Pakistan Journal of Statistics and Operation Research

Evaluating Time Series Imputation Algorithms: Kalman Smoothing vs Seasonal Trend Decomposition Using LOESS on Climate Data

Cevira Dhiya Ayuni[1*], Gustriza Erda[2], Nidia Mindiyarti[3], Felia Rizky Aulia[4]

* Corresponding Author

1. Department of Statistics, Faculty of Mathematics and Natural Sciences, University of Riau, Indonesia, ceviradhiyaayuni16@gmail.com
2. Department of Statistics, Faculty of Mathematics and Natural Sciences, University of Riau, Indonesia, gustrizaerda@lecturer.unri.ac.id
3. The SMART Research Institute (SMARTRI), Indonesia, nidiamindiyarti@gmail.com
4. The SMART Research Institute (SMARTRI), Indonesia, feliarizky27@gmail.com

Abstract

Missing data refer to situations in which part of a dataset is unavailable due to unrecorded information during the data collection process. Such issues must be addressed because they may affect analytical accuracy. This study evaluates the performance of two univariate imputation methods: Kalman Smoothing and Seasonal Trend Decomposition using LOESS (STL Decomposition), in which LOESS stands for Locally Estimated Scatterplot Smoothing. Both methods were applied to time series data collected from a weather station in Lampung Province, Indonesia, covering the period 2001–2024. The analysis incorporates missing data mechanisms, namely MCAR (Missing Completely at Random), MAR (Missing at Random), and MNAR (Missing Not at Random), with missing rates of 5%, 10%, 20%, 30%, 40%, and 50%. The variables examined include average temperature, average relative humidity, total precipitation, average solar radiation, and average wind speed. The results indicate that STL Decomposition outperforms Kalman Smoothing in terms of precision, consistently yielding lower RMSE values, particularly at missing data levels of 5%, 10%, and 20%, and across all missing data mechanisms. Although Kalman Smoothing provides relatively greater stability in preserving temporal dependencies, STL Decomposition demonstrates superior accuracy in imputing missing values, as reflected in its consistently lower RMSE across various missing data scenarios.

Key Words: Missing Data, Imputation, Kalman Smoothing, STL Decomposition, Time Series, RMSE.

1. Introduction

In statistical analysis, data come in various forms and possess distinct characteristics depending on their intended use. One such type is time series data. Time series data consist of a set of quantitative observations arranged in chronological order and are used for various analyses, such as forecasting future outcomes (Kirchgassner & Wolters, 2015). The temporal sequence in time series data may include daily, monthly, or yearly records. One common challenge in time series data is the presence of missing values. The existence of missing data can reduce the accuracy and performance of time series forecasting (Ahn et al., 2021).

Missing data is a common issue that arises when portions of data are unavailable due to unrecorded or omitted values during the data collection or recording process (Austin et al., 2021). While missing data may not pose a significant issue when it comprises only a small portion (e.g., 1%) of the dataset, it can become a critical problem when the proportion of missing values is substantial (Oktavianti & Yanti, 2022). Missing data can lead to biased estimation results, which in turn may produce inaccurate conclusions (Agiwal & Chaudhuri, 2024). Therefore, it is essential to identify the most effective techniques or methods to handle missing data appropriately.

Handling missing data is a fundamental requirement in data management (Irnawan et al., 2021). One of the simplest approaches is to delete or remove the missing values (Wang & Aronow, 2022). However, this method is impractical when the extent of missing data is large, as it may result in the loss of important information, particularly in time series data, which is often used for forecasting (Ahn et al., 2021). Therefore, alternative approaches are needed, such as imputing missing values using appropriate imputation methods.

Imputation is a solution that involves replacing missing values with reasonable estimates (M. R. A. Prasetya et al., 2023). One approach is univariate imputation, which uses only the observed values from a single variable and focuses on its data pattern (Chhabra, 2023). Common univariate imputation methods include Mean, Median, Mode, Kalman Smoothing, and Seasonal Trend Decomposition using LOESS. In time series data, univariate imputation involves leveraging the entire historical series, including both past and future values, to estimate the missing points (Phan, 2020).

Previous studies have compared univariate and multivariate imputation approaches. Ankrah et al. (2019) compared univariate imputation methods in time series data, taking into account data characteristics such as trend and seasonality, as well as missing data mechanisms—namely Missing at Random (MAR) and Missing Completely at Random (MCAR). Their study utilized four real datasets with missing rates ranging from 10% to 90%. Performance was evaluated using Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The results showed that the Seasonal Trend Decomposition using LOESS (Locally Estimated Scatterplot Smoothing), hereafter referred to as STL Decomposition, was the most effective for data exhibiting both trend and seasonality, under both MAR and MCAR conditions. Interpolation was better suited for seasonal data without a trend, while mean and replace imputation methods were more effective for non-trend, non-seasonal data. Another study by Garcia et al. (2022) evaluated several imputation methods for handling missing data in air quality datasets. The methods included Linear Imputation, Predictive Mean Matching, Random Forest, K-Nearest Neighbors, and Kalman Smoothing. The study found that Kalman Smoothing outperformed all other methods, showing the lowest RMSE and MAE values.

The present study evaluates and compares the performance of Kalman Smoothing and STL Decomposition for imputing missing values in climate time series data from Lampung over the period 2001–2024. It also considers various missing data conditions (MCAR, MAR, MNAR) and different missing data proportions, ranging from 5%, 10%, 20%, 30%, 40%, to 50%. The evaluation metric used is the Root Mean Square Error (RMSE) to assess imputation accuracy. This study contributes theoretically by extending the literature on imputation methods in tropical climate data and provides practical guidance for meteorological agencies in selecting appropriate techniques for handling missing observations.

Climate time series data are characterized by long-term trends, recurring seasonal patterns, and strong temporal dependencies between observations. Conventional static imputation techniques applied to climate data with irregular fluctuations may alter or distort the intrinsic properties of the data, potentially affecting the outcomes of time series analyses. Therefore, this study employs Kalman Smoothing and Seasonal-Trend Decomposition using LOESS (STL Decomposition) as imputation approaches, since both methods explicitly account for the underlying temporal structure and seasonal behavior. By preserving the intrinsic dynamic properties of climate data, these techniques provide a more appropriate framework for handling missing observations in time series contexts.

Based on these objectives, the study raises two main research questions. First, how do the Kalman Smoothing and Seasonal Trend Decomposition using LOESS (STL Decomposition) methods perform in addressing the missing data problem in climate time series data in Lampung Province, Indonesia. Missing data frequently occur in climate records due to recording disruptions or technical issues, requiring appropriate imputation methods to improve data quality. Second, which method provides the most accurate imputation results, as measured by RMSE, across different missing data conditions and patterns? The answers to these questions are expected to serve as a basis for determining the most suitable method for handling missing data in climate time series analysis.

2. Data and Methodology

The climate dataset used in this study was obtained from a weather station operated by SMART Research Institute, Tbk, located at SBYE (Sungai Buaya) Estate, Lampung Province, Indonesia. The data were automatically recorded and stored in digital format on a daily basis from 2001 to 2024, comprising a total of 6,787 observations. Of these, 1,473 values were identified as missing, representing approximately 21.71% of the entire dataset. The analyzed variables consist of five weather parameters: average temperature, average relative humidity, total precipitation, average solar radiation, and average wind speed.

