Accepted for publication by *Computational Statistics and Data Analysis*

AN APPLICATION OF MATHEMATICAL PROGRAMMING TO SAMPLE ALLOCATION

Richard Valliant
Bureau of Labor Statistics
Room 4915
2 Massachusetts Ave. NE
Washington DC 20212, USA

James E. Gentle
Institute for Computational Sciences and Informatics
George Mason University
Fairfax VA 22030-4444, USA

November 1996

ABSTRACT

The problem of sample allocation in multipurpose surveys is complicated by the fact that an efficient allocation for some estimates may be inefficient for others. There may also be precision goals that must be met for certain estimates plus constraints on costs and minimum sample sizes for strata to permit variance estimation. These requirements lead to formulating the allocation problem as one of mathematical programming with an objective function and constraints that are nonlinear in the sample size target variables. We discuss a flexible approach for a two-stage sample allocation that uses multicriteria optimization programming. Software was developed to permit survey designers to easily explore alternative problem formulations and to compare the resulting allocations. The method is illustrated using a business establishment survey that estimates the costs to employers of providing wages and benefits to employees and the percentages of employees that receive certain benefits.

Keywords: Nonlinear optimization; multicriteria programming; two-stage sampling; variance component estimation.

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. To reduce costs and produce datasets with richer analysis possibilities, there may also be pressures to consolidate and use the same sample for multiple surveys. This may be particularly true in surveys sponsored by governments. This paper will discuss an application of multicriteria optimization to sample design, using as an illustration, data drawn from the Employment Cost Index (ECI) and the Employee Benefits Survey (EBS) conducted by the Bureau of Labor Statistics (BLS). The two surveys 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. The paper also illustrates some of the practical compromises and approximations that must be employed in the design of a complex sample.

Kish (1988) noted the variety of purposes for which a given survey may be used and why purposes may conflict. A number of different variables may be measured and estimates may be made for diverse domains. Some domains may be straightforward combinations of design strata, like regions, while others may be crossclasses, like occupational groups, that cut across design strata. The ECI, for example, collects the cost to employers of paying for wages and salaries, leave (e.g., holidays, vacations), medical

and life insurance, legally required benefits (e.g., social security, unemployment insurance, workers' compensation), and retirement plans. The EBS collects the numbers of employees who are eligible for or receive various benefits. Estimates are made of an index of change in costs between time periods, of the average cost per employee per hour worked, and of the percentage of employees receiving various benefits within domains, including industry group (e.g., construction, manufacturing, wholesale, services), establishment size, class of worker (e.g., professional, technical, and related occupations, sales occupations, service occupations), and geographic region.

Sample allocation problems, using either variance minimization or cost minimization, can be formulated using *constrained, multicriteria optimization* described, for example, in Narula and Weistroffer (1989), Steuer (1986), or Weistroffer and Narula (1991). An objective function that is 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. A final component of the objective function is a weight times the total cost (or sample size). More formally, the objective function to be minimized in this study is

$$f = \sum_{\ell=1}^{L} w_\ell J_\ell + w_{L+1} c$$

where $\ell = 1, \ldots, L$ are the indexes of the estimators, J_ℓ is the relvariance of estimator ℓ (i.e., the variance of the estimator divided by the square of its expected value), c is the total survey cost, which is implicitly a function of sample size, and $w_1, \ldots, w_{L+1}$ are weights assigned to the relvariances and the cost. Because the relvariance is a unitless measure, we can include estimators in the objective function for variables that are measured on

different scales, e.g., wages and proportions. However, since the relvariances are unitless and the cost is measured in monetary units, they should not simultaneously enter the objective function. Using a reduced gradient programming algorithm, 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 relvariance constraints on individual estimators.

Including both relvariances and cost in the objective function has no special mathematical advantages since, when the relvariances receive nonzero weight, the cost weight w_{L+1} should be zero, and vice versa. However, inclusion of cost and relvariances in f is extremely convenient when using software that flexibly allows weight adjustment. One of the important byproducts of this research was a system, described in section 5, that lets a user interactively change the weights in f, re-optimize, and compare results to previous allocations.

To evaluate the objective function, specific formulas are needed for the relvariances. The ECI/EBS sample design and estimators, like many others in complex surveys, are complicated and appropriate variance formulas cannot be taken directly from a textbook. Sections 2 and 3 sketch the derivation of relvariances needed for this application. Because a 2-stage sample is used, variance components for each stage are derived, and their estimation is discussed in section 4. The software for optimization and the numerical results are covered in sections 5 and 6.

2. SAMPLE DESIGN

This section describes the general sample design used in the ECI and EBS programs. Sample data from this design will be used in estimating variance components that, in turn, will be used in exploring how to allocate the sample. The 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 frame at a particular date. Strata are defined by standard industrial classification (SIC) and employment size. At the second stage, a sample of occupations is selected within each sample establishment. 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. If, for example, a janitor is selected in the systematic sample from an establishment, then all janitors are enumerated. Occupations with more workers are, thus, more likely to be in the sample. The occupation sampling procedure is simple to implement in the field but does allow a particular occupation to be selected or "hit" more than once, introducing some complications into analysis. 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

p_{hi} = inclusion probability of establishment hi

n_h = number of sample establishments in stratum h

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

$\overline{m}_h$ = number of sample occupations assigned to sample establishment hi, which is the same for each sample establishment in stratum h,

s_h = set of sample establishments in stratum *h*, and

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

The same number $\overline{m}_h$ of sample occupations is assigned to each sample establishment in a stratum in order to control work loads. Note that the number of unique occupations obtained in the subsample from an establishment will be less than $\overline{m}_h$ when a particular occupation is hit more than once.

The quantities p_{hi} and $p_{j|hi}$ are general. Specifically for this application, if E_{hi} is the number of employees in the establishment, then the selection probability of establishment *hi* is $p_{hi} = n_h E_{hi}/E_h$ where E_h 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 $E_{hij} > E_{hi}/\overline{m}_h$, then the selection probability of an occupation within the establishment is $p_{j|hi} = \overline{m}_h E_{hij}/E_{hi}$. The overall selection probability of unit *hij* is then $p_{hij} = p_{hi}p_{j|hi} = n_h\overline{m}_h E_{hij}/E_h$. In a case where there are one or more occupations with $E_{hij} > E_{hi}/\overline{m}_h$, the term $p_{j|hi}$ is the expected number of times that occupation *j* is selected given that establishment *hi* is selected. In that situation, p_{hij} is the unconditional expectation of the number of times that the combination *hij* is selected.

To make variance calculations tractable, we later assume that establishments are sampled by stratified simple random sampling *without* replacement and that occupations are sampled with probabilities proportional to size but *with* replacement. The reasoning behind these assumptions is discussed in the next section.

3. THE ECI AND EBS ESTIMATORS

In both the ECI and EBS most published estimates are specific to domains, e.g., the average cost per hour per employee for wages and salaries in goods-producing industries or the percentage of clerical and sales workers who receive paid vacation. 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. Computations will be facilitated by having an indicator for units that are in a particular establishment/occupation domain. Since entire occupations and/or establishments are assigned to a domain or not, let d_{hij} =1 if establishment/occupation *hij* is in the domain and 0 if not. An estimator of the domain total of *y* is

$$\hat{T}_y = \sum_h \sum_{i \in s_h} \sum_{j \in s_{hi}} d_{hij} \frac{g_{j|hi}}{p_{hij}} y_{hij} \ . \qquad (1)$$

where, for U_{hij} the universe of workers in *hij*, $y_{hij} = \sum_{k \in U_{hij}} y_{hijk}$ is the total for occupation *j*, and $g_{j|hi}$ is the number of times that occupation *j* is selected in establishment *hi*. The total y_{hij} is collected by the interviewer after the sample is selected for only those occupations in the sample. The term $g_{j|hi}$ is needed in (1) because of the assumption, discussed later, that occupations are sampled with replacement within an establishment.

