Estimating Population and Design Parameters for NHTSA's New National Automotive Sampling System (NASS)

Rui Jiao, Yumiko Sugawara, Martha Rozsi, Sharon Lohr,
James Green, William Cecere
Westat, 1600 Research Blvd., Rockville, MD 20850

Abstract
The new National Automotive Sampling System (NASS) sample design uses a multivariate optimization method to solve for the sample sizes at the first, second, and third stages of sampling, with a view to minimizing the anticipated variance of the variables of interest for a fixed cost or, alternatively, minimizing the cost for target variances. The anticipated variances were calculated by constructing frames with estimates for key variables, and predicting frame estimates using multiple linear regression models with information from the current NASS General Estimates System and Crashworthiness Data System. After the sample of primary sampling units (PSUs) was selected, additional information was obtained about the secondary sampling units in the sampled PSUs, and that information was used to improve estimates of variability at the second and third stages of sampling.

Key Words: sample allocation, optimization, population estimates

1. Design Goals

The National Automotive Sampling System, NASS, was originally designed in the 1970s. The current NASS, based on the 1970's design, is composed of two modules - the General Estimates System (GES) and the Crashworthiness Data System (CDS). These are based on cases selected from a sample of police accident reports (PARs). The selection of sample crashes in both modules is accomplished in three stages:

1) Selection of Primary Sampling Units (PSU's),
2) Selection of police jurisdictions[1] (PJs) within PSUs, and
3) Selection of PARs.

CDS data focus on passenger vehicle crashes, and are used to investigate injury mechanisms to identify potential improvements in vehicle design. GES data focus on the bigger, overall motor vehicle crash picture, and are used for problem size assessments and tracking trends.

[1] Police jurisdictions for this study include police agencies that respond to motor vehicle crashes and write PARs.

NASS has proven to be a reliable resource for NHTSA and the broader motor vehicle safety research community since its inception. In the time since the last redesign of the survey, however, the distribution of crash types has changed. For example, the number of motor vehicle crashes involving fatalities has dropped from a high of more than 39,000 in 2005 to fewer than 31,000 in each of 2009 to 2012 (NHTSA, 2014). In addition, there have been many changes in road and automobile design, in traffic volume and traffic safety, in population growth and mobility, and in methods for collecting information about crashes. All these changes signaled the need for a re-examination of the NASS survey and study design. A redesign of NASS was undertaken which attempts to meet these new and diverse requirements through expanding its scope and making it more responsive to changing needs.

The new NASS design will have two modules: the Crash Report Sampling System (CRSS), which will replace the GES, and the Crash Investigation Sampling System (CISS), which will replace the CDS. Each of these modules will have a three-stage sampling design using the same units for each stage as in the GES and CDS (see Cecere et al., 2014). Stratification will be performed at each stage of sampling. At the third stage, the PARs will be stratified by crash severity and vehicle model years.

Westat designed a three-stage sample allocation optimization system motivated by three major features desired for the redesign:

1) Due to uncertainty of the future NASS budget, flexibility is needed for sample sizes (see Rozsi et al., 2014);
2) The design is to oversample crashes of specified types – in particular, crashes in which an occupant is seriously injured and crashes involving newer vehicles; and
3) The redesign planning should allow obtaining the anticipated precision for a variety of key estimates.

The optimization system, described in detail in Sugawara et al. (2014), is designed so that the survey design can be adapted to future changing conditions by inputting different precision and/or cost constraints. The optimization program calculates designs from two perspectives: allocating resources so that cost is minimized while achieving precision targets, and allocating resources to minimize a preference-weighted sum of variances of estimates with fixed cost constraints.

