AN APPLICATION OF MATHEMATICAL PROGRAMMING TO SAMPLE ALLOCATION

Richard Valliant, Bureau of Labor Statistics, James E. Gentle, George Mason University
Richard Valliant, Room 4925, 2 Massachusetts Ave. NE, Washington DC 20212

Key Words: Generalized reduced gradient method, multicriteria optimization, multipurpose survey, two-stage sample.

1. INTRODUCTION

Multipurpose surveys produce estimates for many variables and domains. These multiple products complicate the problem of sample allocation since a given allocation will not be efficient for all estimates. This paper will discuss some preliminary results on an application of multicriteria optimization to the sample design of the Employment Cost Index (ECI) and the Employee Benefits Survey (EBS) conducted by the Bureau of Labor Statistics (BLS). The two programs use the same two-stage sample of establishments and occupations to estimate personnel costs and the percentages of employees receiving various benefits. The estimates are made for the population as a whole and for domains. We cover several topics applicable to many sample surveys — estimation of variance components for a multi-stage design and complex estimator, smoothing of variance component estimates to eliminate inconsistencies, and the use of constrained nonlinear programming to optimize the sample allocation.

Bethel (1989) and Kish (1988) noted the variety of purposes for which a given survey may used and why purposes may conflict. A number of different variables may be measured and estimates may be made for diverse domains, complicating the sample allocation problem. The allocation technique used in this paper is *constrained, multicriteria optimization* described, for example, in Narula and Weistroffer (1989), Steuer (1986), and Weistroffer and Narula (1991). A weighted combination of the relvariances of different estimators is formed with each weight being the "importance" of each statistic in the overall survey design. The weighted combination is minimized subject to a variety of constraints, including one on total cost or sample size, minimum and maximum sample size constraints in each stratum, and relative variance constraints on individual estimators.

2. SAMPLE DESIGN

The sample design involves two stages of selection — establishments at the first stage and occupations at the second. First, a sample of establishments is selected within each stratum with probabilities proportional to total employment in each establishment as shown on the Unemployment Insurance (UI) frame at a particular date. In this study of sample allocation, stratification by SIC and size will be used, though the actual sample design is somewhat more elaborate. At the second stage, a sample of occupations is selected within each sample establishment with probabilities proportional to the number of employees in each occupation during the time period specified for sample selection. This is done by selecting a systematic sample of individual employees from a personnel list in each establishment and enumerating all workers in the occupations held by the selected employees. The occupation sampling procedure is simple to implement in the field but does allow a particular occupation to be selected more than once. This point is discussed further at the end of this section.

In order to proceed, we need some notation. Let h denote a stratum defined by SIC/size and i an establishment within the stratum. Define

π_{hi} = selection probability of establishment hi

n_h = number of sample establishments in stratum h

$\pi_{j|hi}$ = expected number of times that occupation j is selected within establishment hi

$\overline{m}_h$ = number of sample occupations selected in sample establishment hi, assumed to be the same for each sample establishment in stratum h, and

s_h = set of sample establishments in stratum h, and

s_{hi} = set of sample occupations within sample establishment hi.

The quantities π_{hi} and $\pi_{j|hi}$ are general. Specifically for ECI/EBS, if E_{hi} is the number of UI employees in the establishment, then the selection probability of establishment hi is

$$\pi_{hi} = n_h E_{hi} / E_{h\cdot}$$

where $E_{h\cdot}$ is the total frame employment in stratum h. If E_{hij} is the number of employees in occupation j in establishment hi and no occupation has $E_{hij} > E_{h\cdot}/\overline{m}_h$, then the selection probability of an occupation within the establishment is $\pi_{j|hi} = \overline{m}_h E_{hij} / E_{hi}$. The overall selection probability of unit hij is then

$$\pi_{hij} = \pi_{hi}\pi_{j|hi} = n_h \overline{m}_h E_{hij} / E_{h\cdot}.$$

In a case where there are one or more occupations with $E_{hij} > E_{h\cdot}/\overline{m}_h$, the term $\pi_{j|hi}$ is the expected number of times that occupation j is selected given that establishment hi is selected. In that situation, π_{hij} is the unconditional expectation of the number of times that the combination hij is selected.

