BAREKENG: Journal of Mathematics and Its Application
December 2022 Volume 16 Issue 4 Page 1373–1384
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss4pp1373-1384

TIME SERIES IMPUTATION USING VAR-IM (CASE STUDY: WEATHER DATA IN METEOROLOGICAL STATION OF CITEKO)

Muhammad E. Rizal [1*], Aji H. Wigena[2], Farit M. Afendi[3]

[1,2,3]Department of Statistics, FMIPA, IPB University
Jl. Meranti IPB University, Bogor, 16680, West Java, Indonesia

Corresponding author's e-mail: [1] muhedyrizal@apps.ipb.ac.id*

Abstract. *Univariate imputation methods are defined as imputation methods that only use the information of the target variable to estimate missing values. While univariate imputation methods are convenient and flexible since no other variable is required, multivariate imputation methods can potentially improve imputation accuracy given that the other variables are relevant to the target variable. Many multivariate imputation methods have been proposed, one of which is the Vector Autoregression Imputation Method (VAR-IM). This study aims to compare the imputation results of VAR-IM-based methods and univariate imputation methods on time series data, specifically on long lag seasonal data such as daily weather data. Three modified VAR-IM methods were studied using simulations with three steps: deletion, imputation, and evaluation. The deletion step was conducted using six different schemes with six missing proportions. The simulations were conducted on secondary daily weather data collected from the meteorological station of Citeko from January 1, 1991, to June 22, 2013. Nine weather variables were examined, that is the minimum, maximum, and average temperatures, average humidity, rainfall rate, duration of solar radiation, maximum and average wind speed, as well as wind direction at maximum speed. The simulation results show that the three modified VAR-IM methods can improve the accuracy in around 75% of cases. The simulation results also show that imputation results of VAR-IM-based methods tend to be more stable in accuracy as the missing proportion increase compared to the imputation results of univariate imputation methods.*

Keywords: *multivariate imputation, VAR-IM, weather data*

Article info:

Submitted: date, month, year *Accepted: date, month, year*

How to cite this article:
M. E. Rizal, A. H. Wigena and F. M. Afendi, "TIME SERIES IMPUTATION USING VAR-IM (CASE STUDY: WEATHER DATA IN METEOROLOGICAL STATION OF CITEKO)", *BAREKENG: J. Math. & App.*, vol. 16, iss. 4, pp. 1373-1384, Dec., 2022.

1. INTRODUCTION

Weather data have been collected since 1659 [1] and is one of the oldest data types still used today. Despite the old origin, weather data nowadays still have many quality problems [2], [3], one of which is missing values. Since weather data is time series data, missing values is a significant issue to address in weather data. Simple methods such as deletion are often avoided as the order of the data will be changed, which can introduce bias. Therefore, imputation, a process of filling missing values using estimation methods [4], becomes one of the most common methods of dealing with missing values in time series data.

Previous studies have proposed many imputation methods. Some simplest methods are mean, median, and mode imputation [5]. However, these methods can lead to bias, especially if the proportion of the missing values is significant [6]. Not to mention that these imputation methods are not explicitly designed for time series imputation, which also needs to consider the time factor. Some of the methods that are more suitable for time series data are interpolation [7], MA [7], Kalman Filters [7], [8], last observation carried forward (loft) [8], or model-based imputation methods such as ARIMA-based method [9]. Some imputation methods are specifically designed for seasonal data, such as seasonally decomposed missing value imputation (seadec) and seasonally split missing value imputation (seasplit) [10]. These methods can only be used to impute one variable at a time; thus, they will be referred to as univariate imputation methods hereafter. These imputation methods are flexible because they do not require any variable other than the target variable itself.

On the other hand, some proposed imputation methods can impute multiple variables at once. These methods will be referred to as multivariate imputation methods. These methods are ideally used when there is multiple relevant time series variable. In such cases, the imputation accuracy may be improved. One of the prevalent examples of multivariate imputation methods is variants of nearest neighbor-based methods [11]–[14]. The main advantage of nearest neighbor-based methods is the ability to deal with categorical data. Some variants of nearest neighbor-based methods can even deal with univariate time series data by introducing covariates using lags and leads [12] or by clustering the trend, seasonality, cyclical, and residual features of the target variable [13]. However, the imputation time tends to be longer the bigger the data are [11]. Neural network-based methods, such as RNN [15], BRITS [16], E^2GAN [17], or NAOMI [18], are also quite popular methods with high accuracy. However, neural network-based methods are complex and can overfit if used in a small dataset.

One of the alternative multivariate imputation methods with a simpler model and shorter imputation time is Vector Autoregression (VAR)-based imputation methods [19], [20]. Liu and Molenaar [19] developed a VAR-based imputation package in R, which they referred to as iVAR. They compared the imputation performance of iVAR with listwise deletion, sample mean imputation, and multiple imputation (MI) method. The simulation results showed that iVAR could produce better estimation than the other three methods. Bashir and Wei [20] also developed an algorithm which they referred to as VAR-IM and compared the imputation performance with listwise deletion, linear regression imputation, Multivariate Auto-Regressive State-Space (MARSS), and Expectation-Maximization algorithm (EM). The results also showed an improvement compared to the other methods. Both studies were also implemented in real-world data, that is, on Electrodermal Activity (EDA) data in [21] and Electrocardiogram (ECG) data in [20]. Neither data is seasonal data with long lags similar to daily weather data.

This study will be focused on daily weather data collected from the Indonesian Agency for Meteorological, Climatological, and Geophysics (*Badan Meteorologi, Klimatologi, dan Geofisika* or simply BMKG). This case study is chosen because of its different characteristics from real-world data used in [19] and [20], especially in its seasonality with long seasonal lags. The other reason is the limited literature on imputation methods that specifically study daily BMKG data. Some of such studies are [22] and [8]. Both of these are imputation studies on BMKG data. However, [22] and [8] use monthly data, which have different characteristics than long seasonal lagged data such as daily weather data. For instance, daily data fluctuate much more and is more prone to missing values. Some weather stations of BMKG have shown multiple cases of consecutive missing values on daily data. There are even some weather stations with consecutive missing values for months.

