Rice Yield Estimation Using Sentinel-2 with Remote Sensing and Machine Learning in Thailand

Pramet K., Jinnawat T., Phukrit S., Yutthaphum K., Sasithon C., Panu, N.*, and Sukji S.

Geo-Informatics and Space Technology Development Agency (GISTDA), Bangkok, Thailand

*panu.n@gistda.or.th (*Corresponding author's email only)

***Abstract*:** *This study explores the effectiveness of various vegetation indices derived from Sentinel-2 satellite data in predicting rice yield through simple and multiple linear regression models. The primary objective was to establish empirical models that relate vegetation indices to rice productivity at different growth stages. A total of ten vegetation indices, including NDVI (Normalized Difference Vegetation Index), GNDVI (Green Normalized Difference Vegetation Index), SAVI (Soil-Adjusted Vegetation Index), EVI (Enhanced Vegetation Index), GCVI (Green Chlorophyll Vegetation Index), DVI (Difference Vegetation Index), RGVI (Red-Green Vegetation Index), NDWI_GAO (Normalized Difference Water Index - Gao), NDWI_MF (Normalized Difference Water Index - Modified), and BSI (Bare Soil Index), were analyzed for their correlation with rice yield. The analysis revealed that NDVI, measured at 70 days after seeding, had the strongest positive correlation (0.83) with rice yield, indicating its effectiveness in reflecting plant health and productivity. SAVI also demonstrated a strong positive correlation (0.80). Conversely, BSI and NDWI_MF, assessed 40 days after tillering, showed significant negative correlations with rice yield, suggesting their relevance in identifying stress conditions. Simple linear regression models showed varying performance across indices, with NDVI_max providing the most accurate predictions (R^2 = 0.68) and GNDVI_max also performing well. However, a multiple linear regression model combining NDVI_max and BSI_min improved predictive accuracy significantly, achieving the highest R^2 (0.72), the lowest RMSE (74.92), and a MAPE of 8.47%. This model's robust performance on both training and test datasets underscores the value of integrating multiple indices for yield prediction. These findings confirm the robust relationship between rice yield and satellite imagery. This study underscores the importance of using remote sensing and machine learning techniques for agricultural monitoring, providing crucial insights to strengthen rice production management within GISTDA's agricultural application "Dragonfly".*

Keywords: Regression model, Yield estimation, Sentinel-2, Rice, Thailand

Introduction

Rice (Oryza sativa L.) is the third most cultivated food crop in the world, following wheat and maize. It is also the staple food for more than half of the world's population (FAO, 2000). Additionally, the livelihoods of millions of farmers depend directly or indirectly on rice production. Asia contributes to 90% of global rice production, with the primary cultivation areas located in the southern and southeastern countries of Asia.

Thailand is one of the leading rice producers, with the sixth largest rice cultivation area in Asia. In 2018, Thailand had a total rice cultivation area of 59.98 million rai, accounting for 49% of the total agricultural land used. The rice production amounted to 25.18 million tons, with rice exports totaling 11.8 million tons, valued at 149.6 billion baht, representing a 30%

share of the global rice market (Office of Agricultural Economics, 2018) .The major importers of Thai rice are China, the United States, the Philippines, and South Africa, respectively. Nakhon Sawan and Suphan Buri provinces, located in the central region of Thailand, are predominantly lowland areas ideal for rice cultivation and serve as major rice-producing regions.

In recent years, initiatives have been undertaken to increase rice production using remote sensing technology. Earth Observation (EO) data enables efficient monitoring of crop growth at the field level due to its high spatial resolution (~10 m) and temporal resolution (every 5 days). For example, ESA's Sentinel-2A mission provides high-frequency, high-resolution data for free, which can be utilized in agriculture. Various Vegetation Indices (VIs), calculated from light reflectance data obtained from satellite images, serve as effective indicators of crop status and are positively correlated with crop yield (Nazir et al., 2021). For instance, the Normalized Difference Vegetation Index (NDVI) is widely used for predicting crop yields and identifying growth stages. Similarly, other indices like NDWI_GAO and BSI are effectively used to monitor water changes during different growth stages.

