Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2008/05

# Ranked Set Sampling Allocation Models for Multiple Skewed Variables: an Application to Agricultural Data

C. Bocci, A. Petrucci, E. Rocco

Università degli Studi
di Firenze

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

# Ranked Set Sampling Allocation Models for Multiple Skewed Variables: an Application to Agricultural Data

Chiara Bocci, Alessandra Petrucci and Emilia Rocco[*] [†]

*Department of Statistics "Giuseppe Parenti", University of Florence*

## SUMMARY

The mean of a balanced ranked set sample is more efficient than the mean of a simple random sample of equal size and the precision of ranked set sampling may be increased by using an unbalanced allocation when the population distribution is highly skewed. The aim of this paper is to use the data of the Italian Fifth Agricultural Census driven in year 2000 and of the Italian Farm Structure Survey driven in year 2003 in order to compare several possible allocation rules and to identify the more appropriate one when several skewed distributed attributes of each sample are of interest. Our study shows that when an auxiliary variable correlated with the study variables is available and is used as ranking variable, a multivariate extension of the univariate unequal allocation models suggested for skew distributions by Kaur *et al.* (1997) may be a good choice.

KEY WORDS: allocation rules; concomitant variables; multiple characteristics; skewness

## 1. INTRODUCTION

Observational economy can be achieved when is possible to identify a large number of sample units that represent the population of interest and yet only a carefully selected sub-sample is examined. This was first recognized by McIntyre (1952) who proposed

[*] Correspondence to: Emilia Rocco, Dipartimento di Statistica "G. Parenti", Università degli Studi di Firenze, viale Morgagni 59, 50134 Firenze, Italy.

[†] E-mail: rocco@ds.unifi.it

ranked set sampling (RSS) to estimate the mean pasture and forage yields. The RSS is a cost efficient alternative to the simple random sampling (SRS) for the population mean estimation. RSS performs better then SRS when the units corresponding to each rank are allocated equally and its performance further improves when appropriate unequal allocation is implemented. The variance of the balanced RSS mean estimator is not greater than the one of the SRS estimator regardless of either ranking errors or the shape of the underlying distribution of the study variable. This has been proven both theoretically and empirically by Dell and Clutter (1972) and Stokes (1977), among others. On the other hand, the unbalanced RSS - if not properly applied - may not produce the expected improvement respect to the balanced RSS and its performance may decrease becoming even worse of that of the SRS. The unbalanced RSS protocols are also sensible to possible errors in judgment ranking.

Another element that may affect the choice between different RSS protocols is the number of attributes of each sample that are of interest. For estimating multiple characteristics using RSS, McIntyre (1952) suggests applying the RSS procedure to a single selected characteristic and taking one's chance regarding the performance of the method for the other characteristics. Several other authors have considered estimating multiple characteristics using RSS and have introduced some possible solutions. This work compares some RSS protocols in order to find the more "advantageous" one in terms of variance reduction when the simultaneous estimation of the means of several skewed variables is of interest.

The analysis is carried out through an empirical study based on the data of the Italian Fifth Agricultural Census driven in year 2000 and the data of the Italian Farm Structure Survey driven in year 2003. The sampling units are ranked on the basis of auxiliary variables provided by the frame.

The next section briefly recalls the basic elements of each of the considered RSS protocols, the data and the empirical study are described respectively in section 3 and 4. Section 5 presents the results. Finally, some concluding remarks and open questions are given in section 6.

## 2. BASIC RANKED SET SAMPLING METHODOLOGY

Originally the RSS methodology was developed considering the easier case in which only one variable is of interest. A ranked set sample can be considered as obtained from $n = \sum_{i=1}^{m} r_i$ sets of random sampling units, each of size *m*, by first ordering the units of each set and then, for $i = 1,2,...,m$, measuring the $i$th ranked order statistics for $r_i$ sets. When all the $r_i$ are equal a balanced ranked set sample is obtained. In the other case several allocation rules have been proposed in literature to determine the $r_i$ values. The sampled units are ordered according to the characteristic of interest, without quantifying it. Ranking could be based on judgment, visual inspection, covariates or any other method that does not require actual measurement of the units.