Note that the sum over $j \in s_{hi}$ in (1) can be replaced by a sum over $j \in U_{hi}$, the universe of occupations in establishment *hi*, since $g_{j|hi}$ is 0 for all occupations not in the sample. The estimated number of employees in the domain is

$$\hat{T}_E = \sum_h \sum_{i \in s_h} \sum_{j \in U_{hi}} d_{hij} \frac{g_{j|hi}}{p_{hij}} E_{hij} .$$

Because $g_{j|hi} = 0$ for nonsample occupations, E_{hij} is needed only for the sample occupations and is collected at the time of sampling by the interviewer. The proportion of employees in the domain who have the characteristic (or the mean per employee if *y* is quantitative) is then estimated as

$$\hat{m} = \hat{T}_y \big/ \hat{T}_E .$$

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

$$\hat{m} - m \cong \left(\hat{T}_y - m\hat{T}_E\right) \big/ T_E$$

$$= \frac{1}{T_E} \sum_h \sum_{i \in s_h} \sum_{j \in s_{hi}} \frac{g_{j|hi}}{p_{hij}} d_{hij} z_{hij} . \qquad (2)$$

where T_y and T_E are finite population totals, $m = T_y / T_E$, and $z_{hij} = y_{hij} - mE_{hij}$.

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 $\hat{m}$ is

$$E_x E_p \left[\left(\hat{m} - m\right) - E_x E_p \left(\hat{m} - m\right) \right]^2 = E_x E_p \left(\hat{m} - m\right)^2 - \left[E_x E_p \left(\hat{m} - m\right) \right]^2 .$$

where E_x denotes expectation with respect to a superpopulation model and E_p is a expectation taken with respect to a sample design. When $\hat{m}$ is design-unbiased (or approximately so), the AV can be written as

$$E_x \operatorname{var}_p(\hat{m}) \cong E_x \operatorname{var}_p\left(\hat{T}_y - m\hat{T}_E\right) \Big/ T_E^2$$

where var_p is the design variance. $E_x \operatorname{var}_p(\hat{m})$ is a variance anticipated at the time the sample is constructed and has several practical advantages for sample design. The design-based variance calculation will allow finite population correction factors to enter the variance in a somewhat easier way than a purely model-based computation would. The superpopulation approach, on the other hand, permits explicit modeling of characteristics, which will prove to be an important advantage when estimating variance components.

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

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

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 often 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 $p_{hi} = n_h / N_h$. Being able to incorporate a finite population correction factor is important in this application because the sampling fraction in some strata is non-negligible. The second-stage sample of occupations within sample establishments is also *pps*. It seems less reasonable to assume that the second stage selection 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. Interviewers typically select the second stage samples from alphabetically sorted personnel files whose order is unrelated to occupation. Consequently, the possibility of having periodicities that can foul the properties of systematic sampling is small.

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 $E\left(g_{j|hi}\right) = p_{j|hi}$, $p_{hij} = p_{hi} p_{j|hi}$, and we assume *srswor* at the first stage, it follows that $E_p\left(\hat{T}_y - m\hat{T}_E \middle| \mathbf{s}_1\right) = \sum_h N_h \bar{\tilde{z}}_{hs}$ where, $\bar{\tilde{z}}_{hs} = \sum_{i \in s_h} \tilde{z}_{hi} / n_h$ with $\tilde{z}_{hi} = \sum_{j \in U_{hi}} \tilde{z}_{hij}$, $\tilde{z}_{hij} = d_{hij} z_{hij}$, 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

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

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

For the second-stage variance component, we need

$$\text{var}_p\left(\hat{T}_y - m\hat{T}_E \,|\mathbf{s}_1\right) = \sum_h \sum_{i\in s_h} \frac{1}{p_{hi}^2}\,\text{var}_p\left(\sum_{j\in U_{hi}} g_{\,j|hi}\,\frac{\tilde{z}_{hij}}{p_{\,j|hi}}\right). \tag{5}$$

Defining $p_{j|hi}^* = p_{\,j|hi}\big/\bar{m}_h$ 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

$$\text{var}_p\left(\sum_{j\in U_{hi}} g_{\,j|hi}\,\frac{\tilde{z}_{hij}}{p_{\,j|hi}}\right) = \frac{1}{\bar{m}_h}\sum_{j\in U_{hi}} p^*_{\,j|hi}\left(\frac{\tilde{z}_{hij}}{p^*_{\,j|hi}} - \tilde{z}_{hi}\right)^2 .$$

Consequently,

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

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

$$T_E^2\,\text{var}_p\left(\hat{m}\right) = \sum_h \frac{N_h^2}{n_h}\left(1 - f_h\right)S_{1h}^2 + \sum_h \frac{N_h}{n_h}\sum_{i\in U_h}\frac{S_{2hi}^2}{\bar{m}_h}. \tag{7}$$

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 $\tilde{z}_{hij}$'s will help solve this problem. First, consider a model for the per employee mean, y_{hij}/E_{hij}, for those establishments that do have employees in a particular occupation. Certain occupations will tend to have larger than average y_{hij}/E_{hij}'s because of their high pay or years of experience. Establishments in some SIC/size strata may also tend to offer more (or less) pay or benefits than others. There is also likely to be residual error associated with a particular establishment/occupation combination. Considering these factors, we will adopt the following model

$$y_{hij}/E_{hij} = m + a_h + b_j + e_{hij},\ a_h \sim (0, s^2_{za}),\ b_j \sim (0, s^2_{zb}),\ e_{hij} \sim (0, s^2_{ze}/E_{hij}) \qquad (8)$$

with the errors a, b, and e being independent. Because SIC/size strata are relatively homogeneous groups, no factor for establishments is included. The model for $z_{hij}/E_{hij} = y_{hij}/E_{hij} - m$ is then $a_h + b_j + e_{hij}$. Since z_{hij}/E_{hij} is a mean, we assume its variance is inversely related to number of employees. Model (8) is undoubtedly an over-simplification since it does not account for the differences in unionized and non-unionized occupations, among other complications. The model may also predict the average pay y_{hij}/E_{hij} to be negative in an unusual case. The model does, however, account for some of the key determinants of wage levels—specifically occupation and industry.

Next, we need to compute the model expectation of the components S^2_{2hi} and S^2_{1h}. There is a considerable amount of algebra involved that is sketched in the Appendix. The final result, expressed in relvariance terms, is

$$J_{\hat{m}} \equiv m^{-2} E_x \operatorname{var}_p(\hat{m}) \cong T_y^{-2} \sum_h N_h \left(\frac{N_h}{n_h} - 1 \right) v_{1h} + T_y^{-2} \sum_h \frac{N_h^2}{n_h \bar{m}_h} v_{2h} \qquad (9)$$

where

$$v_{1h} = s_{za}^2 V_{hE} + s_{zb}^2 \sum_{\text{all } j} V_{hj} + s_{ze}^2 \bar{\tilde{E}}_h \,,\; v_{2h} = s_{za}^2 V_{1hE\tilde{E}} + s_{zb}^2 V_{2hE\tilde{E}} + s_{ze}^2 V_{hEM},$$

$$V_{hE} = \sum_{i \in U_h} \left(\tilde{E}_{hi} - \bar{\tilde{E}}_h \right)^2 \Big/ (N_h - 1),\; V_{hj} = \sum_{i \in U_h} \left(j_{hij} \tilde{E}_{hij} - \sum_{i' \in U_h} j_{hi'j} \tilde{E}_{hi'j} \Big/ N_h \right)^2 \Big/ (N_h - 1),$$

$$V_{1hE\tilde{E}} = \sum_{i \in U_h} \tilde{E}_{hi} \left(E_{hi} - \tilde{E}_{hi} \right) \Big/ N_h \,,\; V_{2hE\tilde{E}} = \sum_{i \in U_h} \left(E_{hi} \tilde{E}_{hi} - \sum_{j \in U_{hi}} \tilde{E}_{hij}^2 \right) \Big/ N_h \,, \text{ and}$$

$$V_{hEM} = \sum_{i \in U_h} \left(E_{hi} M_{hi} - \tilde{E}_{hi} \right) \Big/ N_h$$

with j_{hij} being an indicator, defined in the Appendix, for whether an establishment contains an occupation, $\tilde{E}_{hij} = d_{hij} E_{hij}$, $\tilde{\tilde{E}}_{hi} = \sum_{j \in U_{hi}} \tilde{E}_{hij}$, $\bar{\tilde{E}}_h = \sum_{i \in U_h} \tilde{E}_{hi} \big/ N_h$, 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 the survey.

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 $y_{hij} / E_{hij} = m_h + a_h + b_j + e_{hij}$, with m_h being the fixed mean and a_h, b_j, and e_{hij} being independent random effects with means of 0 and variances s_{ya}^2, s_{yb}^2, and s_{ye}^2, leads to the same form of expression as (9) but with $v_{1h} = \left(s_{ya}^2 + m_h^2 \right) V_{hE} + s_{yb}^2 \sum_{\text{all } j} V_{hj} + s_{ye}^2 \bar{\tilde{E}}_h$ and $v_{2h} = \left(s_{ya}^2 + m_h^2 \right) V_{1hE\tilde{E}} + s_{yb}^2 V_{2hE\tilde{E}} + s_{ye}^2 V_{hEM}$.

