Vector Autoregressive-Moving Average Imputation Algorithm for Handling Missing Data in Multivariate Time Series

I Made Sumertajaya, Embay Rohaeti, Aji Hamim Wigena, and Kusman Sadik

***Abstract*—The issue of missing data is a significant problem in time series. There have been various imputation techniques recommended for handling missing data. Among them, univariate imputation methods are frequently preferred for their adaptability. Nevertheless, in the case of multivariate time series data where the variables are interrelated, these methods tend to eliminate critical information. In such cases, multivariate imputation methods, such as Vector Autoregressive-Imputation Method or more commonly known as VAR-IM, are more suitable since they can enhance imputation accuracy by leveraging information from other variables. To further improve imputation accuracy, a modified VAR-IM called VAR-IMMA was introduced. The objective of this study was to assess and compare the effectiveness of the original VAR-IM method and the newly proposed VAR-IMMA method through the use of simulation studies. The simulation was repeated 100 times at different proportions of missing data from 5% to 30%. The data used were the monthly inflation rate of 82 cities in Indonesia from January 2014 to December 2019. The performances of each methods were evaluated by measuring RMSE and MAE. The results show that VAR-IMMA performs similarly to VAR-IM in data with 5% missing data. But, for larger percentages of missing data, the VAR-IMMA performs better with lower values of RMSE and MAE.**

***Index Terms*—missing data imputation, moving average, VAR-IMMA, vector autoregressive model**

I. Introduction

REAL-LIFE data are rarely clean. One of the most common problems in data quality is missing data, which can be caused by various factors such as respondent non-compliance, technical problems during data recording, unavailability of complete information, or lack of knowledge [1], [2], [3]. Not only is it the most common problem in data quality, but it is also one of the most critical problems, particularly in time series datasets because of the interconnectivity of data points over time. Therefore, time series with missing data generally have to be pre-processed before further analysis [4].

There exist multiple approaches to address missing data, with the listwise deletion technique being among the simplest and most direct techniques [5]. This technique removes any observation with missing values. However, the technique is not a viable option for dynamic modeling, such as multivariate time series data modeling, as the removal of data can reduce the number of variables and the length of the series. In contrast, in dynamic modeling, all data are fundamental to estimating the current value so that they can affect the estimation of the model. For this specific reason, imputation methods are often preferable. In the imputation method, the missing data from all available data is re-entered into the missing data [6].

Previously, numerous imputation methods have been proposed. Among the methods, the relatively simple and adaptive methods are univariate imputation methods. Methods such as interpolation or moving average [7] are categorized as univariate imputation methods since the imputation can only be carried out in one variable at once. While flexibility is generally advantageous, it can be a drawback when handling multivariate time series data, particularly when the data exhibits high correlation among the variables. In such cases, multivariate imputation methods such as Vector Autoregression Imputation Method (VAR-IM) [8] are more likely to have better imputation accuracy.

VAR-IM is an iteration process of handling missing data based on a vector autoregressive model. Therefore, an initial imputation must be performed to form an initial model. One of the techniques to define the initial imputation is a simple imputation [8]. Based on that, this study aimed to compare existing VAR-IM and the newly proposed method which accommodate moving average (MA) in the initial imputation. The newly proposed method is called as VAR-IMMA (Vector Autoregressive Imputation Method Moving Average).

II. Multivariate Time Series Modeling

In this study, the imputation process is based on VAR models. There are five modeling stages used which are stationarity testing using Augmented Dickey-Fuller (ADF) test [9], [10], [11]; cointegration testing [12]; determining the optimum lag [8], [13], [14]; parameter estimation [15], [16], [17]; and evaluation of the best model [18], [19], [20].

Manuscript received September 13, 2022; revised March 11, 2023.

I Made Sumertajaya is an Associate Professor at Department of Statistics, IPB University, Bogor, West Java, 16144, Indonesia. (phone: +62 (251) 8624535, fax: +62 (251) 8624535, e-mail: imsjaya@apps.ipb.ac.id).

Embay Rohaeti is a PhD candidate of IPB University, Bogor, West Java, 16144, Indonesia (e-mail: embay.rohaeti@unpak.ac.id).

Aji Hamim Wigena is a Professor at Department of Statistics, IPB University, Bogor, West Java, 16144, Indonesia (e-mail: aji_hw@apps.ipb.ac.id).

Kusman Sadik is a senior lecturer in Department of Statistic, IPB University, Bogor, West Java, 16144, Indonesia (e-mail: kusmans@apps.ipb.ac.id).

A. Vector autoregressive models

According to [9], [12], [21], [22], [23], [24], Vector Autoregressive model (VAR) is a regression system consisting of equations in which each variable, including itself, is regressed against other variables at the preceding time point. Its general form with order p, VAR(p), is presented in (1).

$$y_t = A_0 + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + u_t \tag{1}$$

Where y_t, y_{t-1} is an $(n \times 1)$ vector, containing n variables included in the model at times $t = 1, 2, \ldots, T$. $i = 1, 2, \ldots, p$ is the lag while p is the order of VAR. $A_0 = (A_{10}, A_{20}, \ldots, A_{n0})'$ is an $(n \times 1)$ intercept vector, A_i is an (n x n) coefficient matrix, and $u_t = (u_{1t}, u_{2t}, \ldots, u_{nt})'$ is an $(n \times 1)$ white noise vector [24].

Dealing with non-stationary data that has one or multiple cointegration relationships is possible using a Vector Error Correction Model (VECM). VECM general form is shown in (2).

$$\Delta y_t = A_0 + \pi y_{t-1} + \sum_{i=1}^{p-1} \Gamma \Delta y_{t-i} + u_t \tag{2}$$

Where Δ is a difference operator, that is $\Delta y_t = y_t - y_{t-1}$, and y_{t-1} is an $(n \times 1)$ first lag variable vector, A_0 is an $(n \times 1)$ intercept, π is an $(n \times k)$ cointegration coefficient matrix, Γ_i is an $(n \times n)$ coefficient matrix of the i-th variable, with $i = 1, 2, \ldots, p-1$, and u_t is an $(n \times 1)$ error vector.