The flowchart in Figure 1 shows the structure of the optimization system. Five input files are required for this system. The first input file specifies the variables of interest (x_i). The second file provides information on the stratification and the population counts at each stage of sampling: N_h is the number of PSUs in PSU stratum h, from $h = 1$ to H; M_{hia} is the number of PJs in PJ stratum a within PSU i of PSU stratum h; and K_{hiajl} is the number of PAR records in PAR stratum l of PJ (a, j) of PSU (h, i). Much of this information is unknown at preliminary design stages---for example, the number of PJs in different PSUs is unknown until the PSUs are selected---so the optimization system allows putting in estimates or a constant value for these quantities. The third input file provides the population variances for each stage of sampling and for each variable listed in the first input file. In the flowchart, *S*1x2, *S*2x2, and *S*3x2 represent the population variances at the PSU, PJ, and PAR levels, and different values of these can be input for different strata at each stage of sampling. The final two input files give cost components (the fixed costs, and the per-unit costs to include an additional PSU, PJ, or PAR in the

sample; these can vary across PSU strata), and Cost of Living Adjustment factors for each PSU stratum. The system outputs the status of a particular run (either optimization achieved or failed); objective function values such as the cost and sum of relative variances over variables of interest; the first, second and third stage sample sizes in each stratum; total sample sizes of PJs and PARs; and a report giving the performance of the design with respect to the target precisions based on the current GES and CDS. From the flowchart in Figure 1, one may observe that this is an iterative process where a user can try different parameters to explore the effects of different assumptions.

The optimization system calls for estimating population counts for each of the variables of interest in every PSU and every PJ, so that the quantities in input files 2 and 3 can be calculated and be fed into the system. These quantities need to be estimated using currently available information. This paper focuses on estimating population parameters needed for the second and third input files. The NASS redesign was optimized in two phases: first, the number of PSUs was determined. Then, after the PSU sample was drawn, additional information was collected about the population of PJs in the sampled PSUs. The optimization system allows the information to be updated as better information becomes available, so the second and third stage designs could be refined after the PJ information was collected for the sampled PSUs.

An important note is that this estimation is only used for design purposes, and at each stage of the design, the estimates are the best projections available at that stage. The estimates produced for use with the optimization system are not meant to be used for analysis, since it is the purpose of the new NASS to obtain data to provide reliable estimates of the quantities of interest.

