FORTHCOMING 62D05-03-21-02

RATIO ESTIMATOR UNDER RANK SET SAMPLING SCHEME USING HUBER M IN CASE OF OUTLIERS

Mir Subzar* and Carlos N Bouza-Herrera**
*Division of Agricultural Statistics SKUAST-K, Shalimar (190025), India
**Facultad de Matemática y Computación, Universidad de La Habana, Cuba

ABSTRACT
Rank Set Sampling (RSS) is an alternative to simple random sampling was proposed by McIntyre (1952): Under such sampling scheme various authors have proposed the ratio rank set estimators in order to estimate the population parameters by using OLS (Ordinary Least Square) method. A big issue emerge that when outliers are present in data in that case all the estimators suggested by different authors can give distorted results as OLS is very sensitive to outliers. So in the present study we mainly focus on this issue and adapted the Huber M estimation Technique on the estimator suggested by Al-Odat (2009) instead of OLS, in order to get precise results in case of presence of outliers in data.

KEYWORDS: Ratio Estimator, Rank Set Sampling, OLS, Huber M. Efficiency.

MSC: 62J05, 62G35

RESUMEN
El Muestreo por Rangos Ordenados (RSS) es una alternativa para el muestreo Simple Aleatorio que fuera propuesta por McIntyre (1952): Usando este esquema de muestreo, varios autores han propuesto estimadores de razón por rangos ordenados para estimar parámetros de la población usando el método de los MCO (Mínimos Cuadrados Ordinarios) . Un gran problema emerge es que, cuando hay outliers en la data, todos los estimadores sugeridos por los diferentes autores pueden proporcionar resultados distorsionados pues MCO es muy sensible a los outliers. Asique en el presente estudio nos enfocamos principalmente en este aspecto y adaptamos la técnica de la M estimación al estimador sugerido por Al-Odat (2009) en vez de usar MCO, para obtener precisos resultados en el caso de haber of outliers en la data.

PALABRAS CLAVE: Estimador de Razón, Muestreo por Rangos Ordenados, MCO, M-Huber M, Eficiencia.

1. INTRODUCTION

It is well-known the importance of the methods of robust estimation. The literature on M-estimation, S-estimation and MM-estimation is vast. M-estimation extends maximum likelihood method by robustifying estimators with respect to the presence of outliers. See Almongy et al. (2018), Andrews and Hampel (2015), Huber (1992): In practice the statistician observes the data and evaluates the suspicion on effects due to the existence of outliers. Then outlier detection tools are used for evaluating the data. Regression methods commonly used ordinary least squares (OLS): Robust regression is an alternative when there are some outliers. They play an important role both for analyzing the data affected by outliers and for developing models that are stout against outliers . Robust regression alternatives eliminate or weaken the influence of outliers. See superb expositions of robust statistics in Maronna, et al. (2006), Jureçková et al. (2019):
As we are familiar, statistical theoretical models consider that when collecting data all the variables are measured in the selected sampled units. If outliers are influencing the behavior of the usual estimators, is reasonable to implement some robust principle for improving the pre-selected or developed estimator Simple random sampling method determines the so-called SRS-design, see Wu-Thompson (2020): Sometimes it becomes too difficult to obtain full data and or though being possible. Then SRS becomes too time consuming and/or expensive. These facts conspire against satisfying the aims and objectives supported by sampling theory models. For this reason, McIntyre (1952) introduced a new sampling scheme known as Rank Set Sampling (RSS): His proposal was a solution for overcoming practical problems appearing usually in sample research in agriculture. In RSS sampling scheme population parameters can be estimated without using all the data in the sample. This property of RSS represents a serious gain both in time and in the involved workload. Thus, the sampling costs are diminished.

