Journal of Air Pollution and Health (Summer 2023); 8(3): 285-298

Original Article

Available online at http://japh.tums.ac.ir

A new imputation method for population mean in the presence of missing data based on a transformed variable with applications to air pollution data in Chiang Mai, Thailand

Natthapat Thongsak[1], *Nuanpan Lawson*[2,*]

[1] State Audit Office of the Kingdom of Thailand, Bangkok, Thailand
[2] Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

ARTICLE INFORMATION

Article Chronology:
Received 26 June 2023
Revised 22 August 2023
Accepted 05 September 2023
Published 29 September 2023

Keywords:
Imputation method; Missing data; Transformed variable; Air pollution data; Mean square error

CORRESPONDING AUTHOR:

nuanpan.n@sci.kmutnb.ac.th
Tel: (+66) 25552000
Fax: (+66) 25874350

ABSTRACT

Introduction: Chiang Mai's air pollution has risen to number one in the world for the highest level of fine particulate matter which further exacerbates the damage to human health. Fine particulate matter can enter the human body and blood circulation, destroying organ systems, increasing the risk for chronic disease and cancer, despite not having smoking habits or other morbidities. The Thai government must sort out this issue before it is too late as the whole nation's health is at risk due to excessive dust levels higher than standard guidelines. Collection of pollution data can help us to come up with solutions and prevent it from turning into a hazardous situation. Unfortunately, pollution data are missing and need to be dealt with before analysis to obtain accurate results.

Materials and methods: A new method of imputation for estimating population mean based on a transformed variable has been suggested under simple random sampling without replacement and the uniform nonresponse mechanism. The bias and mean square error of the proposed estimator are investigated up to the first order of approximation. The performance of the proposed estimator is studied via applications to air pollution data in Chiang Mai, Thailand.

Results: The proposed estimator shows the best performance, giving the least bias and mean square error for all levels of sampling fractions. For the results from application the estimated value of sulfur dioxide from Particulate Matter 2.5 ($PM_{2.5}$), the Percentage Relative Efficiency (PRE) is higher than all existing estimators by at least 16%. For the estimated $PM_{2.5}$ from PM_{10} the PRE is higher than all existing estimators by at least 1600%, an extremely significant difference exhibiting similarity to real values.

Conclusion: The proposed imputation technique based on the transformed auxiliary variable can be helpful for imputing missing values and improving the efficiency of the estimators.

Introduction

Fine particulate matter with small particles less than 2.5 µm ($PM_{2.5}$), less than 10 µm (PM_{10}), and sulfur dioxide have afflicted Thailand on a large-scale, including many sectors. The tourism

Please cite this article as: Thongsak N, Lawson N. A new imputation method for population mean in the presence of missing data based on a transformed variable with applications to air pollution data in Chiang Mai, Thailand. Journal of Air Pollution and Health. 2023;8(3): 285-298.

sector is imperative to Thailand's economy and the beauty and cultural significance of locations is appreciated world-wide. However, sightseers now must be aware of the risks of the extensive air pollution, which deters many tourists and deducts from the economy. One of the most popular places to visit for tourists in Thailand is Chiang Mai, an amazing province with beautiful sites and a fresh atmosphere which benefit the city in terms of investment from tourism and businesses. The growth of tourists and large amounts of revenue due to investment in Chiang Mai have faced an unprecedented crisis caused by the coronavirus pandemic since a few years ago along with worsening air pollution problems. The virus situation is now better with the vaccines distributed to people. However, many areas in Chiang Mai are presently full of smog that is, the accumulation of dust and gases such as sulfur dioxide, carbon monoxide and nitrogen oxide which has spread to the community. The pollution is severely affected by burning of agricultural waste, burning of land, community waste, building fires to keep warm, and so on. Although the Thai government has measures to prevent fire burning in open areas and agricultural areas but this issue is still unresolved. $PM_{2.5}$ PM_{10}, and sulfur dioxide affect the climate and ecosystems and can lead to hazardous chronic diseases including cardiovascular disease, lung disease, and lung cancer with increased risk upon exposure. Furthermore, certain populations are especially vulnerable such as children and the elderly [1].

Chiang Mai's pollution indicates a crisis as the smog has become so thick that Doi Suthep can hardly be seen. Furthermore, the air quality measurement results exceed standard values, which continue to affect health and the situation seems to become even more severe as time passes. Chiang Mai has held the world record for bad air quality rates with the highest $PM_{2.5}$ when compared to other cities around the world, surpassing Bangkok. There are many factors influencing the repercussions of fine particulate matter such as the amount of the received dust particles, the period of time in contact with the dust, especially for specific groups of people like elderly people, young children or people with congenital diseases. $PM_{2.5}$ does not only spread into the respiratory system and blood stream, it can negatively affect the functioning of various organs in the body which can be harmful to human life, and causes life-threatening chronic diseases including heart disease, lung disease, and lung cancer, if the body has been in contact with the dust for a long time.

To help the Thai government in planning and prevention of air pollution problems occurring in Thailand, the Pollution Control Department of Thailand is an organization that records data on air pollution including $PM_{2.5}$, PM_{10}, sulfur dioxide, carbon monoxide, ozone, and so on, but unfortunately some of the air pollution data are missing. Dealing with the missing data in a proper way is ideal to lead to accurate information and planning policies based on that information can follow with good results in preventing the harms of air pollution to human life. Replacing the missing data with plausible values is called imputation and it can assist in solving the problem of item nonresponse. The easiest simple imputation method is mean imputation which replaces the missing values with the sample mean of response of the study variable. The study variable y_i from the mean imputation method after imputation is defined as

$$y_{.i} = \begin{cases} y_i; \ y_i \text{ is observed} \\ \bar{y}_r; \ y_i \text{ is missing}, \end{cases}$$

Where $\bar{y}_r = \sum_{i=1}^{r} y_i / r$ is the sample mean of response of the study variable and r is the number of respondents.

