Songklanakarin J. Sci. Technol.
45 (4), 456–463, Jul. – Aug. 2023

http://www.sjst.psu.ac.th

Original Article

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 Northern Thailand

Chugiat Ponkaew[1] and Nuanpan Lawson[2*]

[1] *Department of Mathematics, Faculty of Science and Technology,*
Phetchabun Rajabhat University, Mueang, Phetchabun, 67000 Thailand

[2] *Department of Applied Statistics, Faculty of Applied Science,*
King Mongkut's University of Technology North Bangkok, Bang Sue, Bangkok, 10800 Thailand

Received: 20 December 2022; Revised: 1 March 2023; Accepted: 10 May 2023

Abstract

The issue of fine particulate matter in Thailand, especially in Northern Thailand, is an urgent problem that needs to be solved because of potential harm to human health. Prior estimates of fine particulate matter help planning how to reduce it. The daily fine particulate matter data reports usually contain some missing values. An improved ratio estimator has been suggested for population total under unequal probability sampling without replacement. The improved estimator is studied under a reverse framework when the nonresponse mechanism is not uniform, called missing at random nonresponse, which is more convenient to apply in practice. The variance and its associated estimator are investigated in theory. Simulation studies are used to assess the suggested estimator's performance. The new estimator is also applied to estimate fine particulate matter in Northern Thailand. The results show that the suggested estimator under the missing at random nonresponse mechanism performs well, as opposed to the existing estimator under missing completely at random assumption. The estimated fine particulate matter levels in Northern Thailand from the proposed estimator give a lesser variance than the existing one.

Keywords: fine particulate matter, ratio estimator, nonresponse mechanism, population total, missing at random

1. Introduction

The world ranking by dust levels in 2019 and 2020 showed that Thailand had the highest dust values, especially in Chiang Mai, which is one of the northern provinces in Thailand. Fine particulate matter 2.5 (PM 2.5) is a type of dust in the air, with particle diameter of 2.5 microns, that is dangerous to human health at levels exceeding the standard limit. Estimating the level of PM 2.5 can help plan policies to prevent pollution and to try to prevent increases in the levels of PM 2.5. Chodjuntug and Lawson (2022a) suggested imputing missing values of PM 2.5 in data from Bangkok, Thailand, and estimated the average PM 2.5 using two constants that gave the least mean square error for the suggested estimator. Later Chodjuntug and Lawson (2022b) studied the estimated values of PM 2.5 in Kanchana Phisek road in Bangkok, Thailand, and developed a new imputation method following Chodjuntug and Lawson (2022a). Their method used the benefit of the response rate and a minimum constant based on the regression estimator. Thongsak and Lawson (2022) proposed to estimate the air pollution data in Nan, Thailand, with two new families of estimators for population mean using the transformation of an auxiliary variable under simple random sampling without replacement. Their suggested transformed estimators assist in reducing bias and mean square error compared to the non-transformed estimators.

The Pollution Control Department of Thailand is an organization that collects pollution data, including those on PM 2.5 as hourly averages. Unfortunately, some of the PM 2.5

*Corresponding author
Email address: nuanpan.n@sci.kmutnb.ac.th

data are missing. Nonresponse or missing data is a serious problem that often arises in sample survey data, and usually occurs in many areas of study including science and engineering. Missing data occur for example when some respondents choose to not answer or accidently skip some questions. Estimating some parameters based on sample survey data may be affected by the nonresponse issue, so it needs to be taken into consideration before conducting further analysis. There are many types of nonresponse mechanisms. For example, missing completely at random (MCAR) or the uniform nonresponse mechanism is the strongest assumption, where missingness does not depend on the value that should be observed nor on the other observed values, and this is unlikely to occur in practice. A more flexible nonresponse mechanism is missing at random (MAR), where the missingness depends on the data that are observed but not on the missing values.

