I.J. Information Engineering and Electronic Business, **2026, 2, 24-36**
Published Online on April 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2026.02.02

Improving Agricultural Commodity Trading through Data Imputation Methods for Price Prediction Accuracy

Pattharaporn Thongnim*
Department of Mathematics, Faculty of Science, Burapha University, Chonburi, 20131, Thailand
E-mail: pattharaporn@buu.ac.th
ORCID iD: https://orcid.org/0000-0001-8904-3979
*Corresponding Author

Sueppong Mueanchamnong
Department of Mathematics, Faculty of Science, Burapha University, Chonburi, 20131, Thailand
E-mail: sueppong.mu@go.buu.ac.th
ORCID iD: https://orcid.org/0009-0008-1277-6988

Received: 04 November, 2025; Revised: 29 December, 2025; Accepted: 02 February, 2026; Published: 08 April, 2026

Abstract: Agricultural price prediction in developing regions faces significant challenges from missing data in Internet of Things (IoT)-based environmental monitoring systems, particularly in tropical fruit cultivation where sensors frequently experience connectivity and operational failures. This study evaluates the impact of missing data imputation methods on agricultural price prediction model performance using environmental and market data from a commercial durian orchard in Chanthaburi Province, Thailand (2023-2024). Three imputation strategies—Linear Interpolation, Prophet, and Kalman Filter—were systematically compared across four machine learning algorithms (Regression Trees, Random Forest, XGBoost, and Artificial Neural Networks) using 10-fold cross-validation. The dataset comprised 182 observations with 28.02% missing environmental data and 68.13% missing price data, representing realistic constraints in developing agricultural economies. Results demonstrated that XGBoost consistently achieved superior performance across all imputation methods, with Kalman Filter combined with XGBoost showing the best testing performance (R^2 = 0.9767, MSE = 0.0013, MAE = 0.0287, MAPE = 1.49%). However, these results require careful interpretation given the limited sample size, high missingness, and potential temporal data leakage from random train-test splitting. Time series visualization revealed distinct characteristics: Linear Interpolation provided computational efficiency but oversimplified data complexity, Prophet captured seasonal patterns but introduced excessive noise, while Kalman Filter offered balanced performance preserving both smoothness and natural variability. Practical price prediction analysis showed substantial variations up to 35 Thai Baht per kilogram between imputation methods. The findings provide methodological evidence for imputation strategy selection in agricultural IoT systems with missing data, though validation with larger multi-site datasets is essential before operational deployment.

Index Terms: Missing Data, Imputation, Machine Learning, Internet of Things, Kalman Filter, Prophet Imputation, Moving Average

1. Introduction

Predicting agricultural prices is very important for farmers, traders and government officials who want to build better farming systems. Thailand is one of the world's top durian producers especially in Chanthaburi Province. However, farmers face big problems when trying to predict market prices because environmental factors and market conditions are very complex. Good price prediction models can help farmers decide when to harvest, manage their supply chains better and make more profit.

Modern farming now uses Internet of Things (IoT) technology to collect data automatically. This technology can help farmers monitor environmental conditions such as temperature, humidity, sunlight, wind and rainfall all the time [1]. These environmental factors directly affect crop quality and how much farmers can produce which then affects market prices. However, real farming data often has missing information because sensors break down, internet

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

connections fail, power cuts happen and equipment needs maintenance [2]. These missing data gaps can make prediction models much less reliable.

Missing data is a serious problem in agricultural forecasting systems. The traditional way of just removing incomplete data can cause a lot of information loss [3,4]. This is especially problematic in farming because collecting data costs a lot of money and takes a long time. Moreover, agricultural data changes over time and observations are naturally connected to each other. This means that smart methods to fill in missing data that preserve the natural patterns and seasonal changes in both environmental and price data are required.

Recent developments in machine learning and time series analysis have created different ways to handle missing data through interpolation techniques [5,6]. Each interpolation method has its own benefits and problems. Linear interpolation is simple to calculate; however, it might be insufficient for complex time patterns [7]. Prophet imputation, created by Facebook, is good at finding seasonal patterns but can introduce excessive variability [8]. Kalman filtering employs rigorous mathematical principles, however, it demands precise parameter tuning [9]. Following the interpolation process, additional smoothing techniques such as moving average can be applied for further data processing, although these methods might smooth out important fluctuations [10,11]. The choice of interpolation method can greatly affect how well prediction models work. Few studies have systematically compared their effects on agricultural price prediction.

Therefore, choosing the right machine learning methods for price prediction is also still being researched. For example, Random Forest and XGBoost have shown good results in farming applications because they can handle complex relationships between different factors [12,13]. Artificial Neural Networks (ANN) can find complicated patterns, although they require careful design [14]. Regression Trees create models that are easy to understand, nevertheless, they might not work well with small datasets [15]. The combination of missing data methods and machine learning algorithms makes choosing and optimizing models even more complicated.

This study addresses these challenges by systematically evaluating three imputation methods for handling missing data: Linear Interpolation, Prophet, and Kalman Filter. Environmental and price data were collected from a commercial durian orchard in Chanthaburi Province, Thailand, during the 2023-2024 growing season, comprising 182 daily observations with 28.02% missing environmental data and 68.13% missing price data. While this sample size is below conventional requirements for robust ML-based forecasting, it reflects realistic data availability constraints in developing agricultural economies where IoT infrastructure is emerging. The research investigates how different imputation methods affect the accuracy of price prediction models using four machine learning algorithms (Regression Trees, Random Forest, XGBoost, and Artificial Neural Networks), with particular attention to the trade-offs between imputation complexity, computational efficiency, and prediction performance.

This research has three main objectives: (1) to measure how different imputation methods affect the quality of time series reconstruction for agricultural data, (2) to evaluate machine learning model performance when applied to imputed datasets, and (3) to identify effective combinations of imputation methods and machine learning algorithms for agricultural price prediction in data-constrained environments. The findings contribute methodological evidence to the precision agriculture literature regarding imputation strategy selection in IoT-based agricultural monitoring systems with missing data. Given the limited sample size and high missingness rate, results should be interpreted as preliminary evidence for imputation method effects rather than as generalizable forecasting capability, with validation across larger multi-site datasets essential before operational deployment.

2. Methodology

This section describes the methodology adopted in this study including data collection, preprocessing steps (imputation strategies and moving average technique), and model building and evaluation. The overall framework aims to assess the robustness of different missing data imputation methods applied to environmental and price variables observed in a durian orchard in Chanthaburi, Thailand (2023–2024).

2.1 Data Collection

The dataset was obtained from a commercial durian orchard located in Chanthaburi Province, Thailand, one of the country's major durian cultivation areas. Primary environmental data were collected using IoT-based sensors that continuously monitored field conditions. The study period covers 2023–2024. The variables considered in this work include temperature, humidity, solar radiation, wind speed, wind direction, rainfall and durian price (secondary market data obtained from official agricultural sources).

The high price data missingness stems from a structural problem in agricultural commodity markets in developing regions. For example, price data are collected through market reports, government agricultural agencies, and trader networks, which often operate inconsistently and lack systematic documentation. This is a real-world constraint that researchers in developing agricultural economies commonly encounter.