3. THE ECI AND EBS ESTIMATORS

In both the ECI and EBS most published estimates are specific to domains. Suppose that D_e is a domain of establishments defined by grouping strata (e.g.,

manufacturing) and D_o is a class of occupations (e.g., clerical and sales). Let y_{hijk} be the variable measured on worker *k* in stratum/establishment/occupation *hij*. For ECI y_{hijk} might be the worker's average hourly wage; for EBS y_{hijk} = 1 if worker *hijk* has a particular characteristic (e.g., receives long-term disability insurance) and 0 if not. An estimator of the total of *y* is

$$\hat{T}_y = \sum_{h\in D_s} \sum_{i\in s_h} \sum_{j\in D_o} \frac{\gamma_{j|hi}}{\pi_{hij}} y_{hij\cdot} \quad (1)$$

where, for U_{hij} the universe of workers in *hij*, $y_{hij\cdot} = \sum_{k\in U_{hij}} y_{hijk}$ is the total for occupation *j*, and $\gamma_{j|hi}$ is the number of times that occupation *j* is selected in establishment *hi*. The term $\gamma_{j|hi}$ is needed in (1) because of the assumption, discussed later, that occupations are sampled with replacement within an establishment. During on-site data collection, the BLS field representative collects both $y_{hij\cdot}$ and E_{hij}.

Since entire occupations and/or establishments are assigned to a domain or not, we define an establishment/occupation indicator

$$\delta_{hij} = \begin{cases} 1 & \text{if establishment / occupation } hij \text{ is in the domain} \\ 0 & \text{if not} \end{cases}$$

The estimated total for the domain can then also be written as

$$\hat{T}_y = \sum_h \sum_{i\in s_h} \sum_{j\in s_{hi}} \delta_{hij} \frac{\gamma_{j|hi}}{\pi_{hij}} y_{hij\cdot} = \sum_h \sum_{i\in s_h} \sum_{j\in U_{hi}} \delta_{hij} \frac{\gamma_{j|hi}}{\pi_{hij}} y_{hij\cdot}.$$

Note that the sum over $j \in s_{hi}$ can be replaced by a sum over $j \in U_{hi}$, the universe of occupations in establishment *hi*, since $\gamma_{j|hi}$ is 0 for all occupations not in the sample. The estimated number of employees in the domain is

$$\hat{T}_E = \sum_h \sum_{i\in s_h} \sum_{j\in U_{hi}} \delta_{hij} \frac{\gamma_{j|hi}}{\pi_{hij}} E_{hij}.$$

The mean per employee in the domain who have the characteristic is then estimated as

$$\hat{\mu} = \hat{T}_y / \hat{T}_E.$$

To approximate the variance of $\hat{\mu}$, use the usual first-order approximation for a ratio

$$\hat{\mu} - \mu \cong \left(\hat{T}_y - \mu \hat{T}_E\right) / T_E \quad (2)$$

where T_y and T_E are finite population totals and $\mu = T_y / T_E$. The key factor in approximation (2) is $\hat{T}_y - \mu \hat{T}_E$ which is equal to

$$\hat{T}_y - \mu \hat{T}_E = \sum_h \sum_{i\in s_h} \sum_{j\in s_{hi}} \delta_{hij} \frac{\gamma_{j|hi}}{\pi_{hij}} \left(y_{hij\cdot} - \mu E_{hij}\right)$$

$$= \sum_h \sum_{i\in s_h} \sum_{j\in s_{hi}} \frac{\gamma_{j|hi}}{\pi_{hij}} \delta_{hij} z_{hij\cdot}$$

where $z_{hij\cdot} = y_{hij\cdot} - \mu E_{hij}$. The term $\hat{T}_y - \mu \hat{T}_E$ is, thus, a type of Horvitz-Thompson estimator.