A powerful estimator is the ratio estimator for population total or population mean, which is almost an unbiased estimator but is more competent than the usual sample mean estimator. It was instigated by Cochran (1977) in the case of a highly positive correlation between a study variable and an auxiliary variable. Later, many works have explored the estimator by Cochran (1977) under simple random sampling without replacement (SRSWOR) (see e.g. Lawson, 2019, 2021; Jaroengeratikun & Lawson, 2019; Soponviwatkul & Lawson, 2017). Some researchers have suggested ratio estimators for population mean or population total using unequal probability sampling without replacement. For example, Bacanli and Kadilar (2008) proposed new ratio estimators by adjusting some ratio estimators under simple random sampling to unequal probability sampling without replacement (UPWOR) using the Horvitz and Thompson (1952) estimator, and the new estimators accomplished more than the existing estimators in consideration of a lower mean square error (see e.g. Ponkaew & Lawson, 2023).

However, when there is nonresponse, the ratio estimator does not work, so Sarndal and Lundstrom (2005) suggested a new population total estimator in a two-phase framework under UPWOR. Lawson (2017) proposed a new population total estimator under UPWOR under the reverse framework. The Lawson estimator considered the missing completely at random case when the sampling fraction can be omitted. Ponkaew and Lawson (2018) brought forward a new ratio estimator for population total following the estimators proposed by Bacanli and Kadilar (2008) and Sarndal and Lundstrom (2005) under UPWOR and the reverse framework. They investigated a case where nonresponse exists in the variable of interest and the nonresponse mechanism is uniform nonresponse, also called missing completely at random, which is quite a restrictive assumption and unlikely to happen in practice.

In this paper, an improved ratio estimator for population total has been developed under UPWOR. We sought to improve the estimator proposed by Ponkaew and Lawson (2018) when nonresponse occurs in the study variable but extending it to a more flexible case when nonresponse mechanism is missing at random under the reverse framework. Simulation studies and an application to fine particulate matter in northern Thailand are used to assess the proficiency of the improved estimator compared to Ponkaew and Lawson’s estimator.

2. Materials and Methods

2.1 Basic setup

Consider $U = \{1, 2, \ldots, i, \ldots, N\}$ that is a finite population of size N, with y_i being the value of the study variable y for the ith population unit. Let $Y=\sum_{i\in U} y_i$ be the population total of y and $X=\sum_{i\in U} x_i$ be the population total of an auxiliary variable x that is correlated with y and assumed to be known. A sample s of size n is chosen under UPWOR. Let π_i and π_{ij} be the first and second order of inclusion probabilities defined by $P(i\in s) = \pi_i$ and $P(i\wedge i\in s) = \pi_{ij}$ for the joint inclusion probability where units i and j are included in the sample s and let I_i be a random variable where I_i = 1if $i \in s$ otherwise $I_i = 0$. Let R_i represent the response indicator variable of y_i and is $R_i = 1$ if y_i is observed otherwise $R_i = 0$. Let $\boldsymbol{R} = (R_1\, R_2 \ldots R_N)'$ be the vector of response indicators. Let p_i represent the response probability that is $p_i = P(R_i = 1)$. In this study, we assume that the nonresponse mechanism is MAR.

2.2 Existing estimator under MCAR nonresponse mechanism

Ponkaew and Lawson (2018) proposed a ratio estimator for population total under UPWOR when the nonresponse mechanism is MCAR. The Ponkaew and Lawson population total estimator is

$$\hat{Y}_{PL}=\frac{\sum_{i\in s}\dfrac{R_i y_i}{\pi_i p}}{\sum_{i\in s}\dfrac{x_i}{\pi_i}}X,=\frac{\hat{Y}_r}{\hat{X}_{HT}}X=X\hat{R}_r \tag{1}$$

where $\hat{Y}_r=\sum_{i\in s}\dfrac{R_i y_i}{\pi_i p}$, $\hat{X}_{HT}=\sum_{i\in s}\dfrac{x_i}{\pi_i}$, $X=\sum_{i\in U} x_i$, $\hat{R}_r=\hat{Y}_r(\hat{X}_{HT})^{-1}$ and the response probability p is constant under MCAR.