Based on this background, this study aimed to compare the imputation results of modified VAR-IM with imputation results of univariate imputation methods, specifically for long seasonal lagged data such as daily weather data. We found that the three modified VAR-IM can improve the imputation accuracy of the six missing data schemes simulated in this study. We hope this study can improve the data quality of daily weather data so that other researchers can improve the quality of their studies.

2. RESEARCH METHODS

2.1 Data

This study uses daily weather data collected from *Stasiun Meteorologi* (Meteorological Station) Citeko. The data were collected from January 1, 1991, to June 22, 2013. The location and time range were chosen because of their completeness since data must be complete to be compared to the later imputation results. The raw data have nine variables. Three of them are temperature-related variables: minimum temperature (*Tn*), maximum temperature (*Tx*), and average temperature (*Tavg*). All of these temperature-related variables are in Celcius. Three other variables are wind-related: maximum wind speed (*ff_x*) in m/s, average wind speed (*ff_avg*) in m/s, and wind direction at maximum speed (*ddd_x*) in degrees. The others are average humidity (*RH_avg*) in percent, rainfall rate (*RR*) in mm, and duration of solar radiation (*ss*) in hour.

2.2 Research Procedure

The study started with data exploration. The goal is to ensure the completeness and study characteristics of the data, such as trend and seasonality. The study was then continued to the pre-simulation stage, a stage to determine variable combinations that will be simulated. Since there are nine variables, there are 255 combinations that can be made. Therefore, to simplify the simulation process, we focused on combinations of rainfall rate (*RR*) and variables significantly affecting it. We specifically choose combinations that involve rainfall rate (*RR*) because of its frequent use in literature. Moreover, rainfall rate (*RR*) is one of the variables that frequently have missing values among the nine variables in BMKG.

The optimal combination was determined using Forecast Error Variance Decomposition (FEVD). FEVD itself was calculated using VAR model that we fit in the ideal condition, that is when the data are complete. FEVD is defined in [23] as Equation (1)

$$\omega_{jk,h} = \frac{\sum_{i=0}^{h-1}\left(e_j'\Theta_i e_k\right)^2}{\sum_{i=0}^{h-1}\sum_{k=1}^{K}\theta_{jk,i}^2} \tag{1}$$

e_j in Equation ((1)

$$\omega_{jk,h} = \frac{\sum_{i=0}^{h-1}\left(e_j'\Theta_i e_k\right)^2}{\sum_{i=0}^{h-1}\sum_{k=1}^{K}\theta_{jk,i}^2}$$

(1) is the j-th column of I_K, an identity matrix of size $K \times K$. $\Theta_i = \Phi_i P$, where P is a lower triangular matrix obtained from decomposition of residual covariance matrix Σ_u using Cholesky decomposition so that $\Sigma_u = PP'$. $\Phi_i = JA_iJ'$, where $J = [I_K \quad 0 \quad \cdots \quad 0]$ of size $K \times Kp$.

There are three stages in the simulation process: randomized deletion stage, imputation stage, and evaluation stage. These three stages are repeated 1000 times.

1. Randomized deletion stage

   In this stage, the data will be deleted randomly with various proportions. The proportions are 7, 30, 90, 180, 270, and 365 days. There are six deletion schemes:

   a. Scheme 1: Consecutive missing values on a variable (*RR*)
   b. Scheme 2: Consecutive missing values on all variables (simultaneous)
   c. Scheme 3: Consecutive missing values on all variables (not simultaneous)
   d. Scheme 4: Random missing values on a variable (*RR*)
   e. Scheme 5: Random missing values on all variables (simultaneous)
   f. Scheme 6: Random missing values on all variables (not simultaneous)

   "Consecutive" means that the selected variable(s) will have consecutive missing values, but the first missing value will still be randomly chosen. On the other hand, "random" means that all missing values will be randomly chosen, thus, will likely not be consecutive. "Simultaneous" means that all variables will have missing values on the same observations, while "not simultaneous" means that the missing values for each variable will be chosen separately.

2. Imputation stage

Imputation was done using VAR-IM algorithm proposed by Bashir and Wei in [20]. The forecasting phase, however, will follow a rolling forecasting process as proposed by Liu and Molenaar in [19]. Rolling forecasting process was chosen because of the nature of missing values in BMKG data, especially for rainfall rate, which often has a long gap of missing values. Three modified VAR-IM methods will be used, that are, the original VAR-IM (VAR), differenced VAR-IM (VARD), and decomposed VAR-IM (VAR Dec). Seasplit, seadec, Moving Average (MA), and linear interpolation are four univariate imputation methods used as comparison methods.

The three modified versions of VAR-IM are defined as the following algorithm:

i. Initial imputation, since VAR model cannot be estimated on data with missing values;

ii. Data preprocessing, using either differencing or decomposition;

iii. Determining VAR order. VAR is defined in [4] as Equation (2)

$$\boldsymbol{Y_t} = \boldsymbol{\mu} + \sum_{i=1}^{p} \boldsymbol{A_i Y_{t-i}} + \boldsymbol{\varepsilon_t} \quad (2)$$

$\boldsymbol{Y_t}$ is a vector of variables at time t, $\boldsymbol{\mu}$ is an intercept vector, A_i is a parameter matrix of size $K \times K$ for order $i = 1,2, \dots, p$, and $\boldsymbol{\varepsilon}_t$ is a residual vector at time t;

iv. Expectation step: impute the missing values;

v. Maximization step: reestimate VAR model using the newest set of data;

vi. If convergence is met, continue to the next step. If convergence is not met, repeat steps ii to iv. Convergence is defined as Equation (3)

$$\boldsymbol{\beta_{r+1}} - \boldsymbol{\beta_r} < \boldsymbol{\zeta} \quad (3)$$

$\boldsymbol{\beta}$ is a coefficient vector, while r is the number of iterations. $\boldsymbol{\zeta}$ is a predefined threshold;

vii. Rolling forecasting for the missing values.