Additionally, research has utilized algorithms based on growth stages and linear regression models to improve the accuracy of assessments, addressing challenges in agricultural activity monitoring and providing reliable information for efficient agricultural management (Martínez-Eixarch, Soriano González, & Alcaraz, 2022; European Space Agency, 2012). Multivariate analysis methods are often employed to support data analysis by using simple linear regression models and machine learning to develop and validate multivariate remote sensing models for yield estimation. These methods enhance the accuracy of crop yield estimates using remote sensing, especially when analyzing the quantitative relationships between remote sensing variables, derived from satellite images during different growth stages of rice, and crop yields. They can also be effectively used to monitor rice growth conditions and predict yields. Therefore, the application of remote sensing technology is a significant step towards increasing rice production and promoting future food security (Zhang et al., 2022a; Yu et al., 2022; Li et al., 2022; Zhang et al., 2022b).

The study focuses on the relationship between rice yield and satellite imagery, combined with remote sensing and machine learning in Thailand. The primary objective of this research is to predict rice yields and examine the relationship between vegetation indices obtained from remote sensing during different growth stages of the crop. Additionally, it

aims to integrate machine learning with Sentinel-2 satellite imagery. By using periodic satellite data, farmers can receive accurate and up-to-date information to make informed decisions, address agricultural issues, and monitor rice production management for farmers in Thailand.

Material and Methodology

a. Study area:

Nakhon Sawan Province is located in the upper part of central Thailand, covering an area of approximately 9,597 square kilometers. The terrain is predominantly lowland, making it suitable for agriculture, with about three-quarters of the province being flat land. The major rivers are the Ping, Yom, and Nan, which converge to form the Chao Phraya River. Additionally, there are small mountains scattered across various districts. Nakhon Sawan shares borders with several provinces: to the north with Phichit and Kamphaeng Phet, to the east with Phetchabun and Lopburi, to the south with Singburi, Chainat, and Uthai Thani, and to the west with Tak.

Suphan Buri Province is located in the western part of central Thailand, at an elevation of 3 to 10 meters above sea level, covering a total area of approximately 5,358 square kilometers. The terrain is predominantly lowland, with some areas being highland. Most of the province's land is used for rice farming, and there are numerous rivers, canals, and ponds scattered throughout the area. The main river running from the northernmost to the southernmost part of the province is the Tha Chin River, also known as the Suphan Buri River.

Both Nakhon Sawan and Suphan Buri provinces in central Thailand are widely known for rice cultivation due to the abundant water resources from the Chao Phraya and Tha Chin rivers, as well as an efficient irrigation system. The lowland areas in both provinces are ideal for rice farming, with fertile soils and a suitable climate. Additionally, government support in developing irrigation systems, promoting agricultural technology, and providing financial assistance to farmers has made rice cultivation in Nakhon Sawan and Suphan Buri a key economic activity, ensuring stable income for farmers in these regions.

Figure 1: Study area

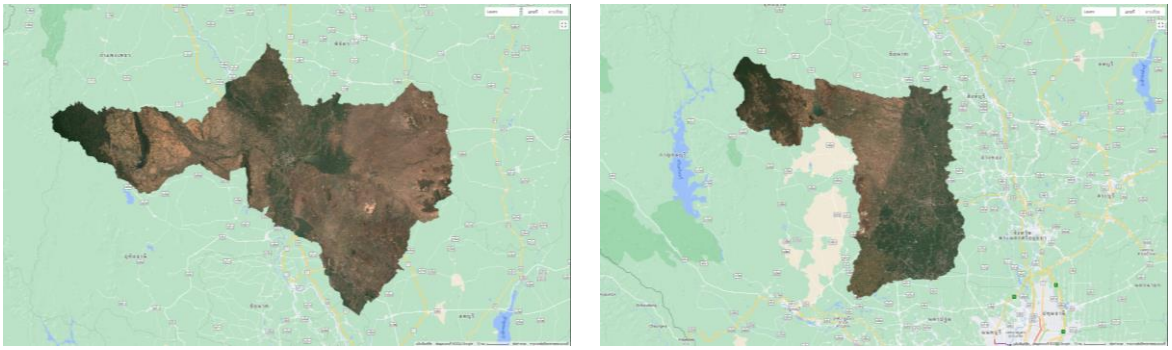
b. Field data and satellite data:

Field data: Field data collection involved gathering yield information for individual plots, along with detailed cultivation information, including the dates of seeding, tillering, flowering, maturity, and harvesting. Data were collected from a total of 50 plots in the central region, with 24 plots from Nakhon Sawan province and 26 plots from Suphan Buri province. The field data for this research were collected during the cultivation period from December 2018 to April 2019, while the harvest data were collected from March to June 2019.