Note that an alternate model for y_{hij} / E_{hij} would be to use m rather than m_h as in (8). This might have led to more stable estimates v_{1h} and v_{2h}; however, we applied some

general smoothing procedures to stabilize all variance component estimates, as described in the next section.

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 datasets differ somewhat from those used for publications from those surveys and estimates here will differ from ones that may have been published by BLS. The total number of strata used here was 360, formed by crossing 72 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, shown in Table 1. The sample design that produced our test data used only *psic*'s as explicit strata. Establishments were sorted within *psic*'s by geographic location and size but those characteristics were not explicitly used as strata in the original design. The 360 strata were, thus, finer subdivisions of the original strata, making the assumption of *srswor* of establishments more tenable.

The variance components s^2_{za}, s^2_{zb}, and s^2_{ze} in (8) were estimated for each of the preceding variables/domains using the MIVQUE0 method (Hartley, Rao, and LaMotte 1978) treating a_h, b_j, and e_{hij} as random effects. The other components of v_{1h} and v_{2h}—V_{hE}, V_{hj}, $V_{1hE\tilde{E}}$, $V_{2hE\tilde{E}}$, and V_{hEM}—were estimated using simple method-of-moments estimators appropriate if observations on establishments were independent and identically distributed in each stratum. Method-of-moments estimation is reasonable from a design-

based point-of-view when size strata are sufficiently narrow that the *pps* sampling actually used in the surveys is approximately equivalent to simple random sampling. A compromise form of V_{hEM}, different from the one defined after expression (9), had to be devised because M_{hi}, the number of occupations in an establishment, was not always available. The method used is sketched in Appendix A.2. Since the contribution of V_{hEM} to v_{2h} was relatively small, this compromise was of little consequence.

The various parts were combined to estimate v_{1h} and v_{2h} for totals in the ECI and proportions in the EBS. As Figure 1 illustrates, these components are related to size in each stratum. The figure shows plots of $\log\left(v_{1h}/\hat{T}_y^2\right)$ and $\log\left(v_{2h}/\hat{T}_y^2\right)$ 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 other variables and domains had similar features. The logs of the relvariance components are generally linearly, or possibly quadratically, 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 $n_h > 4$ are shown as ●'s while points with samples of $n_h \leq 4$ are Δ's.

Estimated variances are themselves subject to variability. 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

$$\log\left(v_{khh'}\right) = a_h + b\log\left(\overline{E}_{hh'}\right) + e_{khh'} \quad (k=1,2) \tag{10}$$

where $v_{khh'}$ is a variance component and $\overline{E}_{hh'}$ is the average population employment per establishment in (*psic h*/size class h'). 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 accommodated the different levels observed for some groups. Fitting models with some curvature would have yielded somewhat better fits, but our major goal was to stabilize component estimates. As long as the relative sizes of stratum component estimates are reasonable, the sample allocations in section 6 will not be sensitive to some misspecification in model (10).

Weighted least squares estimates of a_h and *b* were calculated that minimized the sum of squares $\sum_{h,h'} w_{hh'}\left[\log(v_{khh'}) - a_h - b\log(\overline{E}_{hh'})\right]^2$ with $w_{hh'} = n_{hh'}/\log(\overline{E}_{hh'})$. The weights $w_{hh'}$ were designed to be roughly inversely proportional to the variances of $\log(v_{khh'})$ observed in scatterplots like Figure 1. Strata for estimating the parameters in (10) were limited to those with $n_h \geq 10$, a larger cutoff than in Figure 1, 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. Figure 2 shows the point estimates and smoothed values of $\log\left(v_{1h}/\hat{T}_y^2\right)$ and $\log\left(v_{2h}/\hat{T}_y^2\right)$ plotted versus the log of the average employment size for total compensation in *psic* 61 (Banking, savings and loan, and other credit organizations) and 83 (social services). In 3 of the 4 panels of Figure 2, smoothing leads to a larger component in the 1000+ size class where the point estimates

were based on small sample sizes. This was a common and desirable effect of the smoothing in many *psic*'s having a small sample of large 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 related work, including Bethel (1985, 1989), Chromy (1987), Hughes and Rao (1979), Kokan (1963), Kokan and Khan (1967), and Zayatz and Sigman (1994). These papers deal almost entirely with single-stage sampling. Notable exceptions are Leaver, et.al. (1987, 1996) which discuss a complex, multistage allocation problem for the U.S. Consumer Price Index.

Multicriteria optimization programming is one method for dealing with multivariate allocation. 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—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 (ℓ=1,...,L) and J_ℓ be the anticipated relvariance of the estimator, defined by (9). The total cost is c and its weight in the objective function is w_{L+1}, which is

either 0 or 1. The optimization problem we have formulated for the ECI and EBS is to find $\{n_h, \overline{m}_h; h = 1, \ldots, H\}$ that

$$\text{minimize} \qquad f = \sum_{\ell=1}^{L} w_\ell J_\ell + w_{L+1} c \tag{11}$$

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

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

(5) $J_\ell^{1/2} \le J_{\ell 0}^{1/2}$ 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.

As noted in the Introduction, the cost and relvariances are on different scales and do not enter f simultaneously. Differential stratum costs were not available for these surveys but could be accommodated in our approach.

The number of occupations sampled from an establishment is bounded in two ways with constraints (3) and (4). First, an upper and lower bound on $\overline{m}_h$ is used in each stratum. Second, the average number of occupations sampled per establishment will be bounded above, across some subsets of strata. This approach allows some flexibility in

assigning the number of sample occupations, which may lead to a more efficient allocation, while still restricting the average burden per establishment.

The weights $\{w_\ell\}_{\ell=1}^{L}$ in the objective function 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 called ALLOCATE with a graphical user interface (GUI) to adjust parameters of the optimization problem and then to solve the problem. The PV-WAVE interface calls a C program known as GRG2 (Lasdon and Waren 1978; Windward Technologies Inc. 1994) that solves the nonlinear problem.

Figure 3 shows the main ALLOCATE window that is divided into four sections: action buttons, allocation table, constraints table, and weight slider bars. The allocation table in the upper right shows the stratum population and sample sizes for both stages of sampling. The table also shows two kinds of sample sizes: “trial” and “optimal.” How the trial values are initialized is optional; two possibilities are the lower bounds of the variables or an allocation used in the survey for a previous time period. The trial entries in the table can be modified directly. The constraints table in the lower right shows the constraint bounds and the values of the constraints corresponding to the trial and optimal allocations. The bounds in the table can be also edited directly. Weights for the objective function, i.e., w_ℓ's in (11), are assigned by moving slider bars in the lower left-hand part of

Figure 3. If the objective function consists of only one component (which itself may be a sum of relvariances), the slider bars do not appear in the GUI. Action buttons are in the upper left-hand portion of the window. Clicking any of the first five of these leads to secondary choices, two of which will be sketched briefly.

The buttons labeled “Determine optimum allocation...” and “Compare trial and optimum...” and their secondary choices are pictured in Figure 4. After selecting the former and preparatory to computing an optimum, the user can set some tuning parameters for GRG2 or select starting values for the algorithm which may be the lower bounds, the trial allocation, or the previously computed optimum. After an optimum has been found, the user can compare it with the trial in several ways after pressing “Compare trial and optimum...” One is by plotting the trial and optimal first-stage stratum sample sizes versus each other as shown in Figure 4. Strata to which the points correspond can be identified by clicking them with the mouse. Three of the points in Figure 4 are labeled as illustrations.

The optimization problem can have a nonlinear objective function and nonlinear constraints. In the ECI/EBS problem, the objective function f is a function of n_h^{-1} and $\overline{m}_h^{-1}$ while the constraints depend on n_h and $\overline{m}_h$. The variables (the sample sizes) are restricted to integer values—a restriction that is unimportant in most problems. The sensitivity of using integer solutions that are near to the continuous solution can be investigated by assigning them as trial allocations and seeing whether the algorithm seeks a different solution. In the continuous version of this problem (without the integer restrictions), both the objective function and the constraints are smooth. A variety of algorithms is available for solving this optimization problem (Moré and Wright, 1993).

