Multivariate Sample Design Optimization for NHTSA's New National Automotive Sampling System

Yumiko Sugawara[1], Barnali Das[1], Rui Jiao[1], James Green[1]
[1]1600 Research Boulevard, Rockville, Maryland 20850

Abstract

NHTSA's redesigned National Automotive Sampling System (NASS) is expected to address a large number of research questions and analytic objectives, and therefore required a multi-purpose study and sample design. In addition, the new NASS consists of multiple modules, with future funding levels and precision requirements unknown and subject to change for any given module. A multivariate sample design optimization system was designed and built for NHTSA to address these fluid design requirements, parameters and constraints. The system offers two options: A) Minimize cost subject to variance constraints; B) Minimize a sum of relative variances subject to cost constraints. This paper presents the development, architecture and utility of the sample design optimization system, along with a description of its outputs, and the necessarily iterative process of reviewing those outputs, and modifying design parameters and constraints to arrive at a reasonable sample design.

Key Words: Multivariate sample design, optimization

1. Introduction

The National Automotive Sampling System (NASS) was established in the 1970s by the National Highway Traffic Safety Administration (NHTSA) and it provides nationally representative estimates of motor vehicle crashes. There are two modules in the current NASS - the General Estimates System (GES) and the Crashworthiness Data System (CDS). Both modules use a national sample of crashes selected from police accident reports (PARs) but serve different purposes. GES gives estimates of number of various types of motor vehicle crashes and it requires a large sample of PARs. While GES aims to capture the overall trend of crashes, CDS has its focus on passenger vehicle crashes with at least one car being towed, and greater details about individual crashes are collected where investigators collect data from crash sites and also interview crash victims. CDS data requires a smaller sample and is used to assess the safety standards and understand the injury mechanisms that may change due to improvements made to vehicles. NASS has also previously included other special studies, "the National Motor Vehicle Crash Causation Survey (NMVCCS)," and "the Large Truck Crash Causation Study (LTCCS)." Although additional modules beyond the GES and FOPV were not designed for the redesigned NASS at this time, the sample design optimization system was built for and has the ability to add data and parameters for additional modules and obtain optimization results.

Since its start, there have been many significant changes in the society such as those in road and vehicle design, in population growth and mobility, and in traffic volume and safety. Because of these changes and a concern about a limited sample, NHTSA decided to redesign NASS to meet new, more diverse needs.

NHTSA contracted with Westat to lead the survey modernization effort, and NHTSA and Westat have worked collaboratively to accomplish the redesign of NASS. During the process of redesigning NASS, GES continued to be referred to as GES, while CDS was referred to as the Follow-on Passenger Vehicle (FOPV). In the final redesigned NASS, GES and CDS/FOPV were renamed to Crash Report Sampling System (CRSS) and Crash Investigation Sampling System (CISS) respectively.

Many authors have considered analytical and nonlinear programming solutions to the problem of optimal designs for stratified sampling with multiple responses (see, for example, Bethel (1989), Causey (1983), Díaz-García and Cortez (2008), Green (2000), Green (2001), Khan (2006) Rao (1993), and Valliant and Gentle (1997)). In this paper, we develop a computation-based approach that allows consideration of different costs and variance structures within different strata, describing the sample design optimization system for NHTSA, as well as the specific results obtained for the redesigned NASS GES and FOPV modules.

2. The NASS Redesign Optimization Problem

2.1 NASS Sample Design

The current NASS contains two modules, GES and CDS, and its design is a stratified three stage sample in both modules. The primary sampling units (PSUs) consist of counties or groups of counties, the second sampling units are police jurisdictions (PJs) and the third sample units are police-reported crashes (PARs). This design is carried over to the redesign of NASS. The survey has higher sampling fractions for serious injury, rarer crashes, and there are multiple estimates of interest.

Due to uncertainty in budget, flexibility to accommodate cost constraint changes in the future was desired. Therefore, several “scenarios” were developed under different budget assumptions for each module.

