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

NHTSA

DOT HS 812 706

May 2019

Crash Report Sampling System:
Sample Design and Weighting

DISCLAIMER

Suggested APA Format Citation:

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

Technical Report Documentation Page

1. Report No. DOT HS 812 706	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Crash Report Sampling System: Sample Design and Weighting		5. Report Date May 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 Frances Bents, Jim Green, and the Westat team for their work on CRSS redesign. The authors would also like to thank Phillip Kott for his consultation.			
Abstract As part of the effort to modernize NHTSA's data collection system, NCSA designed a new national crash report probability sampling system, the Crash Report Sampling System (CRSS), to replace the General Estimates System (GES). This report summarizes the sample design and weighting methods used in the new CRSS.			
17. Key Words Crash Report Sampling System, CRSS, CRSS Sample Design, CRSS Weighting, General Estimates System, GES.		18. Distribution Statement 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 63	22. Price

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

Table of Contents

1. Executive Summary 1
2. Introduction 3
3. The Scope of CRSS 6
   - 3.1 NHTSA Data Need 6
   - 3.2 Data Needs of the Public 6
   - 3.3 CRSS Analysis Objectives 7
   - 3.4 CRSS Target Population and Analysis Domains 9
   - 3.5 The Relationship Between the CRSS and the CISS Samples 11
4. An Overview of the CRSS Sample Design 12
5. PSU Sample Selection 13
   - 5.1 PSU Frame 13
   - 5.2 PSU Measure of Size 14
   - 5.3 Minimum PSU Measure of Size 16
   - 5.4 PSU Frame Stratification 18
   - 5.5 PSU Sample Selection 22
     - 5.5.1 Scenario-1 PSU sample 23
     - 5.5.2 Scenario-2 PSU sample 24
     - 5.5.3 Scenario -3 – Scenario-5 PSU samples 25
     - 5.5.4 PSU Sample between Scenarios 26
     - 5.5.5 PSU Sample Selection Summary 28
6. SSU Sample Selection 29
   - 6.1 SSU Measure of Size 29
   - 6.2 SSU Stratification 30
   - 6.3 SSU Sample Selection 30
   - 6.4 EDT PSUs 32
7. TSU Sample Selection 33
   - 7.1 TSU Frame 33
   - 7.2 TSU Sampling Parameters 33
   - 7.3 TSU Sample Selection 35
8. Sample Allocation 37
   - 8.1 Optimization Model 37
   - 8.2 Optimization Results 38

9. Weighting 40
9.1 Design Weights 40
9.2 Non-Response Adjustments 40
9.2.1 Adjustment for Non-Responding PARs 40
9.2.2 Adjustment for Non-Responding PJs 41
9.2.3 Adjustment for Non-Responding PSUs 42
9.3 Adjustment for Duplicates 42
9.4 Post-Stratification (Within PSU Calibration) 43
9.5 Calibration 44
10. Imputation 46
10.1 Item Non-Response 46
10.2 The Sequential Regression Multivariate Imputation 48
10.3 The Univariate Imputation 49
References 51
APPENDIX A. An Example of a PAR A-1
APPENDIX B. Excluded/Included Alaska and Hawaii Counties B-1
APPENDIX C. CRSS PSU Strata for the Five Scenarios C-1

1. Executive Summary

The National Highway Traffic Safety Administration (NHTSA) collects motor vehicle crash data through several systems, including the National Automotive Sampling System (NASS) established in the 1970s to support vehicle and highway safety research, policy making, regulation, and program development.

NASS is comprised of two nested probability sampling systems, the General Estimates System (GES) and the Crashworthiness Data System (CDS). The GES collects general information on traffic crashes only from police crash reports. The CDS collects more detailed information on vehicles and passengers. NHTSA developed and implemented the GES in the 1970s. It is based on a three-stage, stratified probability sample of primary sampling units (PSUs), police jurisdictions (PJs), and police accident reports (PARs). The CDS 24-PSU sample is a subsample of GES 60-PSU sample. The same PSU and PJ samples have been used for GES data collection since the 1980s.

Over the past two decades, however, the general population, vehicles, and highway safety measures have changed dramatically and so have crash characteristic and distributions. 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 need. The U.S. Congress authorized NHTSA to undertake a significant effort to redesign and modernize its crash data collection system. NHTSA identified three major areas for this improvement, redesigning the survey sample, modernizing the information technology infrastructure, and revamping its data collection protocols and technology.

The redesign started in January 2012. The majority of the work was 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 NASS. During this process, two new national probability-based crash sampling systems were designed, the Crash Report Sampling System (CRSS) and the Crash Investigation Sampling System (CISS), to replace GES and CDS, respectively. This report summarizes the sample design and weighting methodology used in the CRSS.

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

CRSS has a stratified three-stage sample design similar to GES: PSU, PJ, and PAR. The PSUs are deeply stratified and selected with probability proportional to the expected number of crash counts based on past experience. In addition, the CRSS PSU sample has been designed to be scalable to accommodate future budgetary fluctuation without completely reselecting the PSU sample.

A new sampling scheme, the Pareto sampling, was used for the PJ sample selection. For PJ sample selection, Pareto sampling produces a scalable PJ sample with selection probabilities approximately proportional to the PJ's crash counts. This scalability can be used to handle PJ frame changes such as the creation, closure or splitting of PJs.

Based on NHTSA's internal and external data need study, to ensure enough sampled cases for estimation on crashes involving motorcycles, commercial vehicles and pedestrians, 10 important analysis domains were identified and used as PAR strata. The overall design weights within each of the 10 important analysis domains are designed approximately equal to improve domain analysis efficiency.

Finally, an optimization technique was applied to find an approximately optimum sample allocation, i.e., the best combination of PSU, PJ, and PAR sample sizes that minimize the variance under a fixed budget.

CRSS design weights are adjusted for non-response at all sampling stages and coverage errors. Estimates using different weight components match with the target benchmarks well.

In summary, the CRSS has been designed as a stratified multi-stage sample survey with scalability built into the first two stages of the sampling process. The sample scalability provides options to adjust for future uncertainties and changes. Also built into this design are protocols that will enable NHTSA to monitor and react to achieve desired sample allocations.

2. Introduction

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

NASS is comprised of two nested systems, GES and CDS, both operated by NHTSA's National Center for Statistics and Analysis (NCSA) to provide national probability samples of crashes.

GES is a survey of PARs (see Appendix A for an example of a PAR). GES collects general information of the traffic crashes from PARs only. GES data provides:

- General and large picture of the crash trends;
- Identification of highway safety problem areas and accesses the size of the problem;
- A basis for regulatory and consumer information initiatives; and
- Form the basis for cost and benefit analyses of highway safety initiatives.

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

While the GES captures general information on all types of traffic crashes, CDS focuses on collecting more detailed information from severe crashes involving passenger vehicles to better understand the crashworthiness of vehicles and consequences to occupants. In addition to PAR information collected, CDS also collects more detailed data about the crashes, vehicles, and occupants through interviews, medical records, vehicle inspections, and scene inspections. CDS data evaluates:

- The overall state of traffic safety and existing and potential traffic safety problems;
- Crash performance, vehicle safety systems, and designs;
- The nature of crash injuries as well as the relationship between the type and seriousness of a crash and the resulting injuries; and
- Traffic safety standards and programs including alcohol and safety belt use programs.

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

Designed in the 1970s and revised in the 1980s, NASS's PSU (a county or a group of counties) sample, and the secondary sampling unit (SSU, a police jurisdiction or a group of police jurisdictions) sample, have not changed for more than 30 years. During this time, however, the NASS sampling frame has had many changes.

- The number and nature of crashes across PSUs
- The population growth and mobility shift

- The 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, however, the highway safety community has been increasingly interested in understanding the factors and reasons leading to a crash in order to develop new crash avoidance safety countermeasures.

The scope of traffic safety studies has also been expanding. With the substantial reductions in passenger vehicle fatalities, more data is needed for crashes involving vehicles and people that CDS currently does not collect detailed data on, such as large trucks, motorcycles, and pedestrians.

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 collection, and reexamining the NASS sample sites and sample sizes.

In the MAP-21 legislation, 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 a part of the effort to modernize NHTSA's data collection system, NCSA designed two new independent national probability crash sampling systems – the Crash Report Sampling System (CRSS) and Crash Investigation Sampling System (CISS) - to replace GES and CDS. This document summaries the sample design and weighting procedures of the CRSS.

In the following chapters, we discuss the scope of CRSS, the formation of the sampling frame, the selection of sample at each sampling stage, the sample size allocation, weighting and imputation. A separate document gives a general guidance and some examples on how to use CRSS data for analysis (see Zhang, Subramanian, Chen, & Noh, 2019).

3. The Scope of CRSS

Data need and research interests have significantly changed since the establishment of NASS. It is critical to identify the data need from NHTSA and public data users and to define the scope of CRSS correctly. This not only includes identifying data elements that are critical to the identification of safety issues, monitoring of trends and evaluation of the effectiveness of countermeasures, but also includes identifying information that is no longer or less useful to the transportation safety research community. To this end, NHTSA conducted two studies to evaluate data need from NHTSA and public data users.

3.1 NHTSA Data Need

In August 2009 NHTSA assembled a project team to conduct a review of the crash data bases and an assessment of current and projected data need. Sixty NHTSA employees, with a broad range of staff expertise and perspective representing all agency offices, were interviewed. The team supplemented the interview data with documented rulemaking and research plans. A report on NHTSA's data needs was submitted to the Congress in 2011 (NHTSA, 2011).

Through this study, NHTSA identified broad goals for the modernized NASS system. These included:

- adding new data elements that support the development of safety countermeasures, especially related to crash avoidance and behavioral safety;
- adding data on motorcycles, commercial vehicles, pedestrians, bicyclists, school buses, and low-speed vehicles;
- providing more data on injuries and on advanced vehicle technologies; and
- enhancing crash analysis through reduced missing data and greater data accessibility, and modifying the research design to better reflect current populations and increased caseload.

3.2 Data Needs of the Public

In order to solicit input 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). 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 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 and remain relevant. The comments covered a wide range of topics:

- 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.

Another key comment is 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. It was also suggested that the new CDS should narrow its scope to collect data on severe crashes along to increase the number of cases of most interest to data users, especially under constrained funding scenarios.

3.3 CRSS Analysis Objectives

The purpose of CRSS is to provide annual, nationally representative estimates of the number, types and characteristics of police-reported motor vehicle crashes. Police Accident Report (PAR) is the sole source of data for CRSS. See Appendix A for an example of a PAR.

Crashes involving motorcycles, commercial vehicles, pedestrians, bicycles, and other road users such as low speed vehicles and ATVs are so called rare populations. Capturing these crashes needs either a very large sample size or a sample design tailored for a particular type of crashes. Motorcycle crashes, for example, are most likely happening in the south and concentrate in a few areas. A sampling system for general passenger vehicle crashes with a small sample size such as CISS will not be able to capture many motorcycle crashes. The most efficient way to study a rare population is to design a special sampling system targets solely on the particular rare population. Therefore, NHTSA decided to capture motorcycle, pedestrian, bicycle and large truck crashes through CRSS since CRSS has a much larger sample size than CISS. If more information about these rare populations is needed, a special study will be designed. This approach will allow both CISS and the special study to be efficient for its own purpose.

The CRSS estimates may then be used for a variety of purposes, including to:

- estimate crash trends;
- identify highway safety problem areas;
- provide a basis for regulatory and consumer information initiatives; and
- form the basis for cost and benefit analysis of highway safety initiatives.

NHTSA's internal and public data need studies also identified the following key estimates and important analysis domains:

- Assessment of the overall state of traffic safety, and identification of existing and potential traffic safety problems.
- The number of police-reported crashes nationwide
- The number of fatalities in police-reported crashes nationwide (based on a 30 day definition of fatality which could be used to compare to FARS[1])
- Vehicle type (passenger car, van, SUV, pickup, medium truck, heavy truck, bus, motorcycle)
- Vehicle age – for example, may be 0-3 years old ("new vehicles"), 4-10 years, and 11+ years
- Counts of crashes by crash severity (fatal injury, incapacitating injury, non-incapacitating injury, property damage only, etc.)
- Counts of vehicles by vehicle type and highest injury severity in the vehicle (or collapsed maximum injury severity to fatality, injured, no injury)
- Impact type (pedestrian, bicyclist, or vehicle)
- Crash type
- Manner of collision: rollover, front, side, rear end
- Single- versus multi-vehicle crashes
- Truck-involved, pedestrian-involved
- Counts of people by age group (from Traffic Safety Facts reports categories = <5, 5-9, 10-15, 16-20, 21-24, 25-34, 35-44, 45-54, 55-64, 65-74, and >74) and injury severity (possibly collapsed injury severity scores)
- Counts of people by person type (person type, possibly collapsed to drivers, occupants, nonoccupants) and injury severity (possibly collapsed injury severity)
- Impact direction (clock direction)
- Vehicle movement (roadway departure, lane/change merge, left turning, etc.)
- Stability of vehicle (jackknife, loss of control)
- Intersection type and traffic control type (might be identifiable from GPS/map data)
- Person type (driver, occupant, pedestrian, cyclist)
- Number of alcohol-related passenger vehicle, motorcycle, pedestrian, and large-truck crashes
- Number of tow away crashes

[1] FARS: Fatality Analysis Reporting System. FARS is a nationwide census of fatal injuries suffered in motor vehicle traffic crashes.

- Number of serious injuries in passenger vehicle, motorcycle, pedestrian, and large-truck crashes

3.4 CRSS Target Population and Analysis Domains

To achieve the CRSS analysis objectives, NHTSA has determined the target population for the CRSS to be all police-reported crashes of motor vehicles (motorcycles, passenger cars, SUVs, vans, light trucks, medium or heavy-duty trucks, buses, etc.) on a traffic way. The CRSS target population is the same as the GES target population.

