approximately equal MOS sizes. The software assists in finding the best-nested stratification scheme for minimizing the between-PSU

[3] Polk is an automotive data and marketing solutions provider. The data set NHTSA purchased from Polk contains vehicle counts, model year, manufacturer, and VMT at the county level.
[4] NHTSA's SDS is a collection of 34 states' computer data files coded from police accident reports. It contains basic crash information coded from PARs and is used to calculate crash counts by year and maximum injury severity.

variance within stratum, while attempting to make the stratum population MOS approximately equal. Stratification variables were identified independently within each primary stratum.

Candidate stratification variables for deeper stratification included:

- ROAD_TYPE_RATE: total highway/primary and secondary road miles divided by MOS;
- TOT_CRASH_RATE: summation of imputed 2008 Injury, imputed 2008 PDO, and 2007-2011 average fatal crashes divided by MOS; and
- VMT_RATE_IMP: imputed HPMS[5] vehicle miles traveled divided by MOS

The search process for the stratification variable maximized the effect of the stratification on the following study variables.

- The average number of fatal crashes across the years 2007-2011
- The sum of the 2008 and 2007 SDS "A" injury crashes (including imputed counts for non-SDS states)
- The sum of the 2008 and 2007 SDS "B" injury crashes (including imputed counts for non-SDS states)
- The number of new registered vehicles in 2011 (from POLK data)

Table 5 describes how the secondary PSU strata were formed within each primary PSU stratum. PSU 7-00 MOS is larger than the MOS of the smallest primary PSU stratum thereby was treated as if it is a stratum by itself. Twenty four secondary strata were formed from the eight primary strata. For example, primary stratum 1 is further partitioned into three secondary strata by the three categories of variable ROAD_TYPE_RATE: 0-225, 225-747, 747-7233. The single PSU stratum 7-00 is treated as a single PSU for PSU sampling purpose. Therefore, a total of 24 secondary PSU strata will be used for PSU sample selection.

[5] HPMS: Highway Performance Monitoring System maintained by DOT. The HPMS contains administrative and extent of system information on all public roads.

Table 5: CISS PSU Strata and PSU Population Counts

Primary Strata	Secondary Strata	VMT_RATE_IM		TOT_CRASH_RATE		ROAD_TYPE_RATE		Number of PSUs	Total Strata MOS
		Lower	Upper	Lower	Upper	Lower	Upper		
1	1-01					0	225	13	0.04841
1	1-02					225	747	21	0.05200
1	1-03					747	7233	61	0.04963
2	2-01	0	5871					30	0.01099
2	2-02	5871	3022					33	0.01122
3	3-01	0	5619			0	490	10	0.03684
3	3-02	0	5619			490	5817	51	0.03623
3	3-03	5619	1924	0.000	0.023			18	0.03630
3	3-04	5619	1924	0.023	0.096			58	0.03633
4	4-01	0	6047					130	0.03685
4	4-02	6047	2767					188	0.03701
5	5-01			0.000	0.024	0	398	12	0.04643
5	5-02			0.000	0.024	398	1530	24	0.04497
5	5-03			0.024	0.026			21	0.05311
5	5-04			0.026	0.032			39	0.04648
5	5-05			0.032	0.042			60	0.04957
5	5-06			0.042	0.138			155	0.04821
6	6-01	0	5774					242	0.06477
6	6-02	5774	4213					372	0.06473
7	7-00							1	0.02491
7	7-01	0	5368					22	0.04122
7	7-02	5368	8298					24	0.04261
7	7-03	8298	1568					22	0.04122
8	8-01			0.000	0.052			46	0.02001
8	8-02			0.052	0.212			131	0.01984
Total	**25**							**1784**	**1**

5.5 PSU Sample Selection

Unknown future funding levels and the need for a stable PSU sample require NHTSA to select a scalable PSU sample in which the PSU sample size can be decreased or increased with minimum impact to the existing PSU sample. To this end, a multi-phase sampling method was used to select the CISS PSU sample by selecting a sequence of nested PSU samples. In this method, a PSU sample larger than actually needed is initially selected. Then from this selected first-phase PSU sample, a smaller subset of PSU sample has been selected. Then from this second-phase PSU sample, another smaller third-phase PSU sample has been selected. This process continued until the PSU sample size reaches unacceptable levels. This way, a sequence of nested PSU samples has been obtained. Each of these PSU samples is a probability sample and can be used for data collection. If a larger or smaller PSU sample is desirable, the appropriate sample could be picked from the nested sequence (Figure 1). This allows us to calculate the selection probabilities and minimizes changes to the PSU sample. The following is a detailed description of how this process has been applied to CISS PSU sampling. For CISS, 7 PSU samples have been selected under the 7 scenarios of number of PSU strata and PSU sample sizes. Table 6 summarizes the CISS PSU sample scenarios which are described in more details in the subsequent sections.

Figure 1: Nested PSU Samples

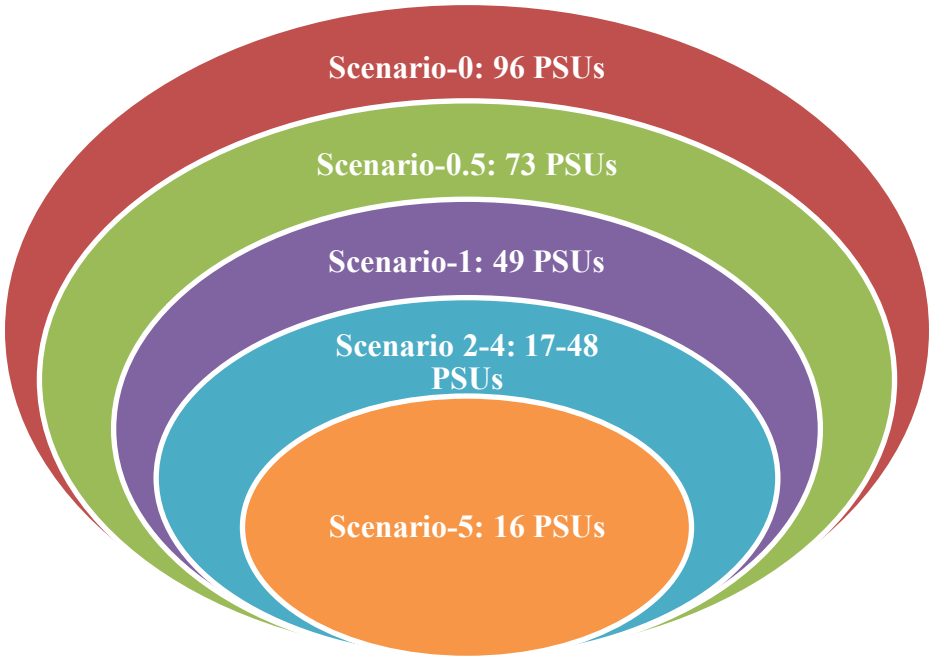


Table 6: Number of PSU Strata and Sampled PSUs for the 7 Scenarios

Scenario	Number of PSU Strata	Number of Certainty PSU	Total Number PSUs	PSU per Stratum
0	24	3	96	4 or 3
0.5	24	1	73	3
1	24	1	49	2
2	20	0	40	2
3	16	0	32	2
4	12	0	24	2
5	8	0	16	2

5.5.1 Scenario-1 PSU Sample

The initial (scenario-1) PSU sample has 48 non-certainty PSUs selected from the 24 secondary strata plus any certainty PSUs. The 24 secondary strata are listed under scenario-1 column of Appendix B. One PSU (7-00) has extraordinary large MOS and has become a certainty PSU under PPS sampling of size 48 from the entire PSU frame without stratification. It was set aside and treated as a stratum (7-00). Variance estimation within such a certainty PSU stratum treats the PJs as primary sampling units.

From each of the 24 secondary PSU strata, 2 PSUs have been selected using randomly-ordered systematic probability proportional to size (PPS) sampling, resulting in a total of 48 PSUs. Selecting two PSUs per stratum allows for both variance estimation within each stratum and near-maximum PSU stratification.

Use $h^{(s)}$ to denote stratum h under scenario-s. Let $S_{h^{(1)}}$ be the PSU population in stratum $h^{(1)}$ for scenario-1 sample selection purposes, and $MOS_{h^{(1)}i}$ be the original MOS assigned to PSU i in scenario-1 stratum h.

The PPS sampling interval for scenario-1 stratum h is its total PSU MOS divided by 2:

$$I_{h^{(1)}} = \frac{\sum_{g \in S_{h^{(1)}}} MOS_{h^{(1)}g}}{2}, \qquad h^{(1)} = 1, 2, \ldots 24.$$

A random number r has been generated to produce a random start and 2 PSUs selected by systematic PPS sampling from each of the 24 secondary PSU strata. In scenario-1, the selection probability for any non-certainty PSU i in stratum $h^{(1)}$ is:

$$\pi^{(1)}_{h^{(1)}i} = \frac{2MOS_{h^{(1)}i}}{\sum_{g \in S_{h^{(1)}}} MOS_{h^{(1)}g}}$$

5.5.2 Scenario-2 PSU Sample

The scenario-2 PSU sample has a smaller sample size of 40. These 40 PSUs have been subsampled from the 48 scenario-1 PSUs. To select the scenario-2 PSU sample, 4 of the 24 scenario-1 PSU *strata* have been collapsed with other strata to form 20 scenario-2 PSU strata. The collapsing of strata follows the following rule:

- Only the secondary strata in the same primary stratum could be collapsed;
- Only the contiguous secondary strata could be collapsed; and
- The resulting strata have similar stratum total MOS.

The resulting 20 strata are listed under the scenario-2 column of Appendix B. With PSU sample size 40, the PSU that constitutes the scenario-1 stratum 7-00 was no longer an overall certainty PSU and has been included in scenario-2 stratum 7-01 as a PSU.

After the 20 scenario-2 PSU strata have been formed, to select a subsample of the scenario-1 PSU sample, the scenario-1 sampled PSUs were used as the PSU population for scenario-2 PSU sample selection. Two (2) PSUs have been selected from each scenario-2 stratum. If a scenario-2 stratum was not the result of collapsed scenario-1 strata, it had only 2 PSUs in it and both of them were selected with certainty. For such a PSU from stratum $h^{(2)} = h^{(1)}$, the selection probabilities are:

$$\pi^{(2)}_{h^{(2)}i} = \pi^{(1)}_{h^{(1)}i} * 1 = \frac{2MOS_{h^{(1)}i}}{\sum_{g \in S_{h^{(1)}}} MOS_{h^{(1)}g}}$$

If a scenario-2 stratum had been collapsed from two scenario-1 strata, then it had 4 PSUs (except scenario-2 stratum 7-01 – there are 5 PSUs because the certainty PSU 7-00 was added), and each of them was assigned a *new* MOS equal to its scenario-1 *stratum* total MOS. That is, if PSU $i \in h^{(1)}$, then:

$$MOS_{h^{(2)}i} = \sum_{g \in S_{h^{(1)}}} MOS_{h^{(1)}g}$$

In general, scenario-2 stratum $h^{(2)}$ is the union of multiple collapsed scenario-1 strata. Let $J_{h^{(2)}}$ be the number of scenario-1 strata collapsed into scenario-2 strata $h^{(2)}$ and $\left\{h_j^{(1)}\right\}_{j=1}^{J_{h^{(2)}}}$ be those corresponding scenario-1 strata. Thus: $h^{(2)} = \cup_{j=1}^{J_{h^{(2)}}} h_j^{(1)}$. Here $J_{h^{(2)}}$ could be 1, 2, or 3 (see Appendix B). Let $n_{h^{(1)}}$ be the scenario-1 stratum $h^{(1)}$ PSU sample size. Since scenario-2 sample had been selected from scenario-1 sample, there were total $\sum_{j=1}^{J_{h^{(2)}}} n_{h_j^{(1)}}$ pooled PSUs available in scenario-2 stratum $h^{(2)}$ for selection. In other words, the stratum $h^{(2)}$ population size for scenario-2 sample selection equaled the sum of the sample sizes over the collapsed scenario-1 strata $\left\{h_j^{(1)}\right\}_{j=1}^{J_{h^{(2)}}}$. From each collapsed stratum, two PSUs have been selected from the pooled $\sum_{j=1}^{J_{h^{(2)}}} n_{h_j^{(1)}}$ PSUs using PPS sampling. The resulting PSU selection probability for $i \in h^{(2)}$ is:

$$\pi_{h^{(2)}i}^{(2)} = \pi_{h^{(1)}i}^{(1)} * \frac{2\sum_{g \in S_{h^{(1)}}} MOS_{h^{(1)}g}}{\sum_{j=1}^{J_{h^{(2)}}} n_{h_j^{(1)}} \sum_{g \in S_{h_j^{(1)}}} MOS_{h_j^{(1)}g}}$$