The point estimator of the population mean is $\hat{\bar{Y}}_S = \frac{1}{n} \sum_{i \in S} y_{.i} = \bar{y}_r.$

The bias and variance of $\hat{\bar{Y}}_S$ are given as

$$Bias\left(\hat{\bar{Y}}_S\right)=0, \quad (1)$$
$$V\left(\hat{\bar{Y}}_S\right)=\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2C_Y^2,$$

Where $\bar{Y}=\frac{\sum_{i=1}^{N}y_i}{N}$ is the population mean of study variable, $C_y=S_y/\bar{Y}$ is the population coefficient of variation of Y, $S_Y^2=\frac{\sum_{i=1}^{N}\left(y_i-\bar{Y}\right)^2}{N-1}$, n and N are the sample size and a population size respectively.

If there is a positive correlation between a study variable Y and an auxiliary variable X, the ratio imputation method can be used to estimate the missing observation. The study variable y_i from the ratio imputation method after imputation is defined as

$$y_{.i}=\begin{cases} y_i,\ y_i \text{ is observed} \\ \left(\frac{\sum_{i=1}^{r}y_i}{\sum_{i=1}^{r}x_i}\right)x_i, y_i \text{ is missing.} \end{cases}$$

The point estimator of the population mean is

$$\hat{\bar{Y}}_{Rat}=\bar{y}_r\frac{\bar{x}_n}{\bar{x}_r},$$

Where $\bar{x}_n=\sum_{i=1}^{n}x_i/n$ and $\bar{x}_r=\sum_{i=1}^{r}x_i/r$ are the sample mean of the auxiliary variable and the sample mean of response of the auxiliary variable respectively.

The bias and mean square error (MSE) of are $\hat{\bar{Y}}_{Rat}$

$$Bias\left(\hat{\bar{Y}}_{Rat}\right)=\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}\left(C_X^2-\rho C_XC_Y\right), \quad (2)$$
$$MSE\left(\hat{\bar{Y}}_{Rat}\right)=\left(\frac{1}{n}-\frac{1}{N}\right)\bar{Y}^2C_Y^2+\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}^2\left(C_Y^2+C_X^2-2\rho C_XC_Y\right),$$

where $\bar{X}=\frac{\sum_{i=1}^{N}x_i}{N}$ is the population mean of auxiliary variable, $C_x=S_x/\bar{X}$ is the coefficient of variation of X, $S_X^2=\frac{\sum_{i=1}^{N}\left(x_i-\bar{X}\right)^2}{N-1}$ and $\rho=\frac{S_{XY}}{S_XS_Y}$. is the population correlation coefficient between X and $S_{XY}=\frac{\sum_{i=1}^{N}\left(x_i-\bar{X}\right)\left(y_i-\bar{Y}\right)}{N-1}$,

A new imputation method called compromised imputation for estimating missing values when the nonresponse mechanism is missing completely at random (MCAR) under simple random sampling without replacement (SRSWOR) was proposed [1]. The mean and ratio imputation methods were outperformed by [1] in terms of a smaller MSE. The study variable from [1] after imputation is defined as

$$y_{.i}=\begin{cases} \alpha\frac{n}{r}y_i+\left(1-\alpha\right)\hat{\beta}x_i\ ;\ y_i \text{ is observed} \\ \left(1-\alpha\right)\hat{\beta}x_i\ ;\ y_i \text{ is missing,} \end{cases}$$

where $\hat{\beta}=\frac{\sum_{i=1}^{r}y_i}{\sum_{i=1}^{r}x_i}$, and α are suitable constants that make the MSE of the estimator minimum.

The point estimator of the population mean is

$$\hat{\bar{Y}}_{SH}=\alpha\bar{y}_r+\left(1-\alpha\right)\bar{y}_r\frac{\bar{x}_n}{\bar{x}_r}.$$

The bias and MSE of $\hat{\bar{Y}}_{SH}$ are

$$Bias\left(\hat{\bar{Y}}_{SH}\right)=\left(1-\alpha\right)\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}\left(C_X^2-\rho C_XC_Y\right),$$
$$MSE\left(\hat{\bar{Y}}_{SH}\right)=\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2C_Y^2+\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}^2\left(\left(1-\alpha\right)^2C_X^2-2\left(1-\alpha\right)\rho C_XC_Y\right).$$

Under the optimum value of $\alpha_{SH}^{opt}=1-\rho\frac{C_Y}{C_X}$, the

MSE of $\hat{\bar{Y}}_{SH}^{opt}$ is

$$MSE\left(\hat{\bar{Y}}_{SH}^{opt}\right)=\bar{Y}^2C_Y^2\left(\left(\frac{1}{r}-\frac{1}{N}\right)-\left(\frac{1}{r}-\frac{1}{n}\right)\rho^2\right). \quad (3)$$

From past literatures based on applications to pollution data, a new imputation method for estimating the average $PM_{2.5}$ in Bangkok, Thailand under MCAR considering the density of ozone as an auxiliary variable was proposed [2]. The idea of minimizing the MSE of their estimator using two constants to get the best estimator for estimating population mean was introduced [2]. A new imputation technique using the benefit of the response rate and the constant that makes MSE optimum and application to study the average amount of $PM_{2.5}$ in Bangkok, Thailand using carbon monoxide as an auxiliary variable in the situation where some parameters are not available was introduced [3]. Recently, a new ratio estimator for estimating population total has been recommended under unequal probability sampling without replacement when nonresponse occurs on the study variable under missing at random [4] . The ratio estimator is used to estimate fine particulate matter in the north of Thailand.

