Estimation of COVID-19 Incidence based on PM2.5 in Chiang Mai, Thailand using New Transformed Estimators in the Presence of Missing Observations

NATTHAPAT THONGSAK[1], NUANPAN LAWSON[2,*]
[1]State Audit Office of the Kingdom of Thailand,
Bangkok, 10400,
THAILAND

[2]Department of Applied Statistics, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok,
1518 Pracharat 1 Road, Wongsawang, Bangsue, Bangkok 10800,
THAILAND

**Corresponding Author*

Abstract: - The COVID-19 pandemic has damaged and taken human lives and still affects their daily routine and way of living. There is a connection between COVID-19 incidence and fine particulate matter which is a type of air pollution that causes issues to human health, especially in Chiang Mai, Thailand. Daily estimation of the incidence of COVID-19 can assist Thailand in planning to cope with the increasing number of COVID-19 incidents. Unfortunately, some COVID-19 data are missing, and as a result, it may yield inaccurate results for planning policies using missing data. A novel class of estimators engaging transformation to transform an auxiliary variable is suggested under simple random sampling without replacement, whilst assuming the population mean of an auxiliary variable is not available under uniform nonresponse. The new estimators are used to estimate the official cases of COVID-19 per day and the total patients diagnosed with pneumonia and are on high-flow oxygen therapy in Chiang Mai, Thailand using fine particulate matter with a diameter of 2.5 microns concentration as the auxiliary variable. The estimators that were brought forward performed well compared to the existing ones with a reduced bias and mean square error. The best-proposed estimator gave the estimated daily confirmed cases of around 101 cases and the total number of patients diagnosed with pneumonia and are on high-flow oxygen therapy around 16 cases. The highest efficiency is above 500 more percentage relative efficiency in contrast to the mean imputation method. The suggested estimators are more practical to use with real-world data as they do not require the population means associated with the auxiliary variable.

Key-Words: - Transformed Auxiliary Variable, Missing Data, Covid-19, Fine Particulate Matter, Auxiliary Variable, Imputation.

Received: October 14, 2024. Revised: November 19, 2024. Accepted: March 6, 2025. Published: May 6, 2025.

1 Introduction

The quality of life of people in Thailand and global people has been afflicted by the ongoing distress caused by the immense impact of the COVID-19 pandemic or SARS-CoV-2 which emerged in 2019 in Wuhan, China. Thailand is not only affected by the pandemic but also the air pollution issue mostly in Bangkok, the capital city of Thailand, and Chiang Mai province one of the northern provinces which is a place for tourists. COVID-19 has killed numerous human lives around the world and also damages the human body and lungs, [1], [2], [3]. Air pollution worsens the ongoing crisis already caused by COVID-19. Nonetheless, current industries' efforts in aiding sustainability have not sufficed and they still inevitably emit an abundance of toxins, aggravating

pollution brought on by fine particulate matter with a diameter of 2.5 microns (PM2.5). Its concentrations have exceeded guidelines persistently for many years, the issue expanding its severity and consequences to the health of the future population, the environment, and so on. It has become an urgent manner for policies to be implemented especially in Chiang Mai, Thailand.

The connection between COVID-19 patients and PM2.5 and air pollution data is elaborated on in an abundance of literature. As an example, the total patients is dependent on on a 7-day lagged effect of PM2.5 levels in Seoul, South Korea [4]. Moreover, the positive correlation between the COVID incidence each day and PM2.5 and humidity in all cities in China can be found using the multivariate Poisson regression model [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], 15].

Missing data usually occurs in COVID-19 incidences. In order to do so, COVID-19 incidences such as daily confirmed cases must be recorded, nevertheless, missing data contained within reports hinders the initiation of a plan to tackle the problem due to unsuitable estimates being derived. In this prevailing situation, it has come to the point where the nation will require action to put an end to this dilemma. Therefore, dealing with missing data is imperative. The imputation methods are suggested to replace the missing values based on plausible observation. For instance, the missing values can be derived from mean imputation by using the sample mean of the study variable, [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31].

Let the auxiliary and study variables be denoted as X and Y, respectively and r be the number of responding units out of the sampled n units chosen through simple random sampling without replacement (SRSWOR) from a population of size N. The point estimator for deriving population mean via the mean imputation technique is:

$$\hat{\bar{Y}}_{\mathrm{S}} = \bar{y}_r, \tag{1}$$

where $\bar{y}_r = \frac{1}{r}\sum_{i=1}^{r} y_i$.

The bias and variance of $\hat{\bar{Y}}_{\mathrm{S}}$ are:

$$Bias\left(\hat{\bar{Y}}_{\mathrm{S}}\right) = 0, \tag{2}$$

$$V\left(\hat{\bar{Y}}_{\mathrm{S}}\right) = \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2 C_y^2, \tag{3}$$

where $C_y = \frac{S_y}{\bar{Y}}$, $S_y^2 = \frac{\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2}{N-1}$.

The ratio method of imputation is a common means to choose when there is information based on an auxiliary variable X, that is correlated to the study variable Y. The point estimator for this technique is:

$$\hat{\bar{Y}}_{\mathrm{Rat}} = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r}, \tag{4}$$

where $\bar{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}$, and $\bar{x}_r = \frac{\sum_{i=1}^{r} x_i}{r}$.