The research questions and analysis objectives mentioned in the previous section also suggest specific important domains of analysis for CRSS. These important analysis domains will be used to stratify the PARs at PAR sample selection stage therefore they are also referred as PAR strata. NHTSA identified these important analysis domains and revised GES PAR strata. In response to data need, pedestrian, motorcycle and late model vehicle strata were added to CRSS PAR strata. The transportation status for the injured passenger and the tow status for the damaged vehicles are no longer used in CRSS PAR stratification because this information is too costly to identify. Detailed CRSS strata are listed in Table 1 along with the desired target percent of sample allocation.

In Table 1, the "Target Percent of Sample Allocation" column specifies the desired distribution of the sampled cases – for example, 9 percent in analysis domain 2 means 9 percent of the sampled cases should be selected from analysis domain 2. The "Estimated Population" column is the estimated population counts for the analysis domains. The "Population Percent" column is the estimated population distribution over analysis domains. If the "Population Percent" is lower than "Target Percent of Sampling Allocation", then the corresponding analysis domain is oversampled.

Table 1: CRSS Analysis Domains, Target Sample Allocation, and Population Sizes

CRSS Analysis Domain	Analysis Domain Description	Target Percent of Sample Allocation	Estimated Population (GES 2011)	Population Percent
1	An in-scope Not-in-Traffic Surveillance (NiTS) crash (take all)*.			
2	Crashes not in Stratum 1 involving: • A killed or injured (includes injury severity unknown) non-motorist	9%	119,579	2.2%
3	Crashes not in Stratum 1 or 2 involving: • A killed or injured (includes injury severity unknown) motorcycle or moped rider	6%	76,513	1.4%
4	Crashes not in Stratum 1-3 in which: • At least one occupant of a late model year** passenger vehicle is killed or incapacitated	4%	22,272	.42%
5	Crashes not in Stratum 1-4 in which: • At least one occupant of an older** passenger vehicle is killed or incapacitated	7%	84,659	1.6%
6	Crashes not in Stratum 1-5 in which: • At least one occupant of a late model year passenger vehicle is injured (including injury severity unknown)	14%	330,619	6.2%
7	Crashes not in Stratum 1-6 involving: • At least one medium or heavy truck or bus (includes school bus, transit bus, and motor coach) with GVWR 10,000 lbs. or more	6%	302,781	5.7%
8	Crashes not in Stratum 1-7 in which: • at least one occupant of an older passenger vehicle is injured (including injury severity unknown)	12%	800,390	15.0%
9.	Crashes not in Stratum 1-8 in which: • At least one late model year passenger vehicle is involved, AND • No person is killed or injured in the crash	22%	1,511,371	28.4%
10	Crashes not in Stratum 1-9: Mostly property-damage-only crashes involving a non-motorists, motorcycles, mopeds, and passenger vehicles that are not late model year, and any crashes not classified in strata 1-9.	20%	2,078,263	39.0%

*: NiTS cases are not in the scope of CRSS. They are identified and set aside here for NiTS analysis. NiTS in-scope cases are police-reported crashes occurring off the traffic way involving a person who was injured or killed. See NHTSA (2014, DOT HS 811 805) for more detailed information on NiTS.

**: Note:

- Late model year passenger vehicle: passenger vehicle that are ≤4 years old
- Older passenger vehicle: passenger vehicle that are 5 years old or older

3.5 The Relationship Between the CRSS and the CISS Samples

By the NASS design, the CDS 24-PSU sample is a subsample of GES 60-PSU sample. In other words, the CDS is nested within the GES. NHTSA reevaluated the possibility of nesting the CISS within the CRSS. The main advantage of this nested design is potential cost savings by sharing resources between the two surveys. For example, the PARs obtained from the same PSU can be used for both survey's sample selection.

The cost saving by nesting the CISS in the CRSS is mainly on the technician's time spent on obtaining the PARs for sample selection. CISS collects time-sensitive information, so the PARs have to be obtained weekly. The extra cost incurred by separating the CISS from the CRSS is the time spent on obtaining the PARs for the CRSS sample selection. CRSS data collection, however, can be performed in a much longer time interval, and in many PSUs PARs can be obtained electronically. Therefore, the cost saving by nesting the two surveys is small.

On the other hand, the main disadvantage of a nested design is the compromises need to be made for both survey designs, since a set of PSUs selected must meet the needs of both surveys. The major differences between CISS and CRSS include:

- CISS and CRSS have different target population: CISS targets the towed passenger vehicles while CRSS targets the entire universe of police-reported crashes and the vehicle involved in them.
- CISS and CRSS have different operational requirements: CISS requires follow-on and potential on-scene investigation therefore to respond quickly to crashes, the PSUs must not exceed certain geographic size while the requirement for CRSS is to primarily select a large quantity of all types of PARs without any sensitivity to response times.

Because of the differences between CISS and CRSS, tailored PSU formation, stratification, PSU measure of size definition and sample selection independently produce efficient samples for both systems. To optimize both CISS and CRSS, NHTSA decided to have the CISS to be independent from CRSS.

4. An Overview of the CRSS Sample Design

The target population of the CRSS is all police-reported motor vehicle crashes on a traffic way resulting in a PAR. A direct selection of a national probability sample of PARs is infeasible because it requires access to more than 5 million PARs in the Nation. Therefore, the CRSS uses three stage sampling method to select a nationally representative sample from the target population. The PSU is a county or a group of counties. The SSU is a PJ or a group of PJs. The TSU is a PAR.

NHTSA's data need studies identified important analysis domains. To meet the data need, some of the rare analysis domains need to be oversampled in order to have enough cases for analysis. Oversampling introduces unequal selection probabilities in the CRSS design. However, although oversampling results in unequal selection probabilities across the analysis domains, it is still possible to achieve approximately equal selection probabilities within each analysis domain identified in Table 1. Equal selection probabilities within an analysis domain leads to equal weights within the analysis domain therefore results in more efficient domain estimates.

Both multi-stage and unequal selection probabilities often inflate the variance. In order to reduce the potential of variance inflation, stratification is desirable and considered at every stage of CRSS sample selection.

Sample allocation and sample size determination are driven by the budget level, which is currently unknown for future years. In addition, budget levels may fluctuate in the future. 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 and the efforts to establish cooperation from PJs and the recruitment and training of technicians. A highly desirable feature for CRSS is 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 considered for CRSS sample selection. This multi-phase sampling method allows for the selection of a deeply stratified and scalable sample, if changes in the future are needed.

The major sample design changes in CRSS over GES include:

- PAR strata are redefined to better address data need.
- PSUs are reformed to allow more equal design weights within PAR strata.
- PSU sample is deeply stratified.
- PSU sample is scalable.
- A composite PSU MOS defined by the new PAR strata is used.
- PJ sample is scalable.
- Sample allocation is guided by mathematical optimization.
- Mathematical method is used to determine PAR sub-listing factor.

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

5. PSU Sample Selection

This chapter introduces the PSU sample selection method that includes the PSU frame, PSU measure of size, minimum PSU MOS, PSU stratification, and the scalable PSU sample selection procedure.

5.1 PSU Frame

Sampling frame refers to a list of the units of the target population through which sample can be selected and accessed. A one stage direct selection of a national probability sample of crash reports requires access to all PARs in the Nation, which is cost prohibited if not impossible. Instead, the country is partitioned into smaller areas called primary sampling units (PSUs) – a county or a group of counties for CRSS – so a probability sample of PSUs can be selected and local PARs can be further selected. Several factors were considered in the formation of the CRSS PSU frame.

First, for operational efficiency, PSUs were formed to be geographically contiguous so that technicians do not need to drive long distances to collect the PARs.

Second, Census region and urbanicity have been used in GES and proved to be effective PSU stratification variables. Therefore, PSUs were formed to respect Census region and urbanicity.

Third, PSUs were formed to have enough crashes by the PAR strata identified in Table 1. A composite measure of size (MOS) of PSU was calculated by the weighted sum of estimated population counts of PAR strata for each PSU. A PSU with larger desirable combination of estimated population counts of all PAR strata has larger MOS. A minimum PSU MOS was determined to ensure enough PARs in PSUs so that PAR sample for each PAR stratum can be sampled at the desired sampling rate specified by the target sampling rate in Table 1. A county with MOS below the minimum MOS is combined with other contiguous counties to meet the minimum MOS requirement. More details on PSU MOS definition and minimum PSU MOS can be found below.

Fourth, the outlying counties that do not containing a city in Alaska and Hawaii were excluded from the PSU frame because they are remote and have few crashes. See Appendix B for a complete list of areas excluded.

Westat's software WesPSU was used to form the CRSS PSU frame with consideration of the above factors. A total of 707 PSUs were formed from 3,117 counties in the Nation.

In summary, the CRSS PSU frame was formed according to the following criteria:

- PSUs were formed as counties or groups of adjacent counties
- PSUs respected region, State and urbanicity status
- PSUs were required to achieve a minimum size (with few exceptions)
- Outlying areas of Alaska and Hawaii were excluded

5.2 PSU Measure of Size

As Table 1 shows, CRSS collects PARs from a spectrum of different PAR strata at different sampling rate. A PSU with a large number of various PARs should have a larger chance to be selected so that there will be enough PARs to be selected from. To this end, a measure of size (MOS) variable is assigned to every PSU in the frame. A PSU with a larger number of various PARs is assigned a bigger MOS. Then a probability proportional to size (PPS) sampling procedure can be applied using this MOS to select a PPS PSU sample. The CRSS PSU MOS was defined as:

$$MOS_i = \sum_{s=2}^{10} \frac{n_{++s}}{n} \frac{N_{i+s}}{N_{++s}}$$

Here
n = the desired total sample size of PARs
n_{++s} = the desired sample size of PARs in the PAR stratum s
N_{++s} = the estimated population counts in the PAR stratum s
N_{i+s} = the estimated population counts in analysis domain s and PSU i.

In the formula, n_{++s}/n is the desired PAR strata sample allocation (the "Target Percent of Sample Allocation" column in Table 1), and N_{i+s}/N_{++s} is the relative estimated population counts of PSU i for domain s. In this way, a PSU with larger desirable combination of estimated population counts of all PAR strata has larger MOS.

Three potential MOSs were created by using different sources to estimate N_{i+s}. The final PSU MOS was determined by comparing correlations between the potential MOS and outcome variables such as FARS fatal crash counts, State Data System (SDS) crash counts, and Census population.

Table 2 and 3 shows the variables used to estimate N_{i+s} for the final PSU MOS. Using these variables and the MOS formula above, the PSU MOS can be expressed as:

MOS=0.09 (ACSPop)/(sum of ACSPop) + 0.06 (MILE_MC)/(sum of MILE_MC)
+ … + 0.20 (ACSPop*PROPOLD)/(sum of ACSPop*PROPOLD)

The MOSs defined as above are very small numbers. For easier interpretation, all PSU MOSs were multiplied by 10,000,000, which makes MOS≥1 for all PSU without changing their relative magnitudes.

Table 2: Variables Used to Estimate Population Counts by PAR Strata

PAR Stratum	Description	Target Sample Allocation	Variables Used to Estimate Population Counts
1	NiTS crashes (take all).		
2	Crashes involve a killed or injured (includes injury severity unknown) non-motorist	9%	ACSPop
3	Crashes involve a killed or injured (includes injury severity unknown) motorcycle or moped rider	6%	MILE_MC
4	Crashes in which: At least one occupant of a late model year passenger vehicle is killed or incapacitated	4%	FATAL5YR x PROPNEW
5	Crashes in which: At least one occupant of an older passenger vehicle is killed or incapacitated	7%	FATAL5YR x PROPOLD
6	Crashes in which: At least one occupant of a late model year passenger vehicle is injured (including injury severity unknown)	14%	ACSPop x PROPNEW
7	Crashes in which: involved at least one medium or heavy truck or bus (includes school bus, transit bus, and motor coach) with GVWR 10,000 lbs. or more	6%	MILE_TRK
8	Crashes in which: At least one occupant of a passenger vehicle is injured (including injury severity unknown)	12%	ACSPop x PROPOLD
9	Crashes: • Involve at least one late model year passenger vehicle, AND • In which no person in the crash is killed or injured	22%	ACSPop x PROPNEW
10	Crashes not classified in strata 1-9.	20%	ACSPop x PROPOLD

Table 3: Variable Source and Description

Variable	Description	Source
ACSPop	U.S. resident population in 2010	ACS
MILE_MC	Total number of miles driven by motorcycles, 2011	POLK
FA-TAL5YR	Number of fatal crashes, 2007 through 2011	FARS
PROPNEW	Proportion of passenger vehicles that are model year 2008 or newer	POLK
PROPOLD	Proportion of passenger vehicles that are model year 2007 or older	POLK
MILE_TRK	Total number of miles driven by medium or heavy-duty trucks, 2011	POLK

ACS - American Community Survey.
POLK - R. L. Polk & Company.
FARS - Fatality Analysis Reporting System.

5.3 Minimum PSU Measure of Size

Minimum PSU MOS is one of the criteria considered for PSU formation. Minimum PSU MOS ensures enough PARs in PSU so that the selected PARs have approximately equal selection probabilities within each PAR stratum. PARs in PAR stratum 1 (Not-in-Traffic Surveillance) are out of CRSS's scope, therefore PAR stratum 1 is not considered for minimum PSU MOS determination. PAR stratum 4 cases are rare and have very high oversampling rate. Imposing equal weight requirement on stratum 4 may result in PSUs so large that they become inefficient to operate. Therefore, the equal weight requirement is not imposed to stratum 4 for PSU formation purpose.

The MOS computed at the county-level was used to determine the minimum PSU MOS. Minimum MOS was determined separately for each of the eight PSU primary strata- combination of Census region (Northeast, West, South, Midwest) and urbanicity (urban and rural) as following.