A transformation method to transform an auxiliary variable in order to increase the efficiency of the estimators was investigated by [5] under Simple Random Sampling without Replacement (SRSWOR). After that there were many researchers who suggested the transformation of variables following [5]. For example, transforming the shape of the auxiliary variable to decrease the bias and MSE of population mean estimators and applying them to estimate carbon monoxide with the $PM_{2.5}$ from the air pollution data in Nan, Thailand under SRSWOR was suggested [6]. Utilization of the transformed variables on the combined estimators for estimating the average $PM_{2.5}$ using nitrogen dioxide pollution data in Chiang Rai under double sampling was also suggested [7]. A higher Percentage Relative Efficiency (PRE) than the single estimators was shown by [7]. A family of estimators using the transformation of an auxiliary variable when the population mean of the auxiliary variable is unknown under double sampling was proposed [8] and it was applied to air pollution in Chiang Rai to estimate nitrogen dioxide with $PM_{2.5}$. Classes of estimators utilizing the transformation on only an auxiliary variable and both the auxiliary and study variables were suggested [9] under double sampling to gain more efficiency for estimating population mean.

There are some limitations of the existing common imputation methods in the past for mean and ratio imputation methods. For example, mean imputation method can lead to biases in standard error and the variance for imputed estimators and also the estimated values in multivariate such as correlation and disregarding distribution. The ratio imputation method can also lead to bias for the imputed estimator. Using the transformation technique can assist in changing the shape of the distribution of an auxiliary variable [6] and is expected to gain more efficiency for the estimators.

In this paper, a new technique to impute missing values based on a transformed auxiliary variable has been proposed under SRSWOR and when the missing data are uniform. Utilizing the transformation method can be used in assisting the efficiency of the new estimator. The bias and MSE are considered using the Taylor series approximation up to the first order. To see how the proposed estimator performs, the proposed estimator is compared with existing estimators using the MSE as a criterion via simulation studies and applications to air pollution data in Chiang Mai, Thailand which is one of the cities that is in a critical situation with the dust problem nowadays.

Materials and methods

Proposed estimator

Utilizing the benefit of the transformation of the auxiliary variable suggested by [5] and following the idea of [1], a new estimator for estimating population mean when the study variable is missing is suggested. We use a new imputation technique based on the transformed auxiliary variable to estimate missing values under SRSWOR and uniform nonresponse. This technique is a renowned method in the field of sample survey for transforming an auxiliary variable which can change its shape of distribution and increase the efficiency of the estimator. Let

$$y_i = \begin{cases} k\bar{y}_r + (1-\alpha)\bar{y}_r\left(\dfrac{\bar{x}^*}{\bar{X}}\right)^k, & y_i \text{ is observed} \\ (1-\alpha)\bar{y}_r\left(\dfrac{\bar{x}^*}{\bar{X}}\right)^k, & y_i \text{ is missing,} \end{cases}$$

where $\bar{x}^* = \dfrac{N\bar{X} - n\bar{x}_n}{N-n} = (1+\pi)\bar{X} - \pi\bar{x}_n; \quad \pi = \dfrac{n}{N-n}$

is a transformed sample mean of the auxiliary variable and is a constant that minimizes the MSE of the proposed estimator and is a constant that is considered in this study such as the sample regression coefficient b and the response rate, which are free from parameters of the auxiliary variable so it is easy to apply in practice.

The population mean estimator after imputation of the missing values is

$$\hat{\bar{Y}}_{\mathrm{N}} = \alpha\bar{y}_r + (1-\alpha)\bar{y}_r\left(\frac{\bar{x}^*}{\bar{X}}\right)^k. \tag{4}$$

To study the bias and MSE of the proposed estimator in Eq. 4,

let $\varepsilon_0 = (\bar{y}_r - \bar{Y})/\bar{Y}, \bar{y}_r = (1+\varepsilon_0)\bar{Y}$, $\varepsilon_1 = (\bar{x}_r - \bar{X})/\bar{X}$,

$\bar{x}_r = (1+\varepsilon_1)\bar{X}$, $\varepsilon_2 = (\bar{x}_n - \bar{X})/\bar{X}, \bar{x}_n = (1+\varepsilon_2)\bar{X}$,

$E(\varepsilon_0) = E(\varepsilon_1) = E(\varepsilon_2) = 0$, $E(\varepsilon_0^2) = \left(\dfrac{1}{r} - \dfrac{1}{N}\right)C_Y^2$,

$E(\varepsilon_1^2) = \left(\dfrac{1}{r} - \dfrac{1}{N}\right)C_X^2$, $E(\varepsilon_2^2) = \left(\dfrac{1}{n} - \dfrac{1}{N}\right)C_X^2$,

$E(\varepsilon_0\varepsilon_1) = \left(\dfrac{1}{r} - \dfrac{1}{N}\right)\rho C_X C_Y$, $E(\varepsilon_0\varepsilon_2) = \left(\dfrac{1}{n} - \dfrac{1}{N}\right)$,

$\rho C_X C_Y$, $E(\varepsilon_1\varepsilon_2) = \left(\dfrac{1}{n} - \dfrac{1}{N}\right)C_X^2$.

