JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 9, No. 1, January 2025, pp. 161-174

Optimizing Long-Term Meteorological Data Completeness in North Aceh, Indonesia: A Comparative Analysis of Interpolation Methods

Novi Reandy Sasmita[1*], Novita Sari Saragih[1], Latifah Rahayu[1], Malfirah[1]
[1]Department of Statistics, Universitas Syiah Kuala, Indonesia
novireandys@usk.ac.id

Article History:
Received : 17-11-2024
Revised : 25-12-2024
Accepted : 28-12-2024
Online : 03-01-2025

Keywords:
Interpolation;
Meteorology Data;
Missing Data;
North Aceh Regency;
Long-term Data.

ABSTRACT

More data in meteorological records is needed to ensure the accuracy of meteorological modeling, particularly in long-term datasets. This study aims to identify the most effective interpolation method for addressing missing data in North Aceh's meteorological dataset from 2010 to 2023, with a focus on the accuracy of methods applied across various meteorological variables. The study analyzed data from North Aceh Regency, Indonesia, comprising 25,565 daily observations of temperature, humidity, rainfall, sunshine duration, and wind speed. Missing values were interpolated using three methods: spline, stineman, and moving average interpolation. Performance was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Squared Logarithmic Error (MSLE) across 10%, 20%, and 30% levels of simulated missing data. All analysis in this study were carried out using R-4.4.2 software. While spline interpolation performed reasonably well, it showed increased variability, especially for high-variance variables like rainfall. Moving average interpolation was less reliable, with error rates increasing alongside higher levels of missing data. In contrast, stineman interpolation consistently achieved the lowest error metrics across all levels of missing data, with MAE ranging from 0.219 to 0.6691, MSLE from 0.035 to 0.109, and RMSE from 1.247 to 2.245, demonstrating superior robustness. Stineman interpolation offers a highly effective approach for managing missing meteorological data in North Aceh's long-term dataset, enhancing data reliability for meteorological modeling and decision-making in meteorological-sensitive sectors. This study provides practical recommendations for selecting optimal interpolation techniques, especially in regions with variable meteorological data quality.

https://doi.org/10.31764/jtam.v9i1.27929

A. INTRODUCTION

Interpolation is a fundamental process in data analysis used to estimate unknown values that fall within the range of known data points. This technique is crucial for ensuring the continuity and completeness of datasets, which is particularly important in fields such as meteorology, where data gaps can significantly affect the interpretation of trends and patterns (Marchi et al., 2020). Reliable meteorology models and predictions depend critically on the completeness and quality of the input data. Therefore, addressing missing data through effective interpolation methods is vital for producing accurate and reliable meteorology analyses (More & Wolkersdorfer, 2024). Meteorological data often contain missing values due to various factors, including equipment malfunctions, sensor failures, and human errors during

data collection (Tierney & Cook, 2023). Variables such as air temperature, humidity, rainfall, sunshine duration, and wind speed are particularly prone to these issues. Incomplete datasets reduce meteorological prediction accuracy, impacting key sectors such as agriculture, transportation, and tourism, which rely heavily on precise meteorology information. Applying robust interpolation methods to fill in missing data is crucial for preserving the integrity and usability of meteorological datasets. Neglecting or inappropriately handling missing data can introduce biases in statistical analyses, emphasizing the importance of robust interpolation methods in meteorological studies (Jones et al., 2023).

Previous studies have explored various interpolation methods, identifying their respective strengths and limitations. Spline interpolation is widely used in meteorological data analysis due to its ability to create smooth curves that accurately represent the underlying data trends (Wicher-Dysarz et al., 2022). This method is particularly effective in scenarios where data points are unevenly distributed, as it minimizes oscillations between points and maintains continuity in both the function and its first derivative. A study has demonstrated that spline interpolation can effectively handle missing data in various applications, including temperature and precipitation modeling (Azizan et al., 2018). Another study on rainfall data fitting using spline interpolation highlighted its capability to produce accurate and reliable estimates, even when faced with incomplete datasets (Yasmin et al., 2024).

Next, stineman interpolation is another valuable technique for addressing missing data in meteorological contexts. This method is designed to preserve the shape of the data, which is crucial when dealing with environmental variables that exhibit specific trends or patterns (Bleidorn et al., 2022). While specific studies on stineman interpolation in meteorology are limited, its general properties suggest that it can yield satisfactory results in maintaining the integrity of the data's form while filling in gaps (Wu et al., 2018). For example, in studies involving temperature variations, stineman interpolation has been employed to effectively reconstruct missing values without introducing significant artifacts or distortions in the data. This shape-preserving property makes stineman interpolation particularly advantageous in meteorology, where the preservation of trends is essential for accurate forecasting (Bleidorn et al., 2022; Horváth et al., 2023).

Furthermore, moving averages interpolation is frequently applied in meteorological analysis to reduce short-term fluctuations and emphasize longer-term trends (Drozd et al., 2022; Leirvik & Yuan, 2021; Portela et al., 2020). This technique involves calculating the average of a subset of data points, which can help mitigate the effects of noise and outliers in the data. In the context of missing data, moving averages can serve as a straightforward imputation method, providing a simple yet effective way to estimate missing values based on surrounding data points. So, moving averages remain a popular choice for preliminary data analysis and trend identification in meteorological studies (Kim et al., 2019). This study addresses the challenge of selecting the optimal interpolation method for accurately filling in missing meteorological data. The general solution proposed involves systematically comparing different interpolation techniques namely spline, stineman, and moving average interpolation, to determine their relative performance across various meteorological variables. This approach is used to identify the methods that provide the most accurate estimates, thereby improving the overall quality of meteorological data.

