A Re-evaluation of the Statistical Learning Approach to Optimal Sample Allocation

Ismael Flores Cervantes [1]
[1]Westat, 1600 Research Blvd, Rockville, MD 20850

Abstract
Statistical learning is both a theory and a group of algorithms for machine learning. These algorithms detect and learn patterns in data that, in turn, can be used to predict outcomes. These methods have become very popular in recent years and have been successfully applied in many fields. One criticism is that these methods are mainly used as black boxes without a good understating of how they work. Clark (2013) developed a novel approach to optimal sample allocation in stratified design based on a statistical learning approach (SL). However, the SL sample allocation is not without problems when compared to simpler allocations. In this paper we expand the research on the SL approach and re-evaluate its performance. We describe the mechanism behind the SL sample allocation for situations not previously considered.

Key Words: Optimal Sample allocation, statistical learning, Neyman allocation, imperfect design data

1. Statistical Learning Methods

Clark (2013) presents a novel approach to sample allocation based on a statistical learning (SL) approach that combines two sets of design data that are used as a check on each other. The approach is very interesting and shows an innovative application of statistical approaches based on training and validation sets. The article is an example of how methods developed in other fields can be extended to survey sampling methodology. In particular, the article provides an alternative solution to the problem of sample allocation in cases when the variances are estimated.

SL, also known as machine learning, refers to a framework that encompasses a large set of tools and algorithms used for understanding data (James, Witten, Hastie, & Tibshirani, 2013). The origins of these tools are many, and they include statistics, functional analysis, computer science, and artificial intelligence among others. SL theory has led to successful applications and is popular for dealing with Big Data. Despite of their success, the SL methods are not without criticism. Because of their complexity, the algorithms, which are readily available software packages, are seen as black boxes. In other words, there is not a clear understanding of the conditions where these algorithms work or the theory behind these methods. In some occasions, the resulting models can be incomprehensible without providing any insights into the data.

Despite of the popularity of the SL methods, their application in survey methodology has been limited. One example is Buskirk & Kolenikov (2015) who used a nonparametric machine learning technique known as random forests to model nonresponse, nonresponse weighting adjustment factors, and stratification. One reason of the limited application of the SL approach in survey methodology is the underlying *iid* assumption. These methods require the data to be independently and identically distributed, which is not the case of

survey data from complex designs. To overcome this limitation, different *ad hoc* modifications are implemented in order to reflect the sample design or the analysis may assume a simple random sample (i.e., equal weights). However, the SL methods with any of these modifications have not been fully evaluated using data from complex designs. It is expected that theoretical and mathematical developments and the advancement of specialized software in the near future to produce survey data versions of these methods.

2. Statistical Learning Sample Allocation

The sample allocation based on SL ideas proposed by Clark (2013) is applicable to one specific situation, that is, the optimal allocation of a sample for a new implementation of a survey with the following conditions:

A The sample is optimally allocated to achieve the smallest variance of $\hat{Y}_3$, the estimate of the total $Y = \sum_{i=1}^{N} y_i$ at time $t = 3$. In other words, there is only one variable of interest. Other variables may benefit of the allocation as long as they are highly correlated with the variable y_i.

B The population can be stratified at two levels. In the first level, the population is stratified into groups g, $g = 1,\ldots, G$. In the second level, these groups or main strata are further stratified into substrata within group. These substrata within group are denoted by h, $h = 1,\ldots, H$. It is assumed that the substrata naturally form into groups. Examples of these groups and substrata are census regions and states within region. For simplicity of notation, we assume the same number of substrata h within a group g.

C The same survey has been previously implemented in two occasions or periods denoted as $t = 1$ and $t = 2$, although the previous implementations are not necessarily used with the same stratification and sample allocation. As an example, the survey was conducted in the two previous years.

D The subpopulations that correspond to stratification by groups and substrata described in B are identifiable in the samples from two previous periods so estimates of population variances for groups, substrata, and total population variance can be computed. These estimates of the different population variances for both periods $t = 1$ and 2 are needed for the algorithm.

E The strata population sizes N_{gh} are known for all substrata and groups in period $t = 3$.

F The total sample size n in period $t = 3$ is fixed, and there are no restrictions in the size that can be allocated in stratum in period $t = 3$.

G The costs of sampling the units are the same for all strata.

H The correlations of the population variances by substratum within groups across periods can be nonzero (i.e., there are changes in variance within stratum across periods).

I Finite population factors can be ignored.

Some of these conditions are very restrictive and limit the application of the type of allocation in most surveys.

The challenge in this problem is finding the best way to combine the information from the two previous periods optimally to allocate the sample in period $t = 3$. In practice, the

population variances for the formulas for optimal allocation are not available so they need to be estimated from the data. These estimates are substituted or "plugged" into the optimal allocation formulas (i.e., plug-in allocation). However, sometimes these estimates are not precise, and the achieved allocation can be less efficient than proportional allocation.

