IJSA (2025), Vol. 9, No. 1, 1-9
DOI: https://doi.org/10.29244/ijsa.v9i1p1-9

Indonesian Journal of Statistics and Its Applications
Available at https://journal-stats.ipb.ac.id/index.php/ijsa
e-ISSN: 2599-0802

Classification of Rice Growth Phase Using Regression Logistic Multinomial Model and K-Nearest Neighbors Imputation on Satellite Data*

Fayyadh Ghaly[1], Yenni Kurniawati[2‡], Nonong Amalita[3], Dina Fitria[4]

[1,2,3,4]Department of Statistics, Universitas Negeri Padang, Indonesia
[‡]corresponding author: yennikurniawati@fmipa.unp.ac.id

Abstract

One of the efforts made by the government to maintain food security is to provide statistical data on rice production through accurate calculation of harvest areas using the area sampling framework approach. Although area sampling framework surveys produce accurate estimates, the costs required are quite high when applying this method. To overcome this problem, one solution that can be applied is to utilize satellite imagery to monitor the greenness index of plants using the enhanced vegetation index. However, in real conditions, the Landsat-8 optical satellite is susceptible to cloud cover, which results in missing data. This study aims to model the phase of rice plants using the regression logistic multinomial model by utilizing Landsat-8 satellites and k-nearest neighbors imputation handling to overcome missing data. The results showed that the model had varying performance in each phase, with an average balanced accuracy of 66.45%. This figure shows that the model can classify the area sampling framework data imputed using the k-nearest neighbors imputation method well. The model shows optimal performance in the late vegetative and generative phases but is less effective in detecting the harvest, puso, and non-rice paddy phases.

Keywords: Area Sampling Framework, Enhanced Vegetation Index, K-Nearest Neighbors Imputation, Landsat-8 Satellite, Multinomial Logistic Regression.

* Received: Dec 2024; Reviewed: Dec 2024; Published: Jan 2025

1. Introduction

Food security is a crucial global issue affecting people's welfare, economic stability, and national security. In Indonesia, rice is a strategic commodity that must be ensured in terms of availability because it serves as a staple food and the primary carbohydrate source, with a 93.8 kg per capita consumption rate in 2023 (BPN, 2024). The dominance of rice in consumption patterns makes paddy cultivation a top priority in maintaining food security. Therefore, effectively monitoring paddy production is essential to ensuring sustainable food availability and supporting national food policies.

To maintain food security, statistical data on rice production is needed so that the government can make the right decisions. Since 2018, the Central Bureau of Statistics (BPS) and the Agency for the Assessment and Application of Technology (BPPT) have used the Area Sampling Framework (ASF) method to estimate the rice harvest area (BPS, 2018). The ASF method adopts a scientific, objective, and measurable approach, offering advantages over traditional methods, which often result in biased and inaccurate estimates (Prasetyo et al., 2020). While ASF provides more accurate estimates, its implementation is costly and time-consuming, particularly in remote areas (Ruslan, 2019). These challenges highlight the need for a faster, more cost-effective, and accurate solution to enhance monitoring efficiency.

To overcome these limitations, the use of remote sensing technology through satellite imagery has emerged as an alternative. Satellite imagery such as Landsat-8 enables faster and more accurate monitoring in estimating the harvest area and identifying the growth phase of rice. Previous studies such as those conducted by Triscowati et al., (2019); Marsuhandi et al., (2020); Kurniawati, (2023) that combining satellite imagery with ASF methods and the use of various spectral indices, including the Enhanced Vegetation Index (EVI) can increase accuracy in monitoring the growth phase of rice. In addition, the use of satellite imagery at various observation periods provides a clearer picture of crop development, thus helping to better plan food production.