The one selected here, GRG2™, is a reduced gradient programming method that iteratively searches for a solution. Termination occurs when the improvements in the objective function are small. Starting values may affect convergence, and it is wise to run any optimization with several initial allocations in order to check results.

6. NUMERICAL RESULTS

To illustrate the flexibility of this approach to allocation and its superiority to "rule-of-thumb" allocations commonly used in sampling, we present results for three ECI/EBS minimization problems. The sample size of establishments in the ECI/EBS dataset in 1992 was about 4,360 which will be maintained as the constraint n_0 in the first two problems. Keeping the same total sample permits us to analyze whether the current sample can be profitably redistributed among the strata. The third problem is the minimization of the number of sample establishments subject to the same set of constraints used in the first two problems.

Table 2 shows the minimum, maximum, and average numbers of sample occupations per sample establishment that will be used for each size class. Each row in the table refers to a size class, and implicitly to all pseudo-SIC's (*psic*) in the size class. For the <50 size class, for instance, the number of occupations assigned to each sample establishment is constrained to be between 2 and 6. The average number of occupations, $\overline{\overline{m}}_{S,\max}$, allocated per establishment across all *psic*'s in the <50 class is constrained to be no more than 4. The constraints on the average $\overline{\overline{m}}_{S,\max}$ in each size class are somewhat larger than the $\overline{m}_h$ = 4, 6, or 8 used in the ECI/EBS in 1992. Thus, the total workload within

establishments is allowed to be more than the current amount but extreme increases are avoided.

Upper bounds on *cv*'s of different estimates are listed in Tables 3-5. The bounds are set to be approximately equal to the *cv*'s achieved for the 1992 ECI/EBS. Table 3 lists bounds on the *cv*'s that were used for ECI estimates of total costs for the full population for six variables. Bounds on *cv*'s of ECI estimates of total compensation costs for major occupational groups (MOGs) and industry divisions are listed in Table 4. Listed in Table 5 are the bounds used for *cv*'s of EBS occupational domain estimates of percentages of employees receiving certain benefits in establishments of 100+ employees (referred to subsequently as EBS/100+). The *cv*'s of EBS estimates of percentages for small establishments (EBS/<100) are unconstrained. Note that the domains for occupational groups, like professional and technical workers and administrative support, are crossclasses that cut across the basic *psic*/size strata, and we, thus, have no direct control over the number of sample employees in those domains.

A five component, weighted objective function was used that had as its parts (1) the relvariance of the estimate of total compensation from the ECI, (2) the sum of the relvariances of ECI estimates of total compensation for major occupational groups A-K in Table 4, (3) relvariance of the ECI estimate of total benefit costs, (4) the sum of the relvariances of the benefit costs of paid leave, insurance, legally required benefits, and retirement and savings, and (5) the total establishment sample size $n = \sum_{h=1}^{H} n_h$. By adjusting the weight slider to 0 for component 5, we formulate a problem for minimizing a function of relvariances; by adjusting the other four sliders to 0 and the last to 1, the problem becomes one of minimizing the total establishment sample size. This is an

extremely handy way for a user to set up a problem because no reprogramming is needed to switch from relvariance minimization to cost minimization.

Selection of weights is subjective but should be guided by the goals and priorities of the survey. Total compensation and benefits are key targets in the ECI and, thus, deserve relatively large weight. In some surveys, sponsors may have difficulties in specifying goals in a way that easily translates into weights. Even in those cases, a set of domains can usually be identified as being important and equal weights assigned to each. The exercise of assigning weights can, in fact, be a useful way of forcing survey designers to think more clearly about their goals.

We report here on the results from three choices of the vector of weights $\mathbf{w} = (w_1, w_2, w_3, w_4, w_5)'$ in the objective function. The first was $\mathbf{w}_1 = (1, 0.5, 1, 0.5, 0)'$ which gives twice the weight to the estimates for total compensation and total benefits as to the other estimates. The second was $\mathbf{w}_2 = (1, 0, 1, 0, 0)$ which eliminates the components for occupational groups and individual benefits. The third was $\mathbf{w}_3 = (0, 0, 0, 0, 1)$ for the sample size minimization problem. The user sets each of the weight vectors with slider bars in Figure 3. Notice, in particular, that total sample size is excluded from $\mathbf{w}_1$ and $\mathbf{w}_2$ by setting component 5 to 0. The relvariances are excluded from $\mathbf{w}_3$ by setting w_1, w_2, w_3, and w_4 to 0. Thus, we do not include sample size and relvariances in the objective simultaneously even though the general form of the objective defined in section 5 incorporates both. Though we report a limited number of weight choices here, one of the major advantages of the software is the facility to easily manipulate the weights to do sensitivity analysis.

The first three rows of Table 6 list the values of the weighted objective function for the optimal allocations for the weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$. Three other allocations are included in rows 4-6 of the table for comparison. Allocation 4 was actually used in the ECI/EBS in 1992; allocation 5 is an allocation of establishments in proportion to the total employment E_h in each stratum; and allocation 6 is in proportion to the total number of establishments N_h in each stratum. For the latter two allocations, we forced at least two sample establishments to be assigned to each stratum and then redistributed the remainder over the other strata in proportion to E_h or N_h—thus, keeping the same $n_{h,\min} = 2$ constraint as in the nonlinear optimization problems. The first two optima (allocation 1 for w_1; allocation 2 for w_2) are substantial improvements over the other allocations in terms of objective function value. The rule-of-thumb allocations—proportional to E_h and proportional to N_h—are notably worse than the optima for either weight vector.

Table 7 shows the averages by size class of the optimum values of the second-stage allocations for the three optimal allocations and the 1992 ECI/EBS allocation. The optimal allocations generally call for more occupations to be sampled per establishment. One might expect that the second-stage allocations would be closer to the $\overline{\overline{m}}_{S,\max}$ bounds in Table 2 since there is no penalty in the optimization to keep those maxima from being achieved. However, the contribution of the second stage of sampling to the relvariance of most estimates is small implying that selecting more occupations did not reduce the overall relvariance much. For example, the median percentage contribution of the second-stage to the total relvariance across the 66 estimates considered here was only about 8% for each of the allocations.

Allocation 2 is the optimum when total compensation and benefits are equally weighted in the objective function. Since these are two of the key estimates from the ECI, it is interesting to see how that allocation compared to the others. The *cv*'s achieved by optimum allocation 2 are compared in Figure 5 to those obtained from allocations 1, 4, and 5 in Table 6. The *cv*'s for 63 estimates are plotted. Forty-three of those were constrained in the optimization while the remaining 20 were not. Allocation 2 generally improves over the others for key estimates—in particular the ECI and EBS/100+. In the lower right-hand panel of Figure 4, the allocation proportional to N_h does produce lower *cv*'s for EBS/<100 estimates because the large number of small establishments leads to larger samples being allocated to the <100 strata under the proportional-to-N_h rule. On the whole, allocation 2 not only substantially reduces the value of the objective function for $\mathbf{w}_2 = (1,0,1,0,0)$ compared to the other allocations, but also reduces the *cv*'s for important individual survey estimates.

Because cost is always an issue in survey design, we prefer an allocation that meets survey goals using the smallest feasible sample size. When the sample size is minimized ($\mathbf{w}_3$) and the same constraints are used as for allocations 1 and 2, we obtain an optimum of n=3,672 (allocation 3). Thus, the constraints can be satisfied for an establishment sample that is 84% (3,672/4,360) of the size used for allocations 1 and 2. The cost of allocation 3 would be somewhat more than 84% of allocations 1 and 2 since the $\overline{m}_h$'s in Table 7 are larger for allocation 3. Figure 6 shows the *cv*'s for the sample size minimizing allocation plotted versus those for allocations 1, 2, and the 1992 allocation. There are

increases in individual *cv*'s incurred with allocation 3 as opposed to allocation 2, but allocation 3 is almost as good as allocation 1 and better than the 1992 allocation.

To summarize, allocations 2 and 3 do well in our example, both in terms of the objective function and the *cv*'s for domain estimates. Allocation 2 will have a cost that is similar or slightly higher than the cost of the 1992 allocation. But, if budget cutting is important, allocation 3, with an establishment sample 84% of the 1992 size, will produce domain estimates almost as precise at lower cost.

Ideally, survey design and sample allocation in a periodic survey like ECI/EBS should be a continuing process. The stability of variance component estimates over time should be studied, along with the sensitivity of the optimal allocations to changes in parameter estimates. Predicted relvariances from expression (9) should also be validated by comparing to relvariances estimated directly by replication or some other method.