There are different ways to approach this problem; however, Clark (2013) observed that the data from the two previous implementations lend themselves naturally to be used as "training" and "validation" data sets under the SL paradigm. Under the SL approach, the sample can be allocated as follows:

Let $\hat{V}_t$, $\hat{V}_{tg}$, and $\hat{V}_{tgh}$ be the estimates of the population variance, group variance, and substratum variance respectively for periods 1 and 2 (i.e., $t = 1$ or $t = 2$). The sample size at period 3 under the SL allocation is the vector $\mathbf{n}_{SL,3} = \left(n_{3,1},...,n_{3,NG}\right)^t \in \mathbb{N}^{GH}$ with values of the stratum sample sizes $\mathbf{n}_{SL,1} = \left(n_{SL,1,1},...,n_{SL,1,NG}\right)^t$ such that

$$\mathbf{n}_{SL,3} = \min_{\mathbf{n}_{SL,1} \in \mathbb{N}^{GH}} \left(\hat{V}_{SL,2}\right);$$
$$\text{subject to } \sum_{g=1}^{G}\sum_{h=1}^{H} n_{SL,1,gh} = n,$$

where $\hat{V}_{SL,2}\left(\hat{Y}_2\right) = \sum_{g=1}^{G}\sum_{h=1}^{H} \frac{N_{gh}^2}{n_{SL,1gh}} \hat{V}_{2gh}$ is the estimate of the variance of the estimate of total of Y_2, $\hat{Y}_2$, at period 2 computed using the stratum samples sizes $\mathbf{n}_{SL,1}$. The components of $\mathbf{n}_{SL,1}$ are computed as $n_{SL,1,gh} = kN_{gh}\sqrt{\lambda_1\hat{V}_{1gh} + \lambda_{21}\hat{V}_{1g} + \lambda_{31}\hat{V}_1}$ subject to $\sum_{i=1}^{3}\lambda_i = 1$, and k is the constant of proportionality defined as $k = n_3 \Big/ \sum_{g=1}^{G}\sum_{h=1}^{H}\left(N_{gh}\sqrt{\lambda_1\hat{V}_{1gh} + \lambda_{21}\hat{V}_{1g} + \lambda_{31}\hat{V}_1}\right)$ where N_{gh} is the population size in stratum gh.

A more intuitive way to describe the SL sample allocation follows. First, we propose values of $\boldsymbol{\lambda} = \left(\lambda_1, \lambda_2, \lambda_3\right)$ such that $\sum_{i=1}^{3}\lambda_i = 1$ and use these values to compute the sample size $\mathbf{n}_{SL,1}$ using the estimates of population variance from period $t = 1$[1]. Then, using these values of $\mathbf{n}_{SL,1}$, we estimate of the variance of the estimate of the total, $\hat{V}_{SL,2}$, using the estimates of stratum variances in period $t = 2$. We repeat the same computations for different values of $\boldsymbol{\lambda}$. The sample sizes for SL allocation $\mathbf{n}_{SL,3}$ for period $t = 3$ correspond to the values of $\mathbf{n}_{SL,1}$ that minimize the value of $\hat{V}_{SL,2}$ among all possible values of $\boldsymbol{\lambda}$. Mathematically, this is a constrained nonlinear optimization problem where an objective function is minimized subject to multiple constraints. Because of the nonlinear nature of the objective function, the sample sizes under the SL sample allocation are solved numerically. Clark (2013) provided a R package called *Robustallocation* that calculates the sample sizes under the SL allocation.

[1] Note that these sample sizes are not those used in $t = 1$ or 2.

One of the core expressions in the algorithm is the formula for the sample sizes $\mathbf{n}_{SL,1}$. The expression of the components of $\mathbf{n}_{SL,1}$ can be rewritten as $n_{SL,1,gh} \propto N_{gh}\sqrt{\hat{V}^{*}_{1gh}} = N_{gh}\hat{S}^{*}_{1gh}$ where $\hat{V}^{*}_{1gh}$ is the weighted average of the estimate of population variance at the substratum gh $\left(\hat{V}_{1gh}\right)$, at the group level g $\left(\hat{V}_{1g}\right)$ or for the complete population $\left(\hat{V}_{1}\right)$. If these were not estimates, then for $\boldsymbol{\lambda} = (1,0,0)$, we have the expression for Neyman allocation (i.e., the theoretical lowest variance for the estimate of population total Y) (Cochran, 1977). Similarly, for $\boldsymbol{\lambda} = (0,0,1)$ the expression is the same as for proportional allocation. Finally, for $\boldsymbol{\lambda} = (0,1,0)$, the sample is allocated so the variance of the total is between the variance from Neyman allocation and the variance from proportional allocation $\left(\text{i.e., } V_{1gh} \le V_{1g} \le V_{1}\right)$. The rationale behind the inclusion of the estimates of population variances $\hat{V}_{1g}$ and $\hat{V}_{1}$, which both cause departures from the Neyman allocation, is that the estimate $\hat{V}_{1gh}$ may not be as precise as $\hat{V}_{1g}$ or $\hat{V}_{1}$ and putting more weight on these estimates of group variance or population variances will produce estimates of totals at least as efficient as proportionally allocating the sample.

3. Evaluation of the Statistical Learning Sample Allocation

As in most SL applications, the SL sample allocation follows a hands-off approach. The practitioner provides the estimates of the population stratum variances and group indicators from the two previous periods, and the desired total sample and the algorithm automatically determine the optimal sample for $t = 3$. To evaluate the properties of the SL allocation, Clark (2013) compared estimates of variance obtained from the SL allocation and other competing allocations through a simulation study. Based on the simulations, he concluded that the SL sample allocation is superior to other allocations, it can handle allocations based on estimates of population variances computed with very small sample sizes, and it can reflect the correlation of the variances across periods. In this section, we examine these claims focusing on both the evaluation and the competing allocations compared to the SL allocation. We also extend the study to include other competing allocations and expand the simulations to include larger sample sizes.

