STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 1–12, DOI 10.21307/stattrans-2019-031
Submitted – 09.09.2019; Paper ready for publication – 12.11.2019

MULTI-DOMAIN NEYMAN-TCHUPROV OPTIMAL ALLOCATION

Jacek Wesołowski[1]

ABSTRACT

The eigenproblem solution of the multi-domain efficient allocation is identified as a direct generalization of the classical Neyman-Tchuprov optimal allocation in stratified SRSWOR. This is achieved through analysis of eigenvalues and eigenvectors of a suitable population-based matrix $\mathbf{D}$. Such a solution is an analytical companion to NLP approaches, which are often used in applications, see, e.g. Choudhry, Rao and Hidiroglou (2012). In this paper we are interested rather in the structure of the optimal allocation vector and relative variance than in such purely numerical tools (although the eigenproblem solution provides also numerical solutions, see, e.g. Wesołowski and Wieczorkowski (2017)). The domain-wise optimal allocation and the respective optimal variance of the estimator are determined by the unique *direction* (defined in terms of the *positive* eigenvector of matrix $\mathbf{D}$) in the space $\mathbb{R}^I$, where I is the number of domains in the population.

Key words: Neyman-Tchuprov allocation, multi-domain allocation, eigenproblem, stratified SRSWOR.

MSC2010 Classification: 62D05

1. Introduction

Consider a stratified SRSWOR in a population U of size N with strata $W_1,\ldots,W_H$, which form a partition of U and let N_h denote the size of the stratum W_h, $h=1,\ldots,H$. For a variable $\mathscr{Y}$ defined on U we denote $y_k=\mathscr{Y}(k)$, $k\in U$. The standard estimator of the total $\tau=\sum_{k\in U}y_k$ has the form $\hat{\tau}_{st}=\sum_{h=1}^H N_h\bar{y}_h$, where $\bar{y}_h=\frac{1}{n_h}\sum_{k\in\mathscr{S}_h}y_k$ with n_h denoting the size of the sample $\mathscr{S}_h$ drawn from W_h, $h=1,\ldots,H$. The variance of $\hat{\tau}_{st}$ is $D^2=\sum_{h=1}^H\left(\frac{1}{n_h}-\frac{1}{N_h}\right)N_h^2S_h^2$, where $S_h^2=\frac{1}{N_h-1}\sum_{k\in W_h}(y_k-\bar{y}_{W_h})^2$ is the population variance in W_h, $h=1,\ldots,H$.

In such a setting one of the main issues is the optimal allocation, $\underline{n}=(n_1,\ldots,n_H)$, 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, e.g. by the total sample size $\sum_{h=1}^H n_h$. An alternative approach is by fixing the total sample size $n=\sum_{h=1}^H n_h$ and minimize the (relative) variance of $\hat{\tau}_{st}$. Both cases are solved through the classical Neyman-Tchuprov optimal allocation procedure (see, e.g. Särndal, Swensson and Wretman, 1992). In particular, it is well known that under the constraint

[1] Statistics Poland and Warsaw University of Technology, Poland. E-mail: wesolo@mini.pw.edu.pl.

$n = n_1 + \ldots + n_H$ the Neyman-Tchuprov optimal allocation is

$$n_h = n\frac{N_h S_h}{\sum_{g=1}^{H} N_g S_g}, \qquad h = 1, \ldots, H. \tag{1}$$

Then, the optimal relative variance assumes the form

$$D_{opt}^2 = \frac{1}{\tau^2}\left[\frac{1}{n}\left(\sum_{h=1}^{H} N_h S_h\right)^2 - \sum_{h=1}^{H} N_h S_h^2\right]. \tag{2}$$

Note that in order for (1) to be a valid solution it is necessary that

$$n < \frac{\left(\sum_{h=1}^{H} N_h S_h\right)^2}{\sum_{h=1}^{H} N_h S_h^2}. \tag{3}$$

Otherwise, (2) gives a non-positive value which is forbidden.