B. The VAR (1) model for two variables

A two-variable VAR(1) model can be written as (3).

$$\begin{aligned} y_{1,t} &= A_{10} + A_{11} y_{1,t-1} + A_{12} y_{2,t-1} + u_{1,t} \\ y_{2,t} &= A_{20} + A_{21} y_{1,t-1} + A_{22} y_{2,t-1} + u_{2,t} \end{aligned} \tag{3}$$

Time index t is $t = 1, 2, \ldots, T$. The first equation can be decomposed as (4).

$$\begin{aligned} y_{1,1} &= A_{10} + A_{11} y_{1,1-1} + A_{12} y_{2,1-1} + u_{1,1} \\ y_{1,2} &= A_{10} + A_{11} y_{1,2-1} + A_{12} y_{2,2-1} + u_{1,2} \\ &\vdots \\ y_{1,T} &= A_{10} + A_{11} y_{1,T-1} + A_{12} y_{2,T-1} + u_{1,T} \end{aligned} \tag{4}$$

Similarly, the second variable can be decomposed as (5).

$$\begin{aligned} y_{2,1} &= A_{20} + A_{21} y_{2,1-1} + A_{22} y_{2,1-1} + u_{2,1} \\ y_{2,2} &= A_{20} + A_{21} y_{1,2-1} + A_{22} y_{2,2-1} + u_{2,2} \\ &\vdots \\ y_{2,T} &= A_{20} + A_{21} y_{1,T-1} + A_{22} y_{2,T-1} + u_{2,T} \end{aligned} \tag{5}$$

Equations (4) and (5) can also be written as matrices (6).

$$\begin{aligned} Y &= \begin{bmatrix} y_{1,1} & y_{2,1} \\ y_{1,2} & y_{2,2} \\ \vdots & \vdots \\ y_{1,T} & y_{2,T} \end{bmatrix}_{T \times 2}, W = \begin{bmatrix} 1 & y_{1,1-1} & y_{2,1-1} \\ 1 & y_{1,2-1} & y_{2,2-1} \\ \vdots & \vdots & \vdots \\ 1 & y_{1,T-1} & y_{2,T-1} \end{bmatrix}_{T \times 3}, \\ A &= \begin{bmatrix} A_{10} & A_{20} \\ A_{11} & A_{21} \\ A_{12} & A_{22} \end{bmatrix}_{3 \times 2}, u = \begin{bmatrix} u_{1,1} & u_{2,1} \\ u_{1,2} & u_{2,2} \\ \vdots & \vdots \\ u_{1,T} & u_{2,T} \end{bmatrix}_{T \times 2} \end{aligned} \tag{6}$$

So that, (4) and (5) can be simplified as (7).

$$Y = WA + u \tag{7}$$

C. Stationary test

The hypothesis that will be tested in the ADF test is as follows.

H_0 : data contains unit root or is not stationary ($A = 0$)

H_1 : data does not contain unit roots or is stationary ($A < 0$)

Meanwhile, the test statistic used in the ADF test is shown in (8).

$$t_{statistic} = \frac{\hat{A}}{\sigma_{\hat{A}}} \tag{8}$$

Where $\hat{A}$ is the estimated value of the intercept A, and $\sigma_{\hat{A}}$ is the standard deviation of $\hat{A}$. H_0 is rejected if $p - value < \alpha = 5\%$, which means that the observed time series data is stationary [9], [24]. Non-stationary data can be converted to stationary data using differencing [11].

D. Optimal lag

Determining the optimal lag of a VAR model is generally done using information criteria, such as the last error prediction (FPE), Hannan-Quinn (HQ), Schwarz (SC), and Akaike (AIC) [8]. Equation (9) to (12) provide the mathematical representation of these criteria.

$$FPE(\rho) = \left[\frac{T + s\rho + 1}{T - s\rho - 1} \right]^k \left| \sum(\rho) \right| \tag{9}$$

$$HQ(\rho) = \ln \left| \sum(\rho) \right| + \frac{2 \ln(\ln T)}{T} \rho s^2 \tag{10}$$

$$SC(\rho) = \ln \left| \sum(\rho) \right| + \frac{\ln T}{T} \rho s^2 \tag{11}$$

$$AIC(\rho) = \ln \left| \sum(\rho) \right| + \frac{2}{T} \rho s^2 \tag{12}$$

Where $\Sigma(\rho)$ is a covariance matrix. T and s are consecutively the numbers of observations and variables, while p is the lag. The lag with the smallest criteria value is the optimal lag [13], [14].

E. The goodness of multivariate time series models

The Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are commonly employed metrics to evaluate a model goodness. Each metric has its characteristics; therefore, combining both metrics will complete each other [18], [19]. RMSE and MAE are calculated based on the formulas in (13) and (14).

$$RMSE = \left(\frac{\sum_{t=1}^{n} \left(y_t - \hat{y}_t \right)^2}{n} \right)^{1/2} \quad (13)$$

$$MAE = \frac{\sum_{t=1}^{n} \left| y_t - \hat{y}_t \right|}{n} \quad (14)$$

Here, t represents the time index, y_t represents the observed value, $\hat{y}_t$ represents the corresponding estimated value, and n denotes the total number of observations.

III. Missing Data Problems

Based on [24], [6], [25], there are three types of missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Missing data in MCAR does not depend on the data value or variable, meaning each observation has the same likelihood of missing. On the other hand, in MAR type, each observation within the same group has the same likelihood of missing, but each observation between different groups has a different likelihood of missing, which means that the missing data in MAR type is related to other observed data. On the contrary, missing data in MNAR type is related to the value itself.

VAR-IM, as an imputation method, uses expectation maximization (EM) algorithm. In estimating parameters, the algorithm works interactively based on maximum likelihood [26] when some variables are incomplete. As detailed in [8], [6], [25], the iterative process of EM consists of two steps: the expectation step (E-step), and the maximization step (M-step). During the E-step, missing data is imputed by calculating expected values based on the available information. In the subsequent M-step, the model parameters are estimated by maximizing the likelihood of the observed and imputed data. Both steps are repeated until convergence, or the maximum iteration is reached.