The use of Machine Learning methods such as Random Forest and Boosting has been widely used to classify the growth phase of rice based on satellite data. Although effective, these methods often face constraints in terms of interpretation of results, making it difficult to understand the factors that specifically affect rice growth. Alternatively, the Regression Logistic Multinomial Model (RLM) can be used to provide a clearer interpretation of the parameters affecting rice growth (Kurniawati et al., 2024). With RLM, the most influential factors on rice growth phases can be better identified, making the analysis results more informative and useful for policymakers.

While Landsat-8 satellite imagery improves the efficiency of agricultural monitoring, one of the main challenges is data loss due to cloud cover, especially in tropical regions like Indonesia. Clouds covering the Earth's surface can obscure satellite sensors, preventing important information on crop growth from being detected. Sinabutar et al., (2020) Mentioned that the more extensive the cloud cover, the greater the amount of information lost from satellite. The loss of data on EVI values can be categorized as Missing Not at Random (MNAR), because the missing data is caused by cloud cover that depends on unobserved information (Kurniawati et al., 2023). To overcome this problem, imputation methods such as K-Nearest Neighbors (KNN) are effective in filling in missing data by utilizing information from similar data (Umar &

Gray, 2023). Handling missing data is crucial to ensure accurate analysis and complete information (Fadlil et al., 2022).

This research aims to examine the utilization of the Regression Logistic Multinomial Model (RLM) in classifying rice growth phases using Landsat-8 satellite, as well as applying the KNN imputation method to overcome the problem of missing data due to cloud cover. By utilizing EVI as the main variable, this research is expected to provide a more accurate and efficient solution for monitoring rice production in Indonesia. The results of this research are expected to support efforts to strengthen national food security through more precise, effective, and sustainable monitoring.

2. Research Methods

This research is categorized as applied research. Applied research focuses on the application of knowledge and methods to solve practical problems in the real world. In the context of this research, the K-Nearest Neighbors (KNN) imputation method will be applied to overcome the problem of missing data in the Enhanced Vegetation Index (EVI) spectral index generated from Landsat-8 satellite. In addition, this study will evaluate the use of Regression Logistic Multinomial Model (RLM) for rice growth phase classification, to improve accuracy and efficiency in monitoring rice production in Indonesia.

The data used is secondary data obtained from an Area Sampling Frame (ASF) survey by the Central Bureau of Statistics in February 2024 in Padang Pariaman Regency. This research involved 61 sampling units and 549 observation units. In addition, remote sensing data from the Landsat-8 satellite will be used for three time periods in 2024. The results of KSA observations can be seen in Figure 1.