2.2 Key Estimates

The key estimates are all of PAR strata and subgroups of interest for each module. PAR strata are mutually exclusive groups of crashes with a desired sampling fraction assigned to each group, whereas subgroups of interest are outcomes that can cut across multiple PAR strata for each module. For GES, the PAR strata are defined by kinds of damage caused by the crash, age of vehicle involved, and types of vehicle. Combining PAR strata and subgroups of interest for GES, there are total of 25 key estimates. For FOPV, the PAR strata are defined by a finer breakdown of model year of a vehicle and severity of injury. Combined, there are 20 key estimates in FOPV. The sample design optimization system aims to achieve improved precision for these estimates compared to the current NASS. In the optimization section, key estimates are referred to as variables of interest.

2.3 Optimization Problems

In sample surveys, one can allocate available resources such that either the cost is minimized while achieving a desired precision, or the variance is minimized while keeping the cost within a budget range.

The optimization of resource allocation becomes complex when there are multiple key estimates and budget can change in the future as is the case for NASS redesign. The target precision was based on the past GES and CDS, and the budgets and cost coefficients at each stage of sampling for the new GES and FOPV were provided by NHTSA. The optimization system for NASS runs either option depending on user-specified option to allow more flexibility.

3. Sample Design Optimization System Development and Implementation

3.1 Optimization

In order to develop the multivariate optimization system, SAS/OR® 9.3 software, which is an Operations Research Software from SAS Institute Inc, was employed (SAS Institute Inc., 2011). Because NASS is a multi-purpose study with several constraints, a powerful software such as SAS/OR® software was needed to compute the solution in a reasonable run time. In our optimization system, PROC OPTMODEL, Multistart NLP option was selected. Optimization models in PROC OPTMODEL include linear, mixed-integer linear, quadratic and general nonlinear models. This is similar to PROC NLP (Non Linear Programming) in version 9.2 with improvements (Huang and Hughes (2010)). The NLP option allows both nonlinear equality and inequality constraints, and with Multistart option, the program starts at several different initial points and chooses the best solution out of a set of locally optimal solutions. This is suitable for problems with many local minima such as the optimization in redesign of NASS. Some of the advantages of using SAS/OR® software in building our system are its power, robustness, and flexibility. SAS/OR® software is capable of solving complex optimization problems such as those seen in NASS redesign in a short period of time, and in order to run the optimization, a user only needs to specify some parameters and change the input files called by the program. It is flexible but also less likely to accidentally modify the optimization program itself unlike some other options considered during the development.

There are some requirements and limitations to our system, however. First, the user must construct an appropriate objective function and enumerate the constraints, which can be challenging to program. For instance, FOPV has a PAR workload requirement of two PARs per researcher per week, where the number of researchers was assumed to vary from one to four researchers per PSU. Therefore, this constraint varied for different PSUs.

Additional programming produced post-solutions reports on cost, variance, performance relative to the current GES and CDS/FOPV, as well as variance components and cost terms at each sampling stage based on the solution.

Finally, SAS/OR® software takes the problem literally, meaning that the program will only take into account what are presented by the objective function and constraints, and does not consider any other potential issues. Because the system can only give the optimal solution based on the input, results must be interpreted in light of assumptions, constraints, options and features of each scenario to determine what is reasonable.

3.2 Architecture of the Optimization System in Redesign of NASS

There are five input files required when using this system for the redesign of NASS; variables of interest (x_i), population counts (N_h, M_h, K_h), variance components (S1x2, S2x2, S3x2), cost components (C0, C1, C2, C3), and cost of living adjustment factors ($COLA_h$). This setup is customized for NASS redesign, but in a general use, some of these parameters can be dropped or set to 0 or 1 as seen in the next section.

For NASS redesign, just one macro is needed for any module (GES or FOPV)-optimization option combinations.

