AStA Advances in Statistical Analysis (2019) 103:563–592
https://doi.org/10.1007/s10182-018-00340-2

ORIGINAL PAPER

Neyman-type sample allocation for domains-efficient estimation in multistage sampling

M. G. M. Khan[1] · **Jacek Wesołowski**[2,3]

Received: 11 September 2016 / Accepted: 30 August 2018 / Published online: 19 September 2018

Abstract

We consider a problem of allocation of a sample in two- and three-stage sampling. We seek allocation which is both multi-domain and population efficient. Choudhry et al. (Survey Methods 38(1):23–29, 2012) recently considered such problem for one-stage stratified simple random sampling without replacement in domains. Their approach was through minimization of the sample size under constraints on relative variances in all domains and on the overall relative variance. To attain this goal, they used nonlinear programming. Alternatively, we minimize here the relative variances in all domains (controlling them through given priority weights) as well as the overall relative variance under constraints imposed on total (expected) cost. We consider several two- and three-stage sampling schemes. Our aim is to shed some light on the analytic structure of solutions rather than in deriving a purely numerical tool for sample allocation. To this end, we develop the eigenproblem methodology introduced in optimal allocation problems in Niemiro and Wesołowski (Appl Math 28:73–82, 2001) and recently updated in Wesołowski and Wieczorkowski (Commun Stat Theory Methods 46(5):2212–2231, 2017) by taking under account several new sampling schemes and, more importantly, by the (single) total expected variable cost constraint. Such approach allows for solutions which are direct generalization of the Neyman-type allocation. The structure of the solution is deciphered from the explicit allocation formulas given in terms of an eigenvector $\underline{v}^*$ of a population-based matrix $\mathbf{D}$. The solution we provide can be viewed as a multi-domain version of the Neyman-type allocation in multistage stratified SRSWOR schemes.

✉ Jacek Wesołowski
wesolo@mini.pw.edu.pl

1 University of South Pacific, Suva, Fiji

2 Politechnika Warszawska, Warsaw, Poland

3 Główny Urząd Statystyczny, Warsaw, Poland

1 Introduction

Consider a stratified SRSWOR in a population U of size N with strata $U_1, \ldots, U_I$, which form a partition of U, and let N_h denote the size of the stratum U_h. For a variable $\mathcal{Y}$ in U, we denote $y_k = \mathcal{Y}(k), k \in U$. The standard estimator of the total $\tau = \sum_{k\in U} y_k$ has the form $\hat{\tau}_{st} = \sum_{h=1}^{I} N_h \bar{y}_h$, where $\bar{y}_h = \frac{1}{n_h} \sum_{k\in\mathcal{S}_h} y_k$ with n_h denoting the size of the sample $\mathcal{S}_h$ drawn from the stratum $U_h, h = 1, \ldots, I$. The variance of this estimator is $D^2(\hat{\tau}_{st}) = \sum_{h=1}^{I} \left(\frac{1}{n_h} - \frac{1}{N_h}\right) N_h^2 S_h^2$, where $S_h^2 = \frac{1}{N_h - 1} \sum_{k\in U_h} (y_k - \bar{y}_h)^2$ is hth stratum population variance.

The basic question for such a setting is the optimal allocation, $\underline{n} = (n_1, \ldots, n_I)$, of the sample among the strata. To this end, one may assign a given (relative) variance of the estimator $\hat{\tau}_{st}$ and minimize the costs expressed, for example, by the total sample size $\sum_{h=1}^{I} n_h$. A related approach is to fix a total sample size $n = n_1 + \cdots + n_I$ and minimize the (relative) variance. Both cases are examples of the classical Neyman optimal allocation procedure which, for example, in the second case results in the allocation $n_h = n \frac{N_h S_h}{\sum_{g=1}^{I} N_g S_g}$, $h = 1, \ldots, I$. In both settings, the result is a simple consequence of minimization using the Lagrange function or can be concluded via the Schwartz inequality.

Recently, we observe a growing interest in more refined allocation methods (also in two-stage sampling) based on nonlinear programming ensuring efficient estimation procedures for the whole population, see, for example, Clark and Steel (2000), Lednicki and Wieczorkowski (2003), Clark (2009), Khan et al. (2010), Münnich et al. (2012), Gabler et al. (2012), Ballin and Barcaroli (2013), Valliant et al. (2013, 2015). Much less is known for allocation procedures which are domains efficient or both population and domains efficient—see, for example, Costa et al. (2004), Longford (2006), Choudhry et al. (2012)—referred to as CRH in the sequel, Molefe and Clark (2015) and Keto and Pahkinen (2017). All of them are again based on nonlinear programming and are designed for single-stage sampling schemes. To the best of our knowledge, the only examples of domains-efficient allocation procedures in two-stage sampling schemes are those related to the eigenproblem approach. Such approach will be explained and discussed in the sequel.

In the stratified SRSWOR, we may treat strata as domains (consequently, we will change the subscript h denoting a stratum into i denoting a domain), that is, we would like to control not only the overall (relative) variance but also (relative) variances in each of the domains. In the context of both multi- and small-area estimations, Longford (2006) suggested to minimize (under a constraint given by the total sample size) the objective function

$$\sum_{i=1}^{I} P_i \, D^2(\bar{y}_i) + G P_+ D^2(\bar{y}_{st}), \tag{1}$$

where $P_i, i = 1, \ldots, I$ are relative preassigned weights which describe "importance" of domains, $P_+ = \sum_{i=1}^{I} P_i$ and G is a weight responsible for a priority for the variance of the population mean estimator. In the context of model-assisted methodology, this

approach has been recently developed in Molefe and Clark (2015). Mathematically, the problem reduces to the Neyman allocation scheme. Similarly, when a given value is assigned for (1), the total sample size is minimized. The weights $(P_i,\ i = 1, \ldots, I)$ are designed in order to cover, at least to some extent, jointly the optimality issue for domains and for the whole population. As pointed out in Friedrich and Münnich (2018), the approach of Gabler et al. (2012) can be used also in this context (actually, they mention the case with $GP_+ = 0$). Since the objective function (1) is a weighted sum of domains and population variances, this approach does not give any convenient tool to control the quality of population and domains means estimators. Moreover, it is not clear how to assess the impact of values of weights $P_i, i = 1, \ldots, I$, and GP_+ on variances $D^2(\bar{y}_i)$, $i = 1, \ldots, I$, and $D^2(\bar{y}_{st})$. These issues are clearly visible in the numerical example given in "Appendix," where such approach is confronted with the one we propose in this paper.

Our approach can be treated as a an alternative to a direct setting of CRH. They proposed an approach, where also both multi- and small-area estimations were considered. CRH minimize the total sample size

$$g(\underline{n}) = n_1 + \cdots + n_I$$

under the constraints for relative variances of estimators of domain totals

$$T_i := N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) \frac{S_i^2}{\tau_i^2} \leq RV_{oi}, \quad i = 1, \ldots, I, \tag{2}$$

where $\tau_i = \sum_{k \in U_i} y_k$ is the total for the domain U_i, $i = 1, \ldots, I$, and the constraint on the relative variance of the estimator of population total

$$S := \frac{1}{\tau^2} \sum_{i=1}^{I} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) N_i^2 S_i^2 = \frac{1}{\tau^2} \sum_{i=1}^{I} T_i \tau_i^2 \leq RV_o. \tag{3}$$

Note that in this approach one specifies conditions for each of the domains and for the whole population separately. The problem was solved under additional box constraints of the form $0 < n_i \leq N_i, i = 1, \ldots, I$, by a nonlinear programming method involving the popular Newton–Raphson algorithm.

The NLP solution, as the one described above, is an efficient tool for applications. Such purely numerical approaches to allocation problems are popular in real surveys. A drawback of such methods is that they gave just numerical values and do not provide any information on the structure of the solution, which, for example, can be important for designing priorities for the domains.

Now we will describe an alternative approach to the problem of domains-overall-efficient allocation in the sampling scheme considered in CRH. The approach will allow to see the analytic form of the solution. The respective expression is based on a unique direction in the space $\mathbb{R}^I$, where the dimension I is equal to the number of domains. The rest of this section is just a warm-up illustration for the eigenproblem methodology we will apply in full swing in several multistage schemes in the main part of the paper.

We would like to minimize each $T_i, i = 1, \ldots, I$, as well as S under the constraint on the total size of the sample. It can be achieved in the following way. To each domain U_i, its (known) priority weight $\kappa_i > 0$ is assigned. These weights describe ratios of relative variances through

$$\frac{T_i}{T_j} = \frac{\kappa_i}{\kappa_j} \quad \forall\, i, j = 1, \ldots, I.$$

Equivalently, we can write

$$\left(\frac{1}{n_i} - \frac{1}{N_i}\right) \frac{N_i^2 S_i^2}{\tau_i^2} = \kappa_i T, \quad i = 1, \ldots, I, \tag{4}$$

where T is an unknown positive constant. Such approach allows to fully control domains variability of (relative) variances of estimators—see the numerical example in "Appendix." Moreover, under the above constraint, the unknown parameter T controls not only relative variances in domains but also the overall relative variance S of the estimator of the population mean. It follows from the fact that under (4), due to (3), the relative overall variance S can be written as

$$S = \left(\frac{1}{\tau^2} \sum_{i=1}^{I} \kappa_i \tau_i^2\right) T.$$

Therefore, when we optimize relative variances within domains, the overall relative variance gets automatically optimized as well. This general rule will hold also for the multistage schemes considered in the sequel.

Upon denoting $\gamma_i^2 = \frac{N_i^2 S_i^2}{\tau_i^2}, i = 1, \ldots, I$, Eq. (4) can be written as

$$\frac{\gamma_i^2}{\kappa_i n_i} - \frac{\gamma_i^2}{\kappa_i N_i} = T, \quad i = 1, \ldots, I. \tag{5}$$

Now we denote $v_i = \frac{n_i \sqrt{\kappa_i}}{\gamma_i}, i = 1, \ldots, I$, and, due to (5), the constraint $\sum_{i=1}^{I} n_i = n$ assumes the form

$$1 = \frac{n}{\sum_{i=1}^{I} \frac{\gamma_i}{\sqrt{\kappa_i}} v_i}. \tag{6}$$

Multiplying (5) by v_i and using (6), we get

$$n^{-1} \frac{\gamma_i}{\sqrt{\kappa_i}} \sum_{r=1}^{I} \frac{\gamma_r}{\sqrt{\kappa_r}} v_r - \frac{\gamma_i^2}{\kappa_i N_i} v_i = T v_i, \quad i = 1, \ldots, I,$$

which is equivalent to

$$\left(\frac{\underline{a}\,\underline{a}^T}{n} - \mathrm{diag}(\underline{c})\right) \underline{v} = T \underline{v}$$

with $\underline{v} = (v_1, \ldots, v_I)^T$, where

$$\underline{a} = \left(\frac{\gamma_i}{\sqrt{\kappa_i}},\ i = 1, \ldots, I\right)^T, \qquad \underline{c} = \left(\frac{\gamma_i^2}{\kappa_i N_i},\ i = 1, \ldots, I\right)^T$$

and diag($\underline{c}$) is a diagonal matrix with the vector $\underline{c}$ being its diagonal. Consequently, by the Perron–Frobenius theorem (for more details, see the Proof of Theorem 2.1), there exists a unique simple positive eigenvalue λ^* of the matrix $\mathbf{D} = \frac{\underline{a}\,\underline{a}^T}{n} - \text{diag}(\underline{c})$ and the respective eigenspace is spanned by a vector $\underline{v}^*$ with all components positive. This vector, $\underline{v}^*$, up to normalization, that is the respective direction in the space R^I is responsible for the efficient allocation. Therefore, in our problem above, $T = \lambda^*$, $\underline{v} = (v_i,\ i = 1, \ldots, I) = \underline{v}^*$ and thus

$$n_i \propto \frac{\gamma_i}{\sqrt{\kappa_i}}\, v_i^*, \quad i = 1, \ldots, I.$$