Firstly, in order to achieve equal selection probabilities within a PAR stratum, the overall PAR selection probability of the three-stage sample selection should satisfy the following equation:

$$\pi_i \pi_{j|i} \pi_{k|ij} = r_s, \quad \text{for all PAR } k \text{ in stratum } s = 2, 3, 5, \dots, 10.$$

Here π_i is the PSU selection probability, $\pi_{j|i}$ is the conditional PJ selection probability, $\pi_{k|ij}$ is the conditional PAR selection probability, r_s is the sampling rate and calculated by $r_s = n_s/N_s$, where n_s is the number of PARs to be selected from PAR stratum s and N_s is the estimated population size of PAR stratum s. Since selection probability $\pi_{j|i} \leq 1$ and $\pi_{k|ij} \leq 1$,

$$\pi_i \geq max\{r_s, s = 2, 3,5, \dots 10\} = r_{max}$$

Secondly, PSUs are to be selected using probability proportional to size (PPS) sampling from each PSU stratum. Therefore, PSU selection probability becomes:

$$\pi_i = \frac{n_h MOS_{hi}}{\sum_{i=1}^{N_h} MOS_{hi}}, h = 1,2, \dots H.$$

Here n_h and N_h are the PSU sample size and population size for PSU stratum h, H is the total number of PSU strata.

By combining two formulas above it becomes:

$$\frac{n_h MOS_{hi}}{\sum_{i=1}^{N_h} MOS_{hi}} \geq r_{max}, h = 1, 2, \dots H.$$

or

$$MOS_{hi} \geq \frac{1}{n_h} r_{max} \sum_{i=1}^{N_h} MOS_{hi},\ h = 1, 2, \dots H.$$

Therefore, the minimum PSU MOS in the primary PSU stratum h is determined as:

$$MOS_{min} = \frac{1}{n_h} r_{max} \sum_{i=1}^{N_h} MOS_{hi}\,, h = 1, 2, \dots H$$

At this stage, it was anticipated that both census region (Northeast, West, South, Midwest) and urbanicity (urban and rural) will be used as the primary PSU stratification variables. The above condition was applied to $H = 4 * 2 = 8$ PSU strata. The further more detailed secondary PSU strata (see next section) depend on the PSU formation therefore are not considered for minimum PSU MOS. Assuming the total PSU sample size is the same as the GES PSU sample size ($n = 60$), n_h was allocated according to the relative PSU primary stratum MOS distribution. The PSU primary stratum MOS was computed by adding all county-level MOS in the corresponding stratum.

A county with MOS below MOS_{min} is combined with adjacent county to meet the minimum MOS requirement with a few exceptions. Table 4 shows the minimum PSU MOS determined by primary PSU strata along with the PAR stratum with the maximum sampling rate, the maximum sampling rates, stratum MOS, and the number of PSUs in a stratum.

Table 4: CRSS Primary PSU Strata and Minimum PSU MOS

Primary Strata	Stratum Description	Total Number of PSUs in Stratum	PAR Stratum with Maximum r_s	r_{max}	Stratum MOS	MIN MOS
1	Northeast, Urban	56	3	0.1547	1,592,889	15,400
2	Northeast, Rural	44	3	0.0499	287,001	4,776
3	Midwest, Urban	66	2	0.1156	1,316,292	11,700
4	Midwest, Rural	110	2	0.0598	760,146	5,678
5	South, Urban	107	2	0.1916	2,648,551	19,521
6	South, Rural	207	2	0.0511	1,279,257	5,030
7	West, Urban	46	5	0.1270	1,690,816	12,632
8	West, Rural	71	5	0.0370	425,012	3,932

5.4 PSU Frame Stratification

Stratification refers to partitioning sampling frame into non-overlapping sub-populations to allow independent sample selection from each sub-population. A careful selection of stratification variables can produce more balanced sample and reduce the variance of estimates of population parameters. Stratification also allows better sample size control for sub-population estimation. An efficient stratification variable forms homogeneous sub-populations, i.e. minimizing the within sub-population variances and maximizing 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. Crosswalk between Census regions and States can be found at www2.census.gov/geo/docs/maps-data/maps/reg_div.txt. In addition, CRSS PSU MOS is distributed fairly unevenly across the regions. Census regions include:

- Northeast
- West
- South
- Midwest

Urbanicity was also used as a PSU stratification variable resulting in a more demographically balanced and representative PSU sample. Urbanicity also produces more efficient stratification because crash rates are correlated with population densities. In CRSS, urbanicity has two categories:

- Urban PSUs – having a population of 250,000 or greater
- Rural PSUs – otherwise

Census region and urbanicity formed eight (4×2) primary CRSS PSU strata. Within each primary CRSS PSU stratum, Westat's proprietary software WesStrat was used to further stratify the PSUs within each primary PSU stratum using the following stratification variables that were considered correlated with traffic crashes:

- VMT_RATE_IMP = imputed HPMS[2] vehicle miles traveled / (PSU MOS×1,000,000)
- TOT_CRASH_RATE = (imputed 2008 injury crashes+ imputed 2008 PDO crashes + 2007-2011 average fatal crashes) / (PSU MOS×1,000,000)
- TRK_MI_RATE = Total truck miles /(PSU MOS×1,000,000)
- ROAD_TYPE_RATE = (highway/primary road miles +secondary road miles) /(PSU MOS×1,000,000)

PSUs were stratified into equal and homogeneous nested strata. Within each primary PSU stratum, PSUs with similar characteristics based on the stratification variables are grouped into nested strata with approximately equal MOS sizes. The software assists in finding the best nested stratification scheme for minimizing the between-PSU variance within stratum, while attempting to make the stratum population MOS approximately equal. Stratification variables used for further stratification were identified independently within each primary stratum.

The stratification maximized the effect on the following evaluation/outcome measures:

- The average number of fatal crashes across the years 2009-2011
- The sum of the 2008 and 2009 State Data System (SDS) incapacitating injury crashes (which includes imputed values for non-SDS reporting States)
- The sum of the 2008 and 2009 SDS non-incapacitating injury crashes (which includes imputed values for non-SDS reporting States)
- The number of insurance claims in 2006 as reported by HLDI[3]
- The total number of truck crashes from years 2009 to 2012

It was anticipated that CRSS will not be able to implement more than 100 sites. Under the PPS sampling with sample size 100, Los Angeles County was identified as a certainty PSU due to its extraordinary large MOS. It was set-aside and treated as a stratum. Since at least 2 PSUs per stratum are needed for variance estimation, 50 secondary strata were allocated to the 8 primary strata so that each secondary stratum has approximate equal stratum MOS. Table 5 lists the 51 PSU strata (including LA County) along with the upper and lower limits of the stratification variables, stratum total MOS, and the number of PSUs.

Table 5 also describes how the secondary PSU strata were formed within each primary PSU stratum. Note that blanks in the table mean that the particular stratum did not rely on that stratification variable. For example, primary stratum 1 (Northeast, Urban) was further stratified into 8 secondary strata (strata 101 – 108). In the Northeast-Urban, the secondary stratum 101consisted of the PSUs for which VMT_RATE_IMP was between 0 and 1800.66 and for which ROAD_TYPE_RATE was between 0 and 358.504, regardless of the values of TOT_CRASH_RATE and TRK_MI_RATE.

[2] Highway Performance Monitoring System
[3] Highway Loss Data Institute

Table 5: CRSS Secondary PSU Strata and PSU Population Counts

PRIMARY STRATA	STRATID	VMT_RATE_IMP		TOT_CRASH _RATE		TRK_MI_RATE		ROAD_TYPE _RATE		Number of PSUs	MOS
		Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower		
1	1-01	1801	0					359	0	5	222,923
1	1-02	4064	1801					359	0	5	183,529
1	1-03	7159	4064					359	0	8	197,557
1	1-04	5791	0	0.028	0	153756	0	2175	359	6	176,868
1	1-05	8040	5791	0.028	0	153756	0	2175	359	7	204,413
1	1-06			0.028	0	249918	153756	2175	359	7	207,205
1	1-07			0.028	0	591241	249918	2175	359	7	200,876
1	1-08			0.039	0.028			2175	359	11	198,297
2	2-01					236701	0			22	138,907
2	2-02					1027526	236701			22	147,852
3	3-01	4135	0			45709	0			3	189,109
3	3-02	7465	4135			45709	0			8	186,036
3	3-03	9898	7465			45709	0			10	185,606
3	3-04					102554	45709			11	198,246
3	3-05	4444	0			339758	102554			13	183,349
3	3-06	6003	4444			339758	102554			11	189,402
3	3-07	11618	6003			339758	102554			10	183,563
4	4-01					66171	0	4345	0	28	191,482
4	4-02	6045	0			565025	66171	4345	0	27	190,434
4	4-03	11623	6045			565025	66171	4345	0	25	187,745
4	4-04							17641	4345	30	189,376
5	5-01	3620	0	0.048	0	125590	0			5	188,584
5	5-02	4530	3620	0.048	0	125590	0			8	194,117
5	5-03	4951	4530	0.048	0	125590	0			6	159,868
5	5-04	5016	4951	0.048	0	125590	0			3	206,325
5	5-05	5277	5016	0.048	0	125590	0			5	223,732
5	5-06	5746	5277	0.048	0	125590	0			6	149,245
5	5-07	6399	5746	0.048	0	125590	0			5	204,319
5	5-08	12826	6399	0.048	0	125590	0			8	205,760
5	5-09	5641	0	0.048	0	210430	125590			6	191,122
5	5-10	8348	5641	0.048	0	210430	125590			7	195,787
5	5-11	13892	8348	0.048	0	210430	125590			10	173,150
5	5-12			0.048	0	358684	210430			8	198,718

PRIMARY STRATA	STRATID	VMT_RATE_IMP		TOT_CRASH _RATE		TRK_MI_RATE		ROAD_TYPE _RATE		Number of PSUs	MOS
		Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower		
5	5-13			0.048	0	877546	358684			13	192,292
5	5-14			0.085	0.048					17	181,098
6	6-01					49854	0			35	211,282
6	6-02	6353	0			162415	49854			34	209,739
6	6-03	14415	6353			162415	49854			35	213,326
6	6-04					250190	162415			33	213,537
6	6-05	5693	0			1156242	250190			35	208,655
6	6-06	16271	5693			1156242	250190			35	211,752
7	7-00									1	286,050
7	7-01	6477	0	0.027	0	104522	0			7	194,314
7	7-02	6921	6477	0.027	0	104522	0			4	234,422
7	7-03	7861	6921	0.027	0	104522	0			5	169,859
7	7-04	5137	0	0.027	0	249358	104522			3	193,052
7	7-05	8070	5137	0.027	0	249358	104522			10	218,728
7	7-06			0.048	0.027	92716	0			9	177,454
7	7-07			0.048	0.027	186409	92716			7	216,070
8	8-01							3938	0	30	206,694
8	8-02							18292	3938	41	205,285

5.5 PSU Sample Selection

A major challenge of CRSS sample design to NHTSA is the uncertainty over future funding. A fixed PSU sample under the current budget level may not be adequate to handle budgetary changes in the future. On the other hand, reselecting the PSU sample in the future would likely change the existing CRSS data collection sites. Changing the CRSS data collection sites is both costly and time-consuming due to the training of new technicians and establishing corporations with local police departments, etc. Therefore, once the CRSS PSU sample is selected and established, it is cost efficient to keep using it for in the long term.

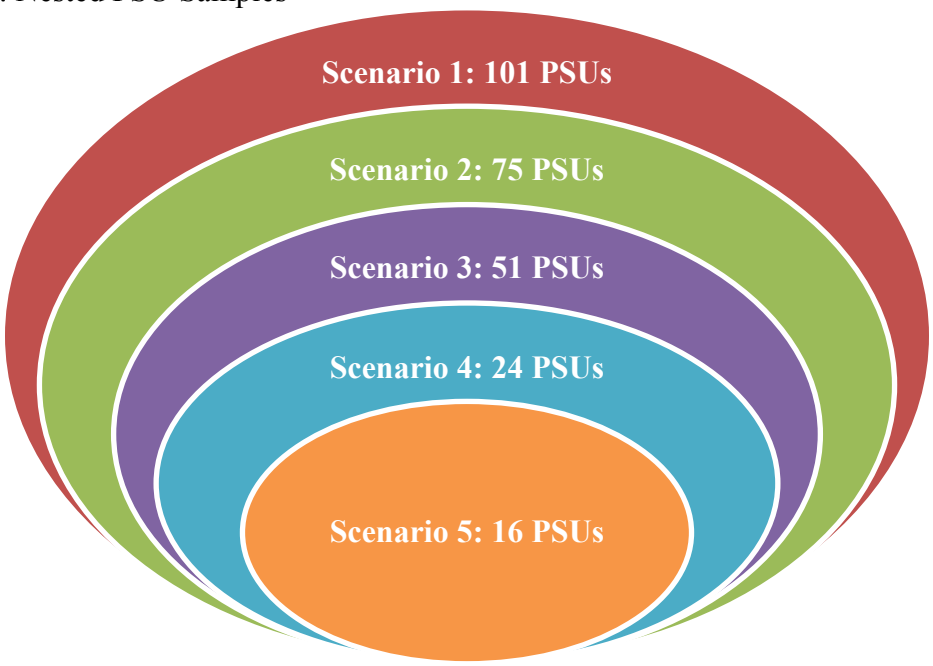
Unknown future funding levels and the need for a stable PSU sample required NHTSA to select a scalable PSU sample so that the PSU sample size can be decreased or increased with minimum impact to the existing PSU sample and the selection probabilities are tractable. To this end, a multi-phase sampling method was used to select the CRSS PSU sample by selecting a sequence of nested PSU samples. This is different from GES where only a single fixed size PSU sample was selected. In this method, a PSU sample larger than actually needed is first selected. Then from this selected first phase PSU sample, a smaller subset PSU sample is selected. Then from this second phase PSU sample, another smaller third phase PSU sample is selected. This process is continued until the PSU sample size reaches unacceptable levels. As Figure 1 shows, this process results in a sequence of nested PSU samples. 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 is picked from the nested sequence. This allows us to track the selection probabilities and minimizes changes to the PSU sample. The following is a detailed description of how this process was applied to CRSS PSU sampling. For CRSS, 5 PSU samples were selected under the 5 scenarios of number of PSU strata and PSU sample sizes. Table 6 summarizes the CRSS PSU sample scenarios.