Typically, $n_{h_j^{(1)}} = 2$ and $\pi_{h^{(1)}i}^{(1)} = 2MOS_{h^{(1)}i} / \sum_{g \in S_{h^{(1)}}} MOS_{h^{(1)}g}$, therefore,

$$\pi_{h^{(2)}i}^{(2)} = \frac{2MOS_{h^{(1)}i}}{\sum_{g \in S_{h^{(1)}}} MOS_{h^{(1)}g}} * \frac{2\sum_{g \in S_{h^{(1)}}} MOS_{h^{(1)}g}}{\sum_{j=1}^{J_{h^{(2)}}} 2\sum_{g \in S_{h_j^{(1)}}} MOS_{h_j^{(1)}g}}$$

$$= \frac{2MOS_{h^{(1)}i}}{\sum_{j=1}^{J_{h^{(2)}}} \sum_{g \in S_{h_j^{(1)}}} MOS_{h_j^{(1)}g}}$$

This is the same selection probability of selecting 2 PSUs directly from the collapsed scenario-1 strata using PPS in one stage.

For example, scenario-2 stratum 3-02 is the result of collapsing two scenario-1 strata: 3-03 and 3-04 (Appendix B), each with 2 sampled PSUs (PSU 12, 44, 19, and 41). The selection probability for each of the 2 PSUs (PSU 12 and 19) selected from scenario-2 stratum 3-02 is:

$$\pi_{h^{(2)}i}^{(2)} = \frac{2MOS_{h^{(1)}i}}{\sum_{g \in S_{3-03^{(1)}}} MOS_{3-03^{(1)}g} + \sum_{g \in S_{3-04^{(1)}}} MOS_{3-04^{(1)}g}}$$

This is equivalent to the selection probability of selecting 2 PSUs directly from the combined scenario-1 strata 3-03 and 3-04 using PPS in one stage.

5.5.3 Scenario-3 5 PSU Samples

Scenario-3-5 PSU samples have been selected in a similar way as the scenario-2 sample. Each scenario PSU sample was a subsample of the PSU sample from the previous scenario. Each PSU stratum was either the same PSU stratum from the previous scenario or collapsed from previous scenario PSU strata. In other words, the scenario PSU samples were nested. The resulting selection probabilities remain PPS in general.

5.5.4 Scenario-0 PSU Sample

Once CISS is established, other studies may utilize CISS infrastructure to collect data. These studies may require more than 49 PSUs. For that purpose, the 49 scenario-1 PSU sample was expanded to a 96 PSU sample (scenario-0). To this end, the same 24 PSU strata for scenario-1 have been used to select 4 PSUs per stratum using PPS with the same MOS but a sampling interval that is exactly half of the scenario-1 sampling interval, as shown below:

$$I_{h^{(0)}} = \frac{I_{h^{(1)}}}{2} = \frac{\sum_{g \in S_{h^{(1)}}} MOS_{h^{(1)}g}}{4}, \qquad h = 1, 2, \dots 24$$

The same random number r used for selecting the 49 PSU sample was multiplied by $I_{h^{(0)}}$ to generate the random start for scenario-0 sampling. In this way, the 49 scenario-1 PSU sample were nested in this 96 PSU sample. During this process, although there are multiple certainty PSUs, none of them was set aside and only one certainty PSU in stratum 3-01 (100_3) was selected twice therefore stratum 3-01 only has 3 PSUs. This resulted in total 96 PSUs: 1 overall certainty PSU (stratum 7-00), 3 PSUs from stratum 3-01 and 4 PSUs from each of the rest 23 strata. Certainty PSUs have selection probability 1. The other PSUs have selection probability:

$$\pi^{(0)}_{h^{(0)}i} = \frac{4MOS_{h^{(1)}i}}{\sum_{g \in S_{h^{(1)}}} MOS_{h^{(1)}g}}$$

In scenario-0 sample selection, in each stratum h there are four systematically and sequentially selected PSUs: $\{u_h^1, u_h^2, u_h^3, u_h^4\}_{h=1}^{24}$. Some of them may be the same PSU if they were selected more than once. The scenario-1 sampled PSU is either in $\{u_h^1, u_h^3\}_{h=1}^{24}$ or in $\{u_h^2, u_h^4\}_{h=1}^{24}$ if selected once, or in both if selected more than once. We can treat scenario-1 sample as a sub-sample of scenario-0 sample by randomly selecting one pair from the two pairs: $\{u_h^1, u_h^3\}_{h=1}^{24}$ and $\{u_h^2, u_h^4\}_{h=1}^{24}$ in each scenario-0 stratum. This random selection has been performed using the same random number r generated for the random start: if the random number r is less than or equal to 0.5 then $\{u_h^1, u_h^3\}_{h=1}^{24}$ is selected, if the random number r is greater than 0.5 but less than or equal to 1 then $\{u_h^2, u_h^4\}_{h=1}^{24}$ is selected. In this way, as a sub-sample of scenario-0 sample, if a scenario-1 PSU was selected only once at scenario-0, then it's selection probability would be:

$$\pi^{(1)}_{h^{(1)}i} = \pi^{(0)}_{h^{(0)}i} * \frac{1}{2} = \frac{2MOS_{h^{(1)}i}}{\sum_{g \in S_{h^{(1)}}} MOS_{h^{(1)}g}}$$

If a scenario-1 PSU was selected more than once at scenario-0, then this PSU must be a scenario-1 certainty PSU as well as a scenario-0 certainty PSU therefore $\pi^{(0)}_{h^{(0)}i} = 1$. Since this PSU were selected more than once, it's scenario-1 selection probability would be:

$$\pi^{(1)}_{h^{(1)}i} = \pi^{(0)}_{h^{(0)}i} * \frac{2}{2} = 1$$

which means it is selected with certainty into scenario-1 sample. In this way, the scenario-1 sample, although actually selected before scenario-0 sample, can be viewed as a subsample of scenario-0 sample.

5.5.5 Scenario-0.5 PSU Sample

Scenario-0.5 is between scenario-0 and scenario-1 and uses the same 24 PSU strata as scenario-0 and scenario-1. Scenario-0.5 sample size is 73: one overall certainty PSU (stratum 7-00) and 3 PSUs from each of the 24 scenario-1 PSU strata. In each scenario-1 stratum, there are 4 PSUs selected as the scenario-0 sample except stratum 3-01 where there are only 3 PSUs selected as scenario-0 sample.

To select the scenario-0.5 sample, we first selected the 49 scenario-1 PSUs from the scenario-0 sample as describe in above subsection. For the remaining 24 PSUs, we first sort the 24 scenario-1 strata by the stratum variances. Then we increased PSU sample size for each of the 24 scenario-1 strata by 1. To select this extra one PSU from each stratum, the remaining PSUs that were in the scenario-0 sample but not selected into the scenario-1 sample were randomly sorted in each stratum and the first one on the list was selected. Let $n_h^{(0)}$ be the number of PSUs in scenario-0 sample, $n_h^{(1)}$ be the number of PSUs in scenario-1 sample for the same stratum h. There are total $n_h^{(0)} - n_h^{(1)}$ PSUs that were in scenario-0 sample but not in scenario-1 sample for scenario-1 stratum h. In this way, the scenario-0.5 PSUs have selection probability:

$$\pi^{(0.5)}_{h^{(0.5)}i} = \pi^{(1)}_{h^{(1)}i} + \left(\pi^{(0)}_{h^{(0)}i} - \pi^{(1)}_{h^{(1)}i}\right) * \frac{1}{n_{h^{(0)}} - n_{h^{(1)}}}$$

5.5.6 Between Scenario PSU Sample

To select any PSU sample with size between two scenario sample sizes, first the scenario PSU samples were randomly sorted in the following sequence (the sorted sample order is in the Appendix B).

1. PSUs #1-16: Random sort of the 16 PSUs in the 16-PSU scenario-5 sample.
2. PSUs #17-24: Random sort of the additional 8 PSUs in the 24-PSU scenario-4 sample.
3. PSUs #25-32: Random sort of the additional 8 PSUs in the 32-PSU scenario-3 sample.
4. PSUs #33-40: Random sort of the additional 8 PSUs in the 40-PSU scenario-2 sample.
5. PSUs #41-49: Random sort of the additional 9 PSUs in the 49-PSU scenario-1 sample.

6. PSUs #50-73: The 24 PSU strata were sorted by a composite stratum variance from largest to smallest so the additional 24 PSUs in the 73-PSU scenario-0.5 sample can be allocated in that order.
7. PSUs #74-96: The additional 24 PSUs in the 96-PSU scenario-0 sample were allocated using the same PSU strata order determined in 6).

The sorting mechanism listed above was used to determine for any given PSU sample size: (1). the strata to be used for the between scenario sample selection and (2). the number of PSUs to be selected from each between scenario stratum.

The strata sorting order in 6) and 7) was determined by the sizes of the 24 PSU strata. First, for each stratum the variances for the four outcome variables (total number of new vehicles, average fatalities in a new passenger vehicle, type-A crashes in 2007 and 2008, and type-B crashes in 2007 and 2008, see Table 4) were calculated. Then the variances were scaled by dividing the variance for each variable by the total for that variable. Finally the scaled variances are summed up across the four variables by strata to give a single sorting variable.

Use $(a-1)\sim a$ to denote between scenario-a and scenario-$(a-1)$. In general, scenario-a stratum $h^{(a)}$ has been collapsed from multiple scenario-$(a-1)$ strata: $h^{(a)} = \bigcup_{j=1}^{J_{h^{(a)}}} h_j^{(a-1)}$. Let $h^{((a-1)\sim a)}$ be the between scenario stratum to be determined. Depending on the sample sizes, $h^{((a-1)\sim a)}$ would be either a scenario-a stratum or scenario-$(a-1)$ strata. To determine $h^{((a-1)\sim a)}$, let $n_{h^{(a-1)}}$ be the scenario-$(a-1)$ stratum $h^{(a-1)}$ PSU sample size, $n_{h^{(a)}}$ be the scenario-a stratum $h^{(a)}$ PSU sample size, and $b_{h^{((a-1)\sim a)}}$ be the number of PSUs in stratum $h^{(a)}$ with sample order (determined by the above sorting) lower than or equal to the given PSU sample size (the total between scenario PSU sample size). In general, $\sum_{j=1}^{J_{h^{(a)}}} n_{h_j^{(a-1)}} \geq b_{h^{((a-1)\sim a)}} \geq n_{h^{(a)}}$. If $\sum_{j=1}^{J_{h^{(a)}}} n_{h_j^{(a-1)}} = b_{h^{((a-1)\sim a)}}$, let $h^{((a-1)\sim a)}$=$h^{(a-1)}$ – i.e. use scenario-$(a-1)$ strata. If $\sum_{j=1}^{J_{h^{(a)}}} n_{h_j^{(a-1)}} > b_{h^{((a-1)\sim a)}}$, then let $h^{((a-1)\sim a)}$=$h^{(a)}$ – i.e. use scenario-a stratum. $b_{h^{((a-1)\sim a)}}$ is the between-scenario sample size for stratum $h^{((a-1)\sim a)}$.

