Comparison of Imputation Methods: Traditional, Machine Learning, and Deep Learning on Multivariate Time Series with MCAR and MNAR

F T T Hakiki[1*], N L Tasbihi[1], A A E Dafi[1], Nurfaudzan[1], A S Muthahharah[2]

[1] Statistics Department, Institut Teknologi Sepuluh Nopember, Kampus ITS Sukolilo, Surabaya 60111, Indonesia
[2] Data Science and Decisions Department, The University of New South Wales, High St, Kensington NSW 2052, Australia

*Corresponding author's email: fferigo36@email.com

Abstract. This study compares the methods of Linear Interpolation, Kalman Filtering, SVR, and RNN-GRU for multivariate time series that exhibit linear trends and seasonality. Synthetic data for three variables were generated for small, medium, and large sample sizes. Missing values were systematically inserted using Missing Completely at Random (MCAR) and Missing Not at Random (MNAR) patterns with proportions of 10%, 20%, and 35%. The accuracy of imputation was evaluated using RMSE, MAPE, and R^2 over 150 simulation repetitions per scenario. The results indicate that each method has advantages under certain conditions. Linear Interpolation is suitable for data with linear trends, small sample sizes, and low to moderate missingness levels, and is effective for both MCAR and MNAR patterns. Kalman Filtering is optimal for medium to large datasets, particularly in handling linear and seasonal trend patterns with high proportions of missing data due to MCAR. SVR excels in large seasonal data scenarios with MNAR missingness patterns. RNN-GRU performs well under low missingness conditions, particularly for small seasonal datasets with MNAR patterns. These findings emphasize that the choice of imputation method should consider data size, trend patterns, and the missing data mechanism to minimize bias and preserve the integrity of the temporal structure.
Keyword: Linear, MCAR, MNAR, Multivariate Time Series Imputation, Seasonality

1. Introduction

Multivariate time series comprises a collection of univariate time series (metrics), each delineating distinct components or characteristics of a complicated entity[1]. This type of data is widely utilized in meteorology (such as recording temperature, humidity, and air pressure), economics and finance (for example, stock prices, market indices, and trading volume), as well as health (such as blood pressure, heart rate, and body temperature of patients)[2]–[4]. The main advantage of multivariate data is its ability to capture the dynamics of relationships and interdependencies among variables within a system that changes over time[5].

Missing values present a major challenge in processing multivariate time series data, since losing one or more values in the time sequence can disrupt the structure of temporal relationships between variables[6]. This can lead to biased estimates, reduced statistical power, and inaccuracies in developing

predictive models[7]. For instance, in financial market analysis, missing trading data during certain periods can lead to incorrect conclusions about market movements. Moreover, in critical applications such as patient monitoring, the loss of vital data, such as heart rate or body temperature, can delay medical intervention, potentially resulting in fatal outcomes[8]. Therefore, handling missing data must be done carefully and systematically to preserve the integrity of the analysis results. Time series data have a temporal nature, meaning there is a dependence between the current value and previous values[9]. In many cases, time series data exhibit long-term trend patterns or seasonal cycles that recur, such as monthly sales patterns or annual weather cycles. Missing data at specific points in time within this sequence can break the flow of important information that forms the basis for predictive modelling. As a result, models relying on historical patterns cannot be optimally constructed[10]. Hence, the presence of missing data in time series cannot be overlooked and requires an approach that aligns with the data's temporal structure.

The two mechanisms for missing data are missing at random (MCAR) and missing not at random (MNAR)[11]. MCAR is the assumption that the occurrence of missing data is independent of both observed and unobserved data[12]. In contrast, MNAR occurs when data loss follows a sequential or block pattern, where the chance of missing data depends on the actual data value or unobserved external factors. Recognizing this difference is crucial because MCAR generally does not cause significant bias, whereas MNAR can considerably skew the results[13].

Imputation represents a viable approach for managing missing data. It substitutes absent values with plausible estimates obtained from observed information, thereby maintaining the dataset's integrity and facilitating comprehensive analysis[13]. The choice of imputation method in this study is driven by the need to address the complex features of multivariate time series data, which include linear and seasonal trends, as well as MCAR and MNAR missing patterns[14]. The main methods used are traditional techniques, machine learning-based methods, and deep learning[13]. From a traditional perspective, linear interpolation is selected for its simplicity and capacity to estimate missing values between neighboring time points linearly. Meanwhile, the Kalman Filter is utilized due to its dynamic updating mechanism, based on a state-space model, making it well-suited to handle fluctuations and system dynamics that vary over time[15]. To overcome the limitations of linearity and capture non-linear patterns, machine learning methods such as Support Vector Regression (SVR) are applied, thanks to their ability to model non-linear relationships and robustness against outliers, due to kernel functions and margin-based loss[16]. For deep learning, Recurrent Neural Networks (RNN) with Gated Recurrent Unit (GRU) architecture are employed to accommodate complex temporal dependencies[17]. Each method has different characteristics, assumptions, and levels of complexity in handling temporal relationships and data loss patterns. Traditional methods excel in simplicity and computational efficiency but are less adaptable to nonlinear patterns. Conversely, Machine Learning and Deep Learning methods are more capable of capturing complex dynamics and nonlinear relationships, although they require large amounts of data and high computational resources. Therefore, a comparison between these three approaches is necessary to assess the effectiveness, efficiency, and stability of imputation performance under various data loss conditions (MCAR and MNAR), so that the most suitable method for multivariate time series data characteristics can be identified. This research aims to thoroughly evaluate the performance of these four imputation methods within the context of multivariate time series data showing linear and seasonal trends, with MCAR and MNAR missing patterns. Through this approach, it is hoped that the most suitable imputation method can be recommended based on data characteristics, the extent of missingness, and the dataset's scale.

2. Literature Review

2.1. Multivariate Time Series Data

Time series data is a sequence of observations recorded in order over a period. A key feature of this data is the presence of temporal dependence, meaning that the value at a specific time is strongly influenced by previous values[18]. Multivariate time series data involves two or more variables recorded at the

same time[19]. The primary benefit of the multivariate approach is its capacity to capture complex relationships and interactions between variables[20].

There are several significant benefits of analyzing multivariate time series data. Firstly, it can identify cross-correlation or inter-variable relationships over time[21]. Secondly, multivariate predictive models such as Vector Autoregression (VAR) and Dynamic Factor Models (DFM) can enhance forecasting accuracy, as they incorporate information from multiple variables that collectively improve prediction performance[22][23]. Thirdly, this approach opens opportunities to uncover causality and feedback within complex systems, as well as support more precise data-driven decision-making[24].

