Current Applied Science and Technology Vol. 23 No. 5 (September-October 2023)

Research article

Estimating the Mean of $PM_{2.5}$ with Missing Data in the Area Around Electricity Generating Authority of Thailand Using the Improved Compromised Imputation Method

Tanart Dachochaiporn and Kanisa Chodjuntug*

Department of Mathematics, Statistics and Computer, Faculty of Science, Ubon Ratchathani University, Ubon Ratchathani, Thailand

Received: 22 July 2022, Revised: 4 November 2022, Accepted: 16 February 2023

DOI: 10.55003/cast.2023.05.23.013

Abstract

Keywords

particulate matter 2.5;

missing data;

mean estimation;

imputation method;

auxiliary variable

Particulate matter with an aerodynamic diameter of less than 2.5 µm or $PM_{2.5}$, is one of the air pollutants that has been found to be at unsafe levels for a number of years in Thailand, leading to public health concerns. In order to lessen the detrimental effects of air pollution, monitoring and analysis of $PM_{2.5}$ concentration are crucial. Following the study of data from the Pollution Control Department Report in the area around the Electricity Generating Authority of Thailand in January 2019, it was found that there was data missing in the $PM_{2.5}$ information. It is well-known that missing data can reduce the accuracy of data analysis. To solve the missing data problem, this paper proposes an improved method of compromised imputation and a corresponding resultant estimator to deal with estimating the mean of $PM_{2.5}$ concentrations in the area. The bias and mean square error of the estimator obtained from the proposed method were derived. The conditions which favor the performance of our estimator over other estimators obtained from the mean, ratio, and compromised imputation methods were obtained using mean square error to apply in the area. The mean of $PM_{2.5}$ concentrations in this case using the proposed estimator was equal to 47.13 µg/m^3, indicating that it did not exceed unsafe levels (≤ 50 µg/m^3) under certain conditions. In order to support more accurate data analysis that will lead to effective management of air pollution problems in the future, this research proposes a new method that is more effective than the existing methods under missing data problem.

*Corresponding author: Tel.: (+66) 45353401
E-mail: kanisa.c@ubu.ac.th

1. Introduction

Many areas in Thailand have experienced air pollution issues such as particulate matter with an aerodynamic diameter of less than 2.5 $(PM_{2.5})$ and carbon monoxide (CO) in recent years. Thailand's air pollution has been found to be at hazardous levels that have concerned public health investigators. $PM_{2.5}$ and CO can cause dizziness, respiratory diseases, cardiovascular and neurodegenerative disorders, miscarriages, and damage to the developing fetus [1- 4]. In January 2018, Bangkok and neighboring business and heavily populated districts were found to have $PM_{2.5}$ concentrations that exceeded the standard of the PCD in Thailand (≤ 50 μg/m^3). Furthermore, the World Health Organization (WHO) determined that $PM_{2.5}$ is a group 1 agent (mixture) that is carcinogenic to humans, accounting for 1 in every 8 premature deaths in globally in 2013. WHO demonstrated that the problems related to air pollution can affect both health and the economy. The main sources of the Thailand's air pollution are vehicle fumes, industrial emissions and crop burning. Therefore, air quality monitoring and analysis are critical for mitigating the negative effects of air pollution. Thailand's Pollution Control Department (PCD) is an agency that monitors air pollution levels. According to a PDC report [5], the area around the Electricity Generating Authority of Thailand in Nonthaburi Province is one of the Bangkok Metropolitan Region's rapidly urbanizing areas, where the government is concerned about air pollution (particulate matter) because the area's air quality index (AQI) was at an unhealthy level for months in 2018-2019. Increasing the area's air pollution is caused by rising emissions from power plants, vehicles, and industry, which when combined with a high rate of power generation, contributes to high levels of air pollution. Furthermore, the area has experienced peak air pollution during the dry season, from January to March, due to consistently low wind speeds, a low rain scavenging rate, and low temperatures [6, 7]. As a result, the government decided to make air pollution a national priority for 2019. In addition, the government has announced plans to address urgent issues in these and surrounding areas by methods including the closing of educational institutions to protect children (who are considered a high-risk group and may suffer long-term health consequences), and the spraying of water into the air from tall buildings or planes in high-traffic areas. Besides, the government requested cooperation from various agencies in the public, private, and general public sectors in order to reduce vehicle use in conjunction with other measures to effectively decrease overall dust levels. Based on the report's study information on air pollutants, we discovered that $PM_{2.5}$ concentration data in January 2019 contained missing values. It is a well-known fact that if data suffers substantially from missing values, the efficiency of data analysis based on that data is reduced. The imputation method is an effective method for handling missing or non-response issues by substituting missing data with other relevant and available data. A corresponding point estimator of the population mean is also obtained from the imputation method. In fact, estimating the population mean is a fundamental statistical tool of data analysis in general research. Furthermore, auxiliary information can improve the precision of the estimator, so several authors including Bahl and Tuteja [8], Singh and Pal [9], Jaroengeratikun and Lawson [10] defined the estimator of population mean using auxiliary information under a simple random sample without replacement scheme (SRSWOR).