3.1 Variance Decomposition

To calculate a variance to be optimized in the allocation of the sample, we use the concept of anticipated variance introduced by Isaki and Fuller (1982). The anticipated variance (AV) of $\hat{\mu}$, which is approximately design unbiased, is

$$E_\xi \operatorname{var}_p(\hat{\mu}) \cong E_\xi \operatorname{var}_p\left(\hat{T}_y - \mu \hat{T}_E\right) / T_E^2$$

where E_ξ denotes expectation with respect to a superpopulation model and E_p is a expectation taken with respect to a sample design.

To evaluate the effects of different sample sizes at the two stages of selection, we need to write the variance as a sum of components associated with establishments and occupations within establishments. The standard approach to deriving variance components is to apply the conditional variance formula

$$\operatorname{var}_p(\hat{\mu}) = \operatorname{var}_p E_p(\hat{\mu}|\mathbf{s}_1) + E_p \operatorname{var}_p(\hat{\mu}|\mathbf{s}_1) \quad (3)$$

where $\mathbf{s}_1$ is the vector of all first-stage stratum samples s_h. Although it is possible to compute variance components under certain probability proportional to size (*pps*) sample designs, the results involve joint probabilities of selection and are difficult or impossible to work with in practice. Särndal, Swensson, and Wretman (1993, ch.4) discuss the design-based methods. If the strata are based on size, as in the ECI/EBS, and are numerous and narrow, a reasonable simplification is to assume that all establishments in a particular stratum have about the same number of employees. In that case, a *pps* sample selected without replacement is equivalent to a simple random sample selected without replacement (*srswor*) in each establishment stratum. The selection probability of establishment *hi* is then $\pi_{hi} = n_h / N_h$. The second-stage sample of occupations within sample establishments is also *pps*. It seems less reasonable to assume that it can be well approximated by equal probability sampling since the number of employees in different occupations varies widely in many companies. In the subsequent development, we assume that *pps* sampling with replacement is used to select occupations. The mechanics of second-stage occupation sampling, that allows an occupation to be hit more than once, is very similar to with-replacement *pps* sampling.

3.2 First and Second-stage Variance Components

The first term on the right-hand side of (3) will generate the between-establishment variance component. Since $E(\gamma_{j|hi}) = \pi_{j|hi}$, $\pi_{hij} = \pi_{hi}\pi_{j|hi}$, and we assume *srswor* at the first stage, it follows that

$$E_p\left(\hat{T}_y - \mu \hat{T}_E|\mathbf{s}_1\right) = \sum_h \sum_{i\in s_h} \sum_{j\in U_{hi}} \frac{E_p\left(\gamma_{j|hi}\right)}{\pi_{hij}} \delta_{hij} z_{hij\cdot}$$

$$= \sum_h \sum_{i\in s_h} \sum_{j\in U_{hi}} \frac{\bar{z}_{hij\cdot}}{\pi_{hi}} = \sum_h N_h \bar{\bar{z}}_{hs}$$

where $\tilde{z}_{hij\cdot} = \delta_{hij} z_{hij\cdot}$, $\bar{\tilde{z}}_{hs} = \sum_{i\in s_h} \tilde{z}_{hi\cdot\cdot}/n_h$ with $\tilde{z}_{hi\cdot\cdot} = \sum_{j\in U_{hi}} \tilde{z}_{hij\cdot}$ and U_{hi} being the universe of occupations in establishment *hi*. From the usual formula for the variance of a stratified total under *srswor* we have

$$\mathrm{var}_p\, E_p\left(\hat{T}_y - \mu\hat{T}_E|\mathbf{s}_1\right) = \sum_h \frac{N_h^2}{n_h}(1-f_h)S_{1h}^2 \qquad (4)$$

where $S_{1h}^2 = \sum_{i\in U_h}\left(\tilde{z}_{hi\cdot\cdot} - \bar{\tilde{z}}_{h\cdot\cdot}\right)^2/(N_h-1)$ with $\bar{\tilde{z}}_{h\cdot\cdot} = \sum_{i\in U_h}\tilde{z}_{hi\cdot\cdot}/N_h$ and U_h the universe of establishments in stratum *h*.