In many studies, including those by Chatfield and Hyndman & Athanasopoulos, it is well established that the primary components of a time series consist of linear trends and seasonal patterns, which indicate long-term directional changes and recurring fluctuations[25]. A linear trend in a time series depicts gradual and steady changes over the long term[26]. This pattern can be either increasing or decreasing over time[27]. In the context of multivariate data, linear trends are observed simultaneously across different variables, such as an increase in the number of motor vehicles alongside rising air pollution in urban areas[28]. Meanwhile, seasonal or seasonal trends refer to variations that recur within specific periods, such as daily, weekly, monthly, or yearly[29]. These seasonal characteristics are typically influenced by natural or social cyclic factors, and such patterns can vary between variables within a multivariate system. Hyndman & Athanasopoulos (2018) emphasize that accurate modelling of seasonality is crucial for detecting outliers, forecasting, and preventing the misinterpretation of long-term trends[30]. During imputation, missing values in the seasonal or strong linear trend components can compromise data quality, so the imputation methods employed must be capable of preserving these patterns consistently[6].

2.2. Missing Value in Time Series Data

Missing values in time series data refer to the absence of one or more observations at specific time points within a chronologically ordered dataset. Each observation in time series has a distinct temporal position, and missing entries not only represent a loss of information but can also disrupt the continuity and temporal dependency structure essential for accurate modeling and forecasting[31]. Losing even a single data point can break this chain of information, leading to a distorted understanding of the underlying process[32]. The mechanism of data loss, as originally classified by Rubin (1976), includes three main categories: Missing Completely at Random (MCAR) and Missing Not at Random (MNAR), which continue to underpin modern studies on missing data in time-dependent systems[33], directly influences the choice of appropriate methods for handling missing data and the potential for bias in the analysis. There are two primary mechanisms of data loss[14]:

2.2.1. Missing Completely at Random (MCAR)

The mechanism of MCAR is that the likelihood of losing an observational value is consistent across all cases or units of observation. Most importantly, this probability does not depend on the actual value of the variable itself (whether observed or missing), nor on the values of other variables in the *dataset* [34]. Simply put, data loss under MCAR happens entirely at random, as if data is lost due to an "accident" unrelated to any information within the *dataset*[35]. This condition implies that a complete set of data (i.e., data without missing values) represents an unbiased random sample of the original dataset, which should be complete[36]. In the case of multivariate time series data, MCAR often appears as data points that disappear sporadically and irregularly at various times across different variables.

2.2.2. Missing Not at Random (MNAR)

Unlike MCAR, MNAR occurs when the chance of an observed value being missing directly depends on the value of the missing variable itself, even after accounting for all other observed variables in the dataset[37]. This means there is a systematic reason why the data is missing, and that reason is directly linked to the value that should be present. MNAR is the most complex and

challenging data loss scenario to handle statistically because key information about the data missingness mechanism is contained within the unobserved values[38]. Ignoring MNAR can lead to serious bias in parameter estimates and research conclusions. In the context of multivariate time series data, MNAR often appears as blocks of data missing sequentially or as systematic stopping points.

2.3. Imputation Method

2.3.1. Linear Interpolation

Linear Interpolation is a straightforward technique that fills in missing values by assuming a linear relationship between two adjacent known data points[36]. For the observation vector $Y_t \in \mathbb{R}^D$ at time t, if the i-th component Y_{t_k}, namely y_i, t_k, is missing, and y_i, t_{k-1} and y_i, t_{k+1} are known, then:

$$y_{i,t_k} = y_{i,t_{k-1}} + \frac{y_{i,t_{k+1}} - y_{i,t_{k-1}}}{t_{k+1} - t_{k-1}}(t_k - t_{k-1}) \quad (1)$$

for each variable (component *i*) in the multivariate time series data separately, the missing y_i, t_k identification. Find the two closest observed data points before (y_i, t_{k-1} on t_{k-1}) and after (y_i, t_{k+1} on t_{k+1}) the missing value. Apply $y_{i,t_k} = y_{i,t_{k-1}} + \frac{y_{i,t_{k+1}} - y_{i,t_{k-1}}}{t_{k+1} - t_{k-1}}(t_k - t_{k-1})$ to calculate the estimate y_{i,t_k}. Replace the missing value y_{i,t_k} with the estimated value[39].

2.3.2. Kalman Filtering

Kalman Filtering models a multivariate time series $y_t \in \mathbb{R}^p$ as an error projection of a latent variable $x_t \in \mathbb{R}^m$[34]. The model can be expressed in linear-Gaussian form[40]:

$$x_t = F_t x_{t-1} + w_t; w_t \sim N(0, Q), \quad y_t = Hx_t + v_t; v_t \sim N(0, R) \quad (2)$$

where,

$$Q = Var(\Delta x_t), \quad R = Var(\Delta x_t); t < t_{missing} \quad (3)$$

with prediction steps [41],

$$\hat{x}_{(t|t-1)} = F\hat{x}_{(t-1|t-1)}, \quad P_{(t|t-1)} = FP_{(t-1|t-1)}F^T + Q \quad (4)$$

and updates [34],

$$K_t = P_{(t|t-1)}H_t^T \left(HP_{(t|t-1)}H_t^T + R_t\right)^{-1}, \quad (5)$$
$$\hat{x}_{(t|t)} = \hat{x}_{(t|t-1)} + K_t\left(y_t - H\hat{x}_{(t|t-1)}\right),$$
$$P_{(t|t)} = (1 - K_t)P_{(t|t-1)}$$

The recursive minimum-variance estimator was first published by Kalman and was thoroughly reformulated for the p, m arbiter case by Durbin & Koopman. The core of the Kalman Filter's imputation mechanism lies in how it responds to the absence of new observations. In the prediction-update cycle, the update step (measurement update) entirely depends on the availability of a new measurement vector y_t. When the observation y_t is unavailable at time *t*, indicating that data is missing, the update step cannot be performed[41]. When the measurement y_t is missing and the update step is skipped, the filter relies entirely on the prediction step to continue. The 'best' estimate of the state at time *k*, with information available up to *k*, is simply the result of the prediction made at the previous step. Mathematically, this means[41]:

- A posteriori state estimation is the same as a priori estimation: x̂t + 1$\hat{x}t + 1\hat{x}_{t|t-1}$
- The covariance of the a posteriori error is equal to the covariance of the a priori: $P_{t|t} = P_{t|t-1}$

Along with that, uncertainty continues to spread and grow. Since there is no new information from measurements to reduce the uncertainty, the error covariance matrix continues to increase

according to the dynamic model and process noise[42]:

$$P_{t+1|t} = F_{t+1}P_{t|t}F_{t+1}^{T} + Q_{t+1} = F_{t+1}P_{t|t-1}F_{t+1}^{T} + Q_{t+1} \quad (6)$$

This behaviour intuitively makes a lot of sense: the longer no new data is received from the real world, the more uncertain one becomes about the true state of the system. The filter quantitatively captures this increase in uncertainty within the P matrix[41].

2.3.3. Support Vector Regression (SVR)

SVR is a non-linear regression algorithm that is not inherently designed for sequential data. Therefore, to use it in time series imputation, the problem must first be transformed into a supervised regression format[43]. Data transformation with lagged values can convert time series data into a feature-target format[16]. For each data point y_t to be predicted (imputed), a series of previous observations ($y_{t-1}, y_{t-2}, \ldots, y_{t-p}$ is used as predictor features[34]. The value p is the number of lagged observations chosen. This process effectively transforms the univariate time series problem into a multivariate regression problem where SVR can learn the relationships between past values and the current value[34].

