NHTSA's Data Modernization Project

Chou-Lin Chen, Rajesh Subramanian, Fan Zhang, Eun Young Noh

National Highway Traffic Safety Administration, Department of Transportation
1200 New Jersey Ave SE, W55-334, Washington, DC 20590

Proceedings of the 2015 Federal Committee on Statistical Methodology (FCSM) Research Conference

1. Introduction

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

NASS is comprised of two nested systems – the General Estimates System (GES) and the Crashworthiness Data System (CDS). Both systems are operated by the NHTSA's National Center for Statistics and Analysis (NCSA) and provide national probability samples of crashes.

GES is a survey of police-reported traffic accident reports (PARs). It collects general information of the traffic crashes from PARs only. GES data are used to provide a general picture of the crash population and trends, identify highway safety problem areas and assess the size of the problem, provide a basis for regulatory and consumer information initiatives, and form the basis for cost and benefit analyses of vehicle regulations. See Shelton (1991) for a detailed discussion of GES sampling and weighting procedures.

While the GES 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 crashes. In addition to the information collected from PARs, CDS also collects more detailed data about the crash, vehicles and occupants through interviews, medical records, vehicle inspections, and scene inspections.

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

Developed in the 1970's and redesigned in the 1980's, NASS's primary data collection sites, also called primary sampling units (PSUs), and the secondary data collection sites, the police jurisdictions (PJs), have not changed for the past 30+ years. During this period of time, the underlying NASS sampling frame has seen many changes, primarily at the PSU and PJ levels. For example, the number and nature of crashes across PSUs, population growth and mobility shift, PJ frame (opening, closing, merging, crash distribution changes among PJs), improvements in vehicle and highway safety.

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

Furthermore, the scope of traffic safety studies has also been expanding with emerging traffic safety issues. Because of the limited CDS sample size, it does not provide enough sampled cases to support detailed domain analysis. While substantial reductions in passenger vehicle fatalities have been realized, data on emerging traffic safety areas such as crashes involving large trucks, motorcycles, and pedestrians are not collected in the current CDS.

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

The United States Congress supported the effort to examine the deficiencies in NASS and to plan for a modernized and comprehensive data system. In "Moving Ahead for Progress in the 21st Century" (MAP-21), Congress instructed the Secretary of the Department of Transportation to submit a report regarding the quality of data collected through NASS. In addition, congress also instructed the Administrator of NHTSA to conduct a comprehensive review of the data element 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.

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

The following sections will discuss the scope, the PSU frames, the sample selection, and the sample allocation of the two surveys. For more detailed information about these two surveys, see Zhang, Noh, Subramanian, and Chen (2015).

2. The Scope of CRSS and CISS

Crash data needs and the focus of traffic safety research have significantly changed since the establishment of NASS in the 1970's. It is critical to identify data user's needs to properly define the scope of CRSS and CISS. 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 relevant to the traffic safety research community. To this end, NHTSA conducted two studies to evaluate internal and public data needs.

2.1 The Data Needs of the U.S. Department of Transportation

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

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

2.2 The Data Needs of the Public

In order to solicit inputs from the broadest possible group of stakeholders, NHTSA published a notice in the Federal Register announcing the survey modernization effort on June 21, 2012 (see NHTSA-2012-0084 at www.regulations.gov). 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 U.S. and abroad. NHTSA also sent the Federal Register Notice to more than 500 interested parties by letters and email. These public stakeholders include:

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

More than 20 organizations and individuals submitted over 300 comments to NHTSA. The comments and suggestions received from data users outside of the NHTSA reflected similar needs to users within NHTSA. Comments regarding the importance and relevance of the various data systems were universally positive. However, data users wanted to see NASS updated to it remains relevant. The comments covered a wide range of topics including:

- 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 alone to increase the number of cases of most interest to data users, especially under constrained funding scenarios.

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.

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

2.3 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.

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

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

NHTSA's internal and public data needs 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, sport utility vehicle (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 persons 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 persons by person type (person type, possibly collapsed to drivers, occupants, non-occupants) 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, large truck crashes
- Number of tow away crashes
- Number of serious injuries in passenger vehicle, motorcycle, pedestrian, large truck crashes

2.4 CISS Analysis Objectives

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

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

- Assessment of the overall state of traffic safety, and identification of existing and potential traffic safety problems.
- Obtaining detailed data on the crash performance of passenger cars, light trucks, vans, and utility vehicles.
- Evaluation of vehicle safety systems and designs.
- Increasing knowledge of crash related injuries, including injury mechanisms.
- Assessment of the effectiveness of motor vehicle and traffic safety program standards.
- Identifying emerging issues in vehicle safety.

2.5 CRSS Target Population and Analysis Domains

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

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 these information are 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% in analysis domain 2 means 9% 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 in which: • Involves a killed or injured (includes injury severity unknown) non-motorist	9%	119,579	2.2%
3	Crashes not in Stratum 1 or 2 in which: • Involves 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 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%	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%

CRSS Analysis Domain	Analysis Domain Description	Target Percent of Sample Allocation	Estimated Population (GES 2011)	Population Percent
9.	Crashes not in Stratum 1-8 in which: • involved at least one late model year passenger vehicle, AND • No person in the crash is killed or injured	22%	1,511,371	28.4%
10	Crashes not in Stratum 1-9: * This includes mostly Property Damage Only Property Damage Only (PDO) crashes involving a non-motorist, motorcycle, moped, 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 and older

2.6 CISS Target Population and Analysis Domains

From the assessment of the CISS analytic objectives, NHTSA has determined the target population for CISS shall be all police-reported motor vehicle crashes on a traffic way involving a passenger vehicle and in which a passenger vehicle is towed from the scene for any reason. This is slightly different from the CDS target population, which required that the vehicle be towed due to damage. This change alleviates issues related to determining the eligibility of crashes in the field.