The variance of $\hat{Y}_{PL}$ is

$$V(\hat{Y}_{PL}) \cong \sum_{i\in U} D_i A_i^2 + \sum_{i\in U} \sum_{i\backslash\{j\}\in U} D_{ij} A_i A_j , \tag{2}$$

where $A_i = y_i - Rx_i$, $R = YX^{-1}$, $D_i = \dfrac{1-\pi_i}{\pi_i}$, $D_{ij} = \dfrac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}$.

2.3 Proposed estimator

We sought to improve a ratio estimator proposed by Ponkaew and Lawson (2018) to be more flexible to when the nonresponse mechanism is MAR. We also propose the variance and its corresponding estimator in a general form when the sampling fraction is not negligible and when it is small and therefore it can be negligible.

An improved ratio estimator for population total under MAR mechanism is

$$\hat{Y}_R = \frac{\sum_{i\in s} \frac{R_i y_i}{\pi_i p_i}}{\sum_{i\in s} \frac{x_i}{\pi_i}} X = \frac{\hat{Y}_r'}{\hat{X}_{HT}} X = X\hat{R}_r' , \tag{3}$$

where $\hat{Y}_r' = \sum_{i\in s} \dfrac{R_i y_i}{\pi_i p_i}$, and $\hat{R}_r' = \hat{Y}_r'(\hat{X}_{HT})^{-1}$.

Under the MAR nonresponse mechanism if p_i is unknown it can be estimated by the probit or logistic regression models. The variance and associated estimators of $\hat{Y}_R$ are discussed in Theorem 1.

Theorem 1. Under reverse framework and nonresponse mechanism is MAR.

(1) The variance of $\hat{Y}_R$ is given by,

$$V\left(\hat{Y}_R\right) \cong \sum_{i\in U} D_i A_i'^2 + \sum_{i\in U} \sum_{i\backslash\{j\}\in U} D_{ij} A_i' A_j' + \sum_{i\in U} E_i' y_i^2 ,$$

where $A_i' = A_i = y_i - Rx_i$, $R = YX^{-1}$and $E_i' = \dfrac{1-p_i}{p_i}$.

(2) The estimator of $V\left(\hat{Y}_R\right)$ is

$$\hat{V}\left(\hat{Y}_R\right) \cong \sum_{i\in s} \hat{D}_i \hat{A}_i'^2 + \sum_{i\in s} \sum_{i\backslash\{j\}\in s} \hat{D}_{ij} \hat{A}_i' \hat{A}_j' + \sum_{i\in s} \hat{E}_i' y_i^2 ,$$

where $\hat{A}_i' = y_i - \hat{R}_r' x_i$, $\hat{R}_r' = \sum_{i\in s} \dfrac{R_i y_i}{\pi_i p_i} \left(\sum_{i\in s} \dfrac{x_i}{\pi_i} \right)^{-1}$, $\hat{E}_i' = \dfrac{R_i E_i}{\pi_i p_i}$, $\hat{D}_i = \dfrac{R_i D_i}{p_i \pi_i}$ and $\hat{D}_{ij} = \dfrac{R_i R_j D_{ij}}{p_i p_j \pi_{ij}}$.

Proof. Let $\hat{Y}_R$ be defined in (3). Therefore, variance of $\hat{Y}_R$ is

$$V(\hat{Y}_R) = V(\hat{R}_r' X) = X^2 V(\hat{R}_r') . \tag{4}$$

The estimator of $V(\hat{Y}_R)$ can be acquired by,

$$\hat{V}(\hat{Y}_R) = X^2 \hat{V}(\hat{R}_r') . \tag{5}$$

Since $\hat{R}_r'$ is nonlinear we have

$$V(\hat{R}_r') = E_R V_S\left(\hat{R}_r' \,\middle|\, \boldsymbol{R}\right) + V_R E_S\left(\hat{R}_r' \,\middle|\, \boldsymbol{R}\right) = V_1 + V_2 , \tag{6}$$

where $V_1 = E_R V_S\left(\hat{R}_r' \,\middle|\, \boldsymbol{R}\right)$, $V_2 = V_R E_S\left(\hat{R}_r' \,\middle|\, \boldsymbol{R}\right)$.