Table 6: CRSS PSU Sample Scenarios: Number of Strata and Sample Size

Scenario	Number of PSU Strata	Number of Non-certainty PSU	Number of Certainty PSU	Total Number of PSU
1	50	97	4	101
2	37	74	1	75
3	25	50	1	51
4	12	24	0	24
5	8	16	0	16

The following is a detail description of the CRSS PSU sampling method in each scenario and between-scenarios.

Figure 1: Nested PSU Samples



5.5.1 Scenario-1 PSU sample

With initial PSU sample size 100, the PSU frame has been stratified into 50 secondary strata so 2 PSUs can be selected from each stratum. One PSU in the frame has extraordinary large MOS bigger than any single stratum total MOS. This PSU has become a certainty PSU. It has been set aside and treated as a separate stratum (7-00). As the result, the scenario-1 PSU sample has 101 PSUs: 100 PSUs from 50 secondary PSU strata and 1 certainty PSU. The original 50 secondary strata and the only overall certainty PSU are listed under the scenario-1 column of Appendix C. Systematic probability proportional to size (PPS) sampling has been used to select the 2 PSUs from each of the 50 secondary PSU strata. Use $h^{(s)}$ to denote stratum h under scenario-s, $S_{h^{(s)}}$ be all the PSUs in $h^{(s)}$ for scenario-s sample selection purpose, $MOS^{(1)}_{h^{(1)}i}$ be the original MOS assigned to PSU i in scenario-1 stratum h. The PPS sampling interval has been set to be the scenario-1 stratum total PSU MOS divided by 2:

$$I_{h^{(1)}} = \frac{\sum_{i \in S_{h^{(1)}}} MOS^{(1)}_{h^{(1)}i}}{2}, \qquad h^{(1)} = 1, 2, \ldots 50.$$

A random start has been generated and 2 PSUs systematically selected using the above interval from each of the 50 secondary PSU strata. The selection probability of non-certainty PSU i within PSU stratum $h^{(1)}$ under scenario-1 is:

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

In this process, three PSUs had selection probabilities greater than one, so these PSUs were selected with certainty from each stratum.

5.5.2 Scenario-2 PSU sample

Scenario-2 PSU sample size is 74. Scenario-2 PSU strata have been reduced to 37 from 50 of the scenario-1 PSU sample so that 2 PSUs can be selected from each stratum. The same certainty PSU stratum (7-00) in scenario-1 sample has been identified as a certainty PSU again before sample selection. Thus, the scenario-2 PSU sample has 75 PSUs: 74 PSUs from 37 PSU strata and 1 overall certainty PSU.

To form the 37 strata, 13 of the 50 scenario-1 PSU strata have been collapsed with other strata. The collapsing of strata followed the following rules:

- Only the secondary strata in the same primary stratum can be collapsed.
- Only the neighboring secondary strata in the sequence listed in Appendix C can be collapsed.
- The resulting strata have similar stratum total MOS within each primary stratum.

The resulting 37 strata are listed under the scenario-2 column of Appendix C.
After the 37 scenario-2 strata were formed, to select a subsample of the scenario-1 PSU sample, the sampled scenario-1 PSUs were used as the PSU population for scenario-2 sample selection.

Two (2) PSUs were 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 $h^{(1)} = h^{(2)}$, the selection probability was:

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

If a scenario-2 stratum was collapsed from two scenario-1 strata, then it has 4 PSUs, 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:

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

Let $J_{h^{(2)}}$ be the number of scenario-1 strata that were collapsed into scenario-2 strata $h^{(2)}$ and $\left\{h^{(1)}_j\right\}_{j=1}^{J_{h^{(2)}}}$ be those corresponding scenario-1 strata. Thus $h^{(2)} = \bigcup_{j=1}^{J_{h^{(2)}}} h^{(1)}_j$. Let $n_{h^{(1)}}$ be the scenario-1 stratum $h^{(1)}$ PSU sample size. Since scenario-2 sample is selected from scenario-1 sample, there are a total of $N_{h^{(2)}} = \sum_{j=1}^{J_{h^{(2)}}} n_{h^{(1)}_j}$ PSUs available in scenario-2 stratum $h^{(2)}$ for selection. From each collapsed stratum, two PSUs were selected from the pooled $N_{h^{(2)}}$ PSUs using PPS sampling. The resulting PSU selection probability was:

$$\pi^{(2)}_{h^{(2)}i} = \pi^{(1)}_{h^{(1)}i} * \frac{2\sum_{g\in S_{h^{(1)}}} MOS^{(1)}_{h^{(1)}g}}{\sum_{j=1}^{J_{h^{(2)}}} n_{h^{(1)}_j} \sum_{g\in S_{h^{(1)}_j}} MOS^{(1)}_{h^{(1)}_j g}}$$

Typically, $n_{h^{(1)}_j} = 2$ and $\pi^{(1)}_{h^{(1)}i} = 2MOS_{h^{(1)}i}/\sum_{g\in S_{h^{(1)}}} MOS^{(1)}_{h^{(1)}g}$, therefore,

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

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

This is the same selection probability that we would get if we had selected 2 PSUs directly from the collapsed scenario-1 strata using PPS.

For example, scenario-2 stratum $1-02^{(2)}$ (see Appendix C Scenario-2 column) is the result of collapsing two scenario-1 strata: $1-02^{(1)}$ and $1-03^{(1)}$, each has 2 sampled PSUs. If a PSU i in the scenario-2 stratum $1-02^{(2)}$ was in the scenario-1 stratum $1-03^{(1)}$, the selection probability of this PSU is:

$$\pi^{(2)}_{1-02^{(2)}i} = \frac{2MOS^{(1)}_{1-02^{(1)}i}}{\sum_{g\in S_{1-02^{(1)}}} MOS^{(1)}_{1-02^{(1)}g}} * \frac{2\sum_{g\in S_{1-02^{(1)}}} MOS^{(1)}_{1-02^{(1)}g}}{2\sum_{g\in S_{1-02^{(1)}}} MOS^{(1)}_{1-02^{(1)}g} + 2\sum_{g\in S_{1-03^{(1)}}} MOS^{(1)}_{1-03^{(1)}g}}$$

$$= \frac{2MOS^{(1)}_{1-02^{(1)}i}}{\sum_{g\in S_{1-02^{(1)}}} MOS^{(1)}_{1-02^{(1)}g} + \sum_{g\in S_{1-03^{(1)}}} MOS^{(1)}_{1-03^{(1)}g}}$$

This is the same selection probability if we had selected 2 PSUs directly from the combined scenario-1 strata 1-02 and 1-03 using PPS.

5.5.3 Scenario -3 – Scenario-5 PSU samples

Scenario-3 to scenario-5 PSU samples have been selected in a similar way as scenario-2 sample – each scenario PSU sample was a subsample of previous scenario sample, each scenario's PSU strata were either the same as previous scenario or were collapsed from multiple previous scenario's strata, and if PSU strata were collapsed, each PSU in the collapsed stratum was assigned a new MOS equal to the summation of MOS over all sampled PSUs in the same stratum before collapsing. In this way, the scenario PSU samples were nested and the resulting selection probabilities remain PPS in general.

5.5.4 PSU Sample between Scenarios

To select any PSU sample of size between two scenarios, first the PSUs between the scenario PSU samples were randomly sorted in the following sequence (see Appendix C for the sorted sample order):

1. PSUs #1-16: Randomly sort the 16-PSUs in the 16-PSU scenario-5 sample.
2. PSUs #17-24: Randomly sort the additional 8 PSUs in the 24-PSU scenario-4 sample.
3. PSUs #25-51: Randomly sort the additional 27 PSUs in the 51-PSU scenario-3 sample.
4. PSUs #52-75: Randomly sort the additional 24 PSUs in the 75-PSU scenario-2 sample.
5. PSUs #75-101: Randomly sort the additional 26 PSUs in the 101-PSU scenario-1 sample.

The sorted PSU sequence would be used to determine which stratum to be used for the between-scenario sample selection and how many PSUs should be selected from each PSU strata for between-scenarios sample selection. The following is a more detailed description.

A between-scenario PSU sample is selected after the 5 scenario samples were selected. In general, scenario-a stratum $h^{(a)}$ was collapsed from multiple scenario-$(a-1)$ strata: $h^{(a)} = \bigcup_{j=1}^{J_{h^{(a)}}} h_j^{(a-1)}$, here $J_{h^{(a)}}$ is the number of strata collapsed into $h^{(a)}$. Use $(a-1)\sim a$ to denote a scenario between scenario $(a-1)$ and scenario a. Depending on the sample size, $h^{((a-1)\sim a)}$ 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 between scenario-$(a-1)$ and scenario-a sample size for $h^{((a-1)\sim a)}$ which equals to 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.

For example, if a sample of total 60 PSUs (between the 51 scenario-3 PSUs and the 75 scenario-2 PSUs) are to be selected, scenario-3 stratum 1-03 was collapsed from 3 scenario-2 strata: 1-03, 1-04, and 1-05. There were $\sum_{j=1}^{J_{h^{(3)}}} n_{h_j^{(2)}}$ =6 PSUs collapsed into scenario-3 stratum 1-03 (PSU 16, 52, 68, 71, 33, 73), $b_{h^{(2\sim 3)}}$ =3 (PSU 16, 52 and 33 have sorting order no more than 60), and

$n_{h^{(3)}}$ =2 (PSU 16 and 33 were selected into scenario-3 stratum 1-03 sample). Because $\sum_{j=1}^{J_{h^{(a)}}} n_{h_j^{(a-1)}} > b_{h^{((a-1)\sim a)}}$, between scenario sample uses stratum-3 stratum 1-03 and the between-scenario sample size is 3.

As another example, scenario-3 stratum 6-02 was collapsed from 2 scenario-2 strata: 6-02 and 6-03. In this case, $\sum_{j=1}^{J_{h^{(3)}}} n_{h_j^{(2)}}$ =4 (PSU 49, 56, 10, and 55 were collapsed into scenario-3 stratum 6-02 from scenario-2 strata), $b_{h^{(2\sim 3)}} = 4$ (PSU 49, 56, 10 and 55 have sorting order no more than 60), $n_{h^{(3)}}$ =2 (PSU 10 and 49 were selected into scenario-3 stratum 6-02 sample). Because $\sum_{j=1}^{J_{h^{(a)}}} n_{h_j^{(a-1)}} = b_{h^{((a-1)\sim a)}}$, scenario-2 stratum 6-02 and 6-03 are used for between scenario sample and 2 PSUs were selected from each stratum.

The between scenario PSU sample 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, the total PSU sample size 60 is between scenario-3 (51 PSUs) and scenario-2 (75 PSUs). Scenario-3 stratum $1-03^{(3)}$ was collapsed from three scenario-2 strata: $1-03^{(2)}$, $1-04^{(2)}$, and $1-05^{(2)}$. There were total 6 PSUs before scenario-3 sample was selected: $\sum_{j=1}^{J_{1-03^{(3)}}} n_{h_j^{(2)}} = 6$. Two PSUs was selected as scenario-3 sample in $1-03^{(3)}$: $n_{1-03^{(3)}} = 2$.

There are three PSUs with sample order less or equal to 60 (PSU 16, 52, and 33): $b_{1-03^{(2\sim3)}} = 3$ is between 6 and 2. Therefore stratum $1-03^{(3)}$ should be used. We first select 2 PSUs as the scenario-3 sample. We then select the first PSU from the remaining 4 PSUs that were not selected into the scenario-3 sample but were randomly sorted. The selection probabilities for these 3 PSUs are:

$$\pi^{(2\sim3)}_{1-03^{(2\sim3)}i} = \pi^{(3)}_{1-03^{(3)}i} + \left(\pi^{(2)}_{h_i^{(2)}i} - \pi^{(3)}_{1-03^{(3)}i}\right) * \frac{1}{4}$$

(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:

$$\pi^{((a-1)\sim a)}_{h^{((a-1)\sim a)}i} = \pi^{(a)}_{h^{(a)}i}$$

Appendix C lists the complete sampling order of all 101 PSUs along with the stratification of each of 5 scenarios. Any one of the PSU samples in the sequence (either one of 5 scenarios or between-scenarios) is a probability sample and can be used for data collection. If a larger or smaller PSU sample is desirable, simply find a bigger or smaller sample from the nested sequence. This scalable PSU sample allows us to adjust PSU sample size without changing the sampled PSUs and also allows us to tract the selection probabilities.

5.5.5 PSU Sample Selection Summary

NHTSA has designed the CRSS PSU sample so that the PSUs can be added (or deleted) as budget changes occur, without having to reselect the entire sample again. Meanwhile, the PSUs were kept deeply stratified to produce efficient estimates while the resulting selection probabilities are still proportional to PSU MOS so that PSUs with large desired proportion of crashes defined by the PAR strata are more likely to be selected.

Currently up to 101 PSUs can be used for CRSS data collection. However, since the original scenario-1 non-certainty PSUs were selected by systematic PPS sampling method within each original scenario-1 PSU stratum, the PSU sample size can be expanded by reducing the systematic sampling interval to half in some or all PSUs strata. This allows us to further expand the CRSS PSU sample size beyond 101 while keep the original 101 PSUs as part of the PSU sample.

6. SSU Sample Selection

PARs are filled out by police officers and reported to the State through a police jurisdiction. For the CRSS, PARs can be obtained from PJs either by visiting the PJs or by electronic transmission. In this way, PJs are viewed as nature clusters of PARs. The CRSS secondary sampling units are police jurisdictions (PJs) that produce PARs for the crashes occurred within the sampled PSUs. In order to create PJ frame, NHTSA collected PJ information for the PJs that reported crash information to the State in the years of 2010 - 2012 for the 75 PSUs of the scenario-2 PSU sample. Among the PJ information, the following 6 types of crash counts were collected to compute SSU MOS:

- Total crashes
- Fatal crashes
- Injury crashes
- Pedestrian crashes
- Motorcycle crashes
- Commercial motor vehicle (CMV) crashes

If multiple PJs in a PSU have the same name and address, which are mostly State police offices, these PJs were combined into one PJ. If a State police office generates PARs for multiple PSUs, the State police office is treated as multiple PJs, each corresponding to the portion of PARs generated for the corresponding PSU.