In the literature concerned with imputation methods under missing data scenarios, many authors like Rubin [11], Kalton *et al.* [12], Rao and Sitter [13], Lee *et al.* [14], Singh and Deo [15], Ahmed *et al.* [16], Kadilar and Cingi [17], Singh and Horn [18], and Singh *et al.* [19] discussed how to make the data complete. In addition, some authors demonstrated that the use of auxiliary information can help improve the imputation method for estimating the population mean. Lee *et al.* [14] suggested using the mean and ratio methods as estimators to apply the imputation method. Their idea is to substitute missing values in the dataset with these estimators. Singh and Deo [15] and Ahmed *et al.* [16] suggested an improved imputation method using a power transformation technique. Kadilar and Cingi [17] suggested the chain technique, which is the application of a regression-type estimator into the ratio imputation method in the presence of missing data under a simple random sampling design. Singh and Horn [18] were the first to propose the compromised imputation method for estimating the study variable's population mean. Their findings revealed that the performance of a compromised imputation method involves selection constants that produce the estimator's minimum mean square error. Later, Singh *et al.* [19] proposed exponential type compromised imputation methods based on the population mean of the auxiliary variable, but the population mean of the auxiliary variable is generally unknown in practice.

In this paper, we propose an improved compromised imputation method and corresponding estimator for estimating population mean of the study variable under the SRSWOR scheme in order to apply the estimated population mean of $PM_{2.5}$ in the area under the unknown population mean of the auxiliary variable. Our method is enhanced by taking in consideration both a transformed auxiliary variable with chain ratio exponential and new weight constants. The properties of the proposed estimator: bias and mean square error of the proposed estimator, were derived using the Taylor series up to the first degree of approximation. In order to identify the optimal conditions in which the proposed estimator performs best, the proposed estimator's performance was compared to that of other estimators produced from mean, ratio, and compromised imputation methods using the mean square error of estimators. To solve the problems, the estimator was used to estimate the population's mean $PM_{2.5}$ concentration as a study variable based on CO concentration as a study auxiliary variable in the area around the Electricity Generating Authority of Thailand in Nonthaburi Province in January 2019.

2. Materials and Methods

In this section, we present the structure and notations, some existing imputation methods, a proposed imputation method, and corresponding point estimators for the population mean under the SRSWOR scheme. Its properties are also shown, including bias and mean square error.

2.1 Structure and notations

Let $U=\{U_1, U_2,\ldots, U_N\}$ be a finite population of size *N*. The aim is to estimate the population mean $\bar{Y}$ of variable y taking values y_i in the existence of auxiliary variable x taking values x_i. Let R and R^c be the set of responding units and non-responding units, respectively. Under the SRSWOR scheme, s of size n with paired variable (x, y) is drawn from U and contains both r responding units and $(n-r)$ non-responding units. The values of y_i are observed for units $i \in R$; meanwhile, the value of y_i is missing for units $i \in R^c$ and imputed values with x_i.

In this work, we have emulated the notation of Rao [20] as follows:

$\bar{Y}=\frac{\sum_{i=1}^{N} y_i}{N}$, $\bar{X}=\frac{\sum_{i=1}^{N} x_i}{N}$: The population mean of variables y and x, respectively.

$\bar{y}_r=\frac{\sum_{i=1}^{r} y_i}{r}$, $\bar{x}_r=\frac{\sum_{i=1}^{r} x_i}{r}$: The response means of variables y and x, respectively.

$\bar{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}$: The sample mean of x.

$S_y^2 = \frac{\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2}{N-1}$, $S_x^2 = \frac{\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2}{N-1}$: The population variance of variables y and x, respectively.

$S_{xy} = \frac{\sum_{i=1}^{N}\left(x_i - \bar{X}\right)\left(y_i - \bar{Y}\right)}{N-1}$: The population covariance between y and x.

$C_y = \frac{S_y}{\bar{Y}}$, $C_x = \frac{S_x}{\bar{X}}$: The population coefficient of variations of variables y and x, respectively.

$\rho = \frac{S_{xy}}{S_x S_y}$: The population correlation coefficient between y and x.

2.2 Existing imputation methods and corresponding estimators

The mean imputation method, which is the imputation scheme, is defined as

$$y_{.i} = \begin{cases} y_i;\ i \in R \\ \bar{y}_r;\ i \in R^c. \end{cases} \tag{1}$$

Under this imputation method, its corresponding point estimator of the population mean becomes

$$\bar{y}_M = \frac{\sum_{i \in r} y_{.i}}{r} = \bar{y}_r. \tag{2}$$

The bias and mean square error of $\bar{y}_M$ are obtained as

$$Bias\left(\bar{y}_M\right) = 0, \tag{3}$$

$$MSE\left(\bar{y}_M\right) = \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2 C_y^2. \tag{4}$$

With the ratio imputation method, the imputation scheme is defined as

$$y_{.i} = \begin{cases} y_i\ \ ;\ i \in R \\ \hat{b}x_i;\ i \in R^c, \end{cases} \tag{5}$$

where $\hat{b} = \dfrac{\sum\limits_{i \in R} y_i}{\sum\limits_{i \in R^c} x_i}$.