The SVR model is trained exclusively on complete data segments (without missing values). This model learns a non-linear function that maps feature vectors (lagged values) to the target value (current value)[44]. The main advantage of SVR at this stage is its ability to handle non-linear relationships through the kernel trick (for example, with the RBF kernel)[36]. Estimation for SVR:

$$f(x) = w^{T}\varphi(x) + b; \min\frac{1}{2}||w||^{2} + C\sum(\alpha_i + \alpha_i^{*}) \quad (7)$$

Once the SVR model is trained, it is used to predict missing values. For each missing data point at time t, a feature vector made up of the available past values is fed into the trained SVR model. The output of this model is the estimated value for the missing point[45].

2.3.4. Recurrent Neural Network Gated Recurrent Unit (RNN-GRU)

Missing data imputation with RNN-GRU is a deep learning technique that utilises the ability of GRU to identify temporal patterns in time series[37]. This approach works by establishing a link between previous observation values and the missing data through a hidden state mechanism[38]. GRU employs two main gates[1]:

Update gate:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (8)$$

Reset gate:

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (9)$$

Then the state candidates are counted:

$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \quad (10)$$

and the final hidden state:

$$h_t = (1 - z_t) \odot h_t - 1 + z_t \odot \tilde{h}_t \quad (11)$$

The prediction output to fill in missing data is obtained from:

$$\hat{y}_t = Vh_t + b_y \quad (12)$$

2.4. Evaluation Matrix

Several evaluation matrices commonly used in comparative studies are Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2)[38]. These

metrics can be used to assess the quality of predictions both individually for each variable and overall within a multivariate framework[38]. A low RMSE value indicates more accurate imputation results and better model performance[38]. The formulas for RMSE, MAE, R^2 are as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (13)$$

$$MAE = \frac{1}{n}\sum_{i-1}^{n}|y_i - \hat{y}_i| \quad (14)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (15)$$

In the context of multivariate three-matrix evaluation, this can be calculated for each variable, and then the values can be averaged to obtain an overall estimate of the quality of imputation across the dataset.

3. Research Method

This study is simulative and aims to assess the performance of various imputation methods for multivariate time series data with missing values. The process involves systematically generating synthetic data, introducing missing data based on specific mechanisms, applying imputation techniques from classical methods, machine learning, and deep learning, and evaluating the imputation outcomes using quantitative metrics such as Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2).

3.1. Synthetic Data Generation

The data used in this study were generated simultaneously using two main scenarios, namely the linear trend scenario and the seasonal scenario. Each scenario produces multivariate time series data with three variables ($p = 3$) that are interrelated. Correlations between variables are established through the Cholesky decomposition of a predefined correlation matrix.

3.1.1. Trend Linier Scenario

In this scenario, each variable follows the following model:

$$\boldsymbol{y_{t,i}} = \boldsymbol{\beta_i} \cdot \boldsymbol{t} + \boldsymbol{\varepsilon_{t,i}} \quad (16)$$

with $\boldsymbol{\beta_i} \sim \boldsymbol{U(0, 20, 0, 30)}$, and $\boldsymbol{\varepsilon_{t,i}}$ are components of multivariate Gaussian noise with correlation between variables[15].

3.1.2. Seasonal Scenario

For seasonal patterns, data is generated using the model:

$$y_{t,i} = \beta_i \cdot t + A_i \cdot \sin\left(\frac{2\pi t}{period} + \phi_i\right) + \varepsilon_{t,i} \quad (17)$$

with the parameter $\boldsymbol{\beta_i}$, $\boldsymbol{A_i}$, $\boldsymbol{\phi_i}$, and is generated from uniform distribution. The components are $\boldsymbol{\varepsilon_{t,i}}$ the same multivariate Gaussian noise as in linear scenarios [46]. Simulations were carried out

on three data size scenarios, namely small: 50 observations, medium: 200 observations and large: 2000 observations[15].

3.2. Missing Value Mechanism

After the data is generated, missing values are inserted based on two mechanisms. MCAR (Missing Completely at Random) occurs when missing values are randomly inserted without a specific pattern, while MNAR (Missing Not at Random) happens when missing values occur sequentially, similar to dropout patterns in observational data. Each mechanism is applied at three levels of missingness: Low: 10%, Medium: 20%, and High: 35%[47].

3.3. Imputation Method Used

After the missing data was inserted, an imputation process was carried out using three groups of methods, namely traditional methods[31], machine learning, and deep learning[46].

3.3.1. Traditional Method

- Linear Interpolation: Filling gaps with data points interpolated between available values.
- Kalman Filtering: A state-space model method for smoothing and predicting values.

3.3.2. Machine Learning Method

- Support Vector Regression (SVR): A regression technique that maximises the margin with non-linear features.

3.3.3. Deep Learning Method

- Recurrent Neural Network (RNN): A deep learning model that accounts for temporal dependence, utilising a GRU (Gated Recurrent Unit) architecture for autoregressive imputation.

3.4. Work Evaluation Matrix

To assess the accuracy of the imputation results, the imputed values are compared with the original values using three evaluation metrics[11]:

- Root Mean Square Error (RMSE): Measures the average squared error and is sensitive to outliers.
- Mean Absolute Percentage Error (MAPE): Shows the error as a relative percentage.
- Coefficient of Determination (R^2): Shows the proportion of variability in the original data that can be explained by the imputation results.

The evaluation was performed with 100 repetitions per scenario to ensure a stable and reliable performance estimate.

3.5. Flowchart

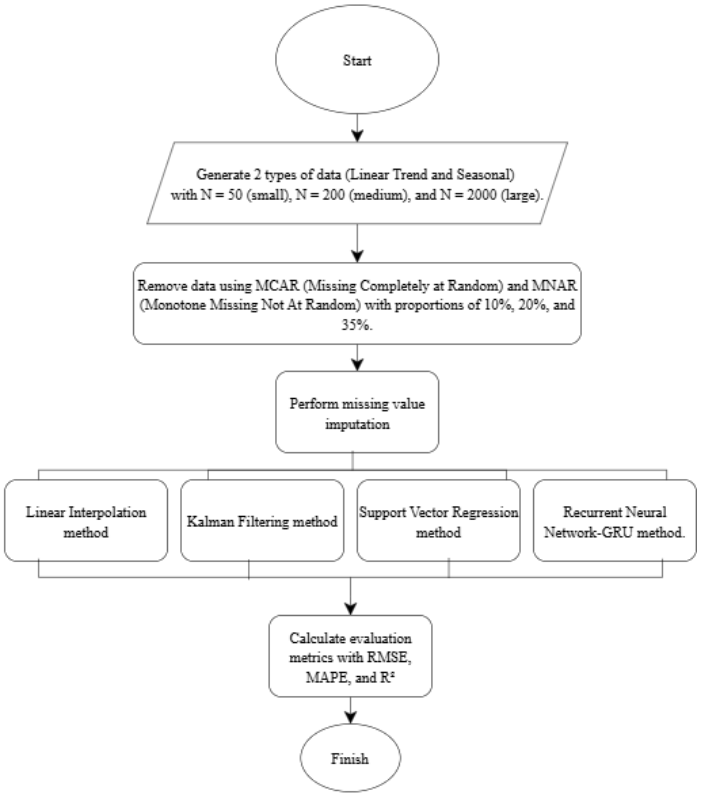


Figure 1. Research flowchart.