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

In Table 2 the "Target Percent of Sample Allocation" column specifies the desired distribution of the sampled cases – for example, '5%' in domain 1 means that 5 percent of the sampled cases should be from domain 1. The "Estimated Population" column is the population count for each analysis domain estimated from current NASS. The "Population Percent" column is the population distribution of each analysis domain estimated from current NASS. If the "Population Percent" is lower than "Target Percent of Sampling Allocation", then the corresponding analysis domain is oversampled relative to its incidence.

Table 2: CISS Analysis Domains, Descriptions, Allocation and Population Sizes

CISS Analysis Domains	Description	Target Percent of Sample Allocation	Estimated Population	Population Percent
1	At least one occupant of towed passenger vehicle is killed	5%	9,576	0.51%
2	Crashes not in Stratum 1 involving: • A recent model year passenger vehicle in which at least one occupant is incapacitated	10%	17,304	0.93%

3	Crashes not in Stratum 1 or 2 involving: • A recent model year passenger vehicle in which at least one occupant is non-incapacitated, possibly injured or injured but severity is unknown.	20%	162,037	8.71%
4	Crashes not in Stratum 1-3 involving: • A recent model year passenger vehicle in which all occupants are not injured	15%	325,332	17.48%
5	Crashes not in Stratum 1-4 involving: • A mid-model year passenger vehicle in which at least one occupant is incapacitated	6%	23,739	1.28%
6	Crashes not in Stratum 1-5 involving: • A mid-model year passenger vehicle in which at least one occupant is non-incapacitated, possibly injured or injured but severity is unknown	12%	210,407	11.31%
7	Crashes not in Stratum 1-6 involving: • A mid-model year passenger vehicle in which all occupants are not injured	10%	418,702	22.51%
8	Crashes not in Stratum 1-7 involving: • A older model year passenger vehicle in which at least one occupant is incapacitated	6%	28,690	1.54%
9	Crashes not in Stratum 1-8 involving: • A older model year passenger vehicle in which at least one occupant is non-incapacitated, possibly injured or injured but severity is unknown.	10%	220,815	11.87%
10	Crashes not in Stratum 1-9 involving: • A older model year passenger vehicle in which all occupants are not injured	6%	443,151	23.83%
Total		100%	1,859,752	100%

Source: Estimated from 2011 CDS data.

Note: This table uses the following definitions:

- Recent MY (or LMY): vehicles that are <= 4 years old.
- Mid MY: 5-9 year old vehicles
- Older MY: vehicles that are 10 years old or older

2.7 The Relationship between the CRSS and the CISS Samples

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

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

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

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

- CISS and CRSS have different target populations: CISS targets towed passenger vehicles while CRSS targets all police-reported crashes and the vehicles involved in them.
- CISS and CRSS have different operational requirements: CISS requires follow-on and potential on-scene investigation and must therefore respond quickly to crashes. To ensure a rapid response the PSUs must not exceed a certain geographic size. The objective of 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, independently tailored PSU formation, stratification, PSU measure of size definitions, and sample selection can produce more efficient samples for both systems. To optimize both CISS and CRSS, NHTSA decided to design CISS independently from CRSS.

3. PSU Sampling Frames

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 crashes requires access to all crashes in the nation, which is cost prohibitive if not impossible. Instead, the country is partitioned into smaller areas called primary sampling units (PSUs) - a county or a group of counties for both CRSS and CISS - so a probability sample of PSUs can be selected and local crashes can be further selected. CRSS and CISS PSU frames were formed independently and tailored to meet operational and estimation needs of each survey.

3.1 CRSS PSU Frame

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 distance to collect the PARs. Second, since regional boundaries and their urbanicity, as specified by the U.S. Census Bureau were found to be effective stratification variables in GES, PSUs in CRSS were also 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, outlying counties in Alaska and Hawaii that do not contain a city were excluded from the PSU frame because they are remote and have few crashes.

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.

3.2 CISS PSU Frame

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

Census Region and urbanicity were identified as effective PSU stratification variables. In the CISS, urbanicity was defined by the census metropolitan statistical area (MSA). Any MSA with 50,000 or more people was considered urban and all others were considers rural. PSUs were formed to respect region and urbanicity for effective PSU stratification. However, PSUs were allowed to cross state lines.

As shown in Table 2, the analysis domain 1 has the lowest population percent (i.e., 0.5% of all eligible crashes in the population). This makes the domain 1 PARs the rarest cases. For domain estimation, it is desirable that the rare cases are selected from as many PSUs as possible. The target sample allocation (desired portion of all sampled crashes) for domain 1 is 5 percent. Therefore, PSUs were formed so that 5 percent of the CISS sample could consist of crashes having at least one fatality in a towed passenger vehicle. We assumed that each PSU will employ at least one technician, and one technician collects about 100 cases per year. Then, PSUs were formed with 90 percent probability of yielding at least 5 fatal crashes involving a passenger vehicle each year.

