Alternative Optimization Techniques for Sample Allocation in Surveys with National and Sub-National Precision Requirements*

John Chesnut[†] and Shawn Baker [‡]

Abstract

Optimal sample allocation can improve the precision of estimates for surveys with fixed budgets. Demographic surveys at the Census Bureau often have precision requirements for estimates at both the national- and sub-national-level. This results in a hierarchical relationship between the national- and sub-national precision requirements. For example, for a fixed budget, sample is allocated such that the sub-national requirements are met then the balance of the remaining sample is allocated such that the national-level requirement is satisfied. The focus of this paper is to explore optimization techniques that satisfy this hierarchical sample allocation problem. Using the Current Population Survey, this research compares four methods for allocating sample units that simultaneously meet the state- and national-level precision requirements. The four methods considered are non-linear optimization, a linear programming algorithm, a maximum sampling interval step reduction, and a greedy heuristic. The results show that all of the methods are capable of satisfying the design requirements with the greedy heuristic resulting in the most efficient allocation for meeting the national-level precision requirement. However, the nonlinear optimization can provide a less greedy sample allocation across states.

Key Words: sample allocation, sample size, precision, nonlinear optimization, greedy heuristic

1. Introduction

Determining a sample allocation for multi-purpose surveys that minimizes sampling error across multiple characteristics and domains of interest for a fixed budget or vice versa has been a point of discussion as early as the 1930's (cf. Neyman, 1934). Neyman allocation is specific to the univariate case however researchers have attempted to adapt this method to the multi-variate case with less than favorable results (Huddleston, Claypool, & Hocking, 1970). In addition, relying on Neyman allocation to provide estimates for multiple domains of interest can produce insufficient precision for small domains (Särndal, Swensson, & Wretman, 1992).

More recently, the literature converges on two general approaches for addressing the problem of optimal sample allocation for multi-purpose surveys. One approach is to define an objective function that is a weighted average of the variances of the characteristics and domains of interest. For example, Valliant and Gentle (1997) developed a flexible approach for two-stage sample allocation that uses a constrained, multi-criteria optimization programming. The objective function is an importance weighted function of the relvariances and cost. A criticism of this approach is that the solution may not be optimal due to the arbitrary choice of importance weights.

[*] This article is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau. DAO approval # CBDRB-FY19-POP001-0073

[†] U.S. Census Bureau, Washington, DC 20233

[‡] Marine Corps Operational Test and Evaluation Activity, Quantico, VA 22134

Alternatively, a survey may establish precision constraints for each of the characteristics and domains of interest. A convex objective function is used in conjunction with the precision constraints. For example, Bethel (1985) developed an algorithm or heuristic for determining the sample allocation for the multi-variate case that minimizes an objective function while satisfying precision constraints for the variables of interest. In this same vein, Srikantan (1963) utilizes nonlinear programming to determine an optimal allocation for the problem of satisfying precision level requirements for multiple domains of interest. Srikantan's method addresses the multi-domain problem which is a focus of our research, however our allocation problem also involves precision requirements for multiple estimate types, e.g. totals and ratio estimates. Additionally, these methods can provide an optimal allocation satisfying the precision constraints but may result in complex analytical solutions requiring numerical methods.

2. Background

2.1 Sample Redesign

This research falls under the umbrella of the Sample Redesign Research Program at Census where the goal of this program is to develop innovative cross-cutting improvements to the sample designs for the following major demographic surveys.

- Current Population Survey (CPS) - labor force characteristics of the U.S. population
- Survey of Income and Program Participation (SIPP) - income and government program participation of individuals and households in the U.S.
- National Crime Victimization Survey (NCVS) - characteristics and consequences of criminal victimization in the U.S.
- Consumer Expenditure Surveys (CES) - two surveys that characterize the buying habits of American consumers
- American Housing Survey (AHS) - collects data on the Nation's housing stock, household characteristics, housing and neighborhood quality, housing costs, and recent movers

These surveys focus on various characteristics of the population and domains of interest, however they face similar challenges in designing samples that meet the goals of the survey. A critical design choice common across the surveys is determining an optimal sample allocation that provides support for desired precision levels. To address this common goal, this research attempts to answer the question of whether the household surveys with multiple objectives are meeting their objectives in an optimal manner. As a first step, we consider the sample allocation problem for the CPS with multiple domain precision requirements at the national- and state-level. We want to compare alternative allocation methods that satisfy the CPS design requirements.