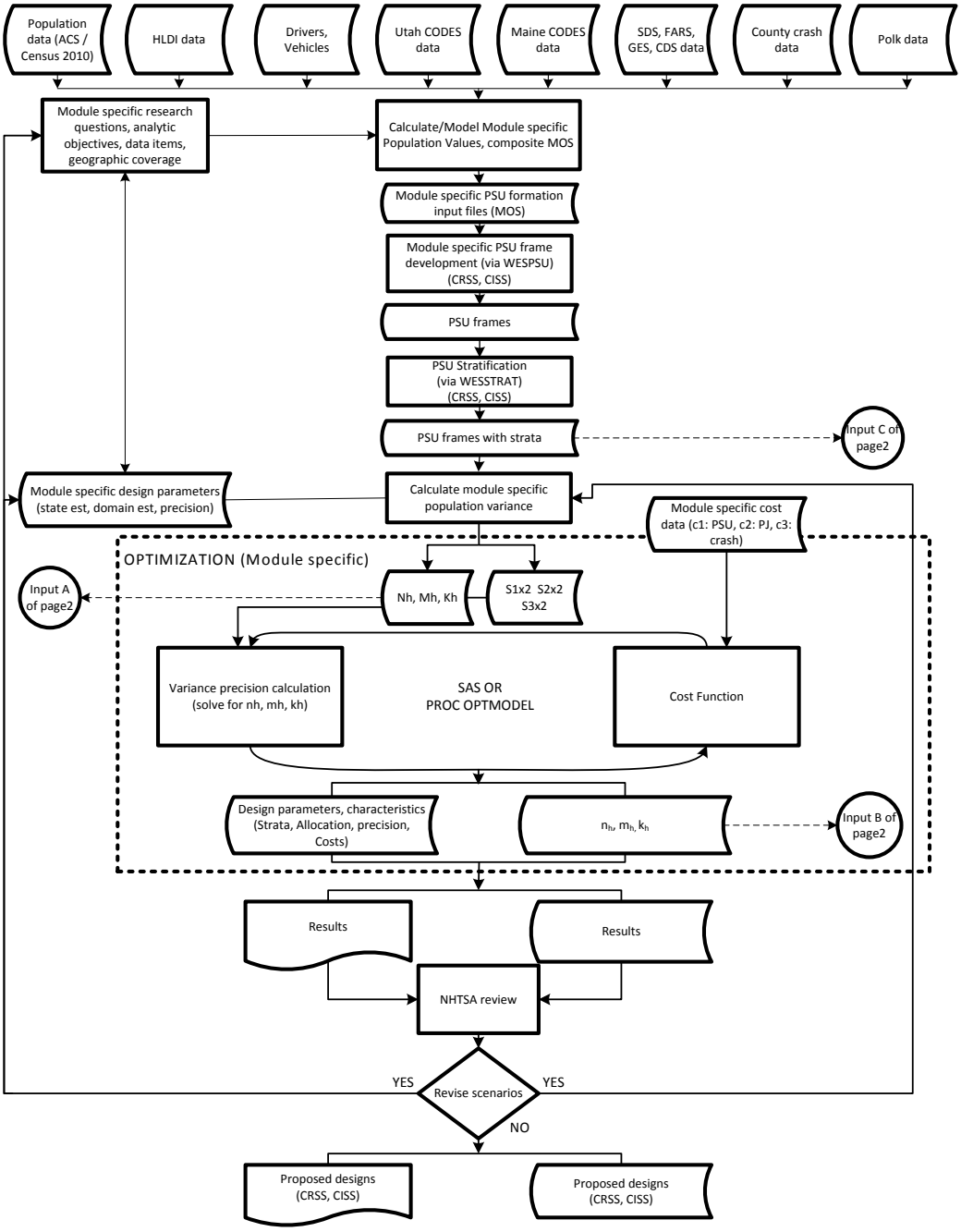

Figure 1: Flowchart of the optimization system architecture

2. Preliminary Estimates for Determining Number of PSUs

In the optimization system, the determination of number of PSUs was driven by the cost model given in Sugawara et al. (2014) and the population variances for each of the estimates of interest. Each NASS module has its own set of key estimates. In order to calculate the variances of those key estimates, we estimated the total crashes for each of the key estimates in every county and aggregated them to the new PSUs of the redesign,

which are described in Cecere et al. (2014). Sections 2.1 and 2.2 describe the procedures of estimating population parameters at the PSU level for the CRSS and CISS modules.

The total crashes of different types were needed for every PJ and PAR stratum as well in order to calculate the variances for key variables. In the initial stages of the design, little information was known at the PJ level. In fact, the number of study-eligible PJs in a PSU was known only for areas in the current GES or CDS, and for many key variables, the GES and CDS samples were the only source of information.

Westat found rough counts of the number of PJs in different counties from the USACOPS® website at www.usacops.com, which lists PJs in each county. The website does not provide information on whether the PJs write PARs and send them to the state, however, and also has no information on motor vehicle crashes in the PJs. In the absence of population information for the PJs in the early phases of the redesign, we used information from other data sets to obtain estimated numbers of crashes and estimated variability at the second and third stages from the 2011 GES and CDS. The remainder of this section describes the modeling that was used to obtain estimates of crash counts and variances for the CRSS and CISS.

2.1 CRSS (formerly GES)

The CRSS module has 9 analytic domains of interest, referred to as CRSS PAR strata. They are shown in Table 1. These strata were defined by severity of injury, type of vehicle, model year of the vehicles involved in the crash, and by the involvement of non-motorists, motorcyclists, or busses or trucks.

Table 1: Descriptions of the nine CRSS PAR strata

PAR stratum number	*PAR stratum description*
Stratum 1	Crashes involving a killed or injured non-motorist
Stratum 2	Crashes not in stratum 1 involving a killed or injured motorcyclist
Stratum 3	Crashes not in stratum 1 or 2 in which at least one occupant of a late model year passenger vehicle is killed or incapacitated
Stratum 4	Crashes not in strata 1-3 in which at least one occupant of an older passenger vehicle is killed or incapacitated
Stratum 5	Crashes not in strata 1-4 in which at least one occupant of a late model year passenger vehicle is injured
Stratum 6	Crashes not in strata 1-5 involving at least one medium or heavy truck or bus
Stratum 7	Crashes not in strata 1-6 in which at least one occupant of an older passenger vehicle is injured
Stratum 8	Crashes not in strata 1-7 involving at least one late model year passenger vehicle (and in which no one is killed or injured)
Stratum 9	Crashes not in strata 1-8.