Rewriting $\hat{\bar{Y}}_N$ in terms of , we get

$$\hat{\bar{Y}}_{\mathrm{N}} = \alpha\bar{y}_r + (1-\alpha)\bar{y}_r\left(\frac{\bar{x}^*}{\bar{X}}\right)^k$$

$$= \alpha(1+\varepsilon_0)\bar{Y} + (1-\alpha)(1+\varepsilon_0)\bar{Y}\left(\frac{(1+\pi)\bar{X} - \pi\bar{x}_n}{\bar{X}}\right)^k$$

$$\approx \alpha(1+\varepsilon_0)\bar{Y} + (1-\alpha)\bar{Y}\left(1 + \varepsilon_0 - k\pi\varepsilon_2 + \frac{k(k-1)}{2}\pi^2\varepsilon_2^2 - k\pi\varepsilon_0\varepsilon_2\right)$$

Up to the first order approximation using Taylor series approximation, terms of powers of more than two are small and considered negligible, the approximation of the bias of $\hat{\bar{Y}}_N$ is

$$Bias\left(\hat{\bar{Y}}_{\mathrm{N}}\right) = E\left[\hat{\bar{Y}}_N - \bar{Y}\right]$$

$$\approx E\left(\bar{Y}\left(\varepsilon_0 - k\pi(1-\alpha)\varepsilon_2 + \frac{k(k-1)}{2}(1-\alpha)\pi^2\varepsilon_2^2 - k\pi\right.\right.$$

$$(1-\alpha)\varepsilon_0\varepsilon_2 = (1-\alpha)\left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}\left(\frac{k(k-1)}{2}\pi^2 C_X^2 -\right.$$

$$\left.\left.k\pi\rho C_X C_Y\right)\right).$$

The approximation of MSE of $\hat{\bar{Y}}_N$ is

$$MSE\left(\hat{\bar{Y}}_{\mathrm{N}}\right) = E\left[\hat{\bar{Y}}_L - \bar{Y}\right]^2$$

$$\approx \bar{Y}^2 E\left(\varepsilon_0 - k\pi(1-\alpha)\varepsilon_2\right)^2$$

$$= \bar{Y}^2\left[\left(\frac{1}{r} - \frac{1}{N}\right)C_X^2 + \left(\frac{1}{n} - \frac{1}{N}\right)\left(k^2\pi^2(1-\alpha)^2 C_X^2 - 2k\pi\right.\right. \tag{5}$$

$$\left.\left.(1-\alpha)\rho C_X C_Y\right)\right].$$

In order to find the minimum MSE, we can find the optimum value of α by taking a partial

derivative of the MSE in Eq. 5 with respect to 1-α and equating it to zero.

$$\frac{\partial MSE\left(\hat{\bar{Y}}_N\right)}{\partial(1-\alpha)}=0$$

$$\bar{Y}^2\left(2k^2\pi^2(1-\alpha)\left(\frac{1}{n}-\frac{1}{N}\right)C_X^2-2k\pi\left(\frac{1}{n}-\frac{1}{N}\right)\rho C_X C_Y\right)=0$$

Therefore, the optimum value of $\alpha(\alpha_{opt})$ is

$$\alpha_{opt}=1-\frac{\rho C_Y}{k\pi C_X}. \quad (6)$$

Plug in the α_{opt} in Eq. 6 to the MSE in Eq. 5, then the approximation of the MSE of the proposed estimator at its optimum is

$$MSE\left(\hat{\bar{Y}}_{\mathrm{N}}^{\mathrm{opt}}\right)\approx\bar{Y}^2C_Y^2\left[\left(\frac{1}{r}-\frac{1}{N}\right)-\left(\frac{1}{n}-\frac{1}{N}\right)\rho^2\right]. \quad (7)$$

Note that: Any values of k can be used in this study to make the MSE of the $\hat{\bar{Y}}_N$ in Eq. 7 minimum.

Efficiency comparison

The MSE is a criterion to compare the performance of the proposed estimators with the existing imputation methods; mean imputation, ratio imputation, and [1]. The minimum MSE of the proposed estimator in Eq. (7) is compared with the MSEs of the existing estimators in Eq. (1)-(3), respectively. The details are shown below.

The proposed estimator performs better than the mean method estimator under the certain condition as follows:

$$MSE\left(\hat{\bar{Y}}_{\mathrm{N}}^{\mathrm{opt}}\right)<MSE(\hat{\bar{Y}}_{\mathrm{S}})$$

$$\bar{Y}^2C_Y^2\left[\left(\frac{1}{r}-\frac{1}{N}\right)-\left(\frac{1}{n}-\frac{1}{N}\right)\rho^2\right]<\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2C_Y^2$$

$$\rho^2>0.$$

The proposed estimator performs better than the ratio method estimator under the certain condition as follows:

$$MSE\left(\hat{\bar{Y}}_{\mathrm{N}}^{\mathrm{opt}}\right)<MSE(\hat{\bar{Y}}_{\mathrm{RAT}})$$

$$\bar{Y}^2C_Y^2\left[\left(\frac{1}{r}-\frac{1}{N}\right)-\left(\frac{1}{n}-\frac{1}{N}\right)\rho^2\right]<\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2C_Y^2+\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}^2\left(C_X^2-2\rho C_X C_Y\right)$$

$$\left(\frac{1}{n}-\frac{1}{N}\right)\bar{Y}^2C_Y^2\rho^2+\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}^2\left(C_X^2-2\rho C_X C_Y\right)>0.$$

The proposed estimator performs better than [1] estimator under the certain condition as follows:

$$MSE\left(\hat{\bar{Y}}_{\mathrm{N}}^{\mathrm{opt}}\right)<MSE\left(\hat{\bar{Y}}_{\mathrm{SH}}^{\mathrm{opt}}\right)$$

$$\left(\frac{1}{n}-\frac{1}{N}\right)\rho^2\Big]<\bar{Y}^2C_Y^2\left(\left(\frac{1}{r}-\frac{1}{N}\right)-\left(\frac{1}{r}-\frac{1}{n}\right)\rho^2\right)$$

$$\left(\frac{N+r}{Nr}\right)n<2.$$

Results and discussion

Simulation studies

Simulation studies are conducted in this study by generating the paired variables *(X ,Y)* with the bivariate normal distribution with the following parameters that are chosen because they are suitable for the conditions specified in efficiency comparison, which is the situation where the proposed estimator is better than existing estimators; $N=2{,}000$, $\bar{X}=40$, $\bar{Y}=50$, $C_X=1.5$, $C_Y=2.5$. The correlation between X and Y are set to be in three levels; ρ = 0.3, 0.5 and 0.8 to see how the estimators perform. The study variable contains 5%, 15% and 30% of missing values under uniform nonresponse and the sample sizes are drawn based on SRSWOR with different sampling fractions f =5%, 10%, 30%, and 50% from a population of size N = 2,000. The simulation is repeated 10,000

times using R program [10].