IV. Methodology

This study used monthly inflation rate data of preserved fish (PF), fresh fish (FF), and vegetables (VG) from 82 cities in Indonesia from January 2014 to December 2019, provided by the Central Bureau of Statistics (BPS) of Indonesia. The simulation is performed using R and repeated 100 times for each proportion and each city. Fig. 1 shows the simulation flowchart.

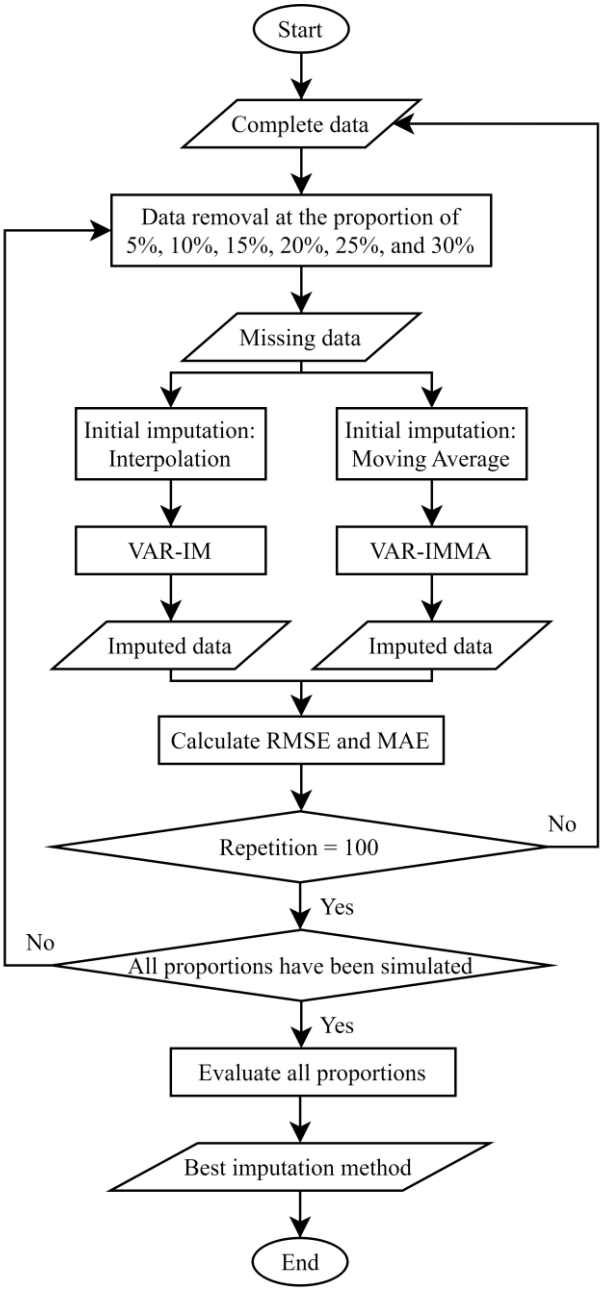

Fig. 1. The simulation flowchart

The simulation process is divided into these three stages:

1) Randomly set some proportions of data become missing which satisfies MCAR assumptions. Data removal is performed on several missing proportions (5%, 10%, 15%, 20%, 25%, and 30%). A city will later be randomly chosen as a study case.
2) Impute the missing data using VAR-IM and VAR-IMMA.
3) Evaluate the imputation accuracy using RMSE and MAE. The performance evaluation is also based on the stability of the performance of the two ways at various proportions of missing data.

V. Result and Discussion

VAR models requires complete observations, so as VAR-IM. Consequently, the initial imputation plays a vital role in VAR-IM accuracy. Different initial values can lead to

different VAR models, which can lead to different imputation results.

Previous studies have proposed various imputation methods, one of which is interpolation. The original article of VAR-IM also used interpolation as in the initial imputation technique [8]. Interpolation may work well as an initial imputation if the data is linearly patterned. However, when the data are volatile, it raises a research question about whether a linear approach can still work well. Based on this question, we proposed a simple development using moving average as the initial imputation method. Moving average is chosen for its ability to follow data fluctuations when the data are volatile. This development will then be referred to as VAR-IMMA.

A. Comparison between linear interpolation and moving average

Linear interpolation is obtained by drawing a straight line from the last observed data to the first observed data after the missing data. Based on [27], imputation by linear interpolation for missing data in period t, y_t, is calculated using (15).

$$y_t = \left(p_t - p_{sb}\right)\frac{\left(y_{st} - y_t\right)}{p_{st} - p_t} + y_{sb} \tag{15}$$

p_t refers to the index of the data that is missing and will be imputed. p_{sb} refers to the index of the most recent observed data. p_{st} refers to the index of the first observed data after the missing data. y_{sb} is the value of the most recent observed data, while y_{st} is the value of the first observed data after the missing data.

On the other hand, moving averages are obtained by averaging values within a time window. We used exponential moving average (EMA) defined by [27] as the initial imputation method, which can be calculated using (16).

$$EMA_t = \frac{\sum_{i=1}^{k}(1-\alpha)^i y_{t-i} + \sum_{i=1}^{k}(1-\alpha)^i y_{t+i}}{2\left(\sum_{i=1}^{k}(1-\alpha)^i\right)} \tag{16}$$

t is the time index to be imputed, α is a constant, which in this study, is equal to $\alpha = 0.5$. k is an integer representing the window length. $k = 2$ means EMA will be calculated using two observations before and two observations after the missing value to be imputed. For cases where data are lost sequentially, EMA is calculated using the non-missing value(s) only. If all values within the time window are missing, the time window is to be expanded, a period before and a period after the former time window, until there are at least two non-missing values [27].

Table I and Table II illustrate the comparison between linear interpolation and EMA. Table I shows data with MCAR values, while Table II shows the imputation results.