A general expression of the ranked set sample mean estimator for a sample with $r_i$ observations at rank *i* is:

$$\hat{\mu}_{rss} = \frac{1}{m}\sum_{i=1}^{m}\frac{T_i}{r_i}, \tag{1}$$

where $T_i$ is the sum of the observations that were assigned to rank *i*. If all the $r_i$ are greater than zero (1) is an unbiased estimator of the population mean. Moreover its variance is:

$$\text{var}(\hat{\mu}_{rss}) = \frac{1}{m^2}\sum_{i=1}^{m}\frac{\sigma^2_{(i:m)}}{r_i}$$

where $\sigma^2_{(i:m)}$ denotes the variance of the $i$th order statistics.

In the case of equal allocation the estimator (1) corresponds to the sample mean.

In the Neyman unequal allocation, the number of sample units for each rank order statistic is allocated proportionally to its standard deviation (SD). Therefore this RSS protocol is the optimal one in terms of variance reduction of the mean estimator, but its implementation becomes impractical in many real applications in which the SD of order statistics are unavailable. If wrong values of the SD are used, the performance of RSS based on Neyman allocation may become worse than that of the SRS. The negative effect of the approximate values for the standard deviation can be aggravate by errors in judgment ranking as an approximate good SD for an order statistics may not be good for a group of units erroneously ascribed to this order statistics.

Kaur *et al.* (1997) suggested to make a “near” optimal allocation for the cases in which the underlying statistical distribution is unknown, and hence the standard deviations of the order statistics. The proposal consists of taking advantage of the knowledge of more easily available characteristics of the population such as skewness, kurtosis and coefficient of variation (CV). In the case of positively skewed or right-tailed distribution in $(0,\infty)$, the variances of order statistics typically increase with the rank orders. According to this consideration Kaur *et al.* (1997) proposed two models of unequal allocation for skewed distributions that are conform with the Neyman's optimal allocation. In the first one, referred as *t-model*, the largest order statistics is quantified $t \geq 1$ times more than the rest. In the second, referred as $(s,t)$*-model* the two largest order statistics are quantified respectively by factors $s$ and $t$ ($1 \leq s \leq t$) more than the rest. The optimal values for the parameters $t$ and $(s,t)$ of these allocation models are function of the variances of the order statistics, which are not generally available. Their unavailability was indeed the motivation of these models. In order to avoid this difficulty Kaur *et al.* (1997) suggest a rule of thumb based on skewness, kurtosis, or CV of the underlying distribution which allows to identify near optimal $t$ or $(s,t)$ values.

If more variables are of interest, the amount of correlation between them affects the choice between different RSS protocols. If all the study variables are correlated, a balanced ranked set sample with respect to one variable may be appropriate for the other variables as well. The efficiency gains for the ranked set sample mean with respect to the mean of a simple random sample for the variables not explicitly considered in the sample selection process is a function of their correlation with the only study variable considered in the sample selection process. Indeed, for each of them the ranking errors will increase and consequently the efficiency gains will decrease as the correlation coefficient decreases.

The application to the multiple characteristics case of the unbalanced RSS based on Neyman allocation is not obvious because the optimal allocation for one variable may be not the optimal one for another variable even if the two variables are correlated.

For correlated positively skewed variables the RSS protocol of Kaur *et al.* (1997) with respect to one variable may be appropriate for the other variables as well. The application of one of the two unequal allocation models proposed by Kaur *et al.* (1997) to one study variable ensures the larger efficiency gains over equal allocation when the parameters of the models are optimal. Moreover, it also performs better than equal allocation when approximate values of the model parameters are used and such values are inside a reasonably broad interval that depends on the underlying distribution. In the multiple variable case efficiency gains may be obtained as well for all the study variables because the chosen $t$- $(s,t)$ parameters with respect to one variable should be reasonably close to the optimal $t$- $(s,t)$ parameters of the others.

In the paper we assume the Kaur *et al.* $(s,t)$ -model but we does not use the method that they suggest to estimate the $(s,t)$ parameters. We consider an alternative allocation rule given that we are dealing with a multivariate problem and that we are not willing to make any assumption on the parameters of the distributions. We propose to equally split a portion of the sample units in all the ranks and to assign the remaining units to the two largest order statistics, more to the largest than to the second largest.

Specific RSS protocols for the simultaneous estimation of the mean of two or more characteristics are present in literature; the bivariate version of balanced RSS proposed by Al-Saleh and Zheng (2002) and the RSS protocol proposed by Ridout (2003) are two of them. Both these two multivariate RSS protocols are referred to situations in which the balanced allocation is suitable. However, even if for skewed variables an unbalanced allocation may be preferable, they are considered for their peculiarity of taking into account explicitly more variables during the ranking process.