On the other hand, we may want to minimize $\sum_{h=1}^{H} n_h$ under the constraint imposed on the variance of $\hat{\tau}_{st}$ of the form

$$\sum_{h=1}^{H}\left(\frac{1}{n_h} - \frac{1}{N_h}\right) N_h^2 S_h^2 = T,$$

where T is given. Then, it is well known that the optimal allocation is given by

$$n_h = \frac{N_h S_h}{T + \sum_{g=1}^{H} N_g S_g^2} \sum_{g=1}^{H} N_g S_g, \quad h = 1, \ldots, H. \tag{4}$$

The optimal size of the sample is

$$n_{opt} = \frac{\left(\sum_{h=1}^{H} N_h S_h\right)^2}{T + \sum_{h=1}^{H} N_h S_h^2}. \tag{5}$$

Note that these two solutions are dual in the following sense: If we insert $n := n_{opt}$ as given in (5) in the formula (1) we obtain (4). Similarly, if we insert $T := \tau^2 D_{opt}^2$ as given in (2) in the formula (4) we obtain (1).

However, even if (3) is satisfied the Neyman solution may still not be satisfactory: it may happen that the formula (1) yields $n_h > N_h$ for some $h \in \{1, \ldots, H\}$. Moreover, n_h as given in (1) typically is not integer-valued. Therefore, in recent years there has been a growing interest in more refined allocation methods, mostly based on nonlinear programming (NLP), see, e.g. the monograph Valliant, Dever and Kreuter (2013) and references given therein (actually, the literature on the subject is more than abundant). Such procedures give remedies for the basic drawbacks of the Neyman allocation, by imposing block constraints of the form $0 < m_h \le n_h \le N_h$, $h = 1, \ldots, H$, on entries of the allocation vector $\underline{n}$. Recently numerical procedures for optimal positive integer solutions also have appeared in the literature, see, e.g. Friedrich, Münnich, de Vries and Wagner (2015) or Wright (2017). Nevertheless, the Neyman-Tchuprov solution remains the only one which gives insight into the analytic structure of the optimal allocation and the optimal variance. For example, it is obvious from (2) that, up to a constant additive term (which is typically small), the optimal (relative) variance is of order $1/n$.

The situation becomes much more complex in the case of multi-domain efficient allocation. In such a setting the population is partitioned into disjoint domains (eventually, domains are further partitioned into strata). The task is to allocate the sample in the domains (eventually in the strata in each domain) in such a way that, simultaneously, the estimators of the total value of a given variable in every domain and in the whole population have minimal variances or relative variances (the precise formulation of the problem is given at the beginning of Section 2). Apparently, such a statement of the allocation problem is natural in many surveys when the goal is to estimate the parameter of interest not only for the whole population but for all the domains the population is partitioned into (e.g. admistration regions in a given country).

NLP procedures are often relatively easily adjustable to multi-domain efficient allocation. One example of such an adjustment is the procedure proposed in Choudhry, Rao and Hidiroglou (2012) (referred to as CRH in the sequel), which is explained in detail later on in this section. A respective useful adjustment of the Neyman-Tchuprov approach seems to be far more challenging.

One example of such an approach is provided by Longford (2006), where the author suggested to minimize (under a constraint given by the total sample size) the objective function

$$\sum_{i=1}^{I} P_i D^2(\bar{y}_i) + GP_+ D^2(\bar{y}_{st}), \tag{6}$$

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. Mathematically, this approach reduces to the Neyman allocation scheme. 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, Münnich and Rupp (2018), the approach of Gabler, Ganninger and Münnich (2012), in which additional box constraints on the strata (or domain) sample sizes are imposed, can be used also in this context. However, within such multi-domain adjustment 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 $D^2(\bar{y}_{st})$. In a numerical example given in the Appendix of Khan and Wesolowski (2019) it is visible that the control on the domain-wise efficiency within this kind of approach is rather problematic.