Outlying counties in Alaska and Hawaii that do not containing a city were excluded because they are remote and have few crashes.

3.3 CRSS PSU Measure of Size

The measure of size (MOS) is a quantity used to assign the selection probability to the PSUs for the unequal selection probability sampling. 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 crash counts, and Census population.

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 all selected with certainty 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.

3.4 CISS PSU Measure of Size

One of the main analysis interest areas of CISS focuses on severe crashes and recent model year passenger vehicles. Since these crashes are rare in the population it is necessary to oversample them. Therefore, PSUs with more high interest crashes should be given a larger MOS so these PSUs are more likely to be selected and more high interest crashes can be selected.

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

with the variables used to estimate domain population counts. The descriptions and source of these variables are listed in Table 4.

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

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

- n = the desired total sample size of PARs in the high interest domains – 1, 2, 3, 4, 5, 6, 8.
- n_{++s} = the desired sample size of PARs in analysis domain s
- N_{++s} = the estimated population counts in analysis domain s
- N_{i+s} = the estimated population counts in analysis domain s and PSU i.

In the formula, n_{++s}/n represent the rescaled target sample allocation. The composite MOS is computed as the weighted sum of scaled population counts (N_{i+s}/N_{++s}) over the high interest domains where the weight is the target sample allocation.

3.5 CRSS PSU Frame Stratification

Census regions were used for the CRSS PSU stratification to produce geographically more balanced and representative PSU sample. Crosswalk between Census regions and states can be found at: http://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 for the CRSS PSU stratification to produce demographically balanced and representative PSU sample. Urbanicity 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

[2] Highway Performance Monitoring System

- The sum of the 2008 and 2009 NHTSA State Data System (SDS) "A" injury crashes (which includes imputed values for non-SDS reporting states)
- The sum of the 2008 and 2009 SDS "B" 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 PSUs. 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 3 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 3 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 1-01 – 1-08). In the Northeast-Urban, the secondary stratum 1-01consisted 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.

Table 3: 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	1800.660	0.000					358.504	0.000	5	222,923.0
1	1-02	4064.065	1800.660					358.504	0.000	5	183,528.6
1	1-03	7159.044	4064.065					358.504	0.000	8	197,557.2
1	1-04	5791.034	0.000	0.028	0.000	153756.114	0.000	2175.024	358.504	6	176,867.6
1	1-05	8040.031	5791.034	0.028	0.000	153756.114	0.000	2175.024	358.504	7	204,412.6
1	1-06			0.028	0.000	249917.616	153756.114	2175.024	358.504	7	207,205.0
1	1-07			0.028	0.000	591240.550	249917.616	2175.024	358.504	7	200,876.3
1	1-08			0.039	0.028			2175.024	358.504	11	198,297.2
2	2-01					236700.660	0.000			22	138,907.1
2	2-02					1027525.695	236700.660			22	147,852.3
3	3-01	4134.622	0.000			45708.732	0.000			3	189,109.0
3	3-02	7465.060	4134.622			45708.732	0.000			8	186,036.2
3	3-03	9897.834	7465.060			45708.732	0.000			10	185,605.7
3	3-04					102553.858	45708.732			11	198,246.3
3	3-05	4443.529	0.000			339758.109	102553.858			13	183,349.1
3	3-06	6002.758	4443.529			339758.109	102553.858			11	189,401.7
3	3-07	11617.975	6002.758			339758.109	102553.858			10	183,563.0
4	4-01					66170.891	0.000	4344.584	0.000	28	191,481.8
4	4-02	6045.032	0.000			565024.725	66170.891	4344.584	0.000	27	190,433.5
4	4-03	11623.151	6045.032			565024.725	66170.891	4344.584	0.000	25	187,744.9
4	4-04							17641.397	4344.584	30	189,375.7
5	5-01	3619.866	0.000	0.048	0.000	125590.275	0.000			5	188,584.1
5	5-02	4529.728	3619.866	0.048	0.000	125590.275	0.000			8	194,116.6
5	5-03	4951.021	4529.728	0.048	0.000	125590.275	0.000			6	159,867.8
5	5-04	5016.203	4951.021	0.048	0.000	125590.275	0.000			3	206,325.1
5	5-05	5277.180	5016.203	0.048	0.000	125590.275	0.000			5	223,731.9
5	5-06	5745.576	5277.180	0.048	0.000	125590.275	0.000			6	149,244.5
5	5-07	6399.201	5745.576	0.048	0.000	125590.275	0.000			5	204,319.4
5	5-08	12826.284	6399.201	0.048	0.000	125590.275	0.000			8	205,760.3
5	5-09	5641.161	0.000	0.048	0.000	210430.008	125590.275			6	191,122.2