The biases and MSEs of the proposed and existing estimators are calculated by the following formula

$$Bias\left(\hat{\bar{Y}}\right)=\frac{1}{10,000}\sum_{i=1}^{10,000}\left|\hat{\bar{Y}}_i-\bar{Y}\right|,$$

$$MSE\left(\hat{\bar{Y}}\right)=\frac{1}{10,000}\sum_{i=1}^{10,000}\left(\hat{\bar{Y}}_i-\bar{Y}\right)^2.$$

Biases and MSEs of the estimators are represented in Tables 1-3.

Table 1 showed the biases and MSEs of the estimators when ρ = 0.3. We can see that the proposed estimators for all values of $k$at its optimum performed the best for all levels of sampling fractions and levels of missing values. The biases do not depend upon the values of $k$as we can see the same results in terms of biases and it is also not much different for the MSEs either. A small amount of missing values leads to smaller biases and MSEs when compared to a big amount of missing values in the study variable, which is inversely correlated to the size of the sampling fraction as the biases and MSEs are smaller for a bigger sampling fraction.

Table 1. Biases and MSEs of the estimators when ρ = 0.3

% Missing	Estimator	Sampling fraction							
		0.05		0.1		0.3		0.5	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
5	Mean imputation	10.19	162.54	6.87	73.70	3.54	19.66	2.42	9.08
	Ratio imputation	10.18	162.01	6.85	73.55	3.55	19.85	2.45	9.33
	Singh and Horn	10.17	161.98	6.86	73.52	3.55	19.74	2.43	9.20
	Proposed $k=1$	9.89	152.96	6.63	68.76	3.42	18.36	2.33	8.44
	Proposed $k=b$	9.89	152.98	6.63	68.78	3.43	18.37	2.33	8.45
	Proposed $k=r/n$	9.89	152.97	6.63	68.76	3.42	18.36	2.33	8.45
15	Mean imputation	10.77	182.96	7.31	84.19	3.91	24.03	2.81	12.27
	Ratio imputation	10.83	184.29	7.31	84.27	3.85	23.35	2.71	11.43
	Singh and Horn	10.70	180.28	7.25	82.70	3.84	23.31	2.74	11.68
	Proposed $k=1$	10.52	174.38	7.10	79.52	3.81	22.78	2.75	11.64
	Proposed $k=b$	10.52	174.41	7.10	79.54	3.81	22.80	2.75	11.65
	Proposed $k=r/n$	10.52	174.40	7.10	79.53	3.81	22.79	2.75	11.64
30	Mean imputation	11.74	217.21	7.94	99.27	4.04	25.82	2.67	11.21
	Ratio imputation	11.95	227.21	8.07	103.31	4.10	26.63	2.73	11.64
	Singh and Horn	11.64	213.30	7.87	97.56	4.00	25.36	2.66	11.03
	Proposed $k=1$	11.49	208.28	7.73	94.63	3.95	24.59	2.60	10.60
	Proposed $k=b$	11.49	208.28	7.73	94.63	3.94	24.59	2.60	10.60
	Proposed $k=r/n$	11.49	208.29	7.73	94.63	3.94	24.59	2.60	10.60

The results from Tables 2-3 showed the biases and MSEs for the estimators when and and we can also see similar results to Table 1. The proposed estimators for all values of k at its optimum performed the best for both bias and MSE for all levels of sampling fractions and all levels of missing values. The mean imputation method seems to perform the worse when ρ is increased compared to others. The increasing of missing values leads to more biases and MSEs. On the other hand, the increase in sampling fraction can reduce the bias and MSE as expected. The higher values of ρ gave more accurate results at all levels of missing values.

Table 2. Biases and MSEs of the estimators when $\rho = 0.5$

% Missing	Estimator	Sampling fraction							
		0.05		0.1		0.3		0.5	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
5	Mean imputation	10.19	162.65	6.87	73.77	3.55	19.71	2.42	9.12
	Ratio imputation	10.14	160.68	6.82	72.88	3.52	19.49	2.40	8.98
	Singh and Horn	10.13	160.35	6.82	72.77	3.52	19.49	2.40	8.98
	Proposed $k=1$	9.12	130.38	6.10	58.36	3.16	15.74	2.16	7.28
	Proposed $k=b$	9.12	130.36	6.10	58.36	3.16	15.74	2.16	7.27
	Proposed $k=r/n$	9.12	130.39	6.10	58.37	3.16	15.74	2.16	7.28
15	Mean imputation	10.77	182.99	7.31	84.22	3.91	24.04	2.81	12.27
	Ratio imputation	10.66	178.57	7.18	81.39	3.75	22.19	2.60	10.54
	Singh and Horn	10.55	174.78	7.12	79.96	3.74	21.98	2.61	10.57
	Proposed $k=1$	9.84	152.40	6.64	69.31	3.58	20.18	2.61	10.44
	Proposed $k=b$	9.84	152.38	6.64	69.30	3.58	20.18	2.61	10.44
	Proposed $k=r/n$	9.84	152.43	6.64	69.33	3.59	20.20	2.61	10.45
30	Mean imputation	11.75	217.50	7.95	99.47	4.04	25.89	2.68	11.23
	Ratio imputation	11.48	208.91	7.77	95.22	3.95	24.65	2.62	10.73
	Singh and Horn	11.36	203.44	7.70	93.26	3.92	24.28	2.60	10.56
	Proposed $k=1$	10.86	186.05	7.29	84.31	3.73	21.96	2.45	9.42
	Proposed $k=b$	10.86	186.03	7.29	84.31	3.73	21.96	2.45	9.42
	Proposed $k=r/n$	10.87	186.07	7.29	84.32	3.73	21.96	2.45	9.41