3. Evaluation stage

Imputation results of the three VAR-IM-based methods and four univariate methods were evaluated using Root Mean Square Error of Prediction (RMSEP). RMSEP is defined as Equation (4)

$$\boldsymbol{RMSEP} = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n}} \quad (4)$$

$\hat{y}_t$ are the imputed values while y_t is the actual values at time t.

3. RESULTS AND DISCUSSION

3.1. Data Exploration

One of the essential characteristics that data for this simulation must have is completeness since imputation results need a comparison to measure the accuracy. Figure 1 shows the time series plots of all variables. If there were any missing values, the date of the missing values would have been marked in the plots. Figure 1 shows that there is no missing value in the data.

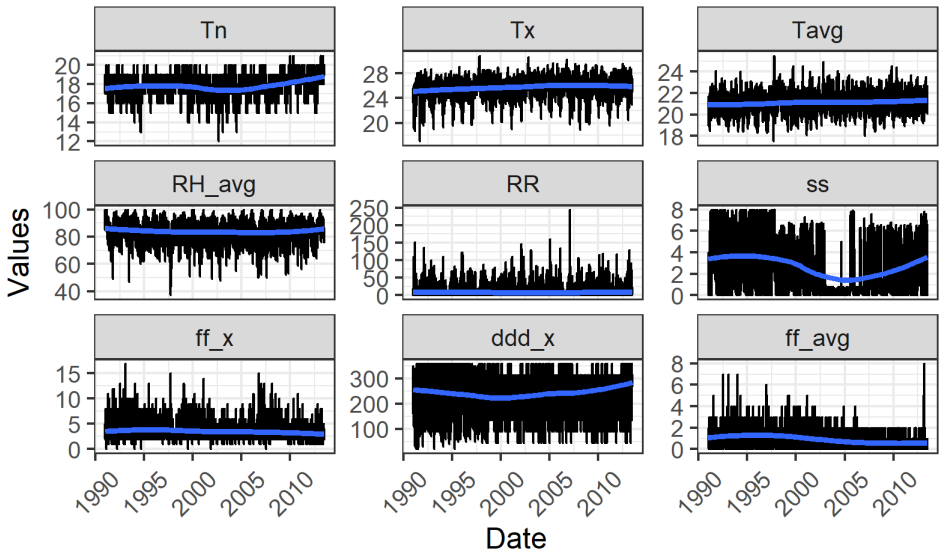

Figure 1. Time Series Plot of all Variables

Figure 1 also shows no trend in all variables, except for duration of solar radiation (ss) with some anomalies around 2004 and 2005. There also seems to be a slight seasonality, except for wind-related variables. This seasonality can also be observed in Figure 2.

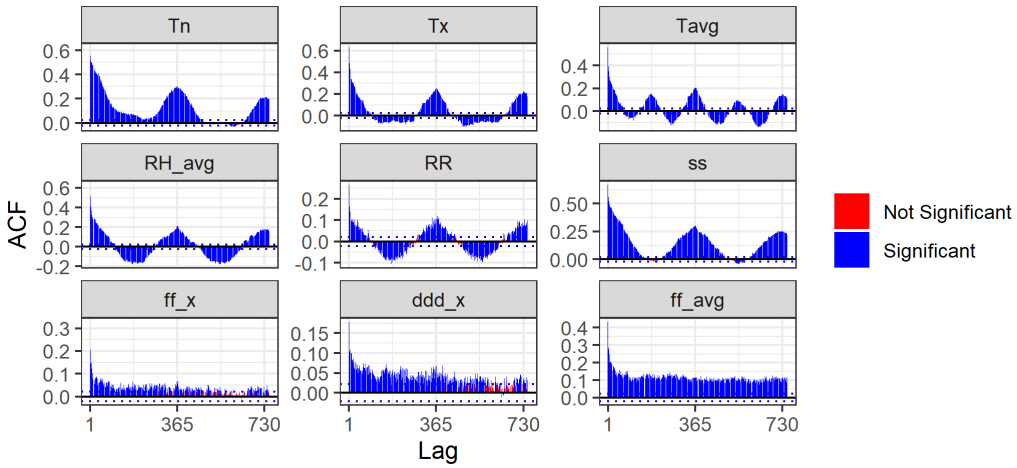

Figure 2. ACF Plot of all Variables

Various approaches can be used to deal with seasonality in data, one of which is by using differencing. Figure 3 shows the ACF plot of differenced data (d = 1) on the non-seasonal component. Figure 3 shows that all variables are cut off on some early lags. Hence there is no seasonality in the differenced data.