This research focuses specifically on a single commercial durian orchard in Chanthaburi Province, Thailand, during the 2023-2024 growing season. This narrow scope was deliberately chosen to enable a detailed investigation of imputation method effects under controlled conditions with consistent data collection protocols. However, this specificity inherently limits generalizability. Findings should be interpreted as applicable primarily to

1. Tropical fruit cultivation in Eastern Thailand's climatic conditions
2. Durian specifically rather than other agricultural commodities
3. IoT-based monitoring systems with similar sensor configurations and missing data patterns
4. Farm-gate price estimation rather than broader market price forecasting

Validation across multiple orchards, diverse crops, different regions, and longer timeframes is essential before conclusions can be extended to agricultural forecasting in general.

2.2 Data Preprocessing

The dataset required careful preprocessing before implementing imputation methods. First, missing values in each variable were identified and their frequencies were calculated to understand the extent of data gaps. Statistical analysis was conducted to detect outliers using established thresholds combined with agricultural domain expertise. When outliers were found, appropriate corrections and removals were applied based on their impact on data quality.

After the outlier stage, missing data were imputed using three primary methods namely Linear Interpolation, Prophet Imputation and Kalman Filter. Following the completion of these interpolation techniques, moving average imputation was applied to smooth the interpolated values.

2.2.1 Imputation Methods

Given the high proportion of missing price data (68.13%), single method imputation is inappropriate and potentially unreliable. Single imputation methods assume imputed values represent fixed ground truth, thereby ignoring the uncertainty inherent in estimation when data loss is severe. Consequently, a multi-method comparative approach was adopted. Three missing data imputation methods were employed and compared in this study.

- Linear Interpolation: Missing value x_t is estimated as a weighted average of the nearest known observations [16].

$$x_t = x_{t_1} + \frac{(t-t_1)}{(t_2-t_1)} \cdot \left(x_{t_2} - x_{t_1}\right) \tag{1}$$

where $t_1 < t < t_2$ and x_{t_1}, x_{t_2} are the nearest observed values before and after t.

- Kalman Filter: State-space model with prediction and update equations as [17].

$$\hat{x}_{t|t-1} = A\hat{x}_{t-1|t-1} + Bu_t \tag{2}$$

$$P_{t|t-1} = AP_{t-1|t-1}A^T + Q \tag{3}$$

$$K_t = P_{t|t-1}H^T\left(HP_{t|t-1}H^T + R\right)^{-1} \tag{4}$$

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t\left(z_t - H\hat{x}_{t|t-1}\right) \tag{5}$$

$$P_{t|t} = (I - K_tH)P_{t|t-1} \tag{6}$$

where $\hat{x}_{t|t-1}$ is the predicted state estimate, $\hat{x}_{t|t}$ is the updated state estimate, $P_{t|t-1}$ is the predicted error covariance, $P_{t|t}$ is the updated error covariance, A is the state transition matrix, B is the control input matrix, u_t is the control vector, Q is the process noise covariance, R is the measurement noise covariance, H is the observation matrix, z_t is the measurement vector, K_t is the Kalman gain and I is the identity matrix.

- Prophet: The model decomposes the time series into trend, seasonality, and holiday components [18].

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{7}$$

where $g(t)$ is piecewise linear or logistic growth trend, $s(t)$ is seasonality modeled with Fourier series, $h(t)$ is holiday effects and ϵ_t is error term.

$$s(t) = \sum_{n=1}^{N}\left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right)\right) \tag{8}$$

where P is the period, N is the number of Fourier terms, a_n and b_n are the Fourier coefficients, n is the harmonic number and t is time.

In this study, the parameters for both the Prophet and Kalman Filter models were set as shown in Table 1. These parameters were selected to ensure appropriate modeling of the time-series characteristics of the data and to support accurate price forecasting.

Table 1. Model Parameters for Prophet and Kalman Filter Imputation Methods Used in This Study.

Model	Parameter	Value
Prophet	Daily seasonality	True
	Weekly seasonality	Auto
	Yearly seasonality	Auto
	Growth	Linear
	Changepoint detection	Enabled
Kalman Filter	Transition matrix	1
	Observation matrix	1
	Initial state mean	0
	Initial state covariance	1
	Observation covariance	1
	Transition covariance	0.01

2.2.2 Post-Imputation Smoothing

Following the imputation stage, a moving average smoothing technique was applied to reduce high frequency noise introduced during imputation [19]. It is important to note that the moving average is not an imputation method but rather a post-processing smoothing technique applied after imputation is complete. The moving average serves to reduce noise artifacts introduced by the imputation process, particularly the high-frequency fluctuations generated by the seasonal decomposition approach.

- Moving Average: The missing value x_t is estimated as the average of its k nearest neighbors

$$x_t = \frac{1}{k}\sum_{i=1}^{k} x_{t-i} \tag{9}$$

where x_t is the missing value at time t, k is the window size (number of previous observations) and x_{t-i} represents the observed values at previous time points.

Each method has specific advantages and limitations. For example, linear interpolation is computationally efficient; however, it may oversimplify nonlinear dynamics while Prophet provides better seasonal adjustment although it is computationally more intensive. Once imputation was finished, the complete dataset was split into training and testing sets using an 80-20 ratio. The training portion was used for model development while 10-fold cross-validation was implemented on the training data to ensure model reliability and minimize overfitting risks. The testing set was reserved for final model evaluation to assess generalization performance.

2.3. Model Building and Evaluation

Four machine learning models were selected to represent different modeling paradigms and complexity levels including Regression Trees (interpretable baseline), Random Forest (robust ensemble), XGBoost (state-of-the-art boosting), and Artificial Neural Networks (neural paradigm). This diversity enables comprehensive evaluation of imputation method effectiveness across varying algorithmic approaches, from simple interpretable models to complex nonlinear methods, while remaining practical for typical agricultural datasets in developing regions.

To evaluate the effectiveness of each imputation method, several machine learning models were applied to the imputed datasets including Regression Trees, Random Forest, XGBoost and Artificial Neural Networks (ANN). Model performance was assessed using 10-fold cross-validation to ensure robustness and reduce bias. Evaluation metrics included the coefficient of determination (R^2), mean squared error (MSE) and mean absolute error (MAE).

The evaluation metrics are defined as follows

- Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{10}$$

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{11}$$

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{12}$$

- Mean Absolute Percentage Error ($MAPE$)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{13}$$

where y_i represents the actual values, $\hat{y}_i$ represents the predicted values, $\bar{y}_i$ is the mean of actual values and n is the number of observations.

3. Results and Discussions

This section presents the experimental results of different imputation strategies applied to environmental and price data from 2023–2024. The analysis includes descriptive statistics and missing data patterns, comparison of three imputation methods, visualization of their impact on reconstructed time series and evaluation of subsequent model performance. The section concludes with a discussion of implications and limitations.

3.1 Descriptive Statistics and Missing Data Patterns

The dataset contains 182 observations. IoT system data including temperature, humidity, solar radiation, wind speed, wind direction and rainfall have 51 missing observations (28.02%) while durian price data have 124 missing observations (68.13%). This substantial difference in missing data patterns between environmental and price variables presents unique challenges for imputation strategies and emphasizes the importance of robust missing data handling techniques in agricultural forecasting applications.