Using again the constraint on the sample size, we see that

$$n_i = n\, \frac{\frac{\gamma_i\, v_i^*}{\sqrt{\kappa_i}}}{\sum_{r=1}^{I} \frac{\gamma_r}{\sqrt{\kappa_r}}\, v_r^*}, \quad i = 1, \ldots, I.$$

Moreover, with this optimal allocation

$$T_i = \kappa_i \lambda^*, \quad i = 1, \ldots, I \qquad \text{and} \qquad S = \frac{\lambda^*}{\tau^2} \sum_{i=1}^{I} \kappa_i \tau_i^2.$$

Remark 1.1 Of course, there is an alternative numerical solution of this problem—see, for example, Lednicki and Wesołowski (1994) (referred to as LW below). From (5), one gets

$$n_i = \frac{\gamma_i^2 N_i}{\kappa_i N_i T + \gamma_i^2}, \quad i = 1, \ldots, I. \tag{7}$$

Now the sample size constraint leads to the equation

$$n = \sum_{i=1}^{I} \frac{\gamma_i^2 N_i}{\kappa_i N_i T + \gamma_i^2} \tag{8}$$

for unknown T. It is obvious that there exists a unique positive solution $T = T^*$, which has to be derived numerically. Then, the allocation is given by (7) with $T = T^*$.

As we mentioned above, there are alternatives for the eigenproblem approach to the (domains-population)-efficient allocation issue in the case of SRSWOR in domains. Except for a possibility mentioned in Remark 1.1, the same allocation can be obtained (up to box constraints) by CRH methodology if $T_i := \kappa_i T^*$ (with the value of T^* as computed in the eigenproblem procedure) and g is minimized through the NLP

procedure. Similarly, each of these three approaches (CRH, LW and eigenproblem) can be applied in the case of stratified SRSWOR in each of the domains. It suffices to start the procedure with the Neyman allocation in every domain.

However, the situation changes drastically when two-stage (or multistage) sampling is taken under account. Then, as it will be explained in the following sections, even in the simplest case of a two-stage sampling with SRSWOR at both stages (and no stratification), the formula relating the sizes of samples at the first and the second stage with variances, an analog of the one which lead to (7), does not allow to get a simple equation, as (8) in Remark 1.1 for the unknown T. Therefore, such direct numerical approach is not possible. To the best of our knowledge, no analogs of the NLP procedure from CRH are available in the literature in the multistage setting. Nevertheless, nonlinearly constrained optimization solvers, for example MINOS, MOSEK or IPOPT, available on the Web through NEOS server can be used as potential tools for NLP answers to the two-stage extension of the original CRH problem.

It appears that in such as well as in a more complicated situation, optimal allocation issue can be conveniently handled through the eigenproblem methodology, which provides insight into the structure of the optimal solutions, though in some non-typical cases it may give not the optimal but only approximately optimal results. It suffers from the same drawbacks as the original Neyman optimal allocation; i.e., the natural box constraints can be violated and the solution typically is not integer valued. The main aim of the present paper is to show how such an eigenproblem approach works in several new settings involving multistage sampling. In Sect. 2, we consider two-stage sampling with stratified SRSWOR on both stages. Special simplified cases are described in Sect. 3. Then, we deal with the situation in which at one of the stages pps sampling with replacement is used while at the other the sample is drawn according to the SRSWOR. Finally, in Sect. 5 we analyze three-stage sampling with SRSWOR at every stage. In all these cases, the allocation problem with the total cost constraint is solved via an eigenproblem for rank-one perturbations of diagonal matrices. The case of the pps sampling with replacement at the first stage and the SRSWOR at the second stage is rather special—then, the eigenproblem is for a matrix of rank one and thus an analytic form of the eigenvector responsible for allocation is available.

The eigenproblem approach to efficient allocation in domains originally was proposed in Niemiro and Wesołowski (2001) (NW in the sequel) and recently developed in Wesołowski and Wieczorkowski (2017) (WW in the sequel). The major difference between the setting of these two papers and our setting is the form of the cost constraints: Here, we consider the single total cost constraint, while two constraints, one on the sample size of the PSUs sample and one on the expected sample size of the SSUs sample, were imposed jointly in these earlier papers. There are important consequences of such a change in the cost constraints. Due to the form of the cost constraint, our solution is a direct generalization of the Neyman-type allocation. In particular, it gives the Neyman-type solution in case when there are no domains (i.e., the whole population is a single domain). At the technical level, the population matrix $\mathbf{D}$, everything is based on, is a rank-one perturbation of a diagonal matrix, while it was a rank-two perturbation of a diagonal matrix in NW and WW. There is also an important difference with NW and WW with respect to the structure of the allocation. The common feature is that there is an eigenvector $\underline{v}^*$ of the matrix $\mathbf{D}$ which plays

important role in the optimal allocation; however, in the case we consider here, it influences only the optimal allocation at the first stage, while in the cases considered in NW and WW the optimal allocation on both stages depends explicitly on respective version of $\underline{v}^*$.

2 Two-stage sampling with stratified SRSWOR at both stages

For any $i = 1, \ldots, I$, the subpopulation $\mathcal{V}_i$ of primary sampling units (PSUs) of ith domain in U is stratified: $\mathcal{V}_i = \bigcup_{h=1}^{H_i} \mathcal{V}_{i,h}$. Let $M_{i,h}$ denote number of PSUs in $\mathcal{V}_{i,h}$. Also every PSU understood as a collection of secondary sampling units (SSUs) is stratified: A PSU j from the stratum $\mathcal{V}_{i,h}$ is stratified into $\bigcup_{g=1}^{G_{j,h,i}} \mathcal{W}_{i,h,j,g}$.

A sample $\mathcal{S}$ is chosen as follows: At the first stage, a PSU's sample $\mathcal{S}_{i,h}^{(I)}$ of size $m_{i,h}$ is selected from $\mathcal{V}_{i,h}$ according to the SRSWOR, $h = 1, \ldots, H_i$, $i = 1, \ldots, I$. At the second stage for each PSU $j \in \mathcal{S}_{i,h}^{(I)}$, an SSU's sample $\mathcal{S}_{i,h,j,g}^{(II)}$ of size $n_{i,h,j,g}$ is selected from $\mathcal{W}_{i,h,j,g}$, according to the SRSWOR, $g = 1, \ldots, G_{i,h,j}$. Let $N_{i,h,j,g}$ denote the number of SSUs in $\mathcal{W}_{i,h,j,g}$. Finally, we set

$$\mathcal{S} = \bigcup_{i=1}^{I} \bigcup_{h=1}^{H_i} \bigcup_{j \in \mathcal{S}_{i,h}^{(I)}} \bigcup_{g=1}^{G_{i,h,j}} \mathcal{S}_{i,h,j,g}^{(II)}.$$

As an example, one can consider a survey of population of students in a given country with parameters to be estimated at the regions level (subpopulations) and at the whole country level as well. Then, SSUs are just students, while PSUs are schools. Schools in each region are stratified into educational districts, and pupils in each school are stratified into grades. That is, U is the population of students, $\mathcal{V}_i$ is subpopulation of schools in ith region, while $\mathcal{V}_{i,h}$ is the stratum of schools in hth district of $\mathcal{V}_i$ and $M_{i,h}$ is the number of schools in $\mathcal{V}_{i,h}$. Moreover, $\mathcal{W}_{i,h,j,g}$ denotes students of grade g of jth school from district $\mathcal{V}_{i,h}$ and $N_{i,h,j,g}$ denotes the number of students in $\mathcal{W}_{i,h,j,g}$. A sample $\mathcal{S}_{i,h}^{(I)}$ of $m_{i,h}$ schools is drawn according to SRSWOR from $\mathcal{V}_{i,h}$, and then a sample $\mathcal{S}_{i,h,j,g}^{(II)}$ of $n_{i,h,j,g}$ students is drawn by SRSWOR from $\mathcal{W}_{i,h,j,g}$ for each school j belonging to $\mathcal{S}_{i,h}^{(I)}$. Here and below in the formulas for variances, a single subscript i refers to region $\mathcal{V}_i$, a double subscript i, h refers to district $\mathcal{V}_{i,h}$, a triple subscript i, h, j refers to jth school from district $\mathcal{V}_{i,h}$ and a quadruple subscript i, h, j, g refers to grade g of jth school in district $\mathcal{V}_{i,h}$.

The variance of π-estimator of the total of $\mathcal{Y}$ over subpopulation U_i has the form, see, for example, Särndal et al. (1992, Ch. 4.3),

$$\sum_{h=1}^{H_i} \left(\frac{1}{m_{i,h}} - \frac{1}{M_{i,h}}\right) M_{i,h}^2 D_{i,h}^2 + \sum_{h=1}^{H_i} \frac{M_{i,h}}{m_{i,h}} \sum_{j \in \mathcal{V}_{i,h}} \sum_{g=1}^{G_{i,h,j}} \left(\frac{1}{n_{i,h,j,g}} - \frac{1}{N_{i,h,j,g}}\right) N_{i,h,j,g}^2 S_{i,h,j,g}^2,$$

where

$$D_{i,h}^2 = \frac{1}{M_{i,h}-1} \sum_{j \in \mathcal{V}_{i,h}} \left(t_j - \bar{t}_{i,h}\right)^2$$

with

$$t_j = \sum_{k \in PSU(j)} y_k \quad \forall \ \mathrm{PSU}(j), \quad \bar{t}_{i,h} = \frac{1}{M_{i,h}} \sum_{j \in \mathcal{V}_{i,h}} t_j;$$

and

$$S_{i,h,j,g}^2 = \frac{1}{N_{i,h,j,g}-1} \sum_{k \in \mathcal{W}_{i,h,j,g}} \left(y_k - \bar{t}_{i,h,j,g}\right)^2$$

with

$$\bar{t}_{i,h,j,g} = \frac{1}{N_{i,h,j,g}} \sum_{k \in \mathcal{W}_{i,h,j,g}} y_k.$$

The actual cost of the survey generated by the sample $\mathcal{S}$ can be modeled by the quantity

$$\sum_{i=1}^{I} \sum_{h=1}^{H_i} m_{i,h} \left(c_{I,i,h}^2 + \sum_{j \in \mathcal{S}_{i,h}^{(I)}} c_{II,i,h,j}^2 \sum_{g=1}^{G_{i,h,j}} n_{i,h,j,g} \right),$$

where $c_{I,i,h}^2 > 0$ and $c_{II,i,h,j}^2 > 0$ are costs generated by a single PSU from hth stratum of PSUs in ith domain (we assume that it is constant within the stratum) and a single SSU from jth PSU of hth stratum of PSUs in ith domain (we assume that it is constant within the PSU), respectively. Obviously, due to randomness of $\mathcal{S}_{i,h}^{(I)}$, the actual cost is a random variable. In such a situation, when one wants to impose a constraint on the total cost, the standard approach is to impose a constraint on its expected variable cost (EVC), see, for example, Ch. 12.8.1 of Särndal et al. (1992), which in the case considered here assumes the form:

$$\sum_{i=1}^{I} \sum_{h=1}^{H_i} c_{I,i,h}^2 m_{i,h} + \sum_{i=1}^{I} \sum_{h=1}^{H_i} \frac{m_{i,h}}{M_{i,h}} \sum_{j \in \mathcal{V}_{i,h}} c_{II,i,h,j}^2 \sum_{g=1}^{G_{i,h,j}} n_{i,h,j,g} = C. \tag{9}$$