6.1 SSU Measure of Size

Similar to the PSU MOS definition, it is sensible to assign larger selection probability to PJs with larger number of crashes in desirable crash composition. 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 strata in Table 1 as follows:

$$MOS_{j|i} = 0.11 \times (Fatal\ crashes) + 0.26 \times (Injury\ crashes) + 0.09 \times (Pedestrian\ crashes) + 0.06 \times (Motorcycle\ Crashes) + 0.06 \times CMV\ Crashes) + 0.42 \times (Total\ crashes - Fatal\ crashes - Injury\ crashes - Pedestrian\ crashes - Motorcycle\ crashes - CMV\ crashes)$$

Strata 4 and 5 were considered as a fatal crash group, and strata 6 and 8 as injury crash group. This rough PJ MOS is used for PJ stratification (see below).
Second, a finer PJ MOS was created for the PJ sample selection. Crash counts of the 9 PAR strata in Table 1 for each PJ in the selected PSUs were estimated based on the 6 types of crash counts collected in the PJ frame and other PJ level information. For PJ j in the PJ frame within the sampled PSU i, the composite SSU MOS is defined as the following:

$$MOS_{ij} = \sum_{k=2}^{10} \frac{n_{++s}}{n} \frac{N_{ijs}}{N_{++s}}$$

where

n = the desired total sample size of crashes
n_{++s} = the desired sample size of crashes in the PAR stratum s
N_{++s} = the estimated population number of crashes in PAR stratum s
N_{ijs} = the estimated population number of crashes in PAR stratum s, PJ j and PSU i

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

6.2 SSU Stratification

PJ MOS varies dramatically within the selected PSUs. To reduce the sampling variance, the PJ frame within each selected PSU was stratified using the coarse PJ MOS.

If a PSU had less than 9 PJs, all PJs were assigned to certainty stratum. Therefore, all PJs in these PSUs were selected with selection probability one.

For other PSUs, certainty PJs were first identified using the following condition:

$$\frac{2MOS_{ij}}{\sum_j MOS_{ij}} \geq 1$$

Here MOS_{ij} is the coarse 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, almost certainty PJs were also identified with the following condition:

$$1 > \frac{2MOS_{ij}}{\sum_j MOS_{ij}} > 0.7$$

All certainty PJs and almost certainty PJs identified through the above process were assigned to the certainty stratum, i.e., they were selected with selection probability one.

The non-certainty PJs were sorted by their PJ MOS within each selected PSU in descending order. Roughly half of PJs with larger PJ MOS were assigned to the large MOS stratum and the other half of PJs were assigned to the small MOS stratum. Therefore, for the PSUs with 9 or more PJs, as many as three PJ strata were formed: the certainty stratum, the large MOS stratum, and the small MOS stratum.

6.3 SSU Sample Selection

One of the major challenges of the SSU sample selection is the PJ frame changing. Unlike PSUs, PJs are relatively unstable as new PJs may emerge or existing PJs may split, merge, or close. The PJ MOS is determined by crash counts that are subject to variation every year and hence the PJ

stratum may also change. In addition, setting up cooperation with the PJs is time consuming and the PJs may refuse to cooperate.

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

Pareto sampling method is applied to the PJ sample selection for each of non-certainty PJ strata (large MOS or small MOS stratum) within the sampled PSU i, as the following:

- Generate a permanent uniform random number $r_{ij} \sim U(0,1)$ for each PJ j in the PJ frame.
- Identify certainty PJs by the 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. And this process is repeated to the remaining PJs with the reduced PJ sample size until there is no more certainty PJs. Let the total number of certainty PJs be m_c.

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

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

- Calculate transformed random numbers:

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

- Sort the transformed random number in ascending order.
- The m_c certainty PJs and 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 this way, the resulting PJ selection probability is approximately PPS (Rosén, 1997). NHTSA conducted a simulation study on the described Pareto sampling strategy to CISS PJ sample selection. The result of this study shows Pareto selection probability is very close to PPS selection probability (Noh & Zhang, 2017).

For the non-certainty PJs, the conditional PJ inclusion probability given PSU selected is:

$$\pi_{j|i} \approx p_{ij}$$

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: $\frac{r_{ij}(1-p_{ij})}{p_{ij}(1-r_{ij})}$ does not change. 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. Therefore, when PJ sample has

to be re-selected, the change to the existing PJ sample under Pareto sampling is much smaller than a regular PPS sampling.

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 SSU strata proportionally to the total stratum PJ MOS (using the finer PJ MOS) with at least one PJ per stratum.

6.4 EDT PSUs

In the process of implementing the CRSS, NHTSA learnt that in 2018 sample year, 14 sampled PSUs provide PARs in electronic formats through NHTSA's Electronic Data Transfer (EDT) system and more PSUs are expected to provide PARs through EDT system in the future. An important feature of EDT is that PARs can be singled out by their crash location electronically. This gives an opportunity for NHTSA to eliminate the second stage, i.e. the PJ, sampling in these EDT PSUs because once all PARs in a sampled PSU can be identified electronically by the crash location, then the PAR sample can be selected directly without selecting the PJs.

There are three major advantages to eliminate the PJ sample selection. First, there is no need for maintaining and monitoring PJ frames changes. Second, it removes the sampling error of PJ sample selection. Third, if there is a PJ sample selection in the EDT PSUs, then the PJs associated with the EDT PARs from the State need to be mapped to the PJs in the PJ frame to determine whether the PARs should be listed/sampled or not. Sometimes this mapping can be difficult and inaccurate. Incorrect PJ mappings causes PARs grouped into the wrong PJs. Eliminating PJ sampling can eliminate these coverage errors.

The PJ sample selection stage is eliminated in these 14 EDT PSUs starting from 2019. In each of these 14 PSUs, all PARs will be assigned to the same pseudo PJ and this pseudo PJ will be selected with certainty (therefore PJ weight equals to one) so that the existing three-stage sampling procedure is nominally preserved. The PAR sampling procedure remains the same as described in the next chapter. In this way, the elimination of PJ sample selection in EDT PSUs can be implemented with minimum changes to the existing process. Since more PSUs transition to electronic PAR format, it is expected to eliminate PJ sample selection stages for more PSUs in the future.

7. TSU Sample Selection

7.1 TSU Frame

The CRSS TSUs are PARs. For each selected SSU (PJ), PARs are periodically obtained by technician's visit to the PJ or by electronic transmission. The date when PARs are obtained is called contact date. The PARs are listed in the order they become available, and stratified by the PAR strata identified in Table 1. In this listing process, PAR sampling frame in each selected PJ are prepared for PAR sample selection.

For a large PJ with too many PARs to be listed, PARs are sub-listed. For example, only PARs with even PAR numbers are listed if a sub-listing factor is 2, or 1 of every 5 PARs is listed if a sub-listing factor is 5. Sub-listing is equivalent to a systematic sampling.

7.2 TSU Sampling Parameters

CRSS PAR sample is selected by a stratified systematic sampling from the listed or sub-listed PARs by PAR stratum within a PJ.

The PAR sampling interval is determined in the following manner. The goal is to achieve an approximately equal inclusion probability for all PARs in the same PAR stratum. Therefore, for all PARs in PAR stratum s, the overall inclusion probability(π_{ijlk}) is set to the sampling rate (r_s) of the PAR stratum s as:

$$\pi_{ijlk} = r_s = \frac{n_s}{N_s}, \qquad \text{for all PAR } k \text{ classified into stratum } s.$$

Here n_s is the desired (or target) PAR sample size and N_s is the estimated total number of PARs in the population for the PAR stratum s.

On the other hand, the overall inclusion probability π_{ijlk} is the result of PSU selection, PJ selection, sub-listing, and PAR sample selection. Therefore,

$$\pi_{ijlk} = \pi_i \pi_{j|i} \pi_{l|ij} \pi_{k|ijl}$$

Here π_i is the selection probability of PSU i, $\pi_{j|i}$ is the selection probability of PJ j given that PSU i is selected, $\pi_{l|ij}$ is the probability that PARs are sub-listed as a cluster, and $\pi_{k|ijl}$ is the selection probability of PAR k given that PARs are sub-listed. By combining two equations above, the selection probability of PAR k from PAR stratum s becomes,

$$\pi_{k|ijl} = \frac{1}{\pi_i \pi_{j|i} \pi_{l|ij}} \times r_s, \qquad \text{for PAR } k \text{ from PAR stratum } s$$

Therefore, the PAR sampling interval, which is the inverse of the PAR selection probability, is determined as

$$w_{k|ijl} = max \left\{ 1, \quad \frac{1}{w_i w_{j|i} w_{l|ij} r_s} \right\}$$

Here, $w_i = 1/\pi_i$ is the PSU weight, $w_{j|i} = 1/\pi_{j|i}$ is the PJ weight, $w_{l|ij} = 1/\pi_{l|ij}$ is the sub-listing factor, and r_s is the sampling rate of PAR stratum s. It is possible that $1/w_i w_{j|i} w_{l|ij} r_s$ is less than one. In that case, PAR k is selected with certainty, and the sampling interval $w_{k|ijl}$ is set to one and equal weight can't be achieved for that certainty case. In CRSS, PAR sampling interval $w_{k|ijl}$ is a real number. The following section describes how the real number interval works in the PAR sample selection.

Although sub-listing reduces the listing cost, the sub-listing factor should be carefully defined so that there are enough listed PARs for the PAR sample selection. From the fact that $\pi_{k|ijl} \leq 1$ we have:

$$w_{k|ijl} = \frac{1}{w_i w_{j|i} w_{l|ij} r_s} \geq 1$$

That is,

$$w_{l|ij} \leq \frac{1}{w_i w_{j|i} r_s}$$

Since PARs are listed for all PAR strata in a given PJ, the above inequality must be met for all PAR strata. Therefore, a preliminary sub-listing factor is determined as:

$$Preliminary\ w_{l|ij} = integer\left\{min\left(\frac{1}{w_i w_{j|i} r_s},\ s = 2,3, \ldots 10\right)\right\}$$

In order to use the last digit of PAR number or PAR ID in the implementation of sub-listing, preliminary determined sub-listing factor is adjusted as shown in Table 7. First, percentage of listed PARs from the preliminary sub-listing factor is rounded up, and translated into the number of listed PARs out of 10 PARs. Then, the final sub-listing factor is determined accordingly.

Table 7: Determination of the Final Sub-Listing Factors

Preliminary		**Rounded-up Percentage of Listed PARs**	**Number of Listed PARs out of 10 PARs**	**Final Sub-listing Factor**
Sub-listing Factor	**Percentage of Listed PARs**			
1	100%	100%	10	1
2	50%	50%	5	2
3	33.3%	40%	4	2.5
4	25%	30%	3	3.333
5	20%	20%	2	5
6	16.7%			
7	14.3%			
8	12.5%			
9	11.1%			
10+	10%	10%	1	10

7.3 TSU Sample Selection

On a contact date, for each sampled PJ, technician first stratifies new PARs accumulated since last contact date into the 9 CRSS PAR strata listed in Table 1 in the order as the PARs become available. If sub-listing is required in this PJ, then the technician only stratifies the sub-listed PARs. This process is referred as "PAR listing." As the result, in each sampled PJ, PARs are sequentially listed under PAR strata. From each list, a systematic PAR sample is selected using the pre-determined sampling intervals. The following is a more detailed description of this process.

Let the pre-determined sampling interval be $Interval_{ijlsk}$ which is a real number equal to or greater than 1, and uniform random number between (0,1) be u_{ijlsk} for PAR i of PSU i, PJ j, sub-list l, and PAR stratum s. Then, the PAR sample is selected in a contact date as following:

(A) For the first contact date,
  (1) For the first listed PAR (i.e., $k = 1$),
    (a) Define the PAR sequence as
    $$SEQ_{ijls1} = 1$$
    (b) Calculate the real number sampling unit, and the integer sampling unit as
    $$RSU_{ijs1} = u_{ijls1} Interval_{ijls1},$$
    $$ISU_{ijs1} = ceiling\left(RSU_{ijs1}\right).$$
    Here the "$ceiling$" function rounds up RSU_{ijs1} to the smallest integer that is greater than or equal to RSU_{ijs1}.
    (c) Define Sampling flag as following, and select the PAR if sampling flag is 1
    $$SampFlag_{ijls1} = \begin{cases} 1, & \text{if } ISU_{ijls1} = SEQ_{ijls1} \\ 0, & \text{Otherwise} \end{cases}$$
  (2) For the following listed PAR k (i.e., $k > 1$), repeat this step until the last listed PAR.
    (a) Define the PAR sequence as
    $$SEQ_{ijlsk} = SEQ_{ijls(k-1)} + 1$$
    (b) Calculate the real number sampling unit and the integer sampling unit as
    $$RSU_{iljsk} = RSU_{ijls(k-1)} + (SampFlag_{ijls(k-1)} \times Interval_{ijslk}),$$
    $$ISU_{ijlsk} = ceiling\left(RSU_{ijlsk}\right).$$
    (c) Define Sampling flag as following, and select PAR k if sampling flag is 1
    $$SampFlag_{ijlsk} = \begin{cases} 1, & if\ ISU_{ijlsk} = SEQ_{ijlsk} \\ 0, & Otherwise \end{cases}$$
(B) For the following contact dates
  (1) List PARs right after the listed PARs from the previous contact date, and do step (A)(2).

CRSS data are collected from PARs. PARs are sampled only if they are available to be listed. Therefore, there is no refusals in CRSS and PAR replacement sample is unnecessary for CRSS. For more details on CRSS PAR sample selection, see Noh et al. (2016): "CRSS PAR Selection Algorithm and the IT Aspects."

From each sampled PAR, approximately 120 data items about crash, event, vehicle and people were coded. See Crash Report Sampling System analytical user's manual 2016 (NHTSA, 2018) for more information about data items coded in CRSS data files.

8. Sample Allocation

CRSS data is collected through three stages of sample selection – a PSU sample, a PJ sample and a PAR sample. In this chapter, we describe how NHTSA found an approximately optimum sample allocation, i.e., the best combination of PSU, PJ and PAR sample sizes that minimizes the variance under fixed budget.