Table 2. Descriptive Statistics for Environmental Variables and Durian Prices Collected from IoT Sensors and Market Data in Chanthaburi Province, Thailand (2023-2024).

Variable	Mean	Min	Max	Std
Temperature	29.0692	23.4610	31.8836	1.4486
Humidity	83.8966	67.1398	93.0932	3.9813
Solar Radiation	23.3484	5.1141	43.3848	7.1054
Wind Direction	198.5286	100.1265	281.0410	45.0652
Wind Speed	17.2367	0.2873	41.7811	12.3735
Rain Fall	3.4153	0.0000	59.0000	9.1359
Price	168.1897	130.0000	255.0000	29.5537

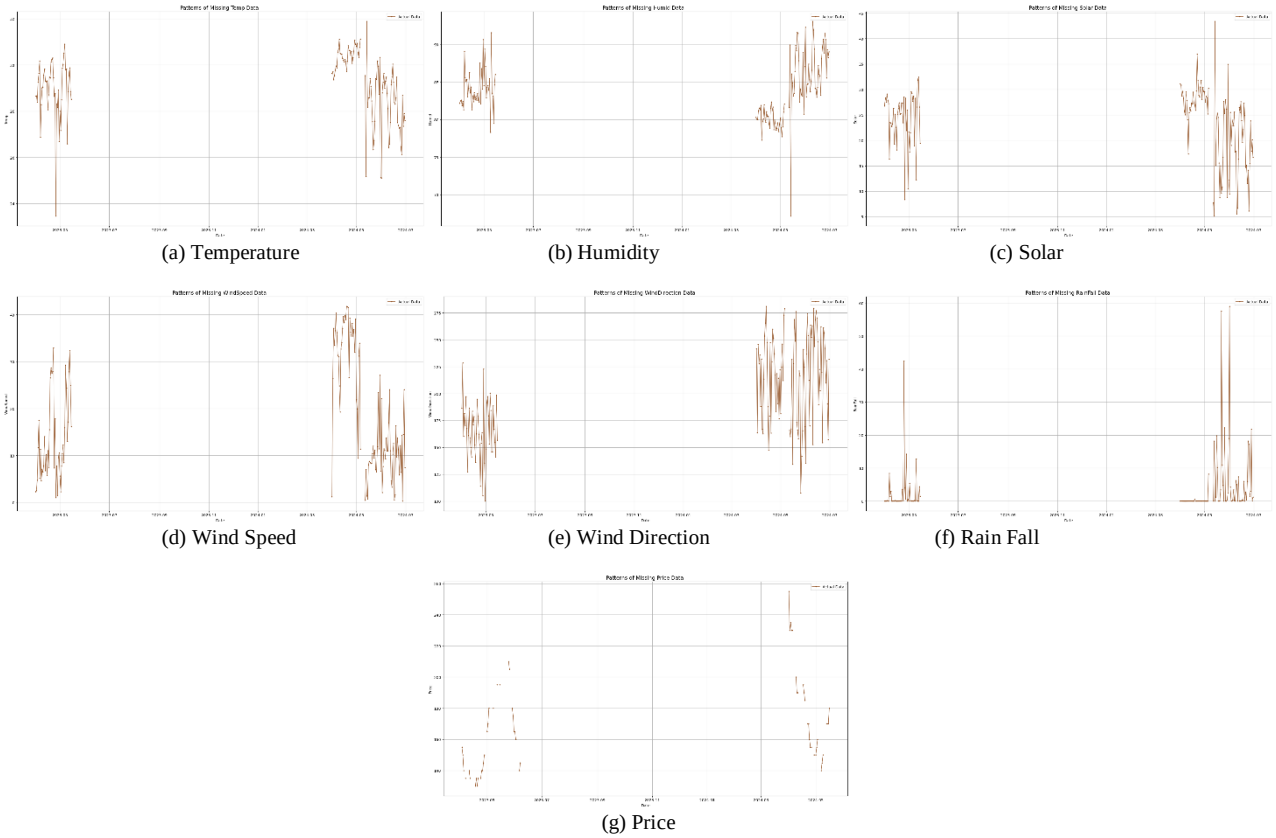
This finding presents the characteristics of the dataset through descriptive statistics summarized in Table 2. and analyzes the distribution of missing values across all variables. Fig. 1. provides a visual representation of missing data patterns to illustrate the extent and nature of data gaps throughout the study period.

Table 2. reveals considerable variation in the measured variables. Temperature shows relatively stable conditions with a mean of 29.07℃ and low variability (standard deviation = 1.45). Humidity levels remain consistently high averaging 83.90% with moderate variation. Solar radiation exhibits wider fluctuations ranging from 5.11 to 43.38 units with substantial variability (standard deviation = 7.11). Wind patterns demonstrate high variability with wind direction spanning from 100.13 to 281.04 and wind speed varying dramatically from near-zero to 41.78 units. Rainfall data shows sporadic patterns with values ranging from 0 to 59 units and high standard deviation (9.14) indicating irregular precipitation events. Durian prices fluctuate between 130 and 255 Thai Baht with a mean of 168.19 and standard deviation of 29.55 reflecting typical market volatility for agricultural commodities.

Table 3. Correlation Coefficients between Environmental Variables and Durian Price.

Variable	Correlation	p-value
Temperature	0.1411	0.3494
Humidity	-0.1722	0.2525
Solar Radiation	0.1627	0.2798
Wind Direction	0.2453	0.1004
Wind Speed	0.4027	0.0055
Rain Fall	-0.1557	0.3016

Table 3. shows correlation between environmental variables and durian price. Most environmental have a positive relationship with durian price, including temperature, solar radiation, wind speed, and wind direction. This means that when these environmental increases occur, durian prices tend to increase as well. In contrast, humidity and rainfall show a negative relationship with durian prices. This indicates that higher humidity and rainfall are linked to lower durian prices. Overall, the relationships of all environmental are weak to moderate.



(a) Temperature (b) Humidity (c) Solar

(d) Wind Speed (e) Wind Direction (f) Rain Fall

(g) Price

Fig. 1. Missing Data Patterns for Environmental Variables and Durian Prices Collected from Commercial Orchard in Chanthaburi Province, Thailand (2023-2024).

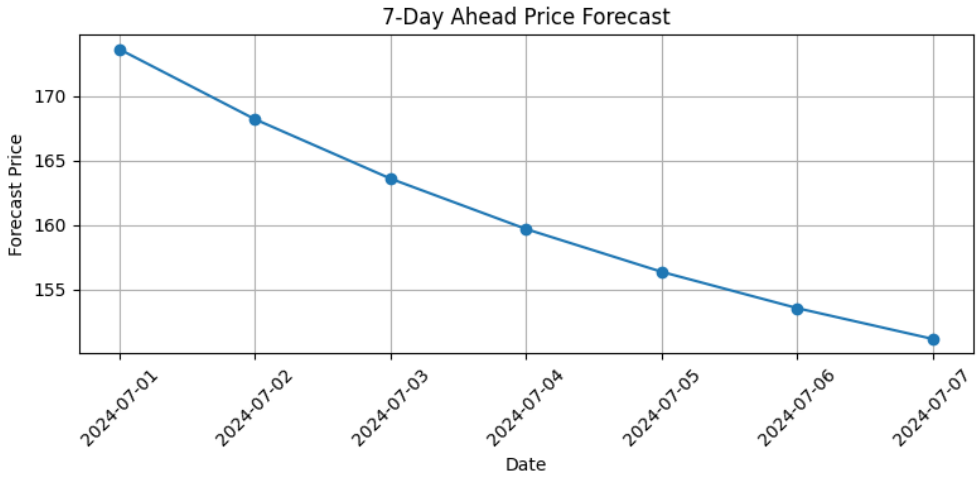


Fig. 2. Seven-Day Ahead Price Forecast Using Multiple Linear Regression (MLR) on Original Dataset Without Imputation.