Step 1: Investigate the formula for $V_1 = E_R V_S\left(\hat{R}_r' \,\middle|\, \boldsymbol{R}\right)$.

The Taylor linearization approach has been applied to find a linear estimator of $\hat{R}_r'$

$$\hat{R}_r' \equiv \text{Constant}+\frac{1}{X}\sum_{i\in s}\frac{\tilde{A}_i'}{\pi_i}, \qquad (7)$$

where $\tilde{A}_i' = \left(\dfrac{R_i y_i}{p_i} - \tilde{R}_r'\ x_i\right)$.

Then $V_1 = E_R V_S\left(\hat{R}_r'\ )\middle|\boldsymbol{R}\right)$ can be approximated by,

$$V_1' \cong E_R V_S\left(\hat{R}_r' \middle|\boldsymbol{R}\right) = E_R V_S\left(\text{Constant}+\frac{1}{X}\sum_{i\in s}\frac{\tilde{A}_i'}{\pi_i}\middle|\boldsymbol{R}\right)$$

$$= \frac{1}{X^2}E_R\left(\sum_{i\in U}D_i\tilde{A}_i'^2 + \sum_{i\in U}\sum_{i\backslash\{j\}\in U}D_{ij}\tilde{A}_i'\tilde{A}_j'\middle|\boldsymbol{R}\right)$$

$$= \frac{1}{X^2}\left(\sum_{i\in U}D_iA_i'^2 + \sum_{i\in U}\sum_{i\backslash\{j\}\in U}D_{ij}A_i'A_j'\right),$$

where $A_i' = E_R V_S\left(\tilde{A}_i'\middle|\boldsymbol{R}\right) = y_i - Rx_i$ and $R = YX^{-1}$.

Therefore.

$$V_1' \cong \frac{1}{X^2}\left(\sum_{i\in U}D_iA_i'^2 + \sum_{i\in U}\sum_{i\backslash\{j\}\in U}D_{ij}A_i'A_j'\right). \qquad (8)$$

Step 2: Investigate formula for $V_2 = V_R E_S\left(\hat{R}_r'^{(1)}\middle|\boldsymbol{R}\right)$.

The formula for $V_2 = V_R E_S\left(\hat{R}_r'^{(1)}\middle|\boldsymbol{R}\right)$ can be approximated by,

$$V_2' \cong V_R E_R\left(\hat{R}_r' \middle|\boldsymbol{R}\right) = V_R E_R\left(\frac{\sum_{i\in s}\dfrac{R_i y_i}{\pi_i p_i}}{\sum_{i\in s}\dfrac{x_i}{\pi_i}}\middle|\boldsymbol{R}\right)$$

$$= V_R\left(\frac{\sum_{i\in U}\dfrac{R_i y_i}{p_i}}{X}\middle|\boldsymbol{R}\right) = \frac{1}{X^2}\sum_{i\in U}\frac{(1-p_i)y_i^2}{p_i} = \frac{1}{X^2}\sum_{i\in U}E_i'y_i^2,$$

where $E_i' = \dfrac{(1-p_i)}{p_i}$.

Then,

$$V_2' \cong \frac{1}{X^2}\sum_{i\in U}E_i'y_i^2. \qquad (9)$$

Step 3: Approximate values of $V(\hat{R}_r')$ and its estimators.

The value of $V(\hat{R}_r')$ is

$$V(\hat{R}_r') \cong V_1' + V_2' = \frac{1}{X^2}\left(\sum_{i\in U}D_iA_i'^2 + \sum_{i\in U}\sum_{i\backslash\{j\}\in U}D_{ij}A_i'A_j' + \sum_{i\in U}E_i'y_i^2\right). \qquad (10)$$

The estimator of $V(\hat{R}_r')$ is

$$\hat{V}(\hat{R}_r') = \frac{1}{X^2}\left(\sum_{i\in s}\hat{D}_i\hat{A}_i'^2 + \sum_{i\in s}\sum_{i\backslash\{j\}\in s}\hat{D}_{ij}\hat{A}_i'\ \hat{A}_j' + \sum_{i\in s}\hat{E}_i'y_i^2\right). \qquad (11)$$