3.1 Competing Sample Allocations

Clark (2013) showed that the SL allocation produced estimates of totals with lower variance (i.e., more efficient) than those from two competing allocations, both based on data from one period. It is not clear why the SL sample allocation that combines data from two periods was compared to allocations that only use one period. Since the SL allocation ingeniously combines the information from two periods using more complex operations, it would be expected to be more efficient than allocations that use only half of the available data. In this re-evaluation, we compare the SL allocation to other allocations that use all available data. The simplest candidates are those allocations based on plain averages of variances from the two periods. These allocations are referred to as "naive" because the sample allocation is not the result of an optimization or any other complex algorithm when combing the data from the two available periods.

A case can be made for the evaluation of naive allocations. First, combining data from multiple periods is what survey practitioners usually do in practice; in particular, when the data are limited (i.e., two data points are better than one). Second, the naive allocations are easily computed and do not require specialized software, complex mathematical operations, searches, or any optimization as the SL allocation.

A second research question is to determine if the optimization in the SL allocation produces estimates that are as efficient as those estimates from a Neyman allocation based on the variances in period 3 (i.e., lowest variance achievable in period 3). Clark (2013) uses the variance from proportional allocation as a benchmark so the proximity of the variance of estimates based on both SL allocation and competing allocations to the theoretical minimum variance could not be evaluated.

As part of the re-evaluation, we extended the simulation in Clark (2013) to include three additional allocations listed in Table 1. The first three allocations in the table were evaluated in the original article. The last three allocations, H12, G12, and H12G12 are proposed in this study. These allocation use simple average of estimated variances from both periods as described in the table. The justification for allocation H12G12 is based on the idea that the instability in the estimates of variances can be reduced by averaging them at all available levels (this is the "*throw in everything but the kitchen sink*" allocation[2]).

Table 1: Sample allocations

Allocation	***Description***
SL	Supervised learning using estimated variances in periods 1 and 2
H2	Neyman plug-in with estimated stratum variances computed using period 2
G2	Neyman plug-in with estimated group variances computed using period 2
H12	Neyman plug-in with the average of estimated stratum variances in periods 1 and 2
G12	Neyman plug-in with the average of estimated group variances in periods 1 and 2
H12G12	Neyman plug-in with the average of estimated stratum variances and estimated group stratum variances in periods 1 and 2

3.2 Simulated Populations

In this study, the same artificial populations found in Clark (2013) were simulated as part of the evaluation of the new allocations. These populations or data files are the Business population, Farm population, and the New Zealand Pacific population (Table 3 in Clark, 2013, and Tables A1 and A2 in the online Appendix). Additional details of the simulation are described in the Clark (2013). Here we expand the description of the artificial populations highlighting both the population characteristics and simulation parameters that explain the results observed in the simulations.

The population variances are generated using a log linear model with parameters for group and stratum effects. These parameters affect the correlation of group variances $\left(\rho_{group}\right)$ and stratum variances $\left(\rho_{stratum}\right)$ across periods. The correlation of group variances was set to 1 for the Farm and Business population and 0.98 for the New Zealand Pacific population. In the Business population, the stratum correlation of the

[2] In a way, the SL allocation throws in everything too because it uses the plug-in Neyman stratum, group, and proportional allocations.

variances between periods is so high $(\rho_{stratum} = 0.99)$ that the variance structure and values are essentially the same in the three periods. Since there are almost no changes in stratum variances and group variances from one period to another, the allocation from either period should converge to the population Neyman allocation in period 3 as the sample size increases. In the second population, Farms, the correlation of the stratum variances from period to period is high $(\rho_{stratum} = 0.89)$, but the variance structure is not exactly the same in the following period despite the modest change. In this case, unless the allocation recognizes these small changes, we know that estimates based on the allocation will not achieve the variance achieved by the Neyman allocation computed using the population in period 3. In the last population, the New Zealand Pacific population, the variance correlation between periods is low $(\rho_{stratum} = 0.44)$.[3] That is, the variance structure in the last period is very different from the previous two. In this case, unless this change is recognized, any allocation will be far from the theoretical lower bound or Neyman variance allocation in period 3. Clark (2013) reported that the SL allocation is able to recognize the variances correlations across periods; that is, the efficiency of this allocation is not only better than the efficiency of the naive allocations but also closer to the Neyman allocation in period $t = 3$. With the results from the additional allocations and expanded samples sizes we revisit these observations in Section 3.

In this paper, we examine the efficiency of allocations under a wider ranges of sample sizes that include larger values than the seven sample sizes studied in Clark (2013) shown in Table 2. In this re-evaluation, 40 simulated effective sample size values between 1 and 100 were used in the Business and Farm populations. For the New Zealand population, the 20 sample sizes between 1 and 100,000 were evaluated. The goal of extending the sample sizes in the simulations is the creation of a plot with curves that describe how the allocations behave as the sample size increases. As the sample size increases, the estimates of variance are more precise and there is no need of the SL allocation. The simulation study was implemented in R version 3.0.0 (R Development Core Team, 2013), using a modification of the code available in the supplemental materials of Clark (2013).