TABLE I
ILLUSTRATION OF DATA WITH MCAR VALUES

X	PF	FF	VG
1	0.60	1.57	0.91
2	0.37	2.33	6.33
3	0.22	1.75	1.46
4	0.42	-0.15	NA*
5	0.88	0.65	-2.94
6	NA*	0.46	-0.40
7	-0.24	NA*	4.97
8	0.11	NA*	0.21
9	0.39	2.60	-4.01
10	1.37	0.03	-0.60

*NA = Missing data

TABLE II
COMPARISON OF IMPUTATION RESULTS
BETWEEN LINEAR INTERPOLATION AND EMA

X	PF	FF	VG
1	0.60	1.57	0.91
2	0.37	2.33	6.33
3	0.22	1.75	1.46
4	0.42	-0.15	-0.74* 0.49**
5	0.88	0.65	-2.94
6	0.32* 0.30**	0.46	-0.40
7	-0.24	1.19* 1.06**	4.97
8	0.11	1.90* 1.43**	0.21
9	0.39	2.60	-4.01
10	1.37	0.03	-0.60

* Linear Interpolation
** EMA

Here are the more detailed calculation that was used to obtain imputation in Table II. Based on (15), linear interpolation on the sixth value in column PF can be calculated as follows.

$$PF_6 = \left(p_6 - p_5\right)\frac{\left(y_7 - y_5\right)}{\left(p_7 - p_5\right)} + y_5$$
$$PF_6 = (6-5)\frac{(-0.24 - 0.88)}{(7-5)} + 0.88$$
$$PF_6 = 0.32$$

Meanwhile, the imputation on the same missing data using EMA can be calculated as follows.

$$PF_6 = \frac{\left(\frac{1}{2}\right)^1 PF_5 + \left(\frac{1}{2}\right)^1 PF_7 + \left(\frac{1}{2}\right)^2 PF_4 + \left(\frac{1}{2}\right)^2 PF_8}{\frac{1}{2}+\frac{1}{2}+\frac{1}{4}+\frac{1}{4}}$$
$$PF_6 = \frac{\frac{1}{2}(0.88) + \frac{1}{2}(-0.24) + \frac{1}{4}(0.42) + \frac{1}{4}(0.11)}{\frac{3}{2}}$$
$$PF_6 = 0.30$$

B. Handling missing data using VAR-IM and VAR-IMMA

Instead of linear interpolation, EMA is used as the initial imputation in VAR-IMMA. EMA is chosen because the method can follow data fluctuations better in data with votality. Fig. 2 shows the flowchart of VAR-IMMA.

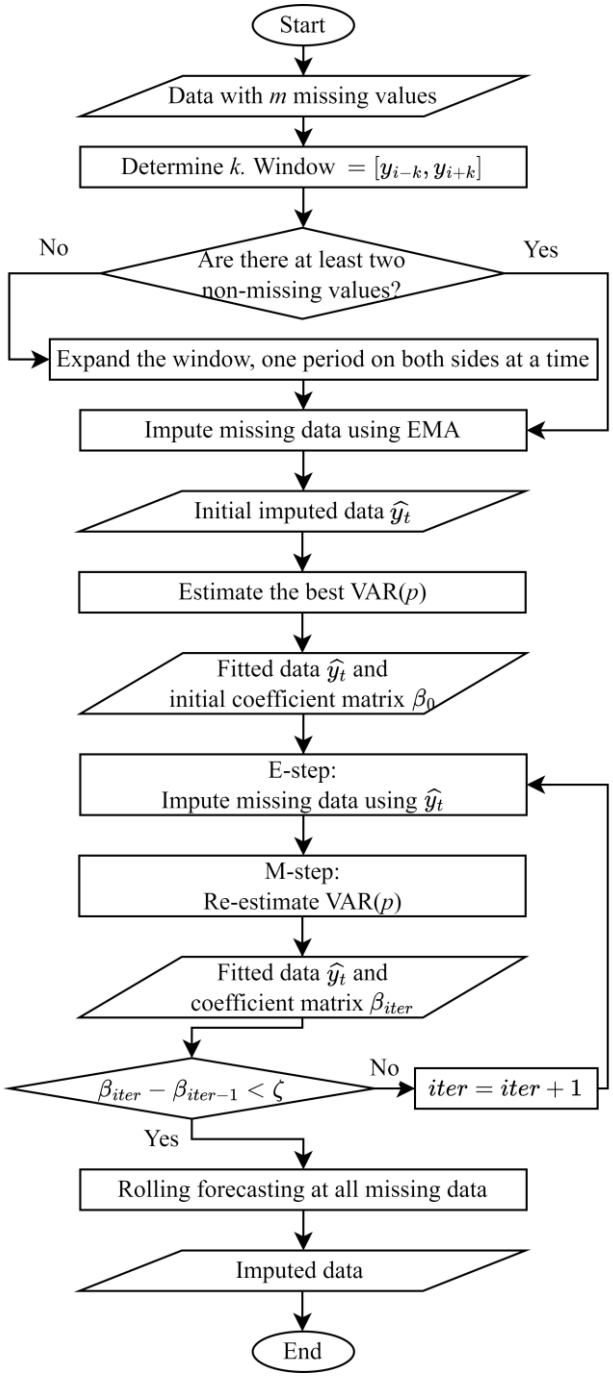

Fig. 2. Flowchart of VAR-IMMA

The algorithm of VAR-IMMA shown in Fig. 2 is as follows:

1) Calculate the initial imputation using EMA. A time window sized 2*k* (*k* previous and *k* after the missing value) has to be defined where *k* is a pre-specified integer. If all values within the time window are missing, expand the time window by 2 observations, 1 before and 1 after the current time window, until there is at least two present value.
2) Using the initial imputed data, find the optimal VAR(*p*) model. The fitted data will be used to initiate Expectation-Maximization (EM) algorithm, while the coefficient matrix will be used to determine convergency.
3) E-step: impute missing data using the fitted data.
4) M-step: reestimate the best VAR(*p*) model based on the updated data from E-step.
5) Repeat steps (2) and (3) until a convergency is reached, that is, if the differences between the current and previous coefficient matrices are less than a pre-specified threshold.
6) If a convergency is reached, perform a forecasting process on the missing data using the convergent model to estimate the final imputed data.

The mean of 100 randomly repeated simulations was calculated and shown in Table III. Table III compares the accuracy of the proposed VAR-IMMA and the original VAR-IM methods.

