ON THE USE OF FAMILIES OF EXPONENTIAL- TYPE ESTIMATORS FOR COMPOSITE IMPUTATION FOR ADJUSTING MISSING DATA

Ajeet Kumar Singh*, Upendra Kumar**, Rajpal*** and V.K. Singh****
*Department of Statistics, University of Rajasthan, Jaipur
**Department of Statistics, U.P.A. College, Varanasi
*** Department of Mathematics and Statistics, MLSU, Udaipur
**** Department of Statistics, Banaras Hindu University, Varanasi

ABSTRACT
The aim of the paper is to suggest some composite methods of imputation (CMI) for filling - in the missing information in the sampled data. These strategies have been developed with the use of an auxiliary variable and with the use of some functions of such a variable in defining some families of exponential type estimators (ETEs). The bias and mean square error of the suggested strategies have been obtained and their particular cases have also been dealt with. A study has been made to compare the performance of all the strategies with each other. Further, in each family of estimators, the optimum estimators have been searched. The results so obtained have been testified on the basis of some empirical data.

KEYWORDS: Bias, mean square error, percentage relative efficiency, imputation.

MSC: 62D05

RESUMEN
El objetivo del artículo es sugerir algunos métodos compuestos de imputación (CMI) para completar la información faltante en los datos muestreados. Estas estrategias se han desarrollado con el uso de una variable auxiliar y con el uso de algunas funciones de dicha variable para definir algunas familias de estimadores de tipo exponencial (ETE). Se ha obtenido el sesgo y el error cuadrático medio de las estrategias sugeridas y también se han abordado sus casos particulares. Se ha realizado un estudio para comparar el rendimiento de todas las estrategias entre sí. Además, en cada familia de estimadores se han buscado los estimadores óptimos. Los resultados así obtenidos han sido atestiguados sobre la base de algunos datos empíricos.

PALABRAS CLAVE: Sesgo, error cuadrático medio, eficiencia relativa porcentual, imputación.

1. INTRODUCTION

The problem of missing value is a common aspect in almost all types of surveys. Indeed, often some sampling units do not respond in the study or refuse to answer all the questionnaires, the interviewer is not able to contact with all sampling units, or they are accidental loss of information caused by unknown factors. Such non - response not only mean less efficient estimates because of reduced sample size, but also traditional survey sampling methods cannot be immediately used to analyse data in hand. Imputation is one of the way of handling non - response where by missing values on one or more study variables are imputed (filled - in) with some substitutes. It is actually applied to compensate for non - response in sample surveys Kalton et al (1981); Sedransk (1985); Rubin (1986). Singh and Horn (2000) suggested CMI. It can be viewed as the imputation that combines at least two different methods to form a new one. Bouza, C. N. and Omari, A , I. AL. (2012); Omari, A, I. AL. and Bouza, C.N. (2014), Bouza, C. N. (2002A and 2002B), Herrera, Bouza et al (2021), Bhushan, S., and Pandey, A.P. (2018), and Singh et al (2022) used information from imputed values for the responding units in addition to non-responding units. Srinath, K. P. (1971) described multiphase sampling in missing problem.

The purpose of the paper is to develop some imputation strategies for filling - in the missing values in the sampled data and to suggest corresponding estimators for estimating the population mean. The suggested strategies have been developed by using the ETEs propounded by Bahl and Tuteja (1991) which utilize the information on an auxiliary variable at the estimation stage. The strategies are in the form of class of estimators which enable us to make a comprehensive study of a number of ratio, product type estimators belonging to the classes. Different properties of the families have been studied. Theoretically the families have been compared for their performances. Motivated by Singh et al (2014 a, 2014b), we develop some efficient imputation strategies on the basis of families of ETEs, which may be looked upon as extensions of the works.

2. INITIATION OF THE PROBLEM AND NOTATIONS

As discussed above, the aim of the work is to suggest some new imputation techniques to substitute the missing values in the sampled data and thereby, making the data set complete. The imputed values have been selected on the basis of observed values and not employing any kind of theoretical models. However, it has been assumed

that the missing pattern in the sample of a given size follows the concept of Missing Completely at Random (MCAR).

Let us assume that a finite population $U = (Y_1, Y_2, ..., Y_N)$ of size N exists which suffers from the non - response. Let the characteristic under study be Y and the information on an auxiliary variable X be available in the population. Let a simple random sample of size n be drawn from the population in order to estimate the population mean. Let the sample consists of r responding units (r < n) and (n - r) non - responding units.

We denote the population by Ω and the sample of size n by s. Further, let the set of responding units be denoted by A and that of non - responding units by A^C such that s = A $\cup$ A^C. For each unit $i \in A$, the value y_i is observed and for the unit $i \in A^C$, the value y_i is missing for which suitable imputed value is to be derived. For this, the values of auxiliary variable are used as a source of imputation.

3. SOME EXISTING STRATEGIES

We have discussed some well-known imputation strategies which has a direct relation with the current work. Let [D, T] be a sampling strategy where D stands for (SRSWOR) scheme and T stands for an estimator. We have discussed followings imputation methods, their corresponding sampling strategies, bias and MSE :

3.1. $[D, \bar{y}_r]$*: Mean Method of Imputation (MMI)*

$$y_{.i} = \begin{cases} y_i & \text{if i} \in \text{A} \\ \bar{\text{y}}_{\text{r}} & \text{if i} \in \text{A}^{\text{C}} \end{cases}. \tag{1}$$

Since the estimator of the population mean is

$$\bar{y}_n = n^{-1}\left[\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} y_i\right], \tag{2}$$

therefore, under $[D, \bar{y}_r]$, the point estimator becomes

$$\bar{y}_M = \bar{y}_r = r^{-1}\sum_{i \in A} y_i \,. \tag{3}$$

The bias, B(.) and MSE, M(.) of $\bar{y}_M$ are

$$B(\bar{y}_M) = 0\,, \tag{4}$$

$$M(\bar{y}_M) = V[\bar{y}_M] = \theta_{r,N}\bar{Y}^2 C_Y^2\,. \tag{5}$$

3.2. $[D, \bar{y}_{RAT}]$*: Ratio Method of Imputation (RMI)*

Under RMI method, the imputation scheme is given as

$$y_{.i} = \begin{cases} y_i & \text{if i} \in \text{A} \\ \hat{\text{b}}\text{x}_{\text{i}} & \text{if i} \in \text{A}^{\text{C}}, \end{cases} \tag{6}$$

where

$$\hat{b} = \frac{\bar{y}_r}{\bar{x}_r}\,. \tag{7}$$

It is clear that, in this method, the imputation is carried out with the aid of an auxiliary variable X, such that the data $x_s = \{x_i, i \in s\}$ are known.