The system outputs the status of a particular run, either optimal or failed, objective function values such as cost and sum of relative variances over variables of interest, first, second and third stage sample sizes, total sample sizes at the second stage (PJs) and at the third stage (PARs), and performance report relative to target precision, which, for this project were based on previous GES and CDS data. From the flowchart in figure 1, one may observe that this is an iterative process where a user can try different parameters to see which design and assumptions are most appropriate.

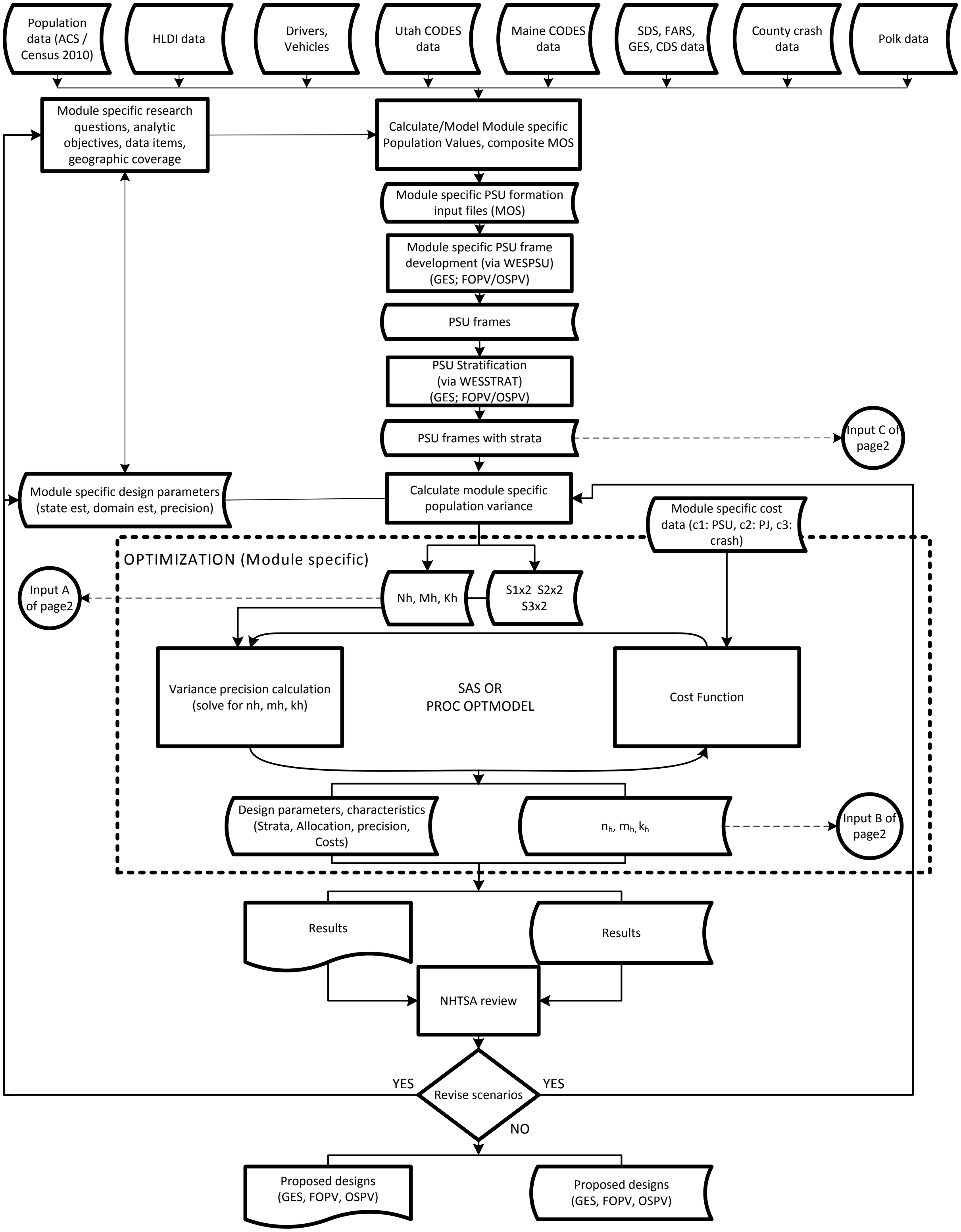
Population data (ACS / Census 2010)
HLDI data
Drivers, Vehicles
Utah CODES data
Maine CODES data
SDS, FARS, GES, CDS data
County crash data
Polk data
Module specific research questions, analytic objectives, data items, geographic coverage
Calculate/Model Module specific Population Values, composite MOS
Module specific PSU formation input files (MOS)
Module specific PSU frame development (via WESPSU) (GES; FOPV/OSPV)
PSU frames
PSU Stratification (via WESSTRAT) (GES; FOPV/OSPV)
PSU frames with strata
Input C of page2
Module specific design parameters (state est, domain est, precision)
Calculate module specific population variance
Module specific cost data (c1: PSU, c2: PJ, c3: crash)
OPTIMIZATION (Module specific)
Input A of page2
Nh, Mh, Kh
S1x2 S2x2 S3x2
Variance precision calculation (solve for nh, mh, kh)
SAS OR PROC OPTMODEL
Cost Function
Design parameters, characteristics (Strata, Allocation, precision, Costs)
n_h, m_h, k_h
Input B of page2
Results
Results
NHTSA review
YES
Revise scenarios
YES
NO
Proposed designs (GES, FOPV, OSPV)
Proposed designs (GES, FOPV, OSPV)

