Transforming Freight Flow Data Collection

**Milestone #2**

Adaptive Sampling Literature Review and Proposed Methodology

Saeed Ghanbartehrani, Sara Akbar Ghanadian, Hoda Rahmani

Ohio University

Athens, OH

Rev. 01 June 2019

Contents

[1. Introduction 3](#_Toc534760225)

[2. Stratified Sampling 3](#_Toc534760226)

[3. Adaptive Sampling 4](#_Toc534760227)

[4. Related Literature 5](#_Toc534760228)

[4.1. Stratification 5](#_Toc534760229)

[4.2. Allocation 8](#_Toc534760230)

[4.3. Joint Stratification-Allocation 9](#_Toc534760231)

[5. Proposed Methodology 10](#_Toc534760232)

[6. Comparison of the Candidate Methodologies 12](#_Toc534760233)

[7. References 14](#_Toc534760234)

# Introduction

The term ‘sampling’ has been used to refer to situations in which a fraction of a population is taken to be representative of the entire population. In other words, sampling is the process of observing selected members to approximate characteristics of the whole population from which they are drawn. By the same token, sampling design is defined as the methodology by which the sample units are chosen (Thompson, 2012). Some common sampling methods are simple random sampling, stratified sampling, and cluster sampling. Amongst those techniques, stratified sampling is the primary focus of this document.

# Stratified Sampling

Stratified sampling is a strategy in which the population is divided into several non-overlapping sub groups referred to as stratum from which the samples are selected (Al-Kateb & Lee, 2014). The main objective of this kind of sampling design is to acquire a sample with the desired level of accuracy while decreasing the sampling error. Stratified sampling can also be used to obtain a smaller sample with maintaining the desired level of accuracy. The variable that the study seeks to measure is assumed to be homogeneous among different strata in stratified sampling (Manly, 2004). Despite the fact that the variable of interest is anticipated to be similar, there might be some disparities across different strata. These potential inconsistencies are rooted in the nature of the sampling design since samples are selected independently in each stratum. To find the variances of estimators for the whole population, the variances of estimators for each stratum can be calculated and added up (Thompson, 2012).

There are two questions that need to be answered in a stratified sampling design (Basoglu, 2014): (i) How the population should be partitioned into a specific number of strata (stratification), (ii) How the sample size for each strata is determined (allocation).

There are two approaches to stratification and allocation in the literature. The first approach solves the problems of stratification and allocation in two separate phases. Stratification can be performed based on various variance reduction techniques such as Naive Monte Carlo Simulation, importance sampling, and stratified sampling, which will be discussed in ‎4.1. When the strata are created, the allocation can be determined using an allocation fractions technique. Allocation techniques are reviewed in ‎4.2.

The second approach employs techniques to solve the problem of stratification and allocation in one phase as a joint stratification-allocation method. Joint stratification-allocation techniques are reviewed in ‎4.3.

# Adaptive Sampling

One significant concept in sampling design is the procedure that alters as the experiment progresses, referred to as ‘adaptive sampling’. Adaptive sampling enables the model to learn from the gathered data during the survey and to be modified accordingly (Lermusiaux, 2007). As a result, adaptive sampling approach can contribute to a more comprehensive understanding of the target population. One of the most remarkable current discussions in adaptive design is adaptive stratification. This strategy seeks to find the strata in which the variability in the variables of interest is abundant. Hence, by collecting more samples from such subgroups in which the variations are inconsistent with the rest of the subgroups (i.e. have significantly larger variability compared to the others), one can have samples that better reflect the larger population (Carpentier & Munos, 2013).

# Related Literature

In this section, related literature in the areas of stratification, allocation, and joint stratification-allocation is reviewed.

## Stratification

A number of techniques have been developed to determine the optimal number of strata along with an efficient allocation approach. Carpentier & Munos (2013) proposed an algorithm referred to as Monte-Carlo Upper Lower Confidence Band (MC-ULCB) which adapts the Monte-Carlo integration algorithm in the stratification methodology. The objective function of this algorithm is based on a ‘noisy function’ which refers to a function involving a random component (i.e. noise). Therefore, the adaptive strategy aims at discovering the areas where the variation of the noisy function is bountiful besides finding the best partitioning method for the sampling procedure. Once the samples are extracted, the algorithm proceeds to get some parameters related to the upper and lower confidence bounds on the variability of the noisy function in each stratum since the function is assumed to be bounded (Carpentier & Munos, 2013). The variability of the variable of interest in each partition is measured in the form of standard deviation. Following this process, the samples are allocated to each stratum proportional to the variability of each stratum multiplied by its upper confidence band. In this study, the authors concluded that the proposed model can obtain more homogenous samples from each stratum as well as maintaining the number of strata it contains as small as possible (Carpentier & Munos, 2013).