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

$$Bias\left(\hat{\bar{Y}}_{\mathrm{Rat}}\right) = \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}\left(C_x^2 - \rho C_x C_y\right), \tag{5}$$

$$MSE\left(\hat{\bar{Y}}_{\mathrm{Rat}}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2 C_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}^2\left(C_y^2 + C_x^2 - 2\rho C_x C_y\right), \tag{6}$$

where $\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}$, $\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N}$, $C_x = \frac{S_x}{\bar{X}}$,

$S_x^2 = \frac{\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2}{N-1}$, and $\rho = \frac{S_{xy}}{S_x S_y}$.

Another transformation process to transform an auxiliary variable for determining the population mean estimator by the dual to ratio estimator under SRSWOR was invented for higher efficiency, [32]. A transformed auxiliary variable is:

$$x_i^* = \left(1 + \pi_{\mathrm{SRS}}\right)\bar{X} - \pi_{\mathrm{SRS}} x_i \;; i = 1, 2, 3, \ldots, N, \tag{7}$$

and the sample mean corresponding to x_i^* is:

$$\bar{x}_{\mathrm{SRS}}^* = \left(1 + \pi_{\mathrm{SRS}}\right)\bar{X} - \pi_{\mathrm{SRS}}\bar{x}, \tag{8}$$

where $\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}$ is the population mean of X and

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is a sample mean belonging to X,

$\pi_{\text{SRS}} = \frac{n}{N-n}$ and a sample of size n is chosen from a population of size N.

More works have been investigated based on [32]. For instance, a class of ratio estimators emerged for the population mean when some parameters of the auxiliary variable known under SRSWOR was proposed using the transformed auxiliary variable, [33]. It was shown that the novel transformed ones worked better than the estimators lacking transformation. The [33] estimator is:

$$\hat{\bar{Y}}_{\text{TL.SRS}} = \bar{y}\left(\frac{A\bar{x}^*_{\text{SRS}} + D}{A\bar{X} + D}\right), \quad (9)$$

where $(A \neq 0, D)$ are real numbers or functions in association with the auxiliary variable. The bias and MSE that are derivatives of the $\hat{\bar{Y}}_{\text{TL.SRS}}$ estimator are:

$$Bias\left(\hat{\bar{Y}}_{\text{TL.SRS}}\right) = -\left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}\pi_{\text{SRS}}\theta\rho C_x C_y, \quad (10)$$

$$MSE\left(\hat{\bar{Y}}_{\text{TL.SRS}}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2\left(C_y^2 + \pi_{\text{SRS}}^2\theta^2 C_x^2 - 2\pi_{\text{SRS}}\theta\rho C_x C_y\right), \quad (11)$$

where $\theta = \frac{A\bar{X}}{A\bar{X} + D}$, $C_y = \frac{S_y}{\bar{Y}}$, $S_y^2 = \frac{\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2}{N-1}$, $\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N}$, $C_x = \frac{S_x}{\bar{X}}$, $S_x^2 = \frac{\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2}{N-1}$, and $\rho = \frac{S_{xy}}{S_x S_y}$. The transformed estimators work better than the untransformed ones, [34], [35], [36].

The process of transformation can be implemented to conditions when missing data present themselves in the study variable. A general class of transformed regression type estimators when the study variable subjects itself to missing data and the population mean of the auxiliary variable could not be acquired, was put forward [37]. The [37] estimator is:

$$\hat{\bar{Y}}_{\text{TL}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{A\bar{x}^* + D}{A\bar{x}_n + D}\right), \quad (12)$$

where $\bar{x}^* = \frac{n\bar{x}_n - r\bar{x}_r}{n-r} = (1+\pi)\bar{x}_n - \pi\bar{x}_r$, $\pi = \frac{r}{n-r}$, $b = \frac{s_{xy}}{s_x^2}$ is the sample regression coefficient and $(A \neq 0, D)$ are real numbers or functions belonging to the auxiliary variable. The bias and MSE that are derivates of the $\hat{\bar{Y}}_{\text{TL}}$ estimator are:

$$Bias\left(\hat{\bar{Y}}_{\text{TL}}\right) = -\pi\theta\left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}\left(\pi\beta KC_x^2 + \rho C_x C_y\right), \quad (13)$$

$$MSE\left(\hat{\bar{Y}}_{TL}\right) = \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2 C_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}^2\left((\theta - \beta K)^2\pi^2 C_x^2 - 2(\theta - \beta K)\pi\rho C_x C_y\right) \quad (14)$$

The [37] estimator performed better than the existing ones suggested when implemented on prior COVID-19 patients total and air pollution data in Chiang Mai, Thailand. The members of the [37] estimator are represented in Table 1.