On contrary, the eigenproblem approach to the domain-optimal allocation gives a full control of the domain-wise efficiency. Moreover, the optimal allocation is given through explicit formulas, not just numerically. This is the essence of the present paper, in which we describe the eigenproblem approach as a generalization of the classical Neyman-Tchuprov methodology to the case of multi-domain optimal allocation. Such eigenproblem setting in the context of the domain-wise efficient allocation originally was proposed in Niemiro and Wesołowski (2001), and developed more recently in Wesołowski and Wieczorkowski (2017), and Khan and Wesołowski (2019). In the first of these papers the authors considered two-stage sampling schemes with SRSWOR and stratification either at the first or at the second stage. The setting considered there imposed jointly two sample size constraints: one on the sample size at the first stage (either in terms of the number of PSUs or SSUs) and one on the sample size at the second stage. Such constraints setting was studied also in the second paper, but for a wider family of sampling schemes: SRSWOR with stratification at both the first and the second stage and the Hartley-Rao scheme at the first stage and stratified SRSWOR at the second stage. Each of these schemes was also considered with additional constraints of equal SSU sample sizes within each of PSUs. The last of three papers dealt with the problem under a single sample size constraint, which was formulated in terms of the expected overall cost. Except of two-stage stratified SRSWOR sampling schemes taken into account in earlier papers, here a combination of *pps* sampling and stratified SRSOWR either at the first or at the second stage was also considered. Finally, the eigenproblem approach was applied in the three-stage sampling scheme with SRSWOR (with no stratification) at each stage. Survey applications and some additional refinements of the eigenproblem approach were given, e.g. in Kozak (2004), Kozak and Zieliński (2005) and Kozak, Zieliński and Singh (2008).

Before we move to a detailed description of the eigenproblem approach, we will first analyze the setting of CRH. These authors consider a population U partitioned into disjoint domains U_i, $i = 1,\dots,I$. In each domain U_i the sample of size n_i is drawn independently according to the SRSWOR, $i = 1,\dots,I$. The aim is to minimize the total sample size

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

under the constraints for relative variances of estimators of the domain totals

$$T_i := \tfrac{1}{\tau_i^2}\left(\tfrac{1}{n_i} - \tfrac{1}{N_i}\right) N_i^2 S_i^2 \le RV_{oi}, \quad i = 1,\dots,I, \tag{7}$$

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

$$\mathrm{S} := \tfrac{1}{\tau^2}\sum_{i=1}^{I}\left(\tfrac{1}{n_i} - \tfrac{1}{N_i}\right) N_i^2 S_i^2 \le RV_o. \tag{8}$$

Note that in this approach one specifies conditions for each of domains and for the whole population separately by assigning (given) upper bounds RV_{oi}, $i = 1,\dots,I$ and RV_o. The problem was solved in CRH under additional box constraints of the form $0 < n_i \le N_i$, $i = 1,\dots,I$, by the NLP method involving the popular Newton-Raphson algorithm. An extension of this approach to the case of stratified SRSWOR in each of the domains is rather straightforward.

Actually, in the case of the problem considered in CRH with constraints restricted to (7), i.e. to those imposed on the relative variances of the estimators of domain totals, the overall sample size is minimized by the trivial solution

$$n_i = \left\lceil \frac{N_i S_i^2}{\tau_i^2 T_i + N_i S_i^2} \right\rceil \in (0, N_i], \quad i = 1, \ldots, I.$$

Of course, it may happen that for such values of n_i's, $i = 1, \ldots, I$, condition (8) may not hold and only then the numerical procedure is needed.

NLP solutions, as the one described in CRH, typically are efficient and rather universal tools for optimal allocation in real surveys, when the practitioners need just numerical values for allocation of the sample in the particular survey. Nevertheless, they have forms of *black boxes*, that is, they are fed with population data (or estimates) and their output gives numbers responsible for allocation. Consequently, such numerical methods do not provide any information on the structure of optimal solutions (the allocation vector and the optimal relative variance), while such structural knowledge is important at the stage of survey design, e.g. for assigning proper efficiency priorities or for strata and/or domains construction.

To shed more light on the structure of optimal solutions we will analyze the eigenproblem approach. As it has been already mentioned, this methodology was developed recently in Wesołowski and Wieczorkowski (2017), referred to as WW in the sequel. To large extent, the results of the present paper depend on a correct interpretation of introductory Th. 2.3 of WW, where stratified SRSWOR in each of domains was analyzed. In comparison with WW, the formulas for domain optimal allocation, which are given in terms of an eigenvector of certain population dependent matrix, will be slightly modified here due to (known) priority weights assigned to each of domains. More importantly, a new analytic formula for the optimal relative variance in terms of this eigenvector will be derived. Combined together, these formulas allow one to conclude that the eigenvector solution is a direct generalization of the classical Neyman-Tchuprov allocation. This is the main message of the present article. In particular, we will see that in the case when there are no domains (i.e. when $I = 1$), the new formulas are reduced directly to (1) and (2). Moreover, in the situation when there are no strata in the domains the eigenvector solution is an analytic alternative to the NLP solution of CRH. Last but not least, let us mention that the analytic formulas we obtain can be also used for computing particular values of the optimal allocation vector (procedures for eigenvectors and eigenvalues are available in many computer packages, e.g. procedure *eigen* in the R package). Typically, numerical values obtained in this way, agree with NLP solutions.