(a) (b)

Figure 2: Field data imagery: (a) Nakhon Sawan province area, (b) Suphan Buri province area.

Satellite data: Satellite imagery was obtained from Sentinel-2A and Sentinel-2B satellites using the Multi-Spectral Imager (MSI) system, which consists of 13 spectral bands covering wavelengths from the visible to shortwave infrared bands. The spatial resolution is 10, 20, and 60 meters depending on the wavelength, and both satellites capture images of the same location every 5 days (European Space Agency, 2012) The details the bands of satellite data as shown in Table 1. For this study, satellite images from Sentinel-2A and Sentinel-2B were collected between January 1, 2019, and June 30, 2019, using Google Earth Engine (GEE) to capture data during the rice growing season.



(a) (b)

Figure 3: Sentinel-2 satellite imagery: (a) Nakhon Sawan province area, (b) Suphan Buri province area.

Table 1: Example of a Table Caption.

Bands	**Spatial resolustion (m)**	**Central wavelength (nm)**	**Bandwidth**
Band 1 (Aerosols)	60	443	20
Band 2 (Blue)	10	490	65
Band 3 (Green)	10	560	35
Band 4 (Red)	10	665	30
Band 5 (Red Edge 1)	20	705	15
Band 6 (Red Edge 2)	20	740	15
Band 7 (Red Edge 2)	20	783	20
Band 8 (NIR)	10	842	115
Band 8A (Red Edge 4)	20	865	20
Band 9 (Water vapor)	60	945	20

Table 1: Example of a Table Caption.

Bands	Spatial resolustion (m)	Central wavelength (nm)	Bandwidth
Band 10 (SWIR-Cirrus)	60	1375	30
Band 11 (SWIR 2)	20	1610	90
Band 12 (SWIR 2)	20	2190	180

All vegetation indices of satellite data are provided in Table 2. The process of analyzing various indices using reflectance data from Sentinel-2 satellite images (Harmonized Level-2A: COPERNICUS/S2_SR_HARMONIZED) was conducted through Google Earth Engine (GEE). Cloud and shadow filtering was applied using the Scene Classification Layer (SCL). The filtered data were then used to analyze various indices and determine their correlation with plot-level rice yield data collected from the field. This was done by calculating the average reflectance values within the plot boundaries using the Zonal Statistics method. Data from January to June 2019 were used, and the most appropriate time frames for rice yield assessment were selected by extracting key growth stage characteristics (phenological stages), including the maximum, minimum, and inflection points of the reflectance indices related to changes in soil, waterlogging, or plant growth stages (Zheng et al., 2016; Liu et al., 2017). This research hypothesized that key rice characteristics would be extracted 70 days after seeding to determine the maximum values of vegetation indices (NDVI , GNDVI, SAVI, EVI, GCVI, DVI, and RGVI(Franch et al., 2021; Nazir et al., 2021; Choudhary et al., 2021; Suksiri, 2015; Geo-Informatics and Space Technology Development Agency (GISTDA), 2024). Additionally, 40 days after tillering, the minimum values of the BSI, NDWI_GAO, and NDWI_MF (Soriano-González et al., 2022). indices were identified.

Table 2: Example of a Table Caption.

Vegetation Indices	Equation
Normalized Difference Vegetation Index : NDVI	$NDVI = \frac{NIR - RED}{NIR + RED}$
Green Normalized Difference Vegetation Index : GNDVI	$GNDVI = \frac{NIR - GREEN}{NIR + GREEN}$
Soil Adjusted Vegetation Index : SAVI	$SAVI = \frac{(NIR - RED)}{(NIR + RED + L)} \times (1 + L)$
Normalized Difference Water Index : NDWI_MF	$NDWI = \frac{GREEN - NIR}{GREEN + NIR}$

Table 2: Example of a Table Caption.

