JIRSS (2021)
Vol. 20, No. 01, pp 307-331
DOI:10.52547/jirss.20.1.307

Sampling of Multiple Variables Based on Partially Ordered Set Theory

Bardia Panahbehagh [1], Rainer Bruggemann [2], and Mohammad Salehi [3]

[1] Department of Mathematics, Faculty of Mathematical Sciences and Computer, Kharazmi University, Tehran, Iran.
[2] Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany.
[3] Department of Mathematics, Statistics and Physics, Qatar University, P. O. Box 2713, Doha, Qatar.

Received: 02/12/2020, Revision received: 15/02/2021, Published online: 03/04/2021

Abstract. We introduce a new method for ranked set sampling with multiple criteria. The method relaxes the restriction of selecting just one individual variable from each ranked set. Under the new method for ranking, units are ranked in sets based on linear extensions in partially order set theory with considering all variables simultaneously. Results will be evaluated by a relatively extensive simulation studies on Bivariate normal distribution and two real case studies on commercial and medicinal use of flowers, and the pollution of herb-layer by Lead, Cadmium, Zinc and Sulfur in some regions of the southwest of Germany.

Keywords. Environmental Pollution, Linear Extension, Medicinal Use of Flowers, Multiple Variables Ranked Set Sampling, Partially Order Set, Theory.

MSC: 62D05, 62P12, 62P10.

Bardia Panahbehagh (panahbehagh@khu.ac.ir)
Rainer Bruggemann (brg_home@web.de)
Corresponding Author: Mohammad Salehi (salehi@qu.edu.qa)

1 Introduction

Ranked set sampling (RSS) design was first introduced by McIntyre (1952) and has been widely used in many applications. The idea behind RSS is appealing particularly to agricultural and environmental scientists where identifying sampling units in the field is straightforward but the exact exploration measurement of the units by measurements is time consuming. Many sampling units can be identified and within them subsets are actually measured. In RSS, the identification of these subsets is based on ranking the units and a selection according to their relative ranks.

The RSS technique briefly involves taking random samples of size m from the population. The sample units are ranked by some quick and easy measure. Then, one unit from each sample is chosen and precisely measured for the character of interest. To take a sample of size m, the unit that has the lowest rank in the first sample (of size m) is chosen, the unit with the second lowest rank is chosen from the second sample, and so on. This process is repeated n times, giving a final sample size, $n_{\cdot} = nm$. Sampling can be balanced or unbalanced where the number of sample units selected in the ranks is not constant. With highly skewed population distributions more units from low (or high) ranks can be selected. Unbalanced designs are similar in concept to the optimal allocation in stratified sampling where strata with bigger variances, take bigger sample fractions. RSS is reported as being more efficient than simple random sampling (Ridout, 2003; Samawi, 1996). See full reviews of RSS by Patil et al.(1999) and Chen et al. (2004).

In this paper, a multivariate RSS based on partially order set thorey (Poset) will be introduced. In some populations there are more than one character of interest, e.g. multiple species in a survey of forest bio-diversity. Patil et al. (1994) have discussed RSS for multiple variables when one of the variables can be defined as a primary variable. Ranking is based on this main primary variable only, and if the other variables are correlated with the main one, the method will perform reasonably well. Norris et al. (1995) have developed two approaches, one using an unbalanced allocation process based on the Neyman allocation for the variable of primary interest, treating this as a concomitant for the other variables of interest and the other using a design based on randomly choosing sample units from the rank list derived from an individual variable. Al-Saleh and Zheng (2002) as well as Chen and Shen (2003) have proposed a two-layer ranked set sampling for the situation in which we have two main variables or two concomitant variables to rank the data. In their methods at the first layer, the data is ranked based on the first variable and a RSS sample is selected. At the second round, the first layer RSS data will be ranked based on the second variable and the RSS data in the second layer will be present as the final sample. One disadvantage of their methods is that they consider the two variables separately, and not simultaneously. Another disadvantage is that they are requiring many initial samples to achieve the needed sample size and also with increasing dimension of the space of variables, the size of the needed sample will increase severely.

We demonstrate and evaluate our suggested sampling technique with two environmental examples:

- The first example deals with the estimation of mean values of "flower dry weight" and "essence" of Matricaria chamomilla in Iran, which is considered as a very important commercial and medicinal plant. The main part of chamomile for medicinal purposes is the flower essence and it is commercially important to maximize the oil yield. It is hardly possible to measure the efficiency of oil yield under all scenarios and all suitable geographical units within Iran. Sampling is therefore necessary and is performed.

- Chemical pollution in the environment is a problem which came into the attention of administration since the early eighties. Chemicals pose a hazard to humans, animals, plants, etc. due to their toxicity. The quantification of the hazard is however extremely difficult as uptake mechanisms, mode of toxic action, the role of chemical speciation and the state of the environmental geographical unit are important. Therefore in almost all nations monitoring programs were installed to observe the chemical pollution spatially and temporarily. The data have mostly the unit mass of chemicals (as total concentration) mass of the target, for example soil.

  These data are thought of as surrogates expressing the hazard potential due to the considered chemicals. It is difficult to obtain for example mean values of concentrations taking into account all geographical units, especially when a temporal trend is to be monitored. Here 59 geographical units are selected by the environmental protection agency taking care of defining the regions as homogeneous as possible with respect to the chemical pollution processes. The sampling technique can be validated, because in that specific case the mean values can also directly obtained from all 59 units for a specified year of observation. When the proposed method is successful then the monitoring process can be simplified, namely to relax the precondition of almost homogeneous geographical units and a more elaborated locally specific monitoring can be applied.

In section 2 we extend the method of Panahbehagh et al. (2018) to multivariate case. In section 3 we introduce stratified sampling using RSS derived from linear extensions (LE) in Poset. Section 4 contains examples, simulations and two real case study to compare the methods and evaluate the results and the paper will be wrapped up in section 5 with a conclusion.

2 Multivariate Virtual Stratified Ranked Set Sampling (MVSR)

In multivariate RSS, we have an R dimensional random variable. We start with the basic idea of multivariate RSS (Patil et al., 1994), ranking according to just one of the variables. Then we adapt the design with an unbalanced ranked set sampling to get more than one sample from each set.