Optimum sample allocation is an optimization problem. We used a non-linear problem to find the optimal PSU sample size n, PJ sample size m, and PAR sample size k by minimizing the overall variance of the proportion estimates of thirteen key estimates under the fixed budget. We also considered variance constraints which ensure the new sample design for the CRSS will be at least as precise as the current GES for the identified key estimates.

In order to build an optimization model for the CRSS, we simplified the complex sample design. The CRSS has a stratified multi-stage probability proportionate to size (PPS) sample design. The deep PSU stratification led to 2 sampled PSUs from each PSU stratum as discussed in Chapter 5. Taking the PSU or PJ stratification into account adds too many constraints to the first two stages and leaves only the PAR sample size to be completely optimized. In addition, taking the unequal PPS selection probabilities into account makes the variance estimation model complicated. Therefore, NHTSA used three stage simple random sampling without replacement in the optimization model for simplicity.

8.1 Optimization Model

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

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

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

$$V\left(\bar{\bar{\bar{y}}}_g\right) = \frac{S_{1,g}^2}{n}\left(1-\frac{n}{N}\right) + \frac{S_{2,g}^2}{nm}\left(1-\frac{m}{M}\right) + \frac{S_{3,g}^2}{nmk}\left(1-\frac{k}{K}\right) \le V_{GES}\left(\bar{\bar{\bar{y}}}_g\right), \quad \text{for } g = 1, \cdots, G.$$

- g: Subscript of the identified key estimate, $g = 1, \cdots G$. Here $G = 13$.
- $\bar{\bar{\bar{y}}}_g$: Proportion estimate of the key variable.
- n, m, k: Optimal sample sizes of PSUs, PJs per PSU, and cases (PARs) per PJ to be determined.
- N: Population size of PSUs (N=707).
- M: Average population size of PJs per PSU (M=27).
- K: Average population size of PARs per PJ (K=1,688).
- $V\left(\bar{\bar{\bar{y}}}_g\right)$: Variance of the identified key estimate $\bar{\bar{\bar{y}}}_g$.
- $S_{1,g}^2,\ S_{2,g}^2,\ S_{3,g}^2$: Variance component at PSU-, PJ-, and PAR-level.
- C, C_0, C_1, C_2, C_3: Total, fixed, PSU-, PJ-, and PAR-level cost coefficients.
- $V_{GES}\left(\bar{\bar{\bar{y}}}_g\right)$: Variance of the identified key estimate $\bar{\bar{\bar{y}}}_g$ in the current system (NASS GES).

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

NHTSA conducted a cost analysis through the GES collection activity. Based on the results of this analysis and other accounting information, the CRSS cost coefficients were estimated. The detail of the cost estimation is in Noh (2014).

Thirteen key variables were identified to be considered in the objective function. These key variables were also used in the variance constraints to ensure the CRSS will produce equal or smaller variance for these variables than GES. The key variables are: fatal crash, incapacitating injury crash, non-incapacitating injury crash, collision with moving vehicle in transport, passenger car, light truck & van, bus, medium/heavy truck, motorcycle, rollover vehicle, impact-front, impact-side, impact-rear.

The 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 thirteen key variables from 3 year GES data (2009~2011). A range of total costs were considered. Five hundred starting points (msnumstarts=500) were used in SAS PROC OPTMODEL to find the global optimum solution. More detailed information on NHTSA's optimization can be found in Noh and Zhang (2016).

8.2 Optimization Results

Table 8 lists the optimization results by rescaled budget levels. In this Table, budget levels were rescaled from \$1 to \$2.5. When budget level increases, SSU sample size m and TSU sample size k tend to be stable. It is mainly the PSU sample size n steadily increasing. This is consistent with the objective function which indicates factor $1/\mathrm{n}$ affects all three terms of the total variance therefore increasing PSU sample size is generally the most effective way of reducing the total variance.

The PAR sample size k was obtained for the general population as one domain. In CRSS, however, separate estimates are made for the 9 PAR strata/domains. Therefore, k PARs needed for each of the 9 PAR strata. The last two columns of Table 8 lists the corresponding PAR counts and estimated total costs.

The objective value is the average variance of the key estimates. The variances of the key estimates under current GES were used as constraints. Therefore, under these sample allocations, we expect these key estimates have equal or smaller variance under CRSS than the corresponding GES estimates. As PSU sample size increases, the average variance decreases, therefore some or all these key estimates will have even smaller variances.

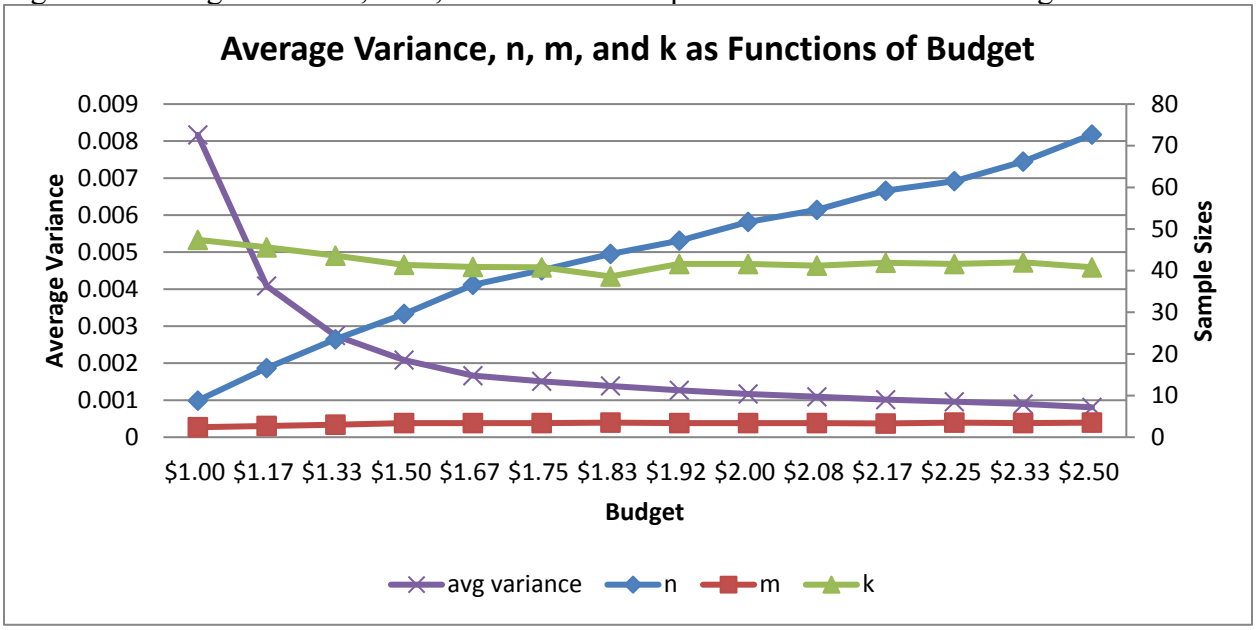
Figure 2 displays the optimization results. As the rescaled budget increases, the PJ sample size m and the PAR sample size k tend to be stable while the PSU sample size n increases and the average variance decreases.

Table 8: CRSS Optimization Results (With Variance Constrains of Individual Key Estimate)

Objective Value	Budget (C)	n	m	k	Sample Size (nmk)	Sample Size (nm9k)	Rescaled Cost (nm9k)
0.0081687	$1.00	8.8	2.4	47.4	995	8,957	$1.19
0.0040891	$1.17	16.6	2.7	45.6	2,028	18,253	$1.56
0.0027454	$1.33	23.5	3	43.6	3,079	27,714	$1.93
0.0020847	$1.50	29.6	3.4	41.4	4,109	36,984	$2.29
0.0016636	$1.67	36.6	3.4	40.9	5,118	46,058	$2.66
0.0015095	$1.75	40.1	3.4	40.8	5,636	50,724	$2.84
0.0013852	$1.83	44	3.5	38.6	5,881	52,928	$2.97
0.0012682	$1.92	47.2	3.4	41.6	6,777	60,992	$3.23
0.0011683	$2.00	51.7	3.4	41.6	7,225	65,024	$3.40
0.0010927	$2.08	54.6	3.4	41.2	7,752	69,772	$3.58
0.0010145	$2.17	59.2	3.3	41.9	8,288	74,590	$3.77
0.0009586	$2.25	61.5	3.5	41.6	8,878	79,906	$3.97
0.0008972	$2.33	66.2	3.4	42	9,375	84,374	$4.15
0.0008072	$2.50	72.7	3.5	40.8	10,268	92,414	$4.49

Note: All costs are rescaled so the lowest cost starts from $1.

Figure 2: Average Variance, PSU, PJ and PAR Sample Size as Functions of Budget



Note: All costs are rescaled so the lowest cost starts from $1.

9. Weighting

The CRSS sample is the result of a probability sampling with sampling features such as stratification, clustering, and unequal selection probabilities. Because of this, CRSS sample is not a simple random sample. CRSS sample should be properly weighted to produce unbiased estimates. Unweighted estimates may be severely biased. This chapter describes how CRSS weights were calculated.

The CRSS weights are created in the following steps:

- Design weights at all three stages
- Non-response adjustments at all three stages
- Duplicate adjustment
- Post-stratification (i.e., within PSU calibration)
- Calibration of case weight

9.1 Design Weights

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

$$w_{ijlk} = w_i * w_{j|i} * w_{lk|ij}$$

Here

- $w_i = \pi_i^{-1}$ is the inverse of PSU i selection probability,
- $w_{j|i} = \pi_{j|i}^{-1}$ is the inverse of PJ j selection probability,
- $w_{lk|ij} = w_{l|ij} * w_{k|ijl}$ is the inverse of PAR selection probability which is decomposed as sub-listing and PAR selection. $w_{l|ij}$ is the sub-listing factor, and $w_{k|ijl}$ is the sampling interval.

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

9.2 Non-Response Adjustments

The CRSS sample suffers from non-responses at all three sampling stages: PAR, PJ and PSU. Estimation without non-response treatment may be severely biased. This section describes how adjustments were made at each sampling stage to mitigate the non-response bias.

9.2.1 Adjustment for Non-Responding PARs

PARs with missing pages and/or non-readable pages cannot be coded therefore are treated as non-responding PARs. On the other hand, in some PJs, some crash reports become available for listing and sampling after the cut-off date of annual sample selection. For these reasons, PAR non-response adjustment is performed by calibrating the estimated PAR counts to the listed PAR counts by PSU and PAR strata.

The following non-responding PAR adjustment factor adj_{is} is calculated for each PAR stratum s of PSU i:

$$adj_{is} = \sum_{j \in r_i} w_{j|i} w_{l|ij} L_{ijls} \Big/ \sum_{j \in r_i} \sum_{k \in r_{ijs}} w_{j|i} w_{lk|ij}$$

Here r_i is the set of responding PJs of PSU i. r_{ijs} is the set of responding PARs, and L_{ijls} is the number of listed PARs in PAR stratum s, PJ j, PSU i. $w_{j|i}$ is the PJ weight, and $w_{lk|ij}$ is the PAR weight.

The adjusted PAR weights are computed by multiplying non-responding PAR adjustment factor to the PAR weights for the responding PARs and by setting the weights for the non-responding PARs set to zero.

$$w_{lk|ij}^{(1)} = \begin{cases} w_{lk|ij} * adj_{is}, & for\ responding\ PARs \\ 0, & for\ nonresponding\ PARs \end{cases}$$

PAR design weights for the listed but non-sampled PARs are unchanged and kept for duplicate PAR adjustment (see section 9.3).

9.2.2 Adjustment for Non-Responding PJs

When a sampled PJ refuses to cooperate, it becomes a non-responding PJ. If PJ non-cooperation is identified prior to the data collection of the sampling year, PJ sample is augmented and replacement PJ is selected. Then PJ non-response adjustment is conducted before sampling parameters are determined. Since this process is conducted in the sample design stage, the adjustment is already considered in the PJ design weight. However, if PJ non-cooperation is identified during the data collection, PJ non-response adjustment is conducted in the weighting stage. The adjustment factor is computed using the weighted PJ MOS response rate for PSU i:

$$adj_i = \sum_{j \in s_i} w_{j|i} MOS_{ij} \Big/ \left(\sum_{j \in s_i} w_{j|i} MOS_{ij} - \sum_{j \in nr_i} w_{j|i} MOS_{ij} \right)$$

Here s_i is the set of sampled PJs (excluding non-responding PJs identified in the design stage), and nr_i is the set of non-responding PJs identified during the data collection in PSU i. MOS_{ij} is the finer PJ MOS.

Then, PJ weight is adjusted by multiplying PJ non-response adjustment factor adj_i as:

$$w_{j|i}^{(1)} = \begin{cases} w_{j|i} * adj_i & \text{for responding PJ} \\ 0 & \text{for non} - \text{responding PJ} \end{cases}$$

9.2.3 Adjustment for Non-Responding PSUs

A non-responding PSU is a sampled PSU that refuses to cooperate. If PSU non-cooperation is identified prior to the data collection of the sampling year, PSU sample is augmented and a replacement PSU is selected. PSU non-response adjustment is then conducted before PAR sampling parameters are determined. Since this process is conducted in the sample design stage, the adjustment is already considered in the PSU design weight. If PSU non-cooperation is identified during the data collection, PSU non-response adjustment is conducted in the weighting stage. The adjustment factor is calculated using the weighted PSU response rate by the urbanicity (urban or rural):

$$Adj_c = \sum_{i \in s_c} w_i \Big/ \sum_{i \in r_c} w_i$$

Here c represents the PSU non-response adjustment cell (i.e., urban or rural). s_c is the set of sample PSUs (excluding non-responding PSUs identified in the design stage), and r_c is the set of responding PSUs in the cell c.