For the second-stage variance component, we need

$$\mathrm{var}_p\left(\hat{T}_y - \mu\hat{T}_E|\mathbf{s}_1\right) = \sum_h\sum_{i\in s_h}\frac{1}{\pi_{hi}^2}\mathrm{var}_p\left(\sum_{j\in U_{hi}}\gamma_{j|hi}\frac{\tilde{z}_{hij\cdot}}{\pi_{j|hi}}\right). \qquad (5)$$

Defining $\pi^*_{j|hi} = \pi_{j|hi}/\bar{m}_h$ to be the 1-draw selection probability of occupation *j* and using Result 2.9.1 of Särndal, Swensson, and Wretman (1993, p.51), the variance on the right-hand side of (5) becomes

$$\mathrm{var}_p\left(\sum_{j\in U_{hi}}\gamma_{j|hi}\frac{\tilde{z}_{hij\cdot}}{\pi_{j|hi}}\right) = \frac{1}{\bar{m}_h}\sum_{j\in U_{hi}}\pi^*_{j|hi}\left(\frac{\tilde{z}_{hij\cdot}}{\pi^*_{j|hi}} - \tilde{z}_{hi\cdot\cdot}\right)^2.$$

Consequently,

$$E_p\,\mathrm{var}_p\left(\hat{T}_y - \mu\hat{T}_E|\mathbf{s}_1\right) = \sum_h\frac{N_h}{n_h}\sum_{i\in U_h}\frac{S_{2hi}^2}{\bar{m}_h} \qquad (6)$$

where $S_{2hi}^2 = \sum_{j\in U_{hi}}\pi^*_{j|hi}\left(\tilde{z}_{hij\cdot}/\pi^*_{j|hi} - \tilde{z}_{hi\cdot\cdot}\right)^2$. Combining (4) and (6), the design-based variance is

$$T_E^2\,\mathrm{var}_p(\hat{\mu}) \cong \sum_h\frac{N_h^2}{n_h}(1-f_h)S_{1h}^2 + \sum_h\frac{N_h}{n_h}\sum_{i\in U_h}\frac{S_{2hi}^2}{\bar{m}_h}. \qquad (7)$$

A troublesome point is that the term S_{2hi}^2 is specific to a particular establishment. To use expression (7) for allocation, a separate variance component would have to be estimated for every establishment and domain of interest.

Use of a reasonable model for the $\tilde{z}_{hij\cdot}$'s will help solve this problem. Note that the sum of $\tilde{z}_{hij\cdot}$ over the full population is zero. Rather than modeling the total $\tilde{z}_{hij\cdot}$ directly, a more reasonable approach is to model the per employee mean, $z_{hij\cdot}/E_{hij}$, for those establishments that do have employees in a particular occupation. The size and sign of the residuals $z_{hij\cdot}/E_{hij}$ will typically depend on both establishment and occupation, and we will adopt the following model

$$z_{hij\cdot}/E_{hij} = \alpha_h + \beta_j + \varepsilon_{hij}$$
$$\alpha_h \sim (0,\sigma_{z\alpha}^2),\ \beta_j \sim (0,\sigma_{z\beta}^2),\ \varepsilon_{hij} \sim (0,\sigma_{z\varepsilon}^2/E_{hij}) \qquad (8)$$

with the errors α, β, and ε being independent. Since $z_{hij\cdot}/E_{hij}$ is a mean, we assume its variance is inversely related to number of employees. Next, we need to compute the model expectation of the components S_{2hi}^2 and S_{1h}^2. There is a considerable amount of algebra involved that is omitted. The final result, expressed in relvariance terms, is

$$\vartheta_{\hat{p}} \equiv \mu^{-2}E_\xi\,\mathrm{var}_p(\hat{\mu}) \cong T_y^{-2}\sum_h N_h\left(\frac{N_h}{n_h}-1\right)v_{1h} + T_y^{-2}\sum_h\frac{N_h^2}{n_h\bar{m}_h}v_{2h} \qquad (9)$$