Three sources provided information at the county level that could be used to estimate population counts of crashes for each county in each of the CRSS PAR strata:

- The Fatality Analysis Reporting System (FARS; see NHTSA, 2014), which is a national census of motor vehicle crashes involving fatalities;
- The State Data System (SDS) records; and
- Information from the R. L. Polk Company.

Because FARS is a census, it provides the number of fatal crashes for every county. SDS contains crashes by year and maximum injury severity for every county in 33 states. The information was imputed for counties in the remaining states. The information from the R. L Polk Company included vehicle counts and vehicle miles travelled by model year and vehicle type for every county.

The FARS and SDS data together give estimates of numbers of crashes in each county that involve either a fatality or an incapacitating injury. The FARS data provide information on model years of the vehicles, but the SDS data do not. Thus, the FARS and SDS data provide information on the total crashes for strata 3 and 4 together, strata 5 and 7 together, and strata 8 and 9 together, but do not provide separate estimates for the strata in each of these pairs. The Polk data were used to help allocate the crashes by vehicle types and the model year of the vehicle, so the counts could be split into individual strata. There was no easy way of estimating the crash counts for strata 2 and 6, so these were based on numbers of registered vehicles (motorcycles, trucks, and busses) from the Polk data, which were poststratified to 2011 GES estimates. Stratum 1 was estimated by using a five year average of fatal crash counts involving nonmotorists taken from the 2007 through 2011 FARS data. The estimates from the FARS data for stratum 1 were then poststratified to 2011 GES estimates.

Besides the nine analytic domains described by the PAR strata in Table 1, an additional 15 subgroups are of interest for CRSS, shown in Table 2. The only relevant data available were the PAR strata counts estimated for every county, and estimates from the previous GES, limited to the sampled PSUs.

Table 2: Additional crash counts of interest for CRSS

1	Any vehicle involved
2	Injury type - fatal injury
3	Injury type - serious injury
4	Vehicle type - passenger cars
5	Vehicle type - light truck vehicle, i.e. truck, van, or SUV
6	Vehicle type - bus
7	Vehicle type - medium/heavy truck
8	Vehicle type - motorcycle
9	Crash type - rollover
10	Crash type - front
11	Crash type - side
12	Crash type - rear
13	Impact type - multiple vehicles
14	Impact type - pedestrian
15	Impact type - bicycle

The quantities in Table 2 were estimated for each frame PSU using multiple linear regression models developed on the PSUs in the 2011 GES, using the estimated CRSS PAR stratum counts from Table 1 as predictors. Because of the skewness of the outcomes, the models were fit using log transformed response variables (both log and square root transformations were considered, and log transformations resulted in better fits as well as predicted counts that were forced to be positive). To reduce multicollinearity and the variance of the predicted values, a reduced set of predictors was used for each response. One possible concern, since the GES has only 60 PSUs available

for developing the model, is that a model selection method could result in a downwardly biased mean squared error (MSE). It was verified that the MSE from each reduced model was at least as large as that from the model with all covariates. If desired, the MSE from a model could be adjusted using methods discussed in Copas and Long (1991).

Two randomly generated normal variables were added to the predicted values in the log-transformed scale for each regression model: one to account for the variability in the estimated regression equation, and a second to account for the residual variability in the model (with variance equal to the MSE). The added noise terms allowed the predictions of counts to reflect the variability in the data. The predicted value with added noise terms was then exponentiated to obtain the predicted crash count for each county in the frame.

The regression models were then used to calculate the subgroup estimates in Table 2 at the county level and the predicted values were poststratified to agree with the corresponding estimated total crashes based on the 2011 GES.

The regression models gave estimates of counts for the CRSS PAR strata and for other key variables at the PSU level. At the early stages of the NASS redesign, little information on crashes was available at the PJ level. Keeping the same approach to create input files 2 and 3 in Figure 1 that was done for PSU would require detailed knowledge of a PJ frame. Instead of generating a PJ frame with the population counts of interest and then calculating the population variance, we estimated the between-PJ variance from the current GES for each key estimate, and applied that value to all PJs. Similarly, the variance at the third stage (PAR level) was estimated from the current GES for each PAR stratum, and that variance was used across the PSUs in the frame.