2.2 Current Population Survey - Sample Design Requirements

To address the allocation problem for the CPS, we first examine its design requirements as stated in Technical Paper 66 (2006) and attempt to frame the problem using methods in the literature. The main function of the CPS design is to produce national- and state-level estimates of the labor force characteristics - specifically unemployment. The official precision requirements require that the sample enable detection of a 0.2 percent month-to-month change in the unemployment rate and that the state-level coefficients of variation (cv) for the annual average monthly unemployment level should not exceed 8 percent. So, this is a multi-domain allocation problem, with global and domain-level precision requirements for a global ratio estimate and domain totals. Because of the mixed use of ratios and totals in the precision requirement, writing the national-level precision requirement as a function of the state level precision requirements is not straight forward. Additional soft constraints require an approximately self-weighted design, reliability for other labor-force characteristics, and a budget constraint of 60,000 sample units.

3. Methodology and Results

3.1 Defining the Sample Allocation Problem

Framing the CPS allocation problem as an optimization problem presents a number of complexities due to the survey design and precision requirements. For example, the national- and state-level precision requirements are defined using different estimate types - the monthly rate of unemployment at the national level and the average monthly unemployment total at the state level. In addition, since the CPS is a longitudinal survey, we need to account for the correlation structure that exists due to repeated measures of sample units.

3.1.1 Objective Function

To define an objective function for allocation, we look to previous work by Rottach and Erkens (2012) where they develop a model that relates the national and state-level design requirements. Both requirements are converted into cv requirements for monthly unemployment totals. The first step in formulating the model is to approximate the cv of the monthly unemployment total $\hat{Y}_t$ for a given month t using the linearization of the the cv of the unemployment rate $\hat{\hat{Y}}_t$. Generically speaking, for a given ratio A/B, we can write the linearization $cv^2(A/B) \cong cv^2(A) - cv^2(B)$. Applying this to the unemployment rate and assuming the coefficient of variation of the civilian labor force (CLF) is negligible, we can write $cv(\hat{\hat{Y}}_t) = cv(\hat{Y}_t)$.

Then, by assuming the estimates for the monthly state employment totals $\{\hat{Y}_{t,s}\}_{s \in States}$ are independent and that the unemployment rates are equal, we can write that the cv^2 of the national unemployment level is equal to the weighted average of the cv^2 values of the state unemployment levels.

$$cv^2(\sum_S \hat{Y}_{t,s}) = \sum_S p_s^2 cv^2(\hat{Y}_{t,s}) \text{ where } p_s = CLF_s/CLF.$$

Furthermore, given direct estimates of the current cv values and leveraging the inversely proportional relationship $cv^2 \propto \frac{1}{n}$, we can equate the new cv to the ratio adjusted observed cv using the current and new sampling intervals (SI) in the ratio adjustment, i.e.,

$$cv^2_{new}(\hat{Y}_{t,s}) = \frac{SI_{new,s}}{SI_{current,s}} cv^2(\hat{Y}_{t,s})$$

Substituting this into our previous function, cv^2 of the national unemployment level equals the weighted average of the SI ratio adjusted state cv^2 values.

$$cv^2(\sum_S \hat{Y}_{t,s}) = \sum_S \left(\frac{CLF_{new,s}}{n_{new,s}}\right)\left(\frac{1}{SI_{current,s}}\right) p_s^2 cv^2(\hat{Y}_{t,s})$$

Expanding the sampling interval under the new design, the decision variables in our objective function are the set of new state sample allocations $\{n_{new,s}\}_S$.

3.1.2 Official Constraints

Using the previous model, Rottach and Erkins (2012) translate the official design requirements into precision requirements for national- and state-level monthly unemployment totals. Deriving the national-level constraint first, the CPS uses a repeated measures design, so the correlations between estimates of unemployment for subsequent months need to be determined. To do this, Rottach and Erkins were able to use a modelling approach to determine the average correlation between unemployment rates for subsequent months. Using the correlation value of 0.41 and assuming a 6 percent unemployment rate and assuming a significance level of 10 percent, they compute that the cv required for the national monthly unemployment total is 1.87 percent.

Next, given the official precision requirement - a state cv for the annual average monthly unemployment totals can not exceed 8 percent - Rottach and Erkins transform this into maximum cv constraints for the state monthly unemployment totals. Utilizing direct estimates of the state between and within PSU variances and model-based estimates of the between and within-month correlation factors over a 12 month period ($\rho_b = 0.71$ and $\rho_w = 0.20$), they derive the following equation for computing the individual state upper bounds for the cv of monthly unemployment.

$$cv^2(\hat{Y}_s) = \frac{.08^2}{0.71\alpha_s + 0.20(1-\alpha_s)} \text{ where } \alpha_s = \frac{V_{b,s}(\hat{Y})}{V_s(\hat{Y})}$$

3.1.3 Soft Constraints

In addition to its official constraints, the CPS sample design also includes soft constraints in attempt to satisfy the needs of achieving reliability for other characteristics of interest, producing an approximately self-weighting sample, and satisfying an approximate fixed budget of 60,000 sample cases. To meet the reliability constraint, a minimum sample size is established across the states. In addition, to attempt to come closer to a self-weighting sample design, a ceiling is established for state sampling intervals to limit the range of values for the state sampling intervals.