In this study, Multiple Linear Regression (MLR) was applied to perform a 7-day ahead forecast using a dataset without missing value imputation, as shown in Fig. 2. The results indicate a continuously decreasing trend in the predicted values. The evaluation metrics are as follows: R^2 = 0.6219, MSE = 324.5169, MAE = 11.9073, and $MAPE$ = 7.00%. Due to the large number of missing values in the dataset, these statistical indicators can exhibit abnormal and unreliable behavior.

3.2 Performance of Imputation Methods

The average cross-validation performance of each imputation method (Linear Interpolation, Prophet and Kalman Filter) combined with different machine learning models (Regression Tree, Random Forest, XGBoost and ANN) is reported. Evaluation metrics include the coefficient of determination (R^2), mean squared error (MSE), mean absolute error (MAE) and mean absolute percentage error ($MAPE$). The results allow for a systematic comparison of method effectiveness across all imputation-model combinations.

Table 4. Cross-Validation Performance Comparison of Three Imputation Methods (Linear Interpolation, Prophet, Kalman Filter) Across Four Machine Learning Models for Durian Price Prediction.

Method	Model	R^2	MSE	MAE	$MAPE$
Linear Interpolation	Regression Tree	0.7380	0.0140	0.0674	4.2927%
	Random Forest	0.8680	0.0070	0.0561	3.7083%
	XGBoost	0.8872	0.0065	0.0440	2.8213%
	ANN	0.8383	0.0083	0.0641	4.1410%
Prophet	Regression Tree	0.7790	0.0080	0.0568	2.9930%
	Random Forest	0.8336	0.0059	0.0473	2.3907%
	XGBoost	0.8335	0.0058	0.0446	2.0863%
	ANN	0.7029	0.0097	0.0716	3.5644%
Kalman Filter	Regression Tree	0.7013	0.0153	0.0695	3.7761%
	Random Forest	0.8324	0.0082	0.0546	3.0874%
	XGBoost	0.8649	0.0069	0.0436	2.4125%
	ANN	0.8250	0.0084	0.0650	3.5946%

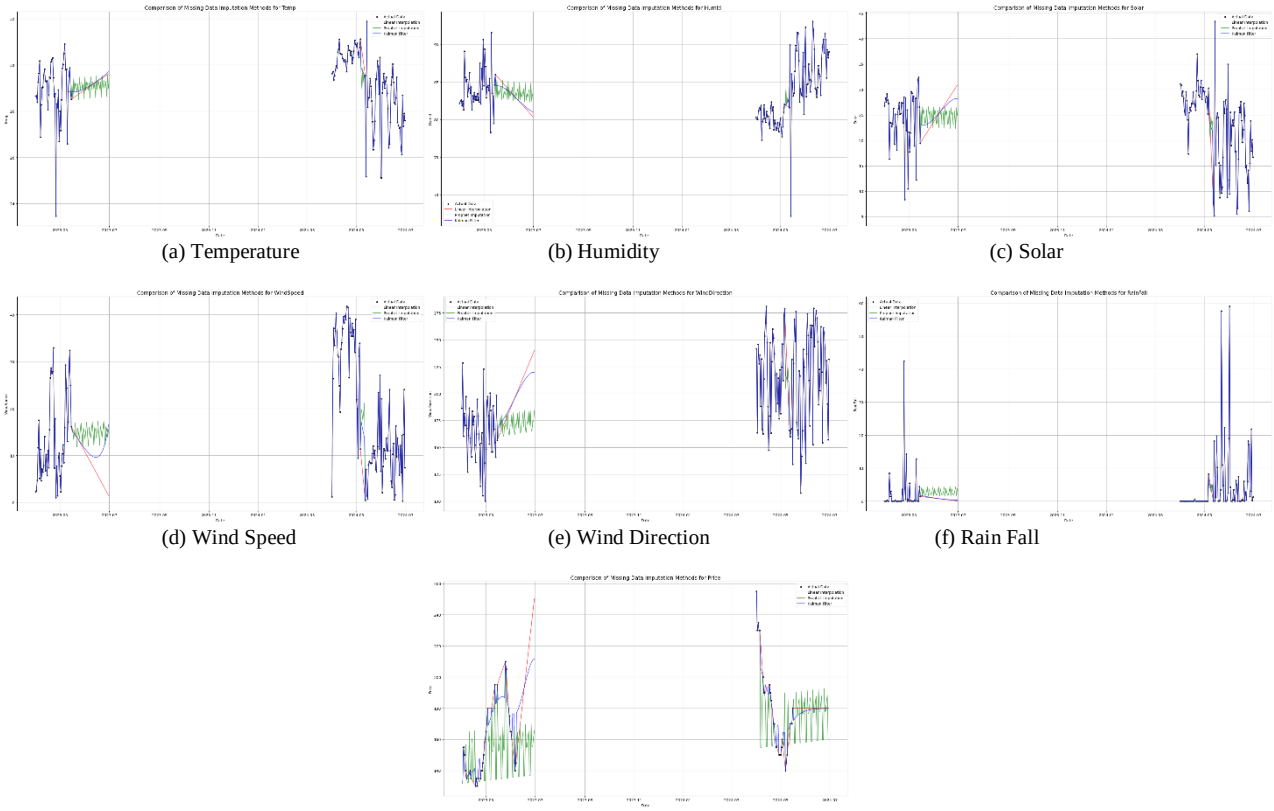
Table 4. shows that XGBoost consistently achieved the best performance across all imputation methods, demonstrating superior predictive accuracy in terms of both coefficient of determination and error metrics. Linear Interpolation combined with XGBoost delivered the highest R^2 (0.8872) with MSE = 0.0065, MAE = 0.0440, and $MAPE$ = 2.82%, indicating strong model fit and relatively low prediction errors. Kalman Filter with XGBoost achieved the second-best performance with R^2 = 0.8649, MSE = 0.0069, MAE = 0.0436, and $MAPE$ = 2.41%, showing comparable accuracy with slightly lower percentage errors. For Prophet imputation, XGBoost again performed best with R^2 = 0.8335, MSE = 0.0058, MAE = 0.0446, and the lowest $MAPE$ of 2.09% among all combinations, though Random Forest showed marginally lower MSE (0.0059) for this imputation method.