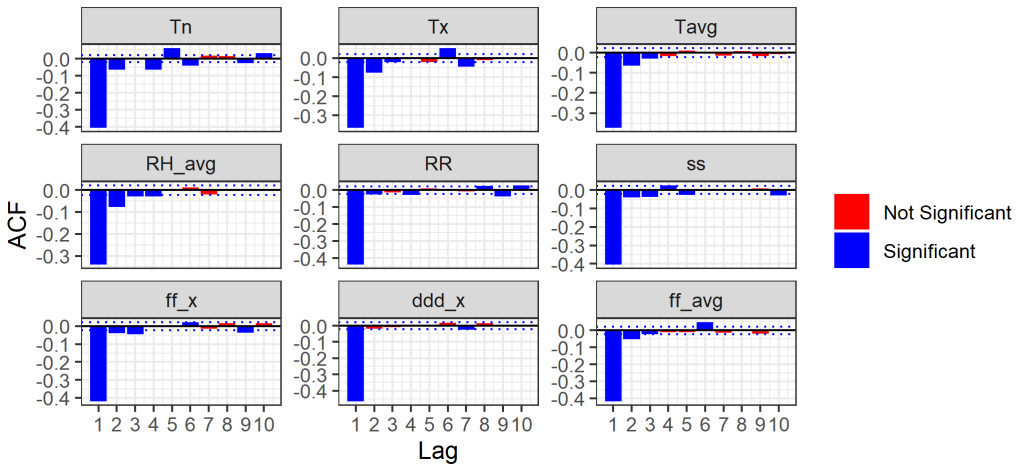

Figure 3. ACF Plot of Differenced Data (d = 1) on Non-Seasonal Component

Another approach that can be used to deal with seasonality is by using decomposition. The decomposition method used in this study was Seasonal and Trend decomposition using Loess (STL). Figure 4 shows the ACF plot of decomposed data using STL decomposition method. Unlike ACF of differenced data shown in Figure 3, ACF plots of decomposed data are slowly decaying. However, the seasonality patterns in Figure 4 are not as visible as those in the original data (Figure 2).

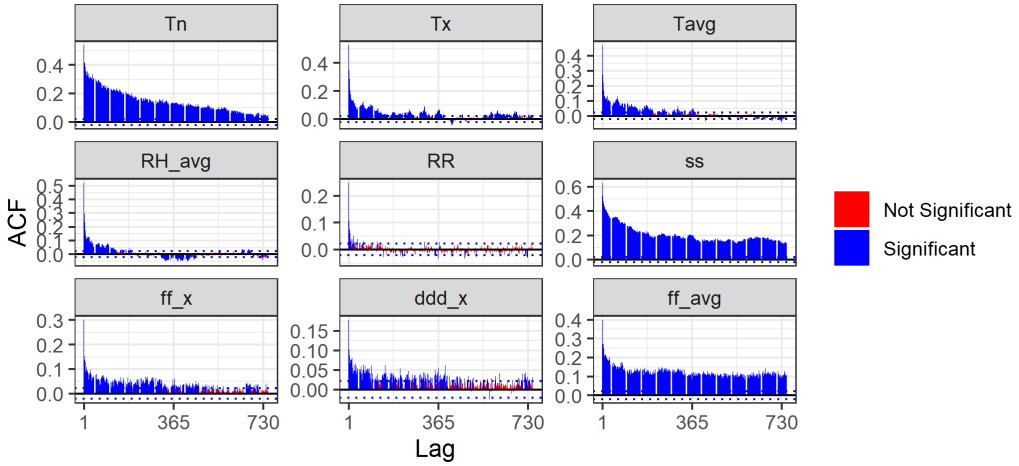

Figure 4. ACF Plot of Decomposed Data Using STL

Based on these ACF plots, three variants of VAR-IM methods will be simulated: the original VAR-IM, differenced VAR-IM, and decomposed VAR-IM. The best variable combinations for each variant were determined using FEVD. Figure 5 shows the FEVD plot. Figure 5 shows that rainfall rate (*RR*), average humidity (*RH_avg*), average temperature (*Tavg*), maximum temperature (*Tx*), and minimum temperature (*Tn*) are five variables that dominantly contribute to rainfall rate (*RR*) variation in the original and decomposed data. On the other hand, only rainfall rate (*RR*), average temperature (*Tavg*), and maximum temperature (*Tx*) dominantly contribute to rainfall rate (*RR*) variation in the differenced data.

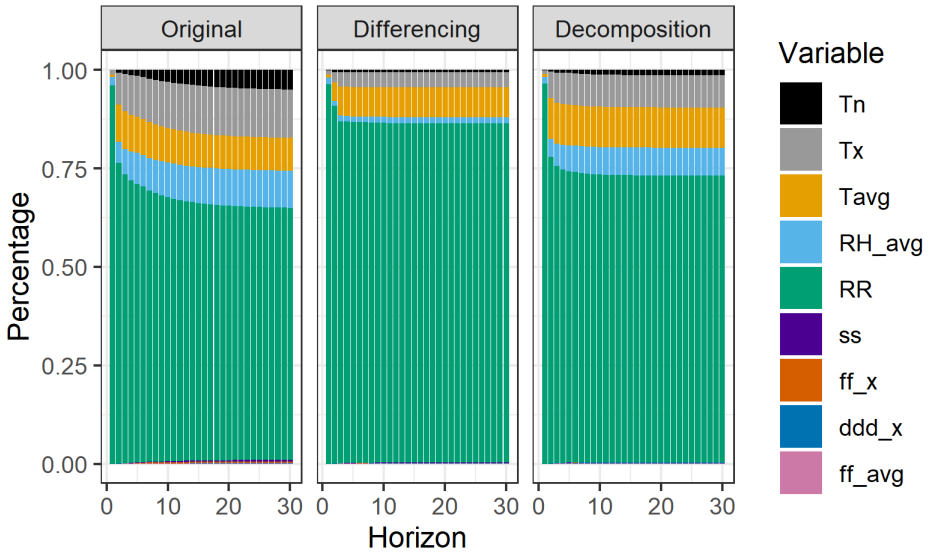

Figure 5. FEVD Plot of Rainfall Rate (*RR*)

3.2. Simulation

As stated in the research procedure, there are three stages for the simulation: deletion stage, imputation stage, and evaluation stage. These are repeated 1000 times with various missing proportions and schemes. Every imputation method had the exact incomplete data for each repetition, so the imputation results can be compared and evaluated.