For example, if a total 30 of PSUs (between 24 scenario-4 PSUs and 32 scenario-3 PSUs) are to be selected, stratum 1-$01^{(4)}$ is the same as 1-$01^{(3)}$, and there are only 2 PSUs (PSU 20 and 6) with sample order 30 or lower. Notice $n_{1-01^{(3)}} = b_{1-01^{(3\sim 4)}} = n_{1-01^{(4)}}$=2 so scenario-3 stratum 1-$01^{(3)}$ should be used. For scenario-4 stratum 2-$01^{(4)}$, notice it was collapsed from two scenario-3 strata: 2-$01^{(3)}$ and 2-$02^{(3)}$, each has 2 PSUs for scenario-3 (sample order 32 or lower). There are 3 PSUs (PSU 13, 26, and 5) with sample order 30 or lower in stratum 2-$01^{(4)}$. $n_{2-01^{(3)}} + n_{2-02^{(3)}} = 4 > b_{2-01^{(3\sim 4)}} = 3$. Therefore scenario-4 stratum 2-$01^{(4)}$ should be used. Also notice 4-$01^{(4)}$= 4-$01^{(3)}$ ∪4-$2^{(3)}$ and there are 4 PSUs (PSU 7, 28, 11, and 27) with sample order lower than or equal to 30 and 32: $n_{4-01^{(3)}} + n_{4-02^{(3)}} = b_{4-01^{(3\sim 4)}} = 4$. Therefore, Scenario-3 strata 4-$01^{(3)}$ and 4-$02^{(3)}$ should be used.

The between scenario PSU samples selection, and the between scenario selection probabilities are then determined by the sizes of three counts: $\sum_{j=1}^{J_{h^{(a)}}} n_{h_j^{(a-1)}}$, $b_{h^{((a-1)\sim a)}}$, and $n_{h^{(a)}}$. There are three different situations:

(1). If $b_{h^{((a-1)\sim a)}} = \sum_{j=1}^{J_{h^{(a)}}} n_{h_j^{(a-1)}}$, then this becomes the exact scenario-$(a-1)$ sample selection. The scenario-$(a-1)$ strata $\left\{h_j^{(a-1)}\right\}_{j=1}^{J_{h^{(a)}}}$ and corresponding sample sizes $\left\{n_{h_j^{(a-1)}}\right\}_{j=1}^{J_{h^{(a)}}}$ should be used, and the between scenario selection probability would be:

$$\pi_{h^{((a-1)\sim a)}_i}^{((a-1)\sim a)} = \pi_{h^{(a-1)}_i}^{(a-1)}$$

(2). If $\sum_{j=1}^{J_{h^{(a)}}} n_{h_j^{(a-1)}} > b_{h^{((a-1)\sim a)}} > n_{h^{(a)}}$, the scenario-$a$ stratum $h^{(a)}$ would be used. And the between scenario sample size is $b_{h^{((a-1)\sim a)}}$. To select these $b_{h^{((a-1)\sim a)}}$ PSUs, the $n_{h^{(a)}}$ PSU scenario-a sample would be first selected from the $\sum_{j=1}^{J_{h^{(a)}}} n_{h_j^{(a-1)}}$ PSUs in $h^{(a)}$. The remaining $b_{h^{((a-1)\sim a)}} - n_{h^{(a)}}$ PSUs are the first $b_{h^{((a-1)\sim a)}} - n_{h^{(a)}}$ PSUs on the randomly sorted list for between-scenario $(a-1)\sim a$ above. This can be viewed as a simple random sample of size $b_{h^{((a-1)\sim a)}} - n_{h^{(a)}}$ selected from the $\sum_{j=1}^{J_{h^{(a)}}} n_{h_j^{(a-1)}} - n_{h^{(a)}}$ PSUs on the list. In this way, a selected PSU would be either selected into the scenario-a sample, or not selected into the scenario-a sample but then selected from the simple random sampling. Therefore, the selection probabilities for these $b_{h^{((a-1)\sim a)}}$ PSUs are:

$$\pi_{h^{((a-1)\sim a)}_i}^{((a-1)\sim a)} = \pi_{h^{(a)}_i}^{(a)} + \left(\pi_{h^{(a-1)}_i}^{(a-1)} - \pi_{h^{(a)}_i}^{(a)}\right) * \frac{b_{h^{((a-1)\sim a)}} - n_{h^{(a)}}}{\sum_{j=1}^{J_{h^{(a)}}} n_{h_j^{(a-1)}} - n_{h^{(a)}}}$$

For example, $b_{2-01^{(3\sim 4)}}$ =3 is between $n_{2-01^{(3)}} + n_{2-02^{(3)}}$ =4 and $n_{2-01^{(4)}}$=2. We first select the 2 PSU scenario-4 sample then select 1 PSU randomly from the remaining 2 PSUs that were not selected into the scenario-4 sample. The selection probabilities for these 3 PSUs are:

$$\pi_{2-01^{(3\sim 4)}_i}^{(3\sim 4)} = \pi_{2-01^{(4)}_i}^{(4)} + \left(\pi_{2-01^{(3)}_i}^{(3)} - \pi_{2-01^{(4)}_i}^{(4)}\right) * \frac{1}{2} = \frac{1}{2}\left(\pi_{2-01^{(4)}_i}^{(4)} + \pi_{2-01^{(3)}_i}^{(3)}\right)$$

(3). If $\sum_{j=1}^{J_{h^{(a)}}} n_{h_j^{(a-1)}} > b_{h^{((a-1)\sim a)}} = n_{h^{(a)}}$, then this becomes the exact scenario-a sample selection. The scenario-a stratum $h^{(a)}$ and sample size $b_{h^{((a-1)\sim a)}} = n_{h^{(a)}}$ would be used. The selection probability is:

$$\pi_{h^{((a-1)\sim a)}_i}^{((a-1)\sim a)} = \pi_{h^{(a)}_i}^{(a)}$$

5.5.7 Between Scenario-1 and Scenario-0.5 PSU Sample

Any PSU sample size more than 49 and less than 73 is a between scenario-1 and scenario-0.5 situation. The PSU strata were the same 24 strata for scenario-1. The sampling method and selection probability are the same as scenario-0.5 except not all 24 PSU strata had 3 PSUs selected – some of them had PSU sample size 2. The sorting order listed earlier determined the sample size for each stratum. To calculate the selection probability, if a stratum had sample size 3,

$$\pi_{h^{(1\sim 0.5)}_i}^{(1\sim 0.5)} = \pi_{h^{(1)}_i}^{(1)} + \left(\pi_{h^{(0)}_i}^{(0)} - \pi_{h^{(1)}_i}^{(1)}\right) * \frac{1}{n_{h^{(0)}} - n_{h^{(1)}}}$$

If a stratum had sample size 2, then it becomes scenario-1 sample selection. Therefore,

$$\pi_{h^{(1\sim0.5)}i}^{(1\sim0.5)} = \pi_{h^{(1)}i}^{(1)}$$

5.5.8 Between Scenario-0.5 and Scenario-0 PSU Sample

Any PSU sample size more than 73 and less than 96 is a between scenario-0.5 and scenario-0 situation. The PSU strata were still the same 24 strata for scenario-1. The sampling method and selection probability are the same as scenario-0.5 except now all 24 PSU strata had at least 3 PSUs selected – some of them had PSU sample size 4. The sorting order listed earlier determined the sample size for each stratum. To calculate the selection probability, if a stratum had sample size 3,

$$\pi_{h^{(0.5\sim0)}i}^{(0.5\sim0)} = \pi_{h^{(1)}i}^{(1)} + \left(\pi_{h^{(0)}i}^{(0)} - \pi_{h^{(1)}i}^{(1)}\right) * \frac{1}{n_{h^{(0)}} - n_{h^{(1)}}}$$

If a stratum had sample size 4, then it becomes scenario-0 sample selection. Therefore,

$$\pi_{h^{(0.5\sim1)}i}^{(0.5\sim1)} = \pi_{h^{(0)}i}^{(0)}$$

6. SSU Sample Selection

6.1 SSU Sampling Frame

PARs are written by police officers and reported to their PJ. To select PARs, technicians need to obtain PARs from the PJs first. Therefore, for CISS purposes, PJs or groups of PJs are viewed as natural clusters of PARs and become the secondary sampling units (SSU). For a large PSU with many SSUs, a probability sample of SSUs can be selected. Technicians need visit only the selected SSUs to obtain the PARs for investigations. Forming SSUs and then selecting a probability sample of them reduces the operational cost of PAR sample selection.

The SSU frame for a selected PSU is the collection of all PJs that report crashes that occur in the selected PSUs to the state. A single state police office may generate PARs for multiple PSUs. In that case, the state police office is treated as multiple PJs that correspond to the portion of PARs generated for the corresponding PSU.

To create the PJ frame, NHTSA collected PJ frame information for the PJs that reported crash data to the States in the years 2010-2012 in the 73 sampled PSUs in scenario-0.5. For the 73 PSUs, NHTSA identified PJ names, PJ addresses, and 6 different PJ level crash counts through NHTSA's Regional Offices, the PJs themselves, the states, and internet research.

The following 6 types of crash counts were collected.

- Total crashes
- Fatal crashes
- Injury crashes
- Pedestrian crashes
- Motorcycle crashes
- Commercial motor vehicle crashes

The PJ frames and the crash counts were reevaluated during the process of establishing PJ cooperation. Discrepancies were corrected to ensure efficient sample selection.

6.2 SSU Measure of Size

Similar to PSU MOS definition, it is sensible to stratify PJs by their sizes to reduce the weight variation and assign a larger selection probability to PJs with a higher incidence of high-interest crashes. To this end, two PJ MOS variables were created.

First, a coarse PJ MOS was created using the six PJ frame crash counts and the target sample allocation by PAR domain in Table 1 as following:

$$MOS_{j|i} = 0.05 \times (Fatal\ crashes) + 0.64 \times (Injury\ crashes) + 0.31 \times (Total\ crashes - Fatal\ crashes - Injury\ crashes - Pedestrian\ crashes - Motorcycle\ crashes - Cormercial\ motorvehicle\ crashes)$$

This rough PJ MOS is used for PJ stratification (see below).

Second, a finer PJ MOS was created for PJ selection probability. The six PJ frame crash counts were used to estimate the ten PAR domain counts in Table 1 for each PJ in the selected PSUs. Then the ten CISS PAR domain counts were estimated from the nine CRSS PAR strata counts using a regression model. The SSU MOS is then defined as follows:

$$MOS_{j|i} = \sum_{s=1}^{10} \frac{n_{++s}}{n} \frac{N_{ijs}}{N_{++s}}$$

n = the desired total sample size (of PARs)

n_{++s} = the desired sample size of crashes in analysis domain s

N_{++s} = the estimated population of crashes in PAR domain s

N_{ijs} = the estimated population number of crashes in domain s, PJ j and PSU i

This finer PJ MOS was used to assign PJ selection probabilities (see below).

6.3 SSU Stratification

PJ MOS varies dramatically within the selected PSUs. To reduce the weight variation, the PJ frame within each selected PSU was stratified by the coarse PJ MOS.

First, certainty PJs were identified as overall certainties using the following condition:

$$\frac{2MOS_{ij}}{\sum_g MOS_{ig}} \geq 1$$

Here MOS_{ij} is the PJ MOS of PJ j in PSU i. The summation is over all PJs in the PSU. After removing the identified certainty PJs, this process was repeated one more time to find certainty PJs. In the second process, secondary certainty PJs was also identified with the following condition:

$$1 > \frac{2MOS_{ij}}{\sum_g MOS_{ig}} > 0.7$$

All overall certainty PJs and secondary certainty PJs identified through the above process were assigned to the certainty stratum, that is, they were selected with certainty.

In addition, if a PSU had less than 5 PJs, all its PJs were assigned to certainty stratum.