[3] Highway Loss Data Institute

5	5-10	8347.640	5641.161	0.048	0.000	210430.008	125590.275			7	195,787.2
5	5-11	13892.020	8347.640	0.048	0.000	210430.008	125590.275			10	173,150.4
5	5-12			0.048	0.000	358684.226	210430.008			8	198,717.9
5	5-13			0.048	0.000	877545.577	358684.226			13	192,291.6
5	5-14			0.085	0.048					17	181,098.3
6	6-01					49853.675	0.000			35	211,282.4
6	6-02	6353.419	0.000			162415.067	49853.675			34	209,739.0
6	6-03	14414.922	6353.419			162415.067	49853.675			35	213,326.3
6	6-04					250189.692	162415.067			33	213,536.6
6	6-05	5693.121	0.000			1156242.173	250189.692			35	208,654.9
6	6-06	16270.533	5693.121			1156242.173	250189.692			35	211,751.8
7	7-00									1	286,050.4
7	7-01	6477.249	0.000	0.027	0.000	104521.554	0.000			7	194,314.4
7	7-02	6920.874	6477.249	0.027	0.000	104521.554	0.000			4	234,421.9
7	7-03	7860.805	6920.874	0.027	0.000	104521.554	0.000			5	169,858.5
7	7-04	5137.293	0.000	0.027	0.000	249358.333	104521.554			3	193,052.2
7	7-05	8069.657	5137.293	0.027	0.000	249358.333	104521.554			10	218,727.9
7	7-06			0.048	0.027	92715.811	0.000			9	177,453.5
7	7-07			0.048	0.027	186409.133	92715.811			7	216,069.5
8	8-01							3938.460	0.000	30	206,693.7
8	8-02							18292.498	3938.460	41	205,284.9
Total	**51**									**707**	**9,987,109**

3.6 CISS PSU Frame Stratification

Census regions were used as PSU stratification variable and produced more geographically balanced and representative PSU sample. Census regions include:

- Northeast
- West
- South
- Midwest

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

- Urban - include at least one Census Metropolitan Statistical Area
- Rural - otherwise

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

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

Candidate stratification variables for further stratification include:

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

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

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

Table 4 describes how the secondary PSU strata were formed within each primary PSU stratum. PSU 7-00 has an extraordinary large MOS and was thereby treated as a stratum by itself. Twenty four secondary strata were formed from the eight primary strata. For example, primary stratum 1 is further partitioned into three secondary strata by the three categories of variable ROAD_TYPE_RATE: 0-225, 225-747, 747-7233.

Table 4: CISS PSU Strata and PSU Population Counts

Primary Strata	Secondary Strata	VMT_RATE_IMP		TOT_CRASH_RATE		ROAD_TYPE_RATE		Number of PSUs	Total Strata MOS
		Lower	Upper	Lower	Upper	Lower	Upper		
1	1-01					0	225	13	0.048414
1	1-02					225	747	21	0.052007
1	1-03					747	7233	61	0.049635
2	2-01	0	5871					30	0.010999
2	2-02	5871	30228					33	0.011220
3	3-01	0	5619			0	490	10	0.036841
3	3-02	0	5619			490	5817	51	0.036230
3	3-03	5619	19240	0.000	0.023			18	0.036300
3	3-04	5619	19240	0.023	0.096			58	0.036337
4	4-01	0	6047					130	0.036853
4	4-02	6047	27671					188	0.037012
5	5-01			0.000	0.024	0	398	12	0.046438
5	5-02			0.000	0.024	398	1530	24	0.044972
5	5-03			0.024	0.026			21	0.053119
5	5-04			0.026	0.032			39	0.046482
5	5-05			0.032	0.042			60	0.049574
5	5-06			0.042	0.138			155	0.048216
6	6-01	0	5774					242	0.064772
6	6-02	5774	42131					372	0.064739
7	7-00							1	0.024913
7	7-01	0	5368					22	0.041221
7	7-02	5368	8298					24	0.042615
7	7-03	8298	15685					22	0.041226
8	8-01			0.000	0.052			46	0.020016
8	8-02			0.052	0.212			131	0.019849
Total	**25**							**1784**	**1**

4. PSU Sample Selection

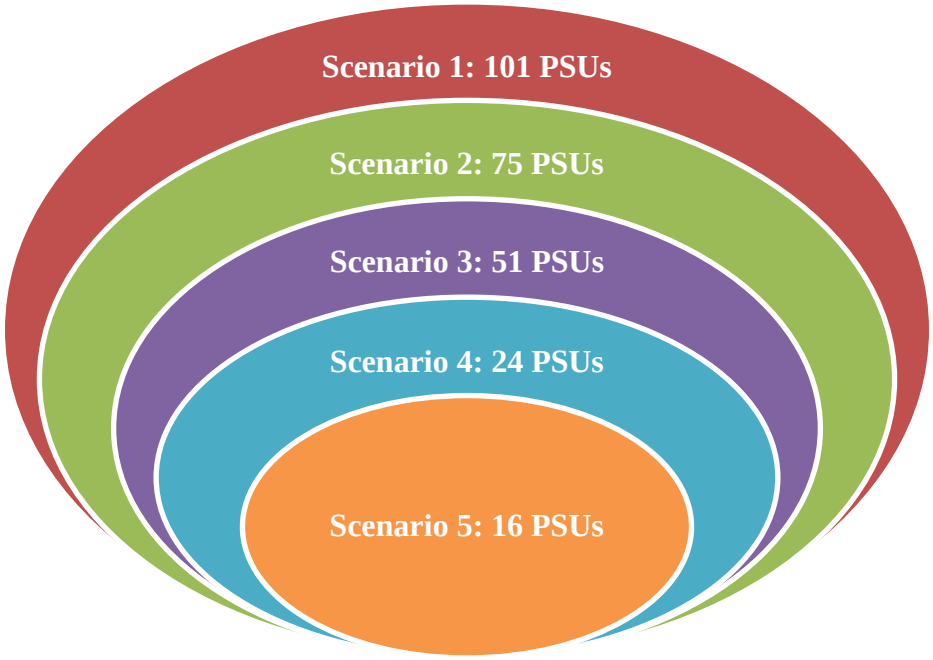
Both CRSS and CISS have a stratified three stage sample design: primary sampling unit sample selection, police jurisdiction sample selection and police accident report sample selection.