Many reviews of closely related literature reveal that while significant progress has been made in understanding the performance of various interpolation methods, more research is still needed, specifically analyzing long-term meteorological data (Haldar et al., 2023; Zhou et al., 2022). Previous Studies provide valuable insights but are limited by their focus on short-term data. This study addresses this gap by conducting a detailed comparison of interpolation methods with spline, stineman, and moving average interpolation on a more than decade-long meteorological dataset, thus providing a more robust framework for future meteorological data analysis.

This research aims to enhance meteorological data quality by determining the most effective interpolation method to fill in missing data using several meteorological variables. The novelty of this study lies in the comprehensive comparison of several interpolation techniques over a period from 2010 to 2023 of the meteorological data of North Aceh Regency, Aceh Province, Indonesia. North Aceh, selected as the study site, is Aceh Province's largest rice producer, making accurate meteorological data essential for optimizing agricultural productivity. The region's varied meteorology patterns and significance in food production underscore the importance of robust data interpolation methods. This approach not only advances our understanding of interpolation methods but also provides practical guidance to improve the integrity of meteorological data, ultimately supporting better decision-making in various sectors.

B. METHODS

1. The Study Area

North Aceh, a significant regency in Aceh Province, Indonesia, is the largest rice-producing area in the Aceh province, making accurate meteorological data critical for agricultural productivity. In 2022, North Aceh Regency had 614,640 population with a population density of 186.43 per km^2. Most of the people of North Aceh work in agriculture. The North Aceh regency is selected by purposive sampling due to its role as the largest rice producer in Aceh Province, with a total production in 2023 of 238,088 tonnes (BPS-Statistics Indonesia, 2023).

2. Data Sources and Variables

The dataset used in this study was sourced from the official BMKG website (www.dataonline.bmkg.go.id) and includes daily meteorological data for North Aceh Regency from 1 January 2010 to 31 December 2023. The study variables encompass temperature (°C), humidity (%), rainfall (mm), sunshine duration (hours), and wind speed (m/s), where each variable had 5,113 observations, totaling 25,565 observations. The data is in the form of time series and falls into the Missing Completely at Random (MCAR) category, where missing data is randomly distributed across variables and not associated with other variables (Heymans & Twisk, 2022).

3. Interpolation Methods

Interpolation serves as a computational technique for generating approximate values across a dataset (Mahmoud et al., 2021). Interpolation results may vary depending on the algorithm used (Pratama & Sam'an, 2023). Some interpolation methods that are compared in this study are spline interpolation, stineman interpolation, and moving average interpolation. Spline interpolation suggests that for $n + 1$ data points, there will be a polynomial interpolation of the value of a function in the interval of data points. A known function (x) in the interval $a \leq x \leq b$ is approximated by another function, $g(x)$ by partitioning the interval $a \leq x \leq b$ into several sub-intervals $a = x_1 < x_2 < \cdots < x_n = b$. The function $g(x)$ obtained is called a spline (Segeth, 2018). Here is the equation used in this method, with $i = 1,2,3, \ldots, n$.

$$y = y_i + \frac{y_{i+1}-y_i}{x_{i+1}-x_i}(x - x_i) + b_1(x - x_i)(x - x_{i+1}) + b_2(x - x_i)(x - x_{i+1})(x - x_{i+2}) \quad (1)$$

where is

$$b_1 = \frac{\left(\frac{y_{i+2}-y_{i+1}}{x_{i+2}-x_{i+1}}\right)-\left(\frac{y_{i+1}-y_i}{x_{i+1}-x_i}\right)}{x_{i+2}-x_i} \quad (2)$$

$$b_2 = \frac{\left[\frac{\left(\frac{y_{i+3}-y_{i+1}}{x_{i+3}-x_{i+2}}\right)-\left(\frac{y_{i+2}-y_{i+1}}{x_{i+1}-x_i}\right)}{x_{i+3}-x_i}\right]-\left[\frac{\left(\frac{y_{i+n+1}-y_{i+n}}{x_{i+2}-x_{i+1}}\right)-\left(\frac{y_{i+n}-y_i}{x_{i+1}-x_i}\right)}{x_{i+3}-x_i}\right]}{x_{i+3}-x_i} \quad (3)$$

Next, stineman interpolation is an interpolation method used to replace missing values with intersecting rational interpolation (Arjasakusuma et al., 2020). The following is the equation used in this method, with $i = 1,2,3, \ldots, n$. If $\Delta y_i, \Delta y_{i+1} > 0$, then the equation is used.

$$y = y_0 + \frac{\Delta y_i \Delta y_{i+1}}{\Delta y_i + \Delta y_{i+1}} \quad (4)$$

If $\Delta y_i, \Delta y_{i+1} < 0$ then the equation is used

$$y = y_0 + \frac{\Delta y_i \Delta y_{i+1}(x - x_i + x - x_{i+1})}{(\Delta y_i + \Delta y_{i+1})(x_{i+1} - x_i)} \quad (5)$$

where is

$$\Delta y_i = y_i + S_i(x - x_i) - y_0 \quad (6)$$

$$\Delta y_{i+1} = y_{i+1} + S_i(x - x_{i+1}) - y_0 \quad (7)$$

$$S_i = \frac{x_{i+1} - x_i}{y_{i+1} - y_i} \quad (8)$$

The last interpolation method is moving average interpolation, which replaces missing data at time t with the average of the k observations before time t (Mohamad et al., 2022). The formula for this interpolation method can be seen in the equation below, with $i = 1,2,3, \ldots, n$.

$$y = \frac{(y_i + y_{i-2} + \cdots + y_0)}{k} \tag{9}$$

4. Performance Evaluation Metrics

Performance evaluation metrics are employed to assess each method's effectiveness in achieving the desired interpolation accuracy. The evaluation metrics used in this study are Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Squared Logarithmic Error (MSLE). The MAE , RMSE, and MSLE formulas are presented in in equation 10-12 respectively, as follows (Bruce et al., 2020).