Figure 1. Flowchart of the Optimization System Architecture

3.3 Cost and Variance Models, Objective Functions and Variables

The cost model used for optimization was the following.

$$C = C_0 + C_1 n + \ C_2 nm + \ C_3 nmk$$

where C is the overall cost, C_0 is the fixed cost, C_1 is the cost coefficient at the first stage, C_2 is the cost coefficient at the second stage, and C_3 is the cost coefficient at the third stage. The decision variables are n, the number of first stage units (PSUs) to sample, m, the number of second stage units (SSUs) to sample in a PSU, and k, the number of third stage units (TSU) to sample in a sampled SSU in a selected PSU. This cost model is written in a general way and it can be varied to fit the survey design of interest. For example, if it is known that there is no fixed cost, then C_0 can be dropped or simply be set to 0. As another example, stratification at the first and second stages can be incorporated in the cost model above. This is shown in section 3.6, which presents the application in NASS redesign.

Variance model used for this optimization can be summarized as

$$V(x_i) = \ V_{x_i}(PSU) + V_{x_i}(SSU) + \ V_{x_i}(TSU)$$

where x_i is a variable of interest (key estimate).

3.4 Optimization Objectives, Options, Outputs and Results

As discussed in section 2.3, two optimization options for NASS redesign were developed. The option A is to minimize cost subject to variance constraints for multiple key estimates, where variance for each variable of interest is smaller than or equal to a preset target variance.

The option B is to minimize the sum of relative variances of all variables of interest subject to cost constraints while keeping the overall cost less than or equal to a preset target budget. Of course, both options require that the sample size at each stage of sampling be between 2 and the appropriate population size.

If desired, additional constraints/features can be added, such as restricting the first stage sample size to 2 per stratum, minimizing total number of second stage units, setting a minimum number of PARs per researcher per week per PSU in FOPV, forcing the overall sample size of cases (PARs) to be larger than or equal to some preset minimum, creating a customized output file with decision variables values (useful for drawing a sample assuming a resulting design is selected), and exempting particular variables of interest from a run to reach a feasible solution.

3.5 General Use of the System

Though the system was originally developed specifically for the redesign of NASS, the system can be used for any three stage designs with multiple estimates of interest, for both options A and B as described in the previous section. The models, constraints and inputs can be changed to accommodate more/less complicated cost and variance models.