The analysis was conducted in several stages. First, data were collected and examined to identify the presence of missing values. Initial imputation was then performed using the Kalman Smoothing and STL Decomposition methods. The imputed datasets were subsequently modeled using the Vector Autoregressive (VAR) approach and evaluated based on the Root Mean Square Error (RMSE). The VAR model with the smallest RMSE was selected to generate simulated complete datasets. These datasets were then used to simulate missing data mechanisms under three scenarios: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR), with missingness proportions of 5%, 10%, 20%, 30%, 40%, and 50%. The imputation procedures were repeated for each missingness scenario and proportion. Finally, the RMSE values were calculated and compared to determine the best-performing method, and conclusions were drawn based on the results.

3. Kalman Smoothing

Kalman Smoothing is a model that operates based on a fundamental structure known as the state-space representation (Umar & Gray, 2023). The model originates from the Kalman Filter concept, which is designed to represent systems evolving over time through a set of equations. Unlike conventional imputation approaches, such as mean substitution or regression-based imputation, which ignore temporal dynamics, Kalman Smoothing enhances the estimates produced by the Kalman Filter by not only utilizing past information but also incorporating future data to generate more optimal predictions at missing data points (Chaudhry et al., 2019). This property makes it particularly suitable for climate data that exhibit serial correlation and gradual structural changes.

Kalman Smoothing offers greater flexibility in handling evolving processes and irregular missing patterns (Slipetz et al., 2023). The method is particularly effective under MCAR and MAR mechanisms, where missing values can be inferred from the observed temporal structure. However, its performance may be limited under MNAR conditions due to the presence of systematic unobserved or latent patterns underlying the missingness process.

Before performing missing value estimation using Kalman Smoothing, an initialization step is required to fill in the missing values. Providing initial values is a crucial first step to ensure the consistency and effectiveness of the model's estimation process (Tsay, 2010). In this study, linear interpolation is used as the method for initial imputation. Linear interpolation estimates missing values based on two known neighboring observations using a straight-line approach (Burden & Faires, 2010). This method was selected due to its simplicity, its lack of reliance on complex statistical assumptions, and its ability to preserve local trends in time series data (Niedzielski & Halicki, 2023). Linear interpolation was used solely as an initial estimate to support the Kalman Smoothing process, thereby minimizing potential bias in the final imputation results.

The imputation process using the Kalman Filter begins by defining two main equations: the State Equation and the Observation Equation. The State Equation represents the model in which the value at a given time point (x_t) is influenced by its preceding value (x_{t-1}) and some random noise (w_t). The general form of the State Equation is as follows (Haykin, 2001):

$$x_t = A_t x_{t-1} + w_t \tag{1}$$

where
x_t = the observed value at time t
x_{t-1} = the observed value at time $(t-1)$
A_t = the transition coefficient representing the influence of the previous observation (x_{t-1}) on the current observation (x_t)
w_t = white noise representing random disturbances

The random disturbance term (w_t) is assumed to follow a normal distribution, $w_t \sim \mathrm{N}(0, \mathrm{Q_t})$, indicating that it has zero mean (unbiased) and a variance of Q_t. In this study, Q_t which represents the error covariance, is set to 0.01 as a small value reflecting low process noise while still allowing flexibility in the system dynamics. This choice is intended to prevent excessive fluctuations in the state estimation process (Commandeur & Koopman, 2001). Consequently, the disturbance component remains stochastic but does not dominate the underlying trend structure of the data.

The Observation Equation is used to link the predicted values to the actual observed values. This equation states that the observed data (z_t) is a combination of the true underlying value and measurement error (w_t). The Observation Equation can be expressed as follows (Haykin, 2001):

$$z_t = H_t x_t + w_t \tag{2}$$

where
z_t = the recorded data at time t
H_t = the coefficient that relates the predicted value to the observed data at time t
x_t = the true value of the observation at time t
w_t = white noise representing random disturbances

After defining both equations, the main stages of the Kalman Filter are carried out, beginning with the Prediction Step, in which the value at a given time is estimated and the level of error is calculated. The Prediction Step is used to estimate the current data point $(\hat{x}_t^-)$ based on the previous data, and to quantify the uncertainty of the prediction result (P_t^-). The equation for the initial estimate in the Prediction Step of the Kalman Filter is given as follows (Brown & Hwang, 2013):

$$\hat{x}_t^- = A_t \hat{x}_{t-1} \tag{3}$$

where
$\hat{x}_t^-$ = the predicted value before the update at time t
$\hat{x}_{t-1}$ = the predicted value after the update at time $(t-1)$
A_t = the transition coefficient representing the influence of the previous observation (x_{t-1}) on the current observation (x_t)

The equation for calculating the initial prediction error covariance is as follows (Brown & Hwang, 2013):

$$P_t^- = A_t P_{t-1} A_t + Q_t \tag{4}$$

where
P_t^- = the error covariance or the uncertainty measure of the prior prediction at time t
P_{t-1} = the error covariance or uncertainty measure of the prediction after the update at time $(t-1)$
A_t = the transition coefficient representing the influence of the previous observation(x_{t-1}) on the current observation (x_t)
Q_t = the error covariance or uncertainty associated with the model

Next is the Update Step, which is the process of refining the state estimation based on the previous prediction and new measurements. In this step, the prior prediction is adjusted by incorporating the Kalman Gain, which serves as a weighting factor that balances the predicted values and the observed data. This stage consists of three main equations: the Kalman Gain, the updated state estimate, and the computation of the error covariance after the update, which are stated as follows (Brown & Hwang, 2013):

$$k_t = \frac{P_t^- H_t}{H_t P_t^- H_t + R_t} \tag{5}$$

$$\hat{x}_t = \hat{x}_t^- + k_t(z_t - H_t \hat{x}_t^-) \tag{6}$$

$$P_t = (1 - k_t H_t) P_t^- \tag{7}$$

where
k_t = the Kalman Gain
P_t^- = the error covariance or uncertainty measure of the prior prediction at time t
R_t = the measurement noise covariance

H_t = the coefficient linking the predicted value to the observed data at time t
$\hat{x}_t^-$ = the predicted state before the update at time t
$\hat{x}_t$ = the updated state estimate at time t
z_t = the observed data at time t
P_t = the updated error covariance or uncertainty measure after the update at time t

The estimates obtained from the Kalman Filter are based on information from past data up to time *t*. To enhance these estimates by also utilizing future information, the Kalman Smoothing technique is employed. Kalman Smoothing operates through a backward pass that involves the calculation of the Backward Gain. The Backward Gain is used to adjust and link the current and previous states by refining the estimates across both time points (Zheng & Wang, 2020). The equation for computing the Backward Gain is as follows (Haykin, 2001):

$$j_t = \frac{P_t A_t}{P_{t+1}^-} \tag{8}$$

where
j_t = the Backward Gain, which corrects the current estimate by incorporating the prediction at the next time step $(t+1)$
$\mathrm{P_t}$ = the error covariance or uncertainty measure of the prediction after the update at time t
A_t = the transition coefficient representing the influence of the previous observation (x_{t-1}) on the current observation (x_t)
P_{t+1}^- = the error covariance or uncertainty of the prior estimate before the update at time $(t+1)$