Different papers, such as Al-Omari et al. (2008), Al-Omari et al. (2009), Al-Omari (2012), Al-Omari and Bouza (2015) and Al-Omari and Al-Nasser (2018), have suggested different type of estimators for particular rank set sampling schemes. They look for getting more precise estimates than the basic ratio rank set estimator. However, applications of RSS methodology are mainly used in Agriculture and ecology and other particular fields, where is needed to deal with the possibility of having outliers. Therefore, using tools based on OLS can distort all the findings. Recently, the papers of Subzar et al. (2019a, 2019b) have suggested different ratio estimators. They used robust regression tools, instead of OLS ones, in the case of observing outliers when the sample is selected using as design Simple Random Sampling Without Replacement (SRSWOR): So, keeping this situation in mind, in the present paper we mainly focused on developing robust alternatives, looking for increasing the accuracy of the estimates in those situations under a Rank Set Sampling selection scheme.
Let

$$(Z_{11},T_{11}),(Z_{12},T_{12}),\ldots,(Z_{1n},T_{1n}),\ (Z_{21},T_{21}),\ldots,(Z_{2n},T_{2n}),\ldots,(Z_{1n},T_{1n}),\ldots,(Z_{nn},T_{nn})$$

be independent bivariate random vectors. They have the same cumulative distribution function (id): Let

$$(Z_{i[1]},T_{i[1]}),\ldots,(Z_{i[n]},T_{i[n]})$$

be the order statistics (OS) of the involved variables

$$Z_{i[1]},Z_{i[2]},\ldots,Z_{i[n]} \text{ and } T_{i[1]},T_{i[2]},\ldots,T_{i[n]}.$$

Let

$$(Z_{1[1]},T_{1[1]}),\ (Z_{2[2]},T_{2[2]}),\ldots,(Z_{n[n]},T_{n[n]})$$

denote the rank set sample where $(Z_{i[i]},T_{i[i]})$ is the $i-th$ order statistics in the $i-th$ sample for variables Z and T . Using the order statistics (OS) is a trick that transfers the properties of the random observations to those of the OS´s. Within this framework mathematical expectations use the distributional properties of OS`s.
The ratio ranked set sampling (RRSS)) estimator, which was suggested by Samawi and Muttlak (1966), is given by

$$\hat{R}_{rss}=\left(\bar{t}_{(i)}/\bar{z}_{(i)}\right),$$

where $\bar{t}_{(i)}$ is the sample mean of the variable of interest and $\bar{z}_{(i)}$ is the sample mean of the ancillary variable. Its variance is given by

$$Var\left(\hat{R}_{rss}\right)=R^2/n\left\{C_{z[n]}^2+C_{t[n]}^2-2\rho C_{z(n)}C_{t(n)}\right\},$$

where

$$C_{z(n)}=\left(S_{z(n)}/\bar{Z}\right),\ C_{t(n)}=\left(S_{t(n)}/\bar{T}\right)$$

This model has been improved in a sequel of papers as Al-Omari et al. (2008), Al-Omari et al. (2009), Al-Omari (2012), Al-Omari and Bouza (2015) and Al-Omari and Al-Nasser (2018): Their authors have suggested different ratio based estimators. We are considering developing. Robust alternative to the proposal of the paper of Al-Odat (2009): In it was suggested the following modified ratio rank set estimator

$$\bar{t}_{pr}=\left(\left(\bar{t}_{i(n)}+\lambda\left(\bar{Z}-\bar{z}_{i(n)}\right)\right)/\bar{z}_{i(n)}\right)\bar{Z}=\hat{R}_{mrss}\bar{Z} \tag{1}$$

where

$$\lambda=\left(S_{zi[n]ti[n]}/S_{zi[n]}^2\right),$$

$S_{zi[n]}^2$ is the sample variance of ancillary variable and $S_{zi[n]ti[n]}$ is the covariance between the ancillary variate and study variate.
The expression for the Mean Squared Error (MSE) of the equation (1) is obtained using the Taylor Series method defined as

$$h(\bar{z}_{in},\bar{t}_{in}) \cong h(\bar{Z},\bar{T}) + \frac{\partial h(u,v)}{\partial u}\Big|_{\bar{z}_{in},\bar{t}_{in}}(\bar{z}_{in}-\bar{Z}) + \frac{\partial h(u,v)}{\partial v}\Big|_{\bar{z}_{in},\bar{t}_{in}}(\bar{t}_{in}-\bar{T}) \tag{2}$$

where

$$h(\bar{z}_{in},\bar{t}_{in}) = \hat{R}_{mrss} \text{ and } h(\bar{Z},\bar{T}) = R$$