7. CONCLUSION

Nonlinear optimization can be a powerful technique in sample allocation in multipurpose surveys, but a number of factors have probably limited its use. Despite there being a number of studies in the literature, optimization algorithms themselves may not be well-known to all survey practitioners. The algorithms for handling nonlinear objective functions and nonlinear constraints are complex and would require significant commitments of time and resources to program from scratch. However, commercial versions of some algorithms are available that can be linked into specially written interface software as was done here. At the time that the ALLOCATE software was developed, the GUI building tools were more limited than they are today, which influenced our choice

of the proprietary PV-WAVE™. Now, however, many other options are available that other developers might find preferable—particularly Java™, PowerBuilder™, and C++. The former two allow cross-platform development.

The functionality and use of our system is spelled out in two Bureau of Labor Statistics manuals—*Allocate User's Guide* and *Allocate Programmer's Guide*. These manuals include many more screenshots and more detailed descriptions of how to use the system and write the C program that calls the GRG2 optimizer. Both of these are available from the authors to anyone wishing to develop their own systems.

Optimizers other than GRG2 are also available, many of which are described and compared by Schittkowski (1985), that use penalty methods or sequential quadratic programming (SQP) techniques. Leaver, et. al. (1987, 1996), for example, applied an SQP method developed by Fiacco and McCormick (1968). Although there are performance differences among algorithms, the particular optimizer used here was less important than illustrating the usefulness of a flexible allocation system for sample design. For organizations that conduct a variety of surveys and must periodically redesign to update and improve existing operations, developing systems to facilitate the use of optimization algorithms seems well worthwhile.

APPENDIX

A.1 Model expectation of the design variance of $\hat{m}$

First, consider $E_x\left(S_{2hi}^2\right)$. Under model (8), we have

$$\begin{aligned}\tilde{z}_{hi} &= \sum_{j\in U_{hi}} \tilde{E}_{hij}\left(a_h + b_j + e_{hij}\right) \\ &= a_h \tilde{E}_{hi} + \sum_{j\in U_{hi}} b_j \tilde{E}_{hij} + \sum_{j\in U_{hi}} e_{hij}\tilde{E}_{hij} \quad .\end{aligned}$$

Using $p^*_{j|hi} = E_{hij} / E_{hi\cdot}$,

$$\frac{\tilde{z}_{hij}}{p^*_{j|hi}} - \tilde{z}_{hi} = a_h\left(d_{hij}E_{hi} - \sum_{j'\in U_{hi}}\tilde{E}_{hij'}\right) + \left(b_j d_{hij}E_{hi} - \sum_{j'\in U_{hi}} b_{j'}\tilde{E}_{hij'}\right).$$
$$+\left(e_{hij}d_{hij}E_{hi} - \sum_{j'\in U_{hi}} e_{hij'}\tilde{E}_{hij'}\right)$$
$$= A_j + B_j + C_j, \text{ say.} \qquad \text{(A.1)}$$

Under model (8), each z_{hij} has expectation 0, so that $E_x\left(\tilde{z}_{hij}/p^*_{j|hi} - \tilde{z}_{hi}\right)^2 = \mathrm{var}_x\left(\tilde{z}_{hij}/p^*_{j|hi} - \tilde{z}_{hi}\right)$. The variance of the first term in (A.1) is then $\mathrm{var}_x(A_j) = s^2_{za}\left(d_{hij}E^2_{hi} - 2E_{hi}\tilde{E}_{hi}d_{hij} + \tilde{E}^2_{hi}\right)$ and $\sum_{j\in U_{hi}} p^*_{j|hi}\,\mathrm{var}(A_j) = s^2_{za}\tilde{E}_{hi}\left(E_{hi} - \tilde{E}_{hi}\right)$. The variances of B_j and C_j follow similarly. Combining results yields

$$E_x\left(S^2_{2hi}\right) = s^2_{za}\tilde{E}_{hi}\left(E_{hi} - \tilde{E}_{hi}\right) + s^2_{zb}\left(E_{hi}\tilde{E}_{hi} - \sum_{j\in U_{hi}}\tilde{E}^2_{hij}\right) + s^2_{ze}\left(E_{hi}M_{hi} - \tilde{E}_{hi}\right). \qquad \text{(A.2)}$$

Substituting this expression for the model-expectation of the second term in (7), leads to the second term on the right-hand side of (9).

Turning to S^2_{1h}, we have

$$\tilde{z}_{hi} - \bar{\tilde{z}}_h = a_h\left(\tilde{E}_{hi} - \bar{\tilde{E}}_h\right) + \left(\sum_{j\in U_{hi}} b_j\tilde{E}_{hij} - \sum_{i'\in U_h}\sum_{j'\in U_{hi}} b_{j'}\tilde{E}_{hi'j'}\Big/N_h\right)$$
$$+ \left(\sum_{j\in U_{hi}} e_{hij}\tilde{E}_{hij} - \sum_{i'\in U_h}\sum_{j'\in U_{hi}} e_{hi'j'}\tilde{E}_{hi'j'}\Big/N_h\right)$$
$$= A_i + B_i + C_i, \text{ say.} \qquad \text{(A.2)}$$

Under model (8), $\sum_{i\in U_h}\mathrm{var}(A_i) = s^2_{za}\sum_{i\in U_h}\left(\tilde{E}_{hi} - \bar{\tilde{E}}_h\right)^2$. To compute $\mathrm{var}(B_i)$, define

$$j_{hij} = \begin{cases} 1 & \text{if establishment } hi \text{ contains occupation } j \\ 0 & \text{if not} \end{cases}.$$

With that definition, $B_i = \sum_{\text{all } j} b_j\left(j_{hij}\tilde{E}_{hij} - \sum_{i'\in U_h} j_{hi'j}\tilde{E}_{hi'j}\Big/N_h\right)$ from which it follows that $\sum_{i\in U_h}\mathrm{var}(B_i) = s^2_b\sum_{\text{all } j} V_{hj}$ where V_{hj} was defined at the end of section 3. Recalling that $\mathrm{var}(e_{hij}) = s^2_{ze}/E_{hij}$, the variance of C_i is

$$\text{var}(C_i) = s_{ze}^2 \left[\{(N_h - 2)/N_h\} \sum_{j \in U_{hi}} \tilde{E}_{hij} + N_h^{-2} \sum_{i' \in U_h} \sum_{j' \in U_{hi'}} \tilde{E}_{hi'j'} \right]$$

and $\sum_{i' \in U_h} \text{var}(C_i) = s_{ze}^2 (N_h - 1) \tilde{\bar{E}}_h$. Combining results for A_i, B_i, and C_i gives

$$= s_{za}^2 V_{hE} + s_{zb}^2 \sum_{\text{all } j} V_{hj} + s_{ze}^2 \bar{\tilde{E}}_h. \qquad \text{(A.3)}$$

Finally, from (7), (A.2), and (A.3) we obtain (9).

A.2 Compromise Form of V_{hEM}

A compromise form of V_{hEM} was used because M_{hi} was not available for all establishments. Let $\tilde{\bar{E}}_h$ be the mean number of employees per occupation per establishment in a domain. Suppose that under model (8) we have $\text{var}(e_{hij}) = s_{ze}^2 / \tilde{\bar{E}}_h$ rather than $\text{var}(e_{hij}) = s_{ze}^2 / E_{hij}$, i.e., substitute an overall stratum mean for the establishment-specific mean. Then $\sum_{i \in U_h} \sum_{j \in U_{hi}} p^*_{j|hi} \text{var}(C_j) = s_{ze}^2 N_h V_{2h\tilde{E}E} / \tilde{\bar{E}}_h$ and $V_{hEM} = V_{2h\tilde{E}E} / \tilde{\bar{E}}_h$. This form of V_{hEM} was estimated by method-of-moments.

REFERENCES

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

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

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

Fiacco, A.V., and McCormick, G., *Nonlinear sequential unconstrained minimization techniques*, (Wiley, New York, 1968).

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

Hughes, E., and Rao, J., Some Problems of Optimal Allocation in Sample Surveys Involving Inequality Constraints, *Communications in Statistics ¾ Theory and Methods*, **A8 (15)** (1979), 1551-1574.