We also assume that priority weights $(\kappa_i,\ i = 1, \ldots, I) \in (0,1)^I$, such that $\sum_{i=1}^{I} \kappa_i = 1$, for relative variances of estimators of means in subpopulations are preassigned, that is

$$T_i = \frac{1}{\tau_i^2} \sum_{h=1}^{H_i} \frac{1}{m_{i,h}} \left(\gamma_{i,h}^2 + M_{i,h} \sum_{j \in \mathcal{V}_{i,h}} \sum_{g=1}^{G_{i,h,j}} \frac{\beta_{i,h,j,g}^2}{n_{i,h,j,g}} \right) - \frac{\sum_{h=1}^{H_i} M_{i,h} D_{i,h}^2}{\tau_i^2} = \kappa_i T, \qquad i = 1, \ldots, I, \tag{10}$$

where

$$\gamma_{i,h}^2 = M_{i,h} \left(M_{i,h} D_{i,h}^2 - \sum_{j \in \mathcal{V}_{i,h}} \sum_{g=1}^{G_{i,h,j}} N_{i,h,j,g} S_{i,h,j,g}^2 \right)$$

and

$$\beta_{i,h,j,g} = N_{i,h,j,g} S_{i,h,j,g}$$

for $\tau_i = \sum_{h=1}^{H_i} \sum_{j \in \mathcal{V}_{i,h}} t_j$, $i = 1, \ldots, I$. We wrote above $\gamma_{i,h}^2$ since we will be assuming that it is nonnegative.

We want to find the allocation that is a set of two tables: a two-way table $\underline{m} = (m_{i,h})$ and a four-way table $\underline{n} = (n_{i,h,j,g})$, which give minimal domain-wise relative variances T_i, $i = 1, \ldots, I$ and minimal relative overall variance S, under the constraints (10) imposed through priority weights and the EVC constraint (9).

The result below says that it can be achieved by searching for positive eigenvalue of a certain matrix based on population quantities and costs coefficients. The allocation is obtained from the respective eigenvector. The approach parallels earlier developments in this setting where, instead of using a single total average cost constraint, the first-stage and second-stage costs were treated separately. In particular, NW in 2001 considered a two-stage scheme with separate constraints for the size of PSUs and SSUs sample and with stratified sampling either at the first or at the second stage. As it has been already mentioned, a similar problem has been recently investigated in WW for two-stage stratified SRSWOR on both stages as well as a scheme with stratified Hartley–Rao scheme at the first stage and stratified SRSWOR at the second stage (also some variations of theses two basic setups were considered there). In that paper, again two constraints were jointly imposed: one for the cost incurred by the PSUs sample size, $\sum_{i=1}^{I} \sum_{h=1}^{H_i} m_{i,h} = m$, and one for the cost generated by the expected SSUs sample size, $\sum_{i=1}^{I} \sum_{h=1}^{H_i} \frac{m_{i,h}}{M_{i,h}} \sum_{j \in \mathcal{V}_{i,h}} \sum_{g=1}^{G_{i,h,j}} n_{i,h,j,g} = n$ (these formulas refer obviously to the stratified SRSWOR on both stages). In meantime, the eigenproblem approach has been further developed in a series of papers: Kozak (2004) (multivariate version of NW was considered with an application to agricultural surveys), Kozak and Zieliński (2005) (the original eigenproblem approach from NW, where it was assumed that relative variances are the same for all domains, was adapted to include priority weights for domains; also an application related to the real forestry survey was given). Only single-stage schemes were considered in both these papers. In the context, we consider here probably the most interesting is the paper (Kozak et al. 2008). These authors were concerned with a two-stage sampling with stratification at the first stage

together with a single cost constraint similar to (9) and domains-related constraints like (10). However, their approach was restricted to the case when SSU's sample sizes are the same for all PSU's in a given stratum of a given domain. Also they did not consider stratification at the second stage. The latter restriction does not seem to be as serious as the former.

In our main result below, we use the notation introduced earlier in this section.

Theorem 2.1 *Assume that* $M_{i,h}D_{i,h}^2 > \sum_{j\in\mathcal{V}_{i,h}} \sum_{g=1}^{G_{i,h,j}} N_{i,h,j,g}\, S_{i,h,j,g}^2$ *for all* $h = 1,\ldots,H_i$, $i = 1,\ldots,I$. *Let* $\mathbf{D} = \frac{\underline{a}\,\underline{a}^T}{C} - \mathrm{diag}(\underline{c})$ *where* $\underline{a} = (a_i,\ i = 1,\ldots,I)^T$, $\underline{c} = (c_i,\ i = 1,\ldots,I)$,

$$a_i = \frac{v_i}{\rho_i}, \quad \textit{with} \quad v_i = \sum_{h=1}^{H_i}\left(c_{I,i,h}\gamma_{i,h} + \sum_{j\in\mathcal{V}_{i,h}} c_{II,i,h,j} \sum_{g=1}^{G_{i,h,j}} \beta_{i,h,j,g}\right), \quad \rho_i = \tau_i\sqrt{\kappa_i}$$

and

$$c_i = \frac{1}{\rho_i^2}\sum_{h=1}^{H_i} M_{i,h}\, D_{i,h}^2, \quad i = 1,\ldots,I.$$

Assume that $\mathbf{D}$ *has a positive eigenvalue* λ^* *with a respective eigenvector* $\underline{v}^* = (v_1^*,\ldots,v_I^*)^T$.

Then, λ^* *is simple and unique and* $\underline{v}^*$ *has all coordinates of the same sign.*

The allocation which minimizes all relative variances in domains T_i, $i = 1,\ldots,I$, *(as well as the relative variance* S *in the whole population) under domain relative variance constraints* (10) *and overall EVC constraint* (9) *is given by*

$$m_{i,h} = C\frac{v_i^*\gamma_{i,h}}{\rho_i c_{I,i,h}\sum_{r=1}^{I} v_r^* v_r/\rho_r} \tag{11}$$

and

$$n_{i,h,j,g} = \frac{c_{I,i,h}M_{i,h}\beta_{i,h,j,g}}{c_{II,i,h,j}\gamma_{i,h}}. \tag{12}$$

Moreover, the minimal relative variances in the domains are $T_i = \kappa_i\, T$, $i = 1,\ldots,I$ *and the overall relative variance is* $S = \frac{T}{\tau^2}\sum_{i=1}^{I}\kappa_i\tau_i^2$ *with*

$$T = \lambda^* = \frac{1}{\sum_{i=1}^{I}\rho_i^2}\left[\frac{1}{C}\left(\sum_{i=1}^{I}\frac{\rho_i}{v_i^*}v_i\right)\left(\sum_{i=1}^{I}\frac{v_i^*}{\rho_i}v_i\right) - \sum_{i=1}^{I}\sum_{h=1}^{H_i} M_{i,h}\, D_{i,h}^2\right]. \tag{13}$$

Remark 2.1 Note that when the condition

$$\sum_{i=1}^{d}\frac{a_i^2}{c_i} > C \tag{14}$$

is satisfied for a matrix of the form $\mathbf{D} = \frac{\underline{a}\,\underline{a}^T}{C} + \mathrm{diag}(\underline{c})$ with $C > 0$ and $\underline{a}, \underline{c} \in (0, \infty)^d$, then $\mathbf{D}$ has a positive eigenvalue (see Prop. 2.1 in WW). Note that in the framework of Theorem 2.1, condition (14) assumes the form

$$\sum_{i=1}^{d} \frac{v_i^2}{\sum_{h=1}^{H_i} M_{i,h}\, D_{i,h}^2} > C.$$

The above assumption as well as the assumption that $\gamma_{i,h}^2 > 0$ is related to convexity of the function being minimized, and as such they are necessary also for the convex NLP methods to provide the unique solution (see also Remark 2.3).

Remark 2.2 Note that the problem we solved in Theorem 2.1 can be formulated equivalently as: Minimize the overall relative variance S under constraints (10) on relative variances T_i in domains ($i = 1, \ldots, I$) and the expected overall cost constraint (9). The reason for validity of such a rephrasing of the original problem is that $S = T\,\frac{1}{\tau^2}\sum_{i=1}^{I} \kappa_i \tau_i^2$ which is a consequence of $T_i = \kappa_i T$, $i = 1, \ldots, I$.

Remark 2.3 The optimal allocation problem in two-stage sampling when no domains efficiency is taken under account has the well-known Neyman-type solution. For example, in case of no stratification on both stages, such solution under EVC constraint is given in Ch. 12.8.1 of Särndal et al. (1992). Our formulas (11) and (12) reduce to (12.8.13) and (12.8.14) of Särndal et al. (1992) in the case when $I = 1$, $H_1 = 1$ and $G_{1,1,j} = 1$, that is when the whole population is a single domain, neither PSUs nor SSUs within PSU are stratified. The optimal allocation in the case of single domain with stratified SRSWOR for PSUs and SRSWOR for SSUs in every PSU from the first-stage sample is considered in Saini and Kumar (2015). The authors provide the NLP solution and then conclude that the same result can be obtained via Neyman-type approach. Actually, they consider a p-variate case. However, their optimal allocation formulas (16) and (17) for $p = 1$ are again special cases of (11) and (12). Note that the assumption $A_h > 0$ needed also for the numerical solution in that paper is (again for $p = 1$) in full agreement with $\gamma_{ih}^2 > 0$, which we assume in Theorem 2.1.

In the case of the population U consisting just of a single domain, i.e., when $I = 1$, the eigenvector cancels out from (11) and formulas (11) and (12) for optimal allocation are immediately reduced to (with the index $i = 1$ suppressed)

$$m_h = C\,\frac{\gamma_h}{c_{I,h}\sum_{\ell=1}^{H}\left(c_{I,\ell}\gamma_\ell + \sum_{j\in\mathcal{V}_\ell} c_{II,\ell,j}\sum_{g=1}^{G_{\ell,j}} \beta_{\ell,j,g}\right)} \quad \text{and} \quad n_{h,j,g} = \frac{c_{I,h}\, M_h \beta_{h,j,g}}{c_{II,h,j}\gamma_h}.$$

Moreover, the optimal relative variance (13) assumes the form

$$D_{opt}^2 = \frac{1}{C}\left[\sum_{h=1}^{H}\left(c_{I,h}\gamma_h + \sum_{j\in\mathcal{V}_h} c_{II,h,j}\sum_{g=1}^{G_{h,j}} \beta_{h,j,g}\right)\right]^2 - \sum_{h=1}^{H} M_h\, D_h^2.$$

Note that these formulas are exact versions of the Neyman optimal allocation and the Neyman optimal variance for two-stage sampling with stratified SRSWOR at both stages.

Remark 2.4 The allocation results given in Theorem 2.1 should be compared to the domain-efficient allocation in the same stratified SRSWOR on both stages but with separate constraints for the size of the first-stage sample and for the expected size of the second-stage sample as given in Theorem 3.3 of WW. The basic difference is that in the latter paper both $m_{i,h}$ and $n_{i,h,j,g}$ depend on the eigenvector $\underline{v}^*$, while in the above result the eigenvector appears only in formula (11) for $m_{i,h}$ and formula (12) is free from $\underline{v}^*$. This is the major, and by no means obvious, structural consequence of the fact that the constraint we consider here is imposed on the expected costs of the first and the second stage jointly.