As shown in Wolter (1985), in equation (2) we may introduce the estimator given in equation (1), so that the MSE equation can be derived as follows. The difference between the estimator and the population ratio may be approximated by

$$\hat{R}_{mrss} - R \cong \frac{\partial\left(\bar{t}_{i[n]} + \lambda\left(\bar{Z} - \bar{z}_{i[n]}\right)\right)\big|\bar{z}_{i[n]}}{\partial \bar{z}_{i[n]}}\Big|_{\bar{Z},\bar{T}}\left(\bar{z}_{i[n]} - \bar{Z}_{i[n]}\right) + \frac{\partial\left(\bar{t}_{i[n]} + \lambda\left(\bar{Z} - \bar{z}_{i[n]}\right)\right)\big|\bar{z}_{i[n]}}{\partial \bar{t}_{i[n]}}\Big|_{\bar{Z},\bar{T}}\left(\bar{z}_{i[n]} - \bar{Z}_{i[n]}\right)$$

$$\cong -\left(\frac{\bar{t}_{i[n]}}{\bar{z}^2_{i[n]}} + \frac{\lambda \bar{Z}_{i[n]}}{\bar{z}^2_{i[n]}}\right)\Bigg|_{\bar{Z},\bar{T}}\left(\bar{z}_{i[n]} - \bar{Z}_{i[n]}\right) + \frac{1}{\bar{z}_{i[n]}}\Big|_{\bar{Z},\bar{T}}\left(\bar{t}_{i[n]} - \bar{T}_{i[n]}\right)$$

The expectation the square of the previuos formula is approximatelly

$$E\left(\hat{R}_{mrss} - R\right)^2 \cong \frac{\left(\bar{T}_{i[n]} + \theta \bar{Z}_{i[n]}\right)^2}{\bar{Z}^4_{i[n]}} V\left(z_{i[n]}\right) - \frac{2\left(\bar{T}_{i[n]} + \theta \bar{Z}_{i[n]}\right)}{\bar{Z}^3_{i[n]}} Cov\left(z_{i[n]}, t_{i[n]}\right) + \frac{1}{\bar{Z}^2_{i[n]}} V\left(t_{i[n]}\right),$$

as $E(\lambda) = \theta$ it is approximatelly

$$E\left(\hat{R}_{mrss} - R\right)^2 \cong \frac{1}{\bar{Z}^2_{i[n]}}\left\{\frac{\left(\bar{T}_{i[n]} + \theta \bar{Z}_{i[n]}\right)^2}{\bar{Z}^2_{i[n]}} V\left(z_{i[n]}\right) - \frac{2\left(\bar{T}_{i[n]} + \theta \bar{Z}_{i[n]}\right)}{\bar{Z}_{i[n]}} Cov\left(z_{i[n]}, t_{i[n]}\right) + V\left(t_{i[n]}\right)\right\}$$

where

$$\theta = \frac{S_{zi[n],ti[n]}}{S^2_{zi[n]}} = \frac{\rho S_{zi[n]} S_{ti[n]}}{S^2_{zi[n]}} = \frac{\rho S_{zi[n]} S_{ti[n]}}{S_{zi[n]}}.$$