3.6 Application in Redesign of NASS

This cost function is a variation of the general cost model in section 3.3. This model reflects the stratification at the first and second stages of sampling as well as an adjustment for the cost of living in each PSU based on its location.

$$C = C_0 + \sum_{h} COLA_h [C_1 n_h + \sum_{i=1}^{n_h} \sum_{a=1}^{A_{hi}} (C_2\, m_{hia} + C_3 m_{hia} k_{hia})]$$

where C is the overall cost, C_0 is the fixed cost, C_1 is the cost at the first stage of sampling (PSU), C_2 is the cost at the second stage of sampling (PJ), and C_3 is the cost at the third stage of sampling (PAR). $COLA_h$ stands for cost of living adjustment in PSU stratum *h*. The decision variables are n_h, the number of first stage units (PSUs) to sample in PSU stratum h, m_{hia}, the number of second stage units (PJs) to sample in PSU stratum *h*, PJ stratum *a* in a given PSU i, and k_{hia}, the number of third stage units (PARs) to sample in PSU stratum h, PJ stratum a in a given PSU i. A_{hi} stands for the number of PJs in a given PSU *i* in PSU stratum *h*, provided by NHTSA.

Variance model used for this optimization can be summarized as

$$V(x_i) = V_{x_i}(PSU) + V_{x_i}(PJ) + V_{x_i}(Case)$$

where x_i is a variable of interest (key estimate). The variance at each stage for each variable of interest was calculated based on the population counts and estimates for NASS redesign. The details of population estimates are illustrated in “Estimating Population and Design Parameters for NHTSA’s New National Automotive Sampling System (NASS)” by Jiao et al.

4. Results

In this section, the GES and FOPV results are presented. In order to obtain these results, a limited number of scenarios were considered, where each scenario was associated with a given overall budget, which in turn suggested an approximate PSU sample size. The bigger the budget, the more PSUs are selected in a scenario.

4.1 GES

In option A, where the goal is to minimize overall cost subject to the precision constraints, one key estimate always seems to be “binding,” driving the resource allocation. For each scenario, table 1 shows the number of PSUs, PJs and PARs to be sampled, the resulting relative cost, and the number of feasible estimates out of 25 key estimates referred to in section 2.1. Because the target precision was based on the 2011 GES with 60 PSUs, the target precision for some estimates cannot be met if the number of PSUs in a new design is too small. For the GES, PJ population information for scenario 1 was not available since scenario 1 was not expected to be implemented in the near future. Therefore, the table only presents results for scenarios 2 through 5. Scenario 4 shows the most expensive design because one of the variables of interest for which the precision was difficult to meet drove the allocation and resulted in taking much bigger second stage and third stage samples compared to the other scenarios. If one were to remove the particular variable, the results would have changed.

Table 1: GES Option A Result

GES Scenario	*PSUs*	*PJs*	*PARs*	*Cost*	*Number of Feasible Estimates*
GES, scenario 2	75	358	7,104	2.19	20 / 25
GES, scenario 3	51	276	8,049	1.84	18 / 25
GES, scenario 4	24	655	141,364	5.14	13 / 25
GES, scenario 5	16	84	1,951	1	6 / 25

Option B minimizes the sum of relative variances of all variables of interest subject to cost constraints. Option B is a more natural option since precision is not fixed but the budget is assumed to be known, which often is the case in reality. Because the PSU sample sizes were different in each scenario, the target budget ranges were varied (roughly in proportion to PSU sample sizes expected) as well. Table 2 below shows the results for option B. Because the program solves to minimize the sum of relative variances subject to a cost constraint in option B, the number of feasible estimates are not displayed as all designs would satisfy requirements for all variables overall.