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{11}$$

$$MSLE = \frac{1}{n}\sum_{t=1}^{n-1}(\log(1 + y_t) - \log(1 + \hat{y}_t))^2 \tag{12}$$

5. Stage of Data Analysis

The data analysis in this study involves six steps. First, descriptive statistics, including minimum, mean, maximum, interquartile range, and standard deviation, are computed to assess data distribution and establish baseline data. Missing values are also identified for each variable. This analysis helps identify patterns and trends (Azharuddin et al., 2023; Sasmita et al., 2023). Second, the baseline data that still contain missing values will be processed using each interpolation technique, namely spline, stineman, and moving average. The interpolation methods were selected for their theoretical and practical relevance to meteorological data. The interpolated datasets produced by each technique serve as the actual data for this study. The actual data will be analyzed using descriptive statistics for tendency central and dispersion analysis.

In the third step, data is randomly deleted with three models of data loss percentage, namely 10%, 20%, and 30% of the total actual data. The deletion of data with various percentages aims to test the robustness and reliability of the interpolation method used as well as to simulate real-world data missingness (Qiu et al., 2024). The fourth step involves re-interpolating the data after data deletion for each percentage of data loss, using the same equations as in the second step, to assess the response of each method to different levels of data incompleteness. At this stage, the interpolated data becomes the prediction data in this study.

In the fifth step, performance evaluation metrics are calculated using MAE, RMSE, and MSLE by comparing actual data and predicted data for each variable at various percentages of data that have been interpolated using each interpolation method. In the sixth step, the average values of MAE, RMSE, and MSLE are calculated from the comparison results of the five variables

used with the same percentage of missing data for each interpolation method. The selection of performance metrics aligns with international standards for data imputation accuracy, focusing on both absolute and proportional errors. Finally, This study prioritizes selecting the optimal interpolation technique by identifying the method with the lowest average MAE, RMSE, and MSLE (Jierula et al., 2021; Putri et al., 2023). Data analysis was conducted using R-4.4.2 software for replicability and precision. R software is powerful due to its extensive statistical tools, efficient data handling, customizable functions, reproducibility, and advanced visualization capabilities (Kruba et al., 2024; Rahayu et al., 2023; Ulhaq et al., 2024). Further, R is open-source nature software, making it ideal for complex and transparent data analysis (Saputra et al., 2024).

C. RESULT AND DISCUSSION

1. Descriptive Statistics for Baseline Data

Descriptive statistics offer a concise summary of key characteristics in North Aceh's meteorological data from 2010-2023, as shown in Table 1. Table 1 presents the summary statistics of the baseline data. Summary statistics reveal distinct trends and distribution patterns across meteorological variables. Temperature exhibited a stable trend, with values from 23.3°C to 30.3°C and an average of 26.7°C, supported by a low standard deviation of 0.911, indicating minimal fluctuation.

Table 1. Summary statistics of baseline data

No	Variable	Min	Mean	Max	IQR	Stdv	Missing Value
1	Temperature	23,30	26,70	30,30	1,20	0,911	38
2	Humidity	57,00	83,99	100,00	6,00	4,315	41
3	Rainfall	0,00	5,60	181,70	4,20	14,220	1.484
4	Sunshine duration	0,00	5,57	11,70	4,80	3,007	281
5	Wind speed	0,00	2,01	5,00	0,00	0,740	33

Humidity showed greater variability, ranging from 57.0% to 100.0% and a mean of 83.99%, with a standard deviation of 4.315, reflecting moderate variability. Rainfall exhibits the most notable trend, characterized by significant variability with values from 0.0 mm to 181.7 mm, a mean of 5.6 mm, and a substantial standard deviation of 14.22, suggesting occasional extreme values. The length of Sunshine has a mean of 5.57 hours, varying from 0.0 to 11.7 hours, with a moderate standard deviation of 3.007, indicating a broad distribution. Wind speed is the most consistent variable, with a mean of 2.01 m/s, minimal variability (standard deviation of 0.74), and an IQR of 0.0, highlighting uniformity in the central distribution. Furthermore, From the table, it can be seen that all meteorological data variables in this study are missing. The variable with the most missing data is the rainfall variable with 1,484 observations and the smallest is the temperature variable with 33 observations.

2. Descriptive Statistics for Actual Data

The baseline meteorological data that previously contained missing values was filled in through each interpolation method. After the interpolation process was completed, descriptive analysis was performed on the actual data, as shown in Table 2.

Table 2. Summary statistics of actual data

No.	Methods	Variable	Min	Mean	Max	IQR	Stdv
1	Spline interpolation	Temperature	23,30	26,70	30,30	1,20	0,912
		Humidity	57,00	84,00	100,00	6,00	4,306
		Rainfall	0,00	45,30	838,50	11,40	152,510
		Sunshine duration	0,00	5,59	14,12	4,80	3,026
		Wind speed	0,00	2,01	5,00	0,00	0,742
2	Stineman interpolation	Temperature	23,30	26,70	30,30	1,20	0,909
		Humidity	57,00	83,99	100,00	6,00	4,303
		Rainfall	0,00	6.59	181,70	7,81	14,172
		Sunshine duration	0,00	5,56	11,70	4,70	2,980
		Wind speed	0,00	2,01	5,00	0,00	0,738
3	Moving average interpolation	Temperature	23,30	26,70	30,30	1,20	0,909
		Humidity	57,00	84,00	100,00	6,00	4,304
		Rainfall	0,00	5,99	181,70	5,70	14,321
		Sunshine duration	0,00	5,56	11,70	4,70	2,960
		Wind speed	0,00	2,01	5,00	0,00	0,738

The data shows that temperature and wind speed exhibit consistent results across all interpolation methods, with identical minimum, mean, and maximum values and minor differences in standard deviation, underscoring their stable distribution. Humidity also maintains a consistent pattern, with marginal differences in standard deviation and IQR across methods. Rainfall displayed significant variability, particularly under spline interpolation, which has a mean of 45.3 mm and a very high standard deviation of 152.51, suggesting the presence of significant outliers or extreme values. Stineman and moving average interpolation methods show reduced variability for rainfall, with means of 6.59 mm and 5.99 mm, respectively, and lower standard deviations (14.172 and 14.321), indicating a more controlled distribution. Sunshine duration demonstrated consistent means across all interpolation methods, but spline interpolation has the highest variability (Stdv of 3.026) while moving average interpolation shows a slightly more stable spread (Stdv of 2.96).