If a PSU had greater than 4 but less than 9 PJs, no further PJ stratum was formed. All non-certainty PJs were grouped into one non-certainty stratum. As a result, this PSU has two PJ strata: a certainty stratum and a non-certainty stratum.

For the remaining PSUs, non-certainty PJs were sorted by their PJ MOS within each selected PSU. Half of the PJs with larger MOS were assigned to the large MOS stratum and the other half of the PJs with

smaller MOS were assigned to the small MOS stratum. As a result, for the PSUs with 9 or more PJs, three PJ strata were formed: the certainty stratum, the large MOS stratum, and the small MOS stratum.

6.4 SSU Sample Selection

One of the major challenges of the SSU sample selection is changes to the PJ frame. Unlike PSUs, PJs are relatively unstable as new PJs may emerge or existing PJs may split, merge or closedown. The PJ MOS is a function of various crash counts of the PJ and the PSU. Therefore, PJ MOS varies every year. In addition, setting up cooperation with the PJs is time consuming and there is a chance that PJs may refuse to cooperate in this effort.

To address these challenges, Pareto sampling was used to select the SSU sample. The Pareto sampling method (see Rosén, 1997) produces an approximate PPS sample and is able to handle the frame changes by controlling changes to the existing sample.

The Pareto sampling method was applied to the PJ sample selection for each of non-certainty PJ strata within the sampled PSU i, as following:

1. Generate a permanent random number (PRN) $r_{ij} \sim U(0,1)$ for each PJ j in the PJ frame.
2. Identify certainty PJs within the non-certainty stratum by condition:

$$\frac{m_i * MOS_{ij}}{\sum_{j=1}^{M_i} MOS_{ij}} \geq 1$$

   Here m_i is the PJ sample size and M_i is the PJ frame size for a PJ stratum within PSU i. MOS_{ij} is the finer PJ MOS. The identified certainty PJs are set aside. This process is repeated with the remaining PJs based on the reduced PJ sample size until there is no more certainty PJs within the non-certainty stratum. Let the total number of certainty PJs be m_c.

3. For the remaining $M_i - m_c$ non-certainty PJs in the frame, calculate each PPS inclusion probability with non-certainty sample size $(m_i - m_c)$:

$$p_{ij} = \frac{(m_i - m_c)MOS_{ij}}{\sum_{j=1}^{M_i - m_c} MOS_{ij}}$$

4. Calculate the transformed random numbers:

$$\left\{\frac{r_{i1}(1-p_{i1})}{p_{i1}(1-r_{i1})}, \frac{r_{i2}(1-p_{i2})}{p_{i2}(1-r_{i2})}, \cdots, \frac{r_{i(M_i-m_c)}(1-p_{i(M_i-m_c)})}{p_{i(M_i-m_c)}(1-r_{i(M_i-m_c)})}\right\}$$

5. Sort the transformed random number in ascending order.
6. The m_c certainty PJs plus the first $m_i - m_c$ non-certainty PJs on the above list are the PJ sample for a PJ stratum within PSU i.

In Pareto sampling, once a permanent random number is assigned to a PJ, it will never change. Therefore, unless the PJ MOS changes, the transformed random number: $\frac{r_{ij}(1-p_{ij})}{p_{ij}(1-r_{ij})}$ will not change either. If an existing PJ is closed, the corresponding transformed random number is dropped from the sorted list. If a new PJ is added to the frame, a new transformed random number is calculated and inserted to the sorted

list according to its magnitude. As a result, when PJ sample must be re-selected, the change to the existing PJ sample under Pareto sampling is likely to be much smaller than under a more conventional method of PPS sampling.

NHTSA conducted a simulation study on the described Pareto sampling strategy. The result of this study shows Pareto selection probability is very close to PPS selection probability for the CISS PJ sample selection (Noh et al., 2017). Therefore, the PPS selection probability is used to approximate the true Pareto selection probability for the non-certainty PJs in CISS.

The number of SSUs selected for data collection was determined by the budget level and the optimum sample allocation. See Chapter 8 for more information about the optimization. First all PJs in the certainty stratum are selected. Then, the SSU sample size determined from the optimization was allocated to the two non-certainty PJ strata proportionally to the total stratum PJ MOS (using the finer PJ MOS) with at least one PJ per stratum.

7. TSU Sample Selection

CISS tertiary sampling units refer to the PARs. In this chapter we describe how the CISS PAR sample was selected.

7.1 TSU Sampling Frame

CISS TSU sampling frame, or PAR sampling frame, consists of all CISS in-scope PARs produced in the sampled PJs. After establishing corporation with the selected PJs, technicians visit the selected PJs weekly to obtain the PARs accumulated since last visit for PAR sample selection. This process is referred as PAR listing. The PARs listed from the sampled PJs are used as the TSU sampling frame. For a few very large PJs with large number of PARs, only a systematic sample of PARs is listed. This process is referred as PAR sub-listing. The sub-listed PARs are systematic sample of all in-scope PARs in the selected PJs. If one of every L PARs is sub-listed in PJ j, PSU i, sub-listing factor is L and the sub-listing probability for PAR l is:

$$\pi_{l|ij} = \frac{1}{L}$$

7.2 TSU Classification

CISS PAR listing and sampling is conducted weekly to prevent the selection of older crashes with a lot of missing data elements. Every week within each PSU, technicians list PARs from the selected PJs. In this listing process, PARs are grouped into 10 CISS PAR domains defined in Table 1. After listing is finished, all the listed PARs from the selected PJs in the same PSU are pooled together for weekly PAR sample selection. In this way, the PAR frame is in fact stratified by the weeks of the year.

Each PSU typically has 1 to 2 technicians and each technician can investigate no more than 2 cases per week. With 1 to 4 cases to be selected per week, it is impossible to stratify the PARs into the 10 PAR domains and select at least one case per domain every week. Therefore, PAR sample is selected by Pareto sampling using PAR measure of size (PAR MOS) without further PAR stratification other than the weeks.

7.3 TSU Measure of Size

To ensure the desired sample allocation for PAR domains in Table 1, PAR MOS needs to be carefully calculated. CISS PAR MOS is determined the similar way as the CDS.

First, PAR MOS factor (f_s) is estimated for each PAR domain by simulation to ensure the target sample distribution in Table 1 is achieved approximately on average. Then, PAR MOS is computed by

multiplying PAR MOS factor (f_s), PJ weight ($w_{j|i}$) and sub-listing factor ($w_{l|ij}$). PAR MOS assigned to each listed PAR is:

$$MOS_{ijlsk} = f_s w_{j|i} w_{l|ij}$$

In this way, all PARs listed in the same PAR domain in the same PJ and PSU have the same PAR MOS. This method generates approximately desirable sample allocation.

7.4 TSU Sample Selection

After the PAR MOS is assigned to listed PARs, the PAR sample is selected using the Pareto sampling method in week v and PSU i as following:

1. For each listed PAR k from PJ j, generate a permanent uniform random number $u_{ijvlk} \sim U(0,1)$. Here subscript l is for sub-listing.
2. Identify certainty PARs in the following steps:
   (a) Calculate the sum of MOS over all listed PARs in week v and PSU i as
   $$MOS_{iv}^{sum} = \sum_j \sum_k MOS_{ijvlk}$$
   (b) Calculate the relative MOS for each listed PAR as
   $$MOS_{ijvlk}^{relative} = \frac{MOS_{ijvlk}}{MOS_{iv}^{sum}}$$
   (c) PAR_{ijvlk} is identified as a certainty if
   $$n_{iv} \times MOS_{ijvlk}^{relative} \geq 1$$
   Here n_{iv} is the PAR sample size (i.e., weekly caseload).
   (d) Set aside the identified certainty PARs, and reduce the PAR sample size by the number of certainty PARs. Then, for the remaining listed PARs, repeat (a)~(d) until there are no more certainty PARs (notice the certainties are removed from the computation of MOS_{iv}^{sum}).
3. Let the number of certainty PARs identified through 2(a)~2(d) be $n_{iv}^{certainty}$. Select these certainty PARs as the PAR sample for the week v and PSU i.
4. For the remaining listed PARs (excluding the certainty PARs), calculate inclusion probability:
   $$p_{ijvlk} = (n_{iv} - n_{iv}^{certainty}) \frac{MOS_{ijvlk}}{\sum_{m=1}^{N_{iv}-n_{iv}^{certainty}} MOS_{ijvlm}}$$
   Then calculate transformed random numbers:
   $$t_{ijvlk} = \frac{u_{ijvlk} \cdot (1 - p_{ijvlk})}{p_{ijvlk} \cdot (1 - u_{ijvlk})}$$
5. Sort the transformed random number in ascending order.
6. Select the first ($n_{iv} - n_{iv}^{certainty}$) PARs on the above sorted list as the non-certainty PAR sample for the week v and PSU i.

7. If a selected PAR turns out to be a non-responding case subject to certain non-response criteria, which is described in detail in the end of this section, a replacement PAR is selected as follows:
   (a) Increase PAR sample size by 1 and let the new sample size be n_{iv}.
   (b) Repeat the sample selection procedure 2~6 with the new PAR sample size.
   (c) Compare the selected sample from 6(b) with the previously selected sample.
      (c1). If the new PAR sample results in no more than one new PAR, then the $n_{iv}^{certainty}$ certainty PARs and the first $(n_{iv} - n_{iv}^{certainty})$ non-certainty PARs on the sorted list become the PAR sample (i.e., one new PAR becomes a replacement PAR for the non-responding PAR). If there are more non-responding PARs to be replaced, repeat 7(a)-7(c) one case a time.
      (c2). If the new PAR sample results in more than one new PAR, then at least one responding case would be replaced. In this case, the previously selected sample is the final PAR sample, and the previous sample size is the PAR sample size. There is no replacement sample selection for the week.

The PPS approximation for the actual Pareto selection probability of the non-certainty PAR k, in week v within PSU i becomes:

$$\pi_{k|ijvl} = p_{ijvlk}$$

The CISS PAR sample selection is conducted weekly. Every Tuesday, in each sampled PSU, PARs are listed and a Pareto sample of PARs is selected. Before Friday, if the case vehicle which defines the case's PAR domain is not available for data collection because it is repaired or removed etc., then the case becomes a non-responding case. For each non-responding case, the PAR sample size is increased by 1, and a new Pareto sample is selected with the new sample size to add one replacement case. If the new PAR sample has more than one new case compared with the original PAR sample, then the original PAR sample is the final PAR sample and there will be no replacement. This prevents the responding cases to be replaced. No replacement is allowed from Friday because of the time constraint. Replacement PARs can never be replaced. Replacement cases increase useful sample size. In 2017, about 12% cases were replaced.

The replaced cases are not investigated therefore have no collected data. They are treated as non-responding cases and are not included in the final analysis file. The replaced cases however, are part of the Pareto PAR sample. Therefore, the weights for the remaining non-replaced cases in the final analysis file should be adjusted.

7.5 CISS Crash Investigation

After PAR sample is selected, CISS technicians collects information on the crash, the CISS eligible vehicles involved in the crash, the occupants involved in the crash. Trained crash investigators obtain data from crash sites, studying crash evidence such as skid marks, fluid spills, broken glass, and bent guard rails. They locate the vehicles involved, photograph them, measure the crash damage, and identify interior locations that were struck by the occupants. The researchers also interview crash victims and review their medical records to determine the nature and severity of injuries.

8. Sample Allocation

As the CISS data collection is labor intensive and therefore costly, it was critical for NHTSA to find an approximately optimal sample allocation, i.e., the best combination of PSU, PJ, and PAR sample sizes that minimizes the variance under a fixed budget.

Determine sample allocation is an optimization problem. A non-linear problem was used to find the optimal PSU sample size n, PJ sample size m, and PAR sample size k by minimizing the overall variance given cost constraints and variance constraints which ensure that the new sample design for CISS will be at least as precise as the CDS for the identified key estimates. To find the optimal sample allocation for CISS, two sets of estimates about CISS are needed: the variance component estimates and the cost coefficient estimates for each sampling stage.