Table 2: GES Option B Result

GES Scenario	*PSUs*	*PJs*	*PARs*	*Cost*	*Objective Function Value*
GES, scenario 2	75	618	28,772	2.63	0.204073438
GES, scenario 3	51	425	18,170	1.95	0.233077041
GES, scenario 4	24	207	8,936	1.21	0.536037246
GES, scenario 5	16	146	6,294	1	0.748608433

Table 3 gives an example of a sample design optimization system performance report for GES. The actual table displays the variance, target variance and difference in variance as well as the ratio, but only the ratio is presented here. This particular table shows the decision variable results, and the ratio of resulting variance and target variance for a subset of variables of interest in scenario 3 option B. A value smaller than 1 represents improvement in precision for the specific variable of interest, relative to the current GES. From the table, this design achieves better precision for many key estimates, which are given dummy names VAR1, VAR2 and so on. The cost is shown on a relative scale compared to the other option B scenarios.

Table 3: GES Performance Report Example

Study Type	GES
Scenario	3
Optimization Option	B
PSUs (n =)	51
SSUs / PJs (nm =)	425
PARs (nmk =)	18,170

Cost	1.95
Variable of Interest	*Ratio of Result vs Target Variance*
VAR 1	0.91798
VAR 2	5.77203
VAR 3	1.55596
VAR 4	0.56028
VAR 5	0.62536
VAR 6	0.79676
VAR 7	0.38302
VAR 8	0.44488
……	……
VAR 23	0.72144
VAR 24	0.56233
VAR 25	1.46286

4.2 FOPV

The FOPV results showed the same theme as the GES results. In option A, one variable of interest seems to govern the allocation, and because precision requirements were based on the 2011 CDS with 24 PSUs, some precision requirements cannot be met if the number of PSUs in a new design is too small. One may wish to use the table below as a guide. For example, it can be viewed as a way to narrow down sample designs given their budget and advantages and disadvantages of each design.

Table 4: FOPV Option A Result

FOPV Scenario	*PSUs*	*PJs*	*PARs*	*Cost*	*Number of Feasible Estimates*
FOPV, scenario 0.5	73	342	6,484	2.74	20 / 20
FOPV, scenario 1	49	286	5,359	2.26	18 / 20
FOPV, scenario 2	40	282	5,545	2.27	18 / 20
FOPV, scenario 3	32	179	3,275	1.56	18 / 20
FOPV, scenario 4	24	137	2,475	1.3	17 / 20
FOPV, scenario 5	16	97	1,675	1	12 / 20

Option B minimizes a sum of relative variances subject to cost constraints and is a more natural fit for FOPV as well because the budget is usually known and precision is not fixed. The budget ranges were varied per scenario due to varying PSU sample sizes. Again, the program solves to minimize the sum of relative variances subject to a cost constraint in option B. Therefore, the number of feasible estimates are not displayed as all designs would satisfy requirements for all variables overall.

Table 5: FOPV Option B Result

FOPV Scenario	*PSUs*	*PJs*	*PARs*	*Cost*	*Objective Function Value*
FOPV, scenario 0.5	73	668	13,418	3.61	0.120866357
FOPV, scenario 1	49	477	8,938	2.51	0.11340313
FOPV, scenario 2	40	389	7,221	2.1	0.120114667
FOPV, scenario 3	32	319	5,719	1.73	0.122451192
FOPV, scenario 4	24	239	4,171	1.32	0.176273714
FOPV, scenario 5	16	157	2,771	1	0.323821495

Table 6 gives an example of a sample design optimization system performance report for FOPV scenario 5 option B. The setup is the same as GES, and the ratio smaller than 1 represents improvements in precision. This particular design improved the precision for many variables, but because the variance targets were set using past CDS data with 24 PSUs, precision requirements for some of the individual key estimates could not be met. Overall, however, our optimization results gave better precision with a lower cost.