Vegetation Indices	Equation
Normalized Difference Water Index : NDWI_GAO	$NDWI = \frac{NIR - SWIR}{NIR + SWIR}$
Enhanced Vegetation Index : EVI	$EVI = G \times \frac{(NIR - RED)}{(NIR + C_1 \times RED - C_2 \times BLUE + L)}$
Green Chlorophyll Vegetation Index : GCVI	$GCVI = \frac{NIR}{GREEN} - 1$
Bare soil index : BSI	$BSI = \frac{(RED + SWIR1) - (NIR + BLUE)}{(RED + SWIR1) + (NIR + BLUE)}$
Difference Vegetation Index : DVI	$DVI = NIR - RED$
Rice Growth Vegetation Index : RGVI	$RGVI = 1 - \frac{(BLUE + RED)}{(NIR + SWIR1 + SWIR2)}$

c. Simple Linear regression analysis:

Linear regression is indeed a powerful and well-established algorithm used in both statistical and machine learning contexts. The linear regression aims to find the best-fitting linear relationship between a dependent variable (often denoted as Y) and one or more independent variables (often denoted as x). The linear regression equation for a simple linear regression model with one independent variable is typically expressed as:

$$Y = mx + b$$

Where Y is the dependent variable (the variable being predicted or explained), x is the independent variable (the variable used to make predictions), m is the slope of the line (the change in Y for a unit change in x), b is the y-intercept (the value of Y when x is zero).

In this research, simple linear regression was used to analyze the relationship between rice yield and satellite imagery data. The goal was to create a model to predict rice yield using remote sensing technology. A multivariate linear regression model was constructed to relate rice yield to surface reflectance values at the pixel level, corresponding to specific indices, in order to find the best model that provides performance indicators compared to actual rice yield data.

When there are multiple variables in the linear regression model, a stepwise selection method was used to identify the best set of independent variables. The stepwise selection method evaluates the partial correlation coefficient between the independent variables and Y, selecting the independent variable that provides the highest partial correlation coefficient with the dependent variable. If $F_0 < F_{out} = F_{a,(1,n-2)}$, or $|t_0| < t_{out} = t_{\frac{a}{2},(n-2)}$, or $p\_value < p_{in}$, it indicates that the next independent variable under consideration is important for predicting Y, given that the previous independent variables are already included in the regression model. The independent variable is then added to the model.

After implementing the linear regression model, the model's performance was evaluated using indicators such as R^2, $MAPE$ and $RMSE$:

Coefficient of Determination: The coefficient of determination (R^2) is a statistic used to measure the performance of a prediction model. The value of R^2 ranges from 0 to 1 and is unitless. If the R^2 value is close to 1, it indicates that the predictive equation has high efficiency. Conversely, if the R^2 value is close to 0, it suggests that the predictive model has low efficiency. The formula is as follows:

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y - \bar{Y}_i)^2}$$

The Root Mean Square Error ($RMSE$): The Root Mean Square Error ($RMSE$) is a metric used to measure the differences between the predicted values and the observed values. It is commonly used to assess the accuracy of a model's predictions. $RMSE$ provides an absolute measure of fit, with lower values indicating better fit and higher accuracy. $RMSE$ is expressed in the same units as the dependent variable, and a lower $RMSE$ value indicates a better performing model. $RMSE$ is calculated using the following formula:

$$RMSE = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n}}$$

The Mean Absolute Percentage Error ($MAPE$): The Mean Absolute Percentage Error ($MAPE$) is a metric used to measure the accuracy of a model in predicting values. It calculates the percentage difference between the predicted and actual values and expresses the average error as a percentage. A lower $MAPE$ value indicates higher model accuracy, as it reflects smaller percentage errors between predicted and actual values. $MAPE$ is commonly used because it is easy to interpret and understand as a percentage. The formula for $MAPE$ is as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \times 100\ \%$$

Where Y_i represents the observed rice yield, $\hat{Y}_i$ represents the predicted value from the model, $\bar{Y}_i$ represents the average of the observed rice yield, and n is the total number of samples in the dataset.