[3] There is a discrepancy with what is stated in Clark (2013) and what is coded in the simulation program available in the supplemental material. The simulation program uses $\rho_{stratum} = 0.28$ and the results match those in Table 3 in the article despite that Table 1 shows the value $\rho_{stratum} = 0.44$. In this study, we assume that $\rho_{stratum} = 0.28$ is the correct correlation. However, independently of the value $\rho_{stratum}$ that was actually used, the point is that the stratification is very different from period to period for this population.

Table 2: Populations, sample sizes, and design effects analysed in Clark

Artificial Population/data	***Nominal Sample Size*** n_h	***Effective Sample Size*** n_h^{eff} ***(Estimated d.f./parameter*** $\hat{d}$ ***)***	***Design effect (Deff)***
Business	6	1.7	3.5
	10	2.9	3.4
	20	6.0	3.3
Farms	6	3.2	1.9
	10	5.6	1.8
	20	7.9	2.5
New Zealand	∞	*N/A*	*N/A*

Source: Clark, R. G. (2013). Sample design using imperfect design data. *Journal of Survey Statistics and Methodology, 1*(1), 6-23.

3.3 Results of the Simulation

The results of the expanded simulation study are shown in the plots in Figures 1, 2, and 3. The plots show lines for the ratios of expected achieved variances of estimates of totals from the different allocations to the same variance achieved using proportional allocation in period 3 with different effective sample sizes (logarithmic scale). These are the same ratios computed by Clark (2013) and mathematically correspond to the estimate of $E\left(\hat{V}_{allocation,3}\right)/V_{proportional,3}$ where this expected value is computed using simulation. The ratio for SL allocation corresponds to the black line in the plots. We are interested in the ratios for the naive allocations H12 (orange line) and H12G12 (purple line) that compute the estimated stratum variances as the average of the estimated stratum variances in periods 1 and 2. All other allocations are included for reference. Better allocations correspond to lines with lower ratios in the plots (i.e., estimates of totals with smaller variances compared to estimates of totals based on proportional allocation).

The plots include three horizontal dotted lines, the first line (red) at 1 is the reference line (i.e., it corresponds to the ratio of variance of proportional allocation to proportional allocation).[4] The second horizontal dotted line (green) corresponds to the ratio of the variance achieved using the group allocation to the variance achieved using proportional allocation. The last horizontal dotted line (black) is the ratio of the variance of theoretical Neyman at period 3. This is the lower bound and no allocation can achieve this value unless they match exactly the value of the population variances at period 3 in each stratum. The line for the ratio of the variances for the group allocation is always between the proportional allocation line and Neyman allocation line.

To verify that the results from the article are reproduced in this re-evaluation, the plots include vertical lines for the sample sizes in Table 3 in the article (i.e., the value of the ratio of variances achieved by the proposed allocation to proportional allocation for nominal sample sizes n_h = 6, 10 and 20 for the Farm and Business data). Although Table 3 in Clark (2013) shows entries for these nominal sample sizes, the simulation program in the supplemental materials does not use these nominal sample sizes directly. Instead, the program uses the effective sample size. Clark (2013) refers to the effective sample size as parameter $\hat{d}$ or estimated d.f. (degrees of freedom) and their values are

[4] The horizontal red dotted line at 1 is not shown in Figure 1 because it outside the vertical axis range.

shown in Table 2 in the article. In other words, Clark (2013) reflects losses in precision of the estimates of the stratum variances by using the smaller effective sample size instead of the nominal sample size in the simulation and in the results in Table 3 in the article. Since the plots show the effective sample sizes, we need to convert the nominal sample size to the equivalent effective size in order to identify the table entries in the plots. The last column in table 3 shows the correspondence between the nominal sample sizes and effective sample for the Business and Farm data.

For example, the table entry for the nominal sample size $n_h = 6$ for Business, corresponds to vertical line with an effective sample size $n_h^{eff} = 1.7$ in the plot in Figure 1 (dotted red line). Using this table, we can identify corresponding effective sample sizes for nominal sample sizes $n_h = 6$, 10, and 20 indicated by the red, blue, and green vertical dotted lines respectively in Figures 1 and 2. Notice that effective sample size (or estimated d.f. as it is called in Clark) for $n_h = 20$ for the Farm data has a value of 7.9. This value is hard-coded in the simulation, and the computed ratio of variances using this value matched the results in Table 3 in the article. However, Table 2 in the article shows a value of 11.7 for the same sample size $n_h = 20$ for Farms. This is likely to be an error in the simulation because the design effect (*Deff*) for $n_h = 20$ (i.e., 2.5) is very different from the design effect for $n_h = 6$ and 10 for Farms (*Deff* = 1.8 and 1.9 respectively, see table above). This difference stands out when we compare the differences among design effects in Business (i.e., *Deff* = 3.5, 3.4. and 3.3). Despite these inconsistencies, we assume that this is a typo in Table 2 in the article and the value used in the simulation is correct.