where

$$v_{1h} = \sigma_{z\alpha}^2 V_{hE} + \sigma_{z\beta}^2\sum_{\text{all } j}V_{hj} + \sigma_{z\varepsilon}^2\bar{\tilde{E}}_h$$
$$v_{2h} = \sigma_{z\alpha}^2 V_{1hE\tilde{E}} + \sigma_{z\beta}^2 V_{2hE\tilde{E}} + \sigma_{z\varepsilon}^2 V_{hEM}$$
$$V_{hE} = \sum_{i\in U_h}\left(\tilde{E}_{hi\cdot} - \bar{\tilde{E}}_h\right)^2/(N_h-1),$$
$$V_{hj} = \sum_{i\in U_h}\left(\varphi_{hij}\tilde{E}_{hij} - \sum_{i'\in U_h}\varphi_{hi'j}\tilde{E}_{hi'j}/N_h\right)^2/(N_h-1),$$
$$V_{1hE\tilde{E}} = \sum_{i\in U_h}\tilde{E}_{hi\cdot}\left(E_{hi\cdot} - \tilde{E}_{hi\cdot}\right)/N_h,$$
$$V_{2hE\tilde{E}} = \sum_{i\in U_h}\left(E_{hi\cdot}\tilde{E}_{hi\cdot} - \sum_{j\in U_{hi}}\tilde{E}_{hij}^2\right)/N_h, \text{ and}$$
$$V_{hEM} = \sum_{i\in U_h}\left(E_{hi\cdot}M_{hi} - \tilde{E}_{hi\cdot}\right)/N_h$$

with φ_{hij} being an indicator for whether an establishment contains an occupation, $\tilde{E}_{hij} = \delta_{hij}E_{hij}$, $\tilde{E}_{hi\cdot} = \sum_{j\in U_{hi}}\tilde{E}_{hij}$, $\bar{\tilde{E}}_h = \sum_{i\in U_h}\tilde{E}_{hi}/N_h$, and M_{hi} being the number of occupations in U_{hi}. The summation over "all *j*" in v_{1h} means sum over all occupations defined for ECI/EBS.

For an estimator of a total $\hat{T}_y$, rather than the mean $\hat{T}_y/\hat{T}_E$, expression (9) must be modified only slightly. The model $y_{hij}/E_{hij} = \mu_h + \alpha_h + \beta_j + \varepsilon_{hij}$, with μ_h being the fixed mean and α_h, β_j, and ε_{hij} being independent random effects with means of 0 and variances $\sigma_{y\alpha}^2$, $\sigma_{y\beta}^2$, and $\sigma_{y\varepsilon}^2$, leads the same form of expression as (9) but with

$$v_{1h} = (\sigma_{y\alpha}^2 + \mu_h^2)V_{hE} + \sigma_{y\beta}^2\sum_{\text{all } j}V_{hj} + \sigma_{y\varepsilon}^2\bar{\tilde{E}}_h \text{ and}$$
$$v_{2h} = (\sigma_{y\alpha}^2 + \mu_h^2)V_{1hE\tilde{E}} + \sigma_{y\beta}^2 V_{2hE\tilde{E}} + \sigma_{y\varepsilon}^2 V_{hEM}.$$

4. VARIANCE COMPONENT ESTIMATION AND SMOOTHING

The variance components in expression (9) were estimated using data from the ECI for the quarter ending in September, 1992, and from EBS for the year 1992. The total number of strata used was 322, formed by crossing 72 2-digit SIC groups with five size classes: <50, 50-99, 100-249, 250-999, and 1000+ employees. The SIC groups will be referred to here as pseudo-SIC's (*psic*'s). Since both the ECI and EBS publish hundreds of statistics quarterly or annually, we made a selection of some of the more important ones to use in this study, as listed below.