After the Backward Gain has been calculated, the next step is to refine the final estimate. This process, known as the State Smoothing Estimate, serves to correct the prior estimates by incorporating future data through a backward pass. It provides improved estimates that are used to fill in the missing values (Wang et al., 2018). The equation for the State Smoothing Estimate is given as follows (Haykin, 2001):

$$\hat{x}_t^s = \hat{x}_t + j_t(\hat{x}_{t+1}^s - \hat{x}_{t+1}^-) \tag{9}$$

where
$\hat{x}_t^s$ = the final smoothed estimate at time t
$\hat{x}_t$ = the updated state estimate at time t
$\mathrm{j_t}$ = the Backward Gain, which adjusts the current estimate by incorporating information from time $(t+1)$
$\hat{\mathrm{x}}_{\mathrm{t+1}}^{\mathrm{s}}$ = the final smoothed estimate at time $(t+1)$
$\hat{\mathrm{x}}_{\mathrm{t+1}}^-$ = the prior state estimate before the update at time $(t+1)$

4. Seasonal Trend Decomposition Using LOESS (Locally Estimated Scatterplot Smoothing)

Seasonal-Trend Decomposition using LOESS (STL Decomposition) is one of the time series analysis methods used to decompose complex data patterns into separate sub-patterns based on specific criteria (Makkulau et al., 2017). This method aims to reveal underlying trends and seasonal variations in the data, as time series data often exhibit multiple seasonal patterns—such as daily, weekly, and others (Zhou et al., 2021).

This method can also handle data with irregular frequency intervals and effectively separates trend, seasonal, and residual components, thereby capturing all underlying patterns in the data (He et al., 2021). Compared to parametric decomposition techniques, STL Decomposition does not impose strict distributional assumptions, making it particularly suitable for environmental data characterized by nonlinearity and irregular variation. In the context of imputation, reconstructing missing values through these separate structural components allows for the preservation of underlying trends and seasonal dynamics.

In the context of imputation, STL Decomposition breaks the data down into three components: trend, seasonal, and residual, through a decomposition process. These components are computed, and missing values are then estimated by taking into account the relevant components. The general form of the STL Decomposition model is expressed as follows (Zhou et al., 2024):

$$y_t = T_t + S_t + R_t \,; t = 1,2, \dots \ , N \tag{10}$$

where
y_t = the observed time series value at time t
T_t = the trend component at time t
S_t = the seasonal component at time t
R_t = the residual (or remainder) component at time t

Before calculating each component, the STL Decomposition model requires complete initial information for the entire dataset. This is because STL Decomposition relies on complete data to accurately perform the decomposition process (Kwok et al., 2023). Therefore, an initial imputation is applied to the missing data using linear interpolation. Once initial values for the missing observations have been estimated, the decomposition of the time series into its components is carried out. The computation of the components in STL Decomposition employs the LOESS (Locally Estimated Scatterplot Smoothing) approach. LOESS is a nonparametric smoothing technique that captures patterns in data without assuming a specific functional form. It works by calculating localized weights and then fitting a regression model within the local neighborhood (Ouyang et al., 2023). The equations for computing each component are given as follows (Kwok et al., 2023):

$$T_t = \text{LOESS}(y_t - S_t) \tag{11}$$

$$S_t = \text{LOESS}(y_t - T_t) \tag{12}$$

$$R_t = y_t - \ T_t - S_t) \tag{13}$$

Once the values of each component have been obtained, they are then used to estimate the missing values. The estimation is calculated by summing the trend and seasonal components. The residual component is excluded from the estimation process because it represents random noise and does not carry meaningful structural information (Qin et al., 2021).

5. Model Vector Autoregressive

The Vector Autoregressive (VAR) model is a statistical model used to analyze interactions among time series variables. This model provides insights into how one variable may influence others (Kirchgassner & Wolters, 2015). One of the advantages of the VAR model is its ability to effectively capture the movement of observed data (Hendajany & Wati, 2020). In climatological data, this method allows for modeling that accounts for both inter-variable relationships and temporal dynamics. A VAR(p) model of order p is generally expressed as follows (Lutkepohl, 2005):

$$\boldsymbol{y}_t = \boldsymbol{v} + \boldsymbol{A}_1\boldsymbol{y}_{t-1} + \boldsymbol{A}_2\boldsymbol{y}_{t-2} + \cdots + \ \boldsymbol{A}_p\boldsymbol{y}_{t-p} + u_t \tag{14}$$

where
$\boldsymbol{y}_t$ = a K x 1 vector of random variables representing the values of the modeled variables at time t
$\boldsymbol{v}$ = a K x 1 constant vector representing the intercept
$\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_\mathrm{p}$ = a K x K coefficient matrices representing the influence of lags 1 through p on y_t.
$\boldsymbol{u_t}$ = a K x 1white noise error vector at time t

the VAR model is estimated using the complete dataset to capture the underlying temporal dynamics and inter-variable relationships. The estimated VAR model is then used as a data-generating process to simulate synthetic multivariate time series data. Artificial missing values are subsequently introduced into the simulated dataset under controlled conditions. This framework enables the true values to be known, allowing an objective evaluation of the imputation methods.

In general, before implementing the Vector Autoregressive (VAR) model, each variable must first be tested for stationarity to ensure that it does not contain a unit root. This can be conducted using the Augmented Dickey–Fuller (ADF) test at a 5% significance level (Guo, 2023). The hypotheses of the ADF test are formulated as follows:
$H_0 : \ \gamma = 0$ (the data are non-stationary or contain a unit root)

$H_1 : \gamma \neq 0$ (the data are stationary or do not contain a unit root)

The decision rule is based on the calculated test statistic $t_{\text{statistic}}$. If the calculated value is smaller than the Dickey–Fuller critical value at the 5% significance level, then H_0 is rejected, indicating that the data are stationary.

Subsequently, the VAR model parameters are estimated based on the selected lag length. Determining the optimal lag is essential to identify how long past values of a variable influence other variables in the system (Febrianti et al., 2021). The optimal lag length in the VAR model can be selected using the Akaike Information Criterion (AIC).

Thereafter, the Granger causality test is conducted to examine the causal relationships among variables, namely whether one variable can be used to predict another (Lutkepohl, 2005). The hypotheses of the Granger Causality test are formulated as follows:
H_0 : one variable does not influence (does not Granger-cause) the other variable.
H_1 : one variable influences (Granger-causes) the other variable.

The decision rule is based on the calculated $F_{\text{statistic}}$. If the calculated F-value is greater than the critical F-value at the 5% significance level, then H_0 is rejected, indicating that one variable Granger-causes the other variable.

6. Missing Data Simulation

The missing data simulation was conducted to evaluate the effectiveness and robustness of the imputation methods. Three missing data mechanisms were applied: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR), each representing a different scenario of data loss.

MCAR (Missing Completely at Random) occurs when the missing values appear completely at random, without depending on the variable itself or any other observed variables (Zhou et al., 2024). In this mechanism, missing values are introduced by randomly selecting observations from the entire dataset and replacing them with NaN.

MAR (Missing at Random) occurs when the missing values in a variable can be predicted from other variables, but do not depend on the variable itself (Zhou et al., 2024). In this mechanism, for each target variable, one other column is randomly selected as a reference. The values in the target variable are then replaced with NaN only for rows where the reference column exceeds its median, and the rows are randomly sampled from these candidates according to the desired number of missing values. This ensures that missingness depends on other variables, but its distribution is not entirely random.