3. Descriptive Statistics for Actual Data Based on Percentage of Missing Data

The removal of some observations based on a percentage of the actual data aims to test the robustness and reliability of the interpolation method. The percentage variations used in data deletion are 10% or 511 observations; 20% or 1,023 observations; and 30% or 1,534 observations on each variable from 5,113 observations and are done randomly. After data deletion, a descriptive analysis of the deleted data was conducted, as shown in Table 3.

Table 3. Summary statistics for actual data based on percentage of missing data

No	Dataset From Methods	Percentage	Variable	Min	Mean	Max	IQR	Stdv
1	Spline interpolation	10%	Temperature	23,30	26,70	30,30	4,20	0,916
			Humidity	58,00	84,00	100,00	19,00	4,297
			Rainfall	0,00	45,78	838,52	838,52	153,147
			Sunshine duration	0,00	5,55	14,12	11,02	3,029
			Wind speed	0,00	2,00	5,00	3,00	0,743
		20%	Temperature	23,30	26,71	30,30	4,20	0,915
			Humidity	59,00	83,99	100,00	19,00	4,297
			Rainfall	0,00	45,90	838,52	838,52	152,901
			Sunshine duration	0,00	5,56	14,12	10,92	3,025
			Wind speed	0,00	2,01	5,00	3,00	0,744
		30%	Temperature	23,30	26,71	30,30	4,20	0,914
			Humidity	59,00	83,97	100,00	19,00	4,291
			Rainfall	0,00	47,81	838,52	838,52	157,299
			Sunshine duration	0,00	5,59	14,12	10,92	3,011
			Wind speed	0,00	2,01	5,00	3,00	0,744
2	Stineman interpolation	10%	Temperature	23,30	26,70	30,30	4,20	0,913
			Humidity	58,00	84,00	100,00	19,00	4,293
			Rainfall	0,00	6,49	181,70	181,70	13,863
			Sunshine duration	0,00	5,52	11,70	8,60	2,984
			Wind speed	0,00	2,00	5,00	3,00	0,739
		20%	Temperature	23,30	26,71	30,30	4,20	0,912
			Humidity	59,00	83,98	100,00	19,00	4,293
			Rainfall	0,00	6,48	181,70	181,70	13,807
			Sunshine duration	0,00	5,53	11,70	8,50	2,981
			Wind speed	0,00	2,01	5,00	3,00	0,740
		30%	Temperature	23,30	26,71	30,30	4,20	0,912
			Humidity	59,00	83,97	100,00	19,00	4,286
			Rainfall	0,00	6,58	181,70	181,70	13,988
			Sunshine duration	0,00	5,56	11,70	8,50	2,969
			Wind speed	0,00	2,01	5,00	3,00	0,740
3	Moving average interpolation	10%	Temperature	23,30	26,70	30,30	4,20	0,913
			Humidity	58,00	84,00	100,00	19,00	4,293
			Rainfall	0,00	6,49	181,70	181,70	13,863
			Sunshine duration	0,00	5,52	11,70	8,60	2,984
			Wind speed	0,00	2,00	5,00	3,00	0,739
		20%	Temperature	23.30	26.71	30.30	4.20	0,912
			Humidity	59.00	83.98	100.00	19.00	4,293
			Rainfall	0.00	6.48	181.70	181.70	13,807
			Sunshine duration	0.00	5.53	11.70	8.50	2,981
			Wind speed	0.00	2.01	5.00	3.00	0,740
		30%	Temperature	23.30	26.71	30.30	4.20	0,912
			Humidity	59.00	83.97	100.00	19.00	4,286
			Rainfall	0.00	6.58	181.70	181.70	13,988
			Sunshine duration	0.00	5.56	11.70	8.50	2,969
			Wind speed	0.00	2.01	5.00	3.00	0,740

The summary statistics in Table 3 show that temperature and humidity remain stable across the percentage removal for each data set used. Temperature's minimum (23.30), maximum (30.30), and mean (26.70 to 26.71) show minimal change, while the interquartile range (IQR) and standard deviation also remain nearly constant, indicating low sensitivity to data removal. Similarly, humidity maintains its range (58.00 to 100.00) with a slight mean drop from 84.00 to 83.97 and a standard deviation from 4.297 to 4.286, suggesting reliability under missing data conditions. However, rainfall shows sensitivity, especially in the dataset from the spline interpolation method, where the mean rises from 45.78 at 10% deletion to 47.81 at 30%, with fluctuating standard deviations. This variability indicates that rainfall's central tendency and spread may shift with missing data, unlike the more consistent sunshine duration and wind speed. Sunshine duration's mean only increases slightly from 5.55 to 5.59, while wind speed remains consistent across all deletion levels, demonstrating resilience to data removal.