Results and Discussion

In this research, a simple linear regression was performed to explore the relationship between vegetation indices and rice yield. This investigation aimed to establish an empirical model elucidating how vegetation indices data can represent rice yield data. The simple linear regression model is represented by equation (1-10):

$$Rice\_Yield_{NDVI_{\max}} = \beta_0 + (\beta_1 \times NDVI_{\max}) \quad (1)$$

$$Rice\_Yield_{GNDVI_{\max}} = \beta_0 + (\beta_1 \times GNDVI_{\max}) \quad (2)$$

$$Rice\_Yield_{SAVI_{\max}} = \beta_0 + (\beta_1 \times SAVI_{\max}) \quad (3)$$

$$Rice\_Yield_{EVI_{\max}} = \beta_0 + (\beta_1 \times EVI_{\max}) \quad (4)$$

$$Rice\_Yield_{GCVI_{\max}} = \beta_0 + (\beta_1 \times GCVI_{\max}) \quad (5)$$

$$Rice\_Yield_{DVI_{\max}} = \beta_0 + (\beta_1 \times DVI_{\max}) \quad (6)$$

$$Rice\_Yield_{RGVI_{\max}} = \beta_0 + (\beta_1 \times RGVI_{\max}) \quad (7)$$

$$Rice\_Yield_{NDWI\_gao_{\min}} = \beta_0 + (\beta_1 \times NDWI\_gao_{\min}) \quad (8)$$

$$Rice\_Yield_{NDWI\_mf_{\min}} = \beta_0 + (\beta_1 \times NDWI\_mf_{\min}) \quad (9)$$

$$Rice\_Yield_{BSI_{\min}} = \beta_0 + (\beta_1 \times BSI_{\min}) \quad (10)$$

* $RY = Rice\_Yield$

Where rice yield ($Y\ or$ dependent variable) represents the observed rice productivity, vegetation index ($X\ or\ in$dependent variable) stands for the calculated value from satellite images, β_0 is the intercept term, indicating the value of rice yield when the vegetation index is zero, and β_1 is the coefficient associated with the vegetation index, representing the change in rice yield for a one-unit change in the vegetation index.

a. Statistical analysis:

During the 2019 rice yield study, the analysis focused on understanding how various vegetation indices, obtained from Sentinel-2 satellite data, correlated with rice yields. The goal was to identify which indices were the most reliable indicators of rice growth at different stages of the crop's life cycle. A total of 10 different vegetation indices were examined: NDVI, GNDVI, SAVI, EVI, GCI, DVI, RGVI, NDWI_GAO, NDWI_MF, and BSI. Each of these indices provides different insights into the health and development of crops, as they are calculated based on the reflectance of light from the Earth's surface in specific wavelengths, as illustrated in Figure 4.

The analysis identified two key timeframes:

- 70 Days After Seeding:

  At this stage, the rice plants are well into their growth, and certain vegetation indices have been shown to strongly reflect their overall health. NDVI (Normalized Difference Vegetation Index), which measures vegetation greenness, had the highest positive correlation with rice yield, with a value of 0.83 (Figure 4). This means that higher NDVI values, typically indicating more vigorous plant growth and healthier biomass, were strongly associated with

higher rice yields. SAVI (Soil-Adjusted Vegetation Index), a modified version of NDVI that reduces the influence of soil brightness, also had a strong positive correlation of 0.80 (Figure 4). This suggests that SAVI is another reliable index for predicting rice yields, especially in conditions where the soil background may affect measurements.

- 40 Days After Tillering:

  During this period, the analysis looked for indices that negatively correlated with rice yield, which could signal stress or less favorable growing conditions. BSI (Bare Soil Index), which measures the presence of bare soil and the condition of the soil, had the strongest negative correlation with rice yield, with a value of -0.72 (Figure 4). This suggests that areas where BSI was high (indicating more exposed soil or poorer soil conditions) were linked to lower rice yields. NDWI_MF (Normalized Difference Water Index - Modified), an index that measures water content in the vegetation, had a negative correlation of -0.63 (Figure 4). A lower NDWI_MF value here might indicate water stress or other issues with water availability that negatively impacted rice yield.

The heat map analysis (Figure 4) showed that during the early stages of rice growth (70 days after seeding), NDVI and SAVI were the most accurate indicators of healthy rice plants and higher yields. Meanwhile, as the rice plants reached the post-tillering stage (40 days after tillering), indices like BSI and NDWI_MF that measure soil and water conditions were most closely related to lower yields. These findings highlight the importance of monitoring specific vegetation indices at different stages of rice growth to better predict final yields.