8.1 Optimization Model

CISS has a stratified multi-stage unequal selection probability sample design. Taking the first two-stage stratification into account requires specifying the number of strata at each stage. The deep PSU stratification leads to at least 2 PSUs selected from each PSU stratum in order to estimate the variances. At the second stage, PJs are also stratified and at least one PJ selected from each stratum. Therefore, to take stratification into account we need to specify the number of strata and impose stratum sample size constraints. This adds too many constraints to the first two stages and leaves only the PAR sample size to be optimized. In addition, taking the unequal PPS selection probabilities into account makes the variance estimation complicated. Therefore, NHTSA used three-stage simple random sampling without replacement in the optimization model for simplicity. Variance components under simple random sampling are much easier to estimate and allow all three-stage sample sizes to be optimized.

The optimization model consists of the objective function, cost constraint, and variance constrains as follows.

$$Minimize: \sum_{g=1}^{G} V_{CISS}(\bar{\bar{\bar{y}}}_g) = \sum_{g=1}^{G} \left\{ \frac{S_{1,g}^2}{n}(1-\frac{n}{N}) + \frac{S_{2,g}^2}{nm}(1-\frac{m}{M}) + \frac{S_{3,g}^2}{nmk}(1-\frac{k}{K}) \right\}$$

$$Subject\ to:\ C = C_0 + nC_1 + nmC_2 + nmkC_3,$$

$$V_{CISS}(\bar{\bar{\bar{y}}}_g) = \frac{S_{1,g}^2}{n}\left(1-\frac{n}{N}\right) + \frac{S_{2,g}^2}{nm}\left(1-\frac{m}{M}\right) + \frac{S_{3,g}^2}{nmk}\left(1-\frac{k}{K}\right)$$

$$\leq V_{CDS}(\bar{\bar{\bar{y}}}_g), \qquad for\ g = 1, \cdots, G.$$

$$mk \geq l.$$

Here

- g: Subscript of the identified key estimate, $g = 1, \cdots G$. Here $G = 7$.
- $\bar{\bar{\bar{y}}}_g$: Identified key proportion estimate.
- n, m, k: Optimal sample sizes of PSUs, PJs, and cases (PARs) to be determined.
- N: Population size of PSUs
- M: Average population size of PJs.
- K: Average population size of PARs
- $V_{CISS}(\bar{\bar{\bar{y}}}_g)$: Variance of the identified key estimate $\bar{\bar{\bar{y}}}_g$ in the CISS.
- $S_{1,g}^2$, $S_{2,g}^2$, $S_{3,g}^2$: Variance component at PSU-, PJ-, and case-level.
- C, C_0, C_1, C_2, C_3: Total, fixed, PSU-, PJ-, and crash-level cost coefficients.
- $V_{CDS}(\bar{\bar{\bar{y}}}_g)$: Variance of the identified key estimate $\bar{\bar{\bar{y}}}_g$ in the NASS CDS.
- l: caseload – the number of cases to be selected per PSU per year.

Note that the summation of variances in the objective function is over all of the key estimates, which indicates we treated all the key estimates equally.

Seven ($G = 7$) key variables were identified to be considered in the objective function. These key variables were also used in the variance constraints to ensure that the CISS will produce estimates with equal or smaller variance than CDS. The key variables are:

- Crash level variables: Rear-end, Head-on, Angle,
- Vehicle level variable: Roll over, and
- Occupant level variables: Fatality, Incapacitating injury, Non-incapacitating injury.

The variance and variance components ($S_{1,g}^2$, $S_{2,g}^2$, $S_{3,g}^2$) at PSU-, PJ-, and case-level were estimated for proportion estimates of the seven key variables based on 3 year CDS data (2009~2011). The variance estimation is described in detail in Noh and Zhang (2016).

NHTSA conducted a time analysis using the CDS data collection activity. Based on the results of this analysis and other accounting information, Noh (2013) estimated the cost coefficients (C, C_0, C_1, C_2, C_3).

Once employed and trained, a data collection technician must collect a certain number of cases every year. This fact imposes a caseload constraint to the model: $mk \geq l$. Depending on how many technicians are to be hired for each PSU, weekly caseload is determined and the annual caseload is calculated by multiplying 49 weeks (3 weeks were excluded due to training and vacation):

- 0.75 weekly caseload (1 technician): $mk \geq 37$ (0.75 cases×49 weeks=36.75 cases)
- 1.5 weekly caseload (1 technician + 1 half time assistant): $mk \geq 74$ (1.5 cases×49 cases=73.5 cases)
    - weekly caseload (1 technician + 1 full time assistant or 2 technicians): $mk \geq 98$ (2.0 cases×49 weeks=98 cases)
- 3.0 weekly caseload (2 technicians+1 half time assistant): $mk \geq 147$ (3.0 cases×49 weeks=147 cases)
    - weekly caseload (2 technicians+1 full time assistant): $mk \geq 196$ (4.0 cases×49 weeks=196 cases)

A range of total cost was considered. Two thousand starting points (MSNUMSTARTS=2,000) were used in SAS PROC OPTMODEL to find an approximate global optimum solution under various budget levels of budgets. More detailed information on NHTSA's optimization can be found in Noh and Zhang (2016).

8.2 Optimization Results

Table 7 lists the optimization results by budget levels and case load. In this Table, budget levels were rescaled from \$1 to \$4.73. As budget level increases, m and k tend to remain stable or change little, while the PSU sample size n steadily increases. This is consistent with the objective function which indicates factor $1/n$ affects all three terms of the total variance. Increasing PSU sample size is generally the most effective way of reducing the total variance when budget increases.

Table 7: Optimum Solutions and Objective Values by Weekly Case Load Options

Budget	Case Load = 0.75				Case Load = 1.5				Case Load = 2.0				Case Load = 3.0				Case Load = 4.0			
	n	m	k	nmk	n	m	k	nmk	n	m	k	nmk	n	m	k	nmk	n	m	k	nmk
$1.00	35.5	5.3	8.4	**1,590**	27.2	6.5	10.5	**1,865**	23.4	6.5	13.6	**2,065**	16.9	6.9	20.5	**2,382**	14.1	6.6	27.5	**2,549**
$1.05	35.6	5.8	8.3	**1,721**	28.4	6.7	10.5	**1,992**	24.8	6.6	13.5	**2,196**	18.8	6.5	20.5	**2,512**	15.0	6.6	27.5	**2,708**
$1.09	33.8	7.1	7.7	**1,861**	30.7	6.5	10.5	**2,097**	26.4	6.5	13.5	**2,323**	19.4	6.7	20.5	**2,675**	16.0	6.5	27.7	**2,869**
$1.14	35.5	7.3	7.6	**1,961**	32.0	6.6	10.5	**2,224**	27.9	6.5	13.5	**2,451**	20.9	6.5	20.5	**2,812**	16.9	6.5	27.5	**3,027**
$1.18	48.3	5.1	7.6	**1,845**	34.1	6.5	10.5	**2,336**	27.6	8.5	10.5	**2,473**	21.0	7.5	18.7	**2,942**	16.6	9.5	19.5	**3,072**
$1.23	53.3	4.7	7.5	**1,885**	34.0	7.5	9.5	**2,423**	28.8	8.6	10.5	**2,600**	23.1	6.5	20.6	**3,112**	18.4	6.6	27.5	**3,352**
$1.27	55.2	4.7	7.7	**2,009**	36.8	6.7	10.5	**2,583**	32.3	6.5	13.5	**2,842**	24.3	6.5	20.5	**3,257**	19.4	6.6	27.5	**3,511**
$1.32	57.7	4.8	7.6	**2,086**	39.0	5.6	13.0	**2,831**	31.8	8.5	10.5	**2,844**	25.5	6.5	20.5	**3,403**	20.4	6.5	27.5	**3,669**
$1.36	61.2	4.7	7.5	**2,153**	40.1	6.7	10.5	**2,821**	33.1	8.5	10.5	**2,973**	24.5	9.5	14.5	**3,392**	20.7	6.8	27.5	**3,845**
$1.41	64.7	4.6	7.5	**2,223**	42.1	6.6	10.5	**2,931**	35.5	6.8	13.5	**3,259**	25.6	8.6	16.5	**3,620**	21.4	6.8	27.5	**4,008**
$1.45	66.6	4.6	7.6	**2,337**	43.8	6.6	10.5	**3,049**	34.3	9.5	9.8	**3,191**	27.1	7.6	18.8	**3,839**	23.1	6.5	27.5	**4,147**
$1.55	69.8	4.8	7.6	**2,551**	47.5	6.6	10.5	**3,277**	40.5	6.6	13.5	**3,634**	28.3	9.5	14.8	**3,983**	23.8	8.5	21.5	**4,354**
$1.64	77.7	4.5	7.7	**2,689**	49.1	6.5	11.4	**3,622**	39.9	9.6	9.5	**3,649**	30.1	9.6	14.8	**4,273**	23.7	13.5	13.5	**4,338**
$1.73	83.2	4.6	7.5	**2,840**	51.4	7.6	9.6	**3,717**	46.6	6.5	13.7	**4,161**	35.2	6.6	20.5	**4,751**	25.9	10.5	17.9	**4,866**
$1.82	84.8	4.8	7.6	**3,099**	62.9	4.6	14.5	**4,168**	49.7	6.6	13.5	**4,405**	33.5	8.9	16.6	**4,967**	26.3	13.8	13.6	**4,936**
$2.00	98.3	4.6	7.6	**3,400**	62.0	6.7	10.9	**4,549**	54.8	6.7	13.5	**4,940**	36.2	11.5	12.7	**5,289**	28.1	16.5	11.6	**5,353**
$2.18	108.1	4.6	7.6	**3,768**	67.9	7.5	9.6	**4,873**	60.6	6.7	13.5	**5,463**	42.2	9.8	14.5	**5,965**	34.9	8.7	21.5	**6,560**
$2.36	114.8	4.9	7.5	**4,185**	76.5	5.9	12.5	**5,661**	62.4	8.5	10.9	**5,774**	45.2	10.0	14.6	**6,567**	37.3	10.6	17.6	**6,985**
$2.55	136.4	4.8	7.5	**4,883**	86.6	5.6	12.7	**6,098**	71.9	6.7	13.6	**6,518**	50.3	8.8	16.5	**7,279**	38.1	14.2	13.5	**7,275**
$2.73	136.4	4.8	7.5	**4,883**	93.1	5.5	12.9	**6,619**	70.8	9.1	10.5	**6,778**	53.1	10.0	14.5	**7,700**	42.9	13.5	13.5	**7,818**
$2.91	149.5	4.6	7.5	**5,193**	94.1	7.5	9.5	**6,717**	81.2	6.9	13.5	**7,606**	59.3	9.5	14.6	**8,227**	44.7	12.5	15.5	**8,666**
$3.09	161.3	4.6	7.5	**5,517**	106.5	5.5	13.0	**7,621**	85.0	8.6	10.5	**7,711**	63.5	9.6	14.5	**8,786**	47.9	9.5	21.5	**9,780**
$3.27	165.9	4.8	7.5	**5,991**	114.8	5.5	12.7	**8,051**	95.5	6.6	13.7	**8,619**	72.6	6.6	20.6	**9,822**	51.8	13.7	13.5	**9,576**
$3.45	170.4	5.5	6.6	**6,177**	116.6	6.5	11.0	**8,368**	102.6	6.6	13.5	**9,079**	69.3	9.8	14.8	**10,017**	60.8	6.7	27.5	**11,235**
$3.64	185.4	4.7	7.7	**6,782**	124.9	6.6	10.5	**8,700**	108.4	6.5	13.6	**9,610**	69.1	14.5	9.6	**9,664**	64.9	6.7	27.5	**11,863**
$3.82	199.4	4.6	7.8	**7,098**	119.1	9.6	7.5	**8,559**	113.0	6.6	13.6	**10,158**	76.9	9.8	14.9	**11,185**	59.2	14.0	13.7	**11,403**
$4.00	204.6	4.9	7.5	**7,468**	140.1	6.5	10.5	**9,613**	109.7	8.9	10.5	**10,253**	80.8	9.6	15.3	**11,817**	63.4	13.7	13.8	**11,968**
$4.18	222.6	4.6	7.5	**7,677**	138.5	7.1	10.5	**10,255**	116.1	8.7	10.7	**10,788**	87.1	9.6	14.7	**12,250**	70.1	10.5	17.8	**13,088**
$4.36	235.0	4.5	7.6	**8,026**	153.9	6.5	10.5	**10,546**	130.5	6.5	13.8	**11,759**	91.3	9.5	14.7	**12,816**	77.4	7.5	24.5	**14,267**
$4.55	237.0	4.8	7.5	**8,517**	160.6	6.5	10.5	**11,023**	130.1	8.5	10.5	**11,664**	97.6	7.7	18.5	**13,900**	75.7	10.8	17.5	**14,290**
$4.73	254.8	4.5	7.7	**8,790**	166.1	6.5	10.7	**11,563**	133.4	8.5	10.8	**12,296**	108.8	6.5	20.5	**14,556**	80.5	10.5	17.5	**14,860**