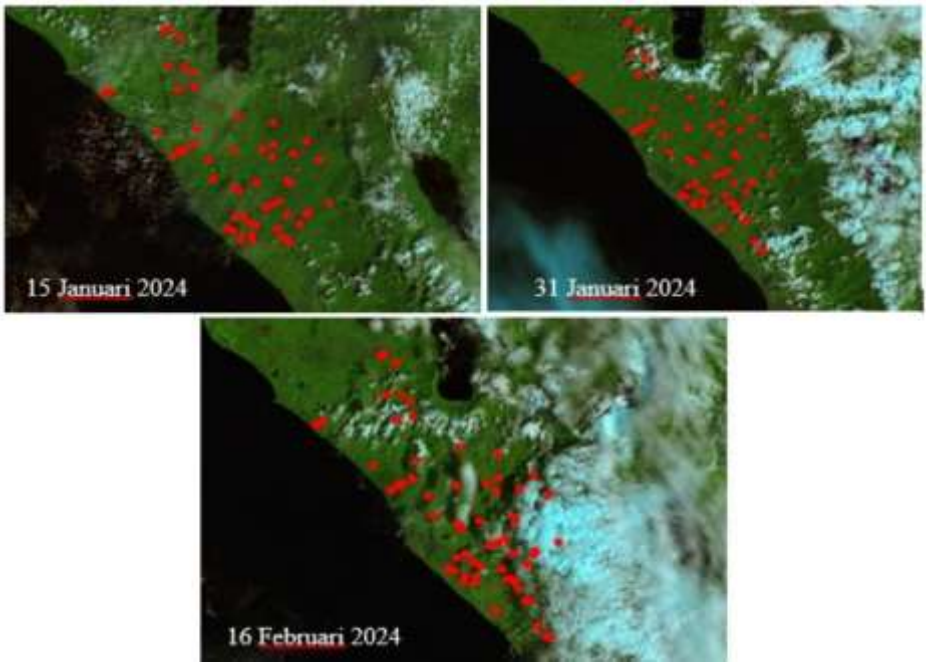

Figure 1: Landsat-8 Padang Pariaman View Period January 15, 2024, January 31, 2024, and February 16, 2024

This study uses three independent variables and one dependent variable. The independent variable consists of the EVI spectral index from three different periods, namely February 16, 2024, January 31, 2024, and January 15, 2024. The dependent variable is the ASF observation class, which includes eight classes. Table 1 provides details on the variables used in this study.

Table 1: Research Variable

No.	Variable	Variable Name	Description
1	Y	ASF Observation	1 : Early Vegetative 2 : Late Vegetative 3 : Generative 4 : Harvest 5 : Land Preparation 6 : Puso 7 : Non-Paddy Rice Field 8 : Not Rice Field
2	X_1	EVI_t	EVI value around survey time (t); t = {February 16, 2024}
3	X_2	EVI_{t-1}	EVI value one period before t; (t-1) = {January 31, 2024}
4	X_3	EVI_{t-2}	EVI value two periods before t; (t-2) = {January 15, 2024}

Data Analysis Technique

1. Prepare the data to be used, namely the Landsat-8 Enhanced Vegetation Index (EVI) spectral index data for 3 periods from January 2024 to February 2024.
2. Exploring data using bar charts for response variables. The use of bar charts is used to see the frequency of values per category of ASF observations.
3. Preprocessing was done to check the completeness of the data by applying the K-Nearest Neighbors (KNN) imputation method to fill in the missing EVI values. This method ensures that the data used in subsequent analysis is complete and free from missing data problems.
4. Divide the data into training data and testing data proportionally so that each segment is represented. The proportion used is 70% for training data and 30% for testing data.
5. Form a Regression Logistic Multinomial (RLM) model.
6. Evaluate the model obtained in step 6 using the confusion matrix by calculating the accuracy, sensitivity, specificity, and balanced accuracy values on the testing data.
7. Interpretation of the results obtained.

3. Result and Discussion

Data exploration was conducted to obtain preliminary information about the data. KSA observation categories that were multiclass were analyzed using bar charts. Of the 549

segments used in the study, more than 100 segments were dominated by the "Generative" class, as shown in Figure 2. In contrast, the "Puso" class was the last segment, because puso conditions rarely occur during the rice planting phase. The figure also shows that the frequency of ASF observation categories is not balanced.