4. Descriptive Statistics for Prediction Data Based on Percentage of Missing Data

After some observations have been removed with a predetermined percentage variation, the next step is to interpolate the data that has been removed. This interpolation process is carried out for each variable using different interpolation methods. The result of the interpolation produces prediction data, which aims to measure the response of each method to variations in data incompleteness. Next, descriptive statistical analysis was applied to the prediction data to identify patterns and trends in each variable, as in Table 4.

Table 4. Summary statistics of prediction data

No	Handling Missing Data with Methods	Percentage	Variable	Min	Mean	Max	IQR	Stdv
1	Spline interpolation	10%	Temperature	23,30	26,70	30,30	1,20	0,922
			Humidity	58,00	84,01	100,00	6,00	4,345
			Rainfall	0,00	45,57	838,52	12,00	152,455
			Sunshine duration	0,00	5,57	14,12	4,86	3,033
			Wind speed	0,00	2,01	5,36	0,00	0,747
		20%	Temperature	22,62	26,70	30,30	1,21	0,935
			Humidity	59,00	84,02	105,07	6,00	4,446
			Rainfall	0,00	45,72	838,52	12,44	152,410
			Sunshine duration	0,00	5,60	19,46	4,80	3,055
			Wind speed	0,00	2,00	5,02	0,00	0,753
		30%	Temperature	22,57	26,71	30,30	1,30	0,948
			Humidity	59,00	84,00	105,30	6,00	4,518
			Rainfall	0,00	45,92	838,52	12,40	152,374
			Sunshine duration	0,00	5,65	15,85	4,74	3,056
			Wind speed	0,00	2,02	6,74	0,03	0,767
2	Stineman interpolation	10%	Temperature	23,30	26,70	30,30	1,20	0,905
			Humidity	58,00	84,00	100,00	6,00	4,253
			Rainfall	0,00	6,62	181,70	8,00	13,937
			Sunshine duration	0,00	5,53	11,70	4,70	2,932
			Wind speed	0,00	2,01	5,00	0,00	0,726
		20%	Temperature	23,30	26,70	30,30	1,20	0,894
			Humidity	59,00	84,00	100,00	6,00	4,216
			Rainfall	0,00	6,55	181,70	8,23	13,524

No	Handling Missing Data with Methods	Percentage	Variable	Min	Mean	Max	IQR	Stdv
			Sunshine duration	0,00	5,54	11,70	4,50	2,884
			Wind speed	0,00	2,00	5,00	0,00	0,711
		30%	Temperature	23,30	26,71	30,30	1,20	0,889
			Humidity	59,00	83,99	100,00	5,26	4,141
			Rainfall	0,00	6,54	181,70	8,12	13,383
			Sunshine duration	0,00	5,57	11,70	4,40	2,828
			Wind speed	0,00	2,01	5,00	0,00	0,700
3	Moving average interpolation	10%	Temperature	23,30	26,70	30,30	1,20	0,897
			Humidity	58,00	84,00	100,00	6,00	4,205
			Rainfall	0,00	6,58	181,70	8,02	13,770
			Sunshine duration	0,00	5,53	11,70	4,50	2,893
			Wind speed	0,00	2,00	5,00	0,00	0,716
		20%	Temperature	23,30	26,71	30,30	1,20	0,879
			Humidity	59,00	83,99	100,00	5,07	4,124
			Rainfall	0,00	6,55	181,70	8,30	13,281
			Sunshine duration	0,00	5,54	11,70	4,30	2,806
			Wind speed	0,00	2,00	5,00	0,00	0,693
		30%	Temperature	23,30	26,71	30,30	1,19	0,868
			Humidity	59,00	83,98	100,00	4,71	4,021
			Rainfall	0,00	6,55	181,70	8,25	13,058
			Sunshine duration	0,00	5,57	11,70	4,08	2,716
			Wind speed	0,00	2,01	5,00	0,00	0,672

In Table 4, Distinct effects of each interpolation method on variable predictions were observed, particularly as missing data levels increased. For temperature, the mean remains steady at around 26.70 across all methods, but spline interpolation shows a gradual increase in Standard Deviation (0.922 to 0.948) as missing data rises, suggesting slightly added variability with this method. Humidity maintains a consistent mean (83.98 to 84.02) across methods, though spline interpolation slightly raises Stdv from 4.345 (10% missing data) to 4.518 (30%). Further, rainfall sees significant differences. Spline Interpolation produces a much higher Mean (45.57 to 45.92) and Stdv (~152.4), while stineman and moving average keep the mean low (around 6.55) with a low Stdv (~13.5), indicating spline interpolation may amplify variance in high-variability data like rainfall. Sunshine duration shows that spline interpolation introduces more spread, with a Stdv of 3.033 versus stineman and moving average (~2.716 to 2.932). Wind speed remains consistent across methods, with minimal changes in mean (2.00 to 2.02) and Stdv (0.672 to 0.767), demonstrating that any interpolation method performs well for this stable variable. In summary, spline interpolation may elevate variability, especially in datasets with high variance data like rainfall. In contrast, stineman and moving average interpolation preserve data consistency across missing data levels, particularly for high-variance variables.

6. Evaluation of Performance Metrics

The evaluation of performance metrics for the three methods compared in this study is illustrated in Figure 1. Figure 1 highlights the differences in accuracy between the spline, stineman, and moving average interpolation methods as missing data increases from 10% to 30% using Mean Absolute Error (MAE), Mean Squared Logarithmic Error (MSLE), and Root

Mean Squared Error (RMSE). Each metric serves as an error indicator, with lower values representing higher accuracy and better performance in handling missing data.

In terms of MAE, which measures the average magnitude of errors without considering their direction, the stineman interpolation consistently shows the lowest error across all percentages, starting at 0.219 at 10% deletion and reaching 0.6691 at 30%. Spline interpolation has slightly higher MAE values (from 0.2671 to 0.8507) but still performs considerably better than the moving average method, which exhibits the highest MAE, rising sharply from 0.4618 at 10% to 0.8897 at 30%. This pattern indicates that stineman Interpolation is the most effective at minimizing absolute error and maintaining accuracy even as missing data increases, as shown in Figure 1.