Table 1. Some components from the [37] estimator

Estimator	A	D
$\hat{\bar{Y}}_{\text{TL1}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{\bar{x}^*}{\bar{x}_n}\right)$	1	0
$\hat{\bar{Y}}_{\text{TL2}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{C_x\bar{x}^* + \beta_1}{C_x\bar{x}_n + \beta_1}\right)$	C_x	β_1
$\hat{\bar{Y}}_{\text{TL3}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{Q_1\bar{x}^* + Q_a}{Q_1\bar{x}_n + Q_a}\right)$	Q_1	Q_a
$\hat{\bar{Y}}_{\text{TL4}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{\beta_1\bar{x}^* + \beta_2}{\beta_1\bar{x}_n + \beta_2}\right)$	β_1	β_2
$\hat{\bar{Y}}_{\text{TL5}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{\beta_1\bar{x}^* + Q_1}{\beta_1\bar{x}_n + Q_1}\right)$	β_1	Q_1
$\hat{\bar{Y}}_{\text{TL6}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{\beta_2\bar{x}^* + Q_2}{\beta_2\bar{x}_n + Q_2}\right)$	β_2	Q_2
$\hat{\bar{Y}}_{\text{TL7}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{Q_d\bar{x}^* + C_x}{Q_d\bar{x}_n + C_x}\right)$	Q_d	C_x

where Q_1 and Q_3 are the first and third quartiles extracted from the auxiliary variable, respectively, $Q_r = Q_3 - Q_1$ is the inter-quartile range taken from the auxiliary variable, $Q_d = \frac{Q_3 - Q_1}{2}$ is the semi-quartile range based on the auxiliary variable, $Q_a = \frac{Q_3 + Q_1}{2}$ is the quartile mean based on the auxiliary variable, β_1 and β_2 are the coefficient of skewness and kurtosis of the auxiliary variable, respectively.

An updated classification of estimators focusing on the auxiliary variable that was transformed is enforced using SRSWOR and the uniform

nonresponse mechanism inspired by [9]. The population mean based on the auxiliary variable is assumed to be inaccessible which is highly prevalent in practice. The bias and MSE of the latest classification of estimators are approximated using the Taylor Series. The suggested estimators are applied to COVID-19 patient incidence and PM2.5 in Chiang Mai, Thailand.

2 Proposed Estimator

Assuming that a population mean coming from an auxiliary variable $(\bar{X})$ is inaccessible, an updated class of estimators utilizing the auxiliary variable's transformation is proposed inspired by [11]. The proposed estimator is defined as:

$$\hat{\bar{Y}}_{\mathrm{N}} = \bar{y}_r \left(\frac{A\bar{x}^* + D}{A\bar{x}_n + D} \right). \tag{15}$$

The MSE and bias of the proposed estimator are gained by the following notations.

$$\varepsilon_0 = \frac{\bar{y}_r - \bar{Y}}{\bar{Y}}, \bar{y}_r = (1+\varepsilon_0)\bar{Y},$$

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

$$, E(\varepsilon_0) = E(\varepsilon_1) = E(\varepsilon_2) = 0,$$

$$E(\varepsilon_0^2) = \left(\frac{1}{r} - \frac{1}{N}\right) C_y^2,\ E(\varepsilon_1^2) = \left(\frac{1}{r} - \frac{1}{N}\right) C_x^2,\ E(\varepsilon_2^2) = \left(\frac{1}{n} - \frac{1}{N}\right) C_x^2,$$

$$E(\varepsilon_0\varepsilon_1) = \left(\frac{1}{r} - \frac{1}{N}\right) \rho C_x C_y,\ E(\varepsilon_0\varepsilon_2) = \left(\frac{1}{n} - \frac{1}{N}\right) \rho C_x C_y,$$

$$E(\varepsilon_1\varepsilon_2) = \left(\frac{1}{n} - \frac{1}{N}\right) C_x^2.$$

Rewriting $\hat{\bar{Y}}_{\mathrm{N}}$ in terms of $e_i's, i = 0,1,2$, we have:

$$\hat{\bar{Y}}_{\mathrm{N}} = \bar{y}_r \left(\frac{A\bar{x}^* + D}{A\bar{x}_n + D} \right)$$

$$= (1+\varepsilon_0)\bar{Y} \left(\frac{A\left[(1+\pi)(1+\varepsilon_2)\bar{X} - \pi(1+\varepsilon_1)\bar{X}\right] + D}{A(1+\varepsilon_2)\bar{X} + D} \right)$$

$$= (1+\varepsilon_0)\bar{Y} \left(\frac{\left(A\bar{X} + D\right) + \left(\varepsilon_2 + \pi\varepsilon_2 - \pi\varepsilon_1\right) A\bar{X}}{\left(A\bar{X} + D\right) + \varepsilon_2 A\bar{X}} \right).$$

Let $\theta = \frac{A\bar{X}}{A\bar{X} + D}$, then

$$\hat{\bar{Y}}_{\mathrm{N}} = (1+\varepsilon_0)\bar{Y} \left(\frac{\frac{A\bar{X}}{\theta} + \left(\varepsilon_2 + \pi\varepsilon_2 - \pi\varepsilon_1\right) A\bar{X}}{\frac{A\bar{X}}{\theta} + \varepsilon_2 A\bar{X}} \right)$$

$$= \bar{Y}(1+\varepsilon_0)\left(1 + \left(\varepsilon_2 + \pi\varepsilon_2 - \pi\varepsilon_1\right)\theta\right)\left(1 + \varepsilon_2\theta\right)^{-1}.$$

Using the Taylor series approximation, we get:

$$\hat{\bar{Y}}_{\mathrm{N}} = \bar{Y}(1+\varepsilon_0)\left(1 + \left(\varepsilon_2 + \pi\varepsilon_2 - \pi\varepsilon_1\right)\theta\right)\left(1 - \varepsilon_2\theta + \varepsilon_2^2\theta^2 - \ldots\right)$$

$$\cong \bar{Y}\left(1 + \varepsilon_0 + \left(-\pi\theta\right)\varepsilon_1 + \left(\pi\theta\right)\varepsilon_2 + \left(-\pi\theta^2\right)\varepsilon_2^2 + \left(-\pi\theta\right)\varepsilon_0\varepsilon_1 + \left(\pi\theta\right)\varepsilon_0\varepsilon_2 + \left(\pi\theta^2\right)\varepsilon_1\varepsilon_2\right).$$