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

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

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

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

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

Leaver, S.G., Weber, W.L., Cohen, M.P., and Archer, K.P., Item-Outlet Sample Redesign for the 1987 U.S. Consumer Price Index Revision, *Proceedings of the 46th Session*, Vol. LII, Book 3 (International Statistical Institute, 1987), 173-185.

Leaver, S.G., Johnson, W.H., Baskin, R., Scarlett, S., and Morse, W., Commodities and Services Sample Redesign for the 1998 Consumer Price Index, *Proceedings of the Section on Survey Methods Research*, (American Statistical Association, 1996), to be published.

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

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

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

Schittkowski, K., NLPQL: a FORTRAN subroutine solving constrained nonlinear programming problems, *Annals of Operations Research*, **5**(6) (1985), 485-500.

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

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

Windward Technologies Inc., *GRG2 User's Guide: Software for Solving Nonlinear Optimization Problems with Nonlinear Constraints*, (Windward Technologies Inc. and Optimal Methods, Inc., Meadows TX, 1994).

Zayatz, L., and Sigman, R., Feasibility Study of the Use of Chromy's Algorithm in Poisson Sample Selection for the Annual Survey of Manufacturers, *Proceedings of the Section on Survey Methods Research* (American Statistical Association, 1994), 641-646.

Figure Titles

Figure 1. Log of stratum relvariance components for total compensation plotted versus log of average establishment employment size. The upper panel shows data for first-stage relvariance components; the lower panel is for second-stage components.

Figure 2. Log of predicted and point estimates of stratum relvariance components for total compensation plotted versus log of average establishment employment size for two SIC groups. Solid circles are point estimates used in smoothing ($n_h \geq 10$); dotted circles are point estimates excluded from estimation ($n_h < 10$).

Figure 3. Main window of the ALLOCATE software.

Figure 4. Two branches a user can take in the ALLOCATE program.

Figure 5. Coefficients of variation for optimal allocation 2 plotted versus *cv*'s for alternative allocations.

Figure 6. Coefficients of variation for optimal allocation 3 (*n*=3,672) plotted versus *cv*'s for alternative allocations (*n*=4,360).

Table 1. Important variables and domains from the ECI/EBS.

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 (used for total compensation)	Domains
Full population	All occupations
9 major occupational groups	Professional, technical, related occupations
7 industries	Clerical, sales occupations
	Production, service occupations
	1 size class (100+)

Table 2. Minimum, maximum, and average numbers of occupations to be assigned to pseudo-SIC's in each size class.

Size Class	$m_{h,\min}$	$m_{h,\max}$	$\overline{\overline{m}}_{S,\max}$
< 50	2	6	4
50-99	2	10	8
100-249	4	12	10
250-999	4	12	10
1000+	6	12	10

Table 3. Bounds on *cv*'s for estimates of different totals of employee costs.

Cost	*cv* bound
Total compensation	.03
Total benefits	.02
Paid leave	.03
Insurance	.03
Legally required benefits	.03
Retirement and savings	.03

Table 4. Bounds on *cv*'s of estimates of total compensation for major occupational groups (MOG's) and industry divisions.

Group	*cv* bound	**Group**	*cv* bound
MOG		**Industry**	
A-Professional, technical	.30	Construction	.10
B-Executive	.25	Manufacturing	.05
C-Sales	none	Transportation	.05
D-Administrative support	.20	Wholesale	.10
E-Precision workers	.40	Retail	.10
F-Machine operators	none	FIRE	.10
G-Transportation	.20	Services	.10
H-Handlers	.10		
K-Service	.20		

Table 5. Bounds on *cv*'s of the estimated percentages of employees in different occupational groups receiving certain benefits for establishments with 100+ employees.

Benefit	All occupations	Professional, technical	Clerical, sales	Production, service
Medical insurance	.015	.015	.015	.015
Life insurance	.010	.010	.010	.010
Retirement, savings	.020	.020	.020	.020
Vacation	.010	.010	.010	.010
Sick leave	.020	.020	.020	.020

Table 6. Values of the weighted objective function for three optimal allocations and three comparison allocations. Optimal allocation 1 is optimal for $\mathbf{w}_1$; optimal allocation 2 is optimal for $\mathbf{w}_2$. The total establishment sample size for allocations 1, 2, 4, 5, and 6 is n=4,360;total size for allocation 3 is n=3,672.

Allocation	$\mathbf{w}_1 = (1, 0.5, 1, 0.5, 0)'$	$\mathbf{w}_2 = (1, 0, 1, 0, 0)$
1. Optimal allocation 1	0.263	1.025×10^{-3}
2. Optimal allocation 2	1.064	0.426×10^{-3}
3. Optimal allocation to minimize total sample size	11.875	1.042×10^{-3}
4. ECI/EBS 1992	3.579	1.345×10^{-3}
5. Proportional to E_h	9.415	1.108×10^{-3}
6. Proportional to N_h	7.484	1.914×10^{-3}

Table 7. Average optimum values, across industries, of the second-stage allocation $\overline{m}_h$ for optimal allocations 1, 2, and 3.

	Allocation			
Size Class	1992	Optimal 1	Optimal 2	Optimal 3
< 50	4	4.0	4.0	3.9
50-99	6	6.9	7.2	8.0
100-249	6	9.4	9.5	10.0
250-999	8	9.9	9.9	10.0
1000+	8	9.2	9.3	10.0

Figure 1. Log of stratum relvariance components for total compensation plotted versus log of average establishment employment size. The upper panel shows data for first-stage relvariance components; the lower panel is for second-stage components.

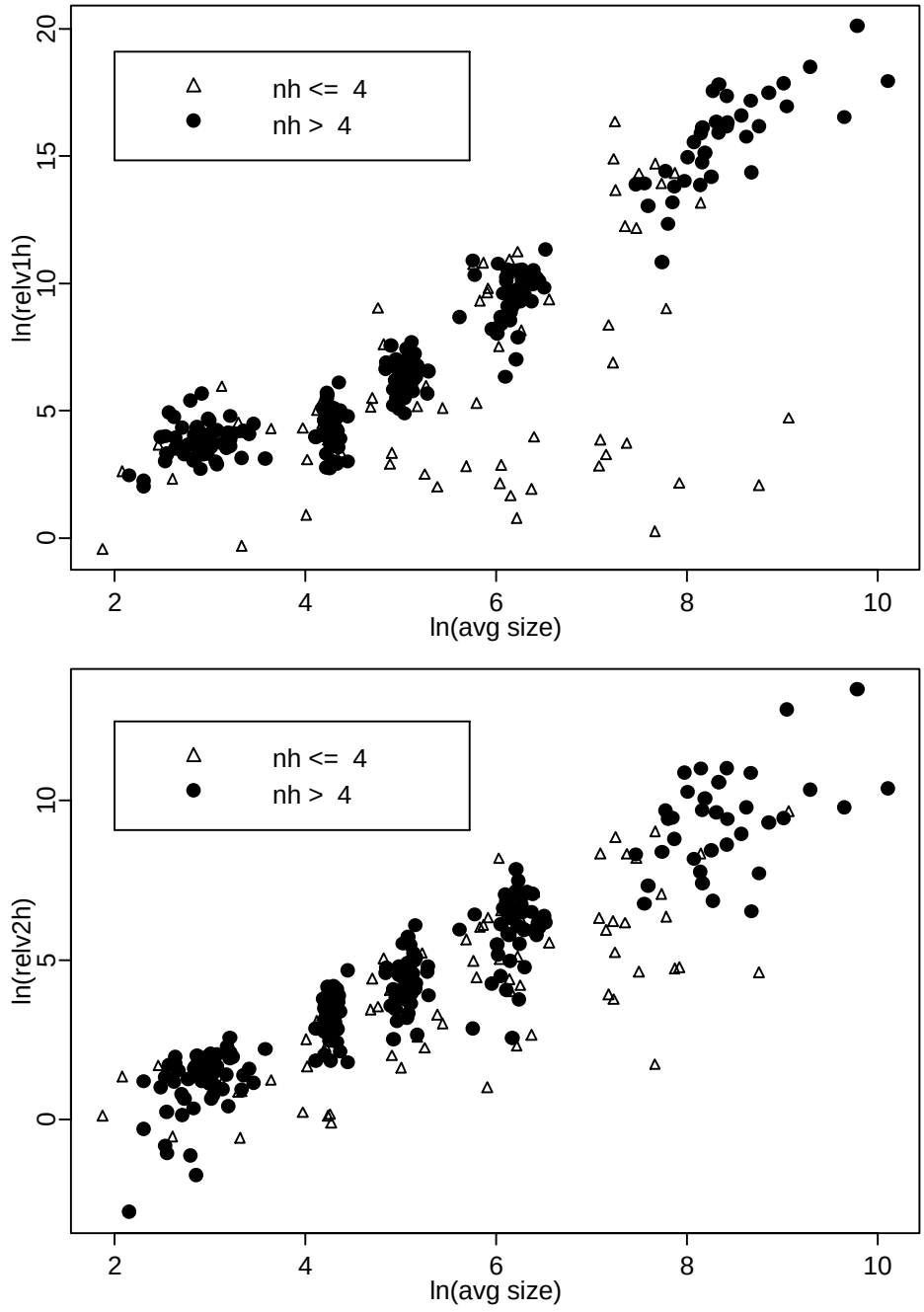

Figure 2. Log of predicted and point estimates of stratum relvariance components for total compensation plotted versus log of average establishment employment size for two SIC groups. Solid circles are point estimates used in smoothing ($n_h \geq 10$); dotted circles are point estimates excluded from estimation ($n_h < 10$).