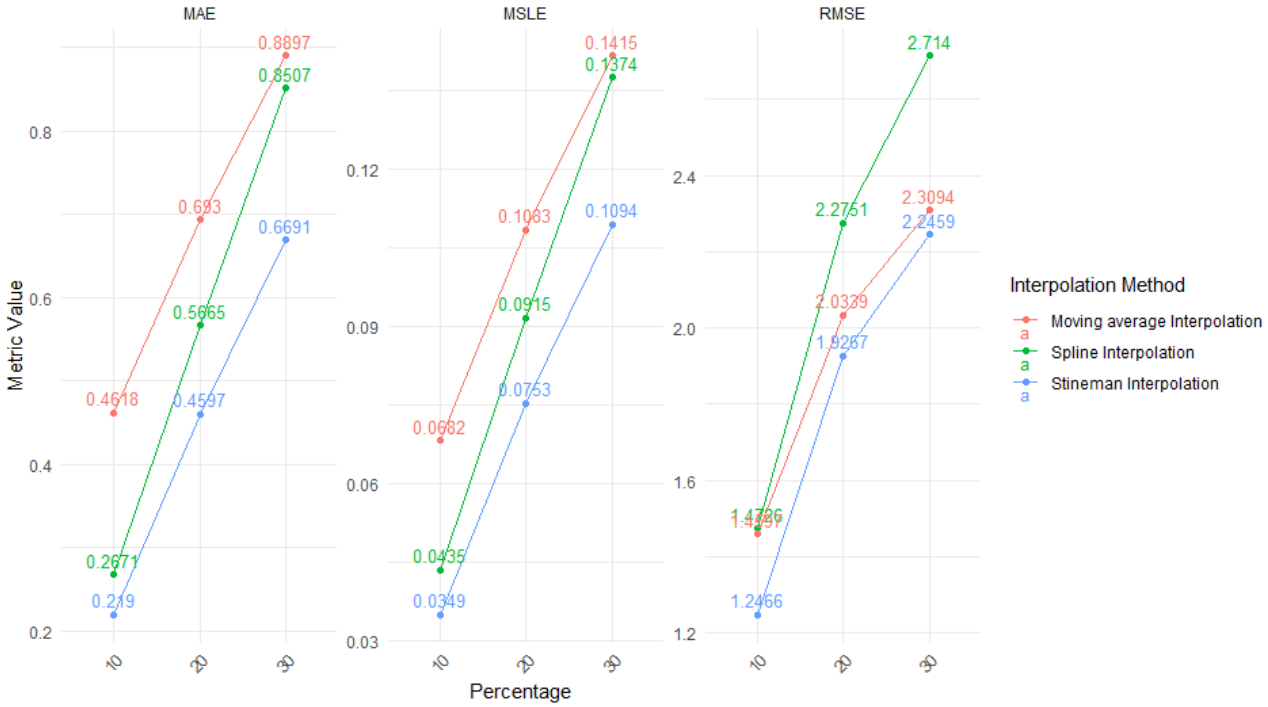

Figure 1. Performance Evaluation Metric by Interpopaltion Method and Percentage

MSLE provides additional insight by penalizing larger differences in logarithmic scale, which is useful for detecting proportional errors. Stineman interpolation again outperforms the other methods with the lowest MSLE values (0.0349 to 0.1094), indicating that it better maintains proportional accuracy as missing data increases. Spline Interpolation has higher MSLE values, ranging from 0.0435 at 10% to 0.1374 at 30% while moving average records the highest MSLE across all levels (0.0682 to 0.1415). It suggests that stineman interpolation is particularly effective at preserving relationships in data with exponential or multiplicative characteristics. This advantage becomes more pronounced as the amount of missing data increases. The RMSE metric, which emphasizes larger errors by squaring them, follows a similar trend. Stineman interpolation has the lowest RMSE values, rising from 1.2466 at 10% deletion to 2.2459 at 30%, showcasing its robustness in minimizing significant errors. Spline interpolation follows with RMSE values from 1.4796 to 2.714, while moving average interpolation shows the most significant RMSE increase, from 1.7336 at 10% to 2.3094 at 30%. These RMSE trends reinforce that stineman interpolation demonstrates superior handling of

extreme deviations, which is crucial when reconstructing datasets with substantial missing data.

From a comparative perspective, the results demonstrate a clear hierarchy in interpolation performance. Stineman interpolation consistently produces the lowest error across all metrics, suggesting it is the most resilient method for missing data among the three tested. Spline interpolation performs moderately well but is outperformed by stineman in all cases. Moving average interpolation, while simpler and computationally less intensive, shows a pronounced decline in accuracy as the percentage of missing data increases, indicating it is less suitable for situations with higher data loss. These findings hold important implications for fields requiring precise data interpolation amidst missing values. Stineman interpolation's ability to maintain low error values across all metrics suggests that it is well-suited for applications where data accuracy is critical and missing values are common. Its superior performance in all metrics implies that it can effectively reconstruct datasets with substantial data loss, preserving the integrity of both absolute and proportional relationships. Spline interpolation, while not as robust as stineman, may still serve as an acceptable alternative when missing data is moderate. In contrast, the moving average interpolation should be used cautiously, especially in high data-loss scenarios, due to its tendency to yield higher errors.