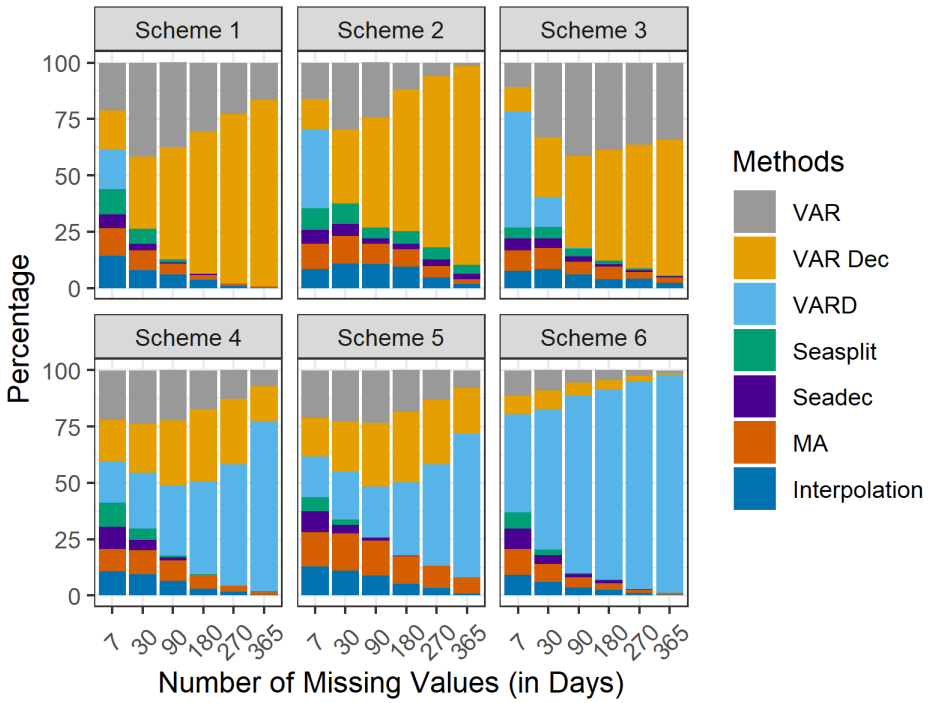

Figure 6. The Number of Times each Imputation Method had the Lowest RMSEP Shown in Percentage

The evaluation stage was started by choosing the best imputation method for each repetition, that is, the imputation method with the lowest RMSEP. **Error! Reference source not found.** shows the number of times each imputation method became the best. The number of times is shown in percentage. As shown in **Figure 6**, on data with 7 missing days, the proportions of each method are pretty balanced, except for schemes 3 and 6, where differenced VAR-IM becomes the best method for more than 50% of the cases. Since schemes 3 and 6 correspond to the "not simultaneous" type of schemes, it shows that differenced VAR-IM can have more accuracy if all variables are not simultaneously missing.

Figure 6 also shows that, as the missing proportion increase in Schemes 4, 5, and 6, the frequencies of differenced VAR-IM becoming the best method also increase. On the contrary, the frequencies decrease for Schemes 1, 2, and 3. As Schemes 1, 2, and 3 are "consecutive" types of schemes, and Schemes 4, 5, and 6 are "random" types of schemes, this shows that differenced VAR-IM has a high accuracy for data with random missing values, but worse accuracy for data with consecutive missing values. It also shows that the accuracy of differenced VAR-IM for data with random missing values tends to be better than the other methods. This tendency is especially true for Scheme 6 (non-simultaneous random missing values on all variables), where differenced VAR-IM becomes the best method in almost all cases.

On the other hand, the original VAR-IM and decomposed VAR-IM also shows good accuracies, especially for scheme 1, 2, and 3. In these schemes, decomposed VAR-IM shows a similar tendency as the one in differenced VAR-IM; decomposed VAR-IM tends to outperform the other methods as the missing proportion increases. Decomposed VAR-IM even outperformed the original VAR-IM in these schemes, which is expected as no seasonality treatment has been done for the original VAR-IM. Combining all these three variants of VAR-IM, it can be concluded that variants of VAR-IM outperform the comparative univariate methods, including seasplit and seadec, which are designed explicitly for seasonal data.

Table 1. Mean RMSEP for all Imputation Methods on Scheme 1

Missing values (days)	VAR	VAR Dec	VARD	Seasplit	Seadec	MA	Interp.
7	10.940	10.930	**10.495***	13.807	16.979	12.585	12.723
30	12.938	**12.750***	2.19E+149	15.568	19.344	14.994	15.028
90	13.955	**13.617***	9.00E+79	16.306	20.564	16.456	16.127
180	14.675	**14.210***	2.68E+70	16.933	21.010	17.166	16.818
270	14.847	**14.361***	2.92E+62	16.995	21.456	17.826	17.673
365	14.961	**14.481***	2.24E+51	17.023	21.325	18.418	18.471

Table 2. Mean RMSEP for all Imputation Methods on Scheme 2

Missing values (days)	VAR	VAR Dec	VARD	Seasplit	Seadec	MA	Interp.
7	6.355	**6.312***	7.073	8.423	9.792	7.576	7.621
30	7.133	**7.000***	2.25E+149	8.971	10.807	8.680	8.666
90	7.640	**7.395***	2.98E+149	9.287	11.433	9.422	9.223
180	8.023	**7.656***	1.64E+143	9.507	11.528	9.799	9.548
270	8.172	**7.716***	8.38E+138	9.500	11.754	10.183	10.053
365	8.244	**7.753***	1.19E+108	9.498	11.652	10.536	10.491

Table 3. Mean RMSEP for all Imputation Methods on Scheme 3

Missing values (days)	VAR	VAR Dec	VARD	Seasplit	Seadec	MA	Interp.
7	6.902	**6.852***	7.021	9.359	10.063	7.937	8.329
30	8.463	**8.376***	1.20E+141	10.819	12.799	10.288	10.326
90	9.198	**9.040***	2.25E+136	10.893	13.632	10.600	10.640
180	9.599	**9.382***	9.16E+108	11.657	15.261	11.729	11.604
270	9.814	**9.560***	2.60E+113	11.586	15.024	11.842	11.783
365	9.996	**9.733***	6.50E+118	11.766	15.539	12.071	12.089