Accordingly, the point estimator for population mean becomes

$$\bar{\text{y}}_{\text{RAT}} = \bar{\text{y}}_{\text{r}}\frac{\bar{\text{x}}_{\text{n}}}{\bar{\text{x}}_{\text{r}}}\,. \tag{8}$$

Further, we get

$$B(\bar{y}_{RAT}) = \theta_{r,n}\bar{Y}\left[C_X^2 - \rho C_X C_Y\right], \tag{9}$$

and

$$M(\bar{y}_{RAT}) = \theta_{r,N}\bar{Y}^2 C_Y^2 + \theta_{r,n}\bar{Y}^2\left[C_X^2 - 2\rho_{XY} C_X C_Y\right]. \tag{10}$$

3.3. $[D, \bar{y}_{COMP}]$: *Compromized Method*

This method was suggested by Singh and Horn (2000). The scheme of imputation is

$$y_{.i} = \begin{cases} p\frac{n}{r}y_i + (1-p)\hat{b}x_i & \text{if } i \in A \\ (1-p)\hat{b}x_i & \text{if } i \in A^C, \end{cases} \tag{11}$$

where p is a unknown constant, such that the variance of $\bar{y}_{COMP}$ is optimum. Meeden (2000) also has developed the idea of adjusting observing values in addition to non-responding values while doing imputation.
Under this scheme, the point estimator of $\bar{Y}$ becomes

$$\bar{y}_{COMP} = p\bar{y}_r + (1-p)\bar{y}_r \frac{\bar{x}_n}{\bar{x}_r}. \tag{12}$$

The bias and MSE of $\bar{y}_{COMP}$ are obtained as

$$B[\bar{y}_{COMP}] = (1-p)\theta_{r,n}\bar{Y}[C_X^2 - \rho C_X C_Y], \tag{13}$$

$$M[\bar{y}_{COMP}] = \theta_{r,N}\bar{Y}^2 C_Y^2 + \theta_{r,n}\bar{Y}^2[(1-p)^2 C_X^2 - 2(1-p)\rho C_X C_Y]. \tag{14}$$

Remark 1: If the MSE of $\bar{y}_{COMP}$ is optimized with respect to the constant p for obtaining its optimum value p_0 ,

$$\text{we get } p_0 = 1 - \rho\frac{C_Y}{C_X}. \tag{15}$$

and then

$$M[\bar{y}_{COMP}]_{\min} = \bar{Y}^2[(\theta_{r,N} - \theta_{r,n}\rho^2)C_Y^2]. \tag{16}$$

3.4. $[D, \bar{y}_{SD}]$: *Singh and Deo (2003)*

The point estimator, B(.) and MSE(.) are

$$\bar{y}_{SD} = \bar{y}_r\left(\frac{\bar{x}}{\bar{x}_r}\right)^{\gamma} \tag{17}$$

$$B(\bar{y}_{SD}) = \theta_{r,n}\bar{Y}\left(\gamma(\gamma-1)\frac{C_X^2}{2} - \gamma\rho C_X C_Y\right) \tag{18}$$

$$MSE(\bar{y}_{SD})_{\min} = MSE(\bar{y}_{RAT}) - \theta_{r,n}S_X^2(B-R)^2 \tag{19}$$

3.5. $[D, \bar{y}_{pr_i}]$: *Kadilar and Cingi (2008) MI for (i=1,2,3)*

The point estimators are

$$\bar{y}_{pr_1} = \frac{\bar{y}_r + b(\bar{X} - \bar{x}_n)\bar{X}}{\bar{x}_n} \tag{20}$$

$$\bar{y}_{pr_2} = \frac{\bar{y}_r + b(\bar{X} - \bar{x}_r)\bar{X}}{\bar{x}_r} \tag{21}$$

$$\bar{y}_{pr_3} = \frac{\bar{y}_r + b(\bar{x}_n - \bar{x}_r)\bar{x}_n}{\bar{x}_r} \tag{22}$$

Biases and MSEs under this MI is

$$B(\bar{y}_{pr_1}) = \theta_{n,N}\bar{Y}C_X^2 \tag{23}$$

$$B(\bar{y}_{pr_2}) = \theta_{r,N}\bar{Y}C_X^2 \tag{24}$$

$$B(\bar{y}_{pr_3}) = \theta_{r,n}\bar{Y}\rho C_X C_Y \tag{25}$$

$$MSE(\bar{y}_{pr_1}) = \theta_{r,N}S_Y^2 + \theta_{n,N}S_X^2(R^2 - B^2) \tag{26}$$

$$MSE(\bar{y}_{pr2}) = \theta_{r,N}(S_Y^2 + R^2S_X^2 - BS_{XY}) \tag{27}$$

$$MSE(\bar{y}_{pr3}) = \theta_{r,N}\, S_Y^2 + \theta_{r,n}\,(B+R)^2 S_X^2 - 2(B+R)S_{XY} \tag{28}$$

4. PROPOSED FAMILIES OF IMPUTATION STRATEGIES

4.1. Motivated by Bahl and Tuteja (1991) and Singh *et al* (2014a), below we define three CMI :

$$\text{(i) } y_{.i} = \begin{cases} k\dfrac{n}{r}\bar{y}_r + (1-k)\bar{y}_r\phi(\alpha, \bar{X}, \bar{x}_n) & if\ \text{i} \in \text{A} \\ (1-k)\bar{y}_r\phi(\alpha, \bar{X}, \bar{x}_n) & if\ \text{i} \in \text{A}^{\text{C}} \end{cases} \tag{29}$$

$$\text{(ii) } y_{.i} = \begin{cases} k\dfrac{n}{r}\bar{y}_r + (1-k)\bar{y}_r\phi(\alpha, \bar{x}_n, \bar{x}_r) & if\ \text{i} \in \text{A} \\ (1-k)\bar{y}_r\phi(\alpha, \bar{x}_n, \bar{x}_r) & if\ \text{i} \in \text{A}^{\text{C}} \end{cases} \tag{30}$$

and

$$\text{(iii) } y_{.i,} = \begin{cases} k\dfrac{n}{r}\bar{y}_r + (1-k)\,\bar{y}_r\,\phi(\alpha, \bar{X}, \bar{x}_r) & if\ \text{i} \in \text{A} \\ (1-k)\,\bar{y}_r\,\phi(\alpha, \bar{X}, \bar{x}_r) & if\ \text{i} \in \text{A}^{\text{C}}, \end{cases} \tag{31}$$

where k and α are constants, the values of which may be either assumed or may be obtained under certain conditions.