Finally, let us mention that while being attractive at the analytical and theoretical level, the eigenproblem apporach has its limitations: e.g. it may give the allocation values which exceed the strata sizes. The NLP black box methods do not have this deficiency. Therefore, it would be plausible to overcome this drawback of the eigenproblem approach. In particular, it would be interesting to study the question whether a recursive version of the proposed methodology, similar to the recursive Neyman approach (see, e.g. Rem. 12.7.1 in Särndal, Swensson and Wretman

(1992)), gives the domain-wise efficient allocation with sample strata sizes within the strata size ranges. At present, this problem is under study. It would be also interesting to investigate possibilites of multivariate extensions of the eigenproblem methodology, since in many applications one would like to allocate the sample taking under account optimality with respect to more than one variable. A step in this direction was made in Kozak (2004).

2. Minimization of domain-wise relative variances

In the case of stratified domains, $U_i = \bigcup_{h=1}^{H_i} W_{i,h}$, $i = 1,\ldots,I$, the domain relative variances are

$$T_i = \frac{1}{\tau_i^2}\sum_{h=1}^{H_i}\left(\frac{1}{n_{i,h}} - \frac{1}{N_{i,h}}\right)N_{i,h}^2 S_{i,h}^2, \quad i = 1,\ldots,I, \tag{9}$$

where $N_{i,h} = \#(W_{i,h})$, $S_{i,h}^2 = \frac{1}{N_{i,h}-1}\sum_{k\in W_{i,h}}(y_k - \bar{y}_{i,h})^2$, with $\bar{y}_{i,h} = \frac{1}{N_{i,h}}\sum_{k\in W_{i,h}} y_k$, $\tau_i = \sum_{k\in U_i} y_k$ and $n_{i,h}$ being the size of the sample in hth stratum of ith domain, $i = 1,\ldots,I$. The relative total variance is

$$\mathrm{S} = \frac{1}{\tau^2}\sum_{i=1}^{I}\sum_{h=1}^{H_i}\left(\frac{1}{n_{i,h}} - \frac{1}{N_{i,h}}\right)N_{i,h}^2 S_{i,h}^2. \tag{10}$$

We will minimize simultaneously all T_i, $i = 1,\ldots,I$, as well as S under the constraint on the total sample size. To this end to each domain U_i a (known) priority weight $\kappa_i > 0$ will be assigned. These weights, describing domain-wise efficiency priorities can be read out e.g. from CRH assignment of the domain-wise relative variance boundary values RV_{oi}, $i = 1,\ldots,I$. That is, for any $i = 1,\ldots,I$, the priority weight κ_i can be taken as $\kappa_i = \frac{RV_{oi}}{RV}$, where $RV = \sum_{i=1}^{I} RV_{oi}$.

Then, (9) can be written as

$$\frac{1}{\tau_i^2}\sum_{h=1}^{H_i}\left(\frac{1}{n_{i,h}} - \frac{1}{N_{i,h}}\right)N_{i,h}^2 S_{i,h}^2 = \kappa_i T, \quad i = 1,\ldots,I, \tag{11}$$

where T is an unknown positive constant. Under (11) the parameter T controls both the relative variances in domains and the overall relative variance S of the estimator of the population mean. To see the latter, note that (10) implies

$$\mathrm{S} = \left(\frac{1}{\tau^2}\sum_{i=1}^{I}\rho_i^2\right)T, \tag{12}$$

where $\rho_i = \tau_i\sqrt{\kappa_i}$, $i = 1,\ldots,I$. Therefore, T will be called the *base* of the relative variance.