Note that we omit the difference $E(\lambda) - \theta$, see a supporting discussion in Cochran (1977):
Let us calculate the MSE of $\bar{t}_{pr}$. It is

$$MSE\left(\bar{t}_{pr}\right) = \bar{Z}^2_{i[n]} E\left[\hat{R}_{mrss} - R\right]^2 \cong \left\{\frac{\left(\bar{T}_{i[n]} + \theta \bar{Z}_{i[n]}\right)^2}{\bar{Z}^2_{i[n]}} V\left(z_{i[n]}\right) - \frac{2\left(\bar{T}_{i[n]} + \theta \bar{Z}_{i[n]}\right)}{\bar{Z}_{i[n]}} Cov\left(z_{i[n]}, t_{i[n]}\right) + V\left(t_{i[n]}\right)\right\}$$

$$\cong \frac{\bar{T}^2_{i[n]} + 2\theta \bar{T}_{i[n]} \bar{Z}_{i[n]} + \theta^2 \bar{Z}^2_{i[n]}}{\bar{Z}^2_{i[n]}} V\left(z_{i[n]}\right) - \frac{2\bar{T}_{i[n]} + 2\theta \bar{Z}_{i[n]}}{\bar{Z}_{i[n]}} Cov\left(z_{i[n]}, t_{i[n]}\right) + V\left(t_{i[n]}\right)$$

$$\cong \frac{1-f}{n}\left\{\left(\frac{\bar{T}^2_{i[n]}}{\bar{Z}^2_{i[n]}} + \frac{2\theta \bar{T}_{i[n]}}{\bar{Z}_{i[n]}} + \theta^2\right) S^2_{zi[n]} - \left(\frac{2\bar{T}_{i[n]}}{\bar{Z}_{i[n]}} + 2\theta\right) S_{zi[n],ti[n]} + S^2_{ti[n]}\right\}$$

$$\cong \frac{1-f}{n}\left(R^2 S^2_{zi[n]} + 2\theta R S^2_{zi[n]} + \theta^2 S^2_{zi[n]} - 2R S_{zi[n],ti[n]} - 2\theta S_{zi[n],ti[n]} + S^2_{ti[n]}\right) \tag{3}$$

Al-Odat (2009) concluded that the condition $\rho \geq \left(2C_{zi}/C_{ti}\right)$ holds if $\rho \geq 0$ and $\rho \leq \left(2C_{zi}/C_{ti}\right)$ if $\rho \leq 0$ are satisfied. This fact sustains the claim that the suggested estimator of Al-Odat (2009) is more efficient than the traditional rank set ratio estimator when the conditions hold.

2. A SUGGESTED ESTIMATOR

M-estimation generalizes the method of maximum likelihood estimation (MLE): In real world datasets is possible to observe non-identical behaviors. Hence, a hypothetical parametric model described by a distribution function F_0 with density f_0, is not valid. The dataset is described, at least , by a contamination of underlying distributions. A solution is to look for a robust estimation

of the parameters of the underlying distribution. Characteristics and properties of the M-estimation method may be obtained in Andrews and Hampel (2015): The M-estimation method is commonly used for the estimation of location and scale parameters. Even when the data are contaminated with a few number of outliers we deal with a serious problem. See an illustrative discussion in Almongy et al. (2018): Robust estimation methods are not overly affected to outliers.
M-estimation was introduced by Huber (1964) generalizing maximum likelihood estimation in the context of location models. M-estimation produces robust estimators for the parameters of a probability density function $f_0(x;\boldsymbol{\theta})$: It looks for minimizing an objective function ρ with respect to $\boldsymbol{\theta}$. The optimization problem is

$$\widehat{\boldsymbol{\theta}} = argmin\left\{\sum_{i=1}^{n}\rho(x_i,\boldsymbol{\theta})\right\}$$

The aim of this study is to obtain robust estimators of a location parameter under RSS using M-estimation based on Huber proposal.
Motivated by the above derived properties of the estimator $\bar{t}_{pr}$ and keeping in mind the existence of outliers, we will mainly focus in getting more precise models, in terms of the resulting MSE . Then the goal of the research was to derive more accurate RSS alternative models when in the sample outliers are present. Our strategy was to use Huber M-Estimation tools instead of OLS in the estimator given by Al-Odat (2009): The expected robustness suggests that the resulting rank set sampling scheme will overcome OLS alternatives.
Note that MLE as a special case of M-estimation method for estimations parameters of a distribution function. As quoted OLS estimation is a MLE problem. It is highly sensitive to outliers (not robust against) . There is not a precise definition of an outlier. They are observations which do not follow the pattern of the other observations. This is not normally a problem when the outlier observation is simply an extreme observation drawn from the tail of a normal distribution $N(\mu,\sigma^2)$: However, if the outlier results from non-normal measurement error or is generated by another normal distribution $N(m,q^2)$ or by other violation of standard OLS assumptions a robust estimation technique should be used.
We propose the M-estimator