It can be seen that with these imputations methods for filling - in the missing sampled values, the corresponding point estimators are obtained as

$$\text{(i)} \qquad T_{E_1} = k\bar{y}_r + (1-k)\bar{y}_r\,\phi(\alpha, \bar{X}, \bar{x}_n)\,, \tag{32}$$

$$\text{(ii)} \qquad T_{E_2} = k\bar{y}_r + (1-k)\bar{y}_r\phi(\alpha, \bar{x}_n, \bar{x}_r)\,, \tag{33}$$

$$\text{(iii)} \qquad T_{E_3} = k\bar{y}_r + (1-k)\bar{y}_r\,\phi(\alpha, \bar{X}, \bar{x}_r) \tag{34}$$

Thus, we have imputation strategies *[* D, T_{E_1}], [D, T_{E_2}] and [D, T_{E_3}].

Remark 2: It is clear that when α = -1 and 1 in (32), (33) and (34), we get compromized imputation strategies with exponential - type product and exponential – type ratio estimators which might be applicable if the population exhibits a negative and positive correlation.

Remark 3: It is further noted that $[D, T_{E_i}]$ for i = 1, 2, 3 reduces to mean method of imputation, $[D, \bar{y}_r]$, if k = 1 and reduces to exponential - type method of imputation, $[D, T_{E_i}]$ if k = 0. Therefore, $[D, T_{E_i}]$ is the generalization of these two imputation strategies.

5. BIAS AND MSE OF THE STRATEGIES $[D, T_{E_i}]$ FOR i = 1, 2, 3

5.1. Biases of the strategies

The bias, B(.) of the suggested strategies $[D, T_{E_i}]$, i=1, 2, 3 can be easily obtained upto the first order of approximations under large sample theory, as has been mentioned in the section A of appendix. Consequently, we have given properties:

Theorem 1: The bias of T_{E_1} upto the order $O(n^{-1})$ is as under

$$B\,[T_{E_1}] = (1-k\ )\bar{Y}\,\theta_{n,N}\left[\frac{\alpha}{4}C_X^2 + \frac{\alpha^2}{8}C_X^2 - \frac{\alpha}{2}\rho C_Y C_X\right] \tag{35}$$

Theorem 2: The bias of T_{E_2} upto the first order of approximation is as under

$$B[T_{E_2}] = (1-k)\bar{Y}\,\theta_{r,n}\left[\frac{\alpha}{4}C_X^2 + \frac{\alpha^2}{8}C_X^2 - \frac{\alpha}{2}\rho C_Y C_X\right] \tag{36}$$

Theorem 3: The bias of T_{E_3} upto the 1st order of approximation is given by

$$B[T_{E_3}] = (1-k)\bar{Y}\theta_{r,N}\left[\frac{\alpha}{4}C_X^2 + \frac{\alpha^2}{8}C_X^2 - \frac{\alpha}{2}\rho C_Y C_X\right] \tag{37}$$

5.2. MSEs of the strategies

Similarly, the MSEs of the strategies $[D, T_{E_i}]$, i=1, 2, 3 can be obtained upto the term $O(n^{-1})$, using the large sample theory, as

Theorem 4: Upto the first order $O(n^{-1})$, the MSE of T_{E_1} is given by

$$M(T_{E_1}) = \theta_{r,N}\bar{Y}^2 C_Y^2 + \theta_{n,N}\bar{Y}^2\left[(1-k)^2\frac{\alpha^2}{4}C_X^2 - \alpha(1-k)\rho C_X C_Y\right]. \tag{38}$$

Further, for the MSEs of T_{E_2} and T_{E_3}, we have the following theorems:

Theorem 5: The MSE of T_{E_2} is obtained as

$$M(T_{E_2}) = \theta_{r,N}\bar{Y}^2 C_Y^2 + \theta_{r,n}\bar{Y}^2\left[(1-k)^2\frac{\alpha^2}{4}C_X^2 - \alpha(1-k)\rho C_X C_Y\right]. \tag{39}$$

Theorem 6: The MSE of T_{E_3} under the large sample of approximation is given by

$$M(T_{E_3}) = \theta_{r,N}\bar{Y}^2 C_Y^2 + \theta_{r,N}\bar{Y}^2\left[(1-k)^2\frac{\alpha^2}{4}C_X^2 - \alpha(1-k)\rho C_X C_Y\right]. \tag{40}$$

The Proof of MSEs of T_{E_i} i=1,2,3 is given in appendix of section A.

6. OPTIMALITY CONDITIONS AND OPTIMUM ESTIMATORS

6.1. It is seen that the estimators T_{E_i} (i =1, 2, 3) are functions of the parameters α and k, whose values may be either known or may be obtained under certain conditions. Obviously, the values of α and k must be chosen in such a way that the corresponding MSEs of the estimators are minimum. Since the parameters are intermingled in the expressions of the MSEs, therefore, minimizing the MSEs simultaneously with respect to both α and k will not yield explicit solutions for the parameters. It is, therefore, advisable to choose a suitable value of one of the parameters and to find the optimum value of the other. We shall now present the optimum value of one parameter, keeping the value of the other, and corresponding optimum MSE of T_{E_i} (i=1, 2, 3).

6.2. Differentiating the expression (38) with respect to k, assuming α to be known and equating the result, so obtained, to zero, we get the minimum value of k (say, $(k_0)_1$) for fixed α as

$$(k_0)_1 = 1 - \frac{1}{\alpha}2\rho\frac{C_Y}{C_X} \tag{41}$$

for which the minimum MSE of the estimator T_{E_1} reduces to

$$M(T_{E_1})_{(k_0)_1.} = \bar{Y}^2 C_Y^2[\theta_{r,N} - \theta_{n,N}\rho^2], \tag{42}$$

which is same as $M[\bar{y}_{T_1}]_{\min}$.