A major challenge 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 data collection sites. Changing the data collection sites is both costly

and time-consuming due to the training of new technicians and establishing cooperation with local police departments, etc. Therefore, once the CRSS and CISS PSU samples are selected and established, it is cost efficient to keep using it for as long as possible.

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 both CRSS and CISS PSU samples by selecting a sequence of nested PSU samples. This is different from GES and CDS 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 easily track the selection probabilities and minimizes changes to the PSU sample.

Figure 1: Nested PSU Samples



4.1 CRSS PSU Sample Selection

For CRSS, 5 PSU samples were selected under the 5 scenarios of number of PSU strata and PSU sample sizes using multiphase sampling method. Table 5 summarizes the CRSS PSU sample scenarios.

The 101 scenario-1 PSUs were selected by stratified PPS sampling. For scenario-2, 13 scenario-1 strata were collapsed with other 13 strata to form 37 scenario-2 strata. The 101 scenario-1 PSUs were used as the sampling frame for scenario-2 sample selection. Each of the 101 scenario-1 PSUs was assigned a new MOS – its scenario-1 stratum total MOS. Then 2 PSUs were selected from each of the 37 scenario-2 stratum using PPS sampling. With one certainty PSU, this resulted in total 75 PSUs for the scenario-2 PSU sample. Typically, the resulting PSU selection probabilities for all phases are PPS.

Other scenario samples were selected in a similar way. PSU sample of size between the scenarios were also selected for CRSS. The PSUs that were not selected for next scenario were randomly sorted into a sequence. This sequence was used to select PSUs between the scenarios. As a result, the 101 PSUs in the scenario 1 sample were arranged in a sequence so that any number of PSUs can be used for data collection.

Table 5 – 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

4.2 CISS PSU Sample Selection

CISS PSU sample was selected in similar way. Table 6 lists the sample sizes for each scenario. CISS PSU sample started with scenario-1 with 49 PSUs, NHTSA later decided to expand the sample size to 96. As the result, any number of PSUs between 16 and 96 can be used for CISS data collection.

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

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

5. PJ Sample Selection

PARs are filled out by police officers and reported to the State through a police jurisdiction. For the CRSS and CISS, PARs can be obtained from PJs either by visiting the PJs or by electronic transmission. In this way, PJs are viewed as natural clusters of PARs. Therefore CRSS and CISS secondary sampling units (SSU) are police jurisdictions (PJs) that produce PARs for the crashes occurring in the PSUs. In order to construct a sampling frame for the PJs, NHTSA collected SSU (PJ) sampling frame 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 for CRSS and the 74 PSUs of the scenario-0.5 for CISS. Among the PJ information, the following 6 types of crash counts (PJ frame crash counts) were collected to compute SSU MOS:

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

For both CRSS and CISS, the PJ frame was created for the sampled PSUs. If multiple PJs in a PSU have the same name and address, which are mostly state police, these PJs were combined to 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.

5.1 PJ Measure of Size

Similar to the PSU MOS definition, it is sensible to assign larger selection probability to PJs with desirable crash composition. For CRSS, 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 the 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_{ijk} = the estimated population number of crashes in PAR stratum s, PJ j and PSU i

For CISS, the six PJ frame crash counts are used to estimate the ten PAR domain counts in Table 2 for each PJ in the selected PSUs. First, the six PJ frame crash counts were split to estimate the nine CRSS PAR strata counts for each PJ. Then the ten CISS PAR domain counts were estimated from the nine CRSS PAR strata counts using a regression model. The SSU MOS is then defined as follows:

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

- n = the desired total sample size (of PARs)
- n_{++s} = the desired sample size of crashes in analysis domain s
- N_{++s} = the estimated population of crashes in PAR domain s
- N_{ijs} = the estimated population number of crashes in domain s, PJ j and PSU i

5.2 PJ Stratification

CRSS and CISS used similar way to stratify PJ frame. PJ MOS varies dramatically within the selected PSUs. To reduce the sampling variance, the PJ frame within each selected PSU might be stratified depending on the PJ frame size. If there were no more than 4 PJs in a selected PSU, then all PJs were assigned to the certainty stratum. If there were more than 4 PJs in a selected PSU, the following condition was first used to identify certainty PJs:

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

Here MOS_{ij} is the 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. All PJs identified in this way were assigned to the certainty stratum.

If there were more than 4 but less than 9 PJs in the selected PSU, no further PJ stratum was formed. Therefore, these PSUs have two PJ strata: certainty stratum and non-certainty stratum.