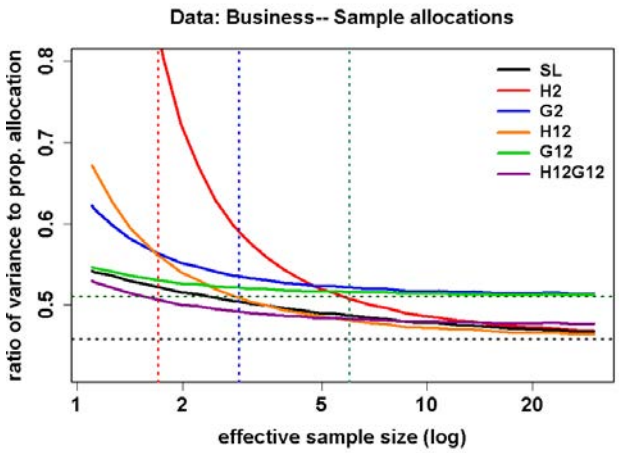

Figure 1: Variance ratios for Business population

As shown in Figure 1, the H12G12 allocation (purple line) does better than the SL allocation (black line) for all table entries for Business data. The figure also shows that combining periods 1 and 2 in H12 allocation (orange line) has a large impact on the ratio of variance of the allocation. Ignoring one period greatly favors those allocations that use both periods. The H12 allocation achieves almost the same variance reduction as the SL allocation except for $n_h = 6$ or $n_h^{eff} = 1.7$ (vertical red dotted line). The plot also shows that those allocations based on groups will never achieve the lowest variance as the sample size increases (blue and green lines). An interesting observation is that the plot shows that very modest sample sizes can achieve variances close to the optimal allocation in the situations similar to Business data (when the variances do not change from period to

period). This is not what is generally shown in the literature as reported in Clark (2013) and it may a consequence of the artificial populations.

Figure 2 shows the same ratios for the allocations in Table 1 for the Farm data where there are modest changes in the stratum variance from period to period (i.e., $\rho_{stratum} = 0.89$). The ratio of the H12 allocation ratio (orange line) is lower than the SL allocation (black line) for all table entries (vertical dotted lines). In the Farm data, despite the modest changes in the stratum variance, none of the allocations achieves the Neyman variance for period 3 even with very large effective sample sizes. However, the H12 allocation that ignores the group population achieves the lowest variance for the larger effective sample sizes shown in the plot. This shows that small changes in the variances throughout the periods have a large impact on the maximum reduction that can be achieved. None of the allocations recognized the correlation of variances across years. Still these reductions are large compared to proportional allocation.

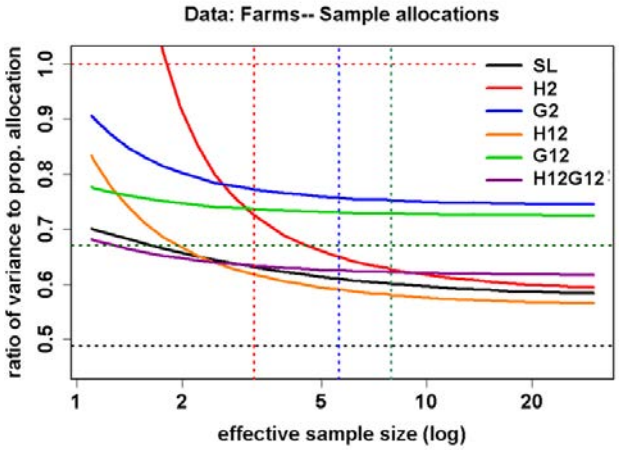

Figure 2: Variance ratios for Farm population

Figure 3 shows the variance ratios for the New Zealand Pacific population and unlike the Farm and Business data, it assumed very large samples sizes per stratum (the simulation program of the article uses a value of Infinity for estimated d.f.). This is the reason why there is only one entry in Table 3 in Clark (2013) without different values of n_h. However, for the re-evaluation, we created a plot using large sample sizes that match the entries of Table 3. The New Zealand data has the largest changes in stratum variance from period to period. Since none of the allocations uses the correlation from period to period, none will achieve the minimum possible variance (horizontal line at 0.6). However, any allocation that takes advantage of the stable correlation for groups will benefit from it (population group variance indicated by horizontal green line at 0.83). The plot in Figure 3 shows that the allocation H12G12 (purple line) does better than the SL allocation (black line). This plot also shows the poor performance of the 1-period Neyman plug-in allocation H2 (red line), which is never competitive. On the other hand, the version that uses the 2 periods (H12, orange line) is a better option.

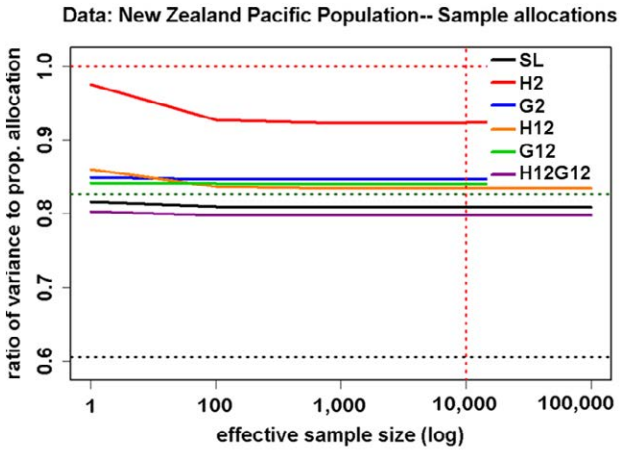

Figure 3: Variance ratios for New Zealand Pacific population

4. Discussion

There are two factors that help the naive allocations achieve similar reductions in the ratios of variances as those observed in the SL allocation in the simulations. First, when the allocations use two periods, the estimates of variance are more precise because the larger sample used in the estimation. Since the simulation on the original article for Business and Farms are based on very small samples, combining two periods makes a big difference. These gains are not observed in larger sample sizes. The second factor, which also helps the SL allocation, is that large reductions in variance were achieved with modest sample sizes in these artificial populations. In other words, the range of sample sizes where these ratios of the allocations are different is not wide. The remaining reductions in variance can be explained by the way the different allocation pays attention to the parameter that remains constant across periods, in this case, the group variances.