Table 7: Optimum Solutions and Objective Values by Case Load Options (Continued)

Square Roots of Objective Function Values: Standard Error for Percent Estimate					
Budget	**Case Load = 0.75**	**Case Load = 1.5**	**Case Load = 2.0**	**Case Load = 3.0**	**Case Load = 4.0**
$1.00	3.78%	3.95%	4.12%	4.60%	4.99%
$1.05	3.68%	3.84%	4.00%	4.42%	4.83%
$1.09	3.62%	3.71%	3.88%	4.32%	4.69%
$1.14	3.53%	3.62%	3.77%	4.18%	4.56%
$1.18	3.35%	3.52%	3.68%	4.09%	4.40%
$1.23	3.26%	3.45%	3.60%	3.97%	4.35%
$1.27	3.19%	3.37%	3.50%	3.88%	4.24%
$1.32	3.12%	3.32%	3.43%	3.79%	4.15%
$1.36	3.05%	3.22%	3.36%	3.69%	4.09%
$1.41	2.98%	3.15%	3.31%	3.64%	4.01%
$1.45	2.93%	3.09%	3.25%	3.59%	3.89%
$1.55	2.82%	2.97%	3.11%	3.43%	3.71%
$1.64	2.72%	2.90%	3.01%	3.32%	3.56%
$1.73	2.63%	2.79%	2.90%	3.21%	3.47%
$1.82	2.55%	2.71%	2.81%	3.15%	3.37%
$2.00	2.41%	2.56%	2.66%	2.97%	3.22%
$2.18	2.29%	2.43%	2.53%	2.79%	3.04%
$2.36	2.19%	2.34%	2.42%	2.69%	2.88%
$2.55	2.09%	2.22%	2.32%	2.57%	2.79%
$2.73	2.01%	2.14%	2.25%	2.47%	2.64%
$2.91	1.94%	2.05%	2.16%	2.36%	2.59%
$3.09	1.87%	2.00%	2.07%	2.28%	2.55%
$3.27	1.81%	1.93%	2.01%	2.22%	2.39%
$3.45	1.76%	1.87%	1.94%	2.16%	2.37%
$3.64	1.71%	1.80%	1.88%	2.10%	2.30%
$3.82	1.66%	1.78%	1.84%	2.05%	2.22%
$4.00	1.62%	1.71%	1.80%	2.00%	2.15%
$4.18	1.58%	1.68%	1.75%	1.93%	2.09%
$4.36	1.54%	1.62%	1.71%	1.88%	2.06%
$4.55	1.50%	1.59%	1.66%	1.86%	2.00%
$4.73	1.47%	1.56%	1.63%	1.80%	1.95%

9. Weighting

The CISS sample is the result of a probability sampling with sampling features such as stratification, clustering, and unequal selection probabilities. Because of this, the CISS sample is not a simple random sample. To produce unbiased estimates, the CISS sample should be properly weighted. Unweighted estimates may be severely biased. The weights to be used for CISS data analysis are the inverse of the PAR selection probabilities adjusted for non-response, calibration, and truncation, etc. This chapter describes how CISS weights are calculated.

The 2017 CISS weights are created in the following steps:

- Calculate design weights at all three stages;
- Adjust for non-responding PJs and PARs;
- Perform post-stratification adjustment using the crash counts collected from sampled and non-sampled PJs;
- Perform calibration to the PSU weights using Census resident population information;
- Truncate large case weights; and
- Create Jackknife replicate weights for variance estimation.

9.1 Design Weights

The design weight is the inverse of the selection probability defined by the sample design. It is the product of PSU weight, PJ weight, sub-listing factor, and PAR weight:

$$w_{ijvlk} = w_i * w_{j|i} * w_{l|ij} * w_{k|ijvl}$$

Here

- $w_i = \pi_i^{-1}$ is the inverse of PSU i selection probability. The CISS PSU sample size may change over the years. Therefore, the PSU weights may also change accordingly. In 2017, the CISS PSU sample size was 24. Weighting was performed only for the cases from 24 PSUs. We denote the PSU weights for the 24 PSU sample as $w_i^{(1)}$.
- $w_{j|i} = \pi_{j|i}^{-1}$ is the inverse of PJ j selection probability.
- $w_{l|ij}$ is the sub-listing factor for a PAR. If there is no sub-listing, $w_{l|ij} = 1$.
- $w_{k|ijvl} = \pi_{k|ijvl}^{-1}$ is the inverse of PAR selection probability for a PAR k at week v in PJ j of PSUi.
- PSU weights, PJ weights, and sub-listing factor may be changed in the middle of year.

The calculation of selection probabilities at all three stages can be found in previous chapters.

9.2 Non-Response Adjustment

In 2017, all 24 sampled CISS PSUs have cooperated with data collection. But the CISS PJ and PAR samples suffered from non-response. Actually, among the 182 sampled PJs, 168 PJs responded. Among the 2,331 selected PARs, 288 PARs were treated as non-responding PARs because their key vehicles were not available for inspection. Estimation made from a sample with non-responding units without treatment may be severely biased. This section describes how adjustments are made at each sampling stage to mitigate the potential non-response bias.

9.2.1 Non-Responding PJ Adjustment

As mentioned earlier, in the CISS, PARs are stratified and sampled by weeks. PJ sample may change in the middle of the year. In addition, it is possible that PJs do not cooperate for the whole year or for some of the weeks. Therefore, PJ non-response adjustment should be conducted by week and PSU.

PJ non-response adjustment factor is calculated for week v and PSU i to adjust PJ non-response:

$$adj_{iv} = \frac{\sum_{j \in s_{iv}} w_{j|i} MOS_{ij}}{\sum_{j \in r_{iv}} w_{j|i} MOS_{ij}} .$$

Here s_{iv} is the set of all sampled PJs and r_{iv} is the set of all responding PJs in PSU i for week v. MOS_{ij} is the finer PJ MOS used in PJ sample selection.

PJ weights are adjusted weekly by multiplying the above adjustment factor adj_{iv}:

$$w_{j|i}^{(1)} = \begin{cases} w_{j|i} * adj_{iv} & \text{for responding PJs} \\ 0, & \text{for nonresponding PJs} \end{cases}$$

9.2.2 Non-Responding PAR Adjustment

The non-responding (replaced) PARs are part of the Pareto PAR sample but they are not kept in the final analysis file because there is no data collected from them. Therefore, the weights for the remaining responding cases should be adjusted.

To adjust the weights of the responding cases, we formed PAR non-response adjustment cells based on the results of NHTSA's CISS non-response study (Smith 2018) and used weighted PAR non-response rate by those cells. Specifically, 71 cells were formed for PAR non-response adjustment as the following.

- 3 PAR domain groups for each PSU (PSU 1~PSU 23):
  - Domain 1, 3, 5, 6, and 8
  - Domain 2, 7, and 9
  - Domain 4 and 10
- 2 PAR domain groups for PSU 24 because the sample size in PSU 24 is too small:
  - Domain 1, 2, 3, 5, 6, 8, and 9
  - Domain 4, 7, and 10

Let q be the subscript for PAR domain groups. Then, calculate the non-responding PAR adjustment factor as the inverse of weighted response rate for each PAR domain group q of PSU i:

$$adj_{iq} = \frac{\sum_{j \in r_i} \sum_{v \epsilon r v_i} \sum_{k \in SPAR_{ijvq}} w_{j|i}^{(1)} w_{l|ij} w_{k|ijvl}}{\sum_{j \in r_i} \sum_{v \epsilon r v_i} \sum_{k \in RPAR_{ijvq}} w_{j|i}^{(1)} w_{l|ij} w_{k|ijvl}}$$

Here r_i is the set of responding PJs, rv_i is the set of weeks in which PAR sample was selected in PSU i. $SPAR_{ijvq}$ is the set of sampled PARs and $RPAR_{ijvq}$ is the set of responding PARs in PAR domain group q of PSU i, PJ j and week v.

Apply adj_{iq} to each responded PAR:

$$w_{k|ijvl}^{(1)} = \begin{cases} w_{k|ijvl} * adj_{iq} & \text{for } responding \text{ PARs} \\ 0, & \text{for nonresponding PARs} \end{cases}$$

With the PSU weight $w_i^{(1)}$ for the 24 PSU sample, non-response adjusted PJ weight, and PAR weight, the within-PSU weight and the case weight are updated as the following:

$$w_{jvlk|i}^{(1)} = w_{j|i}^{(1)} * w_{l|ij} * w_{k|ijvl}^{(1)}$$

$$w_{ijvlk}^{(1)} = w_i^{(1)} * w_{jvlk|i}^{(1)}$$

9.3 Post-Stratification Adjustment for Coverage Error

In the CISS, post-stratification adjustment is used to correct potential coverage error or bias for the following reasons:

- In some weeks, a technical assistant lists PARs but no PAR sample is selected because the technician is not available for crash investigation due to training, annual/sick leave, termination of employment, etc. Those listed PARs are archived and not be used for PAR sample selection. This results in an under-coverage error.
- PPS sampling weights for PAR samples selected by Pareto sampling may have bias because sample size is too small and PAR MOS has big variation.
- Other possible coverage error or non-response bias.

Every year the CISS technician collects crash counts by PAR domain in the non-sampled PJs in each sampled PSU. Combined with the listed (or sub-listed) PARs in the sampled PJs, PAR domain total crash counts are used to correct the potential under-coverage error and bias.

If PSU level PAR domain crash counts are available, i.e., PARs are listed for all sampled PJs and crash counts are collected for all non-sampled PJs (condition A), then a post-stratification factor is computed by the 10 PAR domains and urbanicity (urban or rural) excluding PSUs with at least one non-responding or non-cooperating PJ (non-sampled PJ that does not provide PARs). First, urbanicity level PAR domain total crash counts is estimated:

$$T_{gs} = \sum_{i \in rs_g} \sum_{j \in s_i} \sum_{v \in v_i} w_i^{(1)} w_{l|ij} L_{ijvls} + \sum_{i \in rs_g} \sum_{j \in ns_i} w_i^{(1)} C_{ijs}$$

The post-stratification factor is computed as

$$post_{gs} = T_{gs} / \sum_{i \in rs_g} \sum_{j \in s_i} \sum_{v \in rv_i} \sum_{k \in r_{ijvs}} w_{ijvlk}^{(1)}$$

Here rs_g is the set of sampled PSUs where PSU level PAR domain total crash counts are available in urbanicity g. s_i is the set of sampled PJs and ns_i is the set of non-sampled PJs in PSU i. v_i is set of weeks in which PARs were listed and rv_i is the set of weeks in which PAR sample was selected in PSU i. L_{ijvls} is the number of listed PARs of domain s on week v in PJ j of PSU i, and C_{ijs} is the non-sampled crash count of domain s in PJ j of PSU i. r_{ijvs} is the set of sampled responding PARs in PAR domain s in week v, PJ j, PSU i.