$$\bar{t}_{pr(Huber\,M)} = \left(\left(\bar{t}_{i(n)} + \lambda_{(Huber)}\left(\bar{Z} - \bar{z}_{i(n)}\right)\right)/\bar{z}_{i(n)}\right)\bar{Z} \qquad (4)$$

In this model $\lambda_{(Huber)}$ is obtained by using the so-called Huber M-Estimation method. Due to the robustness of Huber M-Estimation it advantages the least square estimation. As is well known M-estimators are not seriously affected by the presence of outliers. See discussions in Hampel et al. (1986), Almongy et al. (2018), Andrews and Hampel (2015): The knowledge of the role of robustness, for improving the efficiency of statistical models, sustains recommending looking for the presence of outliers in the observed data. That is, using some of the available statistical techniques for outlier detection. When outliers are detected in the data is logic to look for a Huber M-Estimation alternative to tools based on Least Square ones. That is what we are going to do, under a rank set sampling scheme.
M-estimation uses the function $\rho\left(e_{i[n]}\right)$. It is a mesure of the compromise between $e_{i[n]}^2$ and $\left|e_{i[n]}\right|$, where $e_{i[n]}$ is the error term in the regression model

$$t_{i[n]} = \alpha + \lambda z_{i[n]} + e_{i[n]},$$

α is the constant of the model. The Huber´s $\rho\left(e_{i[n]}\right)$ function has the form

$$\rho\left(e_{i[n]}\right) = \begin{cases} e_{i[n]}^2 & -M \le e_{i[n]} \le M \\ 2M\left|e_{i[n]}\right| - M^2 & -M \ or\ M < e_{i[n]} \end{cases}$$

M is a tuning constant that controls the robustness of the estimator. The value of the regression coefficient $\lambda_{(Huber)}$ is obtained by minimizing

$$\sum_{i=1}^{n} \rho\left(t_{i[n]} - \alpha - \lambda z_{i[n]}\right)$$

with respect to α *and* λ, see Birkes and Dodge (1993):
We remark that the MSE equation of the proposed rank set estimator given in (4) has the same form as the MSE equation in (3), but it is clear that θ is replaced by θ_{Huber}. The value of θ_{Huber} is obtained approximately from the approximation to the MSE

$$MSE\left(t_{pr(Huber)}\right) \cong \frac{1-f}{n}\left(R^2 S^2_{zi[n]} + 2\theta_{Huber} RS^2_{zi[n]} + \theta^2_{Huber} S^2_{zi[n]} - 2RS_{zi[n],ti[n]} - 2\theta_{Huber} S_{zi[n],ti[n]} + S^2_{ti[n]}\right)$$

It is well known that

$$E\left(\phi(e_{i[n]}\right) = 0,$$

where

$$\phi(e_{i[n]} = \rho'\left(e_{i[n]}\right)$$

The $e_{i[n]}$´s are identically independent distributed (iid):

3. THEORETICAL EFFICIENCY COMPARISON

In order to check theoretically the efficiency of the suggested estimator and the estimator of Al-Odat (2009) we compare their MSE-expressions given in (3) and (4): We look to establish conditions for satisfying the relationship:

$$MSE\left(t_{pr(Huber)}\right) < MSE\left(t_{pr}\right)$$

That is

$$\left(2\theta_{Huber} RS^2_{zi[n]} + \theta^2_{Huber} S^2_{zi[n]} - 2\theta_{Huber} S_{zi[n],ti[n]}\right) < \left(2\theta RS^2_{zi[n]} + \theta^2 S^2_{zi[n]} - 2\theta S_{zi[n],ti[n]}\right)$$