MNAR (Missing Not at Random) occurs when the probability of a value being missing depends on the value itself (Zhou et al., 2024). In this mechanism, missing values are created based on the variable's own values. For example, values greater than the column median are replaced with NaN.

Each of these missing data mechanisms presents unique challenges for the imputation methods. Additionally, six levels of missing data proportions were applied for each condition: 5%, 10%, 20%, 30%, 40%, and 50% of the total dataset. This variation is used to assess the sensitivity of the imputation methods to the proportion of missing data.

7. Root Mean Square Error

Root Mean Square Error (RMSE) is a fundamental measure of the difference between two corresponding values, commonly used to assess prediction error (Hastomo et al., 2019). RMSE provides an indication of how close predicted results are to the actual values. Mathematically, RMSE is calculated as follows (Kamble & Deshmukh, 2017):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \tag{15}$$

where
x_i = actual value
$\hat{x}_i$ = predicted value

n = number of data points

In evaluating prediction results, RMSE is used as an indicator of the level of error. The method that yields the smallest RMSE value is considered the best, as it indicates the lowest prediction error compared to other methods. Although RMSE is employed as the primary evaluation metric in this study to assess the accuracy of the imputed values, alternative performance measures can provide complementary insights. For instance, Mean Absolute Error (MAE) reflects the average magnitude of errors without assigning greater weight to large deviations, while Mean Absolute Percentage Error (MAPE) allows for a scale-independent assessment of relative error. Nevertheless, RMSE is particularly suitable in this study because it penalizes larger errors more heavily, which is important in the context of climate time series data where extreme deviations may substantially affect model performance (Prasetya et al., 2025). Therefore, RMSE is considered an appropriate and robust primary metric for evaluating imputation accuracy, while MAE and MAPE can provide additional perspective for a more comprehensive assessment.

8. Result

8.1 Descriptive Analysis

Descriptive statistics are utilized to provide an overview of the climate data under study

Table 1. Descriptive Statistics

Variable	Minimum	Maximum	Mean	Median	Standard Deviation
Average Temperature	22.84	31.38	26.37	26.34	0.99
Average Relative Humidity	44.70	118.57	86.00	86.12	0.77
Total Precipitation	0.00	396.40	6.39	0.101	14.76
Average Solar Radiation	0.00	290.33	63.99	61.41	40.33
Average Wind Speed	0.00	7.95	2.35	2.35	1.35

Table 1. shows that average temperature, average relative humidity, and average wind speed exhibit relatively symmetric and stable distributions, as indicated by the close values of the mean and median, as well as small to moderate standard deviations. In contrast, total precipitation and average solar radiation display greater variability, especially total precipitation, which is highly uneven with a median value significantly lower than the mean. This suggests many dry days with occasional extreme total precipitation events. These findings reflect the tropical climate characteristics of Lampung, which tends to experience high fluctuations in certain variables like total precipitation and average solar radiation, while other variables remain relatively consistent throughout the observation period.

Table 2. presents the percentage of missing data in the dataset, where the proportion for each variable is less than 10%. The variable with the lowest percentage of missing data is average solar radiation, with 1.89% missing or 128 missing observations out of the total. The variable with the highest percentage of missing data is average relative humidity, with 6.91% missing, corresponding to 469 missing observations.

Table 2. Missing Data Percentage

Variable	Missing Data
Average Temperature	5.82%
Average Relative Humidity	6.91%
Total Precipitation	4.07%
Average Solar Radiation	1.89%
Average Wind Speed	3.02%

8.2 Kalman Smoothing Imputation

The Kalman Smoothing method was applied to address missing data in each variable of the climate dataset. Unlike simpler techniques that rely solely on past observations, Kalman Smoothing utilizes information from both preceding and subsequent time points within the observation period. This bidirectional framework enables more accurate estimation of missing values by incorporating a more comprehensive representation of the underlying trends and temporal patterns. As a state-space-based approach, it combines prediction and updating steps, thereby enhancing imputation quality compared to methods that treat observations independently.

After implementing the Kalman Smoothing algorithm, a complete set of estimated values was generated for the missing observations in each climate variable. These estimates were subsequently inserted into the dataset to replace missing entries. The resulting imputed dataset preserves temporal coherence and smooth transitions between observed and estimated values, which is particularly important in time series data such as climate records..

The imputation results are presented in the figure, which shows two sets of data points distinguished by color: black represents the observed (non-missing) data, whereas red represents the values imputed using Kalman Smoothing. In Figure 1, the Kalman Smoothing method demonstrates a strong ability to impute missing values effectively. The imputation process follows both the trend and seasonal patterns of the average temperature data, resulting in smooth transitions between the observed and estimated values. In Figure 2, for the average relative humidity variable, missing values in the middle to later periods are estimated smoothly without generating extreme fluctuations. In Figure 3, the total precipitation variable exhibits relatively high variability. Nevertheless, the imputed values continue to follow the overall data pattern without showing significant deviations. In Figure 4, for the average solar radiation variable, the imputation results are generally consistent with the observed trend pattern. Meanwhile, in Figure 5, the average wind speed variable appears relatively fluctuating; however, the imputed values still align consistently with the overall data pattern.