Table 6: FOPV Performance Report Example

Study Type	FOPV
Scenario	5
Optimization Option	B
PSUs (n =)	16
SSUs / PJs (nm =)	157
PARs (nmk =)	2,771
Cost	1
Variable of Interest	*Ratio of Result vs Target Variance*
VAR 1	0.4341
VAR 2	2.42929
VAR 3	0.55632
VAR 4	1.97986
VAR 5	1.07493
VAR 6	0.10359
VAR 7	0.1917
VAR 8	0.33281
......	......
VAR 18	4.15731
VAR 19	4.69582
VAR 20	0.42228

4.3 Additional Results and Decisions

NHTSA selected one scenario for each of the GES and FOPV for initial release and recruitment. Using the solutions from our optimization system as a guide, further modifications were made to the second stage sample size while still employing the 2nd stage stratification as well as the results for the 1st and 3rd stages.

5. Discussion

There are some practical considerations to be mindful of. As mentioned in previous sections, the results need interpretation in light of the assumptions, constraints and features of each run. The optimization system only considers the estimates and constraints provided by the user, and ignores all other potential estimates and constraints. Hence, not only is it crucial that the user include all variables of interest and constraints appropriate for their needs, but also, one must keep in mind that the estimate for any unspecified subgroups may only be protected by having a larger total sample size than suggested by the optimization results.

6. Conclusion

In modernizing NASS, a flexible optimization system which solves a three stage multivariate sample design optimization problem in a reasonable run time was needed. A single macro using SAS/OR® software was developed for this purpose which accommodates all combinations of GES/FOPV, optimization options and flexibility scenarios. Of the two optimization options, option B which minimizes a sum of relative variances while respecting budget constraints is more realistic. If desired, more constraints can be added in the future, and the user may adjust inputs and parameters according to their needs.

This multivariate optimization system can be used for other three stage multivariate sample design optimization problems and it returns sample sizes at three stages of sampling, and cost and variance resulting from those sample sizes. The result then in turn can provide guidance for the overall sample design.

Acknowledgements

The co-authors and reviewers of this paper consisted of James Green, Sharon Lohr, Rui Jiao, Barnali Das and Yumiko Sugawara. We would like to extend our gratitude and appreciation to the NHTSA staff involved in NASS survey modernization project.

References

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, *15*(1), 47-57.

Causey, B. D. (1983). Computational aspects of optimal allocation in multivariate stratified sampling. *SIAM Journal on Scientific and Statistical Computing*, *4*(2), 322-329.

Díaz-García, J. A., and Cortez, L. U. (2008). Multi-objective optimisation for optimum allocation in multivariate stratified sampling. Survey Methodology, Vol. 34, No. 2, pp. 215-222.

Huang, T. and Hughes, E. (2010). Nonlinear optimization in SAS/OR® software: Migrating from PROC NLP to PROC OPTMODEL. Paper presented at SAS Global Forum 2010. Last accessed 9/19/2014 from https://support.sas.com/resources/papers/proceedings10/242-2010.pdf.

Green, J. L. (2000). Mathematical programming for sample design and allocation problems. In *Proceedings of the American Statistical Association, Section on Survey Research Methods,* 688-692.

Green, J. L., Baskin, B., and Lee, KC. (2001). Sample redesign for the drug abuse warning network (DAWN). In *Proceedings of the American Statistical Association, Section on Survey Research Methods.* Last accessed 9/22/2014 from https://www.amstat.org/sections/srms/Proceedings/y2001/Proceed/00508.pdf.

Khan, M. G. M., Chand, Munish A., and Ahmad, Nesar. (2006). Optimum allocation in two-stage and stratified two-stage sampling for multivariate surveys. In *Proceedings of the American Statistical Association, Section on Survey Research Methods, 3215-3220.*

Rao, T. J. (1993). On certain problems of sampling designs and estimation for multiple characteristics. *Sankhyā: The Indian Journal of Statistics, Series B*, 372.

SAS Institute Inc. (2011). *SAS/OR® 9.3 User's Guide: Mathematical Programming.* Cary, NC: SAS Institute, Inc.

Valliant, R., & Gentle, J. E. (1997). An application of mathematical programming to sample allocation. *Computational Statistics & Data Analysis*, *25*(3), 337-360.