Examining all scenarios, there is in most cases a naive allocation based on 2 periods that does better than the SL allocation (i.e., the SL allocation is not consistently better across all scenarios as reported in Clark, 2013). However, in general, the further reductions achieved by the allocations are marginal. The main point of this discussion is that more naive allocations with the same sources of information can be as efficient (or more) as the SL sample allocation. Since the naive allocations are easier to compute without any mathematical complexities and specialized software, they may be a better alternative than the SL allocation for many applications. Furthermore, previous claims such as gains being greatest when the autocorrelations of the true strata variances are weak or the stratum degrees of freedom are small are not inherent to the SL allocation because the same gains are achieved and sometimes surpassed by naive allocations that use all available information.

There are other aspects of the evaluation not discussed in the Clark (2013) where other allocations have advantages over the SL allocation. First, we note that the sample sizes $n_{SL,1,gh}$ in $\mathbf{n}_{SL,3}$ are random variables (i.e., $\hat{\mathbf{n}}_{SL,3}$) and have an associated variability or variance that should be taken into account in the evaluation. This variance of the sample sizes depends on the estimates of population variances for group and substrata from the previous periods. However, rather than computing the variance for each component of

$\hat{\mathbf{n}}_{SL,3}$, we could estimate the variance of the ratio of the reduction of variances achieved in the allocations (i.e., $V\left(\hat{V}_{allocation,3}\left(\hat{Y}\right)\right)/V_{proportional,3}\left(\hat{Y}\right)$). This variance summarizes the effect of the variability in estimating the components of $\hat{\mathbf{n}}_{SL,3}$. Since this variability changes as the sample size increases and it is not symmetric, we instead plot the percentile bands (shaded areas) created using the 2.5th and 97.5th percentiles of the distribution of the ratio of achieved reduction of variances for the different allocations under repeated sampling. The percentile bands for the Business, Farm, and New Zealand populations are shown in Figures 4, 5, and 6. More precise allocations under repeated sampling have narrower percentile bands (i.e., smaller $V\left(\hat{V}_{allocation,3}\left(\hat{Y}_3\right)\right)$). For the Business populations where the correlation of variances across periods is almost one, the bands are very small because the variances do not change across periods and the estimate of the sample size is close to the expected value under repeated sampling. The behavior of these bands is very different for the Farm and New Zealand populations as shown Figures 5 and 6. For Farms in Figure 5, the SL allocation (black shaded area) is less precise (i.e., a wider percentile band) than the naive allocations H12 and H12G12 (orange a purple percentiles bands) for small effective sample sizes. The bands become smaller as the sample size increases. In these cases, it is preferable to use the more naive allocations for, n_h = 6, 10, and 20, which correspond to vertical lines at n_h^{eff} = 3.2, 5.6, and 7.9, respectively. The wider percentile bands can be explained by the fact that more parameters are estimated in the SL allocation. These parameters are $\hat{\boldsymbol{\lambda}} = \left(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3\right)$. In contrast, the naive allocations do not have this source of variability. For the New Zealand population in Figure 6, the bands are constant despite the large sample sizes, that is, the estimates of the sample size are not consistent.

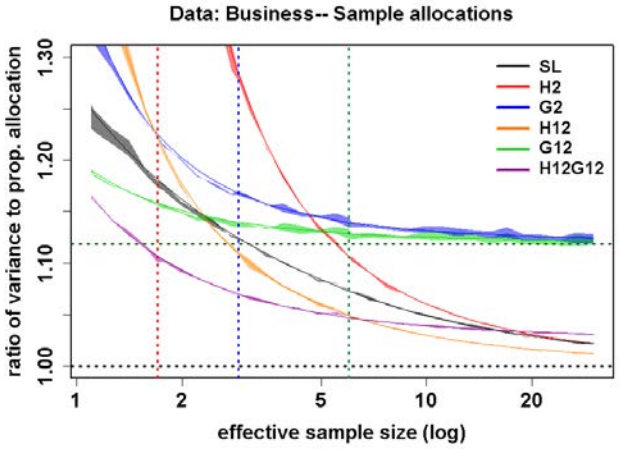

Figure 4: 2.5th and 97.5th percentile bands of achieved reduction of variance for different allocations in the Business population

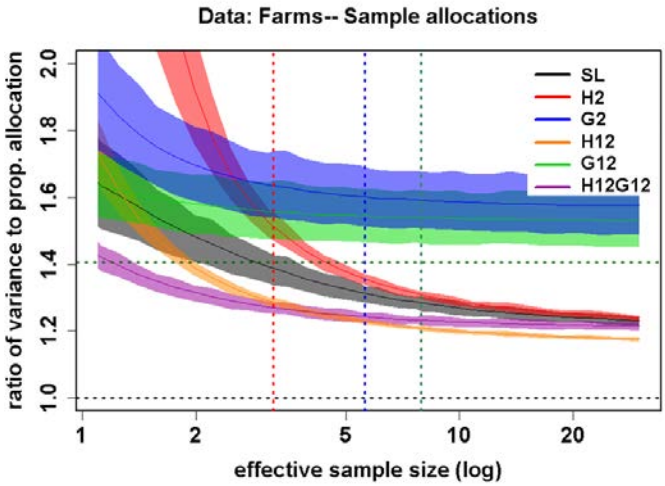

Figure 5: 2.5th and 97.5th percentiles bands of achieved reduction of variance for different allocations in the Farm data