4. Result and Discussion

The data used in this study were generated synthetically using two time series generation functions: make_linear for linear trends and make_seasonal for seasonal trends. Below is a plot of an example of the simulated data generated.

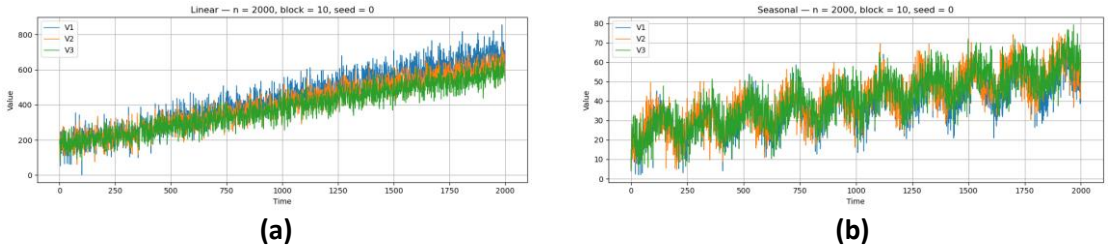


Figure 2. (a) Multivariate Linear Trend Time Series Data (b). Seasonal Multivariate Trend Time Series Data

Once the data is generated, a missing value mechanism is applied with two patterns: random or MCAR - Missing Completely at Random, and sequential or MNAR - Missing Not At Random. The following is a plot illustrating the missing value scenario that has been implemented.

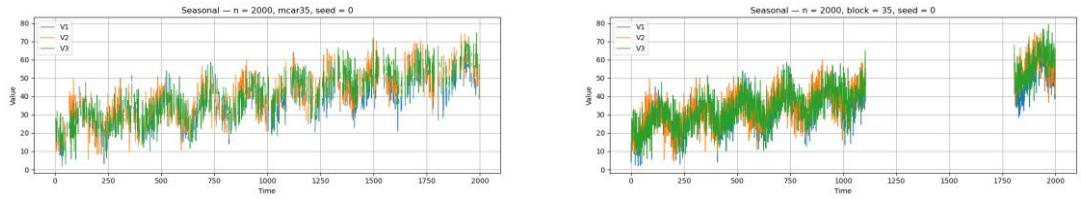


Figure 3. (a) MCAR - Missing Completely at Random, (b) MNAR - Missing Not at Random.

To address the missing values that have been simulated, an imputation process was carried out using four methods commonly used in multivariate time series: Linear Interpolation, Kalman filtering, Support Vector Regression (SVR), and Recurrent Neural Network (RNN-GRU). To obtain reliable results and evaluate the consistency of each method's performance, simulations were run 150 times for each scenario. The evaluation matrix resulting from the imputation process on data with a linear trend pattern and MNAR missing values is shown below.

Table 1. Evaluation metrics for linear data with missing values MNAR.

n	frac	method	rmse_mean	mape_mean	r2_mean
50	0,1	kalman	0,555584	0,017884	0,976204
		linear	0,500531	0,016656	0,982204
		rnn	0,492156	0,016256	0,981659
		svr	0,493473	0,018599	0,980556
	0,2	kalman	1,068379	0,04445	0,91673
		linear	0,74555	0,033013	0,960925
		rnn	0,823771	0,035824	0,946557
		svr	0,769532	0,037447	0,949325
	0,35	kalman	1,995947	0,106264	0,72029
		linear	1,198017	0,069476	0,901376
		rnn	1,91853	0,102366	0,73878
		svr	1,266868	0,081785	0,847274
200	0,1	kalman	2,113286	0,017606	0,980454
		linear	1,974875	0,018283	0,983204
		rnn	2,017423	0,018207	0,982194
		svr	1,862842	0,019701	0,984022
	0,2	kalman	3,80504	0,041951	0,937825
		linear	2,943096	0,036014	0,963654
		rnn	3,561868	0,041416	0,943464
		svr	2,895983	0,041302	0,957709
	0,35	kalman	7,316928	0,100983	0,774011
		linear	4,758655	0,075505	0,904749
		rnn	7,677025	0,11389	0,721641
		svr	4,681705	0,085747	0,874089
2000	0,1	kalman	19,25152	0,014693	0,984698
		linear	19,84746	0,016814	0,983331
		rnn	20,04409	0,016567	0,98268

n	frac	method	rmse_mean	mape_mean	r2_mean
		svr	18,52669	0,017763	0,984653
	0,2	kalman	36,06903	0,035704	0,946313
		linear	30,44653	0,034404	0,961356
		rnn	38,38198	0,041471	0,932435
		svr	28,80188	0,037117	0,959355
	0,35	kalman	69,40682	0,086899	0,801808
		linear	48,04023	0,069021	0,90379
		rnn	69,00949	0,09283	0,781293
		svr	46,03265	0,076311	0,880686

Based on the evaluation results of linear trend data imputation with a Missing Not at Random (MNAR) pattern, it was found that the effectiveness of the imputation method is highly influenced by the data size and the proportion of missing values. For small data sets, n=50, RNN-GRU is recommended when missing data is minimal, as it produces the smallest error, whereas Linear Interpolation is more stable and accurate when missing data is moderate to high. For medium data sets, n=200, linear and SVR perform very well with minimal missing data as they can capture trend patterns, while for moderate to high missing data, Linear Interpolation remains the most reliable method, with SVR as an alternative when higher RMSE precision is required. For large data sets, n=2000, Kalman and SVR are highly effective with minimal missing data, but their performance declines as missing data increases. In this condition, Linear Interpolation again becomes the most consistent method for handling moderate to high missing data. Overall, Linear Interpolation emerges as the most stable method across all scenarios, with SVR as a strong contender, offering competitive performance. Kalman excels with low missing data, and RNN-GRU is only recommended for small datasets with low missing proportions. Below is the diagram of the simulation results, which were conducted 150 times.

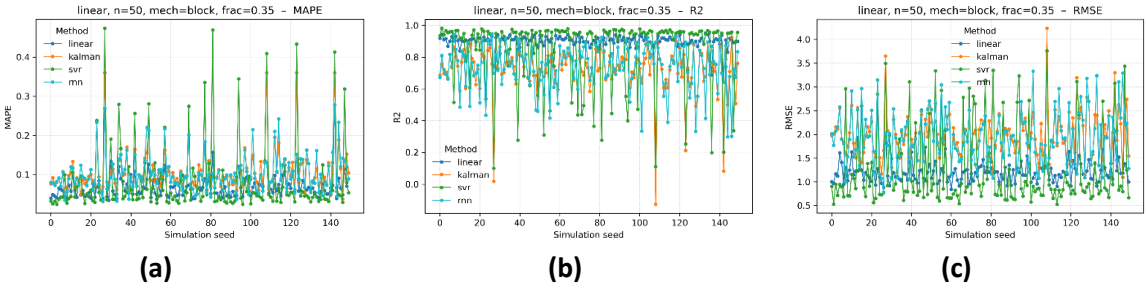


Figure 4. Evaluation metrics for linear data with n = 50 and missing values MNAR using Linear Interpolation, Kalman, SVR, and RNN Methods (a) MAPE, (b) R-Squared, (c) RMSE.