Non-response adjusted PSU weight is computed by multiplying the adjustment factor to the PSU weight:

$$w_i^{(1)} = \begin{cases} w_i * Adj_c, & \text{for responding PSU} \\ 0, & \text{for nonresponding PSU} \end{cases}$$

For example, in 2016, seven out of the sixty originally selected PSUs were non-responding – resulting a final responding PSU sample of size 53. Adjustment was conducted for the seven non-responding PSUs in the weighting. In 2017, six of the seven non-responding PSUs were converted to responding PSUs. The 2017 PSU sample was augmented to 61 PSUs – resulting a final responding PSU sample of size 60. Since adjustment was conducted for one non-responding PSU in design stage, non-response adjustment was not conducted in weighting.

9.3 Adjustment for Duplicates

Police sometimes submit multiple PARs for the same crash with updated information. A crash with multiple PARs have multiple chances to be selected. Assume crash k has n PARs in the listed PARs: $k_1, k_2, \ldots k_n$. Let $\pi_{lk_u|ij} = 1/w^{(1)}_{lk_u|ij}$ $(u = 1, 2, \ldots, n)$ be the inclusion probability of each duplicated PAR. Here $w^{(1)}_{lk_u|ij}$ is the non-response adjusted PAR weight for the PAR k_u. The overall inclusion probability for the crash is: $\pi_{lk|ij} = 1 - \prod_{u=1}^{n} \left(1 - \pi_{lk_u|ij}\right)$. All identified duplicate PARs for the same crash are used to capture the complete and updated information for the crash but only one crash record is kept in the final analysis file. Therefore, the PAR weight adjusted for duplicates becomes:

$$w^{(2)}_{lk|ij} = \begin{cases} w^{(1)}_{lk|ij} & \text{if a crash does not have duplicate PARs} \\ 1/\pi_{lk|ij} & \text{if a crash has duplicate PARs} \end{cases}$$

After non-response adjustments and duplicate adjustment, within-PSU PAR weight and case weight are calculated, respectively as:

$$w_{jlk|i}^{(1)} = w_{j|i}^{(1)} * w_{lk|ij}^{(2)}$$
$$w_{ijlk}^{(1)} = w_i * w_{jlk|i}^{(1)}$$

If PSU non-response was conducted in the weighting stage (as in 2016), $w_i^{(1)}$ is used instead of w_i in the case weight calculation.

In 2016, adjustment for duplicates were conducted after post-stratification (within PSU calibration – see next section) because duplicated PAR information was available only for coded cases.

9.4 Post-Stratification (Within PSU Calibration)

2016 is the first year of CRSS data collection. To better understand the PJ frame used for PJ sample selection, NHTSA verified all the PJs in the PJ frame. Besides the listed cases from the sampled PJs, NHTSA also collected annual total crash counts by PAR strata from all non-sampled PJs in the PJ frame. For the non-responding PSUs in 2016, crash counts by PAR strata from non-sampled PJs were collected in 2017. The crash counts from sampled PJs and non-sampled PJs together can be used for PJ frame updates and within-PSU calibration. NHTSA plans to collect crash counts by PAR strata from non-sampled PJs periodically in the future.

Post-stratification (within-PSU calibration) is conducted as following:
If PSU-level PAR stratum total crash counts, T_{is}, is available (i.e., all sampled PJs and non-sampled PJs are cooperating), the estimated PSU-level PAR stratum total crash counts without duplicates is computed as:

$$T_{is} = \sum_{j \in s_i} w_{l|ij} L_{ijls} + \sum_{j \in ns_i} C_{ijs} - \sum_{j \in s_i} w_{j|i} w_{l|ij} D_{ijls}$$

Here L_{ijls} is the number of listed PARs, C_{ijs} is non-sampled crash counts, and D_{ijls} is the number of duplicates (i.e., cases with PARERROR=8 in the listed case file) in PSU i, PJ j, and PAR stratum s. s_i is the set of sampled PJs and ns_i is the set of non-sampled PJs in PSU i. $\sum_{j \in s_i} w_{j|i} w_{l|ij} D_{ijls}$ is the estimated number of duplicates in PSU i and stratum s. This term was not included in the formula in 2016 CRSS because duplicate PAR information was not available for all listed PARs and duplicate adjustment was conducted after the post-stratification. Then, post-stratification factor $post_{is}$ is calculated for PAR stratum s of PSU i as:

$$post_{is} = T_{is} / \sum_{j \in s_i} \sum_{k \in r_{ijs}} w_{jlk|i}^{(1)}$$

Here r_{ijs} is the set of responded PARs in PAR stratum s in PJ j of PSU i.
If T_{is} is not available (i.e., some PJs are non-cooperating), but PSU level total crash count T_i^{WD} (including all PAR strata, and with duplicates) is available, PSU level total crash count without duplicates is estimated as:

$$T_i = T_i^{WD} - \sum_{j \in r_i} w_{j|i}^{(1)} w_{l|ij} D_{ijl}.$$

Here D_{ijl} is the number of duplicates (i.e., cases with PARERROR=8 in the listed case file) for PSU i and PJ j, and r_i is the set of responding PJs in PSU i. Again, the subtraction part for duplicate PARs were not used in 2016. Post-stratification factor for PSU i is computed as:

$$post_{is} = T_i / \sum_{j \in r_i} \sum_{k \in r_{ij}} w_{jlk|i}^{(1)}$$

Here r_{ij} is the set of responded PARs in PJ j. Notice $post_{is}$ are the same for all s.
If neither PSU level PAR stratum total crash counts T_{is}, nor PSU level total crash count T_i^{WD} is available, post-stratification is not performed:

$$post_{is} = 1.$$

By multiplying post-stratification factor, post-stratified within-PSU PAR weight and case weight are computed, respectively as:

$$w_{jlk|i}^{(2)} = post_{is} * w_{jlk|i}^{(1)}$$
$$w_{ijlk}^{(2)} = w_i * w_{jlk|i}^{(2)}$$

Again, if PSU non-response adjustment was conducted in weighting stage, $w_i^{(1)}$ is used instead of w_i in the case weight calculation.

9.5 Calibration

Case weights are calibrated by benchmarking the Census resident population counts and FARS crash counts simultaneously. For each of eight primary PSU strata[4] (i.e., four Census regions by two urbanicity) g, two benchmarks are computed -- P_g: Census resident population counts and F_g: FARS crash counts. In order to implement case-level calibration, two calibration variables are defined. The first variable, resident population proxy, is:

$$X_{ijlk} = \frac{w_i P_i}{\sum_{j \in r_i} \sum_{k \in r_{ij}} w_{ijlk}^{(2)}}$$

Here w_i is PSU weight (or $w_i^{(1)}$ if PSU non-response adjustment was conducted in weighting stage), and P_i is Census resident population count in PSU i. r_i is the set of responding PJs in PSU i, r_{ij} is the set of responding PARs in PJ j of PSU i, and $w_{ijlk}^{(2)}$ is the post-stratified case weight. Notice that:

$$\sum_{i \in s_g} \sum_{j \in r_i} \sum_{k \in r_{ij}} w_{ijlk}^{(2)} X_{ijlk}$$
$$= \sum_{i \in s_g} \sum_{j \in r_i} \sum_{k \in r_{ij}} w_{ijlk}^{(2)} \frac{w_i P_i}{\sum_{j \in r_i} \sum_{k \in r_{ij}} w_{ijlk}^{(2)}} = \sum_{i \in s_g} w_i P_i$$

[4] In 2016, seven primary PSU strata were used by collapsing Midwest rural and West rural.

Here s_g is the set of sampled PSUs in the primary PSU stratum g. $\sum_{i \in s_g} w_i P_i$ is an unbiased estimate of the resident population counts of primary PSU stratum g (i.e., P_g). In this way, X_{ijlk} is defined as the resident population at case-level. Since it has the same value for all responded PARs in a PSU, it acts as a "resident population proxy."

The second variable, fatal crash identifier, is:

$$Y_{ijlk} = \begin{cases} 1 & \text{if the PAR } k \text{ is a fatal crash} \\ 0 & \text{otherwise} \end{cases}$$

PAR k is defined as a fatal crash if the corresponding imputed coded case is fatal crash.
With two case-level calibration variables and two benchmarks, calibration is simultaneously implemented by primary PSU strata using SUDAAN WTADJUST procedure. The procedure produces calibration factors $Calib_{ijlk}$ and the calibrated case weight $w_{ijlk}^{(3)}$.

10. Imputation

10.1 Item Non-Response

In the CRSS, PARs with missing pages or unreadable pages were treated as non-responding PARs (or unit non-response). Non-responding PARs were dropped from the final analysis file. PAR weights were adjusted to mitigate potential non-response bias caused by unit non-response (see Chapter 9).

For the responding PARs, data are collected from the entries in the PAR (see Appendix A for an example of a PAR form) and from the information interpreted from the crash diagrams and the police officer's written summary of the crash. During this process, some data entries (items) might be found missing and entered as "unknown" or "not reported," resulting missing values (or item non-response) in the CRSS data file.

Estimates using variables with missing values without treatment may be biased. In the 2016 and 2017 CRSS, 27 variables used for NHTSA's publications were treated for item non-response by single imputation – i.e., a single plausible value was plugged in for every missing value. Table 9 shows the variables selected for imputation from the 2016 CRSS accident, vehicle, and person files with corresponding item-missing rates. These are the same variables imputed in GES from 2010 to 2015.

As mentioned in Section 9.4, PAR weights for CRSS fatal crashes are calibrated to FARS fatal crashes using reported and imputed fatal crashes. Since PAR weights use imputed injury severity, the imputation process is completed prior to weighting.

In the CRSS, the missing values (missing entries coded as "unknowns" or "not reported") were imputed by one of the following imputation methods:

- The sequential regression multivariate imputation (SRMI) method (Raghunathan et al., 2001)
- The univariate imputation method
- Logical imputation method

The imputation process starts with some selected accident level variables using the SRMI method. If the SRMI method fails to impute all the missing values, then the univariate method is used to impute the remaining missing values. The vehicle level variable BODY_TYP is imputed solely using the univariate method.

The same process is then repeated to some selected vehicle level variables and all the person level variables: first the SRMI method, then the univariate method until there is no missing value.

Finally, the remaining accident level and vehicle level variables are imputed by the logical imputation method. Logical imputation method derives values from the observed and/or the imputed values of other variables for the missing items.

In the following sections, we briefly describe the first two imputation methods. See Herbert (2019) for more detailed information on how these methods were used to impute missing items in the CRSS.

Table 9: Imputed Variables and Item-Missing Rates for 2016 CRSS

File Name	Variable Name	SAS Label	Item Missing Rate
ACCIDENT	ALCOHOL	Alcohol Involved	14.03%
ACCIDENT	DAY_WEEK	Crash Date (Day of Week)	0.00%
ACCIDENT	HARM_EV	First Harmful Event	0.05%
ACCIDENT	HOUR	Crash Time (Hour)	0.16%
ACCIDENT	LGT_COND	Light Condition	0.66%
ACCIDENT	MINUTE	Crash Time (Minute)	0.16%
ACCIDENT	MAN_COLL	Manner of Collision	0.33%
ACCIDENT	MAX_SEV	Maximum Injury Severity	1.83%
ACCIDENT	NUM_INJ	Number of Injured	1.83%
ACCIDENT	RELJCT1	Relation to Junction – Within Interchange Area	1.98%
ACCIDENT	RELJCT2	Relation to Junction – Specific Location	0.74%
ACCIDENT	WEATHER	Atmospheric Condition	3.40%
VEHICLE	IMPACT1	Area of Impact – Initial Contact Point	2.01%
VEHICLE	BODY_TYP	Body Type	2.54%
VEHICLE	VEH_ALCH	Driver Drinking in Vehicle	9.71%
VEHICLE	HIT_RUN	Hit and Run	0.01%
VEHICLE	MAX_VSEV	Max Injury Severity	4.32%
VEHICLE	MOD_YEAR	Model Year	3.72%
VEHICLE	P_CRASH1	Pre-Event Movement	1.52%
VEHICLE	M_HARM	Most Harmful Event	0.03%
VEHICLE	NUM_INJV	Number Injured in Vehicle	4.32%
PERSON	AGE	Age	5.42%
PERSON	EJECTION	Ejection	2.08%
PERSON	INJ_SEV	Injury Severity	3.45%
PERSON	DRINKING	Police-Reported Alcohol Involvement	29.75%

File Name	Variable Name	SAS Label	Item Missing Rate
PERSON	SEAT_POS	Seating Position	1.29%
PERSON	SEX	Sex	3.41%

10.2 The Sequential Regression Multivariate Imputation

The sequential regression multivariate imputation (SRMI) method first models the variable to be imputed by other covariates in the file and then uses the fitted model to generate values to impute the missing values. The following is a brief description of the SRMI imputation process.

1) From the sample file, identify the variables without missing values: $\{x_1, x_2, \dots, x_p\}$.
2) Sort the variables to be imputed by item missing rates in an ascending order (variable with the lowest item missing rate first): $\{y_1, y_2, \dots, y_k\}$.
3) Regress y_1 on $\{x_1, x_2, \dots, x_p\}$. y_1 may be a continuous, binary, categorical, or count variable. Accordingly, a normal linear regression model, logistic regression model, generalized logit regression model, or Poisson log-linear model was used as the regression model with a flat prior distribution for the parameters. A stepwise regression algorithm automatically selects the covariates.
4) For each missing value in y_1, draw a single value from the posterior distribution specified by the fitted model for y_1 to impute the missing value until there is no missing value in y_1.
5) For i =2 to k, regress y_i on $\{x_1, x_2, \dots, x_p, y_1, \dots, y_{i-1}\}$ assuming a flat prior for the regression parameters. For each missing value in y_i, draw a single value from the posterior distribution specified by the fitted model to impute the missing value.
6) For i =1 to k, regress imputed y_i on $x's$ and all other $y's$ except y_i itself: $\{x_1, x_2, \dots, x_p, y_1, \dots, y_{i-1}, y_{i+1}, \dots y_k\}$. Overwrite each value of y_i previously imputed by redrawing a single value from the posterior distribution specified by the newly fitted model for y_i.
7) Repeat step 6) until either stability in the imputed values or the predetermined number of iterations are reached.

SRMI method was implemented in the free SAS callable software – IVEware[5] . IVEware was used in the CRSS imputation. In addition, IVEware allows the imputation of a variable to:

- Be performed to a subset of observations that meet certainty condition.
- Restrict the imputed values to certain range of values.