To formulate the main result we need to introduce and analyze properties of a population $I \times I$ matrix

$$\mathbf{D} = \frac{1}{n}\underline{a}\,\underline{a}^T - \mathrm{diag}(\underline{c}), \tag{13}$$

where

$$\underline{a} = (a_1,\ldots,a_I)^T = \left(\frac{1}{\rho_i}\sum_{h=1}^{H_i} N_{i,h}S_{i,h},\, i=1,\ldots,I\right)^T, \tag{14}$$

$$\underline{c} = (c_1,\ldots.,c_I)^T = \left(\frac{1}{\rho_i^2}\sum_{h=1}^{H_i} N_{i,h}S_{i,h}^2,\, i=1,\ldots,I\right)^T \tag{15}$$

and $\mathrm{diag}(\underline{c})$ is a diagonal matrix with the vector $\underline{c}$ being its diagonal.

Proposition 2.1 *Assume that*

$$n < \sum_{i=1}^{I} \frac{\left(\sum_{h=1}^{H_i} N_{i,h}S_{i,h}\right)^2}{\sum_{h=1}^{H_i} N_{i,h}S_{i,h}^2}. \tag{16}$$

Then, $\mathbf{D}$ *has the unique, simple and positive eigenvalue* λ^* *and the unique unit eigenvector* $\underline{v}^* \in \mathbb{R}^I$ *associated to* λ^**, which has all coordinates positive.*

The proof of this proposition is given in Section 3.

It appears that the eigenvalue λ^* and the eigenvector $\underline{v}^*$ from Prop. 2.1 are crucial for the multi-domain version of the classical Neyman-Tchuprov allocation, which is the main result of this paper.

Theorem 2.2 *Consider stratified SRSWOR in all domains (as described above) with the total sample size*

$$n = \sum_{i=1}^{I}\sum_{h=1}^{H_i} n_{i,h} \tag{17}$$

and assume that (16) *holds. Let* λ^* *and* $\underline{v}^*$ *be as in Prop. 2.1.*

Then, the multi-domain optimal allocation (with priority weights κ_i*,* $i=1,\ldots,I$*), that is the allocation satisfying* (11) *with the minimal base of relative variance under the sample size constraint* (17) *has the form*

$$n_{i,h} = n\,\frac{v_i^* N_{i,h}S_{i,h}/\rho_i}{\sum_{r=1}^{I} v_r^* \sum_{g=1}^{H_r} N_{r,g}S_{r,g}/\rho_r}, \quad h=1,\ldots,H_i,\ i=1,\ldots,I. \tag{18}$$

For the optimal base of the relative variance T_{opt} *we have* $T_{opt} = \lambda^*$*. Moreover,*

$$T_{opt} = \frac{1}{\sum_{i=1}^{I}\rho_i^2}\left[\frac{1}{n}\left(\sum_{i=1}^{I}\frac{\rho_i}{v_i^*}\sum_{h=1}^{H_i} N_{i,h}S_{i,h}\right)\left(\sum_{i=1}^{I}\frac{v_i^*}{\rho_i}\sum_{h=1}^{H_i} N_{i,h}S_{i,h}\right) - \sum_{i=1}^{I}\sum_{h=1}^{H_i} N_{i,h}S_{i,h}^2\right]. \tag{19}$$

Remark 2.1 *Note that* (18)*, while inserted into* (9)*, implies*

$$T_{i,opt} = \frac{\rho_i}{n\tau_i^2 v_i^*}\sum_{h=1}^{H_i} N_{h,i}S_{h,i}\sum_{r=1}^{I}\frac{v_r^*}{\rho_r}\sum_{g=1}^{H_i} N_{r,g}S_{r,g} - \frac{1}{\tau_i^2}\sum_{h=1}^{H_i} N_{i,h}S_{i,h}^2. \tag{20}$$

The proof of Theorem 2.2 is given in Section 3.

Note that (19) together with (12) implies that, similarly as in the classical Neyman-Tchuprov case, the overall relative variance is of order $1/n$ up to the additive (typically small) constant.

In the boundary case of $I=1$, that is, when there are no domains in U, $\frac{\rho_1}{v_1^*}$ cancels out in (18) and (19). Consequently, these formulas are transformed into the original Neyman-Tchuprov formulas (1) and (2), respectively. Also, (16) becomes (3).