Under this imputation method, its corresponding point estimator of the population mean becomes

$$\bar{y}_{RAT} = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r}. \tag{6}$$

The bias and mean square error of $\bar{y}_{RAT}$ are derived as

$$Bias\left(\bar{y}_{RAT}\right) = \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}C_y^2\left(1 - \rho\frac{C_y}{C_x}\right), \tag{7}$$

$$MSE\left(\bar{y}_{RAT}\right) = \bar{Y}^2\left[\left(\frac{1}{r} - \frac{1}{N}\right)C_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\left(C_x^2 - 2\rho C_x C_y\right)\right]. \tag{8}$$

The compromised imputation method suggested by Singh and Horn [18] is as follows:

$$y_{.i} = \begin{cases} \alpha \dfrac{n}{r} y_i + \left(1 - \alpha\right)\hat{b}x_i \, ; \, i \in R \\ \left(1 - \alpha\right)\hat{b}x_i \qquad\qquad ; \, i \in R^c , \end{cases} \tag{9}$$

where α is a constant and real number. Recently, Audu and Singh [21] proposed $\alpha = \dfrac{r}{n}$, which is a very straightforward into method type compromised imputation method.

Under this imputation method, its corresponding point estimator of the population mean becomes

$$\bar{y}_{COMP} = \alpha\bar{y}_r + \left(1 - \alpha\right)\bar{y}_r \frac{\bar{x}_n}{\bar{x}_r}. \tag{10}$$

The bias and mean square error of $\bar{y}_{COMP}$ are respectively given by

$$Bias\left(\bar{y}_{COMP}\right) = \left(1 - \alpha\right)\left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}C_x^2\left(1 - \rho\frac{C_y}{C_x}\right), \tag{11}$$

$$MSE\left(\bar{y}_{COMP}\right) = \bar{Y}^2\left\{\left(\frac{1}{r} - \frac{1}{N}\right)C_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\left[\left(1 - \alpha\right)^2 C_x^2 - 2\left(1 - \alpha\right)\rho C_x C_y\right]\right\}. \tag{12}$$

2.3 Proposed imputation method and corresponding estimation

We propose a new imputation method that was inspired by Singh and Horn [18] and Singh and Pal [9], employing a modified auxiliary variable as a chain ratio exponential and new constants (w_1, w_2) to enhance the method. Next, we substitute w_1 for α and w_2 for $(1-\alpha)$ into the compromised imputation method as equation (9). After imputation, the data is defined as

$$y_{.i} = \begin{cases} w_1 \dfrac{n}{r} y_i + w_2 \dfrac{\bar{y}_r}{\bar{x}_r} x_i \exp\left(\dfrac{\bar{x}_n - \bar{x}_r}{\bar{x}_n + \bar{x}_r}\right); i \in R \\ w_2 \dfrac{\bar{y}_r}{\bar{x}_r} x_i \qquad ; i \in R^c, \end{cases} \tag{13}$$

where both w_1 and w_2 are constants represented in the set of real numbers, under which w_1 and w_2 are appropriately chosen to achieve the minimum MSE of estimator.

The proposed method's corresponding population mean point estimator is shown below.

$$\bar{y}_{IM} = w_1 \bar{y}_r + w_2 \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} \exp\left(\frac{\bar{x}_n - \bar{x}_r}{\bar{x}_n + \bar{x}_r}\right), \tag{14}$$

Note that if $w_1 = 1$ and $w_2 = 0$ then $\bar{y}_{IM} = \bar{y}_r$ and if $w_1 = 0$ and $w_2 = 1$ then the $\bar{y}_{IM}$ is the analogue of the estimator for the population mean proposed by Singh and Pal [9].

The bias and mean square error of the estimator $\bar{y}_{IM}$ are both obtained by the Taylor series up to the first degree of large sample approximation to achieve its properties.

Let $\bar{y}_r = \bar{Y}(1+e_1)$, $\bar{x}_n = \bar{X}(1+e_2)$, $\bar{x}_r = \bar{X}(1+e_3)$ be defined to easily derive its properties.

Under the SRSWOR scheme, we have the expectation as follows:

$E(e_i) = 0;\ i = 1,2,3,\ E(e_1^2) = M_1 C_y^2,\ E(e_2^2) = M_2 C_x^2,\ E(e_3^2) = M_1 C_x^2,$

$E(e_1 e_2) = M_2 \rho C_x C_y,\ E(e_2 e_3) = M_2 C_x^2$ and $E(e_1 e_3) = M_1 \rho C_x C_y,$

where $M_1 = \dfrac{1}{r} - \dfrac{1}{N}$, $M_2 = \dfrac{1}{n} - \dfrac{1}{N}$ and $M_3 = M_1 - M_2 = \dfrac{1}{r} - \dfrac{1}{n}$.