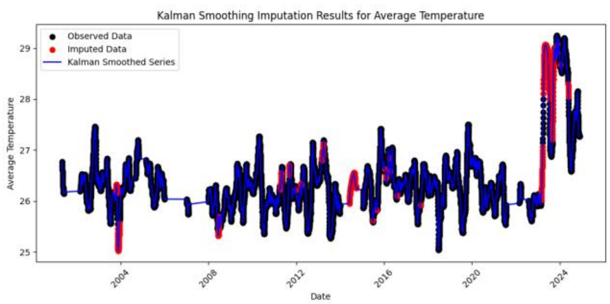

Figure 1. Kalman Smoothing for Average Temperature

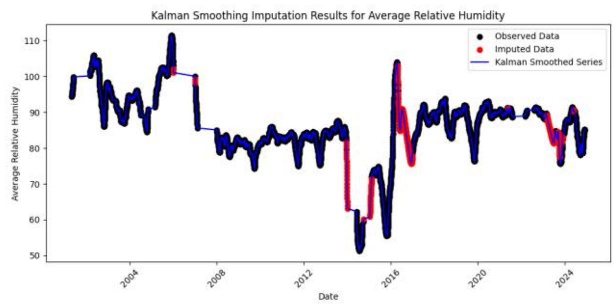

Figure 2. Kalman Smoothing for Average Relative Humidity

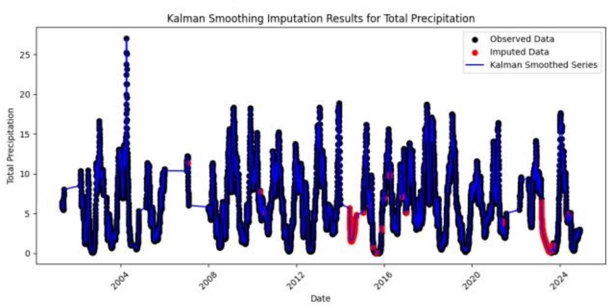

Figure 3. Kalman Smoothing for Total Precipitation

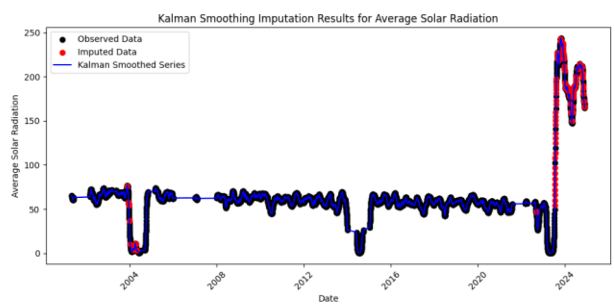

Figure 4. Kalman Smoothing for Average Solar Radiation

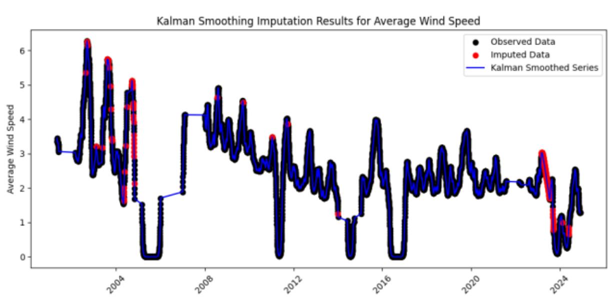

Figure 5. Kalman Smoothing for Average Wind Speed

 DOI: http://dx.doi.org/10.18187/pjsor.v22i2.5011

8.3 Seasonal Trend Decomposition Using LOESS Imputation

The imputation method using STL Decomposition (Seasonal Trend Decomposition using LOESS) was applied by first breaking down the time series data into three fundamental components. This decomposition allows for a clearer understanding of the underlying structure of the data, which is essential when estimating missing values in time series. After decomposition, missing values are estimated by recombining the relevant trend and seasonal components while adjusting for irregularities. The resulting imputed dataset preserves temporal coherence and seasonal patterns, providing a reliable and interpretable approach for handling missing data in climate time series.

In Figure 6 and Figure 7, for the average temperature and average relative humidity variables, respectively, the STL Decomposition method imputes missing values with relatively smooth transitions and without sharp fluctuations, indicating its suitability for relatively stable data. In Figure 8, although the total precipitation variable generally follows the overall trend and seasonal pattern, the imputed values tend to be smoother and fail to fully capture the magnitude of extreme precipitation events. In Figure 9 and Figure 10, for average solar radiation and average wind speed, respectively, the imputation results generally follow the overall data patterns, although minor fluctuations are still observed.

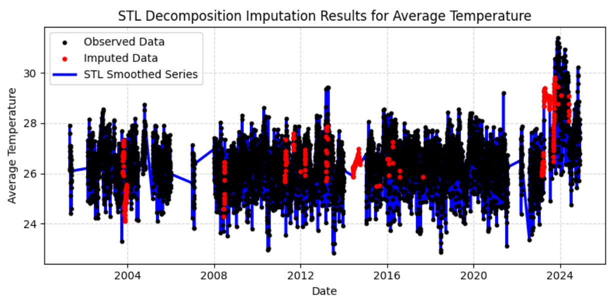

Figure 6. STL Decomposition for Average Temperature

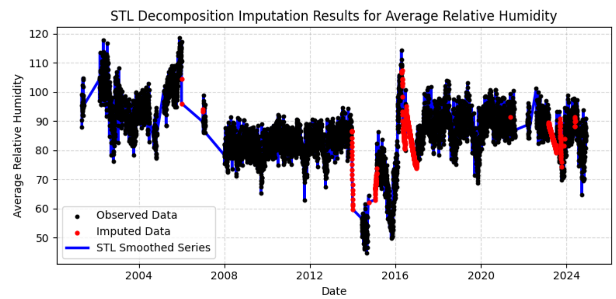

Figure 7. STL Decomposition for Average Relative Humidity

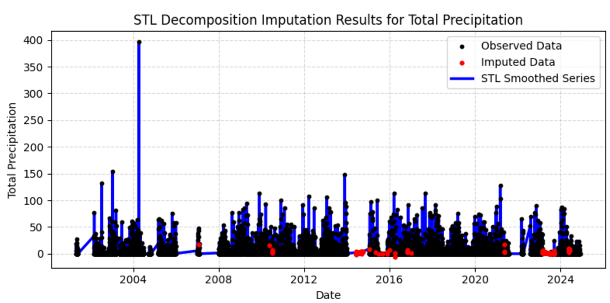

Figure 8. STL Decomposition for Total Precipitation

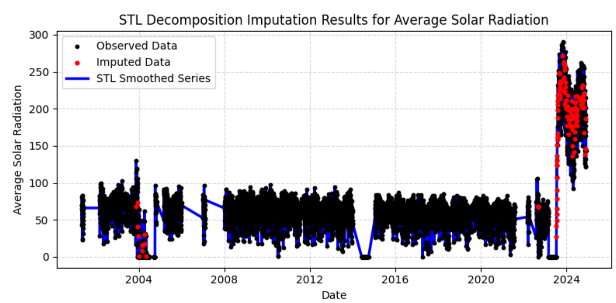

Figure 9. STL Decomposition for Average Solar Radiation

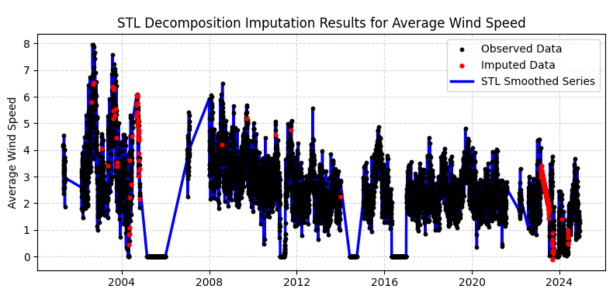

Figure 10. STL Decomposition for Average Wind Speed

8.4 Vector Autoregressive Models

In the Vector Auto Regressive (VAR) modeling process, an Augmented Dickey Fuller (ADF) test was conducted to determine whether the data is stationary or not. The following are the results of the ADF test:

Table 3. Augmented Dickey Fuller Test

Variable	Kalman Smoothing			STL Decomposition		
	t-statistic	critical value	conclusion	t-statistic	critical value	conclusion
Average Temperature	-6.75	-2.86	Stasionary	-6.93	-2.86	Stasionary
Average Relative Humidity	-4.13	-2.86	Stasionary	-4.44	-2.86	Stasionary
Total Precipitation	-8.74	-2.86	Stasionary	-8.77	-2.86	Stasionary
Average Solar Radiation	-2.97	-2.86	Stasionary	-2.88	-2.86	Stasionary
Average Wind Speed	-5.06	-2.86	Stasionary	-5.13	-2.86	Stasionary

Table 3. shows that the five variables under both imputation methods have t-statistics smaller than the critical value $t_{(const,5\%,6787)}$ = -2.86. This indicates that the null hypothesis—that the data are non-stationary—can be rejected. Therefore, there is sufficient evidence that each variable is stationary in both datasets. Following the stationarity test, the next step is to identify the optimal lag order for the VAR model based on the lowest Akaike Information Criterion (AIC) value.