## 3. THE DATA

The data on Italian farms may be considered a natural setting to apply RSS protocols and to investigate how they work when the mean estimation of more skewed variables is of interest.

The Italian Statistical Institute (ISTAT) ten-yearly drives an Agricultural Census and two-yearly drives a sample Farm Structure Survey (FSS). Both in the census and in the FSS the unit of observation is the farm and for each farm are registered the data on the surface areas allocated to different crops. In the FSS, until 2005, the productions of each crop were also observed for each farm.

In our study we assume as population the set of the farms in the province of Florence included in the Fifth Italian Agricultural Census driven in the year 2000 and pick as study variables the variables in the year 2003 (the last year for which, at the time of writing, the data of the FSS are available) related to the main cultivation of the Florentine area, that is the surface areas allocated to grapevines and olives and the corresponding productions. The arable crops surface, even if it is not much cultivated in the area, has been included in the set of study variables because it allows us to evaluate the performance of the RSS solutions with low correlated variables.

We use as ranking criterion known variables for all the farms that are correlated with the study variables: the surface area allocated to olives, grapevines and arable crops and the European size unit (UDE), all registered at 2000 Census.

The census information are, obviously, available for all the farms. The information collected with FSS are available only for the sampled farms. As in our study more RSS protocols are empirically compared using Monte Carlo experiments we need to know the study variables for all the farms, thus for the farms not sampled they are simulated in the following way:

a) using the Census information, the population is first divided into clusters according to the surface area allocated to olives, grapevines and arable crops, the UDE, and the utilized agricultural surface (SAU), all at year 2000, requiring at least one farm from the 2003 FSS in each group;

b) then the missing values of the study variables for the non FSS farms are imputed through a hot deck method based on random donor.

Table 1 summarizes the main characteristics of the study variables with respect to the whole population. It is evident that all the study variables have a highly positively skewed distribution and in all the cases the density is concentrated in the last percentiles.

**Table 1**: *Distribution analysis of the study variables*

| **Distribution Analysis** | **olives surface 2003 (are)** | **olives production 2003 (q)** | **grapevines surface 2003 (are)** | **grapevines production 2003 (q)** | **arable crops surface 2003 (are)** |
|---|---|---|---|---|---|
| **Mean** | 169.007 | 12.451 | 122.672 | 80.678 | 314.103 |
| **CV** | 3.005 | 3.830 | 4.474 | 4.197 | 4.441 |
| **Skewness** | 12.344 | 13.824 | 14.969 | 11.285 | 12.966 |
| **Kurtosis** | 299.035 | 289.876 | 430.826 | 193.962 | 270.470 |
| **Q1 (25%)** | 0 | 0 | 0 | 0 | 0 |
| **Median (50%)** | 60 | 2 | 10 | 7 | 23 |
| **Q3 (75%)** | 176 | 10 | 60 | 40 | 112 |
| **95%** | 560 | 60 | 416 | 320 | 1200 |
| **99%** | 2000 | 154.5 | 2400 | 1452 | 5429 |
| **Max** | 15700 | 1500 | 26477 | 10000 | 49000 |

Table 2 shows the correlation coefficients between the study variables and the ranking variables. The table exhibits a modest correlation between the arable crop cultivation and the other two cultivations. The use of the RSS strategies that use only one variable for ranking and consider it as concomitant for the others may be improper as a consequence of this modest correlation in case all the cultivations are considered equally relevant. From Table 2 we can also note that the correlation coefficients between the UDE and each study variable are in between the correlation coefficients of each study variable and the surface allocated to the same crop at year 2000, and the correlation coefficients between each study variable and the surface allocated to a different crop. Therefore, the use of UDE as ranking variable may be a useful compromise in the case that all the study variables are assumed equally relevant.

**Table 2**: *Correlation among variables*

| **Pearson Correlation Coefficient** | **olives surface 2003** | **olives production 2003** | **grapevines surface 2003** | **grapevines production 2003** | **arable crops surface 2003** |
|---|---|---|---|---|---|
| **UDE 2000** | 0.642 | 0.407 | 0.758 | 0.708 | 0.537 |
| **olives surface 2000** | 0.849 | 0.483 | 0.529 | 0.479 | 0.218 |
| **grapevines surface 2000** | 0.541 | 0.368 | 0.882 | 0.789 | 0.229 |
| **arable crops surface 2000** | 0.244 | 0.149 | 0.362 | 0.379 | 0.895 |

## 4. THE EMPIRICAL STUDY

The use of a RSS sampling design needs choosing how to rank the units, how many ranks to consider and how to allocate the units in the ranks.