Table 4. Mean RMSEP for all Imputation Methods on Scheme 4

Missing values (days)	VAR	VAR Dec	VARD	Seasplit	Seadec	MA	Interp.
7	12.272	**12.199***	38.404	15.219	16.839	12.736	13.566
30	14.022	**13.906***	3.51E+4	16.749	19.437	14.466	15.154
90	14.495	**14.317***	5.02E+10	17.224	20.683	14.890	15.515
180	14.692	**14.504***	5.73E+11	17.384	21.140	15.081	15.649
270	14.728	**14.532***	1.67E+13	17.376	21.112	15.114	15.694
365	14.708	**14.512***	6.95E+3	17.398	21.072	15.087	15.670

Table 5. Mean RMSEP for all imputation methods on scheme 5

Missing values (days)	VAR	VAR Dec	VARD	Seasplit	Seadec	MA	Interp.
7	6.264	**6.239***	23.141	8.911	9.120	7.066	7.464
30	6.832	**6.788***	7015.69	9.414	10.209	7.742	8.081
90	7.001	**6.931***	3.49E+9	9.601	10.767	7.917	8.227
180	7.083	**7.006***	3.98E+9	9.672	10.980	8.001	8.290
270	7.097	**7.016***	7.85E+6	9.663	10.952	8.010	8.307
365	7.096	**7.015***	2317.973	9.674	10.940	7.999	8.298

Table 6. Mean RMSEP for all Imputation Methods on Scheme 6.

Missing values (days)	VAR	VAR Dec	VARD	Seasplit	Seadec	MA	Interp.
7	4.824	**4.799***	6.579	7.605	7.049	5.750	6.071
30	6.098	**6.063***	50.979	8.702	8.774	7.039	7.429
90	6.623	**6.569***	21.722	9.364	9.933	7.636	8.040
180	6.889	**6.827***	9.937	9.495	10.278	7.793	8.137
270	6.969	6.898	**3.261***	9.563	10.526	7.880	8.205
365	7.007	6.936	**1.411***	9.594	10.618	7.905	8.214

Table 1 to Table 6 provide details of RMSEP for each method on each proportion and scheme. Bold and asterisk (*) values indicated the lowest mean RMSEP for a certain missing proportion. Another interesting result, as shown in these tables, is the large mean RMSEP that differenced VAR-IM generally achieved on all schemes, including Schemes 4, 5, and 6, where the majority of the best method is differenced

VAR-IM. These large mean RMSEP are caused by large RMSEP on some cases/repetitions, dragging the overall mean of RMSEP to such large values. These results are also illustrated in Figure 7.

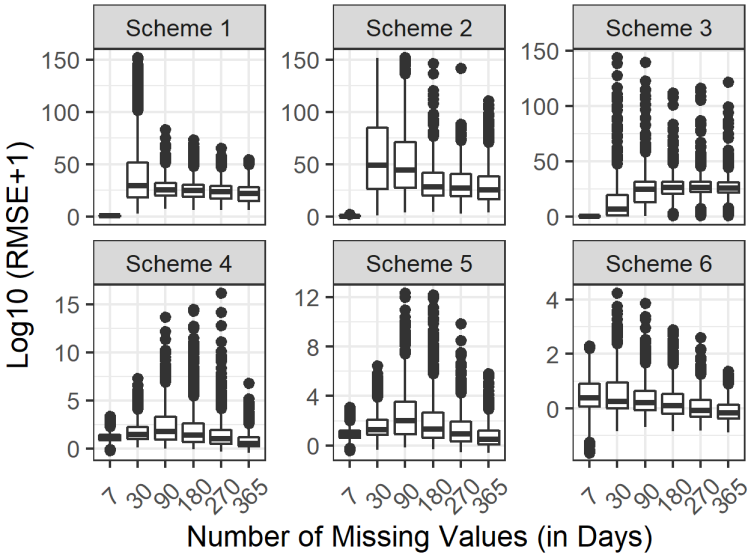

Figure 7. RMSEP Distribution of Differenced VAR-IM

Figure 7 shows the distribution of RMSEP values of 1000 repetitions for differenced VAR-IM. It can be seen clearly that the distributions are highly skewed. Although the RMSEP of schemes 4, 5, and 6 are not as highly skewed as the RMSEP of schemes 1, 2, 3, the RMSEP for schemes 4, 5, and 6 are still too high compared to the other methods, as Figure 7 are shown in $\log_{10}$.

The reason for the skewed RMSEP of differenced VAR-IM can be explained by Figure 8. Figure 8 shows the skewness of residual distribution for each repetition. Differenced VAR-IM is much more skewed than decomposed VAR-IM, especially for Schemes 1, 2, and 3. This skewness explains the large mean RMSEP that differenced VAR-IM achieved in Table 1 to Table 6.