Another boundary case is when there are no strata in domains. Then, from Th. 2.2 we obtain an analytic solution which can be viewed as an alternative to the NLP approach of CRH. In this case (no strata in domains) the matrix $\mathbf{D}$, as defined in (13), has a simple form since then

$$\underline{a}=\left(\tfrac{N_iS_i}{\rho_i},\,i=1,\dots,I\right)^T,\qquad \underline{c}=\left(\tfrac{N_iS_i^2}{\rho_i^2},\,i=1,\dots,I\right)^T.$$

Since $H_i=1$, $i=1,\dots,I$, the inequality (16) is a consequence of the natural assumption $n<N$, where $N=\sum_{i=1}^I N_i$. Let $\underline{v}^*$ be the unique unit eigenvector with positive coordinates for the simplified $\mathbf{D}$ matrix given above (by Prop. 2.1 we know that such vector $\underline{v}^*$ exists).

Corollary 2.3 *In the case of SRSWOR in each of domains (no strata) the optimal domain-wise efficient allocation (with priority weights κ_i, $i=1,\dots,I$) under the sample size constraint*

$$\sum_{i=1}^I n_i = n < N \tag{21}$$

has the form

$$n_i = n\,\frac{v_i^* N_iS_i/\rho_i}{\sum_{j=1}^I v_j^* N_jS_j/\rho_j},\qquad i=1,\dots,I. \tag{22}$$

Then, the optimal base of the relative variance assumes the form

$$T_{opt}=\frac{1}{\sum_{i=1}^I\rho_i^2}\left[\frac{1}{n}\left(\sum_{i=1}^I\frac{\rho_i}{v_i^*}N_iS_i\right)\left(\sum_{i=1}^I\frac{v_i^*}{\rho_i}N_iS_i\right)-\sum_{i=1}^I N_iS_i^2\right]. \tag{23}$$

On the other hand, we may want to minimize the sample size $n=\sum_{i=1}^I\sum_{h=1}^{H_i}n_{i,h}$ under the constraints (9) with given T_i, $i=1,\dots,I$. A straightforward application of the Lagrange multipliers gives the analog of (4) of the form

$$n_{i,h}=N_{i,h}S_{i,h}\,\frac{\sum_{g=1}^{H_i}N_{i,g}S_{i,g}}{\tau_i^2T_i+\sum_{g=1}^{H_i}N_{i,g}S_{i,g}^2},\qquad h=1,\dots,H_i,\ i=1,\dots,I. \tag{24}$$

Therefore,

$$n_{opt}=\sum_{i=1}^I\frac{\left(\sum_{h=1}^{H_i}N_{i,h}S_{i,h}\right)^2}{\tau_i^2T_i+\sum_{h=1}^{H_i}N_{i,h}S_{i,h}^2}. \tag{25}$$

Similarly as in the original Neyman-Tchuprov case the two approaches are dual in the following sense: (18) follows by inserting $T_i := T_{i,opt}$ as given in (20) into (24); dually, we note that (20) (again with $T_i := T_{i,opt}$) can be rewritten as the following relation between elements of the eigenvector $\underline{v}^*$

$$\frac{nv_i^*}{\sum_{j=1}^I v_j^* \sum_{h=1}^{H_i} \frac{N_{j,h}S_{j,h}}{\rho_j}} = \frac{\rho_i \sum_{h=1}^{H_i} N_{h,i}S_{h,i}}{\tau_i^2 T_i + \sum_{h=1}^{H_i} N_{i,h}S_{i,h}^2}, \quad i = 1, \ldots, I$$

and this formula gives (24) when combined with (18).

3. Proofs

The proofs, to some extent, can be read out from Sec. 2 of WW. Nevertheless, to make this article more self-contained we provide most of the arguments referring only to a rather technical Prop. 2.2 from WW. The main new aspect of the argument is related to the formula (19) for the base of relative variances.

Proof of Prop. 2.1 We first refer to Prop. 2.2 of WW, the proof of which was based on the Weyl inequalities (relating eigenvalue of the sum of two matrices to eigenvalues of the summands). Then, see Rem. 2.1 in WW, it follows that there exists a unique, positive eigenvalue of the matrix $\mathbf{D}$, denoted here by λ^*. Moreover, the eigenvalue λ^* is simple, i.e. its eigenspace is one-dimensional.