So the estimation error of $\hat{\bar{Y}}_{\mathrm{N}}$ is:

$$\hat{\bar{Y}}_{\mathrm{N}} - \bar{Y} = \bar{Y}\left(\varepsilon_0 + \left(-\pi\theta\right)\varepsilon_1 + \left(\pi\theta\right)\varepsilon_2 + \left(-\pi\theta^2\right)\varepsilon_2^2 + \left(-\pi\theta\right)\varepsilon_0\varepsilon_1 + \left(\pi\theta\right)\varepsilon_0\varepsilon_2 + \left(\pi\theta^2\right)\varepsilon_1\varepsilon_2\right).$$

Then the bias of $\hat{\bar{Y}}_{\mathrm{N}}$ can be achieved by

$$Bias\left(\hat{\bar{Y}}_{\mathrm{N}}\right) = -\pi\bar{Y}\left(\frac{1}{r} - \frac{1}{n}\right)\theta\rho C_x C_y. \tag{16}$$

Under the assumption terms of ε involving powers greater than two are small enough to be negligible, the approximate MSE of $\hat{\bar{Y}}_{\mathrm{N}}$ is:

$$MSE\left(\hat{\bar{Y}}_{\mathrm{N}}\right) \cong E\left(\bar{Y}\left(\varepsilon_0 + \left(-\pi\theta\right)\varepsilon_1 + \left(\pi\theta\right)\varepsilon_2\right)\right)^2$$

$$= \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2 C_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}^2\left(\pi^2\theta^2 C_x^2 - 2\pi\theta\rho C_x C_y\right). \tag{17}$$

The members of the proposed estimators are represented in Table 2.

Table 2. Members of the proposed estimator

Estimator	A	D
$\hat{\bar{Y}}_{\mathrm{N1}} = \bar{y}_r\left(\frac{\bar{x}^*}{\bar{x}_n}\right)$	1	0
$\hat{\bar{Y}}_{\mathrm{N2}} = \bar{y}_r\left(\frac{C_x\bar{x}^* + \beta_1}{C_x\bar{x}_n + \beta_1}\right)$	C_x	β_1
$\hat{\bar{Y}}_{\mathrm{N3}} = \bar{y}_r\left(\frac{Q_1\bar{x}^* + Q_a}{Q_1\bar{x}_n + Q_a}\right)$	Q_1	Q_a
$\hat{\bar{Y}}_{\mathrm{N4}} = \bar{y}_r\left(\frac{\beta_1\bar{x}^* + \beta_2}{\beta_1\bar{x}_n + \beta_2}\right)$	β_1	β_2
$\hat{\bar{Y}}_{\mathrm{N5}} = \bar{y}_r\left(\frac{\beta_1\bar{x}^* + Q_1}{\beta_1\bar{x}_n + Q_1}\right)$	β_1	Q_1
$\hat{\bar{Y}}_{\mathrm{N6}} = \bar{y}_r\left(\frac{\beta_2\bar{x}^* + Q_2}{\beta_2\bar{x}_n + Q_2}\right)$	β_2	Q_2
$\hat{\bar{Y}}_{\mathrm{N7}} = \bar{y}_r\left(\frac{Q_d\bar{x}^* + C_x}{Q_d\bar{x}_n + C_x}\right)$	Q_d	C_x

3 Efficiency Comparison

The proposed estimator ($\hat{\bar{Y}}_{\mathrm{N}}$) efficiency is measured against the mean imputation estimator ($\hat{\bar{Y}}_{\mathrm{S}}$), ratio imputation estimator ($\hat{\bar{Y}}_{\mathrm{Rat}}$), and [37]'s estimator ($\hat{\bar{Y}}_{\mathrm{TL}}$) using the MSE as criteria.

1) $\hat{\bar{Y}}_{\mathrm{N}}$ outperforms $\hat{\bar{Y}}_{\mathrm{S}}$ if

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

$$\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2C_y^2+\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}^2\left(\pi^2\theta^2C_x^2-2\pi\theta\rho C_xC_y\right)$$

$$<\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2C_y^2$$

$$\pi^2\theta^2C_x^2-2\pi\theta\rho C_xC_y<0$$

$$\rho>\frac{\pi\theta C_x}{2C_y}$$

2) $\hat{\bar{Y}}_{\mathrm{N}}$ performs better than $\hat{\bar{Y}}_{\mathrm{Rat}}$ if

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

$$\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2C_y^2+\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}^2\left(\pi^2\theta^2C_x^2-2\pi\theta\rho C_xC_y\right)$$

$$<\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2C_y^2+\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}^2\left(C_x^2-2\rho C_xC_y\right)$$

$$\pi^2\theta^2C_x^2-2\pi\theta\rho C_xC_y<C_x^2-2\rho C_xC_y$$

3) $\hat{\bar{Y}}_{\mathrm{N}}$ performs better than $\hat{\bar{Y}}_{\mathrm{TL}}$ if