Proof of Theorem 2.1 Note that since $\kappa_i, i = 1, \ldots, I$, are fixed and known, minimizing relative variances $T_i = T_i(\underline{m}, \underline{n}), i = 1, \ldots, I$, is equivalent to minimize T under constraints (9) and (10). Therefore, the Lagrange function has the form

$$F(T, \underline{m}, \underline{n}) = T + \sum_{i=1}^{I} \lambda_i \left(\frac{T_i(\underline{m}, \underline{n})}{\kappa_i} - T \right) + \mu \left(\sum_{i=1}^{I} c_{I,i,h}^2 \sum_{h=1}^{H_i} m_{i,h} + \sum_{j=1}^{J} \sum_{h=1}^{H_j} \frac{m_{j,h}}{M_{j,h}} \sum_{i \in \mathcal{W}_{j,h}} c_{II,i,h,j}^2 \sum_{g=1}^{G_{j,h,i}} n_{j,h,i,g} \right).$$

Note that

$$\frac{\partial F}{\partial n_{i,h,j,g}} = -\frac{\lambda_i M_{i,h} \beta_{i,h,j,g}^2}{\rho_i^2 m_{i,h} n_{i,h,j,g}^2} + \mu \frac{m_{i,h}}{M_{i,h}} c_{II,i,h,j}^2 = 0.$$

Consequently, $\lambda_i > 0$ and

$$m_{i,h} n_{i,h,j,g} = \frac{\sqrt{\lambda_i} M_{i,h} \beta_{i,h,j,g}}{c_{II,i,h,j} \rho_i \sqrt{\mu}}. \tag{15}$$

Moreover,

$$\frac{\partial F}{\partial m_{i,h}} = -\frac{\lambda_i}{m_{i,h}^2 \rho_i^2} \left(\gamma_{i,h}^2 + M_{i,h} \sum_{j \in \mathcal{V}_{i,h}} \sum_{g=1}^{G_{i,h,j}} \frac{\beta_{i,h,j,g}}{n_{i,h,j,g}} \right) + \mu \left(c_{I,i,h}^2 + \frac{1}{M_{i,h}} \sum_{j \in \mathcal{V}_{i,h}} c_{II,i,h,j}^2 \sum_{g=1}^{G_{i,h,j}} n_{i,h,j,g} \right) = 0.$$

The above after multiplication by $m_{i,h}$ can alternatively be written as

$$-\frac{\lambda_i \gamma_{i,h}^2}{m_{i,h}\rho_i^2} - \frac{\lambda_i M_{i,h}}{\rho_i^2} \sum_{j\in\mathcal{V}_{i,h}} \sum_{g=1}^{G_{i,h,j}} \frac{\beta_{i,h,j,g}^2}{m_{i,h}n_{i,h,j,g}}$$
$$+\mu\, c_{I,i,h}^2 m_{i,h} + \mu \frac{1}{M_{i,h}} \sum_{j\in\mathcal{V}_{i,h}} c_{II,i,h,j}^2 \sum_{g=1}^{G_{i,h,j}} m_{i,h}n_{i,h,j,g} = 0.$$

Note that due to (15), the second and fourth terms above cancel out and thus

$$m_{i,h} = \frac{\sqrt{\lambda_i}\gamma_{i,h}}{c_{I,i,h}\,\rho_i\sqrt{\mu}}. \tag{16}$$

Now (12) follows by combining (16) with (15).

To find $m_{i,h}$, we plug (15) and (16) into the cost constraint (9) and obtain

$$\sqrt{\mu} = \frac{1}{C} \sum_{i=1}^{I} \frac{\sqrt{\lambda_i}}{\rho_i} v_i \tag{17}$$

Now let us insert (15) and (16) in the constraint (10). It leads to the equation

$$\sum_{h=1}^{H_i} \frac{\gamma_{i,h}^2 c_{I,i,h}\sqrt{\mu}}{\sqrt{\lambda_i}\gamma_{i,h}} + \sum_{h=1}^{H_i} M_{i,h} \sum_{j\in\mathcal{V}_{i,h}} \sum_{g=1}^{G_{i,h,j}} \frac{\beta_{i,h,j,g}^2 c_{II,i,h,j}\sqrt{\mu}}{\sqrt{\lambda_i} M_{i,h}\beta_{i,h,j,g}} - \frac{1}{\rho_i} \sum_{h=1}^{H_i} M_{i,h}\, D_{i,h}^2 = \rho_i\, T.$$

Multiply its both sides by $v_i := \sqrt{\lambda_i}$, divide by ρ_i and rewrite as

$$\sqrt{\mu}\frac{v_i}{\rho_i} - c_i v_i = T v_i.$$

Expanding now $\sqrt{\mu}$ according to (17), we obtain

$$C^{-1}\frac{v_i}{\rho_i} \sum_{r=1}^{I} \frac{v_r v_r}{\rho_r} - c_i v_i = T v_i,$$

which is valid for any $i = 1, \ldots, I$. Note that in terms of the vector $\underline{a}$ defined in the formulation of the theorem, the above can be rewritten as

$$\frac{a_i \underline{a}^T}{C}\underline{v} - c_i v_i = T v_i, \quad i = 1, \ldots, I,$$

or equivalently $\left(\frac{\underline{a}\,\underline{a}^T}{C} - \mathrm{diag}(\underline{c})\right)\underline{v} = T\underline{v}$.

The final part of the proof follows closely the argument given in WW and is recalled here just for the readers' convenience.

Consider now the matrix $\mathbf{D}$ and let λ^* be its positive eigenvalue. To show that it is simple, unique and the eigenvector $\underline{v}^*$ attached to this eigenvalue has all coordinates of the same sign, we use the celebrated Perron–Frobenius theorem: If $\mathbf{A}$ is a matrix with all strictly positive entries, then there exists a unique positive eigenvalue ν of $\mathbf{A}$; it is simple and such that $\nu > |\lambda|$ for any other eigenvalue λ of $\mathbf{A}$. The respective eigenvector (attached to ν) has all entries strictly positive (up to scalar multiplication)—see, for example, Kato (1981, Th. 7.3 in Ch. 1).

Fix a number $\rho > \max_{1\le i\le I} c_i > 0$. The matrix $\mathbf{D} + \rho\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, has all entries strictly positive. For any eigenvalue δ_j of $\mathbf{D}$ and respective eigenvector $\underline{w}_j$

$$(\mathbf{D} + \rho\mathbf{I})\underline{w}_j = (\delta_j + \rho)\underline{w}_j, \quad j = 1, \ldots, I.$$

That is $\delta_j + \rho$ and $\underline{w}_j$, $j = 1, \ldots, d$, are respective eigenvalues and eigenvectors of the matrix $\mathbf{D} + \rho\mathbf{I}$. By the Perron–Frobenius theorem, there exists j_0 such that $\delta_{j_0} + \rho j \ge |\delta_j + \rho|$ for any $j = 1, \ldots, I$ and respective eigenvector $\underline{w}_{j_0}$ has all entries of the same sign. Consequently, $\delta_{j_0} + \rho j \ge \delta_j + \rho$, and thus $\delta_{j_0} \ge \delta_j$ for any $j = 1, \ldots, I$. Therefore, by assumption that λ^* is the unique positive eigenvalue of $\mathbf{D}$, it follows that $T = \lambda^* = \delta_{j_0}$ and the respective eigenvector $\underline{v}^* = \underline{w}_{j_0}$ has all entries of the same sign.

Now formulas (11) and (12) follow directly from (16) and (15). □

Remark 2.5 Of course, as always when such allocation problems are solved without the natural box constraints: $m_{i,h} \le M_{i,h}$ and $n_{i,h,j,g} \le N_{i,h,j,g}$ (and this is the case of eigenproblem approach), the solution may violate some of them. Then, it is standard to set $m_{i,h} = M_{i,h}$ and $n_{i,h,j,g} = N_{i,h,j,g}$ in all instances of violation of the respective box constraint and then repeat the minimization procedure for reduced population and reduced cost constraint. It may produce solutions which are not optimal (though, typically, close to them). On the other hand, it is known, for example, in the case of the problem of optimal allocation in stratified SRSWOR that it is possible to reduce the population since the optimal solution requires to take $n_h = N_h$ in some strata. Then, minimization can be performed on such reduced population—see, for example, Lemma 1 in Stenger and Gabler (2005). This approach has been developed by introducing box constraints to the numerical procedure of optimal allocation in Gabler et al. (2012); computational aspects of such procedures are analyzed in Münnich et al. (2012) (with further references given in that paper).

We do not consider here also exact optimality with respect to integer solutions. In this context, it is worth to mention again stratified SRSWOR for which an integer-valued optimal allocation has been recently given in Wright (2017) and, another one, even earlier by Friedrich et al. (2015). Here, we are fully satisfied with, for example, random rounding of non-integer allocation, which typically gives solutions close to optimal.

3 Special cases

3.1 Stratification only at the first stage

This is probably the most popular of the two-stage schemes used in practice. In this case, we have $G_{i,h,j} = 1$ for any (i, h, j) and thus it allows for a considerable simplification of the notation used in Sect. 2. The constraints imposed by priority weights for relative variances in domains assume the form

$$T_i = \frac{1}{\tau_i^2} \sum_{h=1}^{H_i} \frac{1}{m_{i,h}} \left(\gamma_{i,h}^2 + M_{i,h} \sum_{j \in \mathcal{V}_{i,h}} \frac{\beta_{i,h,j}^2}{n_{i,h,j}} \right) - \frac{1}{\tau_i^2} \sum_{h=1}^{H_i} M_{i,h} D_{i,h}^2 = \kappa_i T, \qquad i = 1, \ldots, I,$$

where

$$\gamma_{i,h}^2 = M_{i,h} \left(M_{i,h} D_{i,h}^2 - \sum_{j \in \mathcal{V}_{i,h}} N_{i,h,j} S_{i,h,j}^2 \right), \qquad \beta_{i,h,j} = N_{i,h,j} S_{i,h,j}$$

and in jth PSU from $\mathcal{V}_{i,h}$: The number of SSUs is $N_{i,h,j}$, the population variance among SSUs is $S_{i,h,j}^2$ and the sample size is $n_{i,h,j}$. Here, $D_{i,h}^2$ and $M_{i,h}$ have the same definition as in Sect. 2.