In our study we examine several RSS scenarios. In all the scenarios the size of the sample is 300 (roughly equal to the number of units sampled in the 2003 FSS) and the ranking is based on census variables known for all the population units. The rules that we consider are: a) equal allocation, b) Neyman allocation, c) unequal allocation models for skew variables proposed by Kaur *et al.* (1997), d) bivariate allocation model of Al-Saleh and Zheng (2002), e) the allocation rule of Ridout (2003). For each allocation rule different settings are explored.

For the equal allocation scheme we assume each study variable as the most relevant one at the time. When the variable of interest is a surface allocated to a crop we use as ranking variable the same variable at census time and when it is a production of a crop we use the corresponding surface area allocated to the crop at census time. We also investigate what happens using as a ranking variable census UDE that is correlated with all the study variables (see Table 2).

For the Neyman allocation, we approximate the SD of the study variables with the SD of the same variable at census time. Assuming in turn each of the three surface area allocated as more relevant we use the corresponding census variable both as ranking variable and to estimate the SD which is then used to establish the number of units in each rank.

Concerning the allocation models suggested for skewed variables by Kaur *et al.* (1997), in our study we adopt the idea that for a positively skewed variable the variance of the mean estimator may be reduced observing more units at the right tail that is in correspondence of the last or two last order statistics. Unlike Kaur *et al.* however, we assume not to know anything else on the distribution of the study variable besides its not quantified skeweness so we are unable to calculate neither the optimal $t$ or $(s,t)$ parameters or their approximate values. Therefore, we simply test more allocation scenarios in which a different constant number of unit $k$ is allocated in each rank and the remaining units are assigned one-third to the second last rank and two-thirds to the last rank. The scenarios varies also in the number of ranks and in the ranking variable.

For the bivariate allocation model of Al-Saleh and Zheng (2002) four possible scenarios are considered: the simultaneous use for ranking of census olives surface and census grapevines surface with $m$ (the number of rank for each variable) equal to 3 and to 10 and the simultaneous use for ranking of census olives surface and census arable crops surface with $m$ equal to 3 and to 10. The mean estimation is extended to all the study variables even if we use a bivariate allocation model that supposes to deal only with 2 study variables, assuming these as concomitant for all the others. We are aware that the bivariate model can be theoretically generalized to the case of multiple variables but its practical extension may be unfeasible in real situations as the number of units to use for the ranking will be intractable.

The allocation model of Ridout (2003) during the selection process explicitly counts the number of units assigned to each rank for each study variables so does not agree with mean estimation of variables not considered in the selection process.

In our study, we firstly apply the Ridout allocation rule considering as study variables only that concerning olives and grapevines and then adding also the surface allocated to arable crops which mean in the first case is not estimable.

To evaluate the relative performance of mean estimator associated to each RSS scenarios 10.000 samples are selected for each of them as well as for the SRS procedure and for each sample the mean estimate is computed. For each scenario the variance of the mean estimator is empirically estimated using the corresponding set of 10.000 samples and the relative precision respect to the SRS is evaluated on the basis of the relative precision index $\mathrm{RP} = Var(\hat{\mu}_{rss})/Var(\hat{\mu}_{srs})$. All sampling procedures are with replacement and use the same sequence of random values.

## 5. RESULTS

The simulation clearly indicates that a proper application of a RSS procedure based on the use of auxiliary variables can lead to substantial efficiency gains with respect to SRS.

For the equal allocation we use the values *m=3, 10, 15, 30* as the number of ranks. In Table 3 we omit the results for *m=15*, as they follow a pattern similar to the others and lie, as expected, between *m=10* and *m=30*. Its known in literature that the relative variance reduction increases when the number of ranks increases, however in many real situation it is too expensive or not feasible to use a number of ranks even greater than 3. In our case, being the ranking based on variables available from list, choose a “big” value for the number of ranks is not a problem.

As the number of ranks increases, also the ranking errors may increase. This could reduce the efficiency gain attainable by the use of a larger number of ranks. However, the correlation between the study variables and the ranking variables could mitigate this second effect.

Analyzing Table 3 we note that when the ranking variable is UDE we obtain a relative variance reduction for each study variable that is intermediate between the reduction obtained when census variable referred to the same crop is used for ranking and the reduction obtained ranking with a variable referred to another crop. This result confirms that the gain in precision depends on the correlation between the study and ranking variables. This fact is further corroborated by the other results reported in the table.