The surface plot in Figure 1 illustrates the nonlinear relationship of our soft constraints (maximum sampling interval, minimum sample size, and budget) with the national-level cv. The color gradient represents the change in budget or overall sample size. Clearly, this plot demonstrates that a region of values for the soft constraints exist such that the national level cv requirement is satisfied. Therefore, our choice of soft constraints that meet the desired national precision level is not unique, thus allowing the allocation methods we discuss later to relax or tighten these constraints to enable feasible solutions.

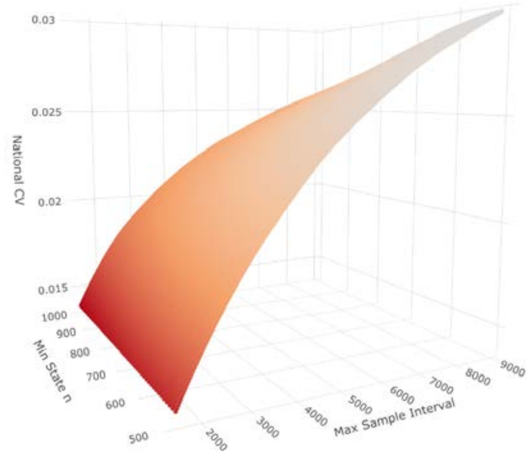


Figure 1: Soft Constraints and National Precision

3.2 Sample Allocation Methods

Having framed the CPS sample allocation problem as an optimization problem with an objective function and constraints, we are now able to apply various optimization methods and compare the performance of these methods in meeting our design constraints. The following list provides a brief description of each method.

- Nonlinear Optimization - constrained optimization using the Augmented Lagrangian algorithm along with the Method of Moving Asymptotes sourced from the NLopt library (Johnson, 2019)
- Maximum Sampling Interval (Max SI) - iteratively decreases a 'ceiling' for the state sampling intervals to prioritize reducing the range of sample weights across states
- Linear Programming (LP) algorithm - sampling intervals defined as the decision variables - iterative algorithm, nonlinear budget constraint requiring computation/checking at each step
- Greedy Heuristic - iteratively adds an additional sample to the stratum with the largest reduction in variance

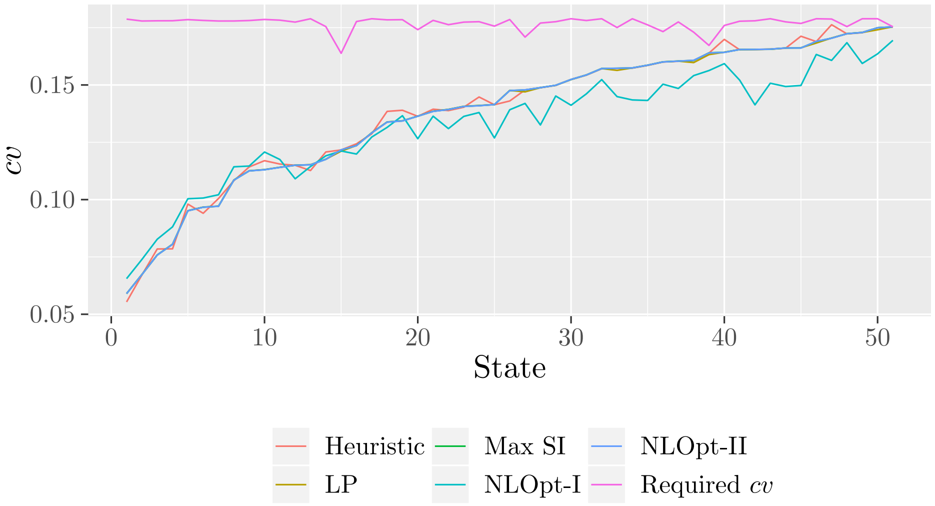
3.3 Comparison of Methods

Table 1 shows for each allocation method, the range of state sample intervals, the national precision or *cv* result for a fixed budget of 60,000 sample units, and the minimal sample size necessary to attain a fixed *cv* equal to 1.9. As a benchmark, we include the results where only the state-level *cv* requirements are satisfied, ignoring the national-level requirement. With the exception of our benchmark and the NLOpt-I method, all methods meet the national-level precision requirement for a fixed budget. The Heuristic method appears to perform the best at meeting the national-level precision requirement with a slightly higher maximum SI. Note that for the NLOpt-I method, we relaxed the Maximum SI constraint. As a result, we have a wider range in our state sampling intervals - which may indicate we achieved better reliability for some states at the expense of a higher national cv.