For the Kalman Smoothing imputed dataset, the optimal lag is found at lag 12 with an AIC value of 9.720. Thus, the VAR model for the Kalman Smoothing method uses lag 12. Meanwhile, for the STL Decomposition imputed dataset, the lowest AIC value is found at lag 14, with an AIC of 9.754. Therefore, the VAR model for the STL Decomposition method uses lag 14.

Table 4. Optimal Lag Results

Lag	AIC	
	Kalman Smoothing	STL Decomposition
0	17.680	17.730
1	10.580	10.550
2	10.149	10.130
3	9.947	9.960
4	9.867	9.880
5	9.797	9.822
6	9.774	9.888
7	9.754	9.788
8	9.745	9.777
9	9.737	9.771
10	9.729	9.763
11	9.728	9.764
12	**9.720**	9.755
13	9.721	9.756
14	9.722	**9.754**
15	9.724	9.757

The VAR model for the average temperature variable in the dataset imputed using the Kalman Smoothing method is described as follows:

$$y_{1t} = 3.361 + 0.347y_{1_{t-1}} - 0.0041y_{2_{t-1}} - 0.0029y_{3_{t-1}} - 0.0020y_{4_{t-1}} - 0.089y_{5_{t-1}} + 0.104y_{1_{t-2}} - 0.0004y_{2_{t-2}} + 0.001y_{3_{t-2}} + 0.0006y_{4_{t-2}} + 0.029y_{5_{t-2}} + \cdots - 0.034y_{5_{t-12}}$$

One of the VAR models, namely the VAR model for y_{1t} (average temperature), indicates that changes in average temperature are positively influenced by the temperature value of the previous day, with a coefficient of 0.347, assuming other variables remain constant. This implies that if the previous day's temperature increases, today's temperature is also likely to increase. However, the effects of relative humidity, precipitation, solar radiation, and wind speed at lag 1 appear to be small and negative. Among these four variables, wind speed has the most notable negative effect, with a coefficient of -0.089, suggesting that stronger winds on the previous day slightly decrease today's average temperature by approximately 0.089°C. Meanwhile, the results of the VAR modeling for the average temperature variable in the dataset imputed using the STL Decomposition method are presented as follows:

$$y_{1t} = 3.107 + 0.3629y_{1_{t-1}} - 0.0033y_{2_{t-1}} - 0.0037y_{3_{t-1}} - 0.0022y_{4_{t-1}} - 0.0892y_{5_{t-1}} + 0.104y_{1_{t-2}} - 0.00058y_{2_{t-2}} + 0.0016y_{3_{t-2}} + 0.00067y_{4_{t-2}} + 0.0318y_{5_{t-2}} + \cdots + 0.0503y_{5_{t-14}}$$

One of the VAR models, specifically the VAR model for y_{2t} (relative humidity), indicates that changes in humidity are positively influenced by the previous period's temperature by 0.3629, assuming other variables are held constant. This means that if the previous day's temperature increases, the current day's relative humidity is also likely to increase. Additionally, the effect of humidity on itself is positive, with a coefficient of 0.6876, indicating that if humidity increased in the previous period, its value in the current period is also likely to increase by approximately 0.6876°C.

From the VAR models for both datasets resulting from Kalman Smoothing and STL Decomposition imputations, the Granger Causality test on the VAR model using the Kalman Smoothing approach shows that average temperature is influenced by precipitation, solar radiation, and wind speed. Relative humidity is influenced by all variables except wind speed. precipitation is influenced by relative humidity. Solar radiation is influenced by temperature, humidity, and wind speed. Wind speed is influenced by all other variables. Meanwhile, for the STL Decomposition approach, the results are generally similar, but with some differences, such as relative humidity not being influenced by precipitation but instead by wind speed, and wind speed not being influenced by relative humidity.

Table 5. Granger Causality Test

Caused Variable	Causing Variable	Kalman Smoothing			STL Decomposition		
		$F_{statistic}$	F_{table}	conc	$F_{statistic}$	F_{table}	conc
Average Temperature	Average Relative Humaidity	0.831	1.753	No	1.048	1.693	No
Average Temperature	Total Precipitation	3.396	1.753	Yes	2.755	1.693	Yes
Average Temperature	Average Solar Radiation	3.101	1.753	Yes	3.022	1.693	Yes
Average Temperature	Average Wind Speed	3.000	1.753	Yes	3.461	1.693	Yes
Average Relative Humaidity	Average Temperature	24.710	1.753	Yes	22.269	1.693	Yes
Average Relative Humaidity	Total Precipitation	2.188	1.753	Yes	1.678	1.693	No
Average Relative Humaidity	Average Solar Radiation	2.883	1.753	Yes	3.376	1.693	Yes
Average Relative Humaidity	Average Wind Speed	1.434	1.753	No	1.912	1.693	Yes
Total Precipitation	Average Temperature	1.401	1.753	No	1.137	1.693	No
Total Precipitation	Average Relative Humaidity	1.858	1.753	Yes	1.830	1.693	Yes
Total Precipitation	Average Solar Radiation	1.160	1.753	No	0.971	1.693	No
Total Precipitation	Average Wind Speed	0.858	1.753	No	0.937	1.693	No
Average Solar Radiation	Average Temperature	13.961	1.753	Yes	13.071	1.693	Yes
Average Solar Radiation	Average Relative Humaidity	4.419	1.753	Yes	3.152	1.693	Yes
Average Solar Radiation	Total Precipitation	0.566	1.753	No	0.662	1.693	No
Average Solar Radiation	Average Wind Speed	1.843	1.753	Yes	1.929	1.693	Yes
Average Wind Speed	Average Temperature	3.556	1.753	Yes	2.928	1.693	Yes
Average Wind Speed	Average Relative Humaidity	2.134	1.753	Yes	1.571	1.693	No
Average Wind Speed	Total Precipitation	3.345	1.753	Yes	3.129	1.693	Yes
Average Wind Speed	Average Solar Radiation	2.887	1.753	Yes	2.294	1.693	Yes

8.5 Evaluation of RMSE on the VAR Model

Table 6. shows that the STL Decomposition imputation method yields a lower RMSE compared to Kalman Smoothing, with a value of 4.5513. This lower RMSE indicates that the estimates produced by STL Decomposition are closer to the actual data and exhibit a lower level of error. Therefore, the VAR model based on the dataset imputed using STL Decomposition is utilized for the generation of new data for further analysis.

Table 6. VAR Model Evaluation

Imputation Method	RMSE Value
Kalman Smoothing	4.8287
STL Decomposition	4.5513

8.6 Evaluation of the Effectiveness of Imputation Methods

Effectiveness testing was conducted using newly generated data derived from the VAR model based on the best-performing dataset, which was obtained through the STL Decomposition imputation method. This generated data was then used to evaluate the effectiveness of the two imputation methods, namely Kalman Smoothing and STL Decomposition. A missing data scheme was applied based on three types of missing data mechanisms: MCAR (Missing Completely at Random), MAR (Missing at Random), and MNAR (Missing Not at Random), each representing different challenges for the imputation methods in estimating missing values. Furthermore, for each mechanism, six levels of missing data percentages were applied: 5%, 10%, 20%, 30%, 40%, and 50% of the total data.