Further, minimizing the MSE of T_{E_1} with respect to α for a given value of k, we get optimum α as

$$(\alpha_0)_1 = \frac{2}{1-k}\rho\frac{C_Y}{C_X}, \tag{43}$$

for which

$$M(T_{E_1})_{(\alpha_0)_1.} = \bar{Y}^2 C_Y^2[\theta_{r,N} - \theta_{n,N}\rho^2]. \tag{44}$$

6.3. Similar treatment with the expression (39) yields the following results:

$$(k_0)_2 = 1 - \frac{2}{\alpha}\rho\frac{C_Y}{C_X}, \tag{45}$$

for which

$$M\left(T_{E_2}\right)_{(k_0)_{2.}} = \bar{Y}^2 C_Y^2[\theta_{r,N} - \theta_{r,n}\rho^2]. \tag{46}$$

$$(\alpha_0)_2 = \frac{2}{1-k}\rho\frac{C_Y}{C_X}, \tag{47}$$

for which

$$M\left(T_{E_2}\right)_{(\alpha_0)_2} = \bar{Y}^2 C_Y^2[\theta_{r,N} - \theta_{r,n}\rho^2]. \tag{48}$$

6.4. For the expression (40), we get the similar results as

$$(k_0)_3 = 1 - \frac{2}{\alpha}\rho\frac{C_Y}{C_X}, \tag{49}$$

which yields

$$M\left(T_{E_3}\right)_{(k_0)_{3.}} = \bar{Y}^2\theta_{r,N}C_Y^2[1-\rho^2]; \tag{50}$$

and

$$(\alpha_0)_3 = \frac{2}{1-k}\rho\frac{C_Y}{C_X}, \tag{51}$$

with

$$M\left(T_{E_3}\right)_{(\alpha_0)_3} = \bar{Y}^2 C_Y^2\theta_{r,N}[1-\rho^2]. \tag{52}$$

Remark 4: It is worthwhile here to mention that the parameter α characterizes the particular family of estimators while the parameter k attaches weights to the used estimators for compromization. Three important ETE(s) are of quite importance, namely, exponential type ratio estimator, for α = 1; exponential type product estimator for α = -1 and sample mean estimator, $\bar{y}_r$ for α = 0. It is, therefore, desirable to utilize any of them at a time for a given population in which the range of the quantity $\rho\frac{C_Y}{C_X}$ is guessed. However, the CMI will be certainly better than other imputation methods if the value of the weight k is wisely and correctly chosen. It is, appropriate to minimize the MSE of the estimators with respect to k after making a choice of α**.** Since, as observed above, the minimum MSE would remain same under k_0 and α_0, the result is unaffected even if k_0 is obtained for a given α**.**

7. COMPARISONS OF DIFFERENT STRATEGIES

7.1. Let us first compare the strategies [D, T_{E_1}] , [D, T_{E_2}] and [D, T_{E_3}] under the optimality condition.

(a) we have $$M\left(T_{E_2}\right)_{(\alpha_0)_2} - M\left(T_{E_1}\right)_{(\alpha_0)_1} = \bar{Y}^2\rho^2 C_Y^2\left(\theta_{n,N} - \theta_{r,n}\right) . \tag{53}$$

Thus, T_{E_1} would be preferable over T_{E_2} if

$$r > \frac{n}{2-f} \quad \text{where } f = \frac{n}{N} \ (0 < f < 1), \tag{54}$$

that is, when the non -response in the sample is observed to be less than fifty percent. In any kind of survey, it generally holds, and, therefore, it could be expected that the strategy T_{E_1} would be fairly better than strategy T_{E_2} under the optimality conditions, in most of the survey situations.

(b) We now compare $M\left(T_{E_1}\right)_{(\alpha_0)_1} - M\left(T_{E_3}\right)_{(\alpha_0)_3} = \theta_{r,n}\rho^2 C_Y^2\,\bar{Y}^2 > 0$

which is always true. This implies that the is better than [D, T_{E_1}]. [D, T_{E_3}] would be always preferable over [D, T_{E_1}]as far as the optimum estimators under both the strategies are concerned.

(c) Similarly comparing the strategy $[D, T_{E_3}]$ with $[D, T_{E_2}]$, we observe that $[D, T_{E_2}]$ - $[D, T_{E_3}]$ is

$$\rho^2 C_Y^2 \, \bar{Y}^2 (\theta_{r,N} - \theta_{r,n}). \tag{55}$$

(55) is always positive, implying that strategy $[D, T_{E_3}]$ would be always better than the strategy $[D, T_{E_2}]$ under optimality conditions.

Considering the results obtained under (a), (b) and (c) above, finally one can conclude that under the optimality condition, we get

$$M[T_{E_3}] \le M[T_{E_1}] \le M[T_{E_2}] \tag{56}$$

if the sample does not have more than 50% information.

7.2. Since for the proposed strategies, the value of the parameter α may assume positive as well as negative values, the comparison of the three strategies for an arbitrary choice of α will include a number of conditions for the choice of α so that one strategy would be preferable over the others. Hence, it would be of no use to compare the strategies theoretically, rather these might be compared on the basis of some empirical data. Due to this reason, for the comparison purpose, it is always preferable to select the choice of α as α_0 everywhere.

7.3. We shall present a comparison of the proposed strategies with $\bar{y}_r$, $\bar{y}_{RAT}$, $\bar{y}_{COMP}$. Since in all the strategies, the correlation coefficient ρ and the parameter, α may assume either positive and negative values, therefore, the comparisons should be made particularly, when (i) $\rho > 0$ and $\alpha > 0$ and (ii) $\rho < 0$ and $\alpha < 0$. For the illustration purpose, we consider the case (i) only.

(a) We observe that $M[T_{E_1}] < M[\bar{y}_M]$ if

$$\theta_{n,N} \bar{Y}^2 \left[(1-k)^2 \frac{\alpha^2}{4} C_X^2 - \alpha (1-k) \rho C_X C_Y \right] < 0 \tag{57}$$

which happens when $\alpha > 0$ and $\alpha < 2(\alpha_0)_1$.

(b) $M[T_{E_1}] < M[\bar{y}_{RAT}]$ if

$$\theta_{n,N} \bar{Y}^2 \left[(1-k)^2 \frac{\alpha^2}{4} C_X^2 - \alpha (1-k) \rho C_X C_Y \right] - \theta_{r,n} [C_X^2 - 2\rho C_Y C_X] < 0 \tag{58}$$

for which a suitable range of α can be deduced for a given set of data. Similarly, ranges of the parameter α (> 0) for positive correlation can be obtained so that $M[T_{E_1}] < M[\bar{y}_{COMP}]$.

(c) With similar steps as mentioned in (a) and (b) above, a comparison of the strategies $[D, T_{E_2}]$ and $[D, T_{E_3}]$ with $[D, \bar{y}_r]$, $[D, \bar{y}_{RAT}]$ and $[D, \bar{y}_{COMP}]$ may be made with a view to find the suitable ranges of α in which the proposed strategies would be preferable over the existing strategies.