The $MAPE$ values across all XGBoost models remained below 3%, indicating that predictions typically deviated less than 3% from actual prices—a practically acceptable error range for agricultural price forecasting applications. In contrast, Regression Tree and ANN models generally exhibited higher $MAPE$ values (ranging from 2.99% to 4.29%), suggesting less reliable predictions. Overall, the results demonstrate that Linear Interpolation combined with XGBoost produced the most accurate predictions among all tested combinations based on R^2 values, while Prophet combined with XGBoost achieved the lowest percentage error ($MAPE$ = 2.09%), highlighting the importance of matching imputation methods with appropriate machine learning algorithms to optimize different performance metrics.

3.3 Performance of Imputation Methods

Fig. 3. illustrates the reconstructed time series for environmental and price variables (Temperature, Humidity, Solar Radiation, Wind Speed, Wind Direction, Rainfall and Price) highlighting the differences between imputation methods. Linear interpolation typically produces smooth and steady trends that are computationally straightforward. However, this approach often oversimplifies the inherent complexity of agricultural data. For instance, in humidity, solar radiation and wind direction variables, the linear method underestimates natural variability and fails to capture local fluctuations that can be critical for accurate modeling.

In contrast, Prophet imputation demonstrates superior capability in identifying seasonal and cyclical patterns particularly evident in humidity, temperature and price variables where it reconstructs more realistic temporal dynamics. Nevertheless, this method occasionally introduces excessive noise that can compromise model stability and complicate subsequent machine learning processes [20]. Kalman filtering provides a balanced approach by maintaining data smoothness while preserving essential temporal variations. This advantage is particularly apparent in rainfall and wind speed variables where Kalman filtering demonstrates better handling of extreme values and underlying trends compared to other methods [21]. However, the method sometimes exhibits oversensitivity to outliers, potentially disrupting temporal continuity in the reconstructed series.



(a) Temperature (b) Humidity (c) Solar

(d) Wind Speed (e) Wind Direction (f) Rain Fall

(g) Price

Fig. 3. Visual Comparison of Time Series Reconstruction Using Three Imputation Methods (Linear Interpolation, Prophet and Kalman Filter) for Environmental and Price Variables.

3.4 Impact on Model Prediction

Four machine learning algorithms were chosen to represent diverse modeling approaches and complexity levels including Regression Trees provide an interpretable baseline, Random Forest offers robust ensemble learning, XGBoost delivers state-of-the-art gradient boosting performance, and Artificial Neural Networks represent neural-based architectures. This selection enables thorough evaluation of imputation method performance across models ranging from simple transparent algorithms to sophisticated nonlinear techniques, while ensuring practical feasibility for agricultural datasets typical of developing countries. More computationally demanding methods, including deep learning architectures and support vector regression, were excluded given the limited dataset size (n=182) and the computational constraints of agricultural monitoring systems in resource-limited settings.

The performance of machine learning models after applying different imputation methods is analyzed in detail. Table 5. presents a comprehensive comparison of training and testing performance across all imputation methods and machine learning models. The results reveal several key patterns and insights regarding model behavior and generalization capability, particularly in terms of error magnitude as reflected by $MAPE$.

XGBoost consistently achieved the highest performance across all imputation methods demonstrating superior predictive accuracy. For training data, XGBoost reached near-perfect performance with R^2 values of 0.9998–0.9999 and extremely low error rates (MSE = 0.0000, MAE = 0.0015–0.0021, and $MAPE$ = 0.0804–0.1328%). However, the testing results show more realistic performance levels indicating appropriate model complexity without severe overfitting, as evidenced by reasonable $MAPE$ values across all imputation methods.

Among the imputation methods, Kalman Filter combined with XGBoost produced the strongest overall results, achieving R^2 = 0.9767, MSE = 0.0013, MAE = 0.0287, and $MAPE$ = 1.4861% on the testing set. Prophet with XGBoost followed closely with R^2 = 0.9679, MSE = 0.0011, MAE = 0.0207, and $MAPE$ = 1.0490%, while Linear Interpolation with XGBoost achieved R^2 = 0.9603, MSE = 0.0025, MAE = 0.0341, and $MAPE$ = 2.2561%.

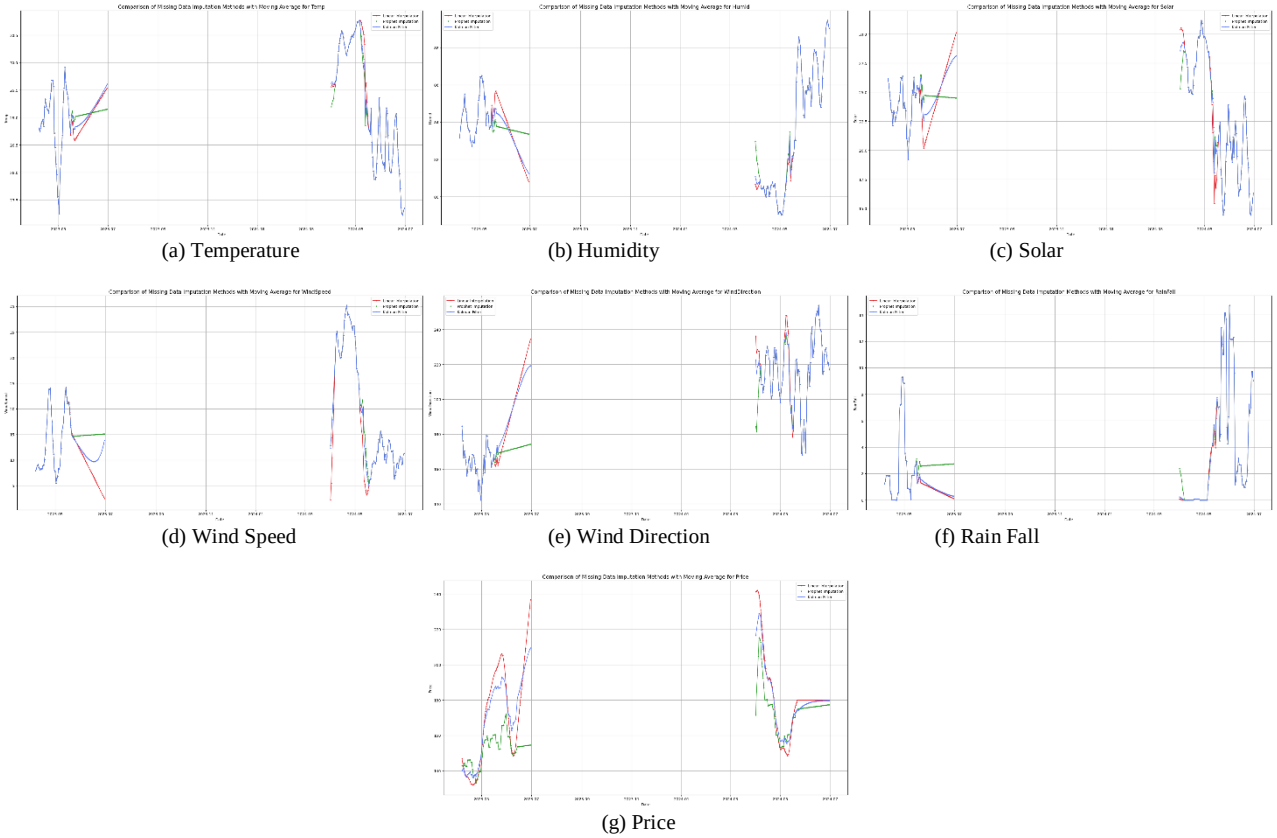
Random Forest demonstrated robust performance across all imputation methods, maintaining good generalization with testing R^2 values ranging from 0.9375 to 0.9432 and $MAPE$ values between 1.8705% and 2.4596%. Notably, Random Forest showed less dramatic differences between training and testing performance compared to XGBoost, suggesting more stable model behavior.