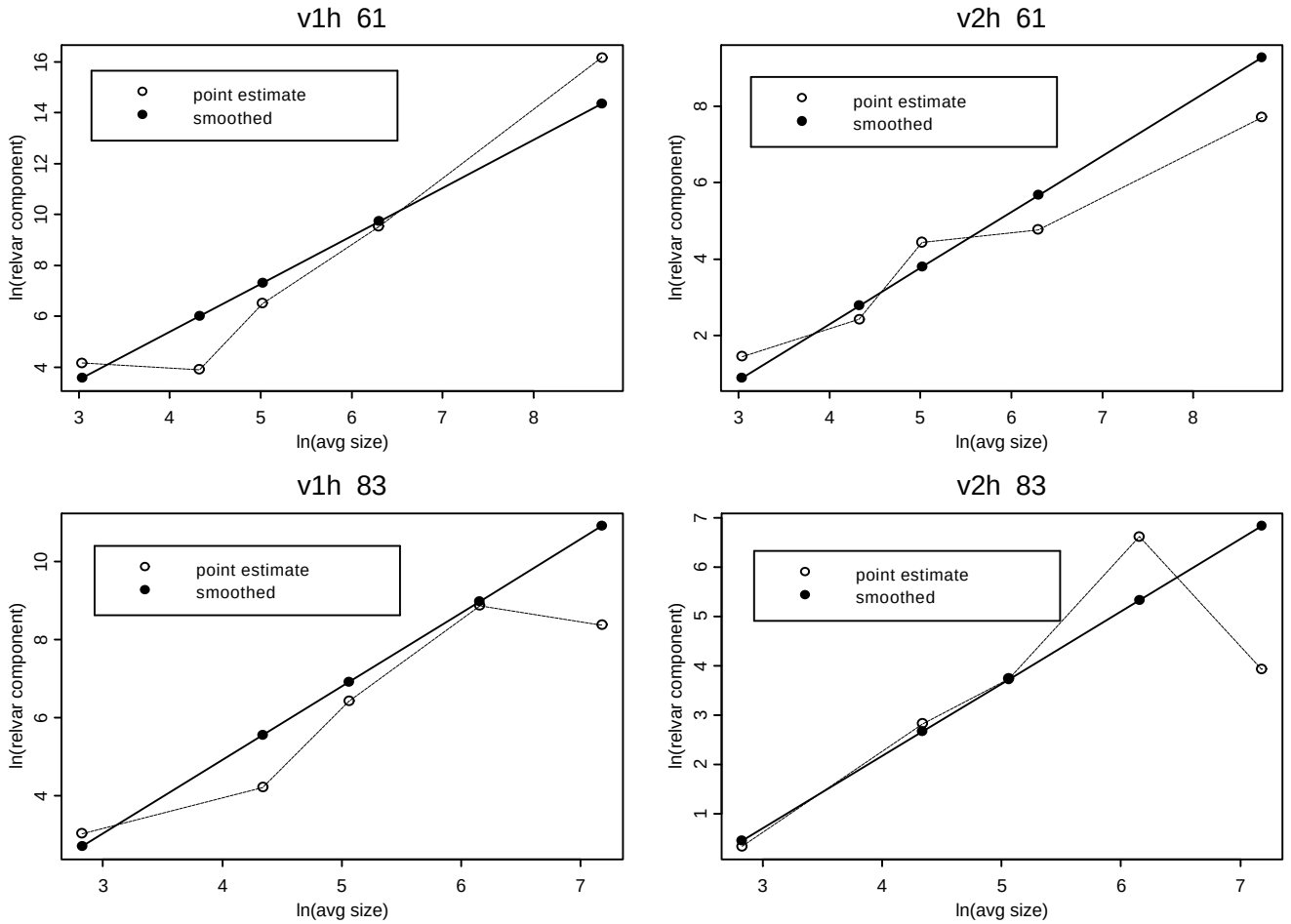

Figure 3. Main window of the ALLOCATE software.

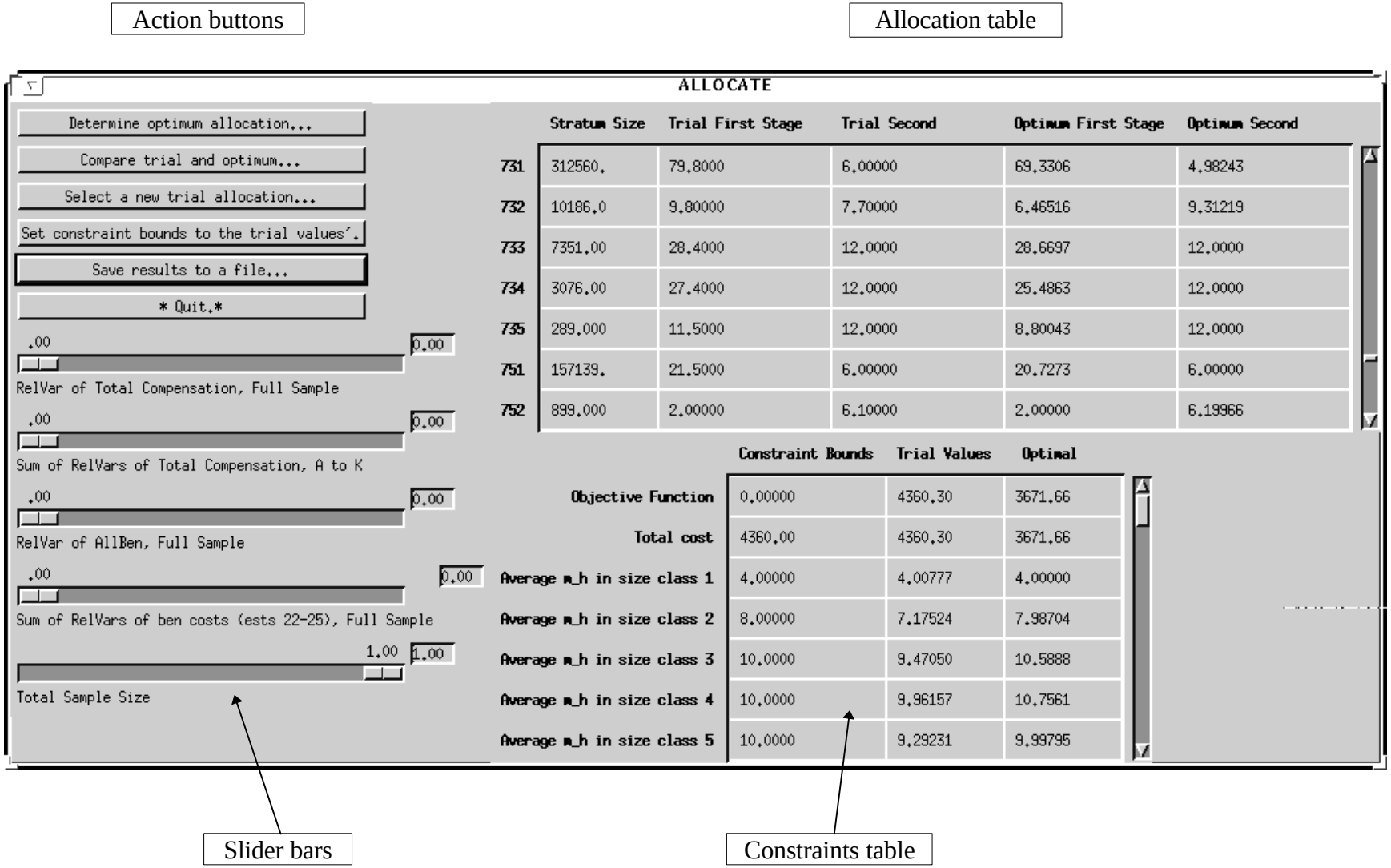

Figure 4. Two branches a user can take in the ALLOCATE program.