Substitute (10) into (4), then the variance of $\hat{Y}_R$ is given by

$$V(\hat{Y}_R) \cong \sum_{i\in U} D_i A_i'^2 + \sum_{i\in U}\sum_{i\backslash\{j\}\in U} D_{ij} A_i' A_j' + \sum_{i\in U} E_i' y_i^2. \tag{12}$$

The estimator of $V(\hat{Y}_R)$ can be obtained by substituting (11) in (5) to get

$$\hat{V}(\hat{Y}_R) \cong \sum_{i\in s} \hat{D}_i \hat{A}_i'^2 + \sum_{i\in s}\sum_{i\backslash\{j\}\in s} \hat{D}_{ij} \hat{A}_i' \ \hat{A}_j' + \sum_{i\in s} \hat{E}_i' y_i^2 \ . \tag{13}$$

In Theorem 1 the variance and its associated estimator are discussed. Next, in Lemma 2 we consider a special case of Theorem 1, in which the sampling fraction can be omitted.

Lemma 2. Under reverse framework when the nonresponse mechanism is MAR and sampling fraction can be disregarded:

(1) The variance of $\hat{Y}_R$ is

$$V\left(\hat{Y}_R\right) \cong \sum_{i\in U} D_i A_i'^2 + \sum_{i\in U}\sum_{i\backslash\{j\}\in U} D_{ij} A_i' A_j' ,$$

where $A_i' = A_i = y_i - Rx_i$, $R = YX^{-1}$.

(2) The estimator of $V\left(\hat{Y}_R\right)$ is

$$\hat{V}\left(\hat{Y}_R\right) \cong \sum_{i\in s} \hat{D}_i \hat{A}_i'^2 + \sum_{i\in s}\sum_{i\backslash\{j\}\in s} \hat{D}_{ij} \hat{A}_i' \ \hat{A}_j' \ ,$$

where $\hat{A}_i' = y_i - \hat{R}_r' x_i$, $\hat{R}_r' = \sum_{i\in s} \frac{R_i y_i}{\pi_i p_i} \left(\sum_{i\in s} \frac{x_i}{\pi_i} \right)^{-1}$.

Next in Lemma 3, we also study the variance and its estimator from Ponkaew and Lawson (2018) in a general situation because they only considered when the sampling fraction is negligible.

Lemma 3. Under reverse framework and the MCAR nonresponse mechanism:

(1) The variance of $\hat{Y}_{PL}$ is given by,

$$V(\hat{Y}_{PL}) \cong \sum_{i\in U} D_i A_i^2 + \sum_{i\in U}\sum_{i\backslash\{j\}\in U} D_{ij} A_i A_j + \sum_{i\in U} E_i y_i^2,$$

where $E_i = \frac{1-p}{p}$ and A is as defined in (2).

(2) The estimator of $V\left(\hat{Y}_{PL}\right)$ is given by,

$$\hat{V}\left(\hat{Y}_{PL}\right) \cong \sum_{i\in s} \hat{D}_i \hat{A}_i^2 + \sum_{i\in s}\sum_{i\backslash\{j\}\in s} \hat{D}_{ij} \hat{A}_i \ \hat{A}_j + \sum_{i\in s} \hat{E}_i y_i^2,$$

where $\hat{A}_i = y_i - \hat{R}_r x_i$, $\hat{R}_r = \sum_{i\in s} \frac{R_i y_i}{\pi_i p} \left(\sum_{i\in s} \frac{x_i}{\pi_i} \right)^{-1}$, $\hat{E}_i = \frac{R_i E_i}{\pi_i p}$, $\hat{D}_i = \frac{R_i D_i}{p\pi_i}$ and $\hat{D}_{ij} = \frac{R_i R_j D_{ij}}{p^2 \pi_{ij}}$.