Suppose that $\mathbf{X} \sim f_{\boldsymbol{\mu}}$ with $E(\mathbf{X}) = \boldsymbol{\mu}$, where $\mathbf{X} = (X^1, X^2, ..., X^R)$ and $\boldsymbol{\mu} = (\mu^1, \mu^2, ..., \mu^R)$ also $Var(X^j) = \sigma^2_j$, $Cov(X^j, X^{j'}) = \rho_{jj'}\sigma_j\sigma_{j'}$ and $E(|X^j|^2) < \infty$ for all j. The main aim is to estimate $\boldsymbol{\mu}$. Our strategy to get a sample of size $n_. = nm$ from the population is to generate an *iid* sample of $\mathbf{X}_i$s of size m from f and sort them according to X^1 (using itself or based on an auxiliary variable) in m columns and repeat this method K times. Then we will have a stratified population, formed in m strata, each of size K (see Table 1), where $\mathbf{X}_{(h)i} = (X^1_{(h)i}, X^2_{[h]i}, ..., X^R_{[h]i})$, and $X^1_{(h)i}$ is the h^{th} order statistics in the i^{th} set with $\mu^1_{(h)}$ and $\sigma^2_{(h)1}$ as the mean and variance respectively, and $X^j_{[h]i}$ for $j = 2, 3, .., R$ are concomitant variables with respect to $X^1_{(h)i}$ in the i^{th} set with $\mu^j_{[h]}$and $\sigma^2_{[h]j}$. Now we get a Simple Random Sampling Without Replacement (SRSWOR) from the h^{th} stratum of size n (an integer smaller than K), say s_h, and we can estimate μ^j by

$$\begin{aligned}
\widehat{\mu}^1_V &= \frac{1}{m}\sum_{h=1}^{m} \bar{X}^1_{(h)}, \\
\widehat{\mu}^j_V &= \frac{1}{m}\sum_{h=1}^{m} \bar{X}^j_{[h]},
\end{aligned}$$

where

$$\begin{aligned}
\bar{X}^1_{(h)} &= \frac{1}{n}\sum_{i\epsilon s_h} X^1_{(h)i}, \\
\bar{X}^j_{[h]} &= \frac{1}{n}\sum_{i\epsilon s_h} X^j_{[h]i},
\end{aligned}$$

Table 1: Virtual strata, using conventional RSS.

1^{st} stratum	2^{nd} stratum	$\cdots$	m^{th} stratum
$\mathbf{X}_{(1)1}$	$\mathbf{X}_{(2)1}$	$\cdots$	$\mathbf{X}_{(m)1}$
$\mathbf{X}_{(1)2}$	$\mathbf{X}_{(2)2}$	$\cdots$	$\mathbf{X}_{(m)2}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathbf{X}_{(1)K}$	$\mathbf{X}_{(2)K}$	$\cdots$	$\mathbf{X}_{(m)K}$

Theorem 1. *In MVSR, $\widehat{\mu}^j_V$ is an unbiased estimator for μ^j and*

$$\begin{aligned} V(\widehat{\mu}^1_V) &= \frac{1}{nm}(\sigma_1^2 - \frac{(1-\frac{n}{K})}{m}\sum_{h=1}^{m}(\mu^1_{(h)} - \mu^1)^2), \\ V(\widehat{\mu}^j_V) &= \frac{1}{nm}(\sigma_j^2 - \frac{(1-\frac{n}{K})}{m}\sum_{h=1}^{m}(\mu^j_{[h]} - \mu^j)^2), \end{aligned}$$

and if we assume that X^1 and X^j are linked with the below linear regression model

$$X_i^j = \mu^j + \rho_{1j}\frac{\sigma_j}{\sigma_1}(X_i^1 - \mu^1) + \varepsilon_i, \tag{2.1}$$

where ε is a random variable independent from X^1, then

$$V(\widehat{\mu}^j_V) = \frac{1}{nm}(\sigma_j^2 - \frac{(1-\frac{n}{K})}{m}\rho_{1j}^2\sum_{h=1}^{m}(\mu^j_{(h)} - \mu^j)^2),$$

and

$$\begin{aligned} \widehat{V}(\widehat{\mu}^1_V) &= \frac{K-1}{m(mK-1)}\sum_{h=1}^{m}\frac{1}{n(n-1)}\sum_{i\epsilon s_h}(X^1_{(h)i} - \bar{X}^1_{(h)})^2 + \frac{1}{m(mK-1)}\sum_{h=1}^{m}(\bar{X}^1_{(h)} - \widehat{\mu}^1_V)^2, \\ \widehat{V}(\widehat{\mu}^j_V) &= \frac{K-1}{m(mK-1)}\sum_{h=1}^{m}\frac{1}{n(n-1)}\sum_{i\epsilon s_h}(X^j_{[h]i} - \bar{X}^j_{[h]})^2 + \frac{1}{m(mK-1)}\sum_{h=1}^{m}(\bar{X}^j_{[h]} - \widehat{\mu}^j_V)^2, \end{aligned}$$

are unbiased estimators for the variance of variables.

The proof of Theorem 1, see Appendix A.

In MVSR one variable is selected as a leading variable to perform a ranking, the other variables are just adjusted which implies some errors. We, therefore, introduce a method of ranking that ranks all variables simultaneously.

3 Ranking Based on Posets

In this section, we first describe Posets theory and then introduce two new versions of multivariate RSS, based on them.

3.1 Posets and Linear Extensions

The application of Poset for ranking has been described by Bruggemann and Patil (2011). In this theory, we have a set containing m units having R variables each, with a binary relation between the units. To compare two units of the set, if all variables of the first unit are equal or bigger (smaller) than the second one, then the first unit is better (≥) (worse (<)) than the second one, otherwise the two units are not comparable. Linear extensions (LEs) are different projections of the partial order into a complete order that respect all the relations in the Poset. I.e. Linear extensions are the result of order preserving mappings. Therefore a relation $x < y$ in a the Poset is preserved in all linear extensions. Also mean height for each unit is defined as the average of ranks of the respective unit in all the LEs.

As an example, assume we have a set with $m = 5$ and $R = 2$ (see Table 2). Then according to the quantities of the variables, it is possible to construct eight LEs as is shown in Table 3. Also mean heights are calculated in Table 4. Here, due to the low number of linear extensions, the mean height of each unit can be easily directly determined from Table 3. Generally, the determination of all linear extensions is computationally cumbersome. However, it is not necessary to determine the set of LEs explicitly, because only the average height is of interest that can be estimated by a sampling technique (Bubley and Dyer, 1999). In this case, there are also pretty good approximations available, see for instance Bruggemann et al. (2004), (2013) or De Loof et al. (2013).