Table 3. Biases and MSEs of the estimators when $\rho = 0.8$

% Missing	Estimator	Sampling fraction							
		0.05		0.1		0.3		0.5	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
5	Mean imputation	10.19	162.62	6.87	73.75	3.55	19.69	2.42	9.11
	Ratio imputation	10.06	158.22	6.76	71.59	3.46	18.88	2.33	8.45
	Singh and Horn	9.99	156.11	6.72	70.71	3.44	18.61	2.30	8.23
	Proposed $k=1$	6.88	74.70	4.54	32.47	2.37	8.91	1.67	4.31
	Proposed $k=b$	6.87	74.43	4.53	32.34	2.36	8.84	1.66	4.26
	Proposed $k=r/n$	6.88	74.72	4.54	32.48	2.37	8.91	1.67	4.31
15	Mean imputation	10.77	182.99	7.31	84.21	3.91	24.04	2.81	12.27
	Ratio imputation	10.38	169.49	6.98	76.72	3.61	20.50	2.45	9.39
	Singh and Horn	10.16	162.20	6.85	73.77	3.52	19.46	2.35	8.62
	Proposed $k=1$	7.86	97.71	5.29	43.85	2.94	13.48	2.24	7.51
	Proposed $k=b$	7.85	97.42	5.28	43.68	2.93	13.38	2.22	7.42
	Proposed $k=r/n$	7.86	97.77	5.30	43.89	2.94	13.51	2.24	7.53
30	Mean imputation	11.75	217.42	7.95	99.42	4.04	25.87	2.68	11.23
	Ratio imputation	10.76	183.40	7.27	83.03	3.69	21.50	2.44	9.29
	Singh and Horn	10.62	177.49	7.19	81.20	3.67	21.16	2.42	9.15
	Proposed $k=1$	9.03	128.92	6.06	57.74	3.08	14.94	2.02	6.39
	Proposed $k=b$	9.03	128.75	6.06	57.67	3.08	14.93	2.02	6.40
	Proposed $k=r/n$	9.04	128.99	6.06	57.77	3.08	14.94	2.02	6.38

Applications to air pollution data

The performance of the proposed estimators are compared with the existing estimators via applications to air pollution data in Chiang Mai, Thailand. Two populations of air pollution data in Chiang Mai from the Pollution Control Department [11] are considered as follows.

Population 1: We consider the hourly concentration of $PM_{2.5}$ and sulfur dioxide between 5 January 2023 and 4 February 2023, the concentration of $PM_{2.5}$ (micrograms per cubic) and sulfur monoxide (part per billion) were used as the auxiliary and study variables, respectively. The population parameters are summarized as

follows;
$N = 720,\ \bar{X} = 38.97,\ \bar{Y} = 1.34,\ C_X = 0.43,$ and $C_Y = 0.36,\ \rho = 0.41$.

Population 2: We consider the daily concentration of PM_{10} and $PM_{2.5}$ between 2019 and 2021, the concentration of PM_{10} (micrograms per cubic) and $PM_{2.5}$ (micrograms per cubic) were used as the auxiliary and study variables, respectively. The population parameters are summarized as follows;
$N = 1,089,\ \bar{X} = 44.42,\ \bar{Y} = 21.63,\ C_X = 0.71,$ and $C_Y = 0.98,\ \rho = 0.98$ .Four levels of sampling fractions are considered: 5%, 10%, 30%, and 50%. The percentage of missing data varied between 1% and 4%. The MSEs of the proposed and existing estimators are represented in Table 4 and the estimated value of sulfur dioxide and $PM_{2.5}$ are given in Table 5.

The results in Table 4 showed that the proposed estimators using all values of k gave the smallest MSEs with respect to all existing estimators for both populations, especially for population 2 we can see a big improvement in terms of MSEs for the proposed estimator. The proposed estimator at its optimum performs better than the existing estimators in these scenarios.

Table 5 illustrated the estimated value of sulfur dioxide for population 1 and the estimated $PM_{2.5}$ for population 2. The proposed estimators using all values of k gave the smallest MSEs with respect to all existing estimators for both populations. The proposed estimators gave closer estimated values with respect to other estimators, and as expected the estimated values for both sulfur dioxide and $PM_{2.5}$ are similar for a big sampling fraction.