The cost constraint (9) changes to

$$\sum_{i=1}^{I} \sum_{h=1}^{H_i} c_{I,i,h}^2 m_{i,h} + \sum_{i=1}^{I} \sum_{h=1}^{H_i} \frac{m_{i,h}}{M_{i,h}} \sum_{j \in \mathcal{V}_{i,h}} c_{II,i,h,j}^2 n_{i,h,j} = C$$

From Theorem 2.1 [if its assumptions, in particular the respective version of (14), are satisfied], we conclude that the optimal allocation at the first stage is:

$$m_{i,h} = C \frac{v_i^* \gamma_{i,h}}{\rho_i c_{I,i,h}^2 \sum_{r=1}^{I} v_r^* v_r / \rho_r} \quad \text{where} \quad v_r = \sum_{s=1}^{H_r} \left(c_{I,r,s} \gamma_{r,s} + \sum_{t \in \mathcal{V}_{r,s}} c_{II,r,s,t} \beta_{r,s,t} \right), \tag{18}$$

$\underline{v}^*$ is the eigenvector (with positive components) of the matrix $\mathbf{D} = \frac{\underline{a}\underline{a}^T}{C} - \mathrm{diag}(\underline{c})$ with

$$a_i = \frac{v_i}{\rho_i}, \quad c_i = \frac{1}{\rho_i^2} \sum_{h=1}^{H_i} M_{i,h} D_{i,h}^2, \tag{19}$$

and the optimal allocation at the second stage is

$$n_{i,h,j} = \frac{c_{I,i,h} M_{i,h} \beta_{i,h,j}}{c_{II,i,h,j} \gamma_{i,h}}. \tag{20}$$

Due to its important role in practice, we chose this setting for presenting the core part of an R-code which produces the domain-efficient allocation. Assume that vectors a:= $\underline{a}$ and c:= $\underline{c}$ have already been computed according to (19) and that the vector of priority weights (κ_i, $i = 1, \ldots, I$) is denoted by kap. Then, to find the respective eigenvector, one may use the following code in R (function eigen being its essence)

```
if (nrow(c)>1) D.matrix<-a%*%t(a)  - diag(c$c)
else D.matrix<-a%*%t(a)  - c$c

eig<-eigen(D.matrix)     # must be unique positive
   eigenvalue
lambda<-eig$values
lambda<-lambda[lambda>0]

if (length(lambda)>1)
   stop("Positive eigenvalue is not unique - solution
     does not exist !")

opt<-eig$values[1]  # maximum eigenvalue; since sorted
  in decreasing order
cat("for domain = ",i," CV optimal (in %) = ",100*sqrt
  (kap[i]*opt[i]),"\n")
v<-(-meig$vectors[,1])  # corresponding eigenvector
```

After computing the eigenvector $\underline{v} := \mathrm{v}$ as given in the last line of the R-code above, one can calculate the optimal sample sizes $m_{i,h}$ and $n_{i,h,j}$ according to (18) and (20), respectively.

The R-code given above was adapted from the full R-code as given in https://github.com/rwieczor/eigenproblem_sample_allocation, which was created (in connection with WW) for optimal fixed precision allocation in subpopulations in two-stage sampling with the stratified Hartley–Rao πps scheme at the first stage and SRSWOR at the second stage and with constraints imposed separately on the size of the sample at the first and on the expected size of the sample at the second stage.

3.2 Stratification only at the second stage

Here, we have $H_i = 1$ for any i. It allows to also simplify the notation of Sect. 2. The constraints imposed by priority weights for relative variances in domains assume the form

$$T_i = \frac{1}{m_i \tau_i^2} \left(\gamma_i^2 + M_i \sum_{j \in \mathcal{V}_i} \sum_{g=1}^{G_{i,j}} \frac{\beta_{i,j,g}^2}{n_{i,j,g}} \right) - \frac{1}{\tau_i^2} M_i \, D_i^2 = \kappa_i \, T, \qquad i = 1, \ldots, I,$$

where

$$\beta_{i,j,g} = N_{i,j,g} S_{i,j,g}, \qquad D_i^2 = \frac{1}{M_i - 1} \sum_{j \in \mathcal{V}_i} \left(t_j - \bar{t}_i\right)^2.$$

and $N_{i,j,g}$ is number of SSUs, $S^2_{i,j,g}$ is the population variance among SSUs and $n_{i,j,g}$ is the sample size, in gth SSU stratum of jth PSU from $\mathcal{V}_i$. Moreover, $G_{i,j}$ is the number of SSUs strata in jth PSU from $\mathcal{V}_i$, $M_i = \#(\mathcal{V}_i)$ and

$$\gamma_i^2 = M_i \left(M_i D_i^2 - \sum_{j \in \mathcal{V}_i} \sum_{g=1}^{G_{i,j}} N_{i,j,g} S^2_{i,j,g} \right).$$

Here, the version of the cost constraint (9) is

$$\sum_{i=1}^{I} c^2_{I,i} m_i + \sum_{i=1}^{I} \frac{m_i}{M_i} \sum_{j \in \mathcal{V}_i} c^2_{II,i,j} \sum_{g=1}^{G_{i,j}} n_{i,j,g} = C.$$

From Theorem 2.1 (if its assumptions are satisfied), it follows that the optimal allocation at the first stage is

$$m_i = C \frac{v_i^* \gamma_i}{\rho_i c_{I,i} \sum_{r=1}^{I} v_r^* \nu_r / \rho_r}, \quad \text{where} \quad \nu_r = c_{I,r} \gamma_r + \sum_{t \in \mathcal{V}_r} c_{II,r,t} \sum_{u=1}^{G_{r,t}} \beta_{r,t,u},$$

$\underline{v}^*$ is the eigenvector (having all components positive) of the matrix $\mathbf{D} = \frac{\underline{a}\,\underline{a}^T}{C} - \mathrm{diag}(\underline{c})$ with

$$a_i = \frac{\nu_i}{\rho_i}, \quad c_i = \frac{M_i D_i^2}{\rho_i^2},$$

and the optimal allocation at the second stage is

$$n_{i,j,g} = \frac{c_{I,i} M_i \beta_{i,j,g}}{c_{II,i,j} \gamma_i}.$$

3.3 No stratification at stage one and two

That is, we assume $H_i = 1$ and $G_{i,h,j} = 1$ for any (i, h, j). In this case, the formulas are further simplified. The constraints imposed by priority weights for relative variances in domains assume the form

$$T_i = \frac{1}{\tau_i^2 m_i} \left(\gamma_i^2 + M_i \sum_{j \in \mathcal{V}_i} \frac{\beta^2_{i,j}}{n_{i,j}} \right) - \frac{1}{\tau_i^2} M_i D_i^2 = \kappa_i T, \qquad i = 1, \ldots, I,$$

where

$$\beta_{i,j} = N_{i,j} S_{i,j}$$

and $N_{i,j}$ is number of SSUs, $S^2_{i,j}$ is the population variance among SSUs and $n_{i,j}$ is the sample size, in jth PSU from $\mathcal{V}_i$. Moreover,

$$\gamma_i^2 = M_i \left(M_i D_i^2 - \sum_{j \in \mathcal{V}_i} N_{i,j,} S^2_{i,j} \right)$$

with M_i and D_i defined as in Sect. 3.2. The cost constraint (9) assumes a simple form

$$\sum_{i=1}^{I} c^2_{I,i} m_i + \sum_{i=1}^{I} \tfrac{m_i}{M_i} \sum_{j \in \mathcal{V}_i} c^2_{II,i,j} n_{i,j} = C.$$

From Theorem 2.1 (if its assumptions are satisfied), we conclude that the optimal allocation at the first stage is

$$m_i = C \frac{v_i^* \gamma_i}{\rho_i c_{I,i} \sum_{r=1}^{I} v_r^* v_r / \rho_r}, \quad \text{where} \quad v_r = c_{I,r} \gamma_r + \sum_{t \in \mathcal{V}_r} c_{II,r,t} \beta_{r,t},$$

$\underline{v}^*$ is the eigenvector (having all components positive) of the matrix $\mathbf{D} = \frac{\underline{a}\,\underline{a}^T}{C} - \mathrm{diag}(\underline{c})$ with

$$a_i = \tfrac{v_i}{\sqrt{\kappa_i}},$$

$\underline{c}$ defined as in Sect. 3.2 and the optimal allocation at the second stage is

$$n_{i,j} = \tfrac{c_{I,i} M_i \beta_{i,j}}{c_{II,i,j} \gamma_i}.$$

4 Two-stage sampling with pps sampling

4.1 pps Sampling at the first stage and SRSWOR at the second stage

We draw the PSUs ordered sample $\mathcal{S}^{(I)} = (K_1, \ldots, K_m)$ using pps sampling, meaning that PSUs are drawn m times with replacement (that is, independently), jth with probability p_j which is proportional to its size, $j \in \mathcal{V}$ (population of PSUs). Then, if jth PSU belongs to $\mathcal{S}^{(I)}$, we draw (by SRSWOR) from it a sample (of size n_j) of SSUs, obtaining in this way the sample $\mathcal{S}_j^{(II)}$, $j \in \mathcal{S}^{(I)}$. Such sampling scheme is considered in Ch. 4.5 of Särndal et al. (1992) (in particular, in Result 4.5.1 the unbiased estimator and its variance are given). A population-efficient allocation procedure for this setup has been given recently in Valliant et al. (2015) as one of the options in the

PracTools R package. Importance of this scheme is due to the fact that when the sample of PSUs is sufficiently small, sampling with or without replacement gives the same results. Consequently, very often in practice, the first-stage variance in πps without replacement sampling is approximated by its pps version. It appears that in such case the eigenproblem methodology we develop here allows for a closed analytic formula for the eigenvector responsible for the domains-efficient allocation. It follows from the fact that the respective population matrix is of rank one. The details are given below.

The unbiased estimator of the population total $\tau = \sum_{k\in U} y_k$ is

$$\hat{\tau} = \frac{1}{m}\sum_{r=1}^{m} \frac{\hat{t}_{K_r}}{p_{K_r}},$$

where $\hat{t}_j = \frac{1}{n_j}\sum_{k\in\mathcal{S}_j^{(II)}} y_k$ for any PSU j. Its variance has the form

$$D^2(\hat{\tau}) = \frac{1}{m}\sum_{j\in\mathcal{V}} p_j\left(\frac{t_j}{p_j} - \tau\right)^2 + \frac{1}{m}\sum_{j\in\mathcal{V}} \frac{D_j^2}{p_j},$$

where for any $j \in \mathcal{V}$ we denote

$$t_j = \sum_{k\in PSU(j)} y_k, \qquad D_j^2 = N_j^2\left(\frac{1}{n_j} - \frac{1}{N_j}\right)S_j^2, \qquad S_j^2 = \frac{1}{N_j - 1}\sum_{k\in PSU(j)}\left(y_k - \frac{t_j}{N_j}\right)^2.$$

To obtain the optimal allocation of the sample at the first and at the second stage in the domains with given priority weights κ_i, $i = 1, \ldots, I$, we need to minimize

$$T_i = \frac{1}{\tau_i^2}\left(\frac{1}{m_i}\sum_{j\in\mathcal{V}_i} p_{i,j}\left(\frac{t_{i,j}}{p_{i,j}} - \tau_i\right)^2 + \frac{1}{m_i}\sum_{j\in\mathcal{V}_i}\frac{D_{i,j}^2}{p_{i,j}}\right), \quad i = 1, \ldots, I,$$

under the constraints given by the priority weights $T_i = \kappa_i T$ and the EVC constraint

$$\sum_{i=1}^{I}\left(m_i c_{I,i}^2 + m_i \sum_{j\in\mathcal{V}_i} c_{II,i,j}^2 p_{i,j} n_{i,j}\right) = C, \tag{21}$$

where C is the total expected cost of the survey, $c_{I,i}$ is the cost incurred by a PSU from V_i (assumed to be constant within the domain) and $c_{II,i,j}$ is the cost incurred by a SSU belonging to the jth PSU from the ith domain.

This setting is somewhat different, actually, simpler than considered earlier. It is due to the fact that in the expression for T_i all summands are multiplied by $1/m_i$.