Now we use the Poset theory to introduce two designs; Ranking based on Posets using complete form (or at least a random sample) of LEs (CPOR) and Ranking based on Posets using just one random selection of LEs (RPOR):

- CPOR: First rank the units according to the mean height of the units and then construct an unequal size population using these mean heights based on the complete LEs.
- RPOR: Select one of the LEs to construct an equal size population.

Table 2: units of a set with their variables.

	X^1	X^2
a	0	1
b	2	1
c	1	2
d	3	3
e	0	4

Table 3: All possible LEs with respect to Posets.

LE1	LE2	LE3	LE4	LE5	LE6	LE7	LE8
d	d	d	e	d	d	d	e
c	c	e	d	b	b	e	d
b	e	c	c	c	e	b	b
e	b	b	b	e	c	c	c
a	a	a	a	a	a	a	a

3.2 CPOR

Here we are going to put each unit of a set into a stratum equal to the nearest integer of its mean height. Following the previous example according to Table 4, we will put the units of the set into 5 virtual strata (see Table 5).

Then, the design proceeds as follows: an *iid* sample of size m (a set) from f_μ will be generated, and according to their variables ($X^j s$) all possible linear extensions will be constructed. We then calculate the mean height either explicitly by determination of the set of all LEs or directly by applying approximations. Finally, using these heights, put the units of the set into the strata and repeat this approach K times. It is obvious that this method can lead to an unequal size stratified population.

Table 4: The Mean height of each unit in all possible LEs.

	mean height	rounded height
a	1	1
b	2.875	3
c	2.875	3
d	4.75	5
e	3.5	4

Table 5: Putting the units of a set in strata.

strata	1	2	3	4	5
	a		b c	e	d

Then instead of an R dimensional variable $\mathbf{X}_{\{h\}i} = (X^1_{\{h\}i}, X^2_{\{h\}i}, ..., X^R_{\{h\}i})$ we have a R+1 dimensional variable $\mathbf{X}_{\{h\}i} = (X^1_{\{h\}i}, X^2_{\{h\}i}, ..., X^R_{\{h\}i}, \mathrm{MH}_{\{h\}i})$ where MH stands for the mean heights of the objects.

We now have a stratified population with unequal size. For the h^{th} stratum we will take a SRSWOR, s_h, of size n_h, proportional to the stratum size, K_h, where $\sum_{h=1}^{m} K_h = Km$ such that $\sum_{h=1}^{m} n_h = n_. = nm$. The stratified population is presented in Table 6.

In Table 6, $\mathbf{X}_{\{h\}i} = (X^1_{\{h\}i}, X^2_{\{h\}i}, ..., X^R_{\{h\}i})$ where $X^j_{\{h\}i}$ is the j^{th} character of an unit that has been fallen into the h^{th} stratum after $i-1$ units, according to its mean height in respective LEs. Now we propose an estimator for μ^j (the expectation of the j^{th} character in f_μ) as

$$\widehat{\mu}^j_P = \sum_{h=1}^{m} W_h \bar{X}^j_{\{h\}},$$

where

$$W_h = \frac{K_h}{Km}, \tag{3.1}$$

and

$$\bar{X}^j_{\{h\}} = \frac{1}{n_h} \sum_{i \epsilon s_h} X^j_{\{h\}i},$$

Table 6: Virtual strata, using Posets ranking.

1^{st} stratum	2^{nd} stratum	$\cdots$	m^{th} stratum
$\mathbf{X}_{\{1\}1}$	$\mathbf{X}_{\{2\}1}$	$\cdots$	$\mathbf{X}_{\{m\}1}$
$\mathbf{X}_{\{1\}2}$	$\mathbf{X}_{\{2\}2}$	$\cdots$	$\mathbf{X}_{\{m\}2}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\vdots$	$\vdots$	$\cdots$	$\mathbf{X}_{\{m\}K_m}$
$\mathbf{X}_{\{1\}K_1}$	$\vdots$		
	$\mathbf{X}_{\{2\}K_2}$		

Theorem 2. *In CPOR, $\widehat{\mu}^j_P$ is an unbiased estimator for μ^j.*

For the proof of Theorem 2 is given in Appendix B.

Here instead of Neyman allocation, the proportional to size allocation is used that is easy to implement and does not need extra information (Sarndal et al. 1992).

3.3 RPOR

RPOR is easier than CPOR to perform. Here it is just enough to select (or construct) one of the LEs in Table 3 randomly and put them in 5 strata and then we will have a stratified population, formed in m strata, each of size K like MVSR (see Table 1). Here we show the vector of the i^{th} variable in the h^{th} stratum with $\mathbf{X}_{[h\}i} = (X^1_{[h\}i}, X^2_{[h\}i}, ..., X^R_{[h\}i})$. Now we get an SRSWOR from the h^{th} stratum of size n (an integer smaller than K), say s_h. Now we propose an estimator for μ^j as

$$\widehat{\mu}^j_R = \frac{1}{m} \sum_{h=1}^{m} \bar{X}^j_{[h\}},$$

where

$$\bar{X}^j_{[h\}} = \frac{1}{n}\sum_{i\in s_h} X^j_{[h\}i}.$$

Theorem 3. *In RPOR, $\widehat{\mu}^j_R$ is an unbiased estimator for μ^j with variance*

$$V(\widehat{\mu}^j_R) = \frac{\sigma_j^2}{Km} + \frac{1}{m^2}\sum_{h=1}^{m}\frac{1-\frac{n}{K}}{n}E_M\left(\frac{1}{Q}\sum_{q=1}^{Q} S^2_{[h\}qjK}\right). \tag{3.2}$$

where $q = 1, 2, ..., Q$ are all the possible combinations of LEs, with the below unbiased estimator of variance

$$\widehat{V}(\widehat{\mu}^j_R) = \frac{1}{nm(Km-1)}\left[\sum_{h=1}^{m}\sum_{i\epsilon s_{[h\}}}(X^j_{[h\}i}-\widehat{\mu}^j_R)^2 + (K-n)\sum_{h=1}^{m} s^2_{[h\}j}\right].$$

where $S^2_{[h\}qjK}$ and $s^2_{[h\}j}$ are variances of h^{th} stratum under q^{th} combination of LEs and sample variance of h^{th} stratum for j^{th} variable respectively.

For the proof of Theorem 3 see Appendix C.

3.4 Negative Correlation

When the correlation between variables is strongly negative, according to Posets theory, it is probable that most of the units in a set are incomparable. This can make it meaningless to stratify the sets (note that in this case most of the units will fall in the middle stratum).