To show that there exists a unit length eigenvector $\underline{v}^*$ (associated with λ^*) with all coordinates positive 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 v of $\mathbf{A}$, it is simple and such that $v > |\lambda|$ for any other eigenvalue λ of $\mathbf{A}$. The respective eigenvector (attached to v) has all entries strictly positive (up to scalar multiplication)* - see, e.g. Kato (1981), Th. 7.3 in Ch. 1.

Fix a number $\alpha > \max_{1 \le i \le I} c_i > 0$. Note that the matrix $\mathbf{D}_\alpha := \mathbf{D} + \alpha \mathbf{Id}$, where $\mathbf{Id}$ is an $I \times I$ identity matrix, has all entries strictly positive. For any eigenvalue λ of $\mathbf{D}$ and the respective eigenvector $\underline{w}$ we have

$$\mathbf{D}_\alpha \underline{w} = (\lambda + \alpha)\underline{w}, \tag{26}$$

that is, $\mu = \lambda + \alpha$ and $\underline{w}$ are eigenvalue and associated eigenvector of $\mathbf{D}_\alpha$, respectively. By the Perron-Frobenius theorem, there exists an eigenvalue μ^* of $\mathbf{D}_\alpha$ such that $\mu^* > |\lambda + \alpha| > \lambda + \alpha$ for any other eigenvalue $\lambda + \alpha$ of $\mathbf{D}_\alpha$. Moreover, the unit eigenvector $\underline{v}^*$ associated with μ^* has all coordinates positive.

We will show that $\lambda^* = \mu^* - \alpha$. Assume not. Then, there exists an eigenvalue $\mu_0 < \mu^*$ of $\mathbf{D}_\alpha$ such that $\lambda^* = \mu_0 - \alpha$. Thus, $\lambda^* < \mu^* - \alpha = \tilde{\lambda}$, where $\tilde{\lambda}$ is an eigenvalue of $\mathbf{D}$. Since λ^* is the unique positive eigenvalue of $\mathbf{D}$, we obtained a contradiction. Therefore, $\lambda^* = \mu^* - \alpha$ and $\mathbf{D}\underline{v}^* = \lambda \underline{v}^*$.

Consequently, λ^* is the unique simple positive eigenvalue of the matrix $\mathbf{D}$ and the respective eigenspace is spanned by the unit vector $\underline{v}^*$ with all components positive.

Now we are ready to prove the main result.

Proof of Theorm 2.2 With $A_{i,h}=\frac{N_{i,h}S_{i,h}}{\rho_i}$ and c_i's defined in (15), equation (11) can be written as

$$\sum_{h=1}^{H_i}\frac{A_{i,h}^2}{n_{i,h}}-c_i=T,\quad i=1,\ldots,I. \tag{27}$$

Consequently, the Lagrange function for the minimization problem assumes the form

$$F(T,\underline{n})=T+\sum_{i=1}^{I}\mu_i\left(\sum_{h=1}^{H_i}\frac{A_{i,h}^2}{n_{i,h}}-c_i\right)+\mu\sum_{i=1}^{I}\sum_{h=1}^{H_i}n_{i,h}.$$

Upon differentiating with respect to $n_{i,h}$ we obtain

$$\frac{\partial F}{\partial n_{i,h}}=\mu-\mu_i\frac{A_{i,h}^2}{n_{i,h}^2}=0$$

which yields $v_i^2:=\mu_i/\mu>0$ and $n_{i,h}=v_iA_{i,h}$, $h=1,\ldots,H_i$, $i=1,\ldots,I$.

Since $a_i=\sum_{h=1}^{H_i}A_{i,h}$, see (14), the constraint (27) assumes the form

$$a_i-c_iv_i=Tv_i,\quad i=1,\ldots,I. \tag{28}$$

Moreover, (17) yields $\frac{1}{n}\sum_{j=1}^{I}v_ja_j=1$. Therefore, (28) can be written in the form

$$\frac{1}{n}\left(\sum_{j=1}^{I}v_ja_j\right)a_i-c_iv_i=Tv_i,\quad i=1,\ldots,I.$$

Equivalently, $\mathbf{D}\underline{v}=T\underline{v}$ with $\mathbf{D}=\frac{1}{n}\underline{a}\underline{a}^T-\mathrm{diag}(\underline{c})$, and $\underline{v}=(v_1,\ldots,v_I)^T$. That is, $\underline{v}$ which is a vector with positive components, is an eigenvector of $\mathbf{D}$ and T is the eigenvalue associated to $\underline{v}$. According to Prop. 2.1, the unique unit vector $\underline{v}$ satisfying positivity requirement is $\underline{v}=\underline{v}^*$ and then $T=\lambda^*$. Consequently,

$$n_{i,h}\propto A_{i,h}v_i^*,\quad h=1,\ldots,H_i,\ i=1,\ldots,I.$$

Using again the constraint (17) we obtain (18).