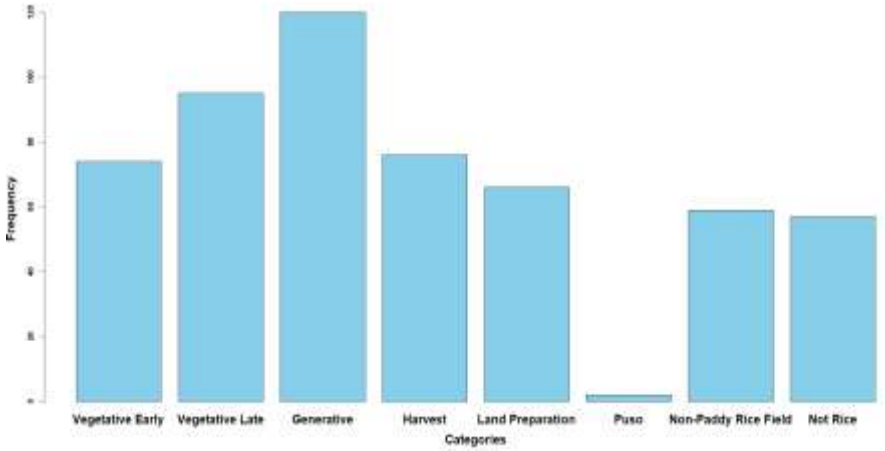

Figure 2: Frequency of ASF Observation Categories

After conducting data exploration, the next step is to ensure the completeness of the data to be used in further analysis. In this study, the process of identifying missing data was carried out on EVI variables generated from Landsat-8 satellite. EVI data were taken from three time periods, namely February 16, 2024 (EVI_t), January 31, 2024 (EVI_{t-1}), and January 15, 2024 (EVI_{t-2}). One of the main causes of EVI data loss is cloud cover at the time of satellite image capture. Landsat-8 satellite images are often unable to penetrate thick clouds, so the surface reflectance information needed to calculate EVI cannot be obtained completely. As a result, there is missing data that needs to be addressed to ensure a more accurate analysis.

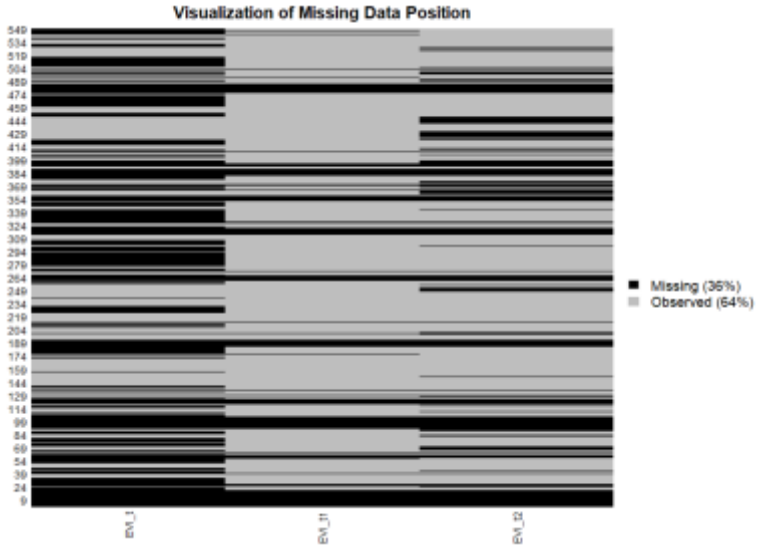

Figure 3: Distribution of Missing Data Positions

The visualization in Figure 3 identifies that 36% of the data in the three observation periods are missing, with the missing pattern randomly distributed. In some rows, missing data occurs simultaneously in all periods, while in other rows it only occurs in one period. This suggests that the missing data is most likely MNAR in nature, caused by cloud cover blocking the capture of satellite. To address this issue, an imputation method using KNN with a value of K = 3 was applied, which successfully filled in the missing data and recovered the entire dataset. A comparison of the missing data before and after imputation can be seen in Table 2.

Table 2: Comparison of Data Before and After Imputation

No	Period	Number of Data	Data Before Imputation		Data After Imputation	
			Number of Missing Data	Percentage	Number of Missing Data	Percentage
1	EVI_t	549	301	54.83%	0	0%
2	EVI_{t-1}	549	113	20.58%	0	0%
3	EVI_{t-2}	549	178	32.42%	0	0%
Amount		1647	592	35.94%	0	0%

After imputation, the data was divided into training data (70%) and testing data (30%) for further analysis. From this division, 389 data for training and 160 data for testing were obtained. The training data was used to train the model, while the testing data was used to test the model's ability to classify the growth phase of rice effectively. Parameter estimates using RLM on the training data are presented in Table 3.