7.4. Now we shall consider the minimum MSEs of the suggested strategies for a comparison of them with the existing strategies $[D, \bar{y}_r]$, $[D, \bar{y}_{RAT}]$ and $[D, \bar{y}_{COMP}]$. We have

(i) $M[T_{E_1}]_{Min} < M[\bar{y}_M]$ if $\theta_{n,N} \rho^2 > 0$ (59)

which is always true.

(ii) $M[T_{E_1}]_{Min} < M[\bar{y}_{RAT}]$ if $r > \dfrac{n}{2-f}$, a condition similar to (54).

Thus, in almost all the practical situations, where the part of universe does not exhibit more than 50% non respondent units, $M[\bar{y}_{T_1}]_{Min} < M[\bar{y}_{RAT}]$.

(iii) Comparing the equations (16) and (52), it is seen that $M[T_{E_1}]_{Min} < M[\bar{y}_{COMP}]_{Min}$. if $r > \dfrac{n}{2-f}$.

(iv) $M[T_{E_2}]_{Min} < M[\bar{y}_M]$ if $-\theta_{r,n} \rho^2 < 0$, which is always true.

(v) $M[T_{E_3}]_{Min} < M[\bar{y}_{RAT}]$ if $\theta_{r,n} \bar{Y}^2 (C_X - \rho C_Y)^2 > 0$, which is always true.

(vi) A comparison of $M[T_{E_2}]_{Min}$ with $M[\bar{y}_{COMP}]_{Min}$ reveals that both are equally efficient.

(vii) $M[T_{E_3}]_{Min} < M[\bar{y}_M]$ if $-\theta_{r,N}\rho^2 < 0$, a condition which always holds.

(viii) $M[T_{E_3}]_{Min} < M[\bar{y}_{RAT}]$ if $\theta_{r,n}(C_X - \rho C_Y)^2 + \theta_{n,N}\rho^2 C_Y^2 > 0$, which always holds.

(ix) $M[T_{E_3}]_{Min} < M[\bar{y}_{COMP}]_{Min}$ if $\theta_{n,N} > 0$, which is trivial.

8. EFFICIENCY COMPARISONS AND EMPIRICAL STUDY

For elaborating the theoretical results obtained in the previous sections, we shall utilize six empirical data which are as follows:

Population I: The data have been taken by Ahmed *et al.* (2006), from the Department of Statistics (Jordan), Healthcare Utilization and Expenditure Survey, 2000. We consider the variables Y(household expenditure) and X (household income). For the given population, we have the following values:

N=8306, $n = 200$, $r = 180$, $\bar{Y}$ =253.75, $\bar{X}$ =343.316, S_X^2 =862017, S_Y^2 =338006, $\rho = 0.5222$, C_X = 2.7044, C_Y = 2.2912.

Population II: We have taken data from the work of Kadilar and Cingi (2008) .the variable Y(level of apple) and X (the number of apple trees) . The values of universe parameters are:

$N = 19$, $n = 10$, $r = 8$, $\bar{Y}$ = 575, $\bar{X}$ =13573.68, S_X =12945.38, S_Y =858.36, $\rho = 0.88$, $C_X = 0.954$, $C_Y = 1.943$.

Population III: This population has been borrowed from the article of Diana and Perri (2010) by a market research company. The variable Y and X represent the sale area (in square meters) and the number of employees respectively. The values of population are:

$N = 2376$, $n = 300$, $r = 100$, $\bar{Y}$ =1701.95, S_Y =2195.52, $\bar{X}$ =40.62, S_X = 95.46, $\rho = 0.90$, $C_X = 2.345$, $C_Y = 1.29$.

Population IV: This population has been borrowed from the work of Shukla *et al* (2011b). Y being the study variable and X being the auxiliary variable. The values of required population parameters were obtained as follows:

N = 200, n =20 and r =15, $\bar{Y}$ =42.485, $\bar{X}$ = 18.515, S_Y = 14.1088, S_X = 6.9669, $\rho = 0.8652$.

Population V: This data has been taken from the paper of Diana and Perri (2010). The household net disposable income Y and the number of household income - earners X were investigated.

N = 8011, n =400 and r =250, $\bar{Y}$ =28229.43, $\bar{X}$ = 1.69, S_Y = 22216.56, S_X = 0.78, $\rho = 0.46$.

Population VI: This population has been taken from Mukhopadhayay (2000). The data are related to the number of labourers, X (in thousands) and quantity of raw materials, Y (in lakhs of bales) in 20 jute mills. We take a random sample of size 7 and took another sample from this sample to constitute the respondent sample of size 5. For the given population, we have the following values:

$\bar{Y}$ =41.5, $\bar{X}$ =441.95, S_X^2 = 10215.21 , S_Y^2 = 95.7368 , $\rho = 0.6521$.

The absolute value of minimum MSE of the suggested strategies and of the strategies $[D, \bar{y}_r]$, $[D, \bar{y}_{RAT}]$, $[D, \bar{y}_{COMP}]$, $[D, \bar{y}_{SD}]$, $[D, \bar{y}_{pr_1}]$,$[D, \bar{y}_{pr_2}]$ and $[D, \bar{y}_{pr_3}]$ are depicted in Table 1. The table also depicts the (PREs) of different strategies with respect to the strategy $[D, \bar{y}_r]$, where $\mathrm{PRE}[D, T] = \dfrac{M[\bar{y}_r]}{M[T]} \times 100$

8.1. Efficiency Comparisons

We have presented a comparison of different imputation strategies with the proposed strategies under respective optimal conditions. Table 1, and 2 present the MSEs and PREs of the strategies for population I, II, III , IV,V and VI and Table 3,4 and 5 present the MSEs and PREs of the strategies for population I,III and V for different non-response rate.

Table 1. MSE and PRE of the Strategies under Optimality Conditions

Strategy	Population I		Population II		Population III	
	M(.)	PRE	M(.)	PRE	M(.)	PRE
$[D, \bar{y}_r]$	1837.977	100	53319.74	100	46174.14	100
$[D, \bar{y}_{RAT}]$	1868.058	98.38969	40124.87	132.8845	47215.68	97.79409

$[D, \bar{y}_{COMP}]$	1786.746	102.8673	39055.64	136.5225	20144.59	229.2137
$[D, \bar{y}_{SD}]$	1867.977	98.39398	39053.93	136.5285	19968.68	231.2328
$[D, \bar{y}_{pr_1}]$	3686.14	49.86183	40530.19	131.5556	81231.33	56.84277
$[D, \bar{y}_{pr_2}]$	4397.042	41.8003	33805.11	157.7269	161355.6	28.61639
$[D, \bar{y}_{pr_3}]$	1838.187	99.98855	46573.78	114.4845	126683.9	36.44831
$[D, T_{E_1}]$	1388.004	132.4187	26293.03	202.7904	34802.64	132.6742
$[D, T_{E_2}]$	1786.746	102.8673	39055.64	136.5225	20144.59	229.2137
$[D, T_{E_3}]$	1336.774	137.4935	12028.93	443.2624	8773.087	526.3158