Theorem 4.1 *Assume that for any* $i = 1, \ldots, I$

$$\sum_{j\in\mathcal{V}_i}\left[p_{i,j}\left(\frac{t_{i,j}}{p_{i,j}} - \tau_i\right)^2 - \frac{N_{i,j}S_{i,j}^2}{p_{i,j}}\right] > 0.$$

Then, the allocation minimizing $T_i = \kappa_i T$, $i = 1, \ldots, I$, (as well as the relative variance S in the whole population) under the cost constraint (21) has the form

$$m_i = C \frac{A_i \left(c_{I,i} A_i + \sum_{j \in V_i} c_{II,i,j} B_{i,j} \sqrt{p_{i,j}} \right)}{c_{I,i} \sum_{r=1}^{I} c_{I,r} A_r + \sum_{s \in V_r} c_{II,r,s} B_{r,s} \sqrt{p_{r,s}}}, \quad i = 1, \ldots, I,$$

and

$$n_{i,j} = \frac{c_{I,i} B_{i,j}}{A_i, c_{II,i,j} \sqrt{p_{i,j}}}, \quad j \in \mathcal{V}_i, \ i = 1, \ldots, I,$$

where

$$A_i^2 = \frac{1}{\tau_i^2 \kappa_i} \sum_{j \in \mathcal{V}_i} \left[p_{i,j} \left(\frac{t_{i,j}}{p_{i,j}} - t_i \right)^2 - \frac{N_{i,j} S_{i,j}^2}{p_{i,j}} \right], \quad \textit{and} \quad B_{i,j}^2 = \frac{N_{i,j}^2 S_{i,j}^2}{p_{i,j}}.$$

Proof Similarly, as in the proof of Theorem 2.1, we consider the Lagrange function

$$F(T, (m_i), (n_{i,j}); (\lambda_i), \mu) = T + \sum_{i=1}^{I} \lambda_i \left(\frac{A_i^2}{m_i} + \sum_{j \in \mathcal{V}_i} \frac{B_{i,j}^2}{m_i n_{i,j}} - T \right) + \mu \sum_{i=1}^{I} m_i \left(c_{I,i}^2 + \sum_{j \in \mathcal{V}_i} c_{II,i,j}^2 p_{i,j} n_{i,j} \right)$$

Again, following the steps of the proof of Theorem 2.1, we arrive at

$$m_i n_{i,j} = \frac{\sqrt{\lambda_i} B_{i,j}}{\sqrt{\mu} c_{II,i,j} \sqrt{p_{i,j}}} \quad \text{and} \quad m_i = \frac{\sqrt{\lambda_i} A_i}{\sqrt{\mu} c_{I,i}}. \tag{22}$$

Thus, the formula for $n_{i,j}$ follows.

Inserting both expressions from (22) into the cost constraints, we obtain

$$\sqrt{\mu} = \frac{\underline{a}^T \underline{v}}{C},$$

where

$$\underline{a} = (a_1, \ldots, a_I)^T \qquad \text{with} \quad a_i = c_{I,i} A_i + \sum_{j \in \mathcal{V}_i} c_{II,i,j} B_{i,j} \sqrt{p_{i,j}}$$

and $\underline{v} = (v_1, \ldots, v_I)$ with $v_i = \sqrt{\lambda_i}$.

On the other hand, plugging formulas (22) into the constraints $T_i = \kappa_i T$, we get

$$T v_i = \sqrt{\mu}\, C\, a_i = \underline{a}^T \underline{v}\, a_i, \quad i = 1, \ldots, I,$$

which is equivalent to

$$\tfrac{1}{C}\,\underline{a}\,\underline{a}^T\,\underline{v} = T\underline{v}.$$

That is, $\underline{v}$ is an eigenvector of the matrix $\mathbf{D} = \frac{1}{C}\underline{a}\,\underline{a}^T$ associated with eigenvalue T. Since the matrix $\mathbf{D}$ is semi-positive definite of rank 1, the number T is its only nonzero simple positive eigenvalue. Moreover, note that $\underline{v}^* := \sqrt{C}\underline{a}$ is the eigenvector of $\mathbf{D}$ associated with eigenvalue $\|\underline{a}\|^2/C$. Finally, from (22), we obtain

$$m_i = C\,\frac{a_i A_i}{c_{I,i}\,\|\underline{a}\|^2} \qquad \text{and} \qquad n_{i,j} = \frac{c_{I,i}\,B_{i,j}}{A_i\,c_{II,i,j}\sqrt{p_{i,j}}}, \quad j \in \mathcal{V}_i,\ i = 1, \ldots, I.$$

□

4.2 SRSWOR at the first stage and pps sampling at the second stage

For completeness of the picture for two-stage sampling involving pps approach, let us consider the situation when the PSUs sample $\mathcal{S}^{(I)}$ is drawn through SRSWOR and the SSUs sample by sampling with replacement with probabilities p_k proportional to the size of kth unit. Here, the simplification of Sect. 4.1 is no longer available. This case falls under the general framework developed in Sect. 3.3.

The standard estimator of the total is

$$\hat{t} = \frac{M}{m} \sum_{j \in \mathcal{S}^{(I)}} \frac{1}{n_j} \sum_{\ell=1}^{n_j} \frac{y_{K_{j,\ell}}}{p_{K_{j,\ell}}},$$

where m is the number of PSUs drawn by the SRSWOR from the total of M PSUs in the population, n_j is the number of "with-replacement" draws from jth PSU, $K_{j,\ell}$ is the SSU drawn from jth PSU in the ℓth draw (with replacement), $j \in \mathcal{V}$ (PSUs population of size M). Evidently, $\hat{t}$ is unbiased for the population total. Its variance is

$$D^2(\hat{t}) = M^2\left(\tfrac{1}{m} - \tfrac{1}{M}\right) S_I^2 + \frac{M}{m} \sum_{j \in \mathcal{V}} \frac{1}{n_j} D_{II,j}^2,$$

where

$$S_I^2 = \frac{1}{M-1} \sum_{j \in \mathcal{V}} \left(t_j - \bar{t}\right)^2, \quad \bar{t} = \frac{1}{M} \sum_{j \in \mathcal{V}} t_j,$$

and for any $j \in V$

$$t_j = \sum_{k \in PSU_j} y_k, \qquad D_{II,j}^2 = \sum_{k \in PSU_j} \left(\frac{y_k}{p_k} - t_j\right)^2 p_k.$$

Consequently, to obtain the optimal allocation of the samples (on the first and second stage) in the domains with given priority weights κ_i, $i = 1, \ldots, I$, we need to minimize

$$T_i = \frac{1}{\tau_i^2}\left(\frac{M_i^2 S_{I,i}^2}{m_i} + \frac{M_i}{m_i}\sum_{j\in\mathcal{V}_i}\frac{1}{n_{i,j}}D_{II,i,j}^2 - M_i S_{I,i}^2\right), \quad i = 1, \ldots, I,$$

under the constraints given by the priority weights $T_i = \kappa_i T$ and the expected cost constraint

$$\sum_{i=1}^{I}\left(c_{I,i}^2 m_i + \frac{m_i}{M_i}\sum_{j\in\mathcal{V}_i} c_{II,i,j}^2 n_{i,j}\right) = C,$$

where C is the total expected cost of the survey, $c_{I,i}$ is the cost incurred by a PSU from V_i (assumed to be constant within the domain) and $c_{II,i,j}$ is the cost incurred by a SSU belonging to the jth PSU from the ith domain.

Since the structure of the problem is exactly the same as for the one considered in Sect. 3.3, we conclude that the optimal allocation has the form

$$m_i = C\frac{v_i^*\gamma_i}{\sqrt{\kappa_i}\,c_{I,i}\,\sum_{r=1}^{I} v_r^*\frac{1}{\sqrt{\kappa_r}}\left(c_{I,r}\gamma_r + \sum_{t\in\mathcal{V}_r} c_{II,r,t}\,\beta_{r,t}\right)}, \quad i = 1, \ldots, I,$$

and

$$n_{i,j} = C\frac{M_i}{m_i}\frac{v_i^*\beta_{i,j}}{\sqrt{\kappa_i}\,c_{II,i,j}\,\sum_{r=1}^{I} v_r^*\frac{1}{\sqrt{\kappa_r}}\left(c_{I,r}\,\gamma_r + \sum_{t\in\mathcal{V}_r} c_{II,r,t}\,\beta_{r,t}\right)}, \quad j \in \mathcal{V}_i,\ i = 1, \ldots, I,$$

where

$$\gamma_i^2 = \frac{M_i S_{I,i}^2}{\tau_i^2}, \qquad \beta_{i,j}^2 = \frac{D_{II,i,j}^2}{\tau_i^2}$$

and $\underline{v}^*$ is the eigenvector (having all components of the same sign) of the matrix $\mathbf{D} = \frac{\underline{a}\underline{a}^T}{C} - \mathrm{diag}(\underline{c})$ with components of $\underline{a}$ of the form

$$a_i = \frac{c_{I,i}\gamma_i + \sum_{j\in\mathcal{V}_i} c_{II,i,j}\,\beta_{i,j}}{\sqrt{\kappa_i}}, \quad \text{and} \quad c_i = \frac{M_i S_{I,i}^2}{\tau_i^2}, \quad i = 1, \ldots, I.$$

5 Three-stage sampling without stratification

In multistage sampling, typically, we do not go beyond three-stage sampling. This scheme is described in detail, for example, in Särndal et al. (1992, Ch. 4.4.2). The optimal allocation of the sample between three stages under the cost constraints, with the additional simplifying assumption that the sizes of SSU and TSU (tertiary

sampling unit) samples do not depend on PSU or SSU, respectively, had been studied already in Cochran (1977, Ch. 10.8) (see also Singh 2003, Ch. 10.4). Recently, the optimal allocation procedure, using a simplified variance formula with the standard constraints regarding the total costs, was designed in Valliant et al. (2015) as a part of their PracTools R package. An application of such a simple three-stage sampling design is given, for example, in Tate and Hudgens (2007).

In this section, we analyze the eigenproblem approach to the domain-efficient allocation of sample in three-stage sampling, but first we recall the Neyman-type optimal allocation in the case of no domains.

It is well known that the variance of the standard estimator $\hat{t}$ of the total of a variable $\mathcal{Y}$ in a population U under three-stage sampling with SRSWOR on every stage has the form

$$D^2 = \left(\tfrac{1}{\ell} - \tfrac{1}{L}\right) L^2 S_I^2 + \tfrac{L}{\ell} \sum_{j=1}^{L} \left(\tfrac{1}{m_j} - \tfrac{1}{M_j}\right) M_j^2 S_{II,j}^2 + \tfrac{L}{\ell} \sum_{j=1}^{L} \tfrac{M_j}{m_j} \sum_{k=1}^{M_j} \left(\tfrac{1}{n_{j,k}} - \tfrac{1}{N_{j,k}}\right) N_{j,k}^2 S_{III,j,k}^2,$$

where L and ℓ denote the number of PSUs, M_j and m_j the number of SSUs in the jth PSU, $N_{j,k}$ and $n_{j,k}$ the number of TSUs in (j,k)th SSU, in population and in the sample, respectively; moreover, S_I^2, $S_{II,j}^2$, $S_{III,j,k}^2$ denote population variances for PSUs in U, SSUs in jth PSU and TSUs in (j,k)th SSU.