TABLE III
COMPARISON OF IMPUTATION ACCURACY

Missing Data Proportion	RMSE		MAE	
	VAR-IM	VAR-IMMA	VAR-IM	VAR-IMMA
0.05	4.79	**4.72**	3.54	**3.49**
0.10	5.45	**5.30**	3.93	**3.82**
0.15	6.12	**5.89**	4.42	**4.23**
0.20	6.97	**6.69**	5.05	**4.83**
0.25	7.93	**7.53**	5.75	**5.45**
0.30	8.91	**8.39**	6.48	**6.10**

Table III shows that, on average, the proposed VAR-IMMA has better accuracy in all proportions. The accuracy of the mean RMSE and MAE values presented in Table III is supported by the consideration of value ranges and outliers in the simulated data, as depicted in Fig. 3. The lower RMSE mean values also indicate that VAR-IMMA can deal with outliers better than the original VAR-IM method. More detailed RMSE and MAE values are shown in Fig. 4.

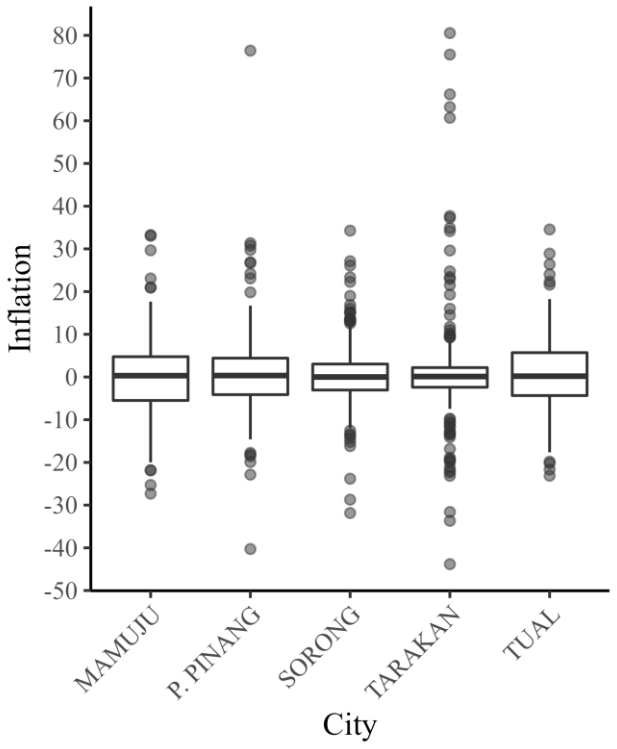

Fig. 3. Illustration of data distribution of five citiers with the highest variance out of the 82 simulated cities

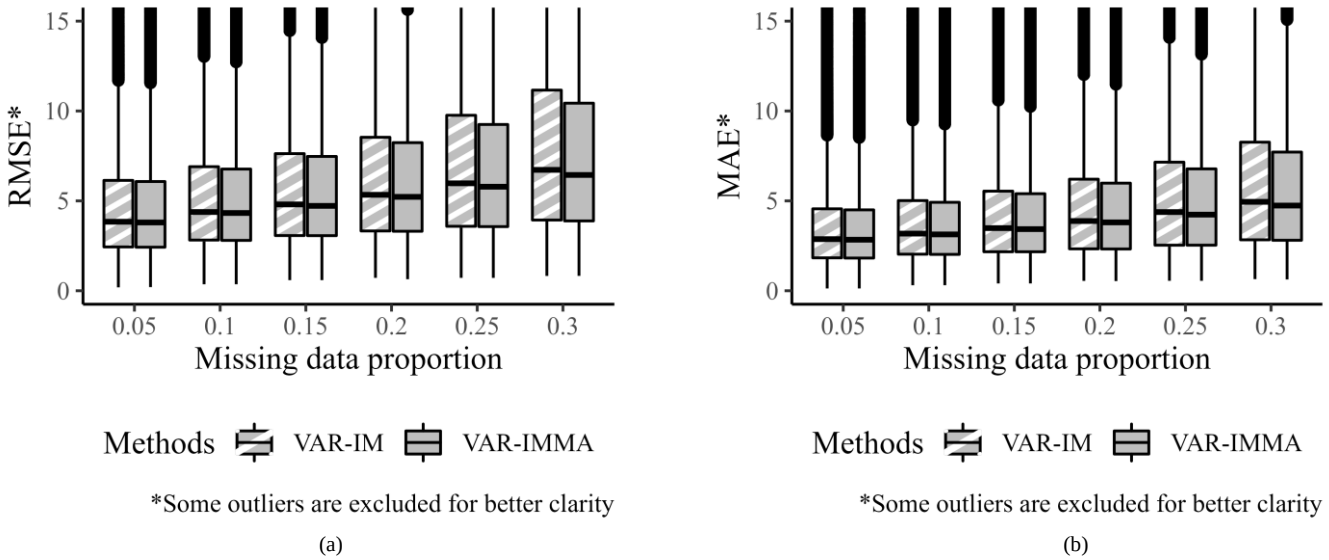

Fig. 4 The accuracy of VAR-IM's and VAR-IMMA's imputations of 100 simulations based on (a) RMSE and (b) MAE