If there are 9 or more PJs in the selected PSU, then three PJ strata were formed. First, the following condition was used to identify more certainty PJs:

$$\frac{2MOS_{ij}}{\sum_j MOS_{ij}} \geq 0.7$$

Here the summation was over all non-certainty PJs. PJs identified by this condition were assigned to the certainty stratum also. The remaining non-certainty PJs were then sorted by their PJ MOS within each selected PSU. The

larger 50% of PJs were assigned to the large MOS stratum and the other 50% of PJs were assigned to the small MOS stratum. Therefore, for the PSUs with 9 or more PJs, total three PJ strata were formed: the certainty stratum, the large MOS stratum, and the small MOS stratum.

5.3 PJ Sample Selection

One of the major challenges of the PJ sample selection is changes to the PJ frame. Unlike PSUs, PJs are relatively unstable as new PJs may emerge or existing PJs may split, merge or close down. The PJ MOS is determined by crash counts that are subject to variation every year and hence the PJ stratum may also change. Also, setting up cooperation with the PJs is time consuming. In addition, there is a chance that PJs may refuse to cooperate in this effort.

Pareto sampling (see Rosén 1997) was used to select the PJ sample. Pareto sampling method produces an approximate PPS sample, handle the frame changes and minimize the changes to the existing sample at the same time.

Pareto sampling method was 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 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 PJ MOS. The identified certainty PJs are set aside. And this process is repeated to the remaining PJs based on 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, let:

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

  be the relative MOS after certainty PJs are removed and calculate transformed random numbers:

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

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

In Pareto sampling, once a permanent random number is assigned to a PJ it will never be changed. Therefore, unless the PJ MOS changes, the transformed random number: $\frac{r_{ij}(1-p_{ij})}{p_{ij}(1-r_{ij})}$ will not change either. If an existing PJ is closed, the corresponding transformed random number is dropped from the sorted list. If a new PJ is added to the frame, a new transformed random number is calculated and inserted to the sorted list according to its magnitude. Therefore, when PJ sample has to be re-selected, the change to the existing PJ sample under Pareto sampling is much smaller than the PPS sampling.

6. PAR Sample Selection

6.1 CRSS PAR Sample Selection

The CRSS tertiary sampling units (TSU) are PARs. The CRSS PAR sample is selected by a stratified systematic sampling in the same way as the current GES. For each selected SSU (PJ), PARs are periodically obtained by either technician's visit to the PJ or electronic transmission. 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 number 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.

After PARs are listed, a PAR sample is selected by systematic sampling from the listed or sub-listed PARs by PAR stratum within a PJ[4].

The PAR sampling interval is determined in the following manner. Our goal is to achieve an approximately equal inclusion probability for all PARs in the same PAR stratum. Therefore, the overall inclusion probability(π_{ijlsk}) of PAR k in PSU i, PJ j, sub-list l, and PAR strata s is set to the same sampling rate (r_s) of the PAR stratum s as:

$$\pi_{ijlsk} = r_s = \frac{n_s}{N_s}, \text{ for all } i, j, l, \text{and } k.$$

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

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

$$\pi_{ijlsk} = \pi_i \pi_{j|i} \pi_{l|ij} \pi_{ks|ijl}$$

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 sub-list l is selected as a cluster, and $\pi_{ks|ijl}$ is the selection probability of PAR k from PAR stratum s given that sub-list l is selected. By combining two equations above, the selection probability of PAR k from PAR stratum s becomes,

$$\pi_{ks|ijl} = \frac{1}{\pi_i \pi_{j|i} \pi_{l|ij}} \times r_s$$

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

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

Here, $w_i = 1/\pi_i$ is the PSU weight, $w_{j|i} = 1/w_{j|i}$ 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. $w_{ks|ijl}$ can be non-integer.

When PAR sample is selected with this sampling interval $w_{ks|ijl}$, all PARs within a PAR stratum have equal weights. However, when there are not enough listed PARs, it is possible that $w_{k|ijl}$s is less than 1. In this case, PAR k is selected with certainty, and the weight is set to one. Therefore, these PARs may not have equal weights.

6.2 CISS PAR Sample Selection

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

[4] Technically, PAR listing and sampling is conducted simultaneously. When a PAR is stratified and inserted into the CRSS IT system through the listing process, the PAR is selected if the PAR hits the specified sampling interval.

In CISS, each PSU typically has 1 to 2 technicians and each technician can investigate no more than 2 cases per week. With 2 to 4 cases to be selected per week, it is impossible to stratify the PARs into the 10 PAR domain defined in Table 2. Therefore, PAR sample is selected by PPS using PAR MOS without further PAR stratification other than the weeks.

To ensure the desired percent of sample allocation for PAR domains in Table 2, PAR MOS needs to be carefully calculated. CRSS PAR measure of size is determined the same way as the current CDS.

First PAR measure of size factors (f_s) are estimated for each PAR domain by simulation and are assigned to all PARs in the same PAR domain. Then, PAR measure of size is computed by multiplying PAR measure of size factor, PJ weight and sub-listing factor as

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

In this way, PAR measure of size is specific to PSU, PJ, and PAR domain. This method generates approximately desirable sample allocation.