An extreme case is when the correlation between two variables is "-1". All the generated units will be incomparable and in the LEs the mean height of all of them will be the same and all will fall in the same stratum. The weight of the stratum (equation (3.1)) will be 1 and the other strata zero. Finally we will take an SRSWOR of size $n_. = mn$ from the stratum and the design will essentially become simple random sampling with replacement.

To overcome this problem, we suggest that if the bivariate correlations between two variables are negative, multiple one of them by "-1" to change the correlations to positive. But if we have more than two variables, sometimes it would not possible to make all the correlations positive. In such cases, it is better to select some more important variables that we are able to make their correlations positive. We then rank the units using Posets theory with this new correlations.

Bruggemann and Patil (2011) explained a procedure, how subsets of variables can be systematically found. The crucial concept is the number of incomparabilities of a Poset. First a sensitivity measure for each variable is to be defined. The sensitivity measures the impact of each variable on the structure of the Poset (roughly: the system of comparabilities within a Poset). Secondly the variables are ordered due to their impact on a Poset. Thirdly considering first the Poset due to the most sensitive variable, then the Poset due to the first two most important variables and so on. The number of incomparabilities is calculated as a function of the merged variables. The resulting curve motivates to find subsets of variables, which constitute mainly the Poset. The remaining variables are considered as fine-tuning, and will be ignored.

4 Simulation Study

To evaluate and compare the efficiency of the designs, we calculate

$$\text{Efficiency}(\widehat{\mu}_{.}) = \frac{V(\bar{y})}{\text{MSE}(\widehat{\mu}_{.})}.$$

where $\bar{y}$ is the sample mean of a simple random sample, and $\widehat{\mu}_{.}$ stands for $\widehat{\mu}_V$ (MVSR design), $\widehat{\mu}_P$ (CPOR design) or $\widehat{\mu}_R$ (RPOR design) and MSE indicate mean square error.

This section contains 3 parts:

- Comparing CPOR and RPOR with MVSR using some simulations
- Comparing CPOR and RPOR with MVSR using a real case study on medical flowers
- Comparing CPOR and RPOR with MVSR using a real case study on environmental pollution.

Also in the simulations, no matter how small were the size and the variance of a particular stratum, at least one sample is allocated to the stratum. All the simulations are done by "R 3.1.2" software. For the Monte Carlo simulation we have used 20000 iterations. Expectations, variances and MSEs of the estimators are computed using Mote Carlo method.

4.1 Comparing CPOR and RPOR with MVSR Using some Simulations

We will investigate efficiency of the introduced design in section 2 and 3, using bivariate normal distribution (with solving negative correlation problem).

4.1.1 Bivariate Normal Distribution with Negative Correlation

Here we performed the simulation assuming normal distribution with negative correlation, with $n = 4, K = 8$ and $m = 3$. As we can see in Table 7, and we asserted in Section 3.4, when the correlation is strongly negative, CPOR and RPOR decline to simple random sampling (efficiency$\simeq 1$). When we convert the correlation to a positive value by changing the sign of one variable, the efficiency is considerably increased(compare the results in the last two columns with the results in the first two columns).

Table 7: Efficiency of the estimators in bivariate normal case $(X^1, X^2) \sim B.N(0, 0, 1, 1, \rho)$ with solving problem of negative correlation.

	ρ=-0.9	ρ=-0.5	ρ=0	ρ=-0.5→0.5	ρ=-0.9→0.9
$\widehat{\mu}_V^1$	1.32	1.32	1.33	1.30	1.31
$\widehat{\mu}_P^1$	1.02	1.02	1.07	1.21	1.31
$\widehat{\mu}_R^1$	0.99	1.02	1.04	1.16	1.28
$\widehat{\mu}_V^2$	1.26	1.08	1.01	1.10	1.25
$\widehat{\mu}_P^2$	1.02	1.00	1.06	1.22	1.31
$\widehat{\mu}_R^2$	0.99	1.02	1.04	1.16	1.28

4.1.2 Bivariate Normal

More complete simulations for bivariate Normal distribution are shown in Table 8. For all the cases we simulated bivariate normal with $\mu^1 = 0$, $\mu^2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 1$ and $\rho = 0.3, 0.5, 0.7, 0.9$.

First note that changing ρ, does not affect the efficiency of the first variable which is confirmed by simulations with less than 0.02 error. As a general point, CPOR and RPOR designs increase the efficiency of the estimator for both variables, simultaneously, whereas the traditional multivariate ranked set sampling just enhances estimation of one of the variables. As the correlations increase, efficiency increase. Unlike MVSR, CPOR and RPOR had good and reasonable efficiency for all the correlations. Also CPOR that uses all information of LEs was more efficient than RPOR.

Table 8: Efficiency of the estimators for different cases for bivariate normal distribution.

m	K	n	ρ	variable	$\widehat{\mu}_V$	$\widehat{\mu}_P$	$\widehat{\mu}_R$
3	12	4	0.3	X^1	1.49	1.16	1.12
				X^2	1.00	1.12	1.11
			0.5	X^1	1.45	1.20	1.17
				X^2	1.05	1.21	1.18
			0.7	X^1	1.47	1.27	1.26
				X^2	1.17	1.30	1.26
			0.9	X^1	1.49	1.41	1.39
				X^2	1.35	1.42	1.41
		6	0.3	X^1	1.31	1.13	1.10
				X^2	1.01	1.10	1.07
			0.5	X^1	1.30	1.16	1.13
				X^2	1.05	1.14	1.11
			0.7	X^1	1.33	1.23	1.19
				X^2	1.13	1.23	1.20
			0.9	X^1	1.32	1.31	1.29
				X^2	1.23	1.31	1.27

4.2 Comparing CPOR and RPOR with MVSR Using a Real Case Study on Medical Flowers

To evaluate the designs in this section we used a real case study data on chamomile flower (Panahbehagh et al. 2018) as medicinal use of flowers. We consider the population mean of the "Flower dry weight" (Fdw) and "Essence" (Esn) as the two main parameters. Because we have no information about them before sampling, and it is expensive to measure them, we used two auxiliary variables, easy to measure with reasonable correlation with the two main variables. For sorting Fdw, we used "Flower height" (Fht) with the correlation of 0.78 and for Esn we used "Number of petals" (Npt) with the correlation of 0.71. The correlation between Fht and Npt is 0.77. Simulation results are in Table 9. As we can see in Table 9, CPOR and RPOR enhance efficiency of both of the estimators simultaneously. The most important factor in efficiency is the portion of K/n and efficiency increased with increasing this factor. For example compare two cases: one $m = 5, K = 7, n = 3$ and two $m = 5, K = 7, n = 5$, although n is larger in the second case, because the portion of K/n is larger for first one, the efficiency of the first case is larger than the second case. Also if the other parameters are equal, m is the other important parameter that affect efficiency and efficiency increased with increasing m. Again CPOR was more efficient than RPOR in almost all the cases.