Table 1 Sampling Interval, National Precision, and Budget Results by Sample Allocation Method

	Sampling Int.			
Method	Min	Max	cv_{nat}	$n_{cv=1.9}$
Required cv_s	473	25,255	3.752	–
NLOpt-I	369	3,394	1.937	60,000
Max SI	405	2,750	1.873	58,700
LP	400	2,752	1.873	58,800
NLOpt-II	405	2,750	1.872	60,000
Heuristic	405	2,945	1.866	58,500

Figure 2 shows the state-level cv values attained under each method. In addition, we included our baseline of required state cv values. Note that the states were ordered by the cv values attained via the Max SI method. Based on this plot, all of our methods tend to perform greedy sample allocations to states with lower variances to simultaneously meet the state and national-level cv requirements. However, we see that the NLOpt-I does perform a somewhat less greedy allocation indicating a more evenly distributed sample across the states. This results in lower state cv values for the majority of states.



**States are ordered by their cv values attained via the Max-SI method of allocation*

Figure 2: State-Level cv Values by Sample Allocation Method

4. Conclusions

The model developed by Rottach and Erkins (2012) relating the state- and national-level precision requirements works well for simultaneously optimizing on both official requirements for the CPS. Most important, the model allows us to represent the national-level precision as a function of state-level precision values. In addition, this model accounts for the correlational structure that is present in the composite estimators for the national monthly unemployment rate and state annual average monthly unemployment levels. Furthermore, this model proves to be a useful tool for assessing the effectiveness of the current sample allocation.

We do highlight some of the model limitations. The model is complex and may not generalize well to other surveys. For example, model- or empirical-based correlation estimates are required as inputs which may be difficult to derive for other surveys. However, depending on the sample design and estimators of interest, surveys may be able derive a simpler model for determining an optimal allocation for multiple domains. Additional limitations include assuming a fixed first stage sample and relying on direct estimates of variance derived from existing survey data. In addition, the global and domain estimates used in the model derivation are assumed to be equal which may not be the case, for example varying unemployment levels across states. Furthermore, surveys may have multiple characteristics of interest that they need to prioritize - this model solution only considers the univariate case with global and domain precision requirements. Cost may vary across states, whereas we assumed cost were equal across domains. Finally, the model did not account for controls on interviewer workloads.

Comparing across allocation methods, all of the methods tend to converge to a 'greedy' allocation meeting desired precision levels while satisfying constraints on budget, self-weighting design, and reliability. In our view, the choice of allocation method depends on the priorities of the survey. If the national level precision is paramount, the best choice here would be the Greedy Heuristic allocation method. This method results in the lowest variance per unit cost at the national level. However, it does sacrifices some control over the self-weighting properties. If the precision levels for the state estimates are a priority, the Non-Linear Optimization method allows more flexibility within the budget to relax the self-weighting constraint and perform less greedy allocations. As a result, this method provides lower state-level variances per unit cost for the majority of states.

References

Bethel, J. (1985). An Optimum Allocation Algorithm for Multivariate Surveys. In *Proceedings of the Joint Statistical Meetings*, Survey Research Methods Section (pp. 209–212). Alexandria, VA: American Statistical Association.

Huddleston, H., Claypool, P., & Hocking, R. (1970). Optimum Sample Allocation to Strata Using Convex Programming. *Applied Statistics*, *19*, 273–278.

Johnson, S. J. (2019). *NLopt.* http://github.com/stevengj/nlopt. GitHub.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, *97*, 558–625.

Rottach, R., & Erkens, G. (2012). Modeling Variances to Determine Sample Allocation for the Current Population Survey October 2012. In *Proceedings of the Joint Statistical Meetings*, Survey Research Methods Section (pp. 4401–4413). Alexandria, VA: American Statistical Association.

Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer.

Srikantan, K. S. (1963). A Problem in Optimum Allocation. *Operations Research*, *11*(2).

U.S. Census Bureau and U.S. Bureau of Labor Statistics. (2006). *Current Population Survey: Design and Methodology, Technical Paper 66.* Retrieved from `https://www.census.gov/prod/2006pubs/tp-66.pdf`

Valiant, R., & Gentle, J. E. (1997). An Application of Mathematical Programming to Sample Allocation. *Computational Statistics and Data Analysis*, *25*, 337-360.