ECI	EBS
Variables	Variables
Total compensation	% workers receiving:
Cost of benefits for	Life insurance
All benefits	Medical insurance
Life insurance	Retirement, savings plans
Legally required benefits	Paid sick leave

Retirement, savings plans	Paid vacation
Domains	Domains
(for total compensation)	All occupations
Full population	Profess., tech., related occs.
9 major occ. groups	Clerical, sales occs
4 regions	Production, service occs
7 industries	2 size classes (<100, 100+)

The variance components $\sigma^2_{z\alpha}, \sigma^2_{z\beta}$, and $\sigma^2_{z\varepsilon}$ in (8) were estimated for each of the preceding variables/domains using the MIVQUE0 method (Hartley, Rao, and LaMotte 1978) treating α_h, β_j, and ε_{hij} as random effects. The other components of v_{1h} and v_{2h} — V_{hE}, V_{hj}, $V_{1hE\bar{E}}$, $V_{2hE\bar{E}}$, and V_{hEM} — were estimated using simple method-of-moments estimators appropriate if simple random samples of establishments were selected in each stratum. The various parts were combined to estimate v_{1h} and v_{2h}. As Figure 1 illustrates, these components are related to size in each stratum. The figure shows a plot of $\log(v_{1h})/\hat{T}_y^2$ versus the log of the average employment size per establishment using total compensation as the variable and the full population as the domain. Plots for $\log(v_{2h})/\hat{T}_y^2$ and other variables and domains had similar features. The logs of the relvariance components are generally linearly related to the log of the average employment size with the exception of strata where the components are poorly estimated due to small establishment sample sizes. In Figure 1 points based on sample sizes of $n_h \le 4$ are shown as •'s while points with samples of $n_h > 4$ are Δ's.

To obtain more stable estimates of variance components, we smoothed the point estimates within each *psic h* across the size classes h' by fitting models of the form

$$\log(v_{khh'}) = a_h + b\log(\bar{E}_{hh'}) + \varepsilon_{khh'} \ (k=1,2). \qquad (10)$$

Based on plots like Figure 1, for a given variable like total compensation, a common slope b for all *psic*'s was reasonable while allowing the intercept a_h to differ among *psic*'s accomodated the different levels observed for some groups. Weighted least squares estimates of a_h and b were calculated using only strata with $n_h \ge 10$, a cutoff that eliminated poor point estimates for virtually all variables and domains. We then used these parameter estimates to compute smoothed, predicted variance components to use as inputs to the optimization algorithm described in the next section. An example of the results of the smoothing is shown in Figure 2 for v_{1h} for banking and credit establishments.

5. THE APPROACH TO OPTIMIZATION

The ECI and EBS publish many estimates that have varying degrees of importance. This makes the problem of sample allocation far more complicated and interesting than Neyman allocation to strata based on a single variable. In addition to the references mentioned in section 1 on multivariate allocation in surveys, there has been a considerable amount of previous, related work, including Bethel (1985), Chromy (1987), Hughes and Rao (1979), Kokan (1963), and Kokan and Khan (1967). Multicriteria optimization programming is one method for dealing with such a situation. Our approach will be to minimize a weighted sum of the relvariances of a number of important statistics subject to various constraints defined below. Because the statistics from these surveys are of disparate types — proportions of employees, total dollar costs, costs per employee per hour — use of relvariances puts the estimates on a comparable scale. To write the optimization problem mathematically, let w_ℓ be a weight associated with estimator ℓ ($\ell = 1,\ldots,L$) and ϑ_ℓ be the anticipated relvariance of the estimator, defined by (9). The optimization problem we have formulated for the ECI and EBS is

$$\text{minimize} \qquad \phi = \sum_{\ell=1}^{L} w_\ell \vartheta_\ell \qquad (11)$$

subject to

(1) $n_{h,\min} \le n_h \le N_h$ for establishment sample sizes n_h,

(2) $n = \sum_h n_h \le n_0$, a bound on the total number of sample establishments,

(3) $m_{h,\min} \le \bar{m}_h \le m_{h,\max}$, i.e. the number of occupations sampled per establishment in stratum h is bounded above and below,

(4) $\dfrac{\sum_{h\in S} n_h \bar{m}_h}{\sum_{h\in S} n_h} \le \bar{\bar{m}}_{S,\max}$, i.e. the average number of occupations sampled per establishment is bounded above in a subset S of strata.

(5) $\vartheta_\ell^{1/2} \le \vartheta_{\ell 0}^{1/2}$ for $\ell \in S_E$, i.e. the coefficient of variation of an estimator ℓ is bounded for all estimators in some set S_E.