Writing $\bar{y}_{IM}$ from equation (14) in term of e_i 's, we have

$$\bar{y}_{IM} = w_1 \bar{Y}(1+e_1) + w_2 \bar{Y}(1+e_1)\left[\frac{\bar{X}(1+e_2)}{\bar{X}(1+e_3)}\right] \exp\left[\frac{\bar{X}(1+e_2) - \bar{X}(1+e_3)}{\bar{X}(1+e_2) + \bar{X}(1+e_3)}\right]. \tag{15}$$

From equation (15), expanding and retaining the terms up to the first degree of approximation, we have

$$\bar{y}_{IM} - \bar{Y} \cong \bar{Y}\left[w_1 + w_2 - 1 + w_1 e_1 + w_2 e_1 + w_2 \begin{pmatrix} \frac{3}{2} e_2 - \frac{3}{2} e_3 + \frac{3}{2} e_1 e_2 - \frac{3}{2} e_1 e_3 \\ -\frac{9}{4} e_2 e_3 + \frac{3}{8} e_2^2 + \frac{15}{8} e_3^2 \end{pmatrix} \right] \tag{16}$$

and then taking the expectation on both sides of equation (16), we get the following $Bias\left(\bar{y}_{IM}\right)$:

$$Bias\left(\bar{y}_{IM}\right) = \bar{Y}\left[\left(w_1 + w_2 - 1\right) + w_2 M_3 \left(\frac{15}{8} C_x^2 - \frac{3}{2} \rho C_x C_y \right) \right]. \tag{17}$$

$MSE\left(\bar{y}_{IM}\right)$ is obtained by squaring both sides of equation (16), expanding the term, taking expectations, and keeping the terms up to the first degree of approximation. Therefore, we get $MSE\left(\bar{y}_{IM}\right)$ as follows:

$$MSE\left(\bar{y}_{IM}\right) = \bar{Y}^2 \begin{bmatrix} \left(w_1 + w_2 - 1\right)^2 + \left(w_1 + w_2\right)^2 M_1 C_y^2 \\ -3\left(w_1 + w_2\right) w_2 M_3 \rho C_x C_y + \frac{9}{4} w_2^2 M_3 C_x^2 \end{bmatrix}. \tag{18}$$

Since $\bar{y}_{IM}$ as given in equation (14), the constants w_1 and w_2 are unknown, so we need to search for w_1 and w_2 optimum values in order to obtained minimum $MSE\left(\bar{y}_{IM}\right)$. We differentiate equation (18) with respect to w_1 and w_2, respectively and equate it to zero as follows:

$$\frac{\partial MSE\left(\bar{y}_{IM}\right)}{\partial w_1} = \bar{Y}^2 \begin{bmatrix} 2\left(w_1 + w_2 - 1\right) + 2\left(w_1 + w_2\right) M_1 C_y^2 \\ -3\left(1 + w_2\right) w_2 M_3 \rho C_x C_y \end{bmatrix} = 0, \tag{19}$$

$$\frac{\partial MSE\left(\bar{y}_{IM}\right)}{\partial w_2} = \bar{Y}^2 \begin{bmatrix} 2\left(w_1 + w_2 - 1\right) + 2\left(w_1 + w_2\right) M_1 C_y^2 \\ -3\left(w_2 + 2w_2\right) M_3 \rho C_x C_y + \frac{9}{2} w_2 M_3 C_x^2 \end{bmatrix} = 0. \tag{20}$$

On solving equations (19) and (20), we get

$$w_1 = \frac{3C_x - 2\rho C_y}{3C_x\left(M_1 C_y^2 - M_3 \rho^2 C_y^2 + 1\right)} = \left(w_1\right)_{opt}, \tag{21}$$

$$w_2 = \frac{2\rho C_y}{3C_x\left(M_1 C_y^2 - M_3 \rho^2 C_y^2 + 1\right)} = \left(w_2\right)_{opt}. \tag{22}$$

To find minimum $MSE\left(\bar{y}_{IM}\right)$, we put equation (21) and equation (22) into equation (18), we get

$$MSE\left(\bar{y}_{IM}\right)_{opt} = \bar{Y}^2\left[1+\frac{1}{C_y^2\left(M_1 - M_3\rho^2\right)}\right]^{-1} = \min MSE\left(\bar{y}_{IM}\right). \tag{23}$$

Typically, we are unable to find the parameters such as S_y^2, S_x^2, S_{xy}, C_y, C_x and ρ. Therefore, we approximate them using $s_y^{*2} = \frac{\sum_{i=1}^{r}\left(y_i - \bar{y}_r\right)^2}{r-1}$, $s_x^{*2} = \frac{\sum_{i=1}^{r}\left(x_i - \bar{x}_r\right)^2}{r-1}$, $s_{xy}^{*} = \frac{\sum_{i=1}^{r}\left(x_i - \bar{x}_r\right)\left(y_i - \bar{y}_r\right)}{r-1}$, $c_y = \frac{s_y^*}{\bar{y}_r}$, $c_x = \frac{s_x^*}{\bar{x}_r}$ and $\hat{\rho} = \frac{s_{xy}^*}{s_x^* s_y^*}$, respectively.

2.4 Efficiency comparison of the proposed estimator

Efficiency comparison of the proposed estimator with the existing estimators is presented by theoretical results. Under the expressions of MSE 's estimators, the following results show conditions under which the proposed estimator performs better than the others.