Table 4. MSEs of the estimators for populations 1 and 2 respectively

Population	Estimator	Sampling fraction			
		0.05	0.1	0.3	0.5
1	Mean imputation	0.00641	0.00304	0.00080	0.00035
	Ratio imputation	0.00653	0.00310	0.00082	0.00036
	Singh and Horn (2000)	0.00637	0.00302	0.00079	0.00035
	Proposed	0.00538	0.00255	0.00067	0.00030
2	Mean imputation	7.99236	3.75132	0.97537	0.42171
	Ratio imputation	7.91713	3.71405	0.96295	0.41424
	Singh and Horn (2000)	7.91164	3.71133	0.96204	0.41370
	Proposed	0.39722	0.18854	0.05195	0.02471

Table 5. Estimated sulfur dioxide and $PM_{2.5}$ from the estimators for populations 1 and 2, respectively

Population	Estimator	Sampling fraction			
		0.05	0.1	0.3	0.5
1	Mean imputation	1.314	1.338	1.312	1.334
	Ratio imputation	1.339	1.359	1.317	1.337
	Singh and Horn	1.321	1.344	1.314	1.335
	Proposed $k=1$	1.355	1.379	1.327	1.333
	Proposed $k=b$	1.355	1.379	1.327	1.333
	Proposed $k=r/n$	1.355	1.379	1.327	1.333
2	Mean imputation	33.640	28.829	27.511	27.004
	Ratio imputation	35.971	30.039	27.913	27.216
	Singh and Horn	36.803	30.512	28.063	27.295
	Proposed $k=1$	23.791	26.050	26.882	27.329
	Proposed $k=b$	23.786	26.049	26.882	27.329
	Proposed $k=r/n$	23.787	26.050	26.882	27.329

Figs. 1 and 2 represented the percentage relative efficiencies of the estimators with respect to the mean imputation method for populations 1 and 2. The results showed that the proposed estimator performed the best and gave a lot higher PREs especially for population 2 when compared to other existing estimators. For the estimated value of sulfur dioxide from $PM_{2.5}$ for population 1, the PRE is higher than all existing estimators by at least 16%. For the estimated $PM_{2.5}$ from PM_{10} for population 2 the PRE is higher than all existing estimators by at least 1600% which are highly significant differences resulting in a huge improvement in estimation.

The results found in this study, at a similar sampling fraction level and a high correlation between the study and auxiliary variables are also similar to a recent one that investigated the reduction of bias and mean square error by changing the shape of the distribution using the transformed auxiliary variable in an application in Nan, Thailand [6]. Their results from an application to estimate carbon monoxide with $PM_{2.5}$ from air pollution data under SRSWOR illustrated a reduction in mean square errors between 91% and 100% with respect to the estimators using the non-transformed auxiliary variable in the case of complete cases of the study variable, which occur when ρ is smaller than a term that is proportionate to the ratio of the coefficient of variation of the auxiliary variable and the coefficient of variation of the study variable, using known ρ and C_x.

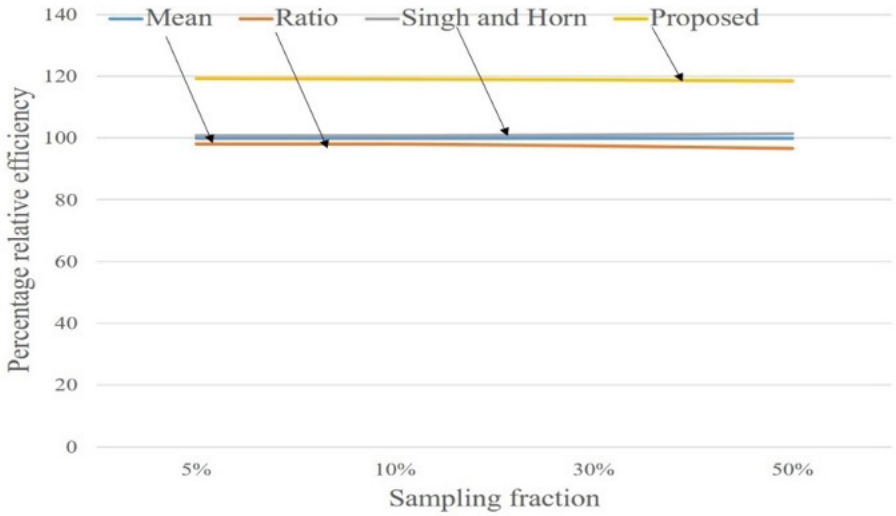

Fig. 1. Percentage relative efficiencies of the estimators for population 1

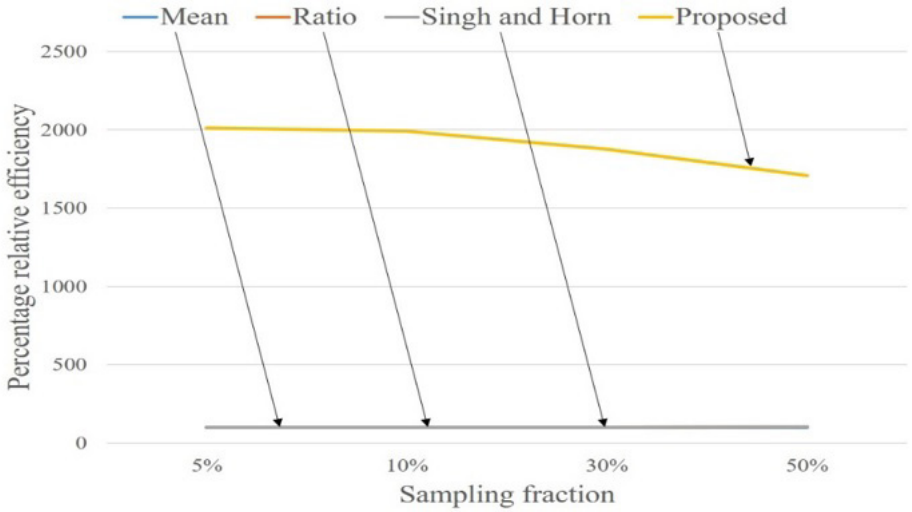

Fig. 2. Percentage relative efficiencies of the estimators for population 2

Some other previous similar studies were based on utilizing the transformed auxiliary variable which results in smaller biases and mean square errors under double sampling. The results based on application to estimate the mean $PM_{2.5}$ using nitrogen dioxide pollution data in Chiang Rai showed that the combined transformed estimators utilizing transformation on both the auxiliary and study variables gave a higher percentage relative efficiency in estimation compared to the single estimators in the case where the population mean of the auxiliary variable is unknown and using available ρ and C_x. [7].