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

$$\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2C_y^2+\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}^2\left(\pi^2\theta^2C_x^2-2\pi\theta\rho C_xC_y\right)$$

$$<\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2C_y^2+\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}^2\left(\left(\theta-\beta K\right)^2\pi^2C_x^2-2\left(\theta-\beta K\right)\pi\rho C_xC_y\right)$$

$$\pi^2\theta^2C_x^2-2\pi\theta\rho C_xC_y<\left(\theta-\beta K\right)^2\pi^2C_x^2-2\left(\theta-\beta K\right)\pi\rho C_xC_y$$

$$\rho<\frac{\pi\left(2\theta-\beta K\right)C_x}{2C_y}.$$

4 Application to Covid-19 Data

The presented estimators are implemented on the COVID-19 dataset from Chiang Mai province, Thailand [38], and daily PM2.5 concentration levels [39] to approximate the COVID-19 case frequency each day and the amount of patients diagnosed with pneumonia on high-flow oxygen support. The data was gathered from 1 April 2022 to 31 July 2022 (with a population size of $N=122$).

Population I

Auxiliary variable (X): the daily PM2.5 concentration (micrograms per cubic meter). Study variable (Y): the daily confirmed cases. The description of the parameters is summarized as mentioned here:

$N=122,\ \bar{X}=18.16,\ \bar{Y}=117.88,\ \mathrm{C}_x=0.85,\ C_y=1.19,$
$\rho=0.79$

Population II

Auxiliary variable (X): the PM2.5 level in the air (micrograms per cubic meter) each day. Study variable (Y): the total patients suffering from pneumonia and requiring high-flow oxygen therapy. The description of the parameters is summarized as follows:

$N=122,\ \bar{X}=18.16,\ \bar{Y}=17.11,\ \mathrm{C}_x=0.85,\ C_y=0.60,$
$\rho=0.62$

A sample $n=36$ is chosen from inside a population $N=122$ through the means of SRSWOR containing 30% missing in the study variable. Estimated values, biases, and MSEs are computed using R program, [40]. The results are illustrated in Table 3. Figure 1 depicts the scatter plot between the daily reported cases of COVID-19 in relation to the quantity of PM2.5 that day and Figure 2 displays the total patients suffering from pneumonia and requiring high-flow oxygen therapy in relation to the daily quantity of PM2.5 in Chiang Mai, Thailand, respectively. Figure 3 shows the percentage relative efficiency of the estimators contrasted to the mean imputation method.

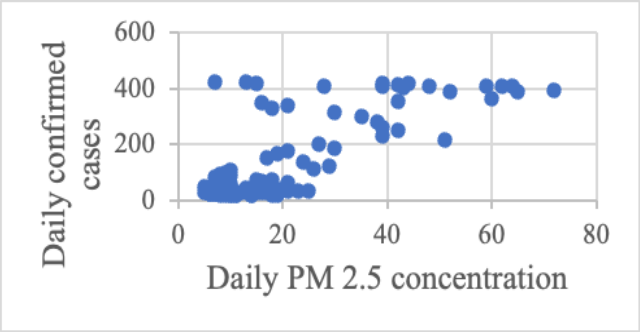

Fig. 1: The scatter plot between the reported cases of COVID-19 per day and the PM2.5 concentration by day in Chiang Mai, Thailand

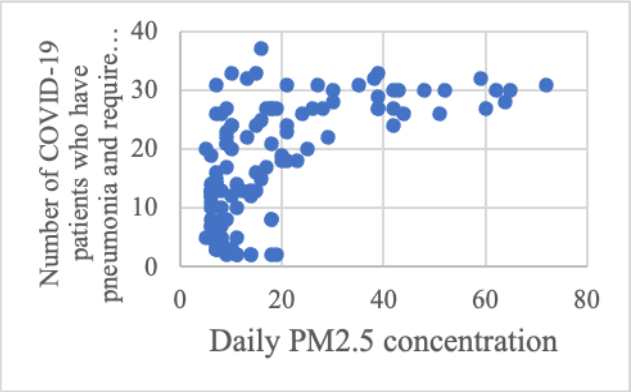

Fig. 2: The scatter plot between the amount of those diagnosed with COVID-19 who also have pneumonia and are on high-flow oxygen and the PM2.5 per day concentration in Chiang Mai, Thailand

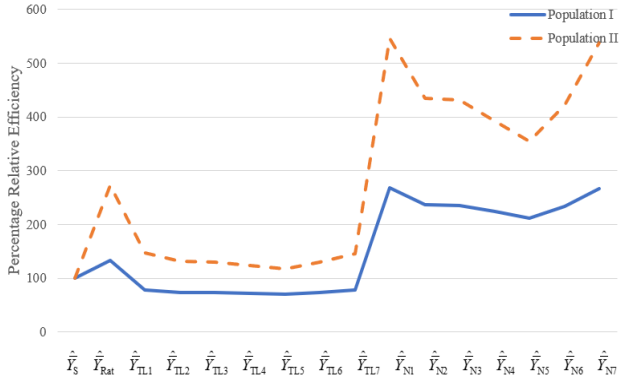

Fig. 3: The percentage relative efficiency of the estimators contrasted to the mean imputation method

Table 3. Estimated values, biases, and MSEs of the estimators when implemented on COVID-19 reports in Chiang Mai province