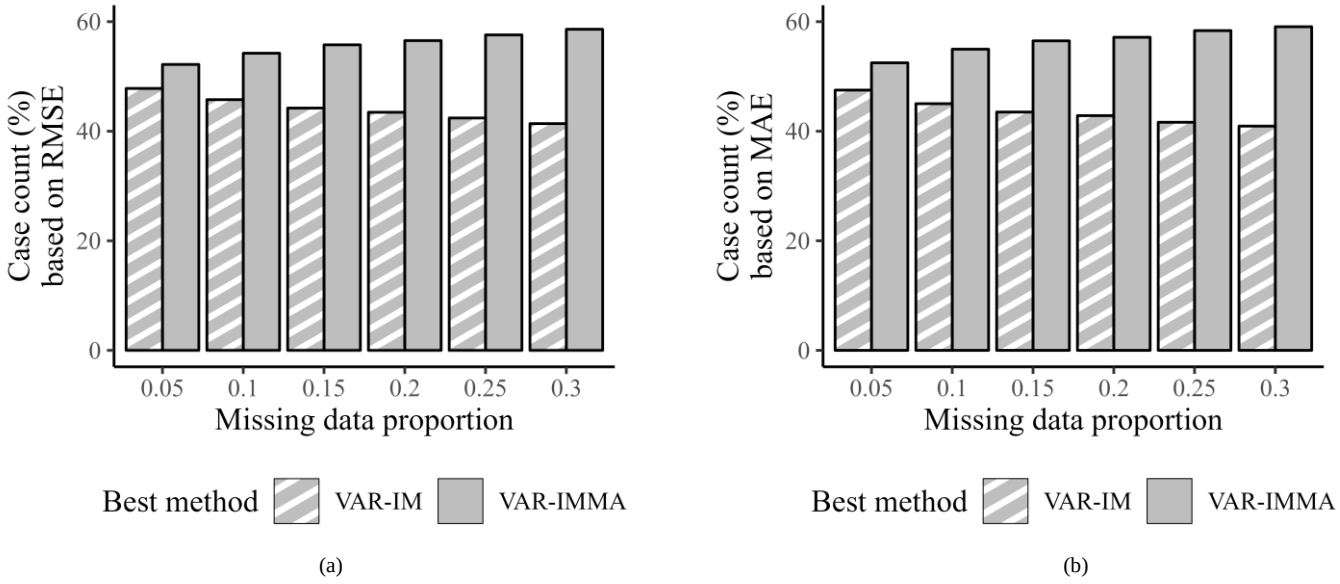

Fig. 5. Frequency of each method as the optimal method based on (a) RMSE and (b) MAE

Fig. 4 shows that RMSE and MAE are more varied as the missing proportion increases. Not only does VAR-IMMA has a lower median on RMSE and MAE distribution, but VAR-IMMA also has less variance in RMSE and MAE distribution than VAR-IM. This result shows that VAR-IMMA has better imputation results in terms of accuracy, either based on MAE or RMSE. Moreover, VAR-IMMA's accuracies are constantly better despite the increase in missing data proportion, indicating more stable imputation results than VAR-IM's.

Fig. 4 only shows the overall comparison between the two methods. Such comparison without further explanation may be misleading since both methods are not compared case by case. It is possible that VAR-IMMA has lower RMSE and MAE distributions but has much lower RMSE and MAE values in some cases, which is not an ideal condition for an imputation method.

Fig. 5 and Fig. 6 complement the evaluation of Fig. 4 by comparing VAR-IM and VAR-IMMA case by case. Fig. 5 presents a graphical representation of the number of times that each method achieves the best performance, as determined by the lowest RMSE or MAE. On the other hand, Fig. 6 illustrates the improvement in accuracy if VAR-IMMA is used instead of the original VAR-IM method. Each boxplot in Fig. 6 represents 100 simulations, excluding some outliers for better clarity.

Fig. 5 indicates that the more significant the missing proportion, the more frequent VAR-IMMA becomes the best method. Roughly 60% of cases have better imputation accuracy if imputed using VAR-IMMA.

While the proportion may not sound that much, the comparison in Fig. 6 shows that the boxplots are positive-skewed, indicating that the proposed VAR-IMMA can increase the imputation accuracy up to around 15% of the original accuracy, especially in data with larger missing proportions. Fig. 6 also shows some cases where the improvements are negative, meaning that VAR-IM is more accurate, but the negative improvements are around 0%.

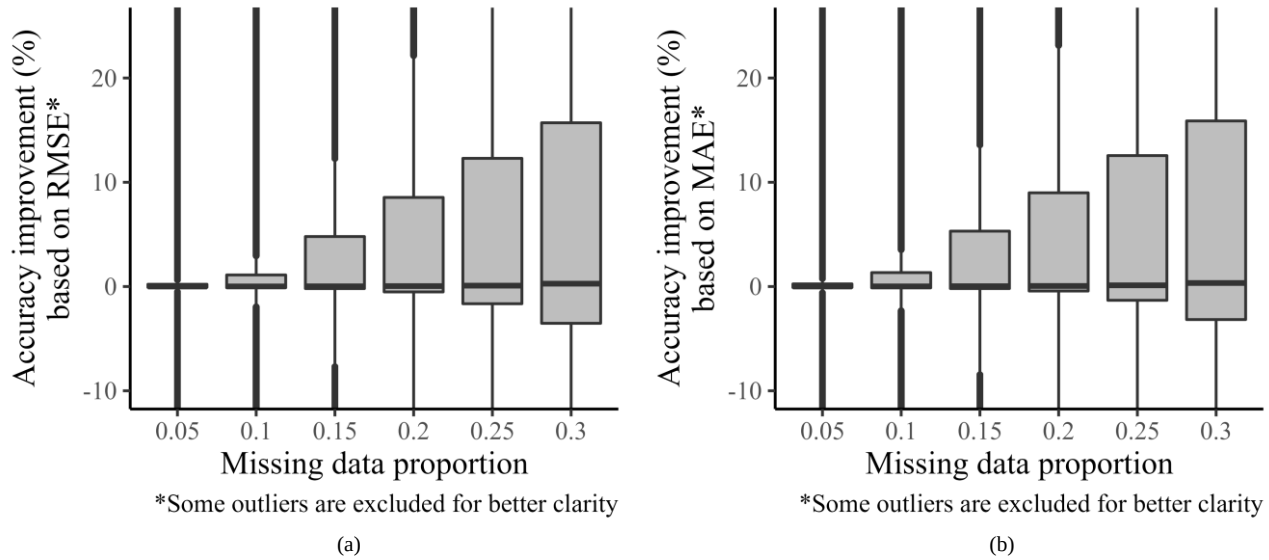

Fig. 6. Improvement (%) of impuation accuracy in each 100 simulations cases based on (a) RMSE and (b) MAE

These simulation results lead to the conclusion that VAR-IMMA outperforms the original VAR-IM method. In the worst cases, VAR-IMMA accuracy can be similar to the accuracy of VAR-IM. However, VAR-IMMA can reduce the RMSE and MAE to around 15% in at least 60% of cases. It can also be concluded that the more significant the missing proportion, the better the accuracy of VAR-IMMA compared to VAR-IM.