**Table 3**: *Relative precision for the equal allocation.*

| | | | Equal Allocation | | | |
|---|---|---|---|---|---|---|
| **Ranking Variable** | **olives surface 2003** | **Olives production 2003** | **grapevines surface 2003** | **grapevines production 2003** | **arable crops surface 2003** | ***Average Relative Precision*** |
| **UDE 2000** | | | | | | |
| m=3 | 0.9611 | 0.9564 | 0.9142 | 0.9445 | 0.9358 | *0.9424* |
| m=10 | 0.8705 | 0.9181 | 0.8241 | 0.8538 | 0.8755 | *0.8684* |
| m=30 | 0.7780 | 0.8912 | 0.7094 | 0.7345 | 0.7662 | *0.7759* |
| **olives surface 2000** | | | | | | |
| m=3 | 0.9614 | 0.9411 | 0.9270 | 0.9540 | 0.9517 | *0.9470* |
| m=10 | 0.8419 | 0.9029 | 0.8717 | 0.9078 | 1.0097 | *0.9068* |
| m=30 | 0.7074 | 0.8798 | 0.8231 | 0.8564 | 0.9387 | *0.8411* |
| **grapevines surface 2000** | | | | | | |
| m=3 | 0.9902 | 0.9691 | 0.9229 | 0.9477 | 0.9621 | *0.9584* |
| m=10 | 0.9266 | 0.9207 | 0.8228 | 0.8434 | 0.9706 | *0.8968* |
| m=30 | 0.8595 | 0.9175 | 0.7037 | 0.7277 | 0.9324 | *0.8282* |
| **arable crops surface 2000** | | | | | | |
| m=3 | 1.0000 | 0.9601 | 0.9441 | 0.9682 | 0.9198 | *0.9592* |
| m=10 | 0.9665 | 0.9432 | 0.8970 | 0.9170 | 0.8333 | *0.9114* |
| m=30 | 0.9343 | 0.9330 | 0.8620 | 0.8695 | 0.6552 | *0.8508* |

The simulation results of the Neyman allocation show that when wrong values of the SD are used its performance - in terms of relative variance reduction - may become worse than that of the SRS. Table 4 reports that, for *m=30*, the Neyman allocation based on an approximate SD works better then the equal allocation for the primary interest variable. For the other variables the performance is often worse than the SRS despite the fact that the correlation coefficient between the variables is more than 0.5 (see Table 2).

Neyman allocation is also affected by ranking errors due to the use of an approximate SD. Keeping the correlation between variables constant, the relative efficiency is lower if the study variables and/or ranking variables have higher variability (see Table 4, ranking with olives surface vs. ranking with grapevines surface). As for the ranking errors, allocation errors tend to increase with the increase of the number of ranks. The total effect of the two types of error tends to dump the efficiency gains due to the use of a greater number of ranks, which could even lead to an overall efficiency loss. Specifically, as documented in Table 4, in our experiment as the number of ranks increases efficiency gets worse and it begins to increase when the number of ranks is greater than 10.

**Table 4**: *Relative precision for the Neyman allocation.*

| | | | **Neyman Allocation** | | | |
|---|---|---|---|---|---|---|
| **Ranking Variable** | **olives surface 2003** | **Olives production 2003** | **grapevines surface 2003** | **grapevines production 2003** | **arable crops surface 2003** | ***Average Relative Precision*** |
| **olives surface 2000** | | | | | | |
| m=3 | 0.6254 | 1.0856 | 0.9389 | 1.1745 | 3.2974 | *1.4244* |
| m=10 | 0.8418 | 1.9133 | 1.9282 | 2.532 | 8.0762 | *3.0583* |
| m=30 | 0.4208 | 1.1550 | 0.8824 | 1.2304 | 4.1720 | *1.5721* |
| **grapevines surface 2000** | | | | | | |
| m=3 | 2.8946 | 4.1282 | 1.0691 | 1.5008 | 7.0176 | *3.3221* |
| m=10 | 4.1621 | 5.5708 | 1.2975 | 1.8727 | 9.4925 | *4.4791* |
| m=30 | 2.2118 | 2.5428 | 0.6321 | 0.9382 | 4.5172 | *2.1684* |
| **arable crops surface 2000** | | | | | | |
| m=3 | 3.6207 | 4.3395 | 2.0937 | 2.2007 | 0.5579 | *2.5625* |
| m=10 | 6.9305 | 7.9763 | 3.7103 | 4.0344 | 0.5156 | *4.6334* |
| m=30 | 2.9656 | 3.4823 | 1.5914 | 1.7393 | 0.2339 | *2.0025* |

In order to get some further insights on the performance of this allocation rule, a Monte Carlo experiment on the Neyman allocation based on exact SD, is also performed. The result, not included in the tables, is obviously optimal for the estimate of the variable used for ranking (for example, the relative precision of olives surface is 0.06). For the other variables the previous argument applies.