In another study, Glasserman et al. (1999) proposed a procedure aiming to reduce the variability in Monte-Carlo simulation using stratification technique. In this scheme, the authors developed a methodology in which some samples are taken from a population that is considered to have a standard normal distribution. In this reserach, the Monte-Carlo technique is performed as well as stratification with a vector along different directions which are defined by the conditional covariance matrix (Glasserman et al., 1999). In order to stratify in different dimensions of the matrix, each dimension is split into multiple intervals with equal width. After obtaining the intervals, a point is uniformly selected from each interval. Finally, samples are selected by calculating the inverse of the cumulative normal distribution for each dimension (Glasserman et al., 1999).

In a study conducted by Yong et al. (2016), the authors attempted to establish an optimal stratification method in the field of medical science using the stratum specific mean. The researchers highlight the importance of employing the baseline information of previous patints in order to predict a variable of interest which is the probability of subject’s response to a certain medicine or procedure. To begin this process, a dataset consisting of the subjects’ predicted values is created. This dataset can provide the researchers with a scoring system. Once the dataset is generated, a regression model is used to associate predicted values to their corresponding actual results. Stratification process is incorporated into this approach to predict future outcomes for the patients. To put it simply, scores are classified into several strata and the average score is calculated as a mean for anticipating the results of future individuals. The authors propose a scheme that can find the best stratifying method which minimizes the prediction error using a loss function.

Etore and Jourdain (2010) presented an adaptive stratified sampling algorithm in which the randomly selected samples from strata converge to an optimal allocation. In their study, the authors claim that the Monte-Carlo estimator in stratified sampling can be used to find the expectation of a target function (of the drawings). It is assumed in the model that the probability of each random variable under study is known and follows the normal distribution. As a result, the expectation of the target function can be computed through multiplying the expectation of each stratum by its probability. In fact, the number of total samples drawn from all strata and the proportion of extracted samples from each stratum are formulated in terms of the expectation of interest. In the same way, the variance can be calculated given the conditional expectations. This framework can contribute to variance reduction if the proportion of samples taken from each stratum is computed appropriately.

Basoglu (2014) applied variance reduction techniques to reduce the size of confidence intervals generated by Monte Carlo simulation used in computing financial risk involved in realistic and complex portfolio models. In this study, an efficient implementation of stratified sampling technique for Monte Carlo simulation problems referred to as Optimal Allocation Stratification and Importance Sampling (OASIS) is proposed. The proposed approach involves an efficient simulation algorithm that combines optimal stratification and importance sampling to estimate multiple conditional loss and gain probabilities for asset portfolios.

Two classes of objective functions are proposed to represent the overall error. The first class of error function minimizes a linear function of the variance-covariance matrix of the stratified estimates. The second class minimizes the maximum of variances weighted with non-negative coefficients. Both objective functions are used in nonlinear optimization models with allocation fractions as decision variables. A closed-form solution is developed for the first class of objective functions. For the second class, an optimal allocation heuristic is utilized to find a near optimal solution. Solutions from these models are used in the sampling phase to minimize quantities such as the mean-squared (relative) error or the maximum absolute (relative) error that represent the overall error of the simulation. The idea of the OASIS algorithm can be used to minimize the overall error of an arbitrary simulation associated with multiple estimates. The numerical results show that the OASIS algorithm is an efficient and flexible method for simulation problems for which we can find efficient stratification functions.

## Allocation

The optimal allocation is a sample allocation method used with stratified sampling, which is designed to provide the best precision (lowest variation) for the least cost (least sample size) (Statistics Dictionary, 2019b). Neyman allocation is a special case of optimal allocation when the total sample size is fixed (Statistics Dictionary, 2019a).

Applying Neyman allocation has been investigated in researches including (Lavallee & Hidirogloui, 1988), (Benedetti et al. 2010) and (Benedetti & Piersimoni, 2012a). Collectively, these methods restrict the number of strata into two or three. In Lavallee & Hidirogloui (1988) method, population is divided into two strata, one of which was used as take-all stratum while the other one was sampled stratum.

In conventional stratified sampling, the fraction of samples to be allocated to strata is typically decided after the stratification is determined, and the focus is on the minimization of variance of the final stratified estimators. However, the optimal allocation in adaptive stratified sampling can be carried out without variance reduction within each stratum (Kawai, 2010). Etore, et al. (2011) proposed an iterative adaptive optimal allocation algorithm. In each iteration, the algorithm adjusts the proportion of further drawings by applying conditional standard deviation estimates. These proportions converge to the optimal allocation fractions. In this method, at least one drawing is allocated to each stratum which is similar to the method proposed in (Etore & Jourdain, 2010) discussed in ‎4.1 which leads to suboptimal allocations in initial iterations.