Table 2. MSE and PRE of the Strategies under Optimality Conditions

Strategy	Population IV		Population V		Population VI	
	M(.)	PRE	M(.)	PRE	M(.)	PRE
$[D, \bar{y}_r]$	12.26868	100	1912690.17	100	14.38837	100
$[D, \bar{y}_{RAT}]$	10.02456	122.3862	1767858.792	108.1925	12.60676	114.1321
$[D, \bar{y}_{COMP}]$	9.786525	125.363	1756029.275	108.9213	12.05754	119.3309
$[D, \bar{y}_{SD}]$	9.786487	125.3635	1756020.12	108.9219	12.05683	119.3379
$[D, \bar{y}_{pr_1}]$	17.06159	71.90819	2067794.888	92.49903	18.95892	75.89232
$[D, \bar{y}_{pr_2}]$	18.83263	65.14586	2165681.232	88.31818	21.76666	66.10277
$[D, \bar{y}_{pr_3}]$	14.04463	87.35499	2010660.293	95.12747	17.20327	83.63737
$[D, T_{E_1}]$	5.566858	220.3879	1664625.826	114.9021	10.60077	135.7295
$[D, T_{E_2}]$	9.786525	125.363	1756029.275	108.9213	12.05754	119.3309
$[D, T_{E_3}]$	3.084702	397.7267	1507964.93	126.8392	8.269937	173.984

Table 3. MSE and PRE of the Strategies for Universe I under different Non Response Rate

Strategy	Non Response Rate (10%)		Non Response Rate (20%)		Non Response Rate (30%)	
	M(.)	PRE	M(.)	PRE	M(.)	PRE
$[D, \bar{y}_r]$	1837.98	100.00	15824.18	100.00	1309431.10	100.00
$[D, \bar{y}_{RAT}]$	1868.06	98.39	15882.04	99.64	1282610.48	102.09
$[D, \bar{y}_{COMP}]$	1786.75	102.87	14378.09	110.06	1280419.83	102.27
$[D, \bar{y}_{SD}]$	1867.98	98.39	14368.32	110.13	1280418.13	102.27
$[D, \bar{y}_{pr_1}]$	3686.14	49.86	50881.37	31.10	1464535.82	89.41
$[D, \bar{y}_{pr_2}]$	4397.04	41.80	55297.60	28.62	1482629.23	88.32
$[D, \bar{y}_{pr_3}]$	1838.19	99.99	20296.94	77.96	1327573.72	98.63
$[D, T_{E_1}]$	1388.00	132.42	4452.68	355.39	1061366.76	123.37
$[D, T_{E_2}]$	1786.75	102.87	14378.09	110.06	1280419.83	102.27
$[D, T_{E_3}]$	1336.77	137.49	3006.59	526.32	1032355.48	126.84

Table 4. MSE and PRE of the Strategies for Universe III under different Non Response Rate

Strategy	Non Response Rate (10%)	Non Response Rate (20%)	Non Response Rate (30%)

	M(.)	PRE	M(.)	PRE	M(.)	PRE
$[D, \bar{y}_r]$	2072.81	100.00	18055.79	100.00	1480811.52	100.00
$[D, \bar{y}_{RAT}]$	2140.50	96.84	18185.99	99.28	1420465.11	104.25
$[D, \bar{y}_{COMP}]$	1957.54	105.89	14802.10	121.98	1415536.15	104.61
$[D, \bar{y}_{SD}]$	2140.31	96.85	14780.11	122.16	1415532.33	104.61
$[D, \bar{y}_{pr_1}]$	3920.98	52.86	53112.99	34.00	1635916.24	90.52
$[D, \bar{y}_{pr_2}]$	4958.85	41.80	63095.98	28.62	1676678.10	88.32
$[D, \bar{y}_{pr_3}]$	2073.29	99.98	28119.51	64.21	1521632.40	97.32
$[D, T_{E_1}]$	1622.84	127.73	6684.30	270.12	1232747.18	120.12
$[D, T_{E_2}]$	1957.54	105.89	14802.10	121.98	1415536.15	104.61
$[D, T_{E_3}]$	1507.57	137.49	3430.60	526.32	1167471.80	126.84

Table 5. MSE and PRE of the Strategies for Universe V under different Non Response Rate

Strategy	Non Response Rate (10%)		Non Response Rate (20%)		Non Response Rate (30%)	
	M(.)	PRE	M(.)	PRE	M(.)	PRE
$[D, \bar{y}_r]$	2374.75	100.00	20925.01	100.00	1701157.77	100.00
$[D, \bar{y}_{RAT}]$	2490.77	95.34	21148.20	98.94	1597706.79	106.47
$[D, \bar{y}_{COMP}]$	2177.14	109.08	15347.25	136.34	1589257.13	107.04
$[D, \bar{y}_{SD}]$	2490.46	95.35	15309.56	136.68	1589250.59	107.04
$[D, \bar{y}_{pr_1}]$	4222.91	56.23	55982.20	37.38	1856262.49	91.64
$[D, \bar{y}_{pr_2}]$	5681.17	41.80	73122.47	28.62	1926169.49	88.32
$[D, \bar{y}_{pr_3}]$	2375.56	99.97	38177.10	54.81	1771136.43	96.05
$[D, T_{E_1}]$	1924.77	123.38	9553.51	219.03	1453093.43	117.07
$[D, T_{E_2}]$	2177.14	109.08	15347.25	136.34	1589257.13	107.04
$[D, T_{E_3}]$	1727.17	137.49	3975.75	526.32	1341192.79	126.84

8.2. Simulation Study

A simulation study depends on the actual selection of large number of samples from the target population and an average value of the deviations of the estimates from the actual value is considered for average bias and average MSE. For this purpose we have chosen only the population IV, with N=200, n = 20 and r = 15.