Regression Tree models exhibited moderate performance with testing R^2 values between 0.9079 and 0.9296 and $MAPE$ values ranging from 1.9624% to 2.6941%, while ANN models showed the most variable results, with testing R^2 ranging from 0.8394 (Prophet) to 0.9251 (Kalman Filter) and MAPE ranging from 2.6118% to 3.0777%. The

substantial gap between training and testing performance for some models highlights the importance of proper model selection and validation in agricultural price prediction applications.

Table 5. Training and Testing Performance of Machine Learning Models Using Different Imputation Methods (Linear Interpolation, Prophet and Kalman Filter) for Agricultural Price Forecasting.

Method	Model	Training				Testing			
		R^2	*MSE*	*MAE*	*MAPE*	R^2	*MSE*	*MAE*	*MAPE*
Linear Interpolation	Regression Tree	0.9706	0.0017	0.0277	1.6881%	0.9296	0.0032	0.0380	2.4290%
	Random Forest	0.9762	0.0014	0.0261	1.7184%	0.9409	0.0024	0.0359	2.4596%
	XGBoost	0.9998	0.0000	0.0021	0.1328%	0.9603	0.0025	0.0341	2.2561%
	ANN	0.8676	0.0075	0.0615	3.9663%	0.9225	0.0042	0.0494	3.0777%
Prophet	Regression Tree	0.9092	0.0033	0.0314	0.0157%	0.9079	0.0031	0.0389	0.0196%
	Random Forest	0.9708	0.0010	0.0203	0.0103%	0.9375	0.0027	0.0367	0.0187%
	XGBoost	0.9998	0.0000	0.0019	0.0010%	0.9679	0.0011	0.0207	0.0105%
	ANN	0.7954	0.0075	0.0636	0.0320%	0.8394	0.0051	0.0584	0.0289%
Kalman Filter	Regression Tree	0.9444	0.0027	0.0320	1.6754%	0.9221	0.0046	0.0518	2.6941%
	Random Forest	0.9720	0.0014	0.0242	1.3581%	0.9432	0.0032	0.0420	2.3027%
	XGBoost	0.9999	0.0000	0.0015	0.0804%	0.9767	0.0013	0.0287	1.4861%
	ANN	0.8734	0.0064	0.0597	3.2917%	0.9251	0.0037	0.0484	2.6118%



(a) Temperature (b) Humidity (c) Solar

(d) Wind Speed (e) Wind Direction (f) Rain Fall

(g) Price

Fig. 4. Performance of Moving Average Imputation for Environmental and Price Variables (2023–2024).

Fig. 4. presents a comparative analysis of the three imputation methods when combined with moving average smoothing for environmental and price variables during 2023-2024. The visualization reveals distinct characteristics and performance patterns for each approach.

Linear interpolation demonstrates a tendency toward excessive smoothing which removes critical short-term fluctuations that can contain important information for price prediction. This over-smoothing effect occasionally results in systematic overestimation or underestimation of values, particularly during periods of rapid environmental change or price volatility. In addition, Prophet imputation exhibits the opposite behavior generating high-frequency fluctuations that sometimes produce values below the actual observed levels. While this method successfully captures seasonal patterns, it occasionally fails to maintain consistency with the underlying long-term trends introducing noise that may complicate subsequent modeling efforts. Moreover, Kalman Filter demonstrates the most balanced performance among

the three methods, effectively preserving both data smoothness and natural variability. This approach produces reconstructed time series that closely approximate the actual data patterns while maintaining temporal consistency and reducing spurious fluctuations. The balanced nature of Kalman filtering makes it particularly suitable for agricultural time series where both trend preservation and noise reduction are essential.

All three imputation methods can be effectively applied to predictive modeling with XGBoost. However, the Kalman Filter consistently delivers the most reliable inputs resulting in the best model performance across both environmental and price variables.

3.5 XGBoost Price Prediction Results Across Imputation Methods

Based on the comprehensive performance evaluation presented in previous sections, XGBoost emerged as the most effective machine learning model across all imputation methods consistently achieving the highest R^2 values and lowest error rates on both training and testing datasets. Given this superior performance, XGBoost was selected for detailed price prediction analysis to demonstrate how different imputation strategies influence actual forecasting outcomes in real-world scenarios.

Table 6. presents specific price predictions generated by XGBoost using the three imputation methods for a representative week in July 2024. This comparison illustrates the practical implications of imputation method selection on agricultural price forecasting applications.

Table 6. Durian Price Forecasts Generated by XGBoost Using Three Different Imputation Strategies (Thai Baht per kilogram).

Date	Linear Interpolation	Prophet	Kalman Filter
7/1/2024	180.0246	177.2117	179.7656
7/2/2024	150.5652	141.7074	141.7500
7/3/2024	179.9705	180.2845	185.5630
7/4/2024	165.1193	163.4789	163.8857
7/5/2024	163.1389	144.4005	179.5250
7/6/2024	146.1558	173.6403	177.3681
7/7/2024	152.3433	172.7687	162.1800

The results demonstrate the practical implications of imputation method selection on actual price forecasting scenarios. The analysis reveals substantial variations in predicted durian prices depending on the chosen imputation strategy with differences ranging from a few Thai Baht to over 35 Baht per kilogram for the same prediction date.

Linear Interpolation generally produced moderate price predictions with values ranging from 146.16 to 180.02 Baht throughout the week. This method showed relatively stable forecasting behavior reflecting its tendency toward smoothing that was observed in the time series reconstruction analysis. Furthermore, Prophet imputation exhibited the most variable predictions with the widest range from 141.71 to 180.28 Baht. Notably, Prophet generated the lowest prediction on July 5th (144.40 Baht) approximately 35 Baht lower than the Kalman Filter prediction for the same date (179.53 Baht). This substantial difference highlights how Prophet's sensitivity to seasonal patterns can significantly impact short-term price forecasts. Kalman Filter demonstrated more consistent prediction patterns with values ranging from 141.75 to 185.56 Baht. This method produced the highest individual prediction (185.56 Baht on July 3rd) but maintained relatively stable forecasting throughout most of the period.