## Joint Stratification-Allocation

Benedetti & Piersimoni (2012b) proposed a multivariate framework as the extension of Hidiroglou (1986) univariate method discussed in ‎4.2. The size of each strata is defined by a set of univariate thresholds for each auxiliary variable present in the sampling frame. Univariate thresholds make the strata to have “box-shaped” boundaries. However, The multivariate framework lifts the limitation of box-shaped partition boundaries by applying a random search algorithm and using simulated annealing to solve a general combinatorial optimization problem (Lisic et al., 2018).

Barcaroli (2014) proposed an optimal stratification and allocation method to minimize the cost while all precision constraints are satisfied. Their method is used for multivariate cases in which the estimate of target variables in strata are available. Since the number of possible alternative stratifications is high, Genetic algorithm is applied to find near optimal stratification in specific iteration.

Lisic et al. (2018) proposed another optimal stratification and allocation method and used simulated annealing to solve the optimization model.

# Proposed Methodology

Two of the methodologies reviewed in the literature were selected to be considered as candidate proposed methodologies. The selected methodologies use an optimization model along with constraints to optimize both stratification and allocation with respect to budget and precision constraints. Both methods have readily available code implemented in R applied to case studies involving real surveys.

The first candidate methodology is proposed by Lisic et al. (2018) which is an optimal stratification and allocation method based on simulated annealing that considers coefficient of variance and fixed sample size constraints. This methodology was developed to create an optimal sample design for the June Area Survey (JAS) under quality (coefficient of variance) and sample size constraints. The JAS is one of the largest annual National Agricultural Statistics Service (NASS) agricultural area survey projects over the contiguous 48 states designed to account for every acre of land, all agricultural activities, and land uses within segment boundaries (National Agricultural Statistics Service, 2018).

The proposed methodology uses an objective function composed of the sum of penalties of deviations from the target CVs as a proxy for quality constraints. This is a soft constraint since it does not prohibit the model from deviating from the target values. On the other hand, hard constraints are introduced to the model through defining nonlinear constraints.

In order to minimize the objective function, simulated annealing heuristic is used. Simulated annealing is a stochastic optimization process that allows the objective function to explore some nonoptimal states with nonzero probabilities. The iterative process starts with a feasible initial stratification and allocation. A primary sampling unit (PSU) is exchanged in each state. In the same way, allocation is performed by choosing a stratum to accept the PSU. The sample size of the stratum which accepts the PSU is increased by 1 and the sample size of the one that loses the PSU is decreased by 1. The algorithm stops after a specified number of iterations or when the threshold is met. In each iteration, a candidate state is randomly generated. Then, a candidate allocation with regards to the new state is created. The inner loop of the algorithm checks whether the new combination of candidate state and allocation improves the objective function.

The next candidate methodology proposed by Barcaroli (2014), is an optimal stratification and allocation method aiming at minimizing the sample cost while satisfying a set of precision constraints. Also, the value of target variables is assumed to be either available in the frame or it is possible to estimate their standard deviation and mean from the same or a previous round of the same survey.

The process initiates with the analysis of the frame data. First, auxiliary variables are identified from current variables. In case the values of auxiliary variables are continuous, they must be converted into categorical variables using the k-means clustering technique. Atomic strata are constructed and characterized based on the categorical auxiliary variables and distributions of the target variables inside the different strata. Then, precision constrains on the target estimates are constructed. These precision constraints are differentiated by domain values. Bethel algorithm is implemented to determine the required number of units to be selected which needs to be reduced in optimization of stratification later. Once the strata and constraints data frames have been prepared, the frame stratification is optimized, and the required sample size and allocation to satisfy the precision constraints are determined. The resulting optimized strata are then analyzed and new labels are assigned to the sampling frame units. Each label reflects the new strata resulting from the optimal aggregation of the atomic strata. Finally, units are selected from the sampling frame based on stratified random sample selection scheme and the optimal solution is evaluated in terms of expected precision and bias.

# Comparison of the Candidate Methodologies

In the model proposed by Lisic et al. (2018) the number of strata and also the total sample size are assumed to be fixed. In Barcaroli's (2014) model, the number of units in each stratum as well as the maximum of the coefficient of variation (CV) for each variable are required. Based on the provided input variables, unknown population characteristics are estimated for both models using Horvitz-Thompson estimator. Moreover, the continuous administrative data is transformed into the categorical data in both sampling methods.