3. Results and Discussion

3.1 Simulation studies

Simulation studies have been used to examine the proficiency of the proposed estimator and its variance estimator. We compared the proposed estimator with that in Ponkaew and Lawson (2018). We generated the study variable y_i by applying the linear model of Lawson and Siripanich (2020) as follows. $y_i = \beta_0 + \beta_1 x_i + \beta_2 k_i + \beta_3 w_i + \varepsilon_i$ with population size N = 10,000 where $x_i \sim N(20000, 1000)$, $k_i \sim N(50, 189)$, $w_i \sim N(44, 625)$, $\varepsilon_i \sim N(0, 1)$, $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)' = (12242, 0.7, -82, 72)'$ and I = 1, 2, . . ., N. The samples of sizes n = 100, 300, 500, 1000, 2000 and 4000 were selected using UPWOR.

The variance formula under UPWOR requires joint inclusion probability in the procedure, but it is usually not available in practice. Midzuno's (1952) scheme for the UPWOR can be used to derive the joint inclusion probability. Under this scheme, the first and second order inclusion probabilities are given by

$$\pi_i = \frac{k_i}{K}\left(\frac{N-n}{N-1}\right) + \frac{n-1}{N-1}, \tag{14}$$

$$\pi_{ij} = \left(\frac{k_i+k_j}{K}\right)\left(\frac{N-n}{N-1}\right)\left(\frac{n-1}{N-2}\right) + \left(\frac{n-1}{N-1}\right)\left(\frac{n-2}{N-2}\right). \tag{15}$$

We consider 85% response rate in the simulation study and repeated it 10,000 times (M=10,000) using the R program (R Core Team, 2021). We consider the case where the true response probabilities are unknown and the logistic regression model is used to estimate response probabilities p_i by $p_i = e^{b_o+b_1w_i} \big/ (1+e^{b_o+b_1w_i})$ where b_0 and b_1 are coefficients from the fitted logistic regression model. The relative bias (RB) and the relative root mean square error (RRMSE) are used as the criteria to compare the competency of the proposed estimator with Ponkaew and Lawson's (2018) estimator. Let, $\hat{Y}_{Ratio}$ be the ratio estimators and $\hat{V}\left(\hat{Y}_{Ratio}\right)$ be their variance estimators. The RB and the RRMSE of the ratio estimator and associated variance estimator are given as follows.

$$\text{RB}(\hat{Y}_{ratio}) = \frac{E(\hat{Y}_{Ratio,m}) - Y}{Y},$$

$$\text{RRMSE}(\hat{Y}_{ratio}) = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{Y}_{Ratio,m} - Y)^2}}{Y},$$

$$\text{RB}(\hat{V}_m\left(\hat{Y}_{ratio}\right)) = \frac{E\ (\hat{V}_m\left(\hat{Y}_{ratio}\right)) - V(\hat{Y}_{ratio})}{V(\hat{Y}_{ratio})},$$

$$\text{RRMSE}(\hat{V}_m\left(\hat{Y}_{ratio}\right)) = \frac{\sqrt{\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{V}_m\left(\hat{Y}_{ratio}\right) - V(\hat{Y}_{ratio})^2\right)}}{V(\hat{Y}_{ratio})}.$$

The results are shown in Tables 1 to 4.

Table 1. The relative biases of the proposed and existing estimators

n	Estimator	
	$\hat{Y}_{PL}$	$\hat{Y}_R$
100	0.0173	0.0012
300	0.0161	0.0011
500	0.0158	0.0006
1000	0.0152	0.0004
2000	0.0109	0.0003
4000	0.0038	0.0001

Table 2. The relative root mean square error of the proposed and existing estimators

n	Estimator	
	$\hat{Y}_{PL}$	$\hat{Y}_R$
100	0.0206	0.0154
300	0.0165	0.0095
500	0.0163	0.0079
1000	0.0150	0.0046
2000	0.0141	0.0027
4000	0.0039	0.0009

Table 3. The relative biases in variance of the proposed and existing estimators

n	Estimator	
	$\hat{V}\left(\hat{Y}_{PL}\right)$	$\hat{V}\left(\hat{Y}_R\right)$
100	0.0361	0.0351
300	0.0304	0.0242
500	0.0176	0.0159
1000	0.0143	0.0103
2000	0.0060	0.0024
4000	0.0024	0.0004