The weights $\{w_\ell\}_{\ell=1}^{L}$ are based on subjective judgments as to the relative importance of each estimator in meeting the goals of the surveys. Because analysts may have different opinions on how the weights should be assigned, we have designed software for solving the optimization problem that flexibly allows the effects of modifying the weights to be explored. We used the PV-WAVE Advantage™ package sold by Visual Numerics™ running under Unix™ and X-Windows™ to develop a program with a graphical user interface (GUI) to adjust parameters of the optimization problem and then to solve the problem.

To use the software expressions must be programmed in a higher-level language (similar to Fortran or C) for

- variances
- constraints
- objective function.

The variances may be similar to those in expression (9), especially if the sampling is two-stage, but that is not necessary.

The programmer can form an objective function in various ways, although the most obvious way would be as a weighted sum of the variances or relvariances, as in (11). To simplify specification of the weights, the program allows the number of components of the objective function to be variable. If the objective function is a single variance, for example, the optimization problem could be the usual one that leads to Neyman allocation to strata.

The bounds on the individual sample sizes are usually easy to set and generally require no programming.

Once the code for the variances, constraints, and objective function has been written, the user may still be able to vary the problem considerably. Because of the common form of the expressions, the user may be able to define a specific problem by supplying key parameters:

- number of stages in sampling design (one or two)
- number of estimators
- number of strata
- number of constraints
- number of components in a weighted objective function.

The kinds of data required are

- strata sizes
- which strata are used for each estimator (in the case of domain estimation)
- the strata variance components for each estimator
- labels for the strata and estimators and any other information required by the specific constraints or objective function

After getting information about the problem, the program opens a window, shown in Figure 3, in which are displayed various tables and action buttons. One important table shows the strata population and sample sizes. The table shows two kinds of sample sizes: "trial" and "optimal." How these are initialized is optional; both are usually initialized to the lower bounds of the variables. The trial entries in the table can be modified directly. Another table shows the constraint bounds and the constraint values corresponding to the trial and optimal allocations. The bounds in the table can be edited directly. There are basically two things the user may do:

(1) Setup the constraint bounds and the weights in the overall objective function, and then determine the optimal allocation.

(2) Fix a trial allocation, and then determine the corresponding values of variances, constraints and objective function.

The action buttons allow the user to perform either of these actions. The buttons also allow the user to save allocations, so that different allocations can be compared.

Constraint bounds are entered by editing a table or by selecting the constraint values corresponding to a trial allocation. Weights for the objective function, i.e., w_t's in (11), are assigned by moving slider bars. If the objective function consists of only one component (which itself may be a sum of variances), the slider bars do not appear in the GUI. A trial allocation can be assigned by editing a table or by choosing an action button that allows a choice of lower or upper bounds, the (rounded) optimum, or a previously saved solution.

The optimization problem has a nonlinear objective function and nonlinear constraints. A variety of algorithms is available for solving this optimization problem (Moré and Wright, 1993). One of the better ones, **GRG2**, due to Lasdon and Waren (1978), was implemented here.

When the user selects the action button to determine an optimal allocation, the program requests selection of "starting values." There are two choices (made by selecting an action button): use lower bounds or use the current trial allocation. The optimization problem is difficult, so selection of the starting values can be important for both speed and convergence. It may require some experimentation to arrive at good starting values. After termination, the user can select allocations that are rounded to integer values as trial allocation, and evaluate the objective function and the constraints.

6. CONCLUSION

The results presented here are preliminary and part of an ongoing project to develop general purpose allocation software for use in BLS sample surveys. More detailed results will be presented in a subsequent paper. In particular, we will explore the effect on allocations of different objective functions and importance weights and will compare optimized allocations to ones based on rules of thumb, e.g., allocate in proportion to stratum employment size.

ACKNOWLEDGEMENTS

Any opinions expressed are those of the authors and do represent policy of the Bureau of Labor Statistics.

REFERENCES

Bethel, J. (1985), "An Optimum Allocation Algorithm for Multivariate Surveys," *Proceedings of the Section on Survey Methods Research*, American Statistical Association, 209-212.

—— (1989), "Sample Allocation in Multivariate Surveys," *Survey Methodology*, **15**, 47-57.