Estimator	Population I			Population II		
	Estimated daily confirmed cases	Bias	MSE	Estimated number of patients who have pneumonia and require high-flow oxygen	Bias	MSE
Mean imputation ($\hat{\bar{Y}}_S$)	90.48	27.40	750.52	14.14	2.98	8.87
Ratio imputation ($\hat{\bar{Y}}_{Rat}$)	94.18	23.69	561.36	15.31	1.80	3.25
[15]'s estimator ($\hat{\bar{Y}}_{TL1}$)	86.93	30.95	957.97	14.66	2.46	6.04
[15]'s estimator ($\hat{\bar{Y}}_{TL2}$)	86.00	31.88	1016.20	14.51	2.60	6.76
[15]'s estimator ($\hat{\bar{Y}}_{TL3}$)	85.97	31.91	1018.03	14.51	2.61	6.79
[15]'s estimator ($\hat{\bar{Y}}_{TL4}$)	85.56	32.32	1044.48	14.44	2.67	7.13
[15]'s estimator ($\hat{\bar{Y}}_{TL5}$)	85.11	32.77	1073.74	14.37	2.74	7.52
[15]'s estimator ($\hat{\bar{Y}}_{TL6}$)	85.88	32.00	1023.86	14.49	2.62	6.86
[15]'s estimator ($\hat{\bar{Y}}_{TL7}$)	86.86	31.02	961.96	14.65	2.47	6.09
Proposed ($\hat{\bar{Y}}_{N1}$)	**101.15**	**16.72**	**279.72**	**15.84**	**1.27**	**1.62**
Proposed ($\hat{\bar{Y}}_{N2}$)	100.07	17.80	316.96	15.69	1.43	2.04
Proposed ($\hat{\bar{Y}}_{N3}$)	100.04	17.84	318.15	15.68	1.43	2.06
Proposed ($\hat{\bar{Y}}_{N4}$)	99.56	18.32	335.48	15.61	1.50	2.26
Proposed ($\hat{\bar{Y}}_{N5}$)	99.04	18.84	354.92	15.53	1.58	2.50
Proposed ($\hat{\bar{Y}}_{N6}$)	99.93	17.94	321.95	15.67	1.45	2.10
Proposed ($\hat{\bar{Y}}_{N7}$)	**101.08**	**16.80**	**282.24**	**15.83**	**1.28**	**1.65**

5 Conclusions

We can see from Figure 1 and Figure 2 that there is a high correlation between both reported cases of COVID-19 per day and those who also have pneumonia and are on high-flow oxygen support and the PM2.5 levels in the air each day in Chiang Mai, Thailand which are equal to 0.79 and 0.62, respectively. From Table 3, it was found that the suggested estimators outperformed the estimators from the mean imputation and ratio imputation processes. In both populations, the proposed estimator $\hat{\bar{Y}}_{N1}$ gave the least bias and MSE in this scenario and the second best is $\hat{\bar{Y}}_{N7}$ which takes advantage of the accessible third quartile of the auxiliary variable. The mean imputation procedure did not deliver in this scenario. The best estimator $\hat{\bar{Y}}_{N1}$ estimated the daily confirmed cases to be around 101 cases and the estimated total patients diagnosed with pneumonia receiving high-flow oxygen support is around 16 cases. The results in Figure 1 and Figure 2 also support the results found in Table 3. Figure 3 showed that the best estimator $\hat{\bar{Y}}_{N1}$, gave a more satisfactory percentage relative efficiency in contrast to the mean

imputation method which is close to the second best $\hat{\bar{Y}}_{N7}$ which is almost 300 and 500 percent in this application for the daily incidence of COVID-19 and the number of patients who have pneumonia and require high-flow oxygen in Chiang Mai, respectively.

The positive relationship between the COVID-19 incidence and daily PM2.5 in Chiang Mai, Thailand can be useful in investigating the estimated average daily number of confirmed cases of COVID-19 incidence and the total patients diagnosed with pneumonia receiving high-flow oxygen support using a novel class of estimators focusing on the transformation of an auxiliary variable when missing data present themselves in the study. The bias and MSE of the updated transformed estimator have been studied and showed that the proposed estimators exceeded the mean and ratio procedures of imputation. The results gained from implementing the COVID-19 data illustrated that the novel estimators excelled by giving a reduced bias and MSE compared to existing ones and can give a more accurate value of the estimated daily confirmed cases and the amount of patients with pneumonia and receiving a high-flow oxygen supply based on PM2.5 concentration on each day. The inaccessible population mean of the auxiliary variable makes the suggested estimators more practical to use with real-world data. The proposed estimators can help in estimating other study variables that contain missing values that must be eliminated before future analysis. The proposed estimators can also be developed in other survey designs such as double sampling, stratified random sampling, and cluster sampling. Nevertheless, reducing PM2.5 may help reduce the daily incidence of COVID-19 and the total amount of cases of pneumonia requiring high-flow oxygen support and therefore can benefit in preventing high COVID-19 incidence.

Acknowledgment:
Thank you to all the referees for their feedback and criticism to help improve this paper.

References:

[1] Nangsue, C., Srisurapanont, K. and Sudjaritruk, T., A comparison of the immunogenicity and safety of an additional heterologous versus homologous COVID-19vaccinationamongnon-seroconvertedimmunocom promised patients after a two-dose primary series of mRNA vaccination: A systematic review and meta analysis, *Vaccines* , Vol. 12, No. 5, 2024, pp. 468. https://doi.org/10.3390/vaccines12050468.