The largest variance reduction on average for all study variables is obtained with our multivariate and approximate interpretation of allocation rule suggested for skewed variables by Kaur *et al.* (1997). Therefore, the skewness of the study variable distribution could significantly influence the performance of RSS even if it is not exactly quantified and the design parameters - specifically the pair $(s,t)$ - are approximated. In particular, from Table 5 we note that using census UDE as ranking variable, for *m=30* as well as *m=15* and *m=10*, it works better on average for each study variables than all the other allocation models. The only exceptions are for *m=10* and *k=3* and for *m=15* and *k=3*. From the table we also note that when *m* increases the best repartition of units occurs when the number of units selected in each rank and the portion of the units that remains to be assigned both decreasing. This probably because when the number of ranks increases the variability within the ranks that are not on the right tail decreases so we observe less units in each one of them but the total number of units assigned to this ranks does not decreases with the same rapidity. We verify that for each value of *m* the better performance corresponds to the repartition of the units nearer to the optimal $(s,t)$ parameters of Kaur *et al.* (1997) for all the study variables, replacing the variances of order statistics with the variances of the “approximate” order statistics obtained using UDE at year 2000 as ranking variable. It is obvious that in real situations is very difficult or impossible to calculate the $(s,t)$ values according to this observation. However, the robustness of this kind of allocation rule with respect to other parameter choices makes it a very appealing solution. Moreover, as can be noted especially from the last four rows of Table 5, this allocation rule provides a significant gain in precision also when the correlation coefficient between the ranking variable and the study variable is not so high (about 0.5) and fails when the correlation is lower (see arable crops vs. olive surface). In the table we show the results for census olives surface as ranking variable only with *m=30* as the simulations with *m=3, 10, 15* follow the same pattern.