Table 4. The relative root mean square error of the proposed and existing estimators

n	Estimator	
	$\hat{V}\left(\hat{Y}_{PL}\right)$	$\hat{V}\left(\hat{Y}_R\right)$
100	0.1557	0.1290
300	0.1296	0.0915
500	0.1263	0.0796
1000	0.0845	0.0602
2000	0.0566	0.0458
4000	0.0411	0.0216

Table 1 shows the RB of the new population total estimator in contrast with the estimator proposed by Ponkaew and Lawson (2018). The results indicate that the proposed estimator $\hat{Y}_R$ gave a lot smaller relative bias when compared to $\hat{Y}_{PL}$ for all levels of sample sizes. The bias of the proposed estimator becomes close to zero when the sample size increases. Similar results are shown in Table 2 when we consider the relative root mean square error of the population total estimator. The proposed estimator gave a lot smaller RRMSE, especially with large sample sizes.

Tables 3 and 4 show the RB and RRMSE of the variance of the proposed estimator against the variance estimator proposed by Ponkaew and Lawson (2018), respectively. Similar to what we found in Tables 2 and 3, the proposed variance estimator performs better than Ponkaew and Lawson's (2018) in terms of a smaller RB and RRMSE.

3.2 An application to PM 2.5 in Northern Thailand

An application to fine particulate matter in northern Thailand was conducted in this study. The data are from the Air Quality and Noise Management Bureau, the Pollution Control Department of Thailand (2022), from October to November 2022. Midzuno's (1952) scheme was applied to select a sample of 13 stations out of the 23 stations. The PM 2.5 (micrograms per cubic meter) data collected on hourly averages on 28 November 2002 were used as a study variable *y*. The average level of PM 2.5 in October 2002 and the air quality index average in October 2002 were considered auxiliary variables *x* and *w*, respectively. The variable *x* was used to construct the proposed ratio estimator, while the variable *w* was used to estimate the response probability by using the logistic regression model under the MAR mechanism. The size variable *k* was the maximum value of PM 2.5 in October 2002. The nonresponse rate was 15.40% in this study. The estimated total and variances of PM2.5 based on the proposed estimator under MAR and the existing estimator under MCAR are shown in Table 5.

Table 5 shows the estimated total values of PM 2.5 from the proposed estimator $\hat{Y}_R$ compared to the existing estimator $\hat{Y}_{PL}$. We see that the estimated total and variance of the estimator of PM 2.5 from $\hat{Y}_R$ under the MAR nonresponse mechanism gave smaller estimates and estimated variance than the existing $\hat{Y}_{PL}$ under the MCAR nonresponse mechanism, which corresponds to the simulation results.

Table 5. The estimated total values and variances of PM2.5

Nonresponse mechanism	Estimate total of PM 2.5	Estimate variance of total estimator of PM2.5
$\hat{Y}_{PL}$	437.7505	1015.343
$\hat{Y}_R$	399.5789	1007.942

4. Conclusions

An improved ratio estimator for population total and its variance estimator have been suggested under UPWOR when nonresponse occurs with the study variable. The proposed estimators were studied under the MAR mechanism, which is more plausible in real life. The improved estimator's variance estimator was also suggested, both when the sampling fraction can be disregarded and when it cannot be disregarded. The improved estimator worked well in the simulation studies and an application to fine particulate matter in Northern Thailand, when compared to the existing estimator in terms of minimum RB and RRMSE. The bias of the proposed estimator approaches zero when the sample size rises. In future research, we can extend the improved ratio estimator to cases with nonresponse in both study and auxiliary variables, and to more complex survey designs. The improved estimator can be useful in applications to real data from many areas of study, not restricted to only air pollution data.

Acknowledgements

This research was funded by the National Science, Research and Innovation Fund (NSRF), and King Mongkut's University of Technology North Bangkok Contract no. KMUTNB-FF-66-56. We would like to thank the anonymous referees for their comments.