Doing some grouping we have

$$2RS^2_{zi[n]}\left(\theta_{Huber} - \theta\right) - 2S_{zi[n],ti[n]}\left(\theta_{Huber} - \theta\right) + S^2_{zi[n]}\left(\theta^2_{Huber} - \theta^2\right) < 0$$

Say that

$$\left(\theta_{Huber} - \theta\right)\left[2RS^2_{zi[n]} - 2S_{zi[n],ti[n]} + S^2_{zi[n]}\left(\theta_{Huber} + \theta\right)\right] < 0$$

For $\left(\theta_{Huber} - \theta\right) > 0$, that is $\theta_{Huber} > \theta$
Hence, we have that

$$\left[2RS^2_{zi[n]} - 2S_{zi[n],ti[n]} + S^2_{zi[n]}\left(\theta_{Huber} + \theta\right)\right] < 0$$

whenever

$$\left(\theta_{Huber} + \theta\right) < -2R + \left(2S_{zi[n],ti[n]} / S^2_{zi[n]}\right)$$

As $\frac{2S_{zi[n],ti[n]}}{S^2_{zi[n]}} > 0$ the inequaity holds only if

$$\theta_{Huber} < \theta - 2R.$$

Similarly, for $\left(\theta_{Huber} - \theta\right) < 0$, that is $\theta_{Huber} < \theta$, we have that

$$\theta_{Huber} > \theta - 2R$$

Consequently, we have the following conditions

$$0 < \theta_{Huber} - \theta < -2R \tag{5}$$

$$-2R < \theta_{Huber} - \theta < 0 \tag{6}$$

When the conditions given in (5) and (6) is satisfied, the proposed rank set estimator is more efficient that the suggested estimator proposed by Al-Odat in (2009) when outliers are present in the data.

4. NUMERICAL EFFICIENCY COMPARISON

The behaviour of the efficiency of the proposal was analysed using different data coming from real life problem. In each case was computed

$$\vartheta(\bar{t}_{pr(Huber)}, \bar{t}_{pr}) = MSE(\bar{t}_{pr(Huber)})/MSE(\bar{t}_{pr})$$

In all the cases the percent of outliers in the samples was denoted %(outliers)

Case 1. Analysis of the time to death of HIV infected persons

Bouza et al. (2019) analyzed a database of the quality of life of a set of 231 persons infected with HIV clustered by the risk-group.

- G1- Drug users
- G2- Bisexual-homo men
- G3- Bisexual-lesbian women
- G4- Hetero men
- G5- Hetero women
- G6- Contaminated by blood transfusions
- G7- Sons of VIH infected

The patients received a treatment for 5 years. The difference between the coefficient in the perceived quality of life of them was measured at the beginning of the retroviral treatment and after five years of continuous use. The results of the efficiency are given in the next table.

Table 1. Efficiency of the M-Estimator in RSS in 8 groups of persons infected with HIV: Variable quality of life

Group	%(outliers)	$\vartheta(\bar{t}_{pr(Huber)}, \bar{t}_{pr})$
1	4,1	0,80
2	2,1	0,73
3	2,4	0,84
4	7,0	0,59
6	9,4	0,67
6	1,2	0,93
7	5,7	0,77
8	11,2	0,24

Note than the efficiency is larger when the % of outliers increases.

Case 2. Bioleaching study

Al-Omari et al. (2016) obtained data on the contains of heavy metals. The data came from a study of leaching of elements from solid waste composts. The grab samples were prepared from multiple grab samples, using coning and quartering methods. The compost was collected from hospitals. The particles were mechanically separated and passed through a fine. Each batch, sent for burning, was evaluated in terms of its estimated toxicity. A qualification in the range 0-100 was given to each batch. For the experiment a sensor was placed in each chimney for measuring the contents in the smoke of plumb, magnesium, cadmium and the rest was classified as "other contaminants". The measurement was made for each batch introduced in the furnaces. The study was developed during six months. In the period were evaluated 1678 batches.

Table 2. Efficiency of the M-Estimator in RSS in 11 hospitals: heavy metals contents in solid compost.