D. CONCLUSION AND SUGGESTIONS

This study evaluates the effectiveness of three interpolation methods, namely spline interpolation, stineman interpolation, and moving average interpolation at different levels of missing data (10%, 20%, and 30%), using three performance metrics: Mean Absolute Error (MAE), Mean Squared Logarithmic Error (MSLE), and Root Mean Square Error (RMSE). The findings show that spline interpolation performs reasonably well in terms of MAE and MSLE but exhibits higher RMSE values, suggesting that it may not be reliable in contexts sensitive to large deviations. Furthermore, Moving average interpolation, although computationally simpler, showed the highest overall error rate, indicating that this method is less suitable for high-precision tasks, especially in cases of significant data loss. Finally, stineman interpolation consistently outperformed the other methods, showing the lowest error values across all metrics at the missing data level. This robustness and precision make it well-suited for applications that demand high accuracy in data recovery, even as data sparsity increases. These results offer a practical framework for selecting interpolation techniques based on data characteristics and accuracy requirements, providing valuable guidance for researchers and practitioners seeking reliable solutions to missing data challenges. Future studies should assess these methods using advanced metrics like Relative Absolute Error (RAE), Residual Standard Error (RSE) or Squared Log Error (SLE) for greater accuracy insights, test their robustness under higher data loss (e.g., 40% or 50%), and explore their applicability to diverse data types such as time-series, spatial, or categorical datasets.

ACKNOWLEDGEMENT

The author would like to thank to Universitas Syiah Kuala funding for supporting this study and publication (Grant Number 509/UN11.2.1/PG.01.03/SPK/PTNBH/2024).

REFERENCES

Arjasakusuma, S., Pratama, A. P., & Lestari, I. (2020). Assessment of Gap-Filling Interpolation Methods for Identifying Mangrove Trends at Segara Anakan in 2015 by using Landsat 8 OLI and Proba-V. *Indonesian Journal of Geography*, *52*(3), 1–9. https://doi.org/10.22146/ijg.50556

Azharuddin, Sasmita, N. R., Idroes, G. M., Andid, R., Raihan, Fadlilah, T., Earlia, N., Ridwan, T., Maya, I., Farnida, & Idroes, R. (2023). Patient Satisfaction And Its Socio-Demographic Correlates In Zainoel Abidin Hospital, Indonesia: A Cross-Sectional Study. *Unnes Journal of Public Health*, *12*(2), 57–67. https://doi.org/10.15294/ujph.v12i2.69233

Azizan, I., Karim, S. A. B. A., & Suresh Kumar Raju, S. (2018). Fitting Rainfall Data by Using Cubic Spline Interpolation. *MATEC Web of Conferences*, *225*(1), 1–9. https://doi.org/10.1051/matecconf/201822505001

Bleidorn, M. T., Pinto, W. de P., Schmidt, I. M., Mendonça, A. S. F., & Reis, J. A. T. dos. (2022). Methodological approaches for imputing missing data into monthly flows series. *Revista Ambiente and Agua*, *17*(2), 1–27. https://doi.org/10.4136/ambi-agua.2795

BPS-Statistics Indonesia. (2023). *Aceh Province Gross Regional Domestic Product by Business Field Quarter 3 in 2023*.

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python* (2nd ed.). O'Reilly Media, Sevastopol.

Drozd, I. D., Gavrikov, A. V, & Stepanenko, V. M. (2022). Comparative characteristics of gap filling methods in high-frequency data of micrometeorological measurements. *IOP Conference Series: Earth and Environmental Science*, *1023*(1), 1–8. https://doi.org/10.1088/1755-1315/1023/1/012009

Haldar, S., Choudhury, M., Choudhury, S., & Samanta, P. (2023). Trend analysis of long-term meteorological data of a growing metropolitan city in the era of global climate change. *Total Environment Research Themes*, *7*(1), 1–12. https://doi.org/10.1016/j.totert.2023.100056

Heymans, M. W., & Twisk, J. W. R. (2022). Handling missing data in clinical research. *Journal of Clinical Epidemiology*, *151*(1), 185–188. https://doi.org/10.1016/j.jclinepi.2022.08.016

Horváth, S., Lukács, D., Farsang, E., & Horváth, K. (2023). Unbiased Determination of Adsorption Isotherms by Inverse Method in Liquid Chromatography. *Molecules*, *28*(3), 1–10. https://doi.org/10.3390/molecules28031031

Jierula, A., Wang, S., OH, T.-M., & Wang, P. (2021). Study on Accuracy Metrics for Evaluating the Predictions of Damage Locations in Deep Piles Using Artificial Neural Networks with Acoustic Emission Data. *Applied Sciences*, *11*(5), 1–20. https://doi.org/10.3390/app11052314

Jones, R. L., Kharb, A., & Tubeuf, S. (2023). The untold story of missing data in disaster research: a systematic review of the empirical literature utilising the Emergency Events Database (EM-DAT). *Environmental Research Letters*, *18*(10), 1–10. https://doi.org/10.1088/1748-9326/acfd42

Kim, T., Ko, W., & Kim, J. (2019). Analysis and Impact Evaluation of Missing Data Imputation in Day-ahead PV Generation Forecasting. *Applied Sciences*, *9*(204), 1–19. https://doi.org/10.3390/app9010204

Kruba, R., Mardalena, S., Rahayu, L., Mahdi, S., Sasmita, N. R., Kesuma, Z. M., Afidh, R. P. F., Misbullah, A., & Ilhamudin, T. (2024). Quantile Regression Neural Network (QRNN) and Hybrid NN – QRNN Model for Forecasting Rice Productivity in Indonesia. In *The 18th IMT-GT International Conference on Mathematics, Statistics and their Applications* (pp. 100–105). Sciendo. https://doi.org/10.2478/9788367405713-018

Leirvik, T., & Yuan, M. (2021). A Machine Learning Technique for Spatial Interpolation of Solar Radiation Observations. *Earth and Space Science*, *8*(4), 1–22. https://doi.org/10.1029/2020EA001527