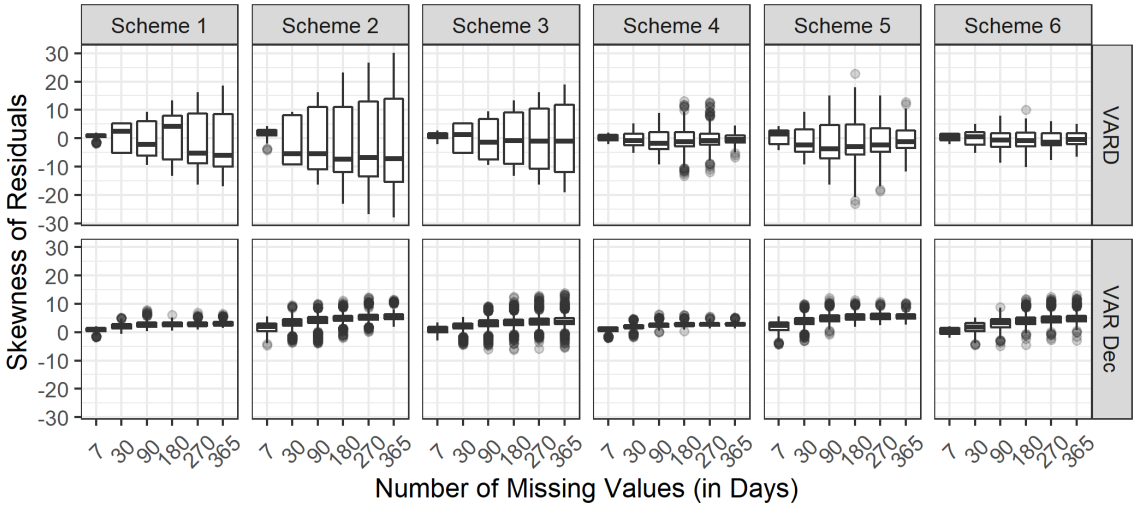

Figure 8. Skewness of Residuals for each Repetition

By comparing Figure 7 and Figure 9, it is clear that decomposed VAR-IM has more stable imputations than differenced VAR-IM. Moreover, it can be seen from Table 1 to Table 6 that decomposed VAR-IM has the best RMSEP on average compared to the other methods. Therefore, it can be concluded that decomposed VAR-IM is the safest imputation method in terms of imputation.

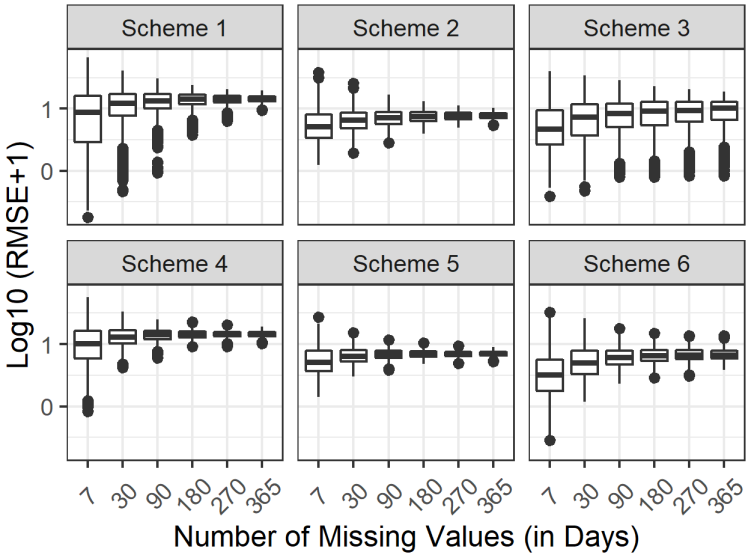

Figure 9. RMSEP distribution of decomposed VAR-IM

Figure 10 shows the comparison of RMSEP between decomposed VAR-IM and each univariate method. The comparison is shown in percentages where negative values indicate lower RMSEP for decomposed VAR-IM. In other words, a negative percentage means that decomposed VAR-IM can improve the imputation accuracies compared to the univariate methods. Figure 10 also shows that decomposed VAR-IM can improve the accuracies up to 20% in around 50% of cases and even more than 20% in another 25% of cases. This improvement is specifically true for data with "consecutive" type missing values, that is, data from Schemes 1, 2, and 3.

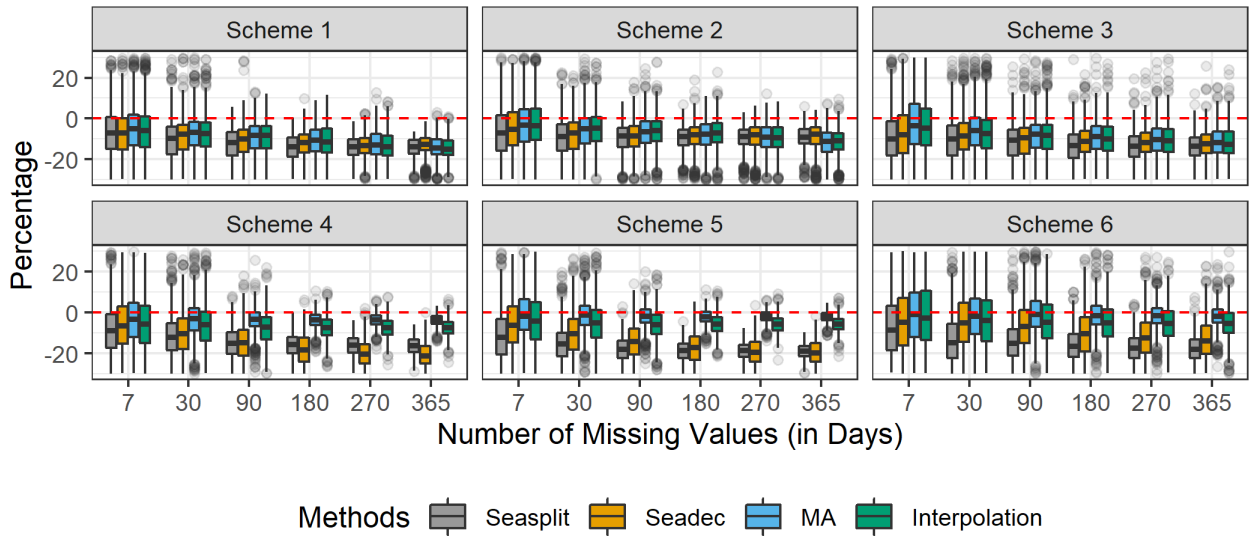

Figure 10. Comparison of RMSEP between Decomposed VAR-IM and each Univariate Imputation Method. Some of the Outliers have been Omitted for the Clarity of the Plot.