**Table 5**: *Relative precision for the rules based on Kaur et al. (1997).*

| **Rule based on Kaur *et al.*** | | | | | | |
|---|---|---|---|---|---|---|
| **Ranking Variable** | **olives surface 2003** | **olives production 2003** | **grapevines surface 2003** | **grapevines production 2003** | **arable crops surface 2003** | ***Average Relative Precision*** |
| **UDE 2000 – m=3** | | | | | | |
| k=3 | 0.8522 | 2.1515 | 0.5704 | 0.7556 | 0.6873 | *1.0034* |
| k=5 | 0.7028 | 1.4934 | 0.5395 | 0.6557 | 0.6289 | *0.8041* |
| k=7 | 0.6481 | 1.2314 | 0.5297 | 0.6173 | 0.5988 | *0.7251* |
| k=9 | 0.6164 | 1.0755 | 0.5255 | 0.5976 | 0.5884 | *0.6807* |
| k=12 | 0.5920 | 0.9454 | 0.5221 | 0.5800 | 0.5728 | *0.6425* |
| k=15 | 0.5846 | 0.8734 | 0.5241 | 0.5755 | 0.5711 | *0.6257* |
| k=20 | 0.5783 | 0.7996 | 0.5328 | 0.5759 | 0.5716 | *0.6116* |
| k=25 | 0.5863 | 0.7571 | 0.5427 | 0.5815 | 0.5776 | *0.6090* |
| k=50 | 0.6542 | 0.7334 | 0.6237 | 0.6506 | 0.6509 | *0.6626* |
| **UDE 2000 – m=10** | | | | | | |
| k=3 | 0.6891 | 1.6058 | 0.3391 | 0.5091 | 0.6841 | *0.7654* |
| k=5 | 0.4768 | 1.0676 | 0.2778 | 0.3806 | 0.4785 | *0.5363* |
| k=7 | 0.4003 | 0.7984 | 0.2428 | 0.3190 | 0.3873 | *0.4296* |
| k=9 | 0.3541 | 0.6777 | 0.2364 | 0.2919 | 0.3526 | *0.3825* |
| k=12 | 0.3417 | 0.5790 | 0.2428 | 0.2843 | 0.3369 | *0.3569* |
| k=15 | 0.3365 | 0.5288 | 0.2674 | 0.3036 | 0.3480 | *0.3569* |
| k=20 | 0.3734 | 0.5103 | 0.3280 | 0.3557 | 0.3799 | *0.3895* |
| k=25 | 0.5100 | 0.6143 | 0.4676 | 0.4810 | 0.4992 | *0.5144* |
| **UDE 2000 – m=15** | | | | | | |
| k=3 | 0.5889 | 1.3397 | 0.3083 | 0.4512 | 0.6624 | *0.6701* |
| k=5 | 0.4150 | 0.8623 | 0.2286 | 0.3184 | 0.4314 | *0.4511* |
| k=7 | 0.3415 | 0.6686 | 0.2046 | 0.2693 | 0.3617 | *0.3691* |
| k=9 | 0.3020 | 0.5389 | 0.2022 | 0.2556 | 0.3388 | *0.3275* |
| k=12 | 0.3045 | 0.4794 | 0.2269 | 0.2618 | 0.3318 | *0.3209* |
| k=15 | 0.3485 | 0.4877 | 0.2914 | 0.3202 | 0.3676 | *0.3631* |
| **UDE 2000 – m=30** | | | | | | |
| k=3 | 0.4747 | 0.9363 | 0.2687 | 0.3729 | 0.5719 | *0.5249* |
| k=5 | 0.3247 | 0.6170 | 0.1951 | 0.2561 | 0.4007 | *0.3587* |
| k=7 | 0.2893 | 0.4832 | 0.1921 | 0.2371 | 0.3392 | *0.3082* |
| k=9 | 0.3822 | 0.5279 | 0.3110 | 0.3430 | 0.4263 | *0.3981* |
| **olives surface 2000 – m=30** | | | | | | |
| k=3 | 0.3990 | 1.0005 | 0.8624 | 1.1196 | 2.1199 | *1.1003* |
| k=5 | 0.2759 | 0.6169 | 0.5428 | 0.6821 | 1.2518 | *0.6739* |
| k=7 | 0.2483 | 0.5178 | 0.4443 | 0.5450 | 0.9129 | *0.5337* |
| k=9 | 0.3276 | 0.5371 | 0.5067 | 0.5713 | 0.8271 | *0.5540* |

In the cases in which all the study variables are equally relevant but moderately correlated, the Ridout method (Table 6) and the bivariate method (Table 7) could be valid alternatives to our modification of the rule of Kaur *et al.* (1997) and the Ridout method should be preferred. Another reason for which they could be preferable may be in case in which not all the study variables are skewed. To investigate this possibility we create a symmetric variable correlated with all the ranking variables (its correlation coefficient varies from 0.4 with arable crops surface to 0.6 with UDE). The simulations implemented adding this study variable confirm that the protocols that generate an equal or nearly equal allocation are the best for this variable. However the gain in precision respect to SRS obtained with our multivariate and approximated Kaur *et al.* model appears to be relevant. For example with *m=30* and using UDE as ranking variable, the RP index varies from 0.7963 when *k=3* and 0.2797 when *k=9*.

**Table 6**: *Relative precision for the Ridout method.*

| | | | Ridout Method | | | |
|---|---|---|---|---|---|---|
| **Ranking Variable** | **olives surface 2003** | **olives production 2003** | **grapevines surface 2003** | **grapevines production 2003** | **arable crops surface 2003** | ***Average Relative Precision*** |
| **olives and grapevines surface 2000** | | | | | | |
| m=3 | 0.8200 | 0.6611 | 0.6173 | 0.6216 | - | *0.6800* |
| m=10 | 0.6005 | 0.5969 | 0.5723 | 0.5660 | - | *0.5839* |
| m=30 | 0.6098 | 0.6304 | 0.5445 | 0.5464 | - | *0.5828* |
| **olives, grapevines and arable crops surface 2000** | | | | | | |
| m=3 | 1.1878 | 0.8358 | 0.8586 | 0.8517 | 0.6480 | *0.8764* |
| m=10 | 0.6495 | 0.5996 | 0.6009 | 0.6044 | 0.5592 | *0.6027* |
| m=30 | 0.6584 | 0.6601 | 0.6800 | 0.6822 | 0.5378 | *0.6437* |

Comparing Table 3 and Table 7, we note that the bivariate equal allocation model performs better than the univariate one based on the same *m*-value. This confirms the statement of Al-Saleh and Zheng (2002) that when the study variables are unrelated the bivariate RSS is equivalent - in terms of efficiency of the mean estimation of each variable - to the univariate unbalanced RSS; otherwise, if the variables are correlated, it works better.