C. Application of the best method

As a study case, VAR-IMMA will now be used for a simulation. A random city is chosen among the 82 cities. The simulation is similar to the previous simulation, consisting of three stages. Fig. 7 compares the imputed data and the actual data.

Fig. 7 illustrates the imputation results of VAR-IMMA, showing that the imputation data often overlap with the actual (missing) data. The imputation results of missing data on each variable seem to have approached the omitted data. Overall, the imputation data are relatively close to the actual data. This relatively close estimation is also indicated by the RMSE and MAE values shown in Table IV.

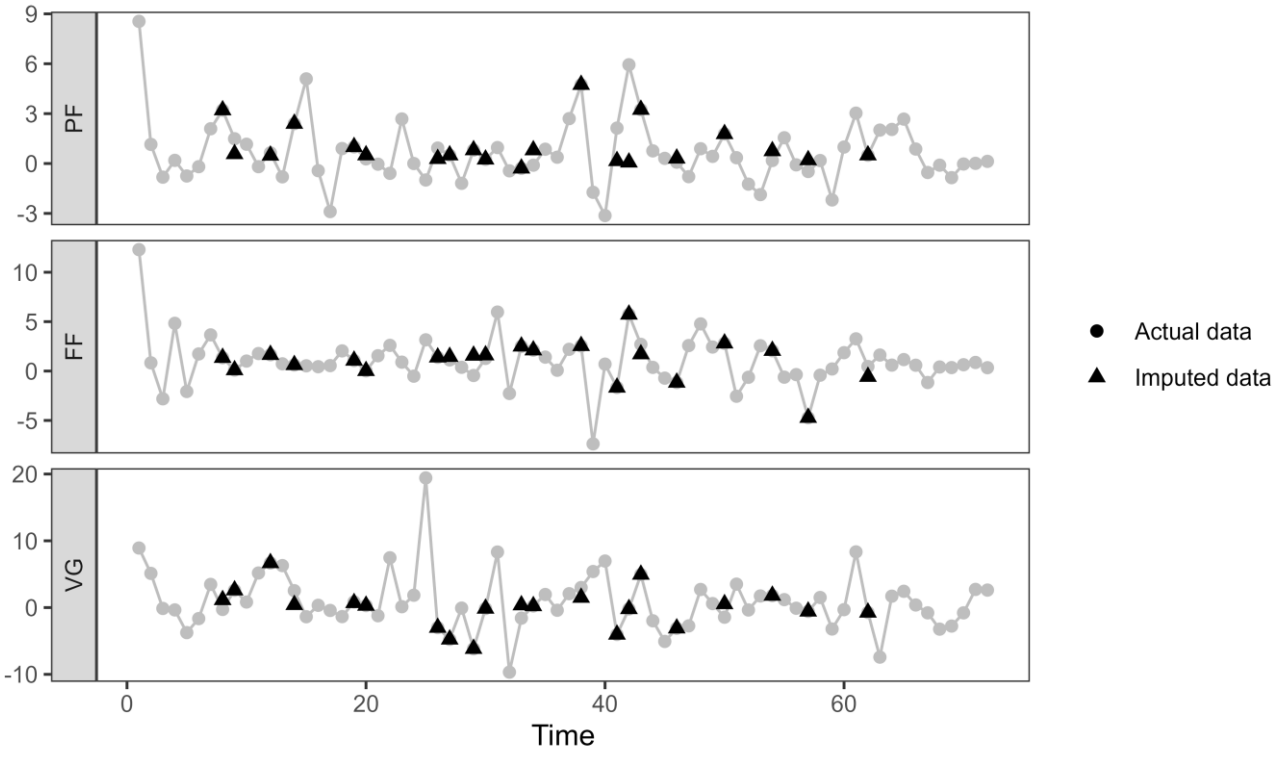

Fig. 7. The comparison between the imputed values and the actual (missing) values of Watampone City.

TABLE IV
IMPUTATION ACCURACY OF THE STUDY-CASE SIMULATION

Method	RMSE	MAE
VAR-IMMA	1.75	1.24

Considering the range of the data, RMSE of 1.75 and MAE of 1.24 are small. These small values show that the imputation results are accurate. Not only are the imputation results accurate, but the imputation results are also stable; that is, the imputation results are consistently close to the actual values. Had the imputation results been inconsistent, that is, the imputation results have small errors only in some data and large errors in others, the RMSE and MAE would have risen sharply.

The convergent multivariate time series model used for VAR-IMMA imputation is a VAR(1) model, mathematically represented as follows.

$$
\begin{aligned}
y_{1,t} &= 0.29 + 0.01y_{1,t-1} + 0.22y_{2,t-1} - 0.03y_{3,t-1} \\
y_{2,t} &= 1.08 - 0.37y_{1,t-1} + 0.11y_{2,t-1} - 0.11y_{3,t-1} \\
y_{3,t} &= 0.60 + 0.39y_{1,t-1} - 0.10y_{2,t-1} + 0.03y_{3,t-1}
\end{aligned}
$$

The final stage is the forecasting stage, where the convergent model above is used to forecast the inflation in the chosen city, Watampone. The forecasting results for the next six months are shown in Table V.

TABLE V
INFLATION FORECASTING OF THE STUDY-CASE SIMULATION

Month	PF	FF	VG
1	0.27	0.79	0.70
2	0.44	0.99	0.65
3	0.48	0.96	0.69
4	0.48	0.93	0.71
5	0.47	0.93	0.71
6	0.47	0.93	0.71

D. Conclusion

The evaluation of the simulation results leads to the conclusion that VAR-IMMA outperforms the original VAR-IM method. The value of RMSE and MAE of the two methods increases as the missing proportion increases. However, the RMSE and MAE of VAR-IMMA increase at a much lower rate than VAR-IM, resulting in more efficiency as the proportion increases. The efficiency refers to how much RMSE and MAE can be reduced after using VAR-IMMA. It means moving average as the initial imputation can follow data fluctuations better than the linear approach. The same results can be observed when VAR-IMMA is applied to study-case data, in this case, to a randomly chosen city, Watampone. The imputation results are not only well performed in terms of the small RMSE and MAE values but also in terms of consistency, where the imputation data are consistently close to the actual data.