The objective function for the Lisic et al.’s methodology consists of two terms; the first term is the weighted norm vector of modeled CVs and the second term is the penalty function for violating the constraints. On the other hand, in the Barcaroli’s objective function, a fixed cost is added to the summation of the products of cost of interviewing each unit and the cost of allocation. Both models take the same approach in defining the constraints. Constraints are on CV and budget. In both methodologies, K-means clustering algorithm is used in the stratification stage and Bethel's (1989) method is employed in the allocation framework.

In Lisic’s et al. method, simulated annealing algorithm is used for optimizing the objective function. The algorithm stops after a given number of iteration or when a threshold is met. Although Barcaroli used genetic algorithm as the solution method, the algorithm terminates after a specific number of iterations or when the value of the objective function reaches a given minimum. Both models rely on historic data as their administrative data variables along with other source of data.

# References

Al-Kateb, M., & Lee, B. S. (2014). Adaptive stratified reservoir sampling over heterogeneous data streams. *Information Systems*, *39*, 199–216. https://doi.org/10.1016/j.is.2012.03.005

Barcaroli, G. (2014). SamplingStrata: An R Package for the Optimization of Stratified Sampling. *Journal of Statistical Software*, *61*(1), 1–24. https://doi.org/10.18637/jss.v061.i04

Basoglu, ismail. (2014). *ismail\_basoglu\_phd\_thesis.pdf*. Bogazici University. Retrieved from http://www.ie.boun.edu.tr/~hormannw/BounQuantitiveFinance/Thesis/ismail\_basoglu\_phd\_thesis.pdf

Bendeti, R., Bee, M., & Espa, G. (2010). A Framework for Cut-off Sampling in Business Survey Design, *26*(4), 651–671.

Benedetti, R., & Piersimoni, F. (2012a). Multivariate Boundaries of a Self Representing Stratum of Large Units in Agricultural Survey Design, *6*, 125–135.

Benedetti, R., & Piersimoni, F. (2012b). Multivariate Boundaries of a Self Representing Stratum of Large Units in Agricultural Survey Design. *Survey Research Methods*, *6*(3), 125–135. https://doi.org/10.18148/srm/2012.v6i3.5127

Carpentier, A., & Munos, R. (n.d.). Toward Optimal Stratification for Stratified Monte-Carlo Integration, 9.

Etore, P., Fort, G., Jourdain, B., & Moulines, E. (2011). On adaptive stratification. *Annals of Operations Research*, *189*(1), 127–154. https://doi.org/10.1007/s10479-009-0638-9

Etore, P., & Jourdain, B. (2010). Adaptive optimal allocation in stratified sampling methods. *Methodology and Computing in Applied Probability*, *12*(3), 335–360. https://doi.org/10.1007/s11009-008-9108-0

Glasserman, P., Heidelberger, P., & Shahabuddin, P. (1999). Asymptotically Optimal Importance Sampling and Stratification for Pricing Path-Dependent Options. *Mathematical Finance*, *9*(2), 117–152. https://doi.org/10.1111/1467-9965.00065

Bethel, J. (1989). Sample allocation in multivariate surveys, *15*, 47–57.

Kawai, R. (2010). Asymptotically optimal allocation of stratified sampling with adaptive variance reduction by strata. *ACM Transactions on Modeling and Computer Simulation*, *20*(2), 1–17. https://doi.org/10.1145/1734222.1734225

Lavallee, P., & Hidirogloui, M. A. (1988). On the Stratification of Skewed Populations, 12.

Lermusiaux, P. F. J. (2007). Adaptive modeling, adaptive data assimilation and adaptive sampling. *Physica D: Nonlinear Phenomena*, *230*(1–2), 172–196. https://doi.org/10.1016/j.physd.2007.02.014

Lisic, J., Sang, H., Zhu, Z., & Zimmer, S. (2018). Optimal Stratification and Allocation for the June Agricultural Survey. *Journal of Official Statistics*, *34*(1), 121–148. https://doi.org/10.1515/jos-2018-0007

Manly, B. F. J. (2004). Using the bootstrap with two-phase adaptive stratified samples from multiple populations at multiple locations. *Environmental & Ecological Statistics*, *11*(4), 367–383.

National Agricultural Statistics Service. (2018, November 9). USDA - National Agricultural Statistics Service - Surveys - June Area. Retrieved January 8, 2019, from https://www.nass.usda.gov/Surveys/Guide\_to\_NASS\_Surveys/June\_Area/

Statistics Dictionary. (2019a). Neyman Allocation: Definition. Retrieved January 8, 2019, from https://stattrek.com/statistics/dictionary.aspx?definition=neyman\_allocation

Statistics Dictionary. (2019b). Optimum Allocation: Definition. Retrieved January 8, 2019, from https://stattrek.com/statistics/dictionary.aspx?definition=optimum\_allocation

Thompson, S. K. (2012). *Sampling* (3rd ed). Hoboken, N.J: Wiley.