On the other hand, we plug $n_{i,h}$, as given in (18), into the formula for the total relative variance (10). Upon cancelations we get (19).

4. Conclusion

The minimization of the common base T of the relative variances in the domains under domain-wise stratified SRSWOR can be achieved analytically through the eigenproblem approach. The formulas for the allocation as well as for the optimal relative variance are explicit in terms of the unique unit eigenvector with positive coordinates of a properly designed population matrix $\mathbf{D}$. Consequently, a direct (but not straightforward) generalization of the classical Neyman-Tchuprov optimal allocation is obtained. Although it has similar drawbacks to those of the Neyman-

Tchuprov allocation, it has its rather unique advantage: it reveals structural properties of the domain-wise optimal allocation. Additionally, in typical situations, the eigenproblem approach gives also numerical solutions which are either identical or close to those obtained through NLP tools. Of course, the NLP procedures allow one to obtain optimal sample strata sizes not exceeding actual strata sizes. The eigenproblem approach may give optimal allocations which do not satisfy such requirements. The proper adjustment of the eigenproblem methodology remains a challenging issue.

REFERENCES

CHOUDHRY, G. H., RAO, J. N. K., HIDIROGLOU, M. A., (2012).On sample allocation for efficient domain estimation, Survey Meth. 38(1) , pp. 23–29.

FRIEDRICH, U., MÜNNICH, R., DE VRIES, S., WAGNER, M., (2015). Fast integer-valued algorithm for optimal allocations under constraints in stratified sampling, Comp. Statist. Data Anal., 92 , pp. 1–12.

FRIEDRICH, U., MÜNNICH, R., RUPP, M., (2018). Multivariate optimal allocation with box-constraints, Austrian J. Statist., 47 , pp. 33–52.

GABLER, S., GANNINGER, M., MÜNNICH, R., (2012). Optimal allocation of the sample size to strata under box constraints, Metrika, 75(2) , pp. 151–161.

KATO, T., (1981), A Short Introduction to Perturbation Theory for Linear Operators, Springer, New York.

KHAN, M. G. M., WESOŁOWSKI, J., (2019). Neyman-type sample allocation for domains-efficient estimation in multistage sampling. Adv. Stat. Anal., 103 , pp. 563–592.

KOZAK, M., (2004). Method of multivariate sample allocation in agricultural surveys. Biom. Collq., 34 , pp. 241–250.

KOZAK, M., ZIELIŃSKI, A., (2005). Sample allocation between domains and strata. Int. J. Appl. Math. Stat, 3 , pp. 19–40.

KOZAK, M., ZIELIŃSKI, A., SINGH, S., (2008). Stratified two-stage sampling in domains: sample allocation between domains, strata and sample stages. Statist. Probab. Lett., 78 , pp. 970–974.

LONGFORD, N. T., (2006). Sample size calculation for small-area estimation. Survey Meth., 32 , pp. 87–96.

NIEMIRO, W., WESOŁOWSKI, J., (2001). Fixed precision allocation in two-stage sampling. Appl. Math., 28 , pp. 73–82.

SÄRNDAL, C.-E., SWENSSON, B., WRETMAN, J., (1992). Model Assisted Survey Sampling, Springer, New York .

VALLIANT, R., DEVER, J.A., KREUTER, F., (2013). Practical Tools for Designing and Weighting Sample Surveys, Springer.

WESOŁOWSKI, J., WIECZORKOWSKI, R., (2017). An eigenproblem approach to optimal equal-precision sample allocation in subpopulations. Comm. Statist. Theory Meth., 46(5) , pp. 2212–2231.

WRIGHT, T., 2017). Exact optimal sample allocation: More efficient than Neyman. Statist. Probab. Lett., 129 (, pp. 50–57.