Based on the comparison of imputation results between Kalman Smoothing and STL Decomposition across different percentages of missing data (5%–50%) and missing data mechanisms (MCAR, MAR, MNAR), STL Decomposition consistently outperformed Kalman Smoothing, yielding lower RMSE values across all variables. At lower levels of missing data (5%–20%), the method demonstrated excellent performance, while at higher levels (40%–50%), its performance declined, particularly under the MNAR condition, but it remained more accurate than Kalman Smoothing. Overall, increasing the proportion of missing data led to higher RMSE values, with MCAR yielding the best performance and MNAR the worst.

Considering the comparison across the five climate variables presented in Table 7, STL Decomposition is recommended as the more effective method for handling missing data in climate datasets in Lampung. This approach better preserves temporal and seasonal patterns, resulting in more accurate and consistent imputations across various missing data scenarios compared to Kalman Smoothing.

Table 7. Comparison of Imputation Methods

Variable	Percentage of Missing Data	Method	RMSE		
			MAR	MCAR	MNAR
Average Temperature	5%	Kalman Smoothing	0.765	0.664	0.711
		STL Decomposition	**0.067**	**0.06**	**0.069**
	10%	Kalman Smoothing	0.684	0.675	0.709
		STL Decomposition	**0.01**	**0.09**	**0.1**
	20%	Kalman Smoothing	0.696	0.744	0.72
		STL Decomposition	**0.14**	**0.14**	**0.15**
	30%	Kalman Smoothing	0.525	0.684	0.712
		STL Decomposition	**0.18**	**0.18**	**0.17**
	40%	Kalman Smoothing	0.725	0.689	0.722
		STL Decomposition	**0.2**	**0.19**	**0.21**
	50%	Kalman Smoothing	0.71	0.692	0.711
		STL Decomposition	**0.22**	**0.22**	**0.23**
Average Relative Humidity	5%	Kalman Smoothing	2.982	3.454	3.526
		STL Decomposition	**0.319**	**0.318**	**0.307**
	10%	Kalman Smoothing	3.422	3.116	3.322
		STL Decomposition	**0.467**	**0.468**	**0.442**
	20%	Kalman Smoothing	3.142	3.688	3.437
		STL Decomposition	**0.63**	**0.644**	**0.639**
	30%	Kalman Smoothing	3.233	3.277	3.636
		STL Decomposition	**0.73**	**0.806**	**0.843**
	40%	Kalman Smoothing	3.263	3.404	3.713

		STL Decomposition	**0.9**	**0.937**	**0.99**
	50%	Kalman Smoothing	3.457	3.404	3.58
		STL Decomposition	**1.02**	**1.013**	**1.078**
Total Precipitation	5%	Kalman Smoothing	12.338	13.369	14.33
		STL Decomposition	**1.49**	**1.54**	**1.74**
	10%	Kalman Smoothing	13.814	13.536	13.457
		STL Decomposition	**2.4**	**2.41**	**3.09**
	20%	Kalman Smoothing	15.267	14.225	13.832
		STL Decomposition	**3.37**	**3.37**	**3.03**
	30%	Kalman Smoothing	14.704	14	14.039
		STL Decomposition	**3.63**	**3.63**	**3.81**
	40%	Kalman Smoothing	14.589	14.297	14.192
		STL Decomposition	**4.28**	**4.28**	**4.44**
	50%	Kalman Smoothing	14.714	15.178	14.814
		STL Decomposition	**4.93**	**4.93**	**5.1**
Average Solar Radiation	5%	Kalman Smoothing	14.824	17.143	17.338
		STL Decomposition	**1.569**	**1.579**	**1.729**
	10%	Kalman Smoothing	14.926	14.625	17.198
		STL Decomposition	**2.441**	**2.441**	**2.317**
	20%	Kalman Smoothing	15.028	15.277	16.235
		STL Decomposition	**3.134**	**3.134**	**3.126**
	30%	Kalman Smoothing	14.783	16.538	16.543
		STL Decomposition	**3.717**	**3.717**	**3.9**
	40%	Kalman Smoothing	15.788	15.891	16.589
		STL Decomposition	**4.361**	**4.36**	**4.638**
	50%	Kalman Smoothing	15.815	16.13	16.469
		STL Decomposition	**5.08**	**5.08**	**5.243**
Average Wind Speed	5%	Kalman Smoothing	0.56	0.571	0.563
		STL Decomposition	**0.055**	**0.048**	**0.055**
	10%	Kalman Smoothing	0.551	0.54	0.568
		STL Decomposition	**0.068**	**0.068**	**0.076**
	20%	Kalman Smoothing	0.558	0.573	0.588
		STL Decomposition	**0.105**	**0.104**	**0.112**
	30%	Kalman Smoothing	0.525	0.564	0.577
		STL Decomposition	**0.131**	**0.131**	**0.135**
	40%	Kalman Smoothing	0.52	0.527	0.571
		STL Decomposition	**0.146**	**0.146**	**0.153**
	50%	Kalman Smoothing	0.525	0.564	0.575
		STL Decomposition	**0.16**	**0.16**	**0.17**

9. Conclusion

Based on the comparison of the imputation methods, Kalman Smoothing and STL Decomposition, it was found that the STL Decomposition method is more effective and accurate. This is indicated by its lower error rate (RMSE) compared to Kalman Smoothing, making it the best method for handling missing values in climate data in Lampung from 2001 to 2024.

Furthermore, based on the effectiveness tests of Kalman Smoothing and STL Decomposition across various missing data conditions and percentages, STL Decomposition consistently showed superior performance, particularly at missing data percentages of 5%, 10%, and 20%, although a decline in accuracy was observed at higher missing percentages. In addition, under MCAR, MAR, and MNAR missing data conditions, STL Decomposition also produced the best results by yielding the lowest RMSE values. While Kalman Smoothing showed relatively stable results, STL Decomposition still outperformed it in terms of precision and estimation accuracy. From this comparison, STL Decomposition is recommended as the preferred imputation method for handling missing climate data in Lampung, Indonesia from 2001 to 2024.