While XGBoost achieved near-perfect performance on training data (R^2 = 0.9998-0.9999), the testing performance showed a notable decline (R^2 = 0.9603-0.9767), indicating overfitting. This gap of approximately 0.02-0.04 in R^2 suggests that the model learned training-specific patterns that did not fully generalize to unseen data. However, several factors warrant consideration. First, despite the overfitting, the testing R^2 values (0.9603-0.9767) remain substantially high and practically useful for agricultural price prediction. The overfitting, while present, does not severely compromise model utility for real-world applications. Second, this overfitting pattern is characteristic of ensemble tree-based methods such as XGBoost when applied to small datasets. The limited sample size (n=182) represents a fundamental limitation of this study. With only 182 observations, the model's capacity to fit complex patterns exceeds what is optimal for the available data, making the model susceptible to learning noise rather than generalizable patterns. This small sample size also constrains the ability to implement more aggressive regularization and larger validation sets without further reducing training data. Despite this limitation, XGBoost still outperformed other models on testing data, suggesting it captures genuine underlying relationships better than alternatives.

Future studies with larger datasets spanning multiple growing seasons and years should employ stricter hyperparameter tuning, including lower learning rates, higher regularization parameters [22]. and more robust cross-validation strategies. The observed overfitting underscores that the findings should be interpreted as preliminary evidence requiring validation with substantially larger datasets before conclusions can be generalized beyond this specific orchard and time period.

The dataset exhibits 68.13% missing price data, which represents a critical limitation that requires careful interpretation. While this high missingness reflects structural characteristics of agricultural commodity markets in developing regions-where price data are typically compiled from multiple independent sources such as government

agricultural agencies, local trader networks, and market reports operating without consistent protocols—it raises fundamental questions about imputation reliability. With more than two-thirds of price data missing, any single imputation method risks introducing substantial bias and artificial patterns. Rather than discarding valuable environmental data, this study addresses this challenge by employing and comparing multiple imputation methods (Linear Interpolation, Prophet, and Kalman Filter). The consistency of XGBoost predictions across all three imputation methods provides some confidence that the approaches preserve meaningful price signals rather than generating pure artifacts. If the imputed values were entirely unreliable, different imputation strategies would yield highly inconsistent results. Nevertheless, the results should be interpreted with appropriate caution given the severity of data missingness.

This study demonstrates that agricultural price forecasting in developing economies is feasible when multiple imputation methods are systematically compared, results are validated against domain expertise, and limitations are transparently reported. However, the 68.13% missing rate remains a substantial limitation that constrains the generalizability and reliability of the findings. Future research should prioritize establishing more robust data collection systems to reduce missingness below 30%, which is generally considered acceptable for statistical analysis [23].

The unusually high R^2values (0.9603-0.9767) observed in this study require careful interpretation and contextualization. These results differ substantially from typical agricultural commodity price forecasting studies, which commonly report R^2 values below 0.70, primarily because this research addresses a fundamentally different and more constrained prediction problem. This study predicts farm-gate prices at a single orchard location using direct on-site environmental measurements, rather than forecasting wholesale, retail, and futures market prices at regional and national scales. Farm-gate prices exhibit substantially lower volatility compared to market prices because they reflect immediate local supply-demand dynamics and direct quality assessment by buyers, while excluding speculative trading, intermediary markups, transportation costs, and external market disruptions that introduce substantial uncertainty in broader market forecasting.

However, a critical methodological limitation is the use of random 80-20 train-test splits applied to time series data from a single growing season (2023-2024). This approach may have introduced temporal data leakage, where models effectively performed interpolation between temporally adjacent observations rather than conducting genuine out-of-sample forecasting. In proper time series forecasting, training should occur exclusively on past data and testing on future data to simulate real-world prediction scenarios. The split data used in this study allows the model to learn from observations both before and after test points, artificially inflating performance metrics and limiting the validity of the results as a true forecasting evaluation. Therefore, additional methodological limitations include: (1) potential overfitting to location-specific patterns from a single orchard, limiting generalizability to other farms and regions, (2) the aforementioned small sample size (n=182) with extremely high missingness (68.13% for price data), constraining statistical power and imputation reliability, and (3) exclusion of critical external market factors such as regional supply fluctuations, export demand, competitor pricing, and agricultural policy changes that would substantially reduce predictive accuracy in real-world forecasting scenarios. Therefore, while this study successfully demonstrates that imputation method selection substantially affects model performance and establishes strong associations between environmental variables and farm-gate prices in this specific localized context, the high R^2 values should not be interpreted as generalizable agricultural price forecasting capability. The primary contribution is methodological—providing empirical evidence for imputation strategy selection in agricultural IoT systems with missing data—rather than demonstrating deployable practical price forecasting systems.

Future research should address these limitations by: (1) collecting data across multiple orchards, growing seasons, and years to increase sample size and representativeness, (2) implementing strict temporal validation where models are trained exclusively on past data and tested on future periods, (3) reducing data missingness through improved IoT infrastructure and systematic price documentation, and (4) incorporating external market variables, including regional production data and export statistics. Only through such comprehensive validation can the practical utility of these methods for real-world agricultural price forecasting be rigorously established.

4. Conclusions

This study conducted a comprehensive evaluation of missing data imputation methods and their impact on agricultural price prediction models using environmental data from a commercial durian orchard in Chanthaburi Province, Thailand. The research systematically compared three imputation strategies (Linear Interpolation, Prophet, and Kalman Filter) across four machine learning algorithms (Regression Trees, Random Forest, XGBoost, and Artificial Neural Networks) to assess their effectiveness in handling missing agricultural data. Results demonstrated that XGBoost consistently achieved superior performance across all imputation methods, with Kalman Filter combined with XGBoost showing the best testing performance (R^2 = 0.9767, MSE = 0.0013, MAE = 0.0287, $MAPE$ = 1.4861%). The practical implications analysis revealed substantial differences between imputation methods, with price variations ranging up to 35 Thai Baht per kilogram for identical forecasting periods. This finding underscores the critical importance of imputation strategy selection in real-world agricultural applications, where prediction accuracy directly impacts farmer decision-making regarding harvest timing, marketing strategies, and financial planning.

The study contributes to the growing body of knowledge in precision agriculture by demonstrating that imputation method selection substantially affects model performance in agricultural time series forecasting. The results suggest that

while Kalman Filter generally provides the most reliable performance, the optimal choice of imputation strategy should consider specific application requirements, data characteristics, and risk tolerance of end users. To address identified limitations, future research should collect multi-site datasets spanning multiple orchards and growing seasons, implement strict temporal validation to eliminate data leakage, reduce missingness below 30%, and incorporate external market variables (regional supply, export demand, policy indicators). Multi-site validation across diverse crops and climatic conditions with substantially larger datasets is essential before operational deployment.

Author Contributions Statement

Pattharaporn Thongnim – Conceptualization, Methodology, Supervision, and Writing original draft.
Sueppong Mueanchamnong– Data preprocessing and Data analysis.
All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

This research was supported by the grant SC05/2568 "Development of Daily Durian Price Prediction Model using IoT-Based Weather Data" from Faculty of Science, Burapha University, Thailand.

Data Availability Statement

The datasets generated and analyzed during this study are publicly available in the Open Science Framework (OSF) repository at https://doi.org/10.17605/OSF.IO/E2U6W.

Ethical Declarations

None

Acknowledgments

We are grateful to the Office of Agricultural Economics Region 6, Chonburi Province, for providing valuable durian market price data that made this research possible.