Thus, in the early stages of the design, the crash counts used in the frame at the PSU level were based on information from FARS and the SDS, but the variance estimates at the PJ and PAR levels were input using estimates from the current GES.

2.2 CISS (formerly CDS)

Similarly to the CRSS, the CISS module has 10 analytic domains referred to as CISS PAR strata. Table 3 describes those domains, defined by the cross classification of severity of injury in a crash and the model year(s) of passenger vehicle(s) involved in the crash. An additional 10 key estimates of crash types were of interest for the CISS redesign, and these are listed in Table 4.

Table 3: CISS PAR strata

PAR stratum number	*PAR stratum description*
Stratum 1	Crashes involving a killed passenger vehicle occupant.
Stratum 2	Crashes not in Stratum 1 involving a recent model year passenger vehicle in which an occupant is incapacitated
Stratum 3	Crashes not in Stratum 1 or 2 involving a recent model year passenger vehicle in which an occupant is possibly injured by severity is unknown
Stratum 4	Crashes not in Stratum 1-3 involving a recent model year passenger vehicle in which all occupants are not injured
Stratum 5	Crashes not in Stratum 1-4 involving a mid-model year passenger vehicle in which an occupant is incapacitated
Stratum 6	Crashes not in Stratum 1-5 involving a mid-model year passenger vehicle in which an occupant is possibly injured
Stratum 7	Crashes not in Stratum 1-6 involving a mid-model year passenger vehicle in which all occupants are not injured
Stratum 8	Crashes not in Stratum 1-7 involving an old model year passenger vehicle in which an occupant is incapacitated
Stratum 9	Crashes not in Stratum 1-8 involving an old model year passenger vehicle in which an occupant is possibly injured
Stratum 10	Crashes not in Stratum 1-9 involving an old model year passenger vehicle in which no occupants are injured

Table 4: Other CISS estimates needed for the design optimization

1	Total crashes in which a vehicle was towed away
2	Total occupants
3	Total vehicles
4	Total occupants by injury type - fatal
5	Total occupants by injury type - severe injury
6	Total occupants by injury type - other injury
7	Total crashes by crash type - rollover
8	Total crashes by crash type - rear-end
9	Total crashes by crash type - head-on
10	Total crashes by crash type - angle

The purpose of the CISS module is to provide annual, national estimates of the number, types and detailed characteristics of crashes in which a passenger vehicle is towed from the scene. This information is collected by trained crash investigators, who visit the crash scene, interview witnesses, and examine medical records. The PSUs formed for the CISS were not the same as the PSUs formed for the CRSS because of the different data requirements (Cecere et al., 2014). Therefore, CISS PSU level population parameters also needed to be estimated.

The CRSS target population is all crashes for which PARs are written, but the target population for the CISS is crashes in which a passenger vehicle is towed. Consequently, with the exception of PAR stratum 1, the CISS PAR stratum counts could not be calculated directly by using the FARS, SDS, and Polk data as was done for the CRSS. Instead, associations between the CISS PAR strata counts and the CRSS PAR strata counts were exploited through regression models fit using the current CDS, which contains information allowing classification of crashes by CRSS PAR strata and by CISS PAR strata. These models were then applied to obtain predicted crash counts for CISS

PAR strata 2-10 for the PSUs in the CISS frame. The FARS data provide county-level data on the number of crashes involving a fatality when a passenger vehicle was towed from the scene, so the PAR stratum 1 count was obtained directly from the FARS.

Multiple regression models were used to obtain the key variable estimates for CISS similarly to the procedure that was used for CRSS. The between-PJ variance and within-PJ variance for CISS were estimated by the same procedure that was used for CRSS.

Once the necessary input files were constructed, the optimization system was run for each of CRSS and CISS. The preliminary runs of the optimization system, together with consideration of budgetary constraints, determined the number of PSUs for the redesign. During the early design stages, the optimization program consistently allocated large numbers of PSUs for some PSU strata. This suggested that a finer stratification could be done, and led to a refined PSU stratification, with smaller variances, for the final design.