In CDS data collection, some PARs, especially PARs of severe crashes, may not be available for sample selection until several weeks after the crash. By then, much of the information (scene, vehicle, driver interview, etc.) may become unavailable when the actual case investigation begins. Currently, CDS technicians proceed to investigate these cases even though these cases are of limited value for analysis because of the missing critical items. Because of this, the sample size of the useful cases is much smaller than the nominal sample size. In general, at least 25 percent of all current CDS cases are missing some critical component such as the vehicle inspection.

To better handle the non-responding cases, sample size scalability was introduced to the CISS PAR sample selection. Every week at each selected PSU, after the PARs are listed for all the selected PJs, the Pareto sampling method will be used to select the PARs. The Pareto sample allows for the selection of PARs with PPS selection probability like the current CDS and it also allows for the selection of replacement PARs if some cases turn out to be non-responding. After the PAR MOS is determined, the PAR sample is selected using the Pareto sampling method (see section 5).

7. Optimum Sample Allocation

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 or CISS will be at least as precise as the current GES or CDS for the identified key estimates.

In order to build an optimization model for the CRSS, we simplified the complex sample design. Both CRSS and CISS have a stratified multi-stage probability proportionate to size (PPS) sample design. The deep PSU stratification led to 2 sampled PSUs from each PSU stratum. 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.

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

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

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

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

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

$$mk \geq l.$$

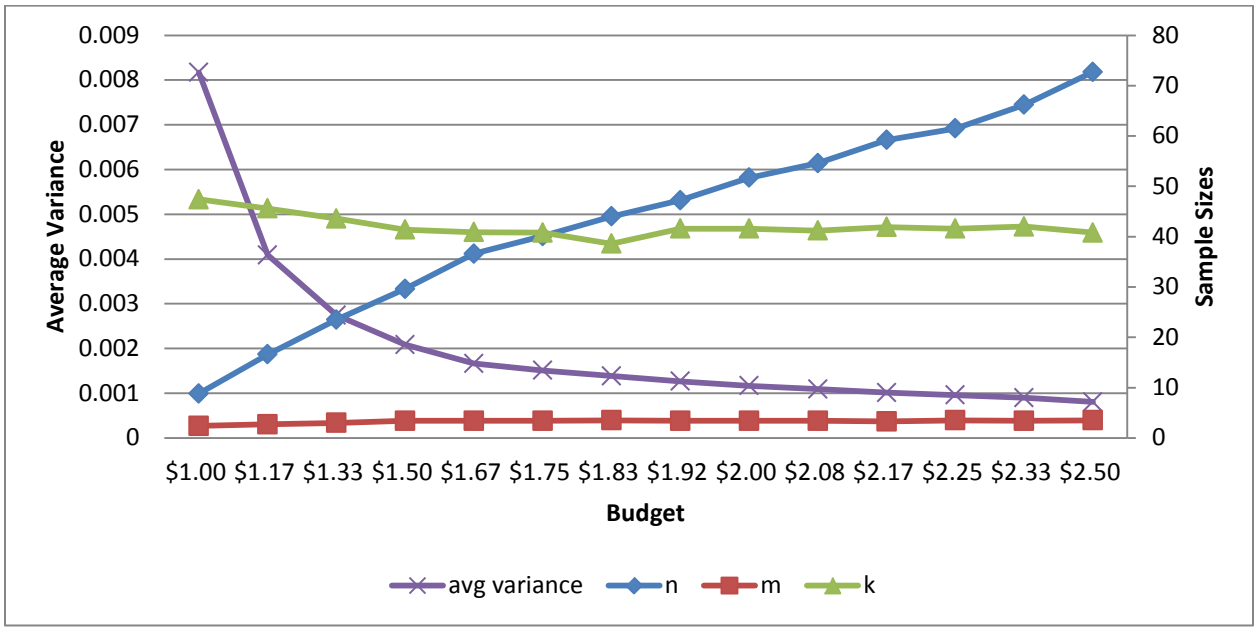
- g: Subscript of the identified key estimate, $g = 1, \cdots G$.
- $\bar{\bar{\bar{y}}}_g$: Identified key proportion estimate.
- n, m, k: Optimal sample sizes of PSUs, PJs, and cases (PARs) to be determined.
- N: Population size of PSUs
- M: Average population size of PJs.
- K: Average population size of PARs
- $V(\bar{\bar{\bar{y}}}_g)$: 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 case-level.
- C, C_0, C_1, C_2, C_3: Total, fixed, PSU-, PJ-, and crash-level cost coefficients.
- $V_{CRSS/CISS}(\bar{\bar{\bar{y}}}_g)$: Variance of the identified key estimate $\bar{\bar{\bar{y}}}_g$ in CRSS or CISS.
- $V_{GES/CDS}(\bar{\bar{\bar{y}}}_g)$: Variance of the identified key estimate $\bar{\bar{\bar{y}}}_g$ in GES or CDS.
- l: known case load.

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

7.1 CRSS Sample Allocation

Figure 2 displays the CRSS 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.