Chromy, J. (1987), "Design Optimization with Multiple Objectives," *Proceedings of the Section on Survey Methods Research*, American Statistical Association, 194-199.

Hartley, H.O., Rao, J.N.K., and LaMotte, L. (1978), "A Simple Synthesis-based Method of Variance Component Estimation," *Biometrics*, **34**, 233-244.

Hughes, E., and Rao, J. (1979), "Some Problems of Optimal Allocation in Sample Surveys Involving Inequality Constraints," *Communications in Statistics — Theory and Methods*, **A8 (15)**, 1551-1574.

Isaki, C. and Fuller, W. (1982), "Survey Design Under the Regression Superpopulation Model," *Journal of the American Statistical Association*, 77, 89-96.

Kish, L. (1988), "Multipurpose Sample Designs," *Survey Methodology*, **14**, 19-32.

Kokan, A.R. (1963), "Optimum Allocation in Multivariate Surveys," *Journal of the Royal Statistical Society A*, **126**, 557-565.

Kokan, A.R., and Khan, S. (1967), "Optimum Allocation in Multivariate Surveys: An Analytical Solution," *Journal of the Royal Statistical Society B*, **29**, 115-125.

Lasdon, L. and Waren, A. (1978), "Generalized Reduced Gradient Software for Linearly and Nonlinearly Constrained Problems," in *Design and Implementation of Optimization Software*, ed. H. Greenberg, Sijthoff and Noordhoff: Alphen aan den Rijn.

More, J. J., and Wright, S. J. (1993), *Optimization Software Guide*, Philadelphia: SIAM.

Narula, S., and Weistroffer, H. (1989), "Algorithms For Multiple Objective Nonlinear Programming Problems," in *Improving Decision Making in Organizations* (A. Lockett and G. Islei, eds.), Berlin: Springer-Verlag, 434-443.

Särndal, C.-E., Swensson, B., and Wretman, J. (1993), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Steuer, R. (1986), *Multiple Criteria Optimization: Theory, Computation, and Application*, New York: Wiley.

Weistroffer, H., and Narula, S. (1991), "The Current State of Nonlinear Multiple Criteria Decision Making," in *Operations Research* (G. Fandel and H. Gehring, eds.), Berlin: Springer-Verlag, 109-119

Figure 1. Log of stratum v1h relvariance components for total compensation plotted versus log of average establishment employment size.

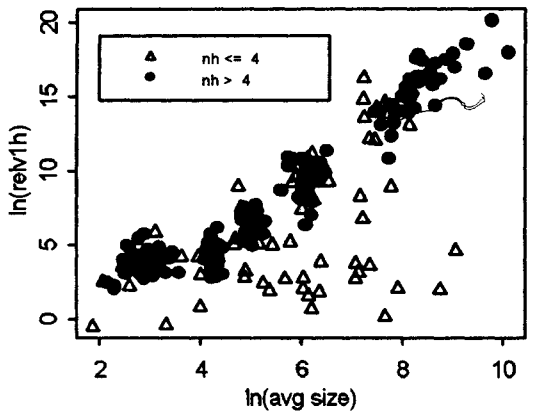

Figure 2. Log of predicted and point estimates of stratum relvar v1h components for total compensation plotted versus log of avg. employment size for banking and credit establishments.

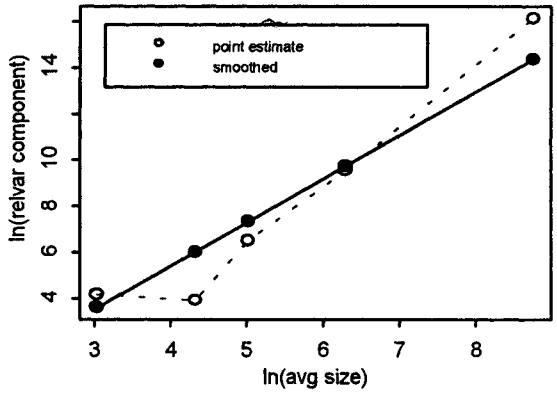

Figure 3. Window from the allocation software with tables and action buttons.