2.4.1 Comparison of $\bar{y}_{IM}$ with $\bar{y}_M$

We obtain $MSE\left(\bar{y}_{IM}\right) - MSE\left(\bar{y}_M\right) < 0$,

$$\text{if} \quad \rho > \frac{\left(w_1 + w_2 - 1\right)^2 + \left[\left(w_1 + w_2\right)^2 - 1\right]M_1 C_y^2 + \frac{9}{4}w_2^2 M_3 C_x^2}{3\left(w_1 + w_2\right)w_2 M_3 C_x C_y}. \tag{24}$$

Therefore, when the above condition is satisfied, the performance of $\bar{y}_{IM}$ is better than $\bar{y}_M$.

2.4.2 Comparison of $\bar{y}_{IM}$ with $\bar{y}_{RAT}$

We obtain $MSE\left(\bar{y}_{IM}\right) - MSE\left(\bar{y}_{RAT}\right) < 0$

$$\text{if} \quad \rho < \frac{\left(w_1 + w_2 - 1\right)^2 + \left[\left(w_1 + w_2\right)^2 - 1\right]M_1 C_y^2 + \left(\frac{9}{4}w_2^2 - 1\right)M_3 C_x^2}{\left[3\left(w_1 + w_2\right)w_2 - 2\right]M_3 C_x C_y}. \tag{25}$$

Therefore, when the above condition is satisfied, the performance of $\bar{y}_{IM}$ is better than $\bar{y}_{RAT}$.

2.4.3 Comparison of $\bar{y}_{IM}$ with $\bar{y}_{COM}$

We obtain $MSE\left(\bar{y}_{IM}\right) - MSE\left(\bar{y}_{COM}\right) < 0$

$$\text{if } \rho < \frac{(w_1+w_2-1)^2+\left[(w_1+w_2)^2-1\right]M_1C_y^2+\left[\frac{9}{4}w_2^2-(1-\alpha)^2\right]M_3C_x^2}{\left[3(w_1+w_2)w_2-2(1-\alpha)\right]M_3C_xC_y}. \tag{26}$$

Therefore, when the above condition is satisfied, the performance of $\bar{y}_{IM}$ is better than $\bar{y}_{COM}$.

3. Results and Discussion

Simulation and case studies were conducted using R software to compare the performance of the proposed estimator with other estimators obtained from imputation methods in order to support the theoretical findings.

3.1 Simulation study

The details of the simulation study are that variables (y,x) were generated from a bivariate normal distribution with population size $N=700$, means $(50,\ 2)$ and variances $(400,\ 0.5)$, while the correlation coefficient was 0.6. A random sample of $n=500$ was selected from this population using SRSWOR. We considered cases where 25% of the data from the variable of y was missing completely at random. The simulation was repeated 10,000 times. The percent relative efficiency (PRE) of each estimator $(\bar{y})$ with respect to $\bar{y}_M$ was used an indicator to consider the performance of estimators. It can be computed from $PRE(\bar{y})=\dfrac{MSE(\bar{y}_M)}{MSE(\bar{y})}\times 100$. The highest PRE indicates the most efficient estimator. The results of the simulation study in the form of the estimator's PRE are reported in Table 1.

Table 1. *PRE* of each estimator based on the simulation study

Estimators	***PRE***
$\bar{y}_M$	100
$\bar{y}_{RAT}$	136.4
$(\bar{y}_{COMP})_{\frac{r}{n}}$	127.6
$(\bar{y}_{IM})_{opt}$	150.0

The results in Table 1 reveal that $(\bar{y}_{IM})_{opt}$ provided the highest PRE . Therefore, the $(\bar{y}_{IM})_{opt}$ is more efficient than the others, meaning that these simulation results are in agreement with the theoretical results, and implying that $(\bar{y}_{IM})_{opt}$ is the best estimator.

3.2 Case study

The $PM_{2.5}$ (μg/m^3) and CO (ppm) data on a time-scale of one per hour (hourly average) in the Electricity Generating Authority of Thailand, Nonthaburi, Thailand was obtained from the PCD website in January 2019. The data contains a population of 744 units. Based on the study of information, we found that 23% of the data for $PM_{2.5}$ concentration was missing, so it was taken as the variable y while the CO data was taken as an auxiliary variable x.

Figure 1 shows the distributions of $PM_{2.5}$ and CO concentrations. It was found that its distribution pattern corresponded to the positive relationship between $PM_{2.5}$ and CO, as seen in Figure 2. Based on response dataset, it was demonstrated that $\hat{\rho}$ is 0.6 as the estimate of population correlation coefficient between $PM_{2.5}$ and CO (ρ) in order to apply the ρ checking requirement of condition in Section 2.4. Moreover, the values were obtained for the considered variables including $N = 744$, $n = 678$, $r = 523$, $\bar{y}_r = 49.12$, $\bar{x}_n = 1.11$, $\bar{x}_r = 1.17$, $s_y^* = 21.49$, $s_x^* = 0.29$.

In addition, we computed the constants $w_1 = 0.34$ and $w_2 = 0.65$, as well as the ρ conditions under which the proposed the proposed estimator $(\bar{y}_{IM})_{opt}$ outperforms the others in order to verify the theoretical conclusions in Section 2.4. The results are shown in Table 2 and 3.