3. Estimating Population Counts at the PJ Level

After the PSU sample for each redesigned module was chosen, NHTSA researchers developed a PJ frame in the sampled PSUs and collected information on six crash types, described in Table 5, for every PJ in the sampled PSUs. With this information it was possible to determine the number of PJs writing PARs in each sampled PSU and to refine the estimated population counts for the key variables at the PJ level. In some sampled PSUs, all PJs would be sampled; in others, the PJs were stratified using information collected for the PJ frame.

Table 5: Categories of PAR counts obtained for each PJ in the sampled PSUs

Category	*PARs involving*
1	Pedestrian
2	Motorcyclist
3	Commercial vehicle
4	Fatal injury
5	Serious or other injury
6	None of the categories above; PARs with property damage only

Categories 1, 2, and 3 can be mapped directly to CRSS PAR strata 1, 2, and 6, respectively. The remaining CRSS PAR strata counts at the PJ level were estimated by apportioning the PJ counts in the different categories to appropriate PAR strata. This was done by using the current GES to estimate the proportion of crashes in that category belonging to each CRSS PAR stratum, and using a multinomial distribution to allocate crash counts using that proportion.

After implementing this procedure, CRSS PAR strata counts were available for each PJ in the population within each sampled PSU. In order to obtain counts for the other CRSS and CISS key estimates, models similar to those described for modeling the PSU-level counts were applied at the PJ level.

With the fine-tuned procedure at the PJ level, the optimization system was able to allocate the number of sampled PJs per PJ stratum, and the number of sampled crashes per PJ, with consideration of the budget and precision level of the key estimates.

4. Discussion

The sample design optimization system for the redesigned NASS described in Sugawara et al. (2014) requires files containing estimates of population counts of sampling units and estimates of variance terms for each key variable at all the three stages of sampling. In this paper, we described how these estimates were calculated at different points in the sample design process, using the information available at that time. In earlier stages of the design, the frame of population counts at the PSU level were based on the information available from FARS, SDS, and the association between some key variables and the auxiliary variables from the current GES / CDS, and variance terms were calculated based on the frame data. At the PJ and PAR levels, the population counts and variance terms were not computed from the frame since little information was known about the PJs. Instead, population counts and variance terms were estimates from the current GES / CDS. In later stages of the design, after the PSU samples were drawn for both CRSS and CISS, the PJ and PAR population was enumerated and estimated within the sampled PSUs, and the variance terms were estimated with this data.

The sample design optimization system used to redesign NASS is highly parameterized and flexible, allowing the user to identify key variables, provide population counts and variance terms obtained through alternative approaches, and specify the per-unit cost coefficients for each study type as well as cost of living adjustments. These parameters and flexibility allow for a relatively large number of designs to be developed and considered in terms of their cost and precision, subject to a number of other constraints and assumptions. The effects of changes in inputs, constraints and assumptions can be discovered through successive iterations. The sample design optimization system could be used for essentially any three stage design, possibly with some modifications, and the system could be modified beyond three stages with moderate changes in its architecture.

References

Cecere, W., Jiao, R., Rozsi, M., Severynse, J., Lohr, S., and Green, J. (2014). Composite measure of size evaluation and primary sampling unit formation for NHTSA's redesign of the National Automotive Sampling System. In press, *Proceedings of the American Statistical Association Section on Survey Research Methods.*

Copas, J.B., and Long, T. (1991). Estimating the residual variance in orthogonal regression with variable selection. *The Statistician*, 40, 51-59.

National Highway Traffic Safety Administration (NHTSA, 2014). Fatality Analysis Reporting System (FARS) Encyclopedia. Available at http://www-fars.nhtsa.dot.gov/Main/index.aspx, last accessed 9/12/2014.

Rozsi, M., Cecere, W., Lohr, S., and Green. J. (2014). Creating a flexible and scalable PSU sample for NHTSA's redesign of the National Automotive Sampling System. In press, *Proceedings of the American Statistical Association Section on Survey Research Methods.*

Sugawara, Y., Das, B., Jiao, R., and Green, J. (2014). Multivariate sample design optimization for NHTSA's new National Automotive Sampling System. In press, *Proceedings of the American Statistical Association Section on Survey Research Methods.*