If PSU level PAR domain crash counts are not available but the PSU level total crash counts can be estimated (condition B), the post-stratification factor is computed at PSU level. PSU level in-scope total crash count is estimated as

$$T_i = \sum_{j \in r_i} \sum_{v \in v_i} w_{l|ij} L_{ijlv} + \sum_{j \in cns_i} C_{ij} + I_i \sum_{j \in nr_i \cup ncns_i} A_{ij}$$

The post-stratification factor is computed:

$$post_i = T_i / \sum_{j \in r_i} \sum_{v \in rv_i} \sum_{k \in r_{ijv}} w_{jvlk|i}^{(1)}$$

Here r_i is the set of responding PJs, nr_i is the set of non-responding PJs, cns_i is the set of cooperating non-sampled PJs, and $ncns_i$ is the set of non-cooperating non-sampled PJs in PSU i. L_{ijlv} is the number of listed PARs of all CISS domains on week v in PJ j of PSU i. C_{ij} is non-sampled crash count of all CISS domains in PJ j of PSU i. A_{ij} is all crash counts (including out-of-scope crashes) of non-responding or non-cooperating PJ j of PSU i. r_{ijv} is the set of sampled responding PARs over all CISS PAR domains on week v in PJ j of PSU i. I_i is the PSU-level in-scope crash rate estimated as

$$I_i = \frac{\sum_{j \in r_i} \sum_{v \in v_i} w_{l|ij} L_{ijlv} + \sum_{j \in cns_i} C_{ij}}{\sum_{j \in r_i} \sum_{v \in v_i} w_{l|ij} AL_{ijlv} + \sum_{j \in cns_i} AC_{ij}}$$

Here AL_{ijlv} is the number of all PARs (including out-of-scope crashes) on week v in PJ j of PSU i and AC_{ij} is non-sample crash count of all crashes (including out-of-scope crashes) in PJ j of PSU i.

If PSU level PAR domain crash counts are not available and the PSU level total crash counts cannot be estimated (condition C), post-stratification is not performed, and post-stratification factor is set to one (i.e., $post_i = 1$).

Finally, post-stratified case weight is compute:

$$w_{ijvlk}^{(2)} = \begin{cases} post_{gs} * w_{ijvlk}^{(1)} & \text{for cases in PSU with condition A} \\ post_i * w_{ijvlk}^{(1)} & \text{for cases in PSU with condition B or C} \end{cases}$$

9.4 PSU Weight Calibration

CISS PSU sample will be used for data collection for many years. Over the years, the resident population may shift between urban and rural area. The crash counts are highly correlated with the resident population counts. To capture population changes, PSU weight calibration is performed by benchmarking the PSU weighted population counts to the known marginal population counts in the urban and rural area. Specifically, the following adjustment factor is calculated for urban PSUs and rural PSUs separately:

$$adj_g = P_g / \sum_{i \in s_g} w_i^{(1)} P_i$$

Here s_g is the set of sampled PSUs in urban or rural area g. P_g and P_i are the Census resident population counts for urban or rural area g and for PSU i, respectively.

With this adjustment, the calibrated case weight is computed as:

$$w_{ijvlk}^{(3)} = adj_g * w_{ijvlk}^{(2)}$$

9.5 Weight Truncation

Because of the small sample size and the unequal selection probability sampling applied in CISS PAR sample selection, large weights are inevitable. For example, the estimated domain 10 population size is about 443,000 (see Table 1) while only 6 percent of the sample is allocated to domain 10. If the total sample size is 2,000, then only about 120 cases selected from domain 10. This results in an average weight of 3,692 for domain 10 with some extremely large weights.

Large weight variation inflates the variances of the weighted estimates and results in abnormal small domain estimates. We used the following procedure to truncate the extremely large weights to reduce the weight variation and calibrate the resulting weights to preserve the total weights:

First, form truncating cells using the 10 PAR domains defined in Table 1. For each PAR k in a PAR domain s, truncate case weight as:

$$w_{ijvlsk}^{(3t)} = \begin{cases} w_{ijvlk}^{(3)} & \text{if } w_{ijvlk}^{(3)} \le 0.03 * \sum_{k \in d_s} w_{ijvlk}^{(3)} \\ 0.03 * \sum_{k \in d_s} w_{ijvlk}^{(3)} & \text{if } w_{ijvlk}^{(3)} > 0.03 * \sum_{k \in d_s} w_{ijvlk}^{(3)} \end{cases}$$

Here d_s is the set of PARs in PAR domain s. The truncation adjustment factor is computed to correct the truncated weights as:

$$adj_s = \frac{\sum_{k \in d_s} w_{ijvlk}^{(3)} - \sum_{k \in (d_s \cap s_{tc})} w_{ijvlk}^{(3t)}}{\sum_{k \in d_s} w_{ijvlk}^{(3)} - \sum_{k \in (d_s \cap s_{tc})} w_{ijvlk}^{(3)}}$$

Here s_{tc} is the set of truncated PARs. Then, the final truncated case weight is calculated by applying the adjustment factor:

$$w_{ijvlk}^{(4)} = \begin{cases} w_{ijvlk}^{(3t)} & \text{if } k \in s_{tc} \\ adj_s * w_{ijvlk}^{(3t)} & \text{otherwise} \end{cases}$$

If any adjusted case weight is greater than the upper bound, $0.03 \times \sum_{k \in d_s} w_{ijvlk}^{(3)}$, truncation process is repeated until all adjusted weight is less than or equal to the upper bound. $w_{ijvlk}^{(4)}$ is the final CISS case weight.

Under this adjustment, the summation of the truncated weights still equals to the summation of the calibrated weights for each PAR domain.

9.6 Replicate Weights for Variance Estimation

Weight adjustments not only mitigate non-response bias and frame coverage errors, but may also affect the variances of the underlying estimators (increasing or decreasing the variance). To capture the effects of weight adjustments to the variance estimates, we created adjusted Jackknife (JK) replicate weights in the following procedure.

1) Calculate PSU replicate weight as:
   - Set the PSU weights to zero for all cases in one PSU from PSU stratum h.
   - Multiply a factor $n_h/(n_h - 1)$ to the PSU weights for all cases in the remaining PSUs of stratum h, here n_h is the PSU sample size of stratum h.
   - PSU weights are not changed for all cases in PSUs in other strata.
2) Calculate PJ and PAR non-response adjusted case weights using the PSU weights from step 1) and non-response adjusted PJ and PAR weights from section 9.2.
3) Perform post-stratification adjustment as described in section 9.3.
4) Perform calibration as described in section 9.4.
5) Perform weight truncation as described in section 9.5. This is one set of the adjusted Jackknife replicate weights.
6) Repeat step 1) through step 5) for all other PSUs.

This procedure results in n (the PSU sample size) sets of adjusted JK replicate weights, each corresponding to one sampled PSU. For 2017 CISS, total 24 sets of adjusted JK replicate weights were created.

Using these adjusted JK replicate weights in estimation helps to take the effects of the following weight adjustments into account:

- The gain in efficiency due to the post-stratification.
- The gain in efficiency due to the calibration.
- The gain in efficiency due to the weight truncation.

PJ non-response adjustment and PAR non-response adjustment are performed within PSU so these adjustments do not affect the adjusted JK replicate weights.

For more details about CISS estimation, computer programs and examples, see Zhang et al (2019, September): Crash Investigation Sampling System: Design Overview, Analytic Guidance, and FAQs (Report No, DOT HS 812 801).

References

Fleming, C. (2010, May). *Sampling and estimation methodologies of CDS* (Report No. DOT HS 811 327). Washington, DC: National Highway Traffic Safety Administration. Available at https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811327

House Report #111-564. Departments of Transportation, Housing, and Urban Development, and Related Agencies Appropriations Bill of 2011.

Noh, E. Y. (2013). *CISS cost component estimation* (Unpublished technical report). Washington, DC: National Highway Traffic Safety Administration.

Noh, E. Y., & Zhang, F. (2016). *Optimum sample allocation of the CISS and CRSS*. (Unpublished technical report). Washington, DC: National Highway Traffic Safety Administration.

Noh, E. Y., & Zhang, F. (2017). *Simulation study of probability proportional to size samplings based on permanent random number in the Crash Investigation Sampling System* (Unpublished technical report). Washington, DC: National Highway Traffic Safety Administration.

Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, *Vol. 62*, pp 159-191.

Senate Report # 111-230. Transportation, Housing and Urban Development, and Related Agencies Appropriations Bill of 2011.

Shelton, T. S. (1991). *National Accident Sampling System, General Estimates System, Technical Note, 1988 to 1990* (Report No. DOT HS 807 796). Washington, DC: National Highway Traffic Safety Administration.

Smith, P. W. (2018). *Crash Investigation Sampling System: A multilevel modeling approach to find covariates that influence the likelihood of a police crash report response* (Unpublished report). Washington, DC: National Highway Traffic Safety Administration.

Westat, Inc. (2012a). *Stakeholder outreach and summary* (Unpublished report to NHTSA). Rockville, MD: Author.

Westat, Inc. (2012b). *Data elements recommendations report* (Unpublished report to NHTSA). Rockville, MD: Author.

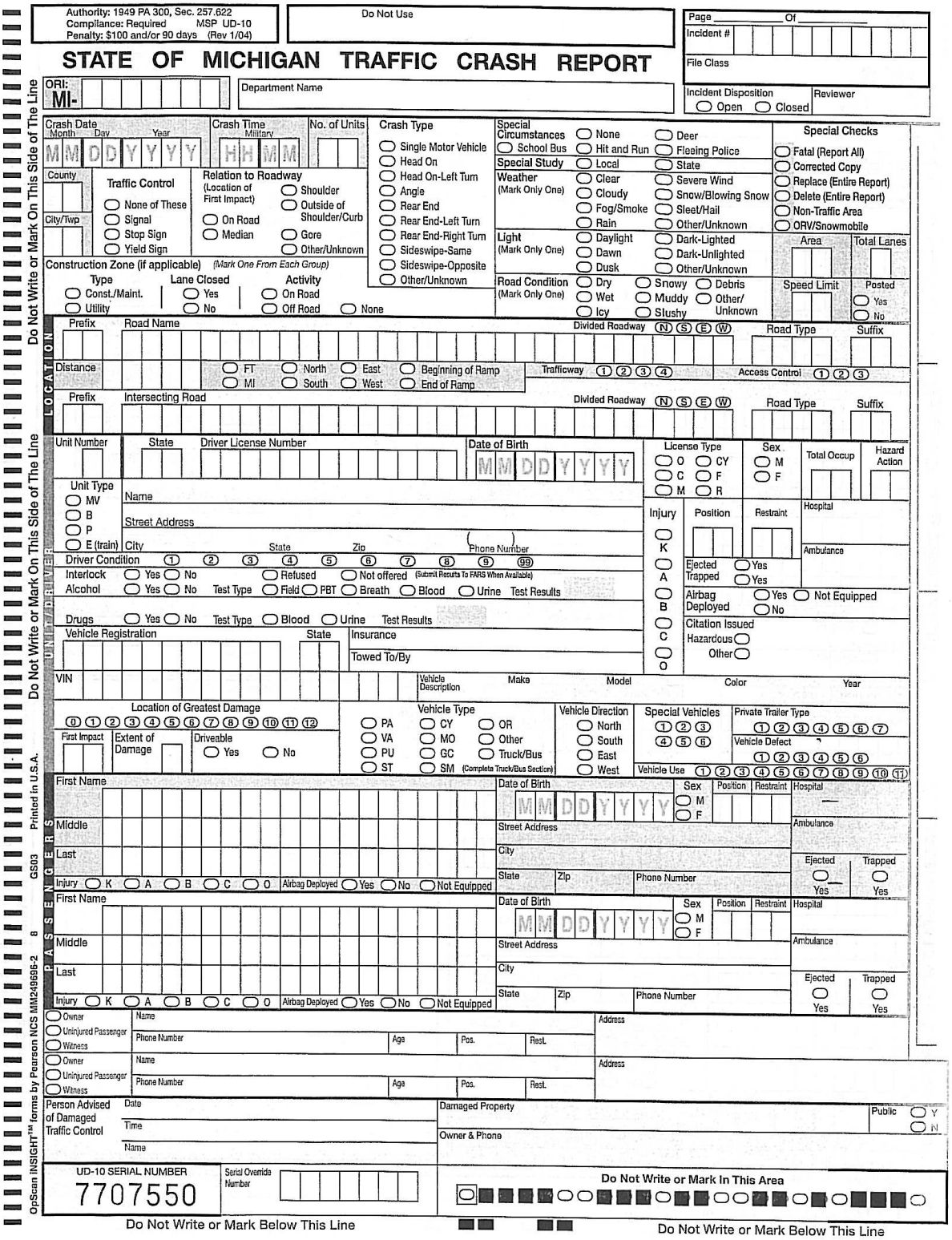
Westat, Inc. (2014). *Survey modernization analysis: Designs for a modernized NASS* (Unpublished report to NHTSA). Rockville, MD: Author.