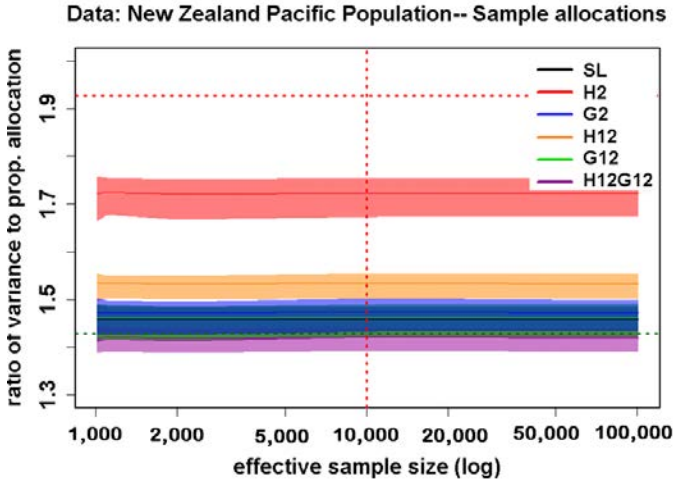

Figure 6: 2.5th and 97.5th percentiles bands of achieved reduction of variance for different allocations in the New Zealand population

Most of the simulation scenarios studied in Clark (2013) have very small samples for computing the stratum variances as indicated in Table 2. This can introduce a bias because the sample size is not a linear function of the estimated variances. As a result, $E\left(\hat{n}_{SL,1,gh}\right) \neq kN_{gh}\sqrt{\lambda_1 E\left(\hat{V}_{1gh}\right)+\lambda_{21}E\left(\hat{V}_{1g}\right)+\lambda_{31}E\left(\hat{V}_1\right)}$, that is, $\hat{n}_{SL,1,gh}$ is biased for small samples. The biases in the sample sizes cause losses in efficiency because they cause departures from the optimal allocation. This is observed in Figures 1 and 2 in the curved lines for the ratios of variance reduction for all allocations. As the sample size increases, the curves converge to the value of the theoretical ratios. The SL and naive allocations do not remove this bias. As observed in the plots for Farm and Business populations, the SL and all other allocations do not account for the bias in small samples. A possible improvement to the SL allocation that addresses the problem of the biased estimates of the population variances estimated with small samples is to compute the sample size as

$\hat{n}^*_{SL,1,gh} = k' N_{gh}\left(\lambda_1 \hat{S}^*_{1gh} + \lambda_{21}\hat{S}^*_{1g} + \lambda_{31}\hat{S}^*_1\right)$, where $\hat{S}^*_{1gh}$, $\hat{S}^*_{1g}$, and $\hat{S}^*_1$ are the bias-corrected estimates of the population standard errors for the substratum, group, and total respectively and k' is the constant of proportionality. The bias-corrected population standard deviation is computed as $\hat{S}^* = \hat{S} \,/\, c_4(m)$ where $c_4(m)$ is the bias correction factor defined as $c_4(m) = \sqrt{2/(m-1)}\,\Gamma(m/2)/\Gamma\left((m-1)/2\right)$ where m is the sample size used to estimate $\hat{S}$. Under normality, the estimate of the population, group, or stratum standard error is unbiased and the bias of $\hat{n}^*_{SL,1,gh}$ should be smaller than the bias of $\hat{n}_{SL,1,gh}$. This change to the SL allocation will be evaluated in a further study. Since it addresses the bias for small samples directly, it expected to perform better than the allocations studied in this paper.

The SL approach produces an allocation but do not necessarily provide insight about how the sample is allocated because of the "black box" methodology. As mentioned before, the components of $\boldsymbol{\lambda}$ in the expression of the SL allocation are weighting factors (i.e., composite factors) that combine the population group variance, stratum variance, and proportional allocation variance. However, the mechanism that assigns the values to these factors is not easily observed (see the ternary composition plots in Figure 2 in Clark, 2013). For example, it is not clear under which conditions one variance receives a larger factor than another. To better understand how the SL allocation works and explore these conditions, we examined the expected values of λ's for the same range of effective sample sizes in the expanded simulations. That is, similar to the sample sizes, we recognize that the composite factors are also random variables, and their expected value $E\left(\hat{\boldsymbol{\lambda}}\right)$ for different sample sizes can be observed using the same simulations. The results of this analysis for the Business, Farm, and New Zealand populations are shown in the plots in Figures 7, 8, and 9.

The plots show that in expectation, the role of the variance from proportional allocation (blue line) in the equation of the SL allocation is null in these artificial populations. The main action is the averaging of the variance of the allocation between the substratum allocation ($E\left(\hat{\lambda}_1\right)$, or red line) and the group allocation ($E\left(\hat{\lambda}_2\right)$, or the black line). In these cases, the SL allocation may perform better if variance of the proportional allocation is removed from the formula in the SL allocation. The plots show that the SL allocation in expectation is averaging the stratum and group variances in a similar way as the naive allocation H12G12 does. One possible advantage of the SL estimator may be that it averages the variances "dynamically," moving the weight from the group variance to the stratum variance as the sample size increases. However, it does not change fast enough because at larger sample sizes, the variances of the more naive allocations are closer to the variance of the Neyman allocation in the Business and Farm populations. In the case of the New Zealand population, the SL allocation essentially averages both variances with factors 0.5 for group variances and 0.4 for stratum variances. The H12G12 uses equal factors that seem to be marginally better. In this case, the SL allocation does not yield a minimum variance and is slightly more inefficient than the H12G12 allocation.