Hospital	%(outliers)	$\vartheta(\bar{t}_{pr(Huber)}, \bar{t}_{pr})$	%(outliers)	$\vartheta(\bar{t}_{pr(Huber)}, \bar{t}_{pr})$	%(outliers)	$\vartheta(\bar{t}_{pr(Huber)}, \bar{t}_{pr})$	%(outliers)	$\vartheta(\bar{t}_{pr(Huber)}, \bar{t}_{pr})$
	plumb		magnesium		cadmium		other	contaminants
1	0,12	0,95	0,22	0,96	0,27	0,10	0,18	0,90
2	0,10	0,74	0,20	0,84	0,24	0,10	0,16	0,88
3	0,19	0,92	0,15	0,88	0,18	0,06	0,16	0,85
4	0,72	0,09	0,45	0,21	0,38	0,33	0,32	0,26
5	0,54	0,07	0,17	0,87	0,15	0,10	0,10	0,90
6	0,18	0,95	0,11	0,87	0,10	0,08	0,06	0,98
7	0,16	0,90	0,10	0,96	0,10	0,06	0,06	0,98

8	0,40	0,15	0,62	0,18	0,54	0,74	0,49	0,71
9	0,48	0,19	0,34	0,21	0,23	0,70	0,11	0,93
10	0,55	0,19	0,28	0,11	0,23	0,89	0,11	0,89
11	0,34	0,22	0,81	0,11	0,11	0,96	0,39	0,25

The study of the heavy metals, of a very different nature from the VIH-study, exhibits a similar behavior. For larger number of outliers, the efficiency of the M-Estimator increases seriously.

Case 3.

Covarrubias et al. (2017) obtained a database provided by a large study of biodiversity of beetles developed in Sierra Madre del Sur México during 2016. Samples were obtained in Disturbed Deciduous Tropical (DDT), Forest of Quercus-Conifers (FQC), Coniferous Forest (CF), Forest of Conifers-Quercus (FCQ) and Quercus Forest (QF): The behavior of 4 biodiversity indexes was studied. The results of the use of the models proposed in this paper are given in table 3.

Table 3. Efficiency of the M-Estimator in RSS in 5 regions: estimation of the biodiversity using 6 indexes.

region	%(outliers)	$\vartheta(\bar{t}_{pr(Huber)},\bar{t}_{pr})$	%(outliers)	$\vartheta(\bar{t}_{pr(Huber)},\bar{t}_{pr})$	%(outliers)	$\vartheta(\bar{t}_{pr(Huber)},\bar{t}_{pr})$	%(outliers)	$\vartheta(\bar{t}_{pr(Huber)},\bar{t}_{pr})$	%(outliers)	$\vartheta(\bar{t}_{pr(Huber)},\bar{t}_{pr})$
	SHANON		SIMPSONT		SIMPSON		FAYER		DCOVA	
DDT	0,38	0,16	0,36	0,59	0,78	0,23	0,21	0,50	0,59	0,34
FQC	0,11	0,91	0,29	0,46	0,26	0,81	0,21	0,60	0,59	0,34
FC	0,39	0,77	0,34	0,27	0,37	0,80	0,30	0,48	0,47	0,65
FCQ	0,44	0,62	0,35	0,24	0,34	0,79	0,29	0,59	0,12	0,87
QF	0,20	0,14	0,14	0,83	0,13	0,98	0,18	0,86	0,26	0,77

The efficacy of the M-method for the indexes estimation evaluated is very dispersed but the fact that high percentages of outliers determines that the proposed method is fairly better than OLS based ones remains its validity.

RECEIVED: FEBRUARY, 2021.
REVISED: MARCH, 2021.

REFERENCES

[1] ALMONGY, H., and ALMETWALY, E. (2018): Comparison between methods of robust estimation for reducing the effect of outliers. **The Egyptian Journal for Commercial Studies**, 4, 1-23.

[2] AL-ODAT, N. A. (2009): Modification in Ratio Estimator Using Rank Set Sampling**. European Journal of Scientific Research**, 29, 265-268.