By using the dataset, the results from Table 2 demonstrate that the conditions are satisfied, thereby confirming the superior performance of the proposed estimator over the existing estimators and supporting the $(\bar{y}_{IM})_{opt}$ application in the case study. From Table 3, we found that the $(\bar{y}_{IM})_{opt}$ has the highest PRE so it outperformed the other estimators. We discovered that the conclusions drawn from the conditions presented in Table 2 were supported by the results from Table 3.

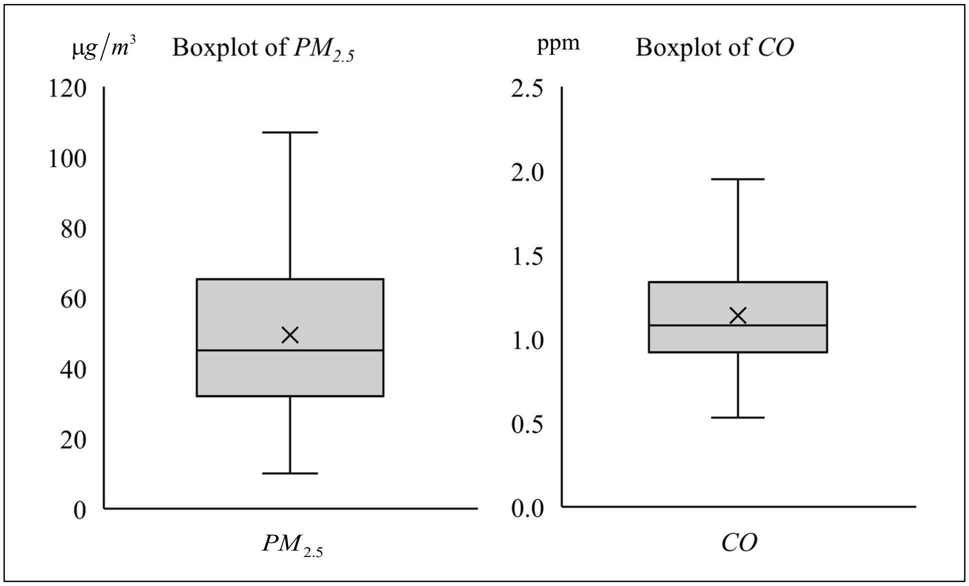

Figure 1. Boxplots of $PM_{2.5}$ (μg/m^3) and CO (ppm)

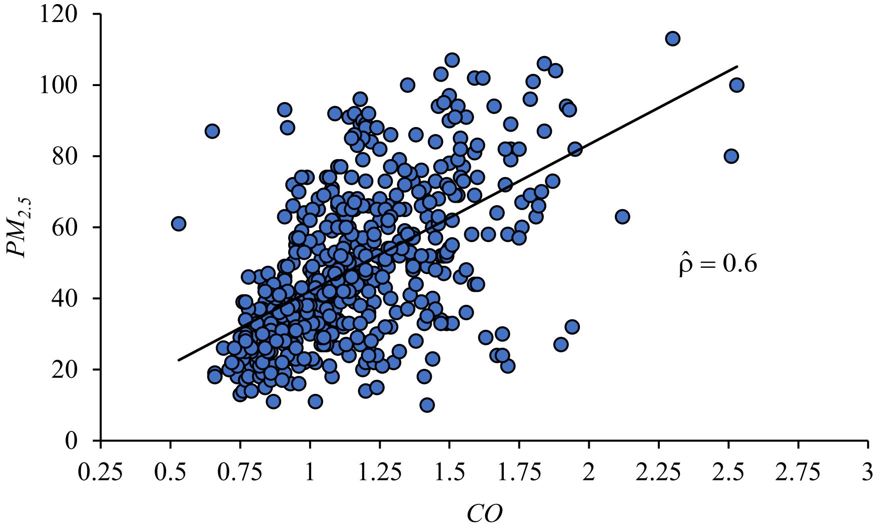

Figure 2. Scatter plot showing relationship between $PM_{2.5}$ (μg/m^3) and CO (ppm)

Table 2. The conditions of ρ for which $(\bar{y}_{IM})_{opt}$ is better than others

Estimators	Condition of ρ for which $(\bar{y}_{IM})_{opt}$ is better than:
$\bar{y}_M$	$\rho > 0.4$
$\bar{y}_{RAT}$	$\rho < 1$
$(\bar{y}_{COMP})_{\frac{r}{n}}$	$\rho < 0.7$

Note: $\rho \approx 0.6$ is estimated by $\hat{\rho}$.

Table 3. *PRE* of each estimator based on the case study

Estimators	*PRE*
$\bar{y}_M$	100
$\bar{y}_{RAT}$	129.4
$(\bar{y}_{COMP})_{\frac{r}{n}}$	115.8
$(\bar{y}_{IM})_{opt}$	146.7

Therefore, it can be concluded that $(\bar{y}_{IM})_{opt}$ is appropriate for estimating the mean of $PM_{2.5}$ with missing data in this case study. The $PM_{2.5}$ mean value by $(\bar{y}_{IM})_{opt}$ is equal to 47.13 μg/m^3 and $MSE(\bar{y}_{IM})_{opt} = 0.15$ (μg/m^3)^2, which is closed to zero indicating the superior quality of $(\bar{y}_{IM})_{opt}$.