4. CONCLUSIONS

Based on the simulation results, it can be concluded that three variants of VAR-IM have better accuracies compared to the four univariate imputation methods. We strongly recommend using any of the three variants of VAR-IM for imputation, especially for long seasonal lagged data such as daily weather data with a significant missing proportion. For "random" type missing data, that is, data with non-consecutive missing values, we recommend using differenced VAR-IM since the methods have shown an outstanding

performance compared to the other methods. However, in terms of stability of imputation accuracies, we recommend using decomposed VAR-IM as the method has shown similar accuracies but far higher stabilities to those of differenced VAR-IM.

REFERENCES

[1] K. K. Tung and J. Zhou, "Using data to attribute episodes of warming and cooling in instrumental records," in *Proceedings of the National Academy of Sciences of the United States of America*, 2013, vol. 110, no. 6, pp. 2058–2063. doi: 10.1073/pnas.1212471110.

[2] G. Pastorello *et al.*, "Observational data patterns for time series data quality assessment," in *Proceedings - 2014 IEEE 10th International Conference on eScience, eScience 2014*, 2014, vol. 1. doi: 10.1109/eScience.2014.45.

[3] S. Hunziker *et al.*, "Identifying, attributing, and overcoming common data quality issues of manned station observations," *Int. J. Climatol.*, vol. 37, no. 11, pp. 4131–4145, 2017, doi: 10.1002/joc.5037.

[4] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction Time Series Analysis and Forecasting*, 2nd ed. John Wiley & Sons, Inc.: Hoboken, New Jersey, 2016.

[5] J. A. Saunders, N. Morrow-Howell, E. Spitznagel, P. Doré, E. K. Proctor, and R. Pescarino, "Imputing missing data: A comparison of methods for social work researchers," *Soc. Work Res.*, vol. 30, no. 1, 2006, doi: 10.1093/swr/30.1.19.

[6] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of Performance of Data Imputation Methods for Numeric Dataset," *Appl. Artif. Intell.*, vol. 33, no. 10, 2019, doi: 10.1080/08839514.2019.1637138.

[7] H. Demirhan and Z. Renwick, "Missing value imputation for short to mid-term horizontal solar irradiance data," *Appl. Energy*, vol. 225, 2018, doi: 10.1016/j.apenergy.2018.05.054.

[8] R. Y. P. Muflihah, "Perbandingan Teknik Interpolasi Berbasis R Dalam Estimasi Data Curah Hujan Bulanan Yang Hilang Di Sulawesi," *J. Meteorol. DAN Geofis.*, vol. 18, no. 3, pp. 107–111, 2017, [Online]. Available: https://puslitbang.bmkg.go.id/jmg/index.php/jmg/article/view/343

[9] Y. Li, Z. Li, and L. Li, "Missing traffic data: Comparison of imputation methods," *IET Intell. Transp. Syst.*, vol. 8, no. 1, 2014, doi: 10.1049/iet-its.2013.0052.

[10] S. Moritz and T. Bartz-Beielstein, "imputeTS: Time Series Missing Value Imputation in R," *R J.*, vol. 9, no. 1, pp. 207–218, 2017, doi: 10.32614/RJ-2017-009.

[11] G. E. A. P. A. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Appl. Artif. Intell.*, vol. 17, pp. 519–533, 2003, doi: 10.1080/713827181.

[12] A. Flores, H. Tito, and C. Silva, "Local Average of Nearest Neighbors: Univariate Time Series Imputation," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 8, 2019, Accessed: Jun. 15, 2022. [Online]. Available: www.ijacsa.thesai.org

[13] N. Savarimuthu and S. Karesiddaiah, "An unsupervised neural network approach for imputation of missing values in univariate time series data," *Concurr. Comput. Pract. Exp.*, vol. 33, no. 9, 2021, doi: 10.1002/cpe.6156.

[14] W. Y. Lai, K. K. Kuok, S. Gato-Trinidad, and K. X. L. Derrick, "A study on sequential K-nearest neighbor (SKNN) imputation for treating missing rainfall data," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 3, pp. 363–368, May 2019, doi: 10.30534/ijatcse/2019/05832019.

[15] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent Neural Networks for Multivariate Time Series with Missing Values," vol. 8, p. 6085, 2018, doi: 10.1038/s41598-018-24271-9.

[16] W. Cao, H. Zhou, D. Wang, Y. Li, J. Li, and L. Li, "BRITS: Bidirectional recurrent imputation for time series," *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, pp. 6775–6785, 2018, Accessed: Jul. 25, 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/734e6bfcd358e25ac1db0a4241b95651-Abstract.html

[17] Y. Luo, Y. Zhang, X. Cai, and X. Yuan, "E 2 GAN: End-to-End Generative Adversarial Network for Multivariate Time Series Imputation," 2019.

[18] Y. Liu, R. Yu, S. Zheng, E. Zhan, and Y. Yue, "NAOMI: Non-autoregressive multiresolution sequence imputation," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, Accessed: Jul. 25, 2021. [Online]. Available: https://github.com/felixykliu/NAOMI

[19] 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, 2014, doi: 10.3758/s13428-014-0444-4.

[20] 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, 2018, doi: 10.1016/j.neucom.2017.03.097.

[21] K. Penny, C. Yozgat, C. İyigün, M. Türkeş, C. Yozgatligil, and S. Aslan, "Comparison of missing value imputation methods in time series: the case of Turkish meteorological data," *Theor. Appl. Climatol.*, vol. 112, pp. 143–167, 2013, doi: 10.1007/s00704-012-0723-x.

[22] D. D. A. Nofianto, A. Djuraidah, and A. Rizki, "Penerapan Algoritme Expectation- Maximization with Bootstrapping (EMB) untuk Pendugaan Data Hilang Curah Hujan Kabupaten Indramayu," IPB University, 2017. [Online]. Available: https://repository.ipb.ac.id/handle/123456789/88373

[23] H. Lütkepohl, *New introduction to multiple time series analysis*. 2005. doi: 10.1007/978-3-540-27752-1.