Declaration of Generative AI in Scholarly Writing

During the preparation of this work, the authors used Claude AI to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] M. Dhanaraju, P. Chenniappan, K. Ramalingam, S. Pazhanivelan, and R. Kaliaperumal, "Smart farming: Internet of Things (IoT)-based sustainable agriculture," *Agriculture*, vol. 12, no. 10, p. 1745, 2022.J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] X. Zou, W. Liu, Z. Hu, S. Wang, Z. Chen, C. Xin, Y. Bai, Z. Liang, Y. Gong, Y. Qian, *et al.*, "Current status and prospects of research on sensor fault diagnosis of agricultural internet of things," *Sensors*, vol. 23, no. 5, p. 2528, 2023.

[3] Y. Zhang and P. J. Thorburn, "Handling missing data in near real-time environmental monitoring: A system and a review of selected methods," *Future Generation Computer Systems*, vol. 128, pp. 63–72, 2022.

[4] J. Du, M. Hu, and W. Zhang, "Missing data problem in the monitoring system: A review," *IEEE Sensors Journal*, vol. 20, no. 23, pp. 13984–13998, 2020.

[5] D. Adhikari, W. Jiang, J. Zhan, Z. He, D. B. Rawat, U. Aickelin, and H. A. Khorshidi, "A comprehensive survey on imputation of missing data in internet of things," *ACM Computing Surveys*, vol. 55, no. 7, pp. 1–38, 2022.

[6] Z. Ding, G. Mei, S. Cuomo, Y. Li, and N. Xu, "Comparison of estimating missing values in IoT time series data using different interpolation algorithms," *International Journal of Parallel Programming*, vol. 48, no. 3, pp. 534–548, 2020.

[7] F. A. N. Declerq, "Interpolation methods for scattered sample data: Accuracy, spatial patterns, processing time," *Cartography and Geographic Information Systems*, vol. 23, no. 3, pp. 128–144, 1996.

[8] M. Weber, M. Turowski, H. K. Çakmak, R. Mikut, U. Kühnapfel, and V. Hagenmeyer, "Data-driven copy-paste imputation for energy time series," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 5409–5419, 2021.

[9] G. Candiani, C. Carnevale, G. Finzi, E. Pisoni, and M. Volta, "A comparison of reanalysis techniques: Applying optimal interpolation and ensemble Kalman filtering to improve air quality monitoring at mesoscale," *Science of the Total Environment*, vol. 458, pp. 7–14, 2013.
[10] A. Ajiono and T. Hariguna, "Comparison of three time series forecasting methods on linear regression, exponential smoothing, and weighted moving average," *International Journal of Informatics and Information Systems*, vol. 6, no. 2, pp. 89–102, 2023.
[11] A. R. Mohammed, K. S. Hassan, and M. A. M. Abdel-Aal, "Moving average smoothing for Gregory–Newton interpolation: A novel approach for short-term demand forecasting," *IFAC-PapersOnLine*, vol. 55, no. 10, pp. 749–754, 2022.
[12] L. Huang, Y. Liu, W. Huang, Y. Dong, H. Ma, K. Wu, and A. Guo, "Combining random forest and XGBoost methods in detecting early and mid-term winter wheat stripe rust using canopy-level hyperspectral measurements," *Agriculture*, vol. 12, no. 1, p. 74, 2022.
[13] O. M'hamdi, S. Takács, G. Palotás, R. Ilahy, L. Helyes, and Z. Pék, "A comparative analysis of XGBoost and neural network models for predicting some tomato fruit quality traits from environmental and meteorological data," *Plants*, vol. 13, no. 5, p. 746, 2024.
[14] I. A. Basheer and M. Hajmeer, "Artificial neural networks: Fundamentals, computing, design, and application," *Journal of Microbiological Methods*, vol. 43, no. 1, pp. 3–31, 2000.
[15] G. De'ath and K. E. Fabricius, "Classification and regression trees: A powerful yet simple technique for ecological data analysis," *Ecology*, vol. 81, no. 11, pp. 3178–3192, 2000.
[16] N. M. Noor, M. M. A. B. Abdullah, A. S. Yahaya, and N. A. Ramli, "Comparison of linear interpolation method and mean method to replace the missing values in environmental data set," in *Proc. Materials Science Forum*, vol. 803, pp. 278–281, 2015.
[17] L. Padilla, B. Lagos-Álvarez, J. Mateu, and E. Porcu, "Space-time autoregressive estimation and prediction with missing data based on Kalman filtering," *Environmetrics*, vol. 31, no. 7, p. e2627, 2020.
[18] S. F. Stefenon, L. O. Seman, V. C. Mariani, and L. dos Santos Coelho, "Aggregating prophet and seasonal trend decomposition for time series forecasting of Italian electricity spot prices," *Energies*, vol. 16, no. 3, p. 1371, 2023.
[19] S. A. Akrami, A. El-Shafie, M. Naseri, and C. A. G. Santos, "Rainfall data analyzing using moving average (MA) model and wavelet multi-resolution intelligent model for noise evaluation to improve the forecasting accuracy," *Neural Computing and Applications*, vol. 25, no. 7, pp. 1853–1864, 2014.
[20] M. A. Jahin, M. S. H. Shovon, J. Shin, A. R. Ridoy, and M. F. Mridha, "Big data–supply chain management framework for forecasting: Data preprocessing and machine learning techniques," *arXiv preprint* arXiv:2307.12971, 2023.
[21] J.-A. Ting, E. Theodorou, and S. Schaal, "A Kalman filter for robust outlier detection," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pp. 1514–1519, 2007.
[22] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data," *Ecological Modelling*, vol. 406, pp. 109–120, 2019.
[23] D. A. Newman, "Missing data: Five practical guidelines," *Organizational Research Methods*, vol. 17, no. 4, pp. 372–411, 2014.

Authors' Profiles



Pattharaporn Thongnim received her B.S. degree in mathematics from Mahidol University, Thailand, in 2006, her M.S. degree in statistics from Chulalongkorn University, Thailand, in 2009, and her Ph.D. degree in statistics and data science from the University of Leicester, Leicester, UK, in 2022. She served as a Lecturer at the Faculty of Science and Arts, Burapha University, Chanthaburi Campus, Thailand, from 2009 to 2023, and as Head of Data Center in Chanthaburi from 2022 to 2023. She currently serves as a Lecturer in the Department of Mathematics, Faculty of Science, Burapha University, Chonburi, Thailand. Her current research interests include data science for agriculture, machine learning, Gaussian process mixture regression models, and the integration of IoT technology with agricultural applications.



Sueppong Mueanchamnong received the B.Sc degree in Statistics from the Department of Mathematics, Faculty of Science, Burapha University, Chonburi, Thailand, in 2024. He completed an internship with Solution Center Co., Ltd., the authorized distributor of Minitab in Thailand and a provider of training services in statistical applications using the Minitab software. His major field of study is statistics, with research interests focusing on forecasting, machine learning, and agricultural data analysis.

How to cite this paper: Pattharaporn Thongnim, Sueppong Mueanchamnong, "Improving Agricultural Commodity Trading through Data Imputation Methods for Price Prediction Accuracy", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.18, No.2, pp. 24-36, 2026. DOI:10.5815/ijieeb.2026.02.02