Table 9: Efficiency of the estimators for estimating the means of Fdw and Esn as the main variables and Fht and Npl as the auxiliary variables with 0.78 and 0.71 correlations.

			$\widehat{\mu}^1_V$	$\widehat{\mu}^2_V$	$\widehat{\mu}^1_P$	$\widehat{\mu}^2_P$	$\widehat{\mu}^1_R$	$\widehat{\mu}^2_R$
K	m	n	Fdw	Esn	Fdw	Esn	Fdw	Esn
5	3	2	1.40	1.11	1.32	1.17	1.28	1.14
		3	1.23	1.07	1.18	1.06	1.18	1.07
		4	1.10	1.04	1.09	1.04	1.09	1.05
	5	2	1.63	1.18	1.45	1.24	1.45	1.23
		3	1.35	1.10	1.26	1.11	1.24	1.13
		4	1.14	1.05	1.11	1.06	1.11	1.06
	7	2	1.77	1.19	1.55	1.27	1.53	1.27
		3	1.40	1.10	1.29	1.12	1.30	1.14
		4	1.17	1.06	1.13	1.07	1.13	1.07
7	3	3	1.36	1.09	1.26	1.15	1.25	1.13
		5	1.15	1.06	1.13	1.07	1.12	1.08
		6	1.09	1.03	1.07	1.03	1.06	1.03
	5	3	1.58	1.16	1.43	1.22	1.40	1.20
		5	1.23	1.07	1.18	1.08	1.17	1.09
		6	1.10	1.03	1.08	1.04	1.08	1.04
	7	3	1.71	1.19	1.53	1.25	1.51	1.24
		5	1.26	1.08	1.22	1.09	1.20	1.11
		6	1.12	1.04	1.09	1.05	1.09	1.05
Average			1.32	1.09	1.24	1.12	1.23	1.12

4.3 Chemical Pollution

The Environmental Protection Agency (EPA) of the German state Baden-Wuerttemberg performed a series of measurements in different targets, for example in the herb layer, in the epiphytic mosses of trees, in fish etc. For this purpose the state Baden Wuerttemberg was divided in 60 more or less homogeneous regions with respect to their natural environment. The regions are not selected according to administrative classification but to get regions as homogeneous as possible with respect to environmental pollution processes.

The task was and is, to protocol the pollution due to industry, traffic, agrarian management with respect to the total concentrations of Lead (Pb), Cadmium (Ca), Zinc (Zn) and Sulfur (S), measured in mg/kg dry mass.

According to the different emission types there are different chemical species, for example SO_2 or solved in atmospheric droplets H_2SO_3, similarly the other metals as for example Pb, which can be bounded in organic chemicals or as oxides.

The different targets, selected by the EPA should help to differentiate among the different transport processes and to be able to trace back the emission source. So, the herb layer is mainly a short range transport indicator, whereas the epiphytic mosses (simply: moss layer) is considered as indicating middle range transports. The herb layer should especially indicate the loading due to the public traffic whereas the moss layer may mainly indicate industrial sources.

An interesting point of geochemical research is as to how far the presence of e.g. Pb implies the presence of Cadmium. A classification approach concerning the pollution of Baden-Wuerttemberg was published by Bruggemann et al. (2013).

4.3.1 Comparing CPOR and RPOR with MVSR Using a Real Case Study on Environmental Pollution

In this study, regions in Baden-Wuerttemberg, South-West of Germany were selected and monitored with respect to total concentrations of the chemical elements Pb, Cd, Zn and S in the herb layer (Environmental Protection Agency Baden-Wurttemberg (Germany) 1994, Signale aus der Natur). The herb layer is one of the targets, selected by the Environmental Protection Agency of Baden-Wuerttemberg. This multi-indicator system with regions as objects and concentrations of the four chemical elements as indicators (Bruggemann and Patil 2011) raises the questions:

- How can we get information about the pollution status?

- What can be said about geochemical relations?

For example does an increase in pollution with respect to one pollutant, for example Pb, always imply the increase of another pollutant, for instance Cd? For an answer from the point of view of applied partial order theory, see Bruggemann and Voigt (2012) (For more details see Bruggemann et al., 1996; Bruggemann et al., 1998; Bruggemann et al., 1999; Bruggemann et al., 2003 and Bruggemann et al., 2013).

Here we examine our methods on this data to see how efficient they are on such situation. For this purpose it is better to give all the correlations a positive value (see subsection 3.4). Then we multiple a "-1" to Cd and Zn. In this part we run two different scenarios:

- Scenario I. Selecting Pb and Zn as the two main variables with high correlation (0.60) and Cd and S as the two main variables with low correlation (0.06). In this scenario we used perfect ranking, and we didn't use auxiliary variables.
- Scenario II. From a chemical point of view, we select Cd and Pb as the two main variables and for sorting them using two auxiliary variables; Zn with 0.48 correlation with Cd and S with 0.27 with Pb. This is a heuristic approach. Basically economical or sociological information or the density of highways could also serve as auxiliary variables.

Results are shown in Table 10 (Scenario I) and Table 11 (Scenario II). In Table 10, the efficiency of estimators for estimating the means of Pb and Zn with correlation of 0.6, and the means of Cd and S with correlation of 0.06 are presented. For two variables with a reasonable correlation (Zn and Pb) MVSR is not bad, because the ranking is just based on the first variable, and the first variable supports the second variable.

For Cd and S in MVSR, because of the weak correlation of about 0.06, the first variable is not able to support the second one. Efficiencies for S in MVSR are around 1. But for CPOR and RPOR results for the second variable are better. With decreasing efficiency of the first variable (Cd) from MVRS to CPOR and RPOR, the efficiency of the second variable (S) raises, reasonably. The average of efficiencies for S in MVSR is around 1.01 but that for CPOR and RPOR are around 1.09. Again, K/n is the most important parameter in efficiency and the next one is m.

In Table 11, we have used two auxiliary variables to rank the main variables. We have used Cd for Zn with correlation of 0.48 and we have used S for Pb with correlation of 0.27. As we can see, MVRS just improves efficiency of the first variable (Cd) and CPOR and RPOR improve the both variables estimations. However, the improvement is not so large because of almost week correlations between auxiliary variables and the main variables (0.48 and 0.27).