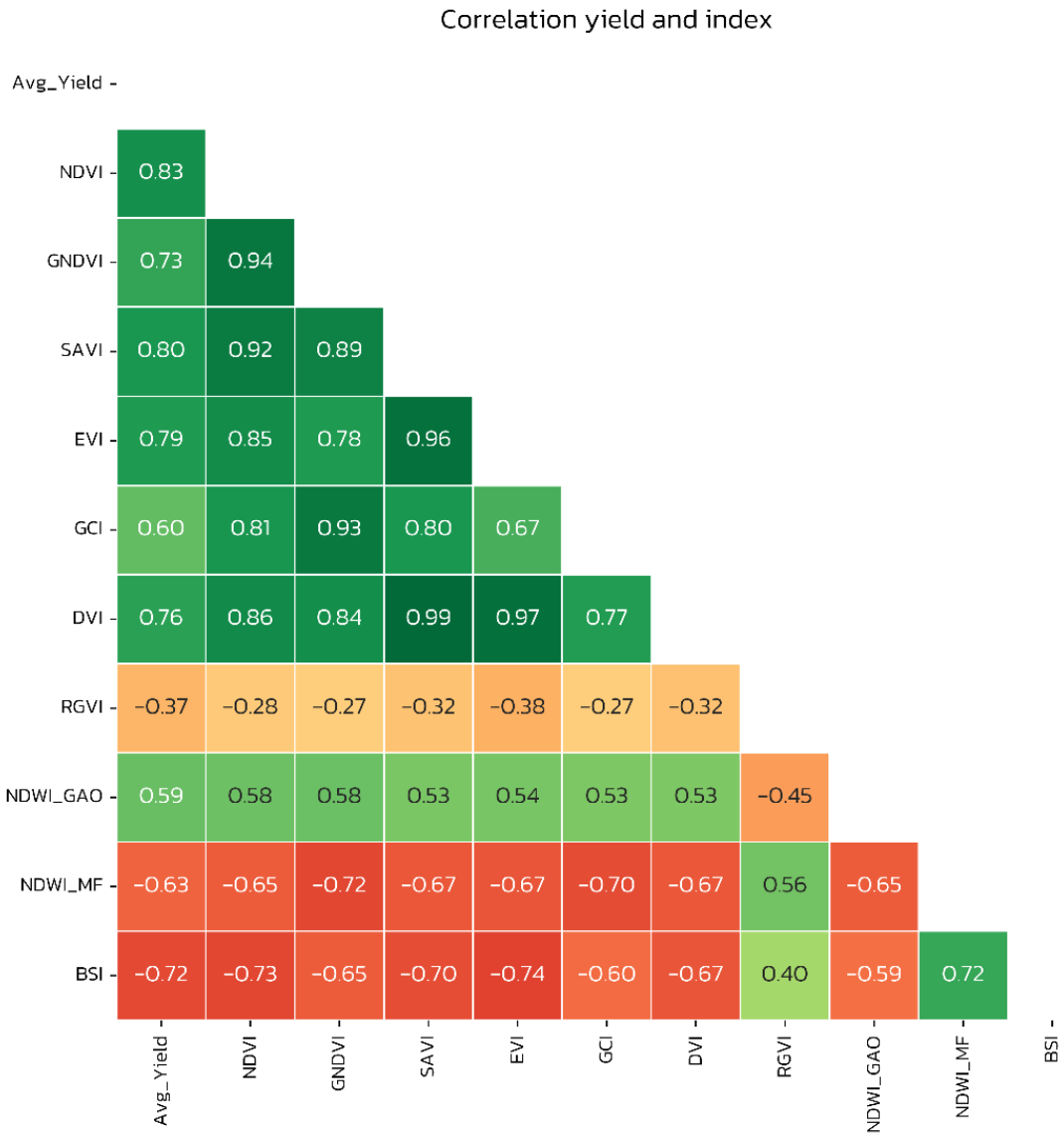


Figure 4: Heat map analysis between vegetation indices and rice yield in this research.

The Figure 5 presents the linear relationships between field yield and predicted yield from various vegetation indices used for yield prediction in