Then, the minimization of D^2 (or D^2/τ^2) under the cost constraints

$$c_I^2 \ell + \tfrac{\ell}{L} \sum_{j=1}^{L} c_{II,j}^2 m_j + \tfrac{\ell}{L} \sum_{j=1}^{L} \tfrac{m_j}{M_j} \sum_{k=1}^{M_j} c_{III,j,k}^2 n_{j,k} = C, \tag{23}$$

where c_I^2, $c_{II,j}^2$ and $c_{III,j,k}^2$ are costs generated by each PSU, each SSU belonging to jth PSU and each TSU belonging to kth TSU from jth PSU of ith subpopulation, while C denotes the overall cost of the survey, obtained through the standard Neyman approach leads to the following optimal allocation solution

$$\ell = \frac{C\gamma}{c_I \left(c_I \gamma + \sum_{j=1}^{L} \left(c_{II,j} \beta_j + \sum_{k=1}^{M_j} c_{III,j,k} \delta_{j,k}\right)\right)}, \tag{24}$$

where $\gamma^2 = L(L S_I^2 - \sum_{j=1}^{L} M_j S_{II,j}^2)$ is assumed to be positive,

$$m_j = \tfrac{c_I L \beta_j}{c_{II,j} \gamma}, \quad j = 1, \ldots, L, \tag{25}$$

where $\beta_j^2 = M_j \left(M_j S_{II,j}^2 - \sum_{k=1}^{j} N_{j,k} S_{III,j,k}^2\right)$ is also assumed to be positive, and

$$n_{j,k} = \tfrac{c_{II,j} M_j \delta_{j,k}}{c_{III,j,k} \beta_j}, \quad k = 1 \ldots, M_j, \ j = 1, \ldots, L, \tag{26}$$

where $\delta_{j,k}^2 = N_{j,k}^2 S_{III,j,k}^2$. The optimal variance assumes the form

$$D_{opt}^2 = \frac{1}{C}\left(c_I\gamma + \sum_{j=1}^{L}\left(c_{II,j}\beta_j + \sum_{k=1}^{M_j} c_{III,j,k}\delta_{j,k}\right)\right)^2 - LS_I^2. \tag{27}$$

Since we will be considering three-stage sampling in subpopulations, all these quantities will be related to a subpopulation by additional subscript $i = 1, \ldots, I$.

Similarly, as in previous sections, we will be interested in minimization of relative variances in subpopulations, provided they satisfy the constraints defined by priority weights $(\kappa_i,\ i = 1, \ldots, I)$, which assume the form

$$\frac{\left(\frac{1}{\ell_i} - \frac{1}{L_i}\right) L_i^2 S_{I,i}^2 + \frac{L_i}{\ell_i}\sum_{j=1}^{L_i}\left(\frac{1}{m_{i,j}} - \frac{1}{M_{i,j}}\right) M_{i,j}^2 S_{II,i,j}^2 + \frac{L_i}{\ell_i}\sum_{j=1}^{L_i}\frac{M_{i,j}}{m_{i,j}}\sum_{k=1}^{M_{i,j}}\left(\frac{1}{n_{i,j,k}} - \frac{1}{N_{i,j,k}}\right) N_{i,j,k}^2 S_{III,i,j,k}^2}{\tau_i^2} = \kappa_i T \tag{28}$$

where T is unknown and has to be minimized under an additional total EVC constraint which in the case of three-stage sampling assumes the form

$$\sum_{i=1}^{I} c_{I,i}^2 \ell_i + \sum_{i=1}^{I}\frac{\ell_i}{L_i}\sum_{j=1}^{L_i} c_{II,i,j}^2 m_{i,j} + \sum_{i=1}^{I}\frac{\ell_i}{L_i}\sum_{j=1}^{L_i}\frac{m_{i,j}}{M_{i,j}}\sum_{k=1}^{M_{i,j}} c_{III,i,j,k}^2 n_{i,j,k} = C, \tag{29}$$

where $c_{I,i}^2$, $c_{II,i,j}^2$ and $c_{III,i,j,k}^2$ are costs generated by each PSU from ith subpopulation, each SSU belonging to jth PSU of ith subpopulation and, each TSU belonging to kth TSU from jth PSU of ith subpopulation, while C denotes the overall cost of the survey.

Therefore, the Lagrange function, up to a constant shift, is of a rather complicated, though regular form:

$$F(T, \underline{\ell}, \underline{m}, \underline{n}) = T + \sum_{i=1}^{I}\frac{\lambda_i}{\tau_i^2}\left(\frac{\gamma_i^2 + L_i\sum_{j=1}^{L_i}\frac{\beta_{i,j}^2 + M_{i,j}\sum_{k=1}^{M_{i,j}}\frac{\delta_{i,j,k}^2}{n_{i,j,k}}}{m_{i,j}}}{\kappa_i\ell_i} - T\right)$$
$$+ \mu\left(\sum_{i=1}^{I}\ell_i\left(c_{I,i}^2 + \frac{1}{L_i}\sum_{j=1}^{L_i} m_{i,j}\left(c_{II,i,j}^2 + \frac{1}{M_{i,j}}\sum_{k=1}^{M_{i,j}} c_{III,i,j,k}^2 n_{i,j,k}\right)\right)\right),$$

where

$$\gamma_i^2 = L_i \left(L_i S_{I,i}^2 - \sum_{j=1}^{L_i} M_{i,j} S_{II,i,j}^2 \right),$$

$$\beta_{i,j}^2 = M_{i,j} \left(M_{i,j} S_{II,i,j}^2 - \sum_{k=1}^{M_{i,j}} N_{i,j,k} S_{III,i,j,k}^2 \right) \quad \text{and} \quad \delta_{i,j,k}^2 = N_{i,j,k}^2 S_{III,i,j,k}^2.$$

Denoting $\rho_i = \tau^2 \sqrt{\kappa_i}$ and differentiating F with respect to:

1. ℓ_i we get

$$-\frac{\lambda_i}{\rho_i^2 \ell_i^2} \left(\gamma_i^2 + L_i \sum_{j=1}^{L_i} \frac{1}{m_{i,j}} \left(\beta_{i,j}^2 + M_{i,j} \sum_{k=1}^{M_{i,j}} \frac{\delta_{i,j,k}^2}{n_{i,j,k}} \right) \right)$$
$$+ \mu \left(c_{I,i}^2 + \frac{1}{L_i} \sum_{j=1}^{L_i} m_{i,j} \left(c_{II,i,j}^2 + \frac{1}{M_{i,j}} \sum_{k=1}^{M_{i,j}} c_{III,i,j,k}^2 n_{i,j,k} \right) \right) = 0, \quad (30)$$

2. $m_{i,j}$ we get

$$-\frac{\lambda_i L_i}{\rho_i^2 \ell_i m_{i,j}^2} \left(\beta_{i,j}^2 + M_{i,j} \sum_{k=1}^{M_{i,j}} \frac{\delta_{i,j,k}^2}{n_{i,j,k}} \right)$$
$$+ \mu \frac{\ell_i}{L_i} \left(c_{II,i,j}^2 + \frac{1}{M_{i,j}} \sum_{k=1}^{M_{i,j}} c_{III,i,j,k}^2 n_{i,j,k} \right) = 0, \quad (31)$$

3. $n_{i,j,k}$ we get

$$-\frac{\lambda_i L_i M_{i,j}}{\rho_i^2 \ell_i m_{i,j} n_{i,j,k}^2} \delta_{i,j,k}^2 + \mu \frac{\ell_i}{L_i} \frac{m_{i,j}}{M_{i,j}} c_{III,i,j,k}^2 = 0. \quad (32)$$

Note that from (32), we get

$$\ell_i m_{i,j} n_{i,j,k} = \frac{\sqrt{\lambda_i} L_i M_{i,j} \delta_{i,j,k}}{\sqrt{\mu} \sqrt{\kappa_i} c_{III,i,j,k}}. \quad (33)$$

Multiply now (32) by $n_{i,j,k}/m_{i,j}$ and insert it into (31). After cancellations, one gets

$$\ell_i m_{i,j} = \frac{\sqrt{\lambda_i} L_i \beta_{i,j}}{\sqrt{\mu} \sqrt{\kappa_i} c_{II,i,j}}. \quad (34)$$

Now multiply (31) by $m_{i,j}/\ell_i$ and insert both into (30). After cancellations, one gets

$$\ell_i = \frac{\sqrt{\lambda_i} \gamma_i}{\sqrt{\mu} \sqrt{\kappa_i} c_{I,i}}. \quad (35)$$

Formulas for $n_{i,j,k}$ and $m_{i,j}$ follow directly from (33) and (34) and from (34) and (35), respectively.

After inserting (33), (34) and (35) into (29), we obtain

$$\sqrt{\mu} = \frac{1}{c} \sum_{i=1}^{I} \sqrt{\lambda_i} \, \frac{c_{I,i}\gamma_i + \sum_{j=1}^{L_i} \left(c_{II,i,j}\beta_{i,j} + \sum_{k=1}^{M_{i,j}} c_{III,i,j,k}\delta_{i,j,k} \right)}{\sqrt{\kappa_i}}. \tag{36}$$

On the other hand, if we plug (33), (34) and (35) into the constraint (28), we obtain

$$\frac{\sqrt{\mu}\sqrt{\kappa_i}\, c_{I,i}\,\gamma_i}{\sqrt{\lambda_i}} + L_i \sum_{j=1}^{L_i} \left(\frac{\sqrt{\mu}\sqrt{\kappa_i}\, c_{II,i,j}\,\beta_{i,j}}{\sqrt{\lambda_i}\, L_i} + M_{i,j} \sum_{k=1}^{M_{i,j}} \frac{\sqrt{\mu}\sqrt{\kappa_i}\, c_{III,i,j,k}\,\delta_{i,j,k}}{\sqrt{\lambda_i}\, L_i M_{i,j}} \right)$$
$$- \frac{L_i S_{I,i}^2}{\kappa_i \tau_i^2} = \kappa_i T.$$

Denote $v_i = \sqrt{\lambda_i}$ and multiply the above equation by v_i/κ_i. Then, after cancellations and upon denoting

$$v_i = c_{I,i}\,\gamma_i + \sum_{j=1}^{L_i} \left(c_{II,i,j}\,\beta_{i,j} + \sum_{k=1}^{M_{i,j}} c_{III,i,j,k}\,\delta_{i,j,k} \right)$$

we have

$$\sqrt{\mu}\,\frac{v_i}{\sqrt{\kappa_i}} - c_i v_i = T v_i,$$

where $c_i = \frac{L_i S_{I,i}^2}{\kappa_i \tau_i^2}$, $i = 1, \ldots, I$. Expanding now $\sqrt{\mu}$ according to (36), we conclude that

$$\frac{\underline{a}\underline{a}^T}{C}\underline{v} - \mathrm{diag}(\underline{c})\underline{v} = T\underline{v},$$

where $\underline{a} = (a_i,\ i = 1, \ldots, I)^T$ with

$$a_i = \frac{v_i}{\sqrt{\kappa_i}}, \quad i = 1, \ldots, I,$$

and $\underline{c} = (c_i,\ i = 1, \ldots, I)$. Consequently, for $\mathbf{D} = \frac{\underline{a}\underline{a}^T}{C} - \mathrm{diag}(\underline{c})$ we have $\mathbf{D}\underline{v} = T\underline{v}$.