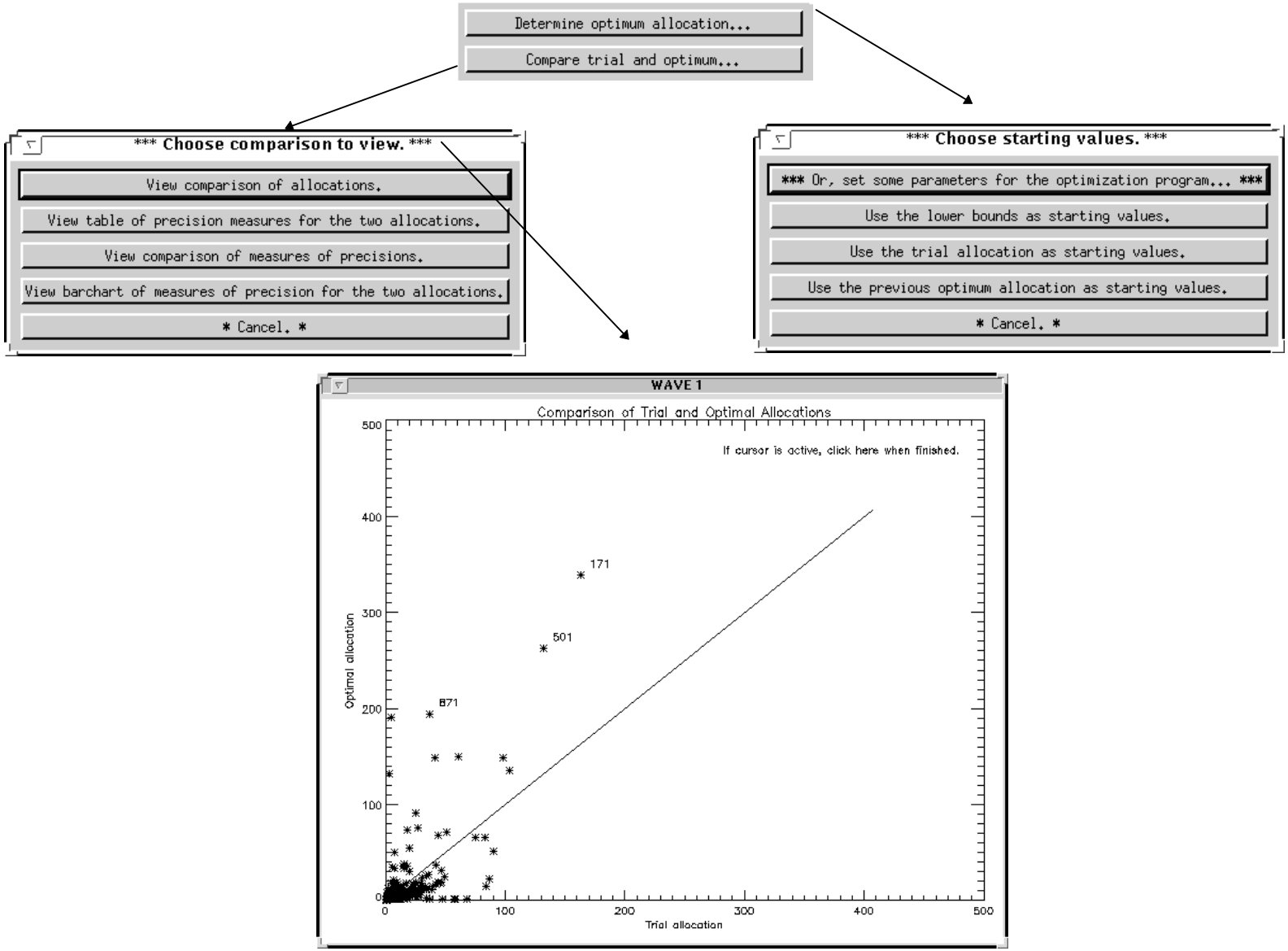

Figure 5. Coefficients of variation for optimal allocation 2 plotted versus cv's for alternative allocations.

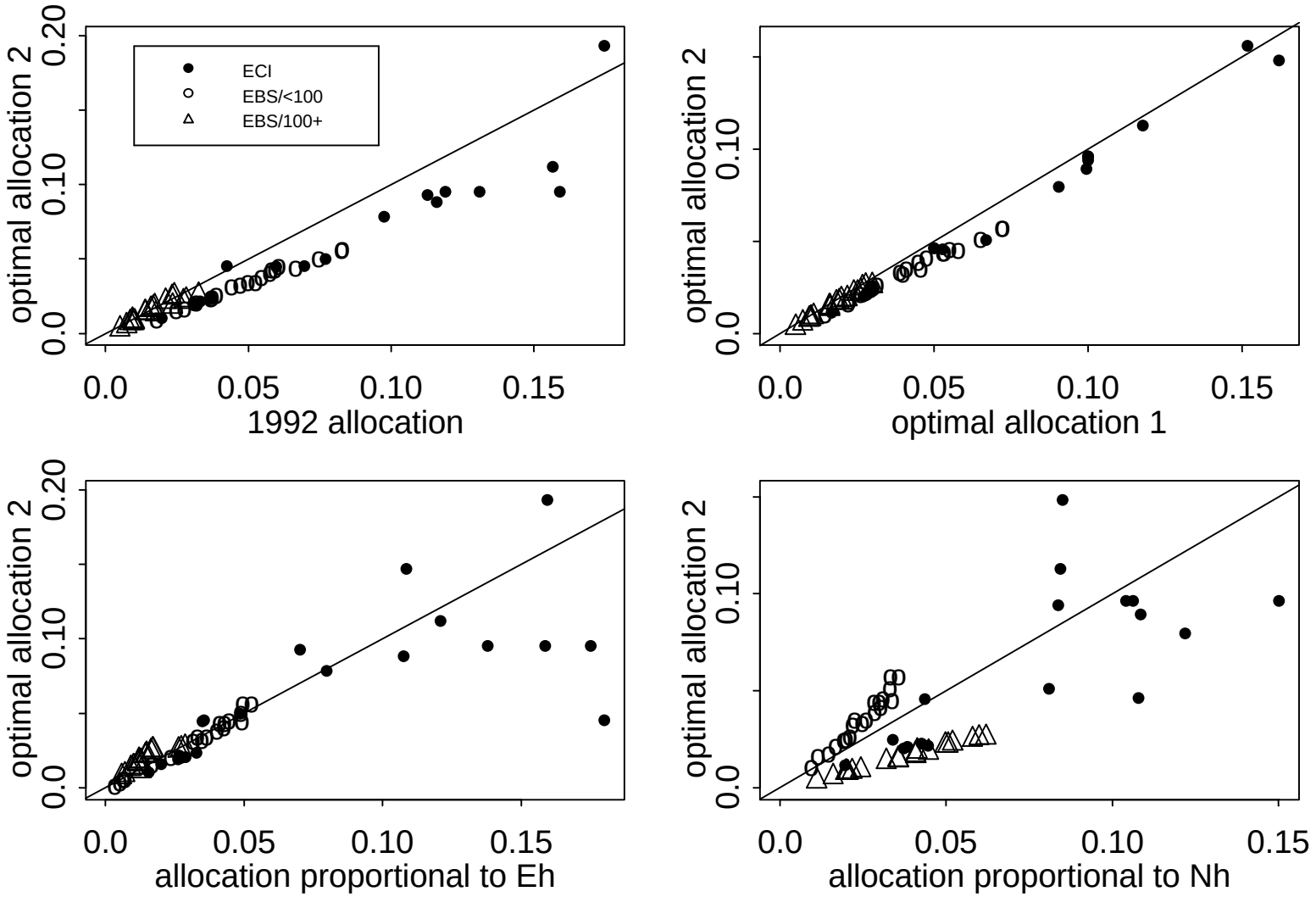

Figure 6. Coefficients of variation for allocation 3 (n=3,672) plotted versus cv's for alternative allocations (n=4,360).