By our sampling technique the mean values referring to a complete set of 59 geographical units are obtained. The regional relation is not taken into account but there is now a number available which can characterize the status of Baden-Württemberg overall. For example, a time series analysis could be carried out to see the general changes with respect to the pollution.

Table 10: Efficiency of the estimators for estimating the means of Pb and Zn with 0.6 correlation and the means of Cd and S with 0.06 correlation. Here we used complete ranking.

			$\widehat{\mu}^1_V$	$\widehat{\mu}^2_V$	$\widehat{\mu}^1_P$	$\widehat{\mu}^2_P$	$\widehat{\mu}^1_R$	$\widehat{\mu}^2_R$	$\widehat{\mu}^1_V$	$\widehat{\mu}^2_V$	$\widehat{\mu}^1_P$	$\widehat{\mu}^2_P$	$\widehat{\mu}^1_R$	$\widehat{\mu}^2_R$
m	K	n	Pb	Zn	Pb	Zn	Pb	Zn	Cd	S	Cd	S	Cd	S
3	5	2	1.32	1.11	1.13	1.21	1.12	1.15	1.36	1.01	1.13	1.09	1.12	1.09
		4	1.11	1.02	1.06	1.03	1.05	1.03	1.11	1.01	1.05	1.00	1.05	1.03
	7	2	1.41	1.11	1.15	1.16	1.10	1.12	1.41	0.99	1.16	1.08	1.11	1.07
		4	1.25	1.08	1.11	1.10	1.09	1.10	1.16	1.01	1.00	1.01	1.03	1.06
	10	2	1.59	1.16	1.24	1.25	1.16	1.17	1.53	1.04	1.22	1.12	1.19	1.11
		4	1.38	1.11	1.15	1.17	1.12	1.14	1.29	1.02	1.16	1.09	1.07	1.07
5	5	2	1.61	1.21	1.23	1.27	1.18	1.23	1.53	1.02	1.19	1.12	1.15	1.10
		4	1.15	1.04	1.05	1.05	1.04	1.06	1.13	1.01	1.05	1.02	1.04	1.02
	7	2	1.69	1.21	1.23	1.30	1.21	1.26	1.62	1.00	1.23	1.11	1.18	1.08
		4	1.35	1.14	1.14	1.15	1.13	1.16	1.33	1.00	1.12	1.04	1.09	1.04
	10	2	1.93	1.28	1.31	1.37	1.24	1.34	1.78	0.99	1.28	1.13	1.21	1.12
		4	1.56	1.21	1.20	1.28	1.17	1.26	1.47	1.03	1.15	1.12	1.13	1.11
7	5	2	1.69	1.21	1.26	1.31	1.21	1.28	1.63	1.04	1.26	1.18	1.21	1.19
		4	1.16	1.10	1.09	1.10	1.07	1.12	1.15	0.99	1.06	1.04	1.06	1.01
	7	2	1.90	1.28	1.27	1.35	1.29	1.32	1.85	1.03	1.30	1.12	1.28	1.09
		4	1.45	1.19	1.19	1.19	1.17	1.20	1.37	1.02	1.17	1.07	1.17	1.07
	10	2	2.20	1.36	1.40	1.49	1.39	1.49	1.99	1.02	1.40	1.19	1.31	1.15
		4	1.66	1.30	1.27	1.36	1.25	1.34	1.62	1.02	1.25	1.13	1.24	1.12
	Average		1.52	1.17	1.19	1.23	1.17	1.21	1.46	1.01	1.18	1.09	1.15	1.09

5 Conclusion

CPOR and RPOR can be used to implement RSS in population surveys where there are multiple variables of interest. CPOR and RPOR enhance the parameters estimation simultaneously with a reasonable sample size, that most of the RSS methods can not do in multiple variables cases. As we see in the real case studies, for CPOR and ROPR there is no need to use perfect ranking using the main variables and it can be done using some variables, easy to measure, having a reasonable correlation with the main variables. The simulation section and real case studies confirmed the assertions in the paper.

For further works, it would be beneficial to find some unbiased estimators for variance of CPOR. Because of randomness of K_h, it is not easy to calculate the variance and an unbiased estimator of variance for CPOR. Since CPOR uses information of all LEs and CPOR was more efficient than RPOR in almost all the cases of the simulation study we suggest to use variance estimator of RPOR as a conservative estimate for variance of CPOR.

Table 11: Efficiency of the estimators for estimating the means of Cd and Pb as the main variables and Zn and S as the auxiliary variables with 0.48 and 0.27 correlations.

			$\widehat{\mu}_V^1$	$\widehat{\mu}_V^2$	$\widehat{\mu}_P^1$	$\widehat{\mu}_P^2$	$\widehat{\mu}_R^1$	$\widehat{\mu}_R^2$
m	K	n	Cd	Pb	Cd	Pb	Cd	Pb
3	5	2	1.31	1.01	1.07	1.03	1.02	1.02
		4	1.08	1.01	0.99	0.98	1.00	1.01
	7	2	1.40	0.99	1.02	1.01	1.04	1.03
		4	1.20	1.00	1.00	1.00	1.02	1.01
	10	2	1.46	0.99	1.04	1.02	1.03	1.03
		4	1.31	0.99	1.02	1.03	1.01	1.01
5	5	2	1.48	1.00	1.07	1.05	1.04	1.05
		4	1.13	1.00	1.00	1.00	1.01	1.01
	7	2	1.63	1.00	1.03	1.05	1.05	1.04
		4	1.31	1.00	1.01	1.04	1.03	1.04
	10	2	1.82	1.01	1.06	1.09	1.07	1.07
		4	1.49	1.00	1.07	1.06	1.04	1.04
7	5	2	1.62	1.02	1.07	1.10	1.06	1.08
		4	1.14	1.00	1.01	1.02	1.02	1.02
	7	2	1.85	1.01	1.07	1.07	1.08	1.09
		4	1.37	1.01	1.03	1.04	1.05	1.05
	10	2	2.05	1.01	1.09	1.08	1.08	1.07
		4	1.61	1.02	1.07	1.10	1.06	1.07
	Average		1.46	1.00	1.04	1.04	1.04	1.04

Acknowledgements

The authors would like to thank the editor and anonymous reviewers for their valuable comments and suggestions.

References

Al-Saleh, M., and Zheng, G. (2002), Estimation of bivariate characterstics using ranked set sampling. *Australian & New Zealand Jurnal of Statistics*, **44**, 221–232.