Table 6. Simulation - Based Results for optimum MSE and PRE of the Strategies for Universe IV

Strategy	Population IV	
	Optimum M(.)	PRE
$[D, \bar{y}_r]$	12.12	100
$[D, \bar{y}_{RAT}]$	9.77	124.13
$[D, \bar{y}_{COMP}]$	9.59	126.44
$[D, \bar{y}_{SD}]$	9.58	126.52
$[D, \bar{y}_{pr_1}]$	17.01	71.27
$[D, \bar{y}_{pr_2}]$	18.43	65.77
$[D, \bar{y}_{pr_3}]$	14.01	86.51

$[D,T_{E_1}]$	5.28	229.57
$[D,T_{E_2}]$	9.58	126.47
$[D,T_{E_3}]$	2.89	419.99

Remark 5: The table reveals the fact that

(i) The proposed estimator T_{E_i} (i = 1, 2, 3) are uniformly better than the existing estimators; the estimator T_{E_2} is as good as the estimator $\bar{y}_{COMP}$. The results hold for all the six populations.

(ii) The strategy $[D,T_{E_3}]$ is the most efficient one among all other strategies.

(iii) All the three proposed strategies, under optimality conditions are uniformly better than the MMI.

(iv) Although the strategy $[D,T_{E_1}]$ is better than the strategy $[D,\bar{y}_{RAT}]$ under certain condition, which is mostly satisfied in sample surveys with presence of non - response; the strategy $[D,T_{E_2}]$ and $[D,T_{E_3}]$ are uniformly better than the strategy $[D,\bar{y}_{RAT}]$ when optimality conditions are satisfied for $[D,T_{E_2}]$ and $[D,T_{E_3}]$.

(v) Strategy $[D,T_{E_1}]$ is better than the strategy $[D,\bar{y}_{COMP}]$, under their optimality conditions, but under certain restriction on the rate of non - response in the sample. Further, strategy $[D,T_{E_2}]$ and $[D,\bar{y}_{COMP}]$ are equally efficient, but $[D,T_{E_3}]$ is uniformly better than the strategy $[D,\bar{y}_{COMP}]$.

It seems, therefore, that the strategy $[D,T_{E_3}]$ is the most powerful strategy amongst the three proposed imputation strategies.

(vi) The **simulated values** of PREs are quite closer to the values depicted in Table 2 for population IV. while comparing the MSEs without simulation, which is obvious as the simulated values are obtained through practically a very large number of samples from the population, thereby representing the characteristics of the population to a greater extent.

9. CONCLUSIONS

This paper is devoted to the development of a number of imputation methods based upon an auxiliary variable through ETEs. The composite method of imputation was used for this purpose. Although, Singh and Horn (2000) propounded the concept of compromized method and developed a point estimator, but throughout in this paper, it was observed that the proposed strategies were more efficient compared to it and that too, with the same amount of knowledge of the population parameters. In this sense, the strategies, suggested here, are more advantageous than Singh and Horn (2000) and Singh and Deo (2003) estimators.

Acknowledgments: The authors are thankful to the referees for their valuable comments which helped in improving the overall work carried out.

RECEIVED: AUGUST 2023.
REVISED: NOVEMBER, 2023

REFERENCES

[1] AHMED, M. S., O. AL-TITI, Z. AL-RAWI, and W. ABU-DAYYEH (2006): Estimation of a population mean using different imputation methods. **Stat. Trans**.; 7, 1247–1264.

[2] AL - OMARI, A. I. and C.N. BOUZA (2014): Ratio estimators of the population mean with missing values using ranked set sampling.**Environmetrics**,DOI:10.1002/env.2286.

[3] . BHUSHAN, S.,and A. P. PANDEY. (2018): Optimality of ratio type estimation methods for population mean in the presence of missing data, **Communications in Statistics: Theory and Methods**. 47, 2576-2589.

[4] BAHL, S. and TUTEJA, R. K. (1991): Ratio and product type exponential estimator, **Journal of Information and Optimization Sciences.** 12, 159-164.

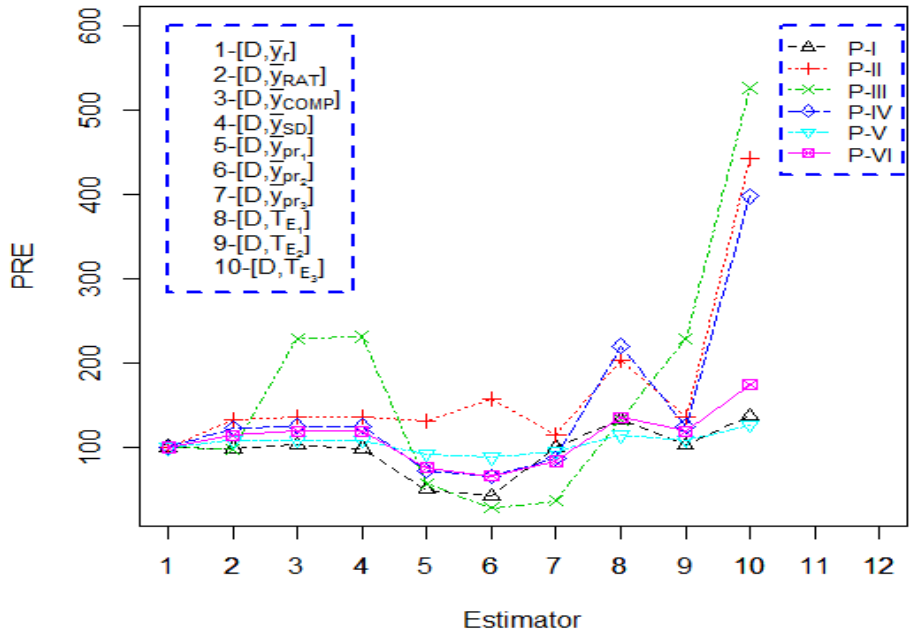
[5] BOUZA, C. N.: (2002A): Estimation of the mean in ranked set sampling with non responses. **Metrika,** 2002A. 56, 171–179.

[6] BOUZA, C. N.(2002B): Ranked set subsampling the non response strata for estimating the difference of mean. **Biometrical Journal**, 44, 903–915.