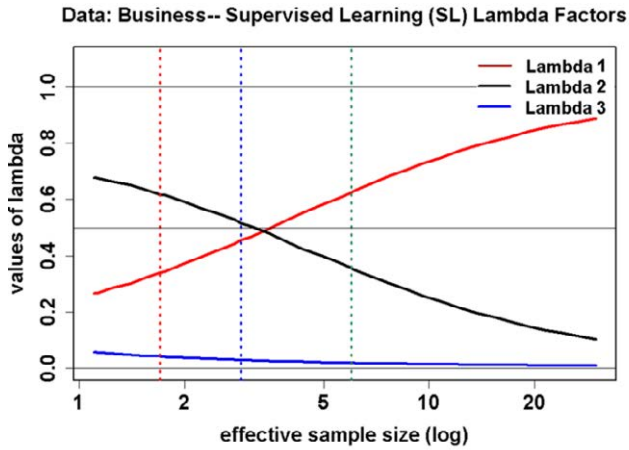

Figure 5: Expected values of the composite factors $\hat{\boldsymbol{\lambda}} = \left(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3\right)$ for the Business data by different sample sizes

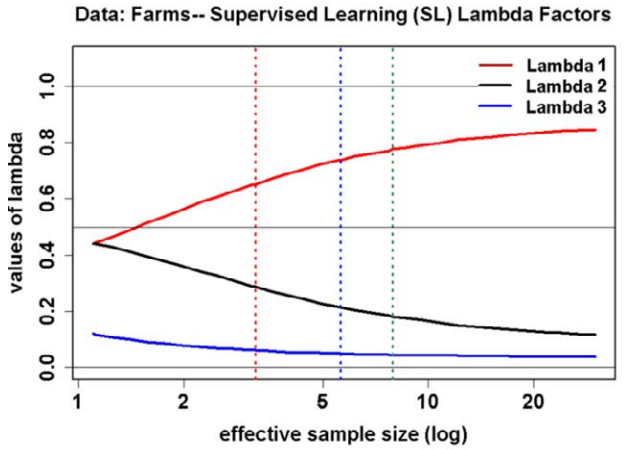

Figure 6: Expected values of the composite factors $\hat{\boldsymbol{\lambda}} = \left(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3\right)$ for the Farm data by different sample sizes

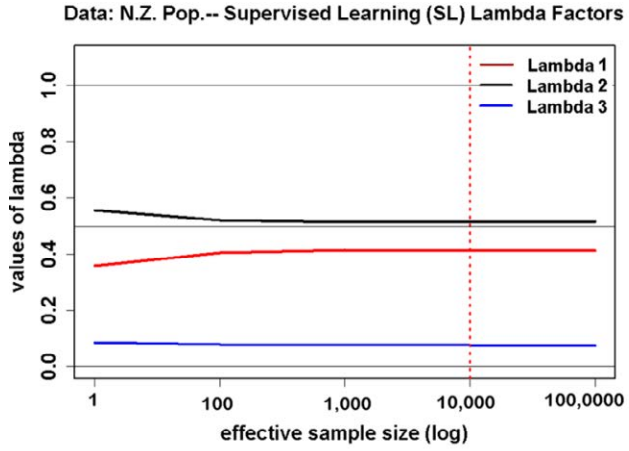

Figure 7: Expected values of the composite factors $\hat{\boldsymbol{\lambda}} = \left(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3\right)$ for the New Zealand population data by different sample sizes

5. Final Comments

The re-evaluation of the SL allocation brings to light several issues in the way new ideas are studied. If the research is based on simulations, special care is needed to determine how it is evaluated, what populations are simulated, and what alternatives are compared. Even if these issues are solved, in rare occasions generalizations based on simulations can be made. We still need to have a good understanding of the theory. This understanding is also needed to confirm our expectations from the simulations and provide ways to improve the research ideas.

We also seem to be drawn to complicated or fancier methods while ignoring simpler approaches. That is precisely the allure of statistical learning. Although they are complex, there are relatively easier to program without the need to have a basic understanding on how and why they work. We are not the first to express these concerns about the statistical and machine learning approaches (Breiman, 2001).

Acknowledgements

I would like to thank Dr. Roger Tourangeau for his comments on this work and Dr. J. Michael Brick for his support.

References

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199--231 doi:10.1214/ss/1009213726.

Buskirk, T. D., & Kolenikov, S. (2015). Finding respondents in the forest: a comparison of logistic regression and random forest models for response propensity weighting and stratification. *Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach.* doi:10.13094/SMIF-2015-00003.

Clark, R. G. (2013). Sample design using imperfect design data. *Journal of Survey Statistics and Methodology, 1*(1), 6-23.

Cochran, W. G. (1977). *Sampling techniques.* (3rd ed.). New York: John Wiley & Sons.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R.* Retrieved June 2, 2015, from http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Fourth%20Printing.pdf.

R Development Core Team. (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. doi:http://www.R-project.org.