Mahmoud, A., Yuan, X., Kheimi, M., & Yuan, Y. (2021). Interpolation Accuracy of Hybrid Soft Computing Techniques in Estimating Discharge Capacity of Triangular Labyrinth Weir. *IEEE Access*, *9*(1), 6769–6785. https://doi.org/10.1109/ACCESS.2021.3049223

Marchi, M., Castellanos-Acuña, D., Hamann, A., Wang, T., Ray, D., & Menzel, A. (2020). ClimateEU, scale-free climate normals, historical time series, and future projections for Europe. *Scientific Data*, *7*(428), 1–9. https://doi.org/10.1038/s41597-020-00763-0

Mohamad, N. B., Lai, A. C., & Lim, B. H. (2022). A case study in the tropical region to evaluate univariate imputation methods for solar irradiance data with different weather types. *Sustainable Energy*

Technologies and Assessments, *2*(50), 101764. https://doi.org/10.1016/j.seta.2021.101764

More, K. S., & Wolkersdorfer, C. (2024). Exploring Advanced Statistical Data Analysis Techniques for Interpolating Missing Observations and Detecting Anomalies in Mining Influenced Water Data. *ACS ES&T Water*, *4*(3), 1036–1045. https://doi.org/10.1021/acsestwater.3c00163

Portela, M. M., Espinosa, L. A., & Zelenakova, M. (2020). Long-Term Rainfall Trends and Their Variability in Mainland Portugal in the Last 106 Years. *Climate*, *8*(12), 1–18. https://doi.org/10.3390/cli8120146

Pratama, H. A., & Sam'an, M. (2023). Implementasi metode interpolasi bilinear untuk perbesaran skala citra. *Jurnal Komputer Dan Teknologi Informasi*, *1*(1), 21–25. https://doi.org/10.26714/.v1i1.11803

Putri, A. L. R., Surarso, B., & SRRM, T. U. (2023). MICE Implementation to Handle Missing Values in Rain Potential Prediction Using Support Vector Machine Algorithm. *JTAM (Jurnal Teori Dan Aplikasi Matematika)*, *7*(4), 1167–1117. https://doi.org/10.31764/jtam.v7i4.16699

Qiu, H., Chen, H., Xu, B., Liu, G., Huang, S., Nie, H., & Xie, H. (2024). Multiple Types of Missing Precipitation Data Filling Based on Ensemble Artificial Intelligence Models. *Water*, *16*(22), 1–21. https://doi.org/10.3390/w16223192

Rahayu, L., Ulfa, E. M., Sasmita, N. R., Sofyan, H., Kruba, R., Mardalena, S., & Saputra, A. (2023). Unraveling Geospatial Determinants : Robust Geographically Weighted Regression Analysis of Maternal Mortality in Indonesia. *Infolitika Journal of Data Science*, *1*(2), 73–81. https://doi.org/10.60084/ijds.v1i2.133

Saputra, A., Sofyan, H., Kesuma, Z. M., Sasmita, N. R., Wichaidit, W., & Chongsuvivatwong, V. (2024). Spatial patterns of tuberculosis in Aceh Province during the COVID-19 pandemic: a geospatial autocorrelation assessment. *IOP Conference Series: Earth and Environmental Science*, *1356*(1), 1–9. https://doi.org/10.1088/1755-1315/1356/1/012099

Sasmita, N. R., Khairul, M., Sofyan, H., Kruba, R., Mardalena, S., Dahlawy, A., Apriliansyah, F., Muliadi, M., Saputra, D. C. E., Noviandy, T. R., & Maula, A. W. (2023). A Statistical Clustering Approach: Mapping Population Indicators Through Probabilistic Analysis in Aceh Province, Indonesia. *Infolitika Journal of Data Science*, *1*(2), 1–9. https://doi.org/https://doi.org/10.60084/ijds.v1i2.130

Segeth, K. (2018). Some splines produced by smooth interpolation. *Applied Mathematics and Computation*, *319*(1), 387–394. https://doi.org/10.1016/j.amc.2017.04.022

Tierney, N., & Cook, D. (2023). Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations. *Journal of Statistical Software*, *105*(7), 1–31. https://doi.org/10.18637/jss.v105.i07

Ulhaq, M. Z., Farid, M., Aziza, Z. I., Nuzullah, T. M. F., Syakir, F., & Sasmita, N. R. (2024). Forecasting Upwelling Phenomena in Lake Laut Tawar: A Semi-Supervised Learning Approach. *Infolitika Journal of Data Science*, *2*(2), 53–61. https://doi.org/10.60084/ijds.v2i2.211

Wicher-Dysarz, J., Dysarz, T., & Jaskuła, J. (2022). Uncertainty in Determination of Meteorological Drought Zones Based on Standardized Precipitation Index in the Territory of Poland. *International Journal of Environmental Research and Public Health*, *19*(23), 15797. https://doi.org/10.3390/ijerph192315797

Wu, L., Liu, X., & Ma, X. (2018). Spatio-temporal temperature -variations in the Chinese Yellow River basin from 1981 to 2013. *Weather*, *73*(1), 27–33. https://doi.org/10.1002/wea.2956

Yasmin, A. A., Azahra, A. S., & Purwani, S. (2024). The application of cubic spline in rainfall modelling in Bogor and its impact on paddy production. *Communications in Mathematical Biology and Neuroscience*, *2024*(1), 1–13. https://doi.org/10.28919/cmbn/8430

Zhou, H., Ren, H., Royer, P., Hou, H., & Yu, X.-Y. (2022). Big Data Analytics for Long-Term Meteorological Observations at Hanford Site. *Atmosphere*, *13*(1), 1–17. https://doi.org/10.3390/atmos13010136