REFERENCES

[1] H. Kang, "The prevention and handling of the missing data," *Korean J. Anesthesiol.*, vol. 64, no. 5, pp. 402–406, 2013.

[2] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu. (2018, April). Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Sci. Rep.* (Online). 8(6085). Available: https://www.nature.com/articles/s41598-018-24271-9

[3] S. Liu and P. C. M. Molenaar, "iVAR: A program for imputing missing data in multivariate time series using vector autoregressive models," *Behav. Res. Methods*, vol. 46, no. 4, pp. 1138–1148, 2014.

[4] A. Salarpour and H. Khotanlou, "An empirical comparison of distance measures for multivariate time series clustering," *Int. J. Eng. Trans. B Appl.*, vol. 31, no. 2, pp. 250–262, 2018.

[5] D. A. Newman, "Missing Data: Five Practical Guidelines," *Organ. Res. Methods*, vol. 17, no. 4, pp. 372–411, 2014.

[6] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*. Hoboken, NJ: J Wiley, 2020, pp. 3-19.

[7] H. Demirhan and Z. Renwick, "Missing value imputation for short to mid-term horizontal solar irradiance data," *Appl. Energy*, vol. 225, pp. 998-1012, 2018.

[8] F. Bashir and H. L. Wei, "Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm," *Neurocomputing*, vol. 276, pp. 23–30, 2018.

[9] W. Enders, *Applied Econometric Time series*, Hoboken, NJ: J Wiley, 2014, pp. 206–290.

[10] J. Van Greunen, A. Heymans, C. Van Heerden, and G. Van Vuuren, "The prominence of stationarity in time series forecasting," *J. Stud. Econ. Econom.*, vol. 38, no. 1, pp. 1–16, 2014.

[11] Z. Hossain, A. Rahman, M. Hossain, and J. H. Karami, "Over-Differencing and Forecasting with Non-Stationary Time Series Data," *Dhaka Univ. J. Sci.*, vol. 67, no. 1, pp. 21–26, 2019.

[12] P. A. V. B. Swamy and P. von zur Muehlen. (2020, September). Cointegration: Its fatal flaw and a proposed solution. *Sustain. Futur.* (Online). 2(100038). Available: https://www.sciencedirect.com/science/article/pii/S2666188820300319

[13] E. Zivot and J. Wang, *Modelling Financial Time Series with S-PLUS*. New York City, NY: Springer, 2006, pp. 385-390.

[14] K. Bulteel, F. Tuerlinckx, A. Brose, and E. Ceulemans. (2016, October). Clustering vector autoregressive models: Capturing qualitative differences in within-person dynamics. *Front. Psychol.* (Online), 7(1540). Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01540/full

[15] H. Mouriño and M. I. Barão. (2013, May). Maximum likelihood estimation of the VAR(1) model parameters with missing observations. *Math. Probl. Eng.* (Online). 2013(1). Available: https://www.hindawi.com/journals/mpe/2013/848120/

[16] H. Lütkepohl, "Estimation of structural vector autoregressive models," *Commun. Stat. Appl. Methods*, vol. 24, no. 5, pp. 421–441, 2017.

[17] Y. Nalita, R. Rahani, E. R. Tirayo, T. Toharudin, and B. N. Ruchjana, "Ordinary least square and maximum likelihood estimation of VAR(1) model's parameters and it's application on covid-19 in China 2020," in *Journal of Physics: Conference Series*, Sanur-Bali, 2021, pp. 1–9.

[18] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014.

[19] U. Khair, H. Fahmi, S. Al Hakim, and R. Rahim, "Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error," in *Journal of Physics: Conference Series*, Medan, 2017, pp. 1–7.

[20] I. Adwan et al., "Predicting Asphalt Pavement Temperature by Using Neural Network and Multiple Linear Regression Approach in the Eastern Mediterranean Region," *J. Eng. Sci. Technol.*, vol. 17, no. 1, pp. 15–32, 2022.

[21] E. G. Elizabeth, C. S. Nicole, K. S. Damian, and F. S. Anne. (2020, August). Using Vector Autoregression Modeling to Reveal Bidirectional Relationships in Gender/Sex-Related Interactions in Mother–Infant Dyads. *Front. Psychol.* (Online). 11(1507). Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01507/full

[22] R. M. Ueda, A. M. Souza, and R. M. C. P. Menezes. (2020, November). How macroeconomic variables affect admission and dismissal in the Brazilian electro-electronic sector: A VAR-based model and cluster analysis. *Phys. A Stat. Mech. its Appl.* (Online). 557(124872). Available:

https://www.sciencedirect.com/science/article/pii/S0378437120304519

[23] A. Suharsono, A. Aziza, and W. Pramesti, "Comparison of vector autoregressive (VAR) and vector error correction models (VECM) for index of ASEAN stock price," in *AIP Conference Proceedings*, Malang, 2017, pp. 1–10.

[24] E. Rohaeti, I. M. Sumertajaya, A. H. Wigena, and K. Sadik, "The Prominence of Vector Autoregressive Model in Multivariate Time Series Forecasting Models With Stationary Problems," BAREKENG J. Ilmu Mat. dan Terap., vol. 16, no. 4, pp. 1313–1324, 2022.

[25] B. J. Washington and L. Seymour, "An adapted vector autoregressive expectation maximization imputation algorithm for climate data networks," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 12, no. 6, pp. 1–15, 2020.

[26] Y. Dong and C. Y. J. Peng. (2013, May). Principled missing data methods for researchers. *Springerplus* (Online). 2(222). Available: https://link.springer.com/article/10.1186/2193-1801-2-222

[27] M. T. Ismail and R. S. Al-Gounmeein, "Overview of Long Memory for Economic and Financial Time Series Dataset and Related Time Series Models: A Review Study," *IAENG Int. J. Appl. Math.*, vol. 52, no. 2, pp. 261–269, 2022.

[28] S. Moritz and T. Bartz-Beielstein, "imputeTS: Time Series Missing Value Imputation in R," *R Journal*, vol. 9, no. 1, pp. 207–218, 2017.