[3] AL-OMARI, A.I. (2012): Ratio estimation of population mean using auxiliary information in simple random sampling and median ranked set sampling. **Statistics and Probability Letters**, 82, 1883-1990.

[4] AL-OMARI, A.I., and BOUZA, C.N. (2015): Ratio estimators of the population mean with missing values using ranked set sampling. **Econometrica**, 26, 67-76.

[5] AL-OMARI, A. I.; BOUZA, C. N.; COVARRUBIAS, D.; and PAL, R. (2016): A New Estimator of the Population Mean: An Application to Bioleaching Studies, **Journal of Modern Applied Statistical Methods,** 15, Article 9.DOI: 10.22237/jmasm/1478002020

[6] AL-OMARI, A.I., IBRAHIM, K. and JEMAIN, A.A. (2009): New ratio estimators of the mean using simple random sampling and ranked set sampling methods. **Revista Investigacion Operacional**, 30, 97-108.

[7] AL-OMARI, A.I., JABER, K. and IBRAHIM, A. (2008): Modified ratio-type estimators of the mean using extreme ranked set sampling, **Journal of Mathematics and Statistics**, 4, 150-155.

[8] AL-OMARI, A.I., and AL-NASSER, A.D. (2018): Ratio estimation using multistage median ranked set sampling approach. **Journal of Statistical Theory and Practice**, 12, 512-529.

[9] ANDREWS D.F, HAMPEL F. R.(2015): **Robust estimates of location: Survey and advances**. Princeton University Press, Princeton.

[10] BIRKES, D. and DODGE, Y.(1993): **Alternative Methods of regression**, John Wiley & Sons,, N. York.

[11] BOUZA. C.N., J. F. GARCÍA, G. K. VISHWAKARMA , S. M. and ZEESHAN (2016): Variance estimation of persons infected with AIDS under ranked set sampling, in "**Ranked Set Sampling. 65 Years Improving The Accuracy In Data Gathering**, 281-290. Elsevier, Amsterdam.

[12] COCHRAN, W. G. (1977): **Sampling Techniques**, 3rd Edition, John Wiley and Sons, New York.
[13] COVARRUBIAS MELGAR, D., C. BOUZA HERRERA and M. I. JEELANI BHAT (2017): An invariant diversity index. **Academic Web Journal of Scientific Research**, 2 1-11.
[14] HUBER, P. J. (1964): Robust Estimation of a Location Parameter. **The Annals of Mathematical Statistics**, 35, 73-101.
[15] HUBER P. J. (1992): Robust estimation of a location parameter. In **Breakthroughs in Statistics**, 492-518.. Springer. New York
[16] JUREÇKOVÁ, J. PICEK, and M**.** SCHINDLER (2019): **ROBUST STATISTICAL METHODS WITH R, 2nd Edition.** Chapman and Hall/CRC.
[17] MARONNA, R.A., MARTIN, R.D. and YOHAI, V.J. (2006): **Robust Statistics: Theory and Methods**. Wiley, New York.
[18] MCINTYRE, G. A. (1952): A Method of Unbiased Selective Sampling Using Ranked Set. **Australian Journal of Agricultural Research,** 3, 385-390.
[19] SAMAWI, H. M. and MUTTLAK, H. A. (1966): Estimation of ratio using Rank set sampling, **Biometrical Journal**, 38, 753-764.
[20] SUBZAR, M., BOUZA, C. N., MAQBOOL, S., RAJA, T. A. and PARA, B. (2019a): Robust ratio type estimators in simple random sampling using Huber M Estimation. **Revista Investigación Operacional**, 40, 201-209.
[21] SUBZAR, M., BOUZA, C. N. and AL OMARI A. I. (2019b): Utilization of different robust regression techniques for estimation of finite population mean in SRSWOR in case of presence of outliers through ratio method of estimation. **Revista Investigacion Operacional**, 40, 600-609.
[22] WOLTER, K. M. (1985): **Introduction to Variance estimation**, Springer-Verlag, Berlin.