References

1. Agiwal, V., & Chaudhuri, S. (2024). Methods and Implications of Addressing Missing Data in Health-care Research. *Current Medical Issues*, (19), 185–187. https://doi.org/10.4103/cmi.cmi
2. Ahn, H., Sun, K., & Kim, K. P. (2021). Comparison of missing data imputation methods in time series forecasting. *Computers, Materials and Continua*, *70*(1), 767–779. https://doi.org/10.32604/cmc.2022.019369
3. Austin, P. C., White, I. R., Lee, D. S., & van Buuren, S. (2021). Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Canadian Journal of Cardiology*, *37*(9), 1322–1331. https://doi.org/10.1016/j.cjca.2020.11.010
4. Brown, R. G., & Hwang, P. Y. C. (2013). introduction to Random signals and Applied Kalman Filtering with MATLAB exercises. In *Journal of Chemical Information and Modeling* (Vol. 53).
5. Burden, R. L., & Faires, J. D. (2010). Numerical Analysis. In *Sustainability (Switzerland)*.
6. Chaudhry, A., Li, W., Basri, A., & Patenaude, F. (2019). A Method for Improving Imputation and Prediction Accuracy of Highly Seasonal Univariate Data with Large Periods of Missingness. *Wireless Communications and Mobile Computing*, *2019*. https://doi.org/10.1155/2019/4039758
7. Chhabra, G. (2023). Comparison of Imputation Methods for Univariate Time Series. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(December 2022), 286–292. https://doi.org/10.17762/ijritcc.v11i2s.6148
8. Commandeur, J. J. ., & Koopman, S. J. (2001). *State Space Time Series Analysis*.
9. Febrianti, D. R., Tiro, M. A., & Sudarmin, S. (2021). Metode Vector Autoregressive (VAR) dalam Menganalisis Pengaruh Kurs Mata Uang Terhadap Ekspor Dan Impor Di Indonesia. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, *3*(1), 23. https://doi.org/10.35580/variansiunm14645
10. Guo, Z. (2023). Research on the Augmented Dickey-Fuller Test for Predicting Stock Prices and Returns. *Advances in Economics, Management and Political Sciences*, *44*(1), 101–106. https://doi.org/10.54254/2754-1169/44/20232198
11. Hastomo, W., Satyo, A., Sudjiran, D., Teknologi, I., Bisnis, D., Dahlan, A., & Jl, J. (2019). Long Short Term Memory Machine Learning Untuk Memprediksi Akurasi Nilai Tukar IDR Terhadap USD. *Universitas Gunadarma Jl. Margonda Raya*, *3*(2), 12140.
12. Haykin, S. (2001). *Kalman Filtering and Neural Networks*.
13. He, H., Gao, S., Jin, T., Sato, S., & Zhang, X. (2021). A seasonal-trend decomposition-based dendritic neuron model for financial time series prediction. *Applied Soft Computing*, *108*. https://doi.org/10.1016/j.asoc.2021.107488
14. Hendajany, N., & Wati, R. (2020). Prediksi indikator makro ekonomi Indonesia dengan model vector autoregressive periode 2019-2023. *Jurnal Ekonomi Dan Bisnis*, *23*(2), 189–202. https://doi.org/10.24914/jeb.v23i1.2878
15. Irnawan, F. D., Hidayah, I., & Nugroho, L. E. (2021). Metode Imputasi pada Data Debit Daerah Aliran Sungai Opak, Provinsi DI Yogyakarta. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *10*(4), 301–310. https://doi.org/10.22146/jnteti.v10i4.2430
16. Kamble, V. B., & Deshmukh, S. N. (2017). Comparision Between Accuracy and MSE,RMSE by Using Proposed Method with Imputation Technique. *Oriental Journal of Computer Science and Technology*, *10*(04), 773–779. https://doi.org/10.13005/ojcst/10.04.11
17. Kirchgassner, G., & Wolters, J. (2015). *Introduction to Modern Time Series Analysis*.
18. Kwok, C. F., Qian, G., & Kuleshov, Y. (2023). Analyzing Error Bounds for Seasonal-Trend Decomposition of Antarctica Temperature Time Series Involving Missing Data. *Atmosphere*, *14*(2), 1–21. https://doi.org/10.3390/atmos14020193
19. Lutkepohl, H. (2005). New Introduction to Multiple Time Series Analysis. In *New Introduction to Multiple Time Series Analysis*. https://doi.org/10.1007/3-540-27752-8
20. Makkulau, Raya, R., & Marlinda, S. (2017). Aplikasi Metode Dekomposisi Pada Peramalan Jumlah Kelahiran. *Teknologi Terapan Berbasis Kearifan Lokal (Snt2Bkl)*, 535–545.
21. Niedzielski, T., & Halicki, M. (2023). Improving Linear Interpolation of Missing Hydrological Data by Applying Integrated Autoregressive Models. *Water Resources Management*, *37*(14), 5707–5724.

https://doi.org/10.1007/s11269-023-03625-7
22. Oktavianti, S. A., & Yanti, T. S. (2022). Penanganan Data Hilang Menggunakan Metode MarkoviChain Monte Carlo (MCMC). *Bandung Conference Series: Statistics*, *2*(2), 409–415. https://doi.org/10.29313/bcss.v2i2.4525
23. Ouyang, Z., Jabloun, M., & Ravier, P. (2023). STLformer: Exploit STL decomposition and Rank Correlation for Time Series Forecasting. *European Signal Processing Conference*, 1405–1409. https://doi.org/10.23919/EUSIPCO58844.2023.10290126
24. Phan, T. T. H. (2020). Machine Learning for Univariate Time Series Imputation. *2020 International Conference on Multimedia Analysis and Pattern Recognition, MAPR 2020*, (October). https://doi.org/10.1109/MAPR49794.2020.9237768
25. Prasetya, M. E., Ryansyah, E., Surya, M. R., & Umaidah, Y. (2025). *Prediksi Curah Hujan Menggunakan Jaringan Syaraf Tiruan Backpropagation pada Software Matlab*. *5*(September). https://doi.org/10.33998/jakakom.v5i2
26. Prasetya, M. R. A., Priyatno, A. M., & Nurhaeni. (2023). Penanganan Imputasi Missing Values pada Data Time Series dengan Menggunakan Metode Data Mining. *Jurnal Informasi Dan Teknologi*, (June), 52–62. https://doi.org/10.37034/jidt.v5i2.324
27. Qin, Y., Ren, G., Zhang, P., Wu, L., & Wen, K. (2021). An imputation method for the climatic data with strong seasonality and spatial correlation. *Theoretical and Applied Climatology*, *144*(1–2), 203–213. https://doi.org/10.1007/s00704-021-03537-9
28. Slipetz, L. R., Falk, A., & Henry, T. R. (2023). *Missing Data in Discrete Time State-Space Modeling of Ecological Momentary Assessment Data: A Monte-Carlo Study of Imputation Methods*.
29. Tsay, R. S. (2010). Analysis of financial time series. In *Analysis of Financial Time Series*. https://doi.org/10.1002/9780470644560
30. Umar, N., & Gray, A. (2023). Comparing Single and Multiple Imputation Approaches for Missing Values in Univariate and Multivariate Water Level Data. *Water (Switzerland)*, *15*(8), 1–21. https://doi.org/10.3390/w15081519
31. Wang, H., Li, H., Zhang, W., Zuo, J., & Wang, H. (2018). Maximum Correntropy Derivative-Free Robust Kalman Filter and Smoother. *IEEE Access*, *6*, 70794–70807. https://doi.org/10.1109/ACCESS.2018.2880618
32. Wang, J. S., & Aronow, P. . (2022). Listwise Deletion in High Dimensions. *Political Analysis*, *31*(1), 149–155. https://doi.org/10.1017/pan.2022.5
33. Zheng, Y., & Wang, M. (2020). Bearing-Only Target Tracking Based on Piecewise Backward Smoothing Kalman Filter. *Journal of Physics: Conference Series*, *1621*(1). https://doi.org/10.1088/1742-6596/1621/1/012018
34. Zhou, Youran, Aryal, S., & Bouadjenek, M. R. (2024). *Review for Handling Missing Data with special missing mechanism*. Retrieved from http://arxiv.org/abs/2404.04905
35. Zhou, Yujue, Jiang, J., Yang, S. H., He, L., & Ding, Y. (2021). MuSDRI: Multi-Seasonal Decomposition Based Recurrent Imputation for Time Series. *IEEE Sensors Journal*, *21*(20), 23213–23223. https://doi.org/10.1109/JSEN.2021.3107836