References

Bacanli, S., & Kadilar, C. (2008). Ratio estimators with unequal probability designs, *Pakistan Journal of Statistics, 24*(3), 167-17.

Chodjuntug, K., & Lawson, N. (2022a). Imputation for estimating the population mean in the presence of nonresponse, with application to fine particle density in Bangkok. *Mathematical Population Studies*. doi:10.1080/08898480.2021.1997466.

Chodjuntug K., Lawson N. (2022b). The chain regression exponential type imputation method for mean estimation in the presence of missing data. *Songklanakarin Journal of Science and Technology, 44*(4), 1109-1118.

Cochran, W.G. (1977). *Sampling techniques*. New York, NY: John Wiley and Sons.

Horvitz, D. F., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, *47*(260), 663-685.

Jaroengeratikun, U., & Lawson, N. (2019). A combined family of ratio estimators for population mean using an auxiliary variable in simple random sampling, *Journal of Mathematical and Fundamental Sciences*, *51*(1), 1–12. doi:10.5614/j.math.fund.sci. 2019.51.1.1.

Kadilar, C., & Cingi, H. (2004). Ratio estimators in simple random sampling, *Applied Mathematicsand Computation*, *151*(3), 893–902. doi:10.1016/S0096-3003(03)00803-8.

Lawson, N. (2017). Variance estimation in the presence of nonresponse under probability proportional to size sampling. *Proceedings of the International Conference on Computational Mathematics, Computational Geometry and Statistics,* 116-119.

Lawson, N. (2019). Ratio estimators of population means using quartile function of auxiliary variable using double sampling, *Songklanakarin Journal of Science and Technology, 41*(1), 117– 122. doi:10. 14456/sjst-psu.2019.14.

Lawson, N. (2021). An alternative family of combined estimators for estimating population mean in finite populations. *Lobachevskii Journal of Mathematics, 42*(13), 3150–3157. doi:10.1134/S1995080222010 115.

Lawson, N., & Siripanich, P. (2020). A new generalized regression estimator and variance estimation for unequal probability sampling without replacement for missing data. *Communications in Statistics - Theory and Methods, 51*(18), 6296-6318. doi:10.10 80/03610926.2020.1860224

Midzuno, H. (1952). On sampling system with probability proportional to sum of sizes, *Annals of the Institute of Statistical Mathematics*, 99-107.

Pollution Control Department. (2022). Thailand's air quality and situation reports. Bangkok, Thailand. Retrieved from http://air4thai.pcd.go.th/webV2/history/.

Ponkaew, C., & Lawson, N. (2018). A new ratio estimator for population total in the presence of nonresponse under unequal probability sampling without replacement. *Thai Journal of Mathematics: Special Issue (ACFPTO2018) on: Advances in Fixed Point Theory towards Real World Optimization Problem*, 417-429.

Ponkaew, C., Lawson N. (2023). New generalized regression estimators using a ratio method and its variance estimation for unequal probability sampling without replacement in the presence of nonresponse. *Current Applied Science and Technology, 23*(2), 1-7 doi:10.55003/cast.2022.02.23.007.

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Sarndal, C. E., & Lundstrom, S. (2005). Estimation in surveys with nonresponse. New York, NY: John Wiley and Sons.

Singh, H. P., & Tailor, R. (2003). Use of known correlation coefficient in estimating the finite population mean. *Statistics in Transition*, 555-560.

Singh, H. P., & Upadhyaya, L. N. (1986). A dual to modified ratio estimator using coefficient of variation of auxiliary variable. *Proceedings of the International Conference on National Academy of Sciences*, 336–340.

Sisodia, B. V., & Dwivedi, V. K. (1981). Modified ratio estimator using coefficient of variation of auxiliary variable. *Journal Indian Society of Agricultural Statistics*, *33*(2), 13–18.

Soponviwatkul, K., & Lawson, N. (2017). New ratio estimators for estimating population mean in simple random sampling using a coefficient of variation, correlation coefficient and a regression coefficient. *Gazi University Journal of Science*, *30*(4), 610–621.