Another result from transforming the auxiliary variable and both the auxiliary and study variables under double sampling to estimate the average yield of rubber in Thailand using the cultivated area also found that the combined transformed estimators gave a smaller mean square error than the existing estimators at least 1.5 times with

respect to existing ones using known ρ and C_x. [9]. In the case that the population mean of the auxiliary variable is unknown therefore it needs to be estimated such as conducting a survey based on double sampling.

In our study, the results found in the application to air pollution data also showed that the transformed estimators reduced the mean square error and gave a lot higher efficiency by at least 1600% compared to existing estimators. We can see that using the benefit of a transformed variable can assist by increasing the performance of the population mean estimator. In addition, the proposed estimators in the current study can be applied in the presence of missing observations which is likely to occur in practice for air pollution data and other data which can be seen in the previous study.

This work studied only when the study variable is missing and non-response occurs uniformly and assumed that the population mean of the auxiliary variable is available. On the other hand, this estimator can be implicated in missing at random and not missing at random. Achievement of the most accurate estimation of pollution induce policymakers to put measures in place to reduce the prevailing pollution and instigate plans to mitigate future problems. Estimation provides valuable information to be utilized extensively and also keep track of progress of pollution reduction projects. The proposed estimators can also be applied in the case of unknown population mean of the auxiliary variable under double sampling and other designs in future work.

Conclusion

A new estimator for estimating population mean based on the transformation of the auxiliary variable has been proposed in the case of missing data under SRSWOR. The transformation method is used to change the shape of the auxiliary variable when missing data appear in the variable of interest. We suggested using the constant α to minimize the MSE of the proposed estimator along with k, such as the sample regression coefficient and the response rate, which does not rely on the known parameters in the study. Upon using the transformed auxiliary variable to impute the missing observations, the biases and mean square errors are studied up to the first order of approximation under the uniform response mechanism. The simulation results and the applications to air pollution data in Chiang Mai, Thailand showed that all the proposed estimators with any values of k gave better results when compared to the existing estimators for both bias and MSE at all levels of missing values and all levels of sampling fractions. Any values of k can be used to get the minimum MSE based on the proposed estimator. A larger sample size can increase the percentage relative efficiency of the estimators. We can see that the proposed imputation technique based on the transformed auxiliary variable can be helpful for imputing missing values and improving the efficiency of the estimators. In future research, the transformation method can be used to transform both the study and auxiliary variables and can also be applied in more complex survey designs. Nevertheless, the proposed estimator can be applied to real world problems where missing data exist and can contribute by imputing the missing observations before processing to further analysis to gain precise results based on sets of data. This method can be applied to an abundance of data in a diverse variety of fields such as other environmental data on water demand, medical data on diseases like COVID-19 or non-communicable diseases, economical, agricultural, and societal data.

Financial supports

This research was funded by King Mongkut's University of Technology North Bangkok, Contract no. KMUTNB-66-BASIC-12.

Competing interests

The authors declare no competing interests.

Acknowledgements

We appreciate the Pollution Control Department for providing the air pollution data.

Ethical considerations

Ethical issues (Including plagiarism, Informed Consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc) have been completely observed by the authors.

References

1. Singh S, Horn S. Compromised imputation in survey sampling. Metrika. 2000 Sep 7;51(3):267–76. Available from: https://doi.org/10.1007/s001840000054.

2. Chodjuntug K, Lawson N. Imputation for estimating the population mean in the presence of nonresponse, with application to fine particle density in Bangkok. Mathematical Population Studies. 2022 Oct 2;29(4):204-25. Available from: https://doi.org/10.1080 /08898480.2021.1997466.

3. Chodjuntug K, Lawson N. A Chain regression exponential type imputation method for mean estimation in the presence of missing data. Songklanakarin Journal of Science and Technology, 2022
Jul; 44(4): 1109-1118.

4. Ponkaew C, Lawson N. A New Estimator for Population Total in the Presence of Missing Data Under Unequal Probability Sampling Without Replacement: A Case Study on Fine Particulate Matter in the North of Thailand. Songklanakarin Journal of Science and Technology. 2023.

5. Srivenkataramana T. A dual to ratio estimator in sample surveys. Biometrika. 1980 April 1; 67(1): 199-204. Available from: https://doi.org/10.2307/2335334

6. Thongsak N, Lawson N. Bias and mean square error reduction by changing the shape of the distribution of an auxiliary variable: application to air pollution data in Nan, Thailand. Mathematical Population Studies. 2023; 30(3): 180–194. Available from: https://doi.org/10.1080 /08898480.2022.2145790

7. Thongsak N, Lawson N. Classes of combined population mean estimators utilizing transformed variables under double sampling: an application to air pollution in Chiang Rai, Thailand. Songklanakarin. Journal of Science and Technology. 2022 Sep;44(5): 1390-1398.

8. Lawson N. An improved family of estimators for estimating population mean using a transformed auxiliary variable under double sampling. Songklanakarin Journal of Science and Technology. 2023.

9. Thongsak N, Lawson N. Classes of Population Mean Estimators using Transformed Variables in Double Sampling. Gazi University Journal of Science. 2023; 36(4): 1834-1852.

10. Team RC. R: The R project for statistical computing. https. www. r-project. org. 2016.

11. Pollution Control Department: Air4Thai [Internet]. Air4thai.pcd.go.th. [cited 2023 Jul 19]. Available from: http://air4thai.pcd.go.th/webV2/history