Bruggemann, R., and Carlsen, L. (2011), An Improved Estimation of Averaged Ranks of Partial Orders. *MATCH Comm.Math.Comput.Chem.* **65**, 383–414.

Brüggemann, R., Kaune, A., and Voigt. K. (1996), Vergleichende ökologische Bewertung von Regionen in Baden- Württemberg. Pages 455-467 in Landesanstalt für Umweltschutz Baden-Württemberg, ed. 4.Statuskolloquium, Projekt "Angewandte Ökologie" Nr. 16. Präzis-Druck Karlsruhe, Karlsruhe.

Bruggemann, R., Mucha, H. J., and Bartel, H.G. (2013), Ranking of Polluted Regions in South West Germany Based on a Multi-indicator System. *MATCH Commun. Math. Comput. Chem.*, **69**,433–462.

Bruggemann, R., and Patil, G. P. (2011), *Ranking and Prioritization with Multi- Indicator Systems, Introduction to Partial Order and Its Applications*, Springer, New York.

Bruggemann, R., Pudenz S., Voigt K., Kaune A., and Kreimes K. (1999), An algebraic/graphical tool to compare ecosystems with respect to their pollution. IV: Comparative regional analysis by Boolean arithmetic. *Chemosphere*, **38**, 2263–2279.

Bruggemann, R., Sorensen, P. B., Lerche, D., and Carlsen, L. (2004), Estimation of Averaged Ranks by a Local Partial Order Model. *J. Chem. Inf. Comp. Sc.* **44**, 618–625.

Bruggemann, R., Voigt, K., Kaune, A., Pudenz, S., Komossa, D., and Friedrich, J. (1998), Vergleichende ökologische Bewertung von Regionen in Baden- Württemberg GSF-Bericht 20/98. GSF, Neuherberg.

Bruggemann, R., Welzl, G., and Voigt, K. (2003), Order Theoretical Tools for the Evaluation of Complex Regional Pollution Patterns. *J. Chem. Inf. Comp. Sc.*, **43**, 1771–1779.

Bubley, R., and Dyer, M. (1999), Faster random generation of linear extensions. *Discr. Math.*, **201**, 81–88.

Chen, Z., Bai, Z., and Sinha, B. (2004), *Ranked set sampling: theory and applications. Lecture Notes in Statistics*, Springer, New York.

Chen, Z., and Shen, L. (2003), Two-layer ranked set sampling with concomitant variables. *Journal of Statistical Planning and Inference*, **115**, 45–57.

David, H. A., and Nagaraja, H. N. (2003), *Order Statistic*, third ed. Wiley, New York.

De Loof, K., De Baets, B., and De Meyer, H. (2011), Approximation of Average Ranks in Posets. *MATCH Commun. Math. Comput. Chem.*, **66**, 219–229.

Environmental Protection Agency Baden-Wurttemberg, (1994), Signale aus der Natur 10 Jahre Okologisches Wirkungskataster Baden-Wurttemberg. Kraft Druck GmbH, Ettlingen.

McIntyre, G. A. (1952), A method of unbiased selective sampling. using ranked sets. *Australian Journal of Agricultural Research*, **3**, 385-390.

Norris, R. C., Patil, G. P., and Sinha, A. K. (1995), Estimation of multiple characteristics by ranked set sampling methods. *Coenoses*, **10**, 95–111.

Panahbehagh, B., Bruggemann R., Parvardeh, A., Salehi, M., and Sabzalian, M. R. (2018), An unbalanced ranked set sampling to get more than one sample from each set. *Journal of Survey Statistics and Methodology*, **6**(3), 285–305.

Patil, G. P., Sinha, A. K., and Taillie, C. (1994), Ranked set sampling for multiple characteristics. *International Journal of Ecology and Environmental Sciences*, **20**, 94–109.

Patil, G. P., Sinha, A. K., and Taillie, C. (1999), Ranked set sampling: A bibliography. *Environmental and Ecological Statistics*, **6**, 91–98.

Patil, G. P., Sinha, A. K., and Taillie, C. (1994), *Ranked set sampling*, in Handbook of Statistics, Environmental Statistics, Vol. 12, G.P. Patil and C.R. Rao, eds, NorthHolland, Amsterdam.

Ridout, M. S. (2003), On ranked set sampling for multiple characterestics. *Environmental and Ecological Statistics*, **10**, 225–262.

Samawi, H. M. (1996), Stratified ranked set sample. *Pakistan Journal of Statistics*, **12**, 9–16.

Sarndal, C. E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York, Springer.

Yang, S. S. (1977), General distribution theory of the concomitants of order statistics. *The Annals of Statistics*, **5**, 996–1002.

Appendix

A. Proof of Theorem 1

Proof of the theorem is similar to Panahbehagh et al. (2018) and just note that here $E(I_{[h]i}) = E(I_{(h)i})$.

B. Proof Theorem 2

Here according to the sampling strategy, (i) taking an *iid* sample from f (a model) and (ii) taking a stratified finite population sampling from the selected sample (a design), we have a Model-Design based sampling, let indexes of "M" and "D", mean "according to the Model and the Design" respectively. Then with

$$\widehat{\mu}^j_P = \sum_{h=1}^{m} W_h \bar{X}^j_{\{h\}},$$

where

$$I_{\{h\}i} = \begin{cases} 1 & \text{if } X^j_{\{h\}i} \text{ is in the } s_{\{h\}}, \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$\begin{aligned} E(\widehat{\mu}^j_P) &= E_M[E_D(\widehat{\mu}^j_P|\mathbf{X}_{Km})] = E_M[\sum_{h=1}^{m} \frac{K_h}{Km}\frac{1}{n_h}\sum_{i=1}^{K_h} X^j_{\{h\}i} E_D(I_{\{h\}i}|\mathbf{X}_{Km})] \\ &= E_M(\sum_{h=1}^{m} \frac{K_h}{Km}\frac{1}{n_h}\sum_{i=1}^{K_h} X^j_{\{h\}i}\frac{n_h}{K_h}) = \frac{1}{Km}E_M(\sum_{h=1}^{m}\sum_{i=1}^{K} X^j_{hi}) = \mu^j, \end{aligned}$$

where $\mathbf{X}_{Km}$ indicates whole sample of size Km.