[2] Lee, A.R.Y.B., Wong, S. Y. , Chai, L.Y.A., Lee, S.C., Lee, M.X., Muthiah, M.C. ,Tay, S.H., Teo, C.B., Tan, B.K.J, Chan, Y.H., Sandar, R. and Soon Y.Y., Efficacy of COVID-19 vaccines in immunocompromised patients: Systematic review and meta-analysis. *BMJ*, 2022, pp. 376:e068632. https://doi.org/10.1136/bmj-2021-068632.

[3] Petrone, L., Sette, A., de Vries, R.D. and Goletti, D., The importance of measuring SARS-CoV-2-specific T-cell responses in an ongoing pandemic*, Pathogens*, , Vol.12, 2023, pp. 862. https://doi.org/10.3390/pathogens12070862.

[4] Lym, Y. and Kim, K.J.B. , Exploring the effects of PM2.5 and temperature on COVID-19 transmission in Seoul, South Korea, *Environmental Research* , Vol. 203, 2022, pp. 111810. https://doi.org/10.1016/j.envres.2021.111810.

[5] Jiang, Y. Wu, X.J. and Guan, Y.J. , Effect of ambient air pollutants and meteorological variables on COVID-19 incidence, *Infection Control & Hospital Epidemiology,* Vol. 41, No. 9, 2020, pp. 1011-1015, https://doi.org/10.1017/ice.2020.222.

[6] Liu, C., Peng, J., Liu, Y., Peng, Y., Kuang, Y., Zhang, Y. and Ma, Q., Causal relationship between particulate matter 2.5 (PM2.5), PM2.5 absorbance, and COVID-19 risk: A two-sample Mendelian randomisation study, *Journal of Global Health,* , 2023, Vol. 13, 06027. https://doi.org/10.7189/jogh.13.06027.

[7] Martelletti, L. and Martelletti, P. , Air pollution and the novel Covid-19 disease: A putative disease risk factor, *SN Comprehensive Clinical Medicine,* , Vol. 2, No.4, 2020, pp. 383–387. https://doi.org/10.1007/s42399-020-00274-4.

[8] Zhao, M., Liu, Y. and Gyilbag, A., Assessment of meteorological variables and air pollution affecting COVID-19 cases in urban agglomerations: evidence from China, *International Journal of Environmental*

Research and Public Health, , Vol. 19, No. 1, 2022, pp. 531. https://doi.org/10.3390/ijerph19010531.

[9] Adhikari, A. and Yin, J., Short-term effects of ambient ozone, PM2.5, and meteorological factors on COVID-19 confirmed cases and deaths in Queens, New York, *International Journal of Environmental Research and Public Health*, Vol. 17, No. 11, 2020, pp.4047. https://doi.org/10.3390/ijerph17114047.

[10] Zhu, Y., Xie, J., Huang, F. and Cao, L., Association between short-term exposure to air pollution and COVID-19 infection: evidence from China, *Science of The Total Environment*, Vol. 727, 2020, pp. 138704. https://doi.org/10.1016/j.scitotenv.2020.138704

[11] Meo, S.A. Al-khlaiwi, T. and Ullah, C.H., Effect of ambient air pollutants PM2.5 and PM10 on COVID-19 incidence and mortality: observational study, *European Review for Medical and Pharmacological Sciences*, Vol. 25, No. 23, 2021, pp. 7553-7564. https://doi.org/10.26355/eurrev_202112_27455.

[12] Sharma, G., Upadhyay, E., Kulkarni, A. and Sagalgile, A. , COVID-19 transmission due to interplay between PM2.5 and weather conditions, *Journal of Associated Medical Sciences*, Vol. 57, No. 1, 2024, pp. 104-111. https://doi.org/10.12982/JAMS.2024.012.

[13] Nguyen, T.T.N., Le, T. C., Sung, Y.T., Cheng, F.Y., Wen, H.C., Wu, C.H., Aggarwal, S.G. and Tsai, C.J., The influence of COVID-19 pandemic on $PM_{2.5}$ air quality in Northern Taiwan from Q1 2020 to Q2 2021,*Journal of Environmental Management,* Vol. 343, 2023, pp. 118252. https://doi.org/10.1016/j.jenvman.2023.118252.

[14] Lawson, N. and Thongsak, N., A class of population mean estimators in stratified random sampling: a case study on fine particulate matter in the north of Thailand, *WSEAS Transactions on Mathematics*, Vol. 23, 2024, pp. 160-166. https://doi.org/10.37394/23206.2024.23.19.

[15] Ponkaew, C. and Lawson, N. New product estimators for population mean under unequal probability sampling with missing data: a case study on the number of new COVID-19 patients, *Thailand Statistican*, Vol. 22, No. 3, 2024, pp. 634-656.

[16] Nangsue, N, Adjusting for nonresponse in the analysis and estimation of sample survey data for cluster designs. University of Southampton, Social Sciences, Doctoral Thesis, 2014, 145pp.