Theorem 5.1 *Assume that*

$$L_i S_{I,i}^2 - \sum_{j=1}^{L_i} M_{i,j} S_{II,i,j}^2 > 0, \quad i = 1, \ldots, I,$$

and

$$M_{i,j}S^2_{II,i,j} - \sum_{k=1}^{M_{i,j}} N_{i,j,k}S^2_{III,i,j,k} > 0, \quad j = 1, \ldots, L_i, \ i = 1, \ldots, I.$$

Assume that the matrix $\boldsymbol{D}$ *has a positive eigenvalue* λ^**. Then, it is unique and simple and the respective eigenvector* $\underline{v}^*$ *has all coordinates of the same sign.*

The allocation $\underline{\ell}$*,* $\underline{m}$ *and* $\underline{n}$ *which minimizes all relative domain-wise variances* T_i*,* $i = 1, \ldots, I$*, (as well as the relative variance* S *in the whole population) under the constraints* $T_i = \kappa_i T$*,* $i = 1, \ldots, I$*, and under the EVC constraint* (29) *has the form*

$$\ell_i = C \frac{v_i^* \gamma_i}{\sqrt{\kappa_i}\, c_{I,i} \sum_{r=1}^{I} v_r^* \frac{1}{\sqrt{\kappa_r}} v_r}, \tag{37}$$

$$m_{i,j} = \frac{c_{I,i} L_i \beta_{i,j}}{\gamma_i c_{II,i,j}} \tag{38}$$

and

$$n_{i,j,k} = \frac{c_{II,i,j} M_{i,j} \delta_{i,j,k}}{\beta_{i,j} c_{III,i,j,k}} \tag{39}$$

for any $k = 1, \ldots, M_{i,j}$*,* $j = 1, \ldots, L_i$*,* $i = 1, \ldots, I$*.*

Moreover, the minimal relative variances in the domains are $T_i = \kappa_i, T$*,* $i = 1, \ldots, I$*, where* T *the base of the relative variance has the form*

$$T = \lambda^* = \frac{1}{C} \left(\sum_{i=1}^{I} \frac{v_i \tau_i \sqrt{\kappa_i}}{v_i^*} \right) \left(\sum_{i=1}^{I} \frac{v_i v_i^*}{\tau_i \sqrt{\kappa_i}} \right) - \sum_{i=1}^{I} L_i S^2_{I,i}. \tag{40}$$

Remark 5.1 Note that in the case of no domains, i.e., when $I = 1$, the allocation formulas (37)–(39) as well as the formula for the optimal variance (40) are simplified to the Neyman-type allocation and optimal variance formulas as given in (24)–(26) and (27), respectively.

Note also that only the allocation of the first-stage sample and the optimal base of the variance depend on the eigenvector $\underline{v}^*$. Formulas (38) and (39) for the allocation of the second- and third-stage samples are given directly in terms of population quantities with no reference to the eigenvector $\underline{v}^*$.

6 Conclusions

In this paper, we search for Neyman-type solutions to domains-efficient allocation in multistage stratified sampling. Such a solution can be seen as an alternative to the purely numerical one proposed in CRH for stratified single-stage scheme. We develop the eigenproblem method originating in NW and use eigenvalues and eigenvectors for allocation which, under specified priority coefficients for the constraints on the domains relative variances, assures optimal estimation both in the whole population

and in the domains. In particular, we consider two- and three-stage sampling. The novelty of the solutions we provide here, with respect to what is known for eigenproblem approach to domains-efficient allocation, is with respect to several aspects. The most important is that, in contrast to earlier situations, as, for example, in WW, a single total cost constraint is taken under account. In previous papers instead, two constraints related to (expected) samples sizes of the PSUs and SSUs, respectively, were jointly imposed. In those papers, the two-stage sampling with SRSWOR (or Hartley–Rao) schemes with stratification either at the first or the second stage was considered. Here, we apply the eigenproblem methodology also to new sampling schemes: stratified SRSWOR at both stages as well as pps sampling with replacement and SRSWOR either at the first or the second stage and to the three-stage sampling with SRSWOR at each stage. In each of these cases, the allocation which assures optimality (under given domain priority weights) of estimators of domain totals is given in terms of eigenvectors of a population-dependent matrix (which typically is rank-one perturbations of a diagonal matrix). Moreover, the standard errors of the estimates in the domains and in the whole population are given in terms of the respective eigenvalue. The latter allows to interpret the solution as a direct generalization of Neyman-type optimal allocation to the multi-domain case. Another important consequence of the approach we use here is that through the analytic formulas, we obtained, the structure of the optimal allocation can be seen. For example, it is visible that only the first-stage optimal allocation is influenced by the eigenvector $\underline{v}^*$ of the population matrix $\mathbf{D}$.

Acknowledgements We are very thankful to two anonymous referees whose remarks allowed us to improve presentation of the paper. We are also grateful to R. Münnich for interesting discussions on different aspects of optimal allocation problems. Thanks to R. Wieczorkowski for help with computations regarding the example in "Appendix."

7 Appendix

Here, we compare the optimal multi-domain allocation obtained by Neyman-type approach to minimization of the weighted L_2-norm for the vector of relative variances of domains

$$\sum_{i=1}^{I} w_i T_i$$

with the eigenproblem approach to minimization of relative domain variances with priority weights

$$T_i = \kappa_i T, \quad i = 1, \ldots, I.$$

We consider only the easiest situation of stratified SRSWOR in domains.

The numerical example is based on data from the survey of "Turnover in the Trade Sector in Poland." In this survey, a population of 493,863 units was divided into 12 domains and stratified into 66 strata. A sample of 3600 units is to be allocated.

We consider two choices of $(w_i)_{i=1,\ldots,12}$ and two choices of $(\kappa_i)_{i=1,\ldots,12}$.

In L_2-norm approach, we consider either $w_i = 1$, $i = 1, \ldots, 12$ (labeled L1) or $w_1 = \cdots = w_6 = 4$, $w_7 = \cdots = w_{12} = 12$ (labeled L2).

In the eigenproblem approach, we consider either $\kappa_i = 1$, $i = 1, \ldots, 12$ (labeled E1) or $\kappa_1 = \cdots = \kappa_6 = 4$, $\kappa_7 = \cdots = \kappa_{12} = 1$ (labeled E2).

The results are given in the table below:

i	H_i	E1:n_i	E1:cv_i	L1:n_i	L1:cv_i	E2:n_i	E2:cv_i	L2:n_i	L2:cv_i
1	9	389	9.56	722	6.97	146	15.68	1189	5.44
2	3	358	9.56	249	11.47	134	15.68	209	12.53
3	3	387	9.56	222	13.10	159	15.68	186	14.42
4	3	53	9.56	341	3.75	20	15.68	286	4.09
5	3	219	9.56	183	10.45	82	15.68	154	11.42
6	3	157	9.56	181	8.88	59	15.68	152	9.72
7	3	100	9.56	186	6.98	148	7.84	156	7.63
8	6	152	9.56	124	10.59	226	7.84	104	11.58
9	9	137	9.56	183	8.26	202	7.84	153	9.03
10	12	182	9.56	367	6.71	270	7.84	308	7.34
11	3	735	9.56	423	12.67	1077	7.84	355	13.86
12	9	731	9.56	416	12.70	1079	7.84	349	13.88

In the table above, H_i, n_i and cv_i denote, respectively, the number of strata, the (rounded) number of units to be drawn, cv of the estimator in the domain i, $i = 1, \ldots, I$. Note that, as designed, E1:cv_i are identical throughout domains, while E2:cv_i are twice larger for the first six domains. Note also that in both cases of L_2-norm minimization the cv_i's fluctuate and do not seem to be controlled easily.

References

Ballin, M., Barcaroli, G.: Joint determination of optimal stratification and sample allocation using genetic algorithm. Survey Methodol. **39**(2), 369–393 (2013)

Choudhry, G.H., Rao, J.N.K., Hidiroglou, M.A.: On sample allocation for efficient domain estimation. Survey Methodol. **38**(1), 23–29 (2012)

Clark, R.G.: Sampling of subpopulations in two stage surveys. Stat. Med. **28**(29), 3697–3717 (2009)

Clark, R.G., Steel, D.G.: Optimum allocation of sample to strata and stages with simple additional constraints. J. R. Stat. Soc. D **49**, 197–207 (2000)

Cochran, W.G.: Sampling Techniques, 3rd edn. Wiley, New York (1977)

Costa, A., Satorra, A., Ventura, E.: Using composite estimators to improve both domain and total area estimation. SORT **19**, 69–86 (2004)

Friedrich, U., Münnich, R., de Vries, S., Wagner, M.: Fast integer-valued algorithm for optimal allocations under constraints in stratified sampling. Comput. Stat. Data Anal. **92**, 1–12 (2015)

Friedrich, U., Münnich, R., Rupp, M.: Multivariate optimal allocation with box-constraints. Aust. J. Stat. **47**, 33–52 (2018)

Gabler, S., Ganninger, M., Münnich, R.: Optimal allocation of the sample size to strata under box constraints. Metrika **75**(2), 151–161 (2012)

Kato, T.: A Short Introduction to Perturbation Theory for Linear Operators. Springer, New York (1981)
Keto, M., Pahkinen, E.: Sample allocation for efficient model-based small area estimation. Survey Methodol. **43**(1), 93–106 (2017)
Khan, M.G.M., Maiti, T., Ahsan, M.J.: An optimal multivariate stratified sampling design using auxiliary information: an integer solution using goal programming approach. J. Off. Stat. **26**(4), 695–708 (2010)
Kozak, M.: Method of multivariate sample allocation in agricultural surveys. Biom. Colloq. **34**, 241–250 (2004)
Kozak, M., Zieliński, A.: Sample allocation between domains and strata. Int. J. Appl. Math. Stat. **3**, 19–40 (2005)
Kozak, M., Zieliński, A., Singh, S.: Stratified two-stage sampling in domains: sample allocation between domains, strata and sampling stages. Stat. Probab. Lett. **78**, 970–974 (2008)
Lednicki, B., Wesołowski, J.: Localization of sample between subpopulations. Wiad. Statyst. **39**(9), 2–4 (1994). (**in Polish**)
Lednicki, B., Wieczorkowski, R.: Optimal stratification and sample allocation between subpopulations and strata. Stat. Trans. **6**(2), 287–305 (2003)
Longford, N.T.: Sample size calculation for small-area estimation. Survey Methodol. **32**, 87–96 (2006)
Molefe, W., Clark, R.G.: Model-assisted optimal allocation for planned domains using composite estimation. Survey Methodol. **41**(2), 377–387 (2015)
Münnich, R., Sachs, E.W., Wagner, M.: Numerical solution to optimal allocation problems in stratified sampling under box constraints. Adv. Stat. Anal. **96**(3), 435–450 (2012)
Niemiro, W., Wesołowski, J.: Fixed precision allocation in two-stage sampling. Appl. Math. **28**, 73–82 (2001)
Saini, M., Kumar, A.: Optimum allocation in stratified two stage design by using double sampling for multivariate surveys. Probab. Stat. Forum **8**, 19–23 (2015)
Särndal, C.-E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer, New York (1992)
Singh, S.: Advanced Sampling Theory with Applications. Kluwer, Dordrecht (2003)
Stenger, H., Gabler, S.: Combining random sampling and census strategies: justification of inclusion probabilities equal to 1. Metrika **61**, 137–156 (2005)
Tate, J.I., Hudgens, M.G.: Estimating population size with two- and three-stage sampling designs. Am. J. Epidemiol. **165**(11), 1314–1320 (2007)
Valliant, R., Dever, J.A., Kreuter, F.: PracTools: computations for design of finite population samples. R J. **7**(2), 163–176 (2015)
Valliant, R., Dever, J.A., Kreuter, F.: Practical Tools for Designing and Weighting Survey Samples. Springer, Berlin (2013)
Wesołowski, J., Wieczorkowski, R.: An eigenproblem approach to optimal equal-precision sample allocation in subpopulations. Commun. Stat. Theory Methods **46**(5), 2212–2231 (2017)
Wright, T.: Exact optimal sample allocation: more efficient than Neyman. Stat. Probab. Lett. **129**, 50–57 (2017)