C. Proof of Theorem 3

Here the design affected by two sources of variations; variation from selecting one of the LEs and variation from selecting the sample from the fixed form of the stratified population conditional on the result of the LEs which we indicate them with "D_1" and "D_2" respectively. Therefore here based on LEs assume we have $\mathbf{X}_{Km.q}; q = 1, 2, ..., Q$ and

$\mathbf{X}_{Km.q}$ may happen with probability L_q. Please note that $L_q = \frac{1}{Q}$ because all combinations of LEs happen with equal probability. Then we have

$$E(\widehat{\mu}^j_R) = E_M E_{D_1} E_{D_2}(\widehat{\mu}^j_R),$$

now as

$$E_{D_2}(I_{[h\}i}) = \frac{n}{K},$$

and

$$E_{D_1} E_{D_2}(\widehat{\mu}^j_R) = E_{D_1}(\bar{X}^j_{Km}) = \bar{X}^j_{Km},$$

we have

$$E(\widehat{\mu}^j_R) = E_M[\frac{1}{mK}\sum_{h=1}^{m}\sum_{i=1}^{K} X^j_{hi}] = \mu^j,$$

then $E(\widehat{\mu}^j_R) = \mu^j_R$.

For variance we have

$$V(\widehat{\mu}^j_R) = V_M E_{D_1} E_{D_2}(\widehat{\mu}^j_R) + E_M V_{D_1} E_{D_2}(\widehat{\mu}^j_R) + E_M E_{D_1} V_{D_2}(\widehat{\mu}^j_R).$$

It is easy to see that

$$V_M E_{D_1} E_{D_2}(\widehat{\mu}^j_V) = \frac{\sigma^2_j}{Km},$$

and then as $V_{D_1} E_{D_2}(\widehat{\mu}^j_R) = 0$ (because $E_{D_2}(\widehat{\mu}^j_R) = \bar{X}^j_{Km}$ is not variable respect to D_1) we have

$$E_{D_1} V_{D_2}(\widehat{\mu}^j_R) = \frac{1}{m^2}\sum_{h=1}^{m}\frac{1-\frac{n}{K}}{n}\frac{1}{Q}\sum_{q=1}^{Q} S^2_{[h\}qjK},$$

and therefore

$$V(\widehat{\mu}^j_R) = \frac{\sigma^2_j}{Km} + \frac{1}{m^2}\sum_{h=1}^{m}\frac{1-\frac{n}{K}}{n} E_M(\frac{1}{Q}\sum_{q=1}^{Q} S^2_{[h\}qjK}),$$

where $\frac{1-n/N}{n} S^2_{[h\}qjK} = V_{D_2}(\frac{1}{K}\sum_{i=1}^{K} X^j_{hi})$ and $\frac{1}{Q}\sum_{q=1}^{Q} S^2_{[h\}qjK} = E_{D_1}(S^2_{[h\}qjK})$.

For the unbiased estimator of the variance first note that as we take an *iid* sample for each set and rank them in m ranks then rank for each unit is distributed uniformly in vector $(1, 2, ..., m)$ and therefore we have

$$\mu^j = E(X_1^j) = EE(X_1^j|rank(X_1^j)) = \frac{1}{m}\sum_{h=1}^{m} E(X_1^j|rank(X_1^j) = h) = \frac{1}{m}\sum_{h=1}^{m} \mu^j_{[h\}},$$

$$\begin{aligned}
\sigma_j^2 &= VE(X^j|rank(X^j)) + EV(X^j|rank(X^j)) \\
&= V[\sum_{h=1}^{m} \mu^j_{[h\}} I(rank(X^j) = h)] + E[\sum_{h=1}^{m} \sigma^2_{[h\}j} I(rank(X^j) = h)] \\
&= \frac{1}{m}\sum_{h=1}^{m}(\mu^j_{[h\}} - \mu)^2 + \frac{1}{m}\sum_{h=1}^{m} \sigma^2_{[h\}j}, \qquad \text{(C.1)}
\end{aligned}$$

where $rank(X_1)$ indicates rank of X_1 in its selected set and $I(rank(X_1^j) = h)$ is an indicator function which takes 1, if $rank(X_1^j) = h$.
Then

$$E(\widehat{V}(\widehat{\mu}_R^j)) = \frac{1}{nm(Km-1)}[E(\sum_{h=1}^{m}\sum_{i\epsilon s_{[h\}}}(X^j_{[h\}i} - \widehat{\mu}_R^j)^2) + (K-n)E(\sum_{h=1}^{m} s^2_{[h\}j})].$$

Now as

$$E(\sum_{h=1}^{m} s^2_{[h\}j}) = \sum_{h=1}^{m} E_M(\frac{1}{Q}\sum_{q=1}^{Q} S^2_{[h\}qjK}),$$

and

$$\begin{aligned}
& E(\sum_{h=1}^{m}\sum_{i\epsilon s_{[h\}}}(X^j_{[h\}i} - \widehat{\mu}_R^j)^2) \\
= \; & E(\sum_{h=1}^{m}\sum_{i\epsilon s_{[h\}}}(X^j_{[h\}i})^2) - nmE(\widehat{\mu}_R^j)^2 \\
= \; & E(\sum_{h=1}^{m}\sum_{i=1}^{K}(X^j_{[h\}i})^2 I_{[h\}i}) - nmV(\widehat{\mu}_R^j) - nmE^2(\widehat{\mu}_R^j) \\
= \; & \sum_{h=1}^{m}\frac{n}{K}K(\sigma^2_{[h\}j} + (\mu^j_{[h\}})^2) - \frac{n}{K}\sigma_j^2 - \frac{1}{m}\sum_{h=1}^{m}(1-\frac{n}{K})E_M(\frac{1}{Q}\sum_{q=1}^{Q} S^2_{[h\}qjK}) - nm\mu^2
\end{aligned}$$

$$\begin{aligned}
&= nm(\frac{1}{m}(\sum_{h=1}^{m}\sigma^2_{[h\}j}+\sum_{h=1}^{m}(\mu^j_{[h\}}-\mu)^2))-\frac{n}{K}\sigma^2_j-\frac{1}{m}\sum_{h=1}^{m}(1-\frac{n}{K})E_M(\frac{1}{Q}\sum_{q=1}^{Q}S^2_{[h\}qjK}) \\
&= (nm-\frac{n}{K})\sigma^2_j-\frac{1}{m}\sum_{h=1}^{m}(1-\frac{n}{K})E_M(\frac{1}{Q}\sum_{q=1}^{Q}S^2_{[h\}qjK}),
\end{aligned}$$

where the last equation is based on C.1, we have

$$E(\widehat{V}(\widehat{\mu^j_R}))=V(\widehat{\mu^j_R}).$$