The simulation diagram images above show how the imputation method performs across different data types, all within the same model structure. The three evaluation diagrams (a) MAPE, (b) R^2, and (c) RMSE demonstrate the consistency and reliability of each method when handling data variations generated through the simulation seed.

Then, based on the evaluation results of linear trend data imputation with a Missing Completely at Random (MCAR) pattern, for small data sets (n = 50), the Linear Interpolation method showed the most stable and accurate performance across all levels of missing data (10%, 20%, 35%) with the lowest RMSE and MAPE values and the highest R^2. It was followed by SVR as an alternative with high precision. For medium-sized data (n = 200), the Kalman Filter consistently outperformed others, especially when missing data was minimal to substantial, due to its ability to effectively capture temporal patterns, as indicated by low RMSE and high R^2. For large data sets (n = 2000), the Kalman Filter

remained the most effective method at all levels of missing data, maintaining the lowest RMSE and MAPE and the highest R^2 compared to other methods. Overall, Linear Interpolation is recommended for small data sets because of its stability, while the Kalman Filter is the primary choice for medium to large data sets due to its ability to more accurately recognise temporal dynamics under various missing data conditions. The following diagram illustrates the results of the simulation process for imputing linear trend data with a Missing Completely at Random (MCAR) pattern.

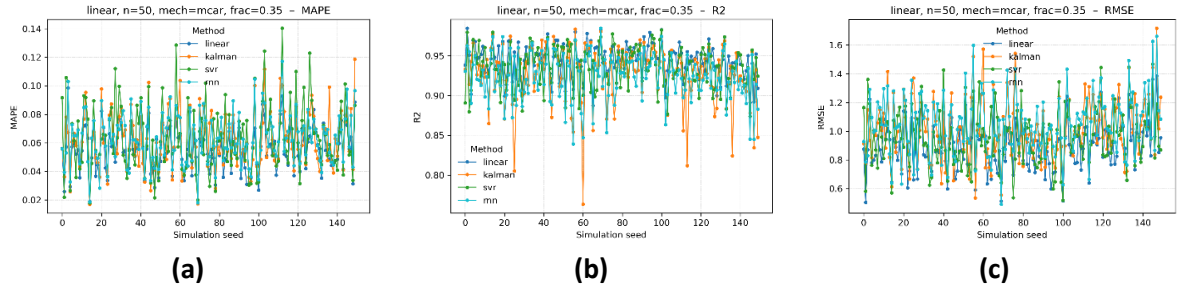


Figure 5. Evaluation metrics for linear data with n = 50 and missing values MCAR using Linear, Kalman, SVR, and RNN Methods (a) MAPE, (b) R-Squared, (c) RMSE.

Next, the evaluation results of the seasonal linear trend data with a Missing Not at Random (MNAR) pattern show that for small data sets (n = 50), RNN-GRU provides the best results when missing data is minimal (10%), with the lowest RMSE and MAPE and the highest R^2. However, its performance declines significantly as missing data increases, making Linear and Kalman more stable choices for moderate to large levels of missing data. For medium-sized data (n = 200), RNN-GRU and SVR perform best when the proportion of missing data is low (10%), with smaller RMSE and MAPE and the highest R^2 (around 0.95), making them the most accurate methods for this scenario. When missing data rises to 20%, SVR remains superior with the best performance in terms of precision (RMSE = 4.20 and R^2 = 0.89), while other methods, such as Linear and Kalman, show a decline in performance. At high missing levels (35%), SVR again demonstrates relatively better performance compared to other methods, although overall accuracy decreases, with an R^2 of 0.79. Therefore, for medium-sized data, SVR is the most consistent and superior method across various levels of missing data, with RNN-GRU being the best option only at low missing levels. For large data sets (n = 2000), SVR consistently remains the most accurate method at all levels of missing data, with the lowest RMSE and MAPE and the highest R^2, indicating its effectiveness in managing complex seasonal patterns. Overall, SVR is recommended for large and medium data sets, while RNN-GRU is advisable for small and medium data with low missing data.

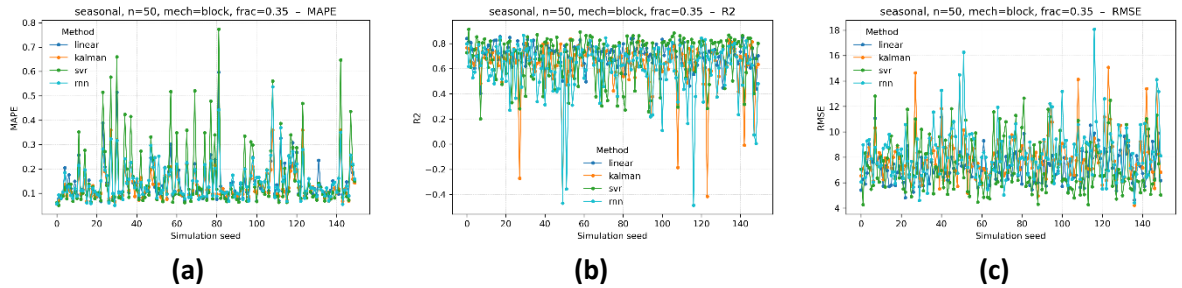


Figure 6. Evaluation metrics for seasonal data with n = 50 and missing values MNAR using Linear, Kalman, SVR, and RNN Methods (a) MAPE, (b) R-Squared, (c) RMSE.

The final imputation method for seasonal trend data with a missing pattern of MCAR on small datasets (n = 50) is Linear Interpolation, which emerges as the best method across all levels of missingness, with the lowest RMSE and MAPE values and the highest R^2. Meanwhile, the Kalman Filter ranks second with consistent performance. For medium-sized datasets (n = 200), a similar trend is observed, where Linear and Kalman continue to perform well, maintaining R^2 above 0.90 at 20% missingness and remaining above 0.82 at 35% missingness. In large datasets (n = 2000), the Kalman Filter prevails and demonstrates the greatest stability; for example, at 10% missingness, it records RMSE = 2.29 and R^2 = 0.969, surpassing other methods. Although RNN-GRU and SVR remain reasonably strong (R^2 > 0.84 at high missingness), both tend to decline sharply as the proportion of missing data increases. Overall, Linear Interpolation and the Kalman Filter are the most dependable options for handling completely random missing data in seasonal datasets, while RNN-GRU appears less robust in managing such patterns.