Table 3: RLM Parameter Estimator

Logit	Intercept	X_1	X_2	X_3
Early Vegetative	13.3549	-28.0619	-13.1234	-37.8288
Late Vegetative	11.5423	-2.9223	10.7817	-74.4195
Generative	5.2768	-20.6547	15.7196	-18.0567
Harvest	6.1487	-23.9272	16.9682	-23.5309
Land Preparation	11.3199	-31.2112	-13.5324	-20.2366
Puso	2.8016	-33.6943	-6.4883	7.4386
Non-Paddy Rice Field	5.9407	-14.3257	2.7705	-17.3771

To assess the feasibility of the model obtained, a confusion matrix was used to calculate the classification accuracy of the testing data, which included 160 observations. Model evaluation is done by calculating several main criteria from the confusion matrix, namely accuracy, sensitivity, specificity, and balanced accuracy. These criteria provide an overall picture of the model's performance in classifying rice growth phases, by measuring the extent to which the model can identify the correct category and distinguish between different categories. The values of the three confusion matrix criteria are presented in Table 4.

Table 4: RLM Confusion Matrix Criteria Value

ASF Observation	Criteria		
	Sensitivity	**Specificity**	**Balanced Accuracy**
Early Vegetative	69.23%	90.48%	79.85%
Late Vegetative	71.64%	91.28%	81.46%
Generative	85.71%	72.04%	78.88%
Harvest	00.00%	99.10%	49.55%
Land Preparation	36.17%	92.67%	64.42%
Puso	00.00%	100.00%	50.00%
Non-Paddy Rice	00.00%	99.42%	49.71%
Not Rice Field	60.00%	95.40%	77.70%
Average	**40.34%**	**92.55%**	**66.45%**

Based on Table 4, the evaluation results of the Regression Logistic Multinomial (RLM) model, Balanced Accuracy averaged 66.45%, indicating that the model has a fairly good ability to differentiate between the various rice growth phases as well as other categories. The vegetative phase, especially Late Vegetative and Early Vegetative, showed the best performance with a Balanced Accuracy of 81.46% and 79.85% respectively, meaning that the model was able to detect the vegetative phase of rice with high accuracy. This result is important because the vegetative phase is one of the key stages in the rice growth cycle that greatly affects production yield. The strong performance in this phase reflects the potential of the model to be applied in effectively monitoring the early development of rice plants.

However, the model faces significant challenges in detecting the Harvest and Puso phases, which is evident from the low Balanced Accuracy of 49.55% and 50.00%, respectively. This low value is mainly due to the Sensitivity reaching zero, indicating that the model fails to fully detect both the harvest phase and puso (crop failure) cases. This is a critical weakness as these phases are crucial for monitoring agricultural yields and crop failure-related disasters. However, the high Specificity in these two phases shows that the model can recognize non-harvest and non-desert cases well, but needs to be improved to detect actual events. Similar challenges were also seen in the Non-Paddy Rice Field category, which only recorded a Balanced Accuracy of 49.71%, signaling the difficulty in distinguishing non-paddy fields from other categories.

On the other hand, the Generative phase performed quite well with a Balanced Accuracy of 78.88%, although Specificity in this phase is relatively lower than the vegetative phase. The model's performance in detecting the generative phase, which is the stage where rice starts to form seeds, remains important for monitoring crop productivity. Meanwhile, the Not Rice Field category also showed satisfactory results with a Balanced Accuracy of 77.70%, indicating that the model is quite effective in distinguishing areas that are not rice fields. This ability to identify non-rice fields is relevant in the context of land management and agricultural area mapping.

Overall, while the RLM model performed well in the vegetative phase and in distinguishing non-fields, there were significant weaknesses in detecting the Harvest and Puso phases. The results of this evaluation suggest that data enhancement or enrichment in certain phases is required to improve the prediction accuracy of the model, especially for phases that have low Sensitivity. Improving the model's performance in detecting critical phases such as harvest and puso will be crucial to improving overall agricultural monitoring, particularly in the context of yield management and crop failure risk mitigation