**Table 7**: *Relative precision for the bivariate method.*

| Bivariate Method | | | | | | |
|---|---|---|---|---|---|---|
| **Ranking Variable** | **olives surface 2003** | **olives production 2003** | **grapevines surface 2003** | **grapevines production 2003** | **arable crops surface 2003** | ***Average Relative Precision*** |
| **olives and grapevines surface 2000** | | | | | | |
| m=3 | 0.9218 | 0.9656 | 0.8958 | 0.9099 | 0.9804 | *0.9347* |
| m=10 | 0.7151 | 0.8613 | 0.6357 | 0.6750 | 0.9314 | *0.7637* |
| **olives and arable crops surface 2000** | | | | | | |
| m=3 | 0.9348 | 0.9720 | 0.9441 | 0.9595 | 0.9382 | *0.9497* |
| m=10 | 0.7959 | 0.9219 | 0.8226 | 0.8292 | 0.7451 | *0.8229* |

## 6. FINAL REMARKS

The results of this study clearly indicate that the use of ranking information that may be readily available allows to improve the efficiency of the mean estimator. The efficiency gain depend both on the correlation between the study and ranking variables and on the allocation rule. For the unequal allocation the efficiency improvements also depend upon the knowledge of the distribution parameters used for the allocation itself and the robustness of the model respect to the use of approximate parameter values.

Taking into account the skewness in the allocation model may lead to benefits in the case of a positively skewed study variable; even if the degree of skewness is unknown.

The number of study variables, the correlation between them, and the relevance assigned to each one of them are relevant in the choice between different allocation rules when more variables are of interest. The novel method proposed in this work as an extension of the rule of Kaur *et al.* (1997) obtains the best performance in terms of average reduction of the relative precision when most variables of interest are skewed.

If we compare the relative precision coefficients of each variable, the examined specific RSS protocols for multivariate cases may be sometimes a valid alternative.

Moreover, our proposal based on that suggested by Kaur *et al.* (1997) is robust respect to the use of not optimal values for its parameters. We are not aware of any theoretical justification for this property. Perhaps the allocation of some units in all the ranks protects from possible ranking errors which could increase the variability within each rank. At the same time preserving a portion of units for the two last ranks agree with the increasing of the variances of order statistics with the rank orders for positively skewed variables.

We believe that this robustness property makes the procedure useful when we are not able to allocate the units precisely and is therefore an appealing alternative to other methods that require exact knowledge of the study variable distribution.

## References

Al-Saleh MF, Zheng G. 2002. Estimation of bivariate characteristics using ranked set sampling. *Australian & New Zealand Journal of Statistics*, **44**: 221-232.

Barnett V. 1999. Ranked set sample design for environmental investigations. *Environmental and Ecological Statistics* **6**: 59-74.

Dell TR, Clutter JL. 1972. Ranked set sampling theory with other statistics background. *Biometrics* **28**: 545-555.

Husby CE, Stasny EA, Wolfe DA. 2005. An application of ranked set sampling for mean and median estimation using USDA crop production data. *Journal of Agricultural, Biological, and Environmental Statistics* **10:** 354-373.

Kaur A, Patil GP, Taillie C. 1997. Unequal allocation models for ranked set sampling with skew distributions. *Biometrics* **53**: 123-130.

McIntyre GA. 1952. A method for unbiased selective sampling, using ranked sets. *Australian journal of agricultural research* **3**: 385–390.

Patil GP. 2002. Ranked set sampling. In *Encyclopedia of Environmetrics*, El-Shaarawi AH, Piegorsch WW (eds.); John Wiley & Sons; Chichester; **3**: 1684–1690.

Patil GP, Sinha AK, Taillie C. 1994. Ranked set sampling. In *Handbook of Statistics, Volume 12: Environmental Statistics*, Patil GP, Rao CR (eds.); North-Holland; Amsterdam; 167-200.

Ridout MS. 2003. On ranked set sampling for multiple characteristics. *Environmental and Ecological Statistics* **10:** 255-262.

Stokes SL. 1977. Ranked set sampling with concomitant variables. *Communications in Statistics – Theory and Methods* **A6:** 1207-1211.

Takahasi K, Wakimoto K. 1968. On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute Statistical Mathematics* **20**: 1-31.