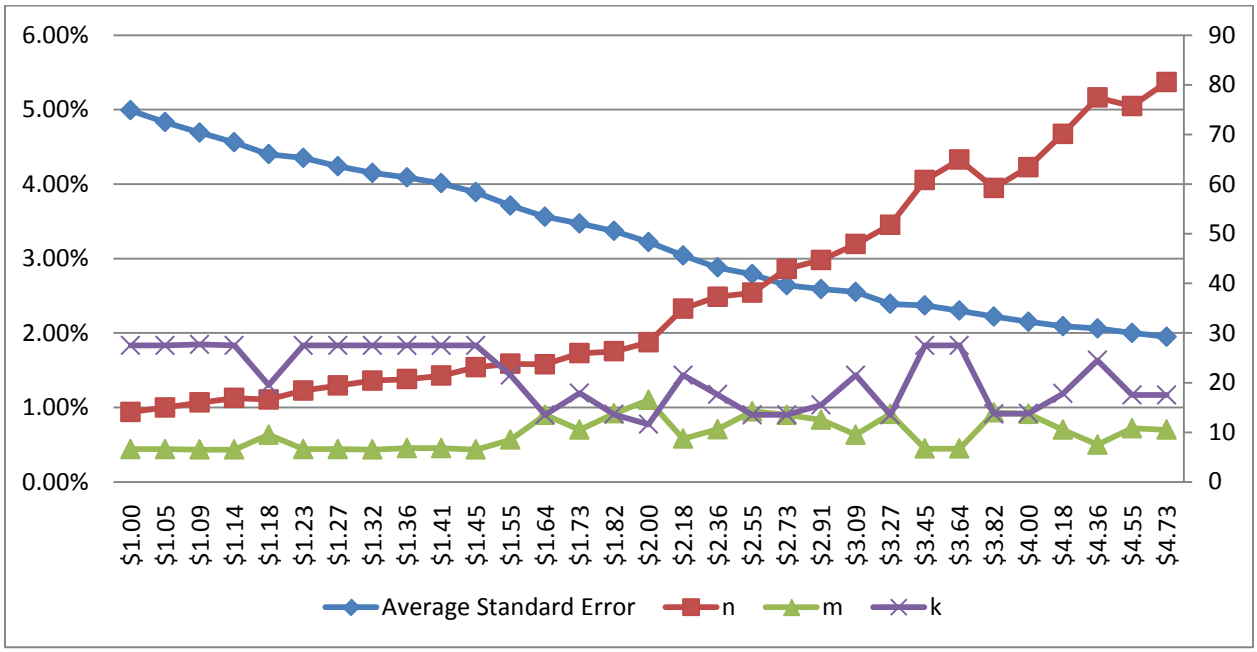
Figure 2: Average Variance, PSU, PJ and PAR Sample Size as Functions of Budget



7.2 CISS Sample Allocation

Figure 3 displays the CISS 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.

Figure 3: Average Variance, PSU, PJ and PAR Sample Size as Functions of Budget



References

Fleming, Charles (2010): *Sampling and Estimation Methodologies of CDS, Technical Report*, DOT-HS-811-327, National Highway Traffic Safety Administration, 1200 New Jersey Avenue, S.E. Washington, D.C. 20590.

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

Noh, Eun Young (2013): *CISS Cost Component Estimation*, unpublished technical report, National Highway Traffic Safety Administration, 1200 New Jersey Avenue, S.E. Washington, D.C. 20590.

Noh, Eun Young and Zhang, Fan (2013): *Sample Allocation for Crash Investigation Sampling System*, unpublished technical report, National Highway Traffic Safety Administration, 1200 New Jersey Avenue, S.E. Washington, D.C. 20590.

Noh, Eun Young and Zhang, Fan (2014): *An Empirical Study of Sequential Poisson Sampling*, unpublished technical report, National Highway Traffic Safety Administration, 1200 New Jersey Avenue, S.E. Washington, D.C. 20590.

Ohlsson, Esbjörn (1998): *Sequential Poisson Sampling, Journal of Official Statistics*, Vol.14, No.2, 1998. pp. 149–162.

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

Shelton, Terry S. (1991): *National Accident Sampling System, General Estimates System, Technical Note, 1988 to 1990*, DOT-HS-807-796, National Highway Traffic Safety Administration, 1200 New Jersey Avenue, S.E. Washington, D.C. 20590.

Westat (2012 A): *Stakeholder Outreach and Summary*, unpublished report to NHTSA. Rockville, MD: Westat.

Westat (2012 B): *Data Elements Recommendations Report*, unpublished report to NHTSA. Rockville, MD: Westat.
Westat (2014): *Survey Modernization Analysis: Designs for a Modernized NASS*, unpublished report to NHTSA. Rockville, MD: Westat.

Zhang, Fan and Chen, Chou-Lin (2013): *NASS-CDS: Sample Design and Weights, Technical Note*, DOT-HS-811-807, National Highway Traffic Safety Administration, 1200 New Jersey Avenue, S.E. Washington, D.C. 20590.

Zhang, Fan; Noh, Eun Young; Subramanian, Rajesh; Chen, Chou-Lin (2014a): *Crash Records Sampling System*, technical report, National Highway Traffic Safety Administration, 1200 New Jersey Avenue, S.E. Washington, D.C. 20590

Zhang, Fan; Noh, Eun Young; Subramanian, Rajesh; Chen, Chou-Lin (2014b): CISS PAR Selection Algorithm and the IT Aspects, unpublished technical report, National Highway Traffic Safety Administration, 1200 New Jersey Avenue, S.E. Washington, D.C. 20590.