4. Conclusion

This study successfully demonstrated that the Regression Logit Multinomial Model (RLM) coupled with the K-Nearest Neighbors (KNN) imputation method to fill in missing data due to cloud cover can be effectively used to classify rice growth phases using Landsat-8 and Enhanced Vegetation Index (EVI) satellite data. The RLM model shows a good capability with an average Balanced Accuracy of 66.45%, especially in detecting the vegetative phase, which is an important stage in the rice growth cycle. The Late Vegetative and Early Vegetative phases performed the best, with Balanced Accuracy of 81.46% and 79.85%, respectively. However, significant weaknesses emerged in the Harvest and Puso phases, which had low Balanced Accuracy due to zero Sensitivity, indicating the model was unable to accurately detect these phases. This is a critical weakness that needs to be corrected given the importance of these phases in agricultural yield monitoring. To improve the accuracy of the model, data enrichment is required, especially in the harvest and puso phases, which are important to support agricultural yield monitoring and mitigate the risk of crop failure.

References

BPN. (2024). Situasi Konsumsi Pangan Nasional Tahun 2023. In *Badan Pangan Nasional*. Jakarta Selatan: Badan Pangan Nasional.

BPS. (2018). *Upaya Perbaikan Data Padi Dengan Metode Kerangka Sampel Area (KSA) 2018*.

Fadlil, A., Herman, & Praseptian M, D. (2022). K Nearest Neighbor Imputation Performance on Missing Value Data Graduate User Satisfaction. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(4): 570–576. https://doi.org/10.29207/resti.v6i4.4173

Kurniawati, Y. (2023). *Penduga Area Kecil Berhirarki untuk Luas Panen Padi Berbasis Survei KSA-BPS dengan Memanfaatkan Citra Satelit LANDSAT 8*. 1–130.

Kurniawati, Y., Wijayanto, H., Kurnia, A., Dirgahayu D, D., & Susetyo, B. (2024). Rice phenology monitoring via ensemble classification for an extremely imbalanced multiclass dataset of hybrid remote sensing. *Remote Sensing Applications: Society and Environment*, *35*: 101246. https://doi.org/10.1016/J.RSASE.2024.101246

Kurniawati, Y., Wijayanto, H., Kurnia, A., Domiri, D. D., & Susetyo, B. (2023). Selection

of Multinomial Logit Models Based on Accuracy Reclassification of the Area Sampling Frame Labels. *Science and Technology Asia*, *28*(2): 18–30. https://doi.org/10.14456/scitechasia.2023.23

Marsuhandi, A. H., Soleh, A. M., Wijayanto, H., & Domiri, D. D. (2020). Pemanfaatan Ensemble Learning Dan Penginderaan Jauh Untuk Pengklasifikasian Jenis Lahan Padi. *Seminar Nasional Official Statistics*, *2019*(1): 188–195. https://doi.org/10.34123/semnasoffstat.v2019i1.247

Prasetyo, O. R., Kadir, & Amalia, R. R. (2020). A pilot project of area sampling frame for maize statistics: Indonesia s experience. *Statistical Journal of the IAOS*, *36*(4): 997–1006. https://doi.org/10.3233/SJI-200743

Ruslan, K. (2019). Improving Indonesia's Food Statistics through the Area Sampling Frame Method. *Center for Indonesian Policy Studies*, 34.

Sinabutar, J. J., Sasmito, B., & Sukmono, A. (2020). Studi Cloud Masking Menggunakan Band Quality Assessment, Function of Mask Dan Multi-Temporal Cloud Masking Pada Citra Landsat 8. *Jurnal Geodesi Undip Agustus*, *9*(3): 51–60.

Triscowati, D. W., Sartono, B., Kurnia, A., Domiri, D. D., & Wijayanto, A. W. (2019). *Multitemporal remote sensing data for classification of food crops plant phase using supervised random forest*. *11311*: 10. https://doi.org/10.1117/12.2547216

Umar, N., & Gray, A. (2023). Comparing Single and Multiple Imputation Approaches for Missing Values in Univariate and Multivariate Water Level Data. *Water (Switzerland)*, *15*(8): 1–21. https://doi.org/10.3390/w15081519