According to the findings, the relationship between $PM_{2.5}$ and CO had a positive correlation which was confirmed by Fu *et al*. [22], and this meant that $PM_{2.5}$ and CO were trending in the same direction. Using the proposed strategy, the mean value of $PM_{2.5}$ in the January 2019 case study did not exceed Thailand's average PCD standard $\left(\leq 50\,\mu g/m^3\right)$. However, it was nearly 50 μg/m^3 and exceeded the WHO annual guideline value for $PM_{2.5}$ $\left(\leq 25\,\mu g/m^3\right)$. The findings are consistent with the research of Chinwetkitvanich *et al*. [23], who reported that $PM_{2.5}$ levels in this area exceeded the WHO guideline during January. In addition, the $PM_{2.5}$ concentrations under consideration were collected during the summer in January, and the research of Sun *et al*. [24] and Zhao *et al*. [25] confirmed that summertime $PM_{2.5}$ concentrations were higher than other seasons. Due to the dry climate with little wind and rain during the season, there is insufficient humidity and no wind speed to reduce the $PM_{2.5}$ concentration [26, 27]. Furthermore, excessive fuel combustion is the primary contributor to $PM_{2.5}$ issues in this area, most of which comes from the power plant and heavy traffic. People's health can suffer if they are exposed for an extended period of time in the area. Therefore, our finding can support the government's plan to reduce the air pollution levels and thus reduce the impact that air pollution will have on the residents in the future.

4. Conclusions

We proposed an improved method of compromised imputation by taking into consideration both the auxiliary variable transformation as a chain ratio exponential technique and the two constants $\left(w_1, w_2\right)$ whose conditions can be written as $w_1 + w_2 \neq 1$ and $w_1 + w_2 = 1$. The traditional compromise imputation method, on the other hand, is limited to the condition $w_1 + w_2 = 1$. In order to obtain an estimator for estimating the population mean in the presence of missing data, the proposed method was utilized. We offered the optimum values of constants to make the minimum mean square error. Moreover, the minimum mean square error as a formula was presented to measure the estimator quality. As a result, it is easy to implement for research. The conditions under which the performance of our estimator over other estimators obtained from mean, ratio, and compromised imputation methods were derived. The case study findings supported the obtained conditions. It was shown that the performance of our method was better than the other estimators when conditions were satisfied. To estimating the mean of $PM_{2.5}$, the calculated $PM_{2.5}$ concentration was 47.13 μg/m^3, which was almost at an unsafe level capable of affecting public health in the future because pollution concentrations tend to increase due to the continuous expansion of the industrial sector. Therefore, there are higher amounts of various toxins as well. It is very important to monitor pollution levels and validate the analysis of $PM_{2.5}$ concentrations. However, in this study, CO was used as an auxiliary variable for estimating mean $PM_{2.5}$. In the event that CO data cannot be collected, PM_{10}, O_3, or other quantitative data correlated with $PM_{2.5}$ may be used for estimating the mean value of $PM_{2.5}$. In addition, in the future, further study on the different relationships between variables y and x, which can affect the efficiency of the estimator should be performed. Moreover, researchers can apply our work as a guideline to solve the missing data problem in order to provide a complete amount of data. Our method can also save time and money for researchers. The proposed method has a limitation in that it only employs a simple random sampling design, and only one variable has missing data, and it is the study variable. Therefore, we recommend the proposed method to estimate the population mean when only the study variable contains missing data under the SRSWOR scheme in practice.

References

[1] Zhang, Y., Wang, S.G., Ma, Y.X., Shang, K.Z., Cheng, Y.F., Li, X., Ning, G.C., Zhao, W.J. and Li, N.R., 2015. Association between ambient air pollution and hospital emergency admissions for respiratory and cardiovascular diseases in Beijing: a time series study. *Biomedical and Environmental Sciences*, 28(5), 352-363, DOI: 10.3967/bes2015.049.

[2] Xing, Y.F., Xu, Y.H., Shi, M.H. and Lian, Y.X., 2016. The impact of $PM_{2.5}$ on the human respiratory system. *Journal of Thoracic Disease,* 8(1), 69-74, DOI: 10.3978/j.issn.2072-1439.2016.01.19.

[3] Rose, J.J., Wang, L., Xu, Q., Mctiernan, C.F., Shiva, S., Tejero, J. and Gladwin, M.T., 2017. Carbon monoxide poisoning: pathogenesis, management and future directions of therapy. *American Journal of Respiratory and Critical Care Medicine*, 195(5), 596-606, DOI: 10.1164/rccm.201606-1275CI.

[4] Yu, Y., Yao, S., Dong, H., Wang, L., Wang, C., Ji, X., Ji, M., Yao, X. and Zhang, Z., 2019. Association between short-term exposure to particulate matter air pollution and cause-specific mortality in Changzhou, China. *Environment Research*, 170, 7-15.