[17] Nangsue, N. Adjusted ratio and regression type estimators for estimation of population mean when some observations are missing, *International Journal of Mathematical and Computational Sciences*, Vol.3, No.5, 2009, pp. 334-337. https://doi.org/10.5281/zenodo.1082821

[18] Van Buuren, S., Brand, J.P., Groothuis-Oudshoorn, C.G. and Rubin, D.B., Fully conditional specification in multivariate imputation, *Journal of Statistical Computation and Simulation,* Vol. 76, No. 12, 2006, pp. 1049–1064. https://doi.org/10.1080/10629360600810434

[19] Shao, J. and Wang, H., Sample correlation coefficients based on survey data under regression imputation, *Journal of the American Statistical Association*, Vol. 97, No. 458, 2002, pp. 544–552, [Online]. https://www.jstor.org/stable/3085670 (Accessed Date: November 10, 2024).

[20] Skinner, C. J., Calibration weighting and non-sampling errors, *Research in Official Statistics,* Vol. 2, 1999, pp. 33–43.

[21] Skinner, C. J. and Coker, O., Regression analysis for complex survey data with missing values of a covariate,*Journal of the Royal Statistical Society Series A: Statistics in Society,* ., Vol.159, No. 2, 1996, pp. 265–274. https://doi.org/10.2307/2983173.

[22] Skinner, C. J. and Darrigo, J., Inverse probability weighting for clustered nonresponse, *Biometrika*, Vol. 98, No. 4, 2011, pp. 953–966. https://doi.org/10.1093/biomet/asr058.

[23] Skinner, C. J. and Rao, J. N. K., Jackknife variance estimation for multivariate statistics under hot-deck imputation from common donors, *Journal of Statistical Planning and Inference,* Vol. 102, No. 1, 2002, pp. 149–167. https://doi.org/10.1016/S0378-3758(01)00185-9.

[24] Roderick, J. L. and Rubin, D. B., *Statistical analysis with missing data*, Wiley, Hoboken, NJ, 2020.

[25] Andridge, R. R. and Little, R. J. A., A review of hot deck imputation for survey non-response, *International Statistical Review*, Vol. 78, No. 1, 2010, pp. 40–64. https://doi.org/10.1111/j.1751-5823.2010.00103.x.

[26] Y. G. Berger, J. N. K. Rao, Adjusted jackknife for imputation under unequal probability sampling without replacement, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Vol. 68, No. 3, 2006, pp. 531–547. https://doi.org/10.1111/j.1467-9868.2006.00555.x.

[27] Brick, J. M. and Kalton, G., Handling missing data in survey research, *Statistical Methods in Medical Rese arch*, Vol. 5, No. 3, 1996, pp.215–238. https://doi.org/10.1177/096228029600500302.

[28] Groves, R. M., Dilman, D. A., Eltinge, J. L. and Little, R. J. A., *Survey nonresponse*, New York: John Wiley and Sons, Inc, 2002.

[29] Haziza, D. and Rao, J., A nonresponse model approach to inference under imputation for missing survey data, *Survey Methodology*, Vol. 32, No. 1, 2006, pp.53–64.

[30] Yuan, Y. C., *Multiple imputation for missing data: Concepts and new development*, SAS Institute Inc., Rockville, 2000.

[31] Singh, S. and Horn, S., Compromised imputation in survey sampling, *Metrika*, Vol. 51, No. 3, 2000, pp. 267-276. https://doi.org/10.1007/s001840000054.

[32] Srivenkataramana, T., A dual to ratio estimator in sample surveys, *Biometrika*, 1980, Vol. 67, No. 1, 1980, pp. 199–204.

[33] Thongsak, N. and Lawson, N., Classes of dual to modified ratio estimators for estimating population mean in simple random sampling, *Proceedings of the 2021 Research, Invention and Innovation Congress*, Bangkok, Thailand; 2021 Sep 1-2; Bangkok, Thailand, 2021.

[34] Upadhyaya, L. N., Singh, G.N. and Singh, H.P., Use of transformed auxiliary variable in the estimation of population ratio in sample survey, *Statistics in Transition*, Vol. 4, 2000, pp. 1019-1027.

[35] Onyeka, A. C., Nlebedim, V.U. and Izunobi, C.H., Estimation of population ratio in simple random sampling using variable transformation, *Far East Journal of Theoretical Statistics*, Vol. 13, 2013, pp. 56-65.

[36] Onyeka, A. C., Nlebedim, V.U. and Izunobi, C.H., A class of estimators for population ratio in simple random sampling using variable transformation, *Open Journal of Statistics*, Vol.4, 2014, pp.284-291. https://doi.org/10.4236/ojs.2014.44029.

[37] Thongsak, N. and Lawson, N., Transformed regression type estimators in the presence of missing observations: case studies on COVID-19 incidence in Chiang Mai, Thailand, *WSEAS Transactions on Biology and Biomedicine*, Vol. 21, 2024, pp.131–137. https://doi.org/10.37394/23208.2024.21.13.

[38] Chiangmai, Covid-19 situation in Chiang Mai Province, [Online]. https://www.chiangmai.go.th/covid19/index.html (Accessed Date: November 10, 2024).

[39] Pollution Control Department, Daily PM2.5 concentration; 2022, [Online]. http://air4thai.pcd.go.th/webV2/history/ (Accessed Date: November 10, 2024).

[40] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2021, [Online]. https://www.R-project.org/ (Accessed Date: November 5, 2024).

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed to the present research, at all stages from the formulation of the problem to the final findings and solution

Sources of funding for research presented in a scientific article or scientific article itself

This work was funded by King Mongkut's University of Technology North Bangkok, Thailand. Contract number sci-680035.

Conflict of Interest

The authors have no conflicts of interest to declare.