[7] BOUZA C.N and A, I. AL - OMARI (2012): Estimating the population mean in the case of missing data using simple random sampling. **Statistics**, 46, 279-290

[8] DIANA, G. and PERRI, P. F.(2010): Improved estimators of the population mean for missing data, **Communications in Statistics: Theory and Methods**. 39, 3245-3251.
[9] Herrera, Bouza, C.N. and VIADA, CARMEN E. (2021): Imputation of Individual Values of A Variable Using Product Predictors, **Journal of Revista Investigacion Operacional**, (2021), 42 , 321-325.
[10] KADILAR, C. and CINGI, H.(2008): Estimators for the population mean in the case of missing data, **Communication in Statistics: Theory and Methods.** 37, 2226-2236.
[11] KALTON, G., KASPRZYK, D. and SANTOS, R. (1981): **Issues of Non-Response and Imputation in the Survey of Income and Programme Participation, in Current Topics in Survey Sampling** (eds. D. Krewski, R. Platek and J. N. K. Rao), Academic Press, New York.
[12] MEEDEN, G. (2000): A decision theoretic approach to imputation in finite population sampling, **Journal of the American Statistical Association**, 95, 586-595.
[13] MULHOPADHYAY, P.(2000): **Theory and Methods of Survey Sampling**, Prentice-Hall of India, New Delhi.
[14] RAO, J. N. K. and SITTER, R. R. (1995): Variance estimation under two phase sampling with application to imputation for missing data, **Biometrika,** 82, 453 – 460.
[15] Rubin, D. B.(1986) : Basic ideas of multiple imputation on nonresponse, **Survey Methodology**.12(1), 37-47.
[16] SEDRANSK, J. (1985): The objective and practice of imputation, Proceeding of the First Annual Research Conference, **Bureau of the Census, Washington, D .C**. 445-452.
[17] SHUKLA, D., SINGHAI, R., and THAKUR, N.S.(2011B): A new imputation method for missing attribute values in data mining, **Journal of Applied Computer Science and Mathematics**, 10, 14-19.
[18] SINGH, S., HORN, S. (2000): Compromised imputation in survey sampling. **Metrika,** 51, 267-276.
[19] SINGH, S., DEO, B. (2003): Imputation by power transformation. **Statistical Papers**, 44, 555-579.
[20] SINGH, A. K., SINGH, PRIYANKA. and SINGH, V. K. (2014A): Exponential-type compromised imputation in Survey Sampling, **Journal of the Statistics Applications and Probability**, 3, 211-217.
[21] SINGH, A.K, SINGH, PRIYANKA and SINGH, V. K. (2014B): Imputation methods of missing data for estimating the population mean using simple random sampling**, Global Journal of Advanced Research,** 1. 253-263.
[22] SINGH, A.K, SINGH, PRIYANKA and SINGH, V. K. (2022): Estimation of Mean with Imputation of Missing data using Exponential-Type Estimators, **Journal of the Statistics Applications and Probability and Letters,** 9, 1-9.
[23] SRINATH, K. P. (1971) : Multiphase Sampling in Nonresponse Problems**, Journal of American Statistical. Association**, 66, 583-589.

Figure 1



Appendix
Section A

We have

$$\bar{y}_{T_{E_1}} = k\bar{y}_r + (1-k)\,\bar{y}_r\,\phi(\alpha, \bar{X}, \bar{x}_n),$$
$$\bar{y}_{T_{E_2}} = k\,\bar{y}_r + (1-k\,)\bar{y}_r\phi(\alpha, \bar{x}_n, \bar{x}_r),$$

and

$$\bar{y}_{T_{E_3}} = k\,\bar{y}_r + (1-k\,)\bar{y}_r\phi(\alpha, \bar{X}, \bar{x}_r)$$

Where $\phi(\alpha, \bar{X}, \bar{x}_n) = \exp\left[\alpha\left(\frac{\bar{X}-\bar{x}_n}{\bar{X}+\bar{x}_n}\right)\right]$;

$\phi(\alpha, \bar{x}_n, \bar{x}_r) = \exp\left[\alpha\left(\frac{\bar{x}_n-\bar{x}_r}{\bar{x}_n+\bar{x}_r}\right)\right]$;and

$\phi(\alpha, \bar{X}, \bar{x}_r) = \exp\left[\alpha\left(\frac{\bar{X}-\bar{x}_r}{\bar{X}+\bar{x}_r}\right)\right]$

Using the idea of double sampling and MCAR concept of Rao and Sitter (1995) mechanism, we have
$\bar{y}_r = \bar{Y}(1+\varepsilon)$, $\bar{x}_r = \bar{X}(1+\delta)$ and $\bar{x}_n = \bar{X}(1+\eta)$ such that $E(\varepsilon) = E(\delta) = E(\eta) = 0$ and
$E(\varepsilon^2) = V(\varepsilon) = \theta_{r,N}C_Y^2$; $E(\delta^2) = V(\delta) = \theta_{r,N}C_X^2$; $E(\eta^2) = V(\eta) = \theta_{n,N}C_X^2$;
$E(\varepsilon\delta) = \theta_{r,N}\rho C_Y\,C_X$; $E(\varepsilon\eta) = \theta_{n,N}\rho C_Y\,C_X$; $E(\delta\eta) = \theta_{n,N}C_X^2$.

The estimators $\bar{y}_{T_{E_i}}$ (i=1, 2, 3) is a function of ε, δ and η up to first order of approximation, could be expressed as:

$$\bar{y}_{T_{E_1}} = \bar{Y}\left[(1+k\varepsilon) + (1-k\,)\left(\frac{\alpha\eta^2}{4} + \frac{\alpha^2\eta^2}{8} - \frac{\alpha\varepsilon\eta}{2} - \frac{\alpha\eta}{2} + \varepsilon\right)\right] \quad \text{(A1)}$$

$$\bar{y}_{T_{E_2}} = \bar{Y}\left[(1+k\varepsilon) + (1-k\,)\left(\frac{\alpha(\eta-\delta)}{2} - \frac{\alpha(\eta^2-\delta^2)}{4} + \frac{\alpha^2(\eta-\delta)^2}{8} + \frac{\alpha(\varepsilon\eta-\varepsilon\delta)}{2} + \varepsilon\right)\right] \quad \text{(A2)}$$

$$\bar{y}_{T_{E_3}} = \bar{Y}(1+k\varepsilon) + (1-k\,)\bar{Y}\left(\varepsilon - \frac{\alpha\delta}{2} + \frac{\alpha\delta^2}{4} + \frac{\alpha^2\delta^2}{8} - \frac{\alpha\varepsilon\delta}{2}\right) \quad \text{(A3)}$$

The expression (A1), (A2) and (A3), found with assumption that $|\varepsilon|<1, |\eta|<1 \text{and} |\delta|<1$. Applying expectation of (A1), (A2) and (A3) of both the sides. That is $B[T_{E_i}] = E(T_{E_i}) - \bar{Y}, i = 1,2,3$ .We have the equations (35), (36) and (37).
Similarly, squaring the expression (A1), (A2) and (A3), neglecting the terms of $\varepsilon, \delta \text{ and } \eta$ greater than two and realising that

$$M\left(T_{E_i}\right) = E[T_{E_i}{}^2] + \bar{Y}^2 - 2\bar{Y}E[T_{E_i}]\ , \qquad i = 1,2,3$$

The expressions (38), (39) and (40) could be obtained.