IVEware imputation may terminate prematurely upon reaching the predetermined number of iterations or other convergence criteria before all missing values are imputed. Additionally, some

[5] Raghunathan, T., Solenberger, P., Berglund, P., & van Hoewyk, J. (2016). IVEware: Imputaion and Variance Estimation Software (version 0.3), www.src.isr.umich.edu/software/

imputed values may not conform to the permissible values of the data element. For example, "Sleet" may be assigned to a missing value of variable WEATHER for a crash happened in June in the South. An editing and consistency checking process is performed after IVEware imputation. When the SRMI method fails or consistency checks fail, the univariate imputation is used to impute the remaining missing values.

10.3 The Univariate Imputation

The univariate imputation method is also known as the simple random sampling with-replacement imputation. The following is a description of the univariate imputation method:

1. First, the values of the variable to be imputed, y, are grouped into two groups: the non-missing values (observed values or values already imputed by the SMRI method), and the missing values (non-observed values that have not been imputed by SMRI method).
2. For each missing value, randomly select one value from the non-missing group with replacement and assign the selected non-missing value to the missing value.

This univariate imputation method preserves the non-missing values' sample distribution but ignores any correlation with other variables. Variable BODY_TYP is the only variable only imputed by the univariate imputation method.

After the univariate imputation, data inconsistencies may still occur. These anomalies are reviewed and corrected.

The imputed maximum injury severity variables, MAXSEV_IM in the accident file and MAXVSEV_IM in the vehicle file, were both derived from INJSEV_IM in the person file. The imputed police reported alcohol involvement variables, ALCHL_IM in the accident file and V_ALCH_IM in the vehicle file, were both derived from PERALCH_IM in the person file. The imputed number of injured variables, NO_INJ_IM in the accident file and NUMINJ_IM in the vehicle file were both derived based on INJSEV_IM.

The imputed variables were named by their original names plus suffix "_IM", e.g., if the original variable is AGE, then the imputed variable is AGE_IM. Table 10 lists the names and labels of the imputed variables.

Table 10: Names and Labels for the Imputed Variables

File Name	Original Variable Name	Imputed Variable Name	SAS Label
ACCIDENT	ALCOHOL	ALCHL_IM	Imputed Drinking in Crash
ACCIDENT	DAY_WEEK	WKDY_IM	Imputed Day of the Week
ACCIDENT	HARM_EV	EVENT1_IM	Imputed First Harmful Event
ACCIDENT	HOUR	HOUR_IM	Imputed Hour
ACCIDENT	LGT_COND	LGTCON_IM	Imputed Lgt Condition
ACCIDENT	MINUTE	MINUTE_IM	Imputed Minute
ACCIDENT	MAN_COLL	MANCOL_IM	Imputed Manner of Collision
ACCIDENT	MAX_SEV	MAXSEV_IM	Imputed Maximum Injury Severity
ACCIDENT	NUM_INJ	NO_INJ_IM	Imputed Number Injured in Crash
ACCIDENT	RELJCT1	RELJCT1_IM	Relation to Junction – Within Interchange Area
ACCIDENT	RELJCT2	RELJCT2_IM	Imputed Relation to Junction - Junction
ACCIDENT	WEATHER	WEATHR_IM	Imputed Weather Condition
VEHICLE	IMPACT1	IMPACT1_IM	Imputed Area of Impact-Initial
VEHICLE	BODY_TYP	BDYTYP_IM	Imputed Body Type
VEHICLE	VEH_ALCH	V_ALCH_IM	Imputed Driver Drinking in Vehicle
VEHICLE	HIT_RUN	HITRUN_IM	Imputed Hit and Run
VEHICLE	MAX_VSEV	MXVSEV_IM	Imputed Maximum Injury in Vehicle
VEHICLE	MOD_YEAR	MDLYR_IM	Imputed Model Year
VEHICLE	P_CRASH1	PCRASH1_IM	Imputed Vehicle P_Crash1
VEHICLE	M_HARM	VEVENT_IM	Imputed Most Harmful Event
VEHICLE	NUM_INJV	NUMINJ_IM	Imputed Number Injured in Vehicle
PERSON	AGE	AGE_IM	Imputed Age
PERSON	EJECTION	EJECT_IM	Imputed Ejection
PERSON	INJ_SEV	INJSEV_IM	Imputed Injury Severity
PERSON	DRINKING	PERALCH_IM	Imputed Police Rep. Alcohol Inv.
PERSON	SEAT_POS	SEAT_IM	Imputed Seating Position
PERSON	SEX	SEX_IM	Imputed Sex

References

Fleming, C. (2010, May). Sampling and Estimation Methodologies of CDS, Technical Report, DOT HS 811-327, Washington, DC: National Highway Traffic Safety Administration. Available at https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811327

Herbert, G. C. (in press). Crash Report Sampling System: Imputation. Washington, DC: National Highway Traffic Safety Administration.

H. Rept. 111-564, Departments of Transportation, and Housing and Urban Development, and Related Agencies Appropriations Bill, 2011.

Kott, P. (2012). An introduction to calibration weighting for establishment surveys. Rockville, MD: RTI International. Available at https://ww2.amstat.org/meetings/ices/2012/papers/302286.pdf.

National Highway Traffic Safety Administration. (2011, August). Report to Congress: NHTSA's NASS data needs (Report No. DOT HS 811 889). Washington, DC: National Highway Traffic Safety Administration.

National Highway Traffic Safety Administration. (2018, March). Crash Report Sampling System analytical user's manual 2016. (Report No. DOT HS 812 510). Washington, DC: National Highway Traffic Safety Administration. Available at https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812510

National Highway Traffic Safety Administration (2014). Not In Traffic Surveillance (NITS) Nontraffic Crash Injuries and Fatalities 2008-2011 Analytical User's Manual (Report No. DOT HS 811 805). Washington, DC: National Highway Traffic Safety Administration. Available at https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811805

Noh, E. Y. (2014). Estimation of cost components (Unpublished technical report). Washington, DC: National Highway Traffic Safety Administration.

Noh, E. Y., & Zhang, F. (2016). Optimum sample allocations of Crash Investigation Sampling System and Crash Reporting Sampling System (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.

Noh, E. Y., Zhang, F., Subramanian, R., & Chen, C.-L. (2016). CRSS PAR selection algorithm and the IT aspects (Unpublished technical report). Washington, DC: National Highway Traffic Safety Administration.

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

Raghumathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputation missing values using a sequence of regression models. *Survey Methodology, Vol 27*, No. 1, pp. 85-95.

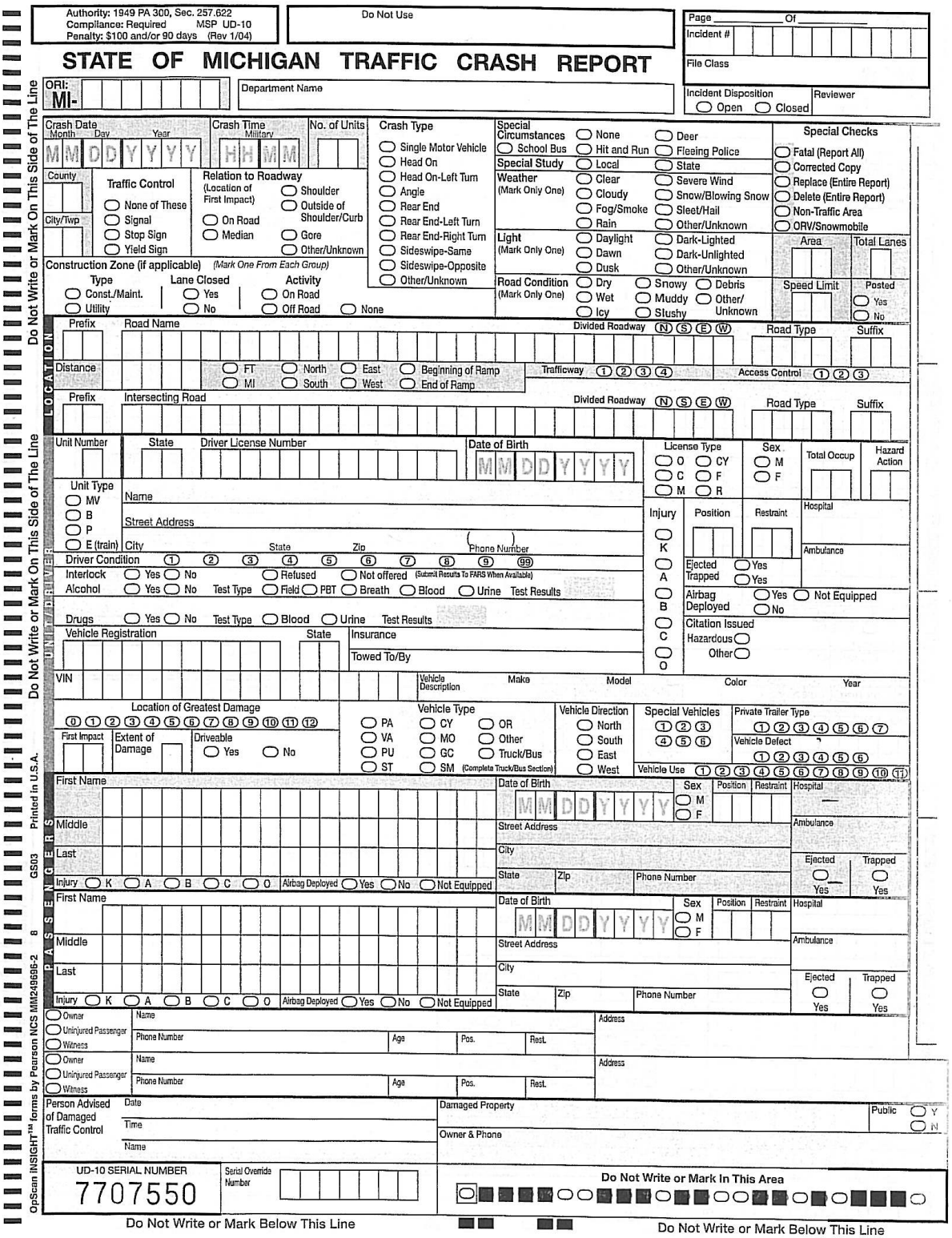
Sen. Rept. 111-230, Transportation, and Housing and Urban Development, and Related Agencies Appropriations Bill, 2011.

Shelton, T. S. (1991). National Accident Sampling System General Estimates System, Technical Note, 1988 to 1990 (Report No. DOT HS 807796). Washington, DC: National Highway Traffic Safety Administration.

Zhang, F., & Chen, C.-L. (2013, July). 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., Subramanian, R. Chen, C.-L., & Noh, E. Y. (2019, April). Crash Report Sampling System: Design Overview, Analytic Guidance, and FAQs (Report No. DOT HS 812 688). Washington, DC: National Highway Traffic Safety Administration. Available at https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812688

APPENDIX A. An Example of a PAR



APPENDIX B. Excluded/Included Alaska and Hawaii Counties

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

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

APPENDIX C. CRSS PSU Strata for the Five Scenarios

Scenario	Sample Order	Scenario-1	Scenario-2	Scenario-3	Scenario-4	Scenario-5
#Strata		51	38	26	12	8
Strata	17	1-01	1-01	1-01	1-01	Northeast Urban
	27					
	24	1-02	1-02	1-02		
	83					
	50	1-03				
	93					
	16	1-04	1-03	1-03	1-02	
	52					
	68	1-05	1-04			
	71					
	33	1-06	1-05			
	73					
	6	1-07	1-06	1-04		
	84					
	40	1-08				
	100					
	13	2-01	2-01	2-01	2-01	Northeast Rural
	29					
	4	2-02	2-02	2-02		
	38					
	14	3-01	3-01	3-01	3-01	Midwest Urban
	89					
	45	3-02				
	91					
	48	3-03	3-02	3-02		
	67					
	41	3-04	3-03			
	75					
	7	3-05	3-04	3-03		
	90					
	65	3-06				
	78					
	44	3-07	3-05			
	63					
	25	4-01	4-01	4-01	4-01	Midwest Rural
	37					
	59	4-02	4-02	4-02		
	86					
	11	4-03				
	98					
	8	4-04	4-03			
	69					
	12	5-01	5-01	5-01	5-01	SouthUrban
	74					
	39	5-02	5-02			
	57					
	22	5-03	5-03	5-02		
	82					
	30	5-04				
	77					

CRSS PSU Strata for the Five Scenarios (continued)

Scenario	Sample Order	Scenario-1	Scenario-2	Scenario-3	Scenario-4	Scenario-5
#Strata		**51**	**38**	**26**	**12**	**8**
Strata	42	5-05	5-04	5-03	5-02	SouthUrban
	79					
	21	5-06				
	88					
	20	5-07	5-05	5-04		
	61					
	35	5-08	5-06			
	64					
	58	5-09	5-07	5-05	5-03	
	66					
	15	5-10	5-08			
	76					
	43	5-11				
	97					
	23	5-12	5-09	5-06		
	60					
	51	5-13	5-10			
	95					
	72	5-14				
	101					
	36	6-01	6-01	6-01	6-01	South Rural
	96					
	1	6-02				
	99					
	49	6-03	6-02	6-02		
	56					
	10	6-04	6-03			
	55					
	26	6-05	6-04	6-03		
	92					
	46	6-06				
	81					
	32	7-00 (LA)	7-00 (LA)	7-00 (LA)	7-01	West Urban
	62	7-01	7-01	7-01		
	80					
	9	7-02				
	94					
	19	7-03	7-02			
	54					
	2	7-04	7-03	7-02	7-02	
	85					
	47	7-05				
	87					
	28	7-06	7-04	7-03		
	70					
	18	7-07	7-05			
	53					
	3	8-01	8-01	8-01	8-01	West Rural
	34					
	5	8-02	8-02	8-02		
	31					

DOT HS 812 706
May 2019

U.S. Department
of Transportation

**National Highway
Traffic Safety
Administration**

14186-051019-v5