[5] Pollution Control Department, 2019. *Thailand's Air Quality and Situation Report*. [online] Available at: http://air4thai.pcd.go.th/webV2/history/.

[6] Chirasophon, S. and Pochanart, P., 2020. The long-term characteristics of PM_{10} and $PM_{2.5}$ in Bangkok, Thailand. *Asian Journal of Atmospheric Environment*, 14(1), 73-83, DOI: 10.5572/ajae.2020.14.1.073.

[7] Peng-in, B., Sanitluea, P., Monjatturat, P., Boonkerd, P. and Phosri, A., 2022. Estimating ground-level over Bangkok Metropolitan Region in Thailand using aerosol optical depth retrieved by MODIS. *Air Quality, Atmosphere and Health*, 5, 2091-2102, DOI: 10.1007/s11869-022-01238-4.

[8] Bahl, S. and Tuteja, R.K., 1991. Ratio and product-type exponential estimator. *Information and Optimization Sciences*, 12(1), 159-163.

[9] Singh, H.P. and Pal, S.K., 2015. A new chain ratio-ratio-type exponential estimator using auxiliary information in sample surveys. *International Journal of Mathematics and Its Applications*, 3, 37-46.

[10] Jaroengeratikun, U. and Lawson, N., 2019. A combined family of ratio estimators for population mean using an auxiliary variable in sample random sampling. *Journal of Mathematical and Fundamental Sciences*, 51(1), 1-12.

[11] Rubin, R.B., 1976. Inference and missing data. *Biometrika*. 63(3), 581-592.

[12] Kalton, G., Kasprzyk, D. and Santos, R., 1981. Issues of nonresponse and imputation in the survey of income and program participation. In: D. Krewski, R. Platek and J.N.K. Rao, eds. *Current Topics in Survey Sampling*. New York: Academic Press, pp. 455-480.

[13] Rao, J.N.K. and Sitter, R.R., 1995. Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82(2), 453-60.

[14] Lee, H., Rancourt, E. and Sarndal, C.E., 1994. Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

[15] Singh, S. and Deo, B., 2003. Imputation by power transformation. *Statistical Papers*, 44, 555-579.

[16] Ahmed, M.S., Al-Titi, O., Al-Rawi, Z. and Abu-Dayyeh, W., 2006. Estimation of a population mean using different imputation methods. *Statistics in Transition*, 7(6), 1247-1264.

[17] Kadilar, C. and Cingi, H., 2008. Estimators for the population mean in the case missing data. *Communications in Statistics-Theory and Methods,* 37, 2226-2236.

[18] Singh, S. and Horn, S., 2000. Compromised imputation in survey sampling. *Metrika*, 51, 267-276.

[19] Singh, A.K., Singh P. and Singh, V., 2014. Exponential-type compromised imputation in survey sampling. *Journal of Statistics Applications and Probability*, 3(2), 211-217.

[20] Rao, P.S.R.S, 1988. Ratio and regression estimators. In: P.R. Krishnaiah and C.R. Rao, eds. *Handbook of Statistics. Vol. 6*. Amsterdam: Elsevier Science, pp, 449-468.

[21] Audu, A. and Singh, R.V.K., 2020. Exponential-type regression compromised imputation class of estimators. *Journal of Statistics and Management Systems*, 30, DOI: 10.1080/09720510.2020.1814501.

[22] Fu, H., Zhang, Y., Liao, C., Mao, L., Wang, Z., and Hong, N., 2020. Investigating $PM_{2.5}$ responses to other air pollutants and meteorological factors across multiple temporal scales. *Scientific Reports*, 10, DOI: 10.1038/s41598-020-72722-z.

[23] Chinwetkitvanich, S., Ngamsritrakul, T. and Panyametheekul, S., 2021. New normal role in PM2.5 reduction in Bangkok. *International Journal of Environmental Science and Development*, 12(4), 100-106.

[24] Sun, Y.L., Wang, Z.F., Fu, P.Q., Yang, T., Jiang, Q., Dong, H.B., Li, J. and Jia, J.J., 2013. Aerosol composition, sources and processes during wintertime in Beijing, China. *Atmospheric Chemistry and Physics*, 2013, 13, 4577-4592.

[25] Zhao, P.S., Dong, F., He, D., Zhao, X.J., Zhang, X.L., Zhang, W.Z., Yao, Q. and Liu, H.Y., 2013. Characteristics of concentrations and chemical compositions for $PM_{2.5}$ in the region of Beijing, Tianjin, and Hebei, China. *Atmospheric Chemistry and Physics*, 13, 4631-4644.

[26] Liu, Y., Zhou, Y. and Lu, J., 2020. Exploring the relationship between air pollution and meteorological conditions in china under environmental governance. *Scientific Reports*, 10, DOI: 10.1038/s41598-020-71338-7.

[27] Yang, H., Peng, Q., Zhou, J., Song, G. and Gong, X., 2020. The unidirectional causality influence of factors on $PM_{2.5}$ in Shenyang City of China. *Scientific Reports*, 10, DOI: 10.1038/s41598-020-65391-5.