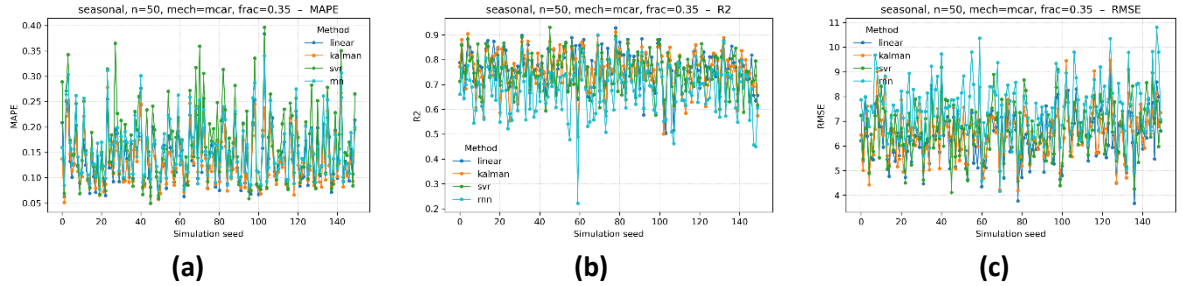
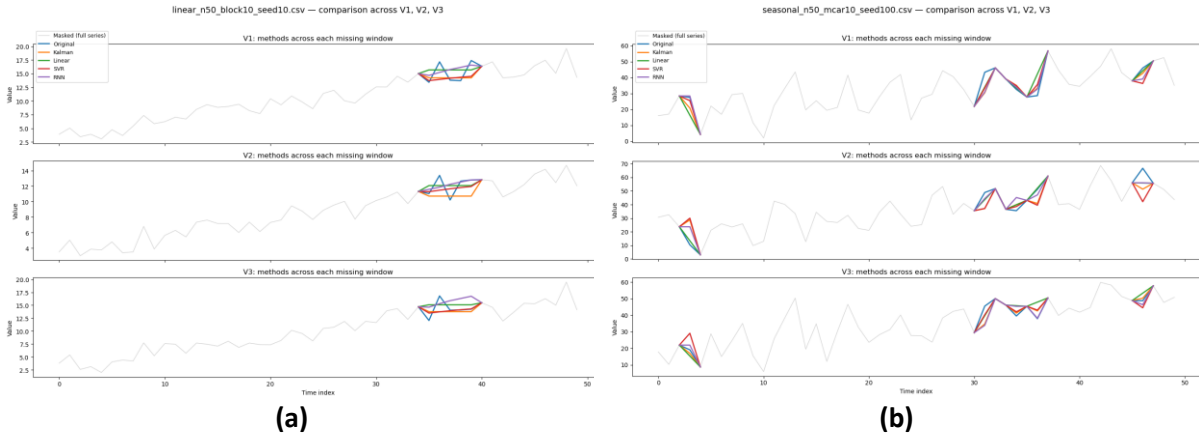


Figure 7. Evaluation metrics for seasonal data with n = 50 and missing values MCAR using Linear, Kalman, SVR, and RNN Methods (a) MAPE, (b) R-Squared, (c) RMSE.

The following presents the results of visualising missing data imputation, both for linear trend and seasonal patterns, with MCAR and MNAR patterns, to provide an overview of how the four imputation methods replace missing values.



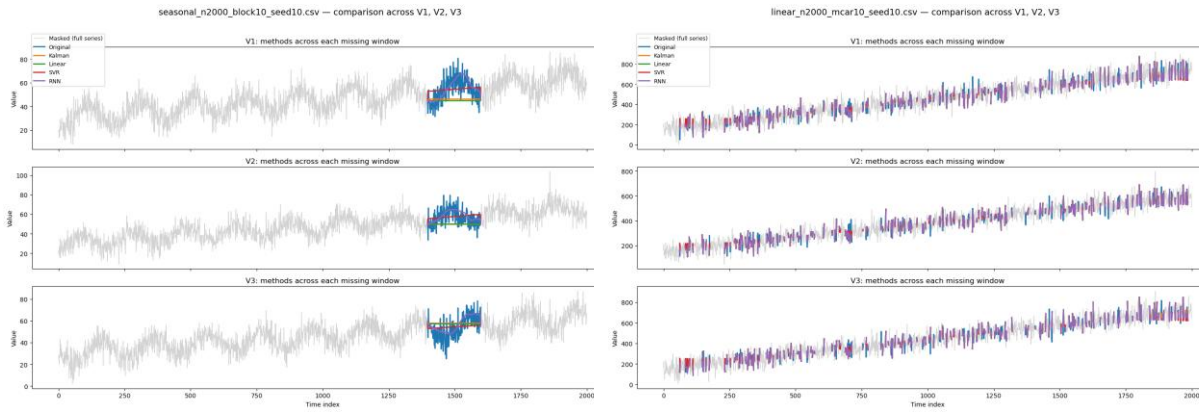


Figure 8. Results of data imputation visualisation (a) Linear MNAR 50 data, (b) Linear MCAR 50 data, (c) Seasonal MNAR 2000 data, (d) Linear MCAR 2000 data.

Figure 8 shows the visualisation of missing data imputation results across four different scenarios: (a) Linear Trend MNAR with 50 data points, (b) Seasonal Trend MCAR with 50 data points, (c) Seasonal Trend MNAR with 2000 data points, and (d) Linear Trend MCAR with 2000 data points. In each graph, the blue line indicates where data has been deliberately removed (missing), while the other coloured lines represent imputation results from four methods: Linear Interpolation, Kalman Filter, SVR, and RNN-GRU. The grey line shows the original complete data before missing values, serving as a reference to assess the accuracy of the imputation. The closer the imputed data line is to the original data line, the more accurate the method is in filling in the missing values. This visualisation offers a clear overview of how each method adapts to different data characteristics and missing data patterns. Additionally, these visual results have been validated through quantitative analysis using evaluation metrics such as RMSE, MAPE, and R^2, discussed earlier. The validation indicates that the performance of the imputation methods depicted aligns with the evaluation calculations, supporting conclusions about the appropriateness and suitability of each method in different contexts.

5. Conclusion

This study assesses the performance of imputation methods: Linear Interpolation, Kalman, SVR, and RNN-GRU, in managing missing data in synthetic multivariate time series with two trend patterns (linear and seasonal) and two missing mechanisms (MCAR and MNAR). It considers three data sizes (n = 50, 200, 2000) and three missing proportion levels (10%, 20%, 35%). To offer a more systematic and concise overview of each method's effectiveness under various condition combinations, Table 2 is provided below, summarising the best imputation methods based on data scenarios, sample sizes, and data loss levels.

Table 2. Summary of the best imputation methods based on data scenarios, sample size, and missing proportion.

Data Scenario	n	*Missing* (10%)	*Missing (20%)*	*Missing* (35%)	Results
Linear Trend - MCAR	50	Linear	Linear	Linear	Highly stable linear suitable at all levels missing
	200	Kalman	Kalman	Kalman	Kalman excels at capturing temporal patterns
	2000	Kalman	Kalman	Kalman	Kalman is dominant in big data with random misses

Data Scenario	n	*Missing* (10%)	*Missing (20%)*	*Missing* (35%)	Results
Linear	50	RNN	Linear	Linear	RNN excels in low misses
Trend -	200	Linear	Linear	Linear	Linear stable at all levels missing
MNAR	2000	Kalman	Linear	Linear	Linear remains consistent in large misses
Seasonal	50	RNN	Kalman	Linear	RNN only excels in low missing.
Trend -	200	RNN	SVR	SVR	RNN decreases as missing increases
MNAR	2000	SVR	SVR	SVR	SVR consistently excels on all missing
Seasonal	50	Linear	Linear	Linear	Linear stable, Kalman is quite competitive
Trend -	200	Linear	Linear	Kalman	Kalman excels in high misses
MCAR	2000	Kalman	Kalman	Kalman	Kalman is very stable in all types of missing

Overall, the simulation results indicate that no method is completely superior in all situations. The choice of the best imputation method largely depends on the data's characteristics, the missing data mechanism, and the level of missingness.