Zhang, F. and Chen, C.-L. (2013). *NASS-CDS: Sample design and weights* (Report No. DOT HS 811 807*)*. Washington, DC: National Highway Traffic Safety Administration. Available at https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811807

Zhang, F., Noh, E. Y.; Subramanian, R., & Chen, C.-L. (2019, April). *Crash Report Sampling System: Sample design and weighting* (Report No, DOT HS 812 706). Washington, DC: National Highway Traffic Safety Administration.

Zhang, F., Subramanian, R., Chen, C.-L., & Young Noh, E. Y. NHTSA (2019, September). Crash Investigation Sampling System: Design overview, analytic guidance, and FAQs (Report No. DOT HS 812 801). Washington, DC: National Highway Traffic Safety Administration.

Appendix A: An Example of PAR



Authority: 1949 PA 300, Sec. 257.622
Compliance: Required MSP UD-10
Penalty: $100 and/or 90 days (Rev 1/04)

Do Not Use

Page ____ Of ____
Incident #
File Class
Incident Disposition ○ Open ○ Closed
Reviewer

STATE OF MICHIGAN TRAFFIC CRASH REPORT

ORI: MI-
Department Name

Crash Date Month Day Year (MM DD YYYY)
Crash Time Military (HH MM)
No. of Units

Crash Type
○ Single Motor Vehicle
○ Head On
○ Head On-Left Turn
○ Angle
○ Rear End
○ Rear End-Left Turn
○ Rear End-Right Turn
○ Sideswipe-Same
○ Sideswipe-Opposite
○ Other/Unknown

Special Circumstances ○ None ○ Hit and Run ○ School Bus ○ Deer ○ Fleeing Police
Special Study ○ Local ○ State
Weather (Mark Only One) ○ Clear ○ Cloudy ○ Fog/Smoke ○ Rain ○ Severe Wind ○ Snow/Blowing Snow ○ Sleet/Hail ○ Other/Unknown
Light (Mark Only One) ○ Daylight ○ Dawn ○ Dusk ○ Dark-Lighted ○ Dark-Unlighted ○ Other/Unknown
Road Condition (Mark Only One) ○ Dry ○ Wet ○ Icy ○ Snowy ○ Muddy ○ Slushy ○ Debris ○ Other/Unknown

Special Checks
○ Fatal (Report All)
○ Corrected Copy
○ Replace (Entire Report)
○ Delete (Entire Report)
○ Non-Traffic Area
○ ORV/Snowmobile

County
City/Twp

Traffic Control
○ None of These ○ Signal ○ Stop Sign ○ Yield Sign

Relation to Roadway (Location of First Impact)
○ On Road ○ Shoulder ○ Median ○ Outside of Shoulder/Curb ○ Gore ○ Other/Unknown

Construction Zone (if applicable) (Mark One From Each Group)
Type ○ Const./Maint. ○ Utility
Lane Closed ○ Yes ○ No
Activity ○ On Road ○ Off Road ○ None

Area
Total Lanes
Speed Limit
Posted ○ Yes ○ No

LOCATION

Prefix Road Name Divided Roadway (N S E W) Road Type Suffix
Distance ○ FT ○ MI ○ North ○ South ○ East ○ West ○ Beginning of Ramp ○ End of Ramp
Trafficway (1 2 3 4)
Access Control (1 2 3)
Prefix Intersecting Road Divided Roadway (N S E W) Road Type Suffix

UNIT / DRIVER

Unit Number
State
Driver License Number
Date of Birth (MM DD YYYY)
License Type ○ O ○ CY ○ C ○ F ○ M ○ R
Sex ○ M ○ F
Total Occup
Hazard Action

Unit Type ○ MV ○ B ○ P ○ E (train)
Name
Street Address
City State Zip (Phone Number)
Driver Condition (1 2 3 4 5 6 7 8 9 99)
Interlock ○ Yes ○ No ○ Refused ○ Not offered (Submit Results To FARS When Available)
Alcohol ○ Yes ○ No Test Type ○ Field ○ PBT ○ Breath ○ Blood ○ Urine Test Results
Drugs ○ Yes ○ No Test Type ○ Blood ○ Urine Test Results

Injury ○ K ○ A ○ B ○ C ○ O
Position
Restraint
Hospital
Ambulance
Ejected ○ Yes
Trapped ○ Yes
Airbag Deployed ○ Yes ○ No ○ Not Equipped
Citation Issued
Hazardous ○
Other ○

Vehicle Registration State
Insurance
Towed To/By

VIN
Vehicle Description Make Model Color Year

Location of Greatest Damage (0 1 2 3 4 5 6 7 8 9 10 11 12)
First Impact
Extent of Damage
Driveable ○ Yes ○ No

Vehicle Type ○ PA ○ CY ○ OR ○ VA ○ MO ○ Other ○ PU ○ GC ○ Truck/Bus ○ ST ○ SM (Complete Truck/Bus Section)
Vehicle Direction ○ North ○ South ○ East ○ West
Special Vehicles (1 2 3 4 5 6)
Private Trailer Type (1 2 3 4 5 6 7)
Vehicle Defect (1 2 3 4 5 6)
Vehicle Use (1 2 3 4 5 6 7 8 9 10 11)

PASSENGERS

First Name
Middle
Last
Date of Birth (MM DD YYYY)
Sex ○ M ○ F
Position
Restraint
Hospital
Ambulance
Street Address
City
State Zip Phone Number
Ejected ○ Yes
Trapped ○ Yes
Injury ○ K ○ A ○ B ○ C ○ O
Airbag Deployed ○ Yes ○ No ○ Not Equipped

First Name
Middle
Last
Date of Birth (MM DD YYYY)
Sex ○ M ○ F
Position
Restraint
Hospital
Ambulance
Street Address
City
State Zip Phone Number
Ejected ○ Yes
Trapped ○ Yes
Injury ○ K ○ A ○ B ○ C ○ O
Airbag Deployed ○ Yes ○ No ○ Not Equipped

○ Owner ○ Uninjured Passenger ○ Witness
Name
Phone Number Age Pos. Rest.
Address

○ Owner ○ Uninjured Passenger ○ Witness
Name
Phone Number Age Pos. Rest.
Address

Person Advised of Damaged Traffic Control
Date
Time
Name
Damaged Property
Owner & Phone
Public ○ Y ○ N

UD-10 SERIAL NUMBER 7707550
Serial Override Number

Do Not Write or Mark In This Area

Do Not Write or Mark Below This Line

Do Not Write or Mark Below This Line

Do Not Write or Mark On This Side of The Line

Do Not Write or Mark On This Side of The Line

OpScan INSIGHT™ forms by Pearson NCS MM249696-2 8 GS03 Printed in U.S.A.

Appendix B: Nested Scenario Strata

Scenario	Sample Order	Scenario-1	Scenario-2	Scenario-3	Scenario-4	Scenario-5
#Strata		**24**	**20**	**16**	**12**	**8**
Strata	20	1-01	1-01	1-01	1-01	Northeast MSA
	33					
	53					
	76					
	6	1-02	1-02			
	36					
	55					
	78					
	16	1-03	1-03	1-02	1-02	
	18					
	60					
	83					
	13	2-01	2-01	2-01	2-01	Northeast non-MSA
	26					
	71					
	94					
	5	2-02	2-02	2-02		
	31					
	73					
	96					
	21	3-01	3-01	3-01	3-01	Midwest MSA
	47					
	52					
	2	3-02				
	45					
	61					
	84					
	12	3-03	3-02	3-02	3-02	
	44					
	62					
	85					
	19	3-04				
	41					
	67					
	90					
	7	4-01	4-01	4-01	4-01	Midwest non-MSA
	28					
	68					
	91					
	11	4-02	4-02	4-02		
	27					
	69					
	92					

Nested Scenario Strata (continued)

Scenario	Sample Order	Scenario-1	Scenario-2	Scenario-3	Scenario-4	Scenario-5
#Strata		24	20	16	12	8
	22	5-01	5-01	5-01	5-01	South MSA
	42					
	51					
	75					
	35	5-02				
	48					
	50					
	74					
	15	5-03	5-02			
	37					
	58					
	81					
	24	5-04	5-03	5-02	5-02	
	34					
	54					
	77					
	1	5-05	5-04			
	39					
	65					
	88					
	38	5-06	5-05			
	40					
	66					
	89					
	10	6-01	6-01	6-01	6-01	South non-MSA
	30					
	59					
	82					
	3	6-02	6-02	6-02		
	32					
	64					
	87					
	46	7-00	7-01	7-01	7-01	West MSA
	14	7-01				
	49					
	63					
	86					
	23	7-02				
	43					
	57					
	80					
	4	7-03	7-02	7-02	7-02	
	17					
	56					
	79					

Nest Scenario Strata (continued)

Scenario	Sample Order	Scenario-1	Scenario-2	Scenario-3	Scenario-4	Scenario-5
#Strata		24	20	16	12	8
	8	8-01	8-01	8-01	8-01	West non-MSA
	25					
	72					
	95					
	9	8-02	8-02	8-02		
	29					
	70					
	93					

Appendix C: Excluded AK and HI Counties

Excluded Counties AK	
FIPS Code	**County Name**
02013	Aleutians East Borough
02016	Aleutians West Census Area
02050	Bethel Census Area
02060	Bristol Bay Borough
02070	Dillingham Census Area
02100	Haines Borough
02105	Hoonah-Angoon Census Area
02110	Juneau City and Borough
02130	Ketchikan Gateway Borough
02150	Kodiak Island Borough
02164	Lake and Peninsula Borough
02180	Nome Census Area
02185	North Slope Borough
02188	Northwest Arctic Borough
02195	Petersburg Census Area
02198	Prince of Wales-Hyder Census Area
02220	Sitka City and Borough
02230	Skagway Municipality
02261	Valdez-Cordova Census Area
02270	Wade Hampton Census Area
02275	Wrangell City and Borough
02282	Yakutat City and Borough
02290	Yukon-Koyukuk Census Area

Included Counties AK		
FIPS Code	**County Name**	**PSU MOS**
02020	Anchorage Municipality	10033.67
02170	Matanuska-Susitna Borough	
02068	Denali Borough	2812.77
02090	Fairbanks North Star Borough	
02240	Southeast Fairbanks Census Area	
02122	Kenai Peninsula Borough	1489.59

Excluded Counties HI	
FIPS Code	**County Name**
15005	Kalawao County
15007	Kauai County
15009	Maui County

Included Counties HI		
FIPS Code	**County Name**	**PSU MOS**
15001	Hawaii County	7624.65
15003	Honolulu County	33361.12

DOT HS 812 804
September 2019

U.S. Department
of Transportation

**National Highway
Traffic Safety
Administration**

14424-091019-4