References

[1] F. Wang, Y. Jiang, R. Zhang, A. Wei, J. Xie, and X. Pang, "A Survey of Deep Anomaly Detection in Multivariate Time Series: Taxonomy, Applications, and Directions," *Sensors*, vol. 25, no. 1, p. 190, Jan. 2025, doi: 10.3390/S25010190.

[2] P. Wang, X. He, H. Feng, and G. Zhang, "A Multivariate Short-Term Trend Information-Based Time Series Forecasting Algorithm for PM2.5 Daily Concentration Prediction," *Sustainability*, vol. 15, no. 23, p. 16264, Nov. 2023, doi: 10.3390/SU152316264.

[3] U. Ahmed, J. C. W. Lin, and G. Srivastava, "Multivariate time-series sensor vital sign forecasting of cardiovascular and chronic respiratory diseases," *Sustain. Comput. Informatics Syst.*, vol. 38, p. 100868, Apr. 2023, doi: 10.1016/J.SUSCOM.2023.100868.

[4] V. Papastefanopoulos, P. Linardatos, T. Panagiotakopoulos, and S. Kotsiantis, "Multivariate Time-Series Forecasting: A Review of Deep Learning Methods in Internet of Things Applications to Smart Cities," *Smart Cities*, vol. 6, no. 5, pp. 2519–2552, Sep. 2023, doi: 10.3390/SMARTCITIES6050114.

[5] X. Wang, H. Liu, J. Du, X. Dong, and Z. Yang, "A long-term multivariate time series forecasting network combining series decomposition and convolutional neural networks," *Appl. Soft Comput.*, vol. 139, p. 110214, May 2023, doi: 10.1016/J.ASOC.2023.110214.

[6] A. K. Tripathi, P. K. Gupta, H. Saini, and G. Rathee, "MVI and Forecast Precision Upgrade of Time Series Precipitation Information for Ubiquitous Computing," *Inform.*, vol. 47, no. 5, pp. 83–94, 2023, doi: 10.31449/INF.V47I5.4152.

[7] S. Lin, X. Wu, G. Martinez, and N. V Chawla, "Filling Missing Values on Wearable-Sensory Time Series Data," in *Proceedings of the 2020 SIAM International Conference on Data Mining (SDM 2020)*, 2020, pp. 46–54, doi: 10.1137/1.9781611976236.6.

[8] M. Shi and S. Mollah, "NeTOIF: A Network-based Approach for Time-Series Omics Data Imputation and Forecasting," 2021, doi: 10.1101/2021.06.05.447209.

[9] A. S. AlSalehy and M. Bailey, "Improving Time Series Data Quality: Identifying Outliers and Handling Missing Values in a Multilocation Gas and Weather Dataset," *Smart Cities*, vol. 8, no. 3, p. 82, May 2025, doi: 10.3390/SMARTCITIES8030082.

[10] Z. Magyari-Sáska, I. Haidu, and A. Magyari-Sáska, "Experimental Comparative Study on Self-Imputation Methods and Their Quality Assessment for Monthly River Flow Data with Gaps: Case Study to Mures River," *Appl. Sci.*, vol. 15, no. 3, Feb. 2025, doi: 10.3390/APP15031242.

[11] M. W. Heymans and J. W. R. Twisk, "Handling missing data in clinical research," *J. Clin. Epidemiol.*, vol. 151, pp. 185–188, Nov. 2022, doi: 10.1016/J.JCLINEPI.2022.08.016.

[12] K. J. Lee, J. B. Carlin, J. A. Simpson, and M. Moreno-Betancur, "Assumptions and analysis planning in studies with missing data in multiple variables: moving beyond the MCAR/MAR/MNAR classification," *Int. J. Epidemiol.*, vol. 52, no. 4, p. 1268, Aug. 2023, doi: 10.1093/IJE/DYAD008.

[13] A. A. Mir, K. J. Kearfott, F. V. Çelebi, and M. Rafique, "Imputation by feature importance (IBFI): A methodology to envelop machine learning method for imputing missing patterns in time series data," *PLoS One*, vol. 17, no. 1, p. e0262131, Jan. 2022, doi: 10.1371/JOURNAL.PONE.0262131.

[14] G. Chhabra, "Comparison of Imputation Methods for Univariate Time Series," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 2s, pp. 2321–8169, 2023, doi: 10.17762/ijritcc.v11i2s.6148.

[15] N. A. Zamri, M. I. Jaya, I. D. Irawati, T. H. Rassem, Rasyidah, and S. Kasim, "Comparative Analysis of Imputation Methods for Enhancing Predictive Accuracy in Data Models," *Int. J. Informatics Vis.*, vol. 8, no. 3, pp. 1271–1276, 2024, doi: 10.62527/JOIV.8.3.1666.

[16] L. R. Slipetz, A. Falk, and T. R. Henry, "Missing Data in Discrete Time State-Space Modeling of Ecological Momentary Assessment Data: A Monte-Carlo Study of Imputation Methods," Jan. 2023, doi: 10.1080/00273171.2025.2469055.
[17] A. O.A., O. O.C., and A. S., "Kalman Filter Algorithm versus Other Methods of Estimating Missing Values: Time Series Evidence," *African J. Math. Stat. Stud.*, vol. 4, no. 2, pp. 1–9, May 2021, doi: 10.52589/AJMSS-VFVNMQLX.
[18] S. A. Rahman, Y. Huang, J. Claassen, and S. Kleinberg, "Imputation of Missing Values in Time Series with Lagged Correlations," 2015.
[19] F. Schlembach, E. Smirnov, I. Koprinska, and M. H. M. Winands, "Conformal multistep-ahead multivariate time-series forecasting," *Mach. Learn.*, vol. 114, no. 7, pp. 1–51, Jul. 2025, doi: 10.1007/S10994-024-06722-9/FIGURES/40.
[20] Y. Shu and V. Lampos, "DeformTime: Capturing Variable Dependencies with Deformable Attention for Time Series Forecasting," Jun. 2024, Accessed: Oct. 11, 2025. [Online]. Available: https://arxiv.org/pdf/2406.07438.
[21] C. Zeng, Z. Liu, G. Zheng, L. Kong, and S. Jiao, "CMamba: Channel Correlation Enhanced State Space Models for Multivariate Time Series Forecasting," Jun. 2024, Accessed: Oct. 11, 2025. [Online]. Available: https://arxiv.org/pdf/2406.05316.
[22] M. Löf and P. Stockhammar, "Staff memo Indicators for short-term fore-casting," 2024.
[23] D. Franjic and K. Schweikert, "Predictor Preselection for Mixed-Frequency Dynamic Factor Models: A Simulation Study With an Empirical Application to GDP Nowcasting," *J. Forecast.*, vol. 44, no. 2, pp. 255–269, Mar. 2025, doi: 10.1002/FOR.3193.
[24] Z. Zhang, S. Ren, X. Qian, and N. Duffield, "Learning Flexible Time-windowed Granger Causality Integrating Heterogeneous Interventional Time Series Data," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 1, pp. 4408–4418, Aug. 2024, doi: 10.1145/3637528.3672023.
[25] X. Chu, W. Cui, S. Xu, L. Zhao, H. Guan, and Y. Ge, "Multiscale Time Series Decomposition for Structural Dynamic Properties: Long-Term Trend and Ambient Interference," *Struct. Control Heal. Monit.*, vol. 2023, no. 1, p. 6485040, Jan. 2023, doi: 10.1155/2023/6485040.
[26] L. Jendges and J. M. Brockmann, "Refining linear trend estimates from one dimensional time series data with autoregressive covariance modelling: an application to GRACE total water storage time series data," *Stoch. Environ. Res. Risk Assess.*, vol. 39, no. 9, pp. 3813–3825, Jun. 2025, doi: 10.1007/S00477-025-03038-5/FIGURES/7.
[27] C. Cao, R. Debnath, and R. M. Alvarez, "Physics-based deep learning reveals rising heating demand heightens air pollution in Norwegian cities," May 2024, Accessed: Oct. 11, 2025. [Online]. Available: https://arxiv.org/pdf/2405.04716.
[28] C. D. Le, H. V. Pham, D. A. Pham, A. D. Le, and H. B. Vo, "A PM2.5 concentration prediction framework with vehicle tracking system: From cause to effect," Dec. 2022, Accessed: Oct. 11, 2025. [Online]. Available: https://arxiv.org/pdf/2212.01761.
[29] A. Zainuddin, M. A. Hairuddin, A. I. M. Yassin, Z. I. A. Latiff, and A. Azhar, "Time Series Data and Recent Imputation Techniques for Missing Data: A Review," in *2022 International Conference on Green Energy, Computing and Sustainable Technology, GECOST 2022*, 2022, pp. 346–350, doi: 10.1109/GECOST55694.2022.10010499.
[30] "Forecasting: Principles and Practice (2nd ed)," 2018. https://otexts.com/fpp2/ (accessed Oct. 12, 2025).
[31] N. Niako, J. D. Melgarejo, G. E. Maestre, and K. P. Vatcheva, "Effects of missing data imputation methods on univariate blood pressure time series data analysis and forecasting with ARIMA and LSTM," *BMC Med. Res. Methodol.*, vol. 24, no. 1, pp. 1–32, Dec. 2024, doi: 10.1186/S12874-024-02448-3/FIGURES/26.
[32] S. Goel *et al.*, "An Enhanced Integrated Method for Healthcare Data Classification with Incompleteness," 2024, doi: 10.32604/cmc.2024.054476.
[33] A. A. Toye, A. Celik, and S. Kleinberg, "Benchmarking Missing Data Imputation Methods for Time Series Using Real-World Test Cases," in *Proceedings of Machine Learning Research (2025) 287*, Jul. 2025, pp. 480–501, Accessed: Oct. 11, 2025. [Online]. Available: https://proceedings.mlr.press/v287/toye25a.html.
[34] A. Becker, *Kilman filter : from the ground up*. KilmanFilter.NET, 2023.
[35] A. R. Alsaber, J. Pan, and A. Al-Hurban, "Handling Complex Missing Data Using Random Forest Approach for an Air Quality Monitoring Dataset: A Case Study of Kuwait Environmental Data (2012 to 2018)," *Int. J. Environ. Res. Public Heal. 2021, Vol. 18, Page 1333*, vol. 18, no. 3, p. 1333, Feb. 2021, doi: 10.3390/IJERPH18031333.
[36] J. Park *et al.*, "Long-term missing value imputation for time series data using deep neural networks," *Neural Comput. Appl.*, vol. 35, no. 12, pp. 9071–9091, Apr. 2023, doi: 10.1007/S00521-022-08165-6.
[37] C. Shoko and C. Sigauke, "Short-term forecasting of COVID-19 using support vector regression: An application using Zimbabwean data," *Am. J. Infect. Control*, vol. 51, no. 10, pp. 1095–1107, Oct. 2023, doi: 10.1016/J.AJIC.2023.03.010.
[38] A. Flores, H. Tito-Chura, O. Cuentas-Toledo, V. Yana-Mamani, and D. Centty-Villafuerte, "PM2.5 Time Series Imputation with Moving Averages, Smoothing, and Linear Interpolation," *Computers*, vol. 13, no. 12, p. 312, Nov. 2024, doi: 10.3390/COMPUTERS13120312.
[39] Y. Gao, M. Taie Semiromi, and C. Merz, "Efficacy of statistical algorithms in imputing missing data of streamflow discharge imparted with variegated variances and seasonalities," *Environ. Earth Sci.*, vol. 82, no. 20, pp. 1–25, Oct. 2023, doi: 10.1007/S12665-023-11139-Z/FIGURES/16.
[40] "Perbandingan Metode LSTM dan BiLSTM untuk Prediksi Data Curah Hujan."

https://www.researchgate.net/publication/391449921_Perbandingan_Metode_LSTM_dan_BiLSTM_untuk_Prediksi_Data_Curah_Hujan (accessed Jun. 13, 2025).
[41] T. Lacey, “Tutorial: The Kalman Filter,” pp. 133–140.
[42] A. Chhabra, J. R. Venepally, and D. Kim, “Measurement Noise Covariance-Adapting Kalman Filters for Varying Sensor Noise Situations,” *Sensors*, vol. 21, no. 24, p. 8304, Dec. 2021, doi: 10.3390/S21248304.
[43] L. R. Slipetz, A. Falk, and T. R. Henry, “Missing Data in Discrete Time State-Space Modeling of Ecological Momentary Assessment Data: A Monte-Carlo Study of Imputation Methods,” Jan. 2023, doi: 10.1080/00273171.2025.2469055.
[44] C. Linroth, “Statistical analysis of wave heights using Kalman Filtering methods,” 2014.
[45] X. Chen *et al.*, “Multi-Task Data Imputation for Time-Series Forecasting in Turbomachinery Health Prognostics,” *Machines*, vol. 11, no. 1, p. 18, Dec. 2022, doi: 10.3390/MACHINES11010018.
[46] Z. R. Firdhani, “Penerapan Metode Seasonally Decomposed Missing Value Imputation pada Pemodelan Hybrid Machine Learning untuk Peramalan Kualitas Udara di Kota Surabaya,” Feb. 2023.
[47] S. Hassankhani Dolatabadi, I. Budinská, R. Behmaneshpour, and E. Gatial, “Closing the Data Gap: A Comparative Study of Missing Value Imputation Algorithms in Time Series Datasets,” pp. 77–90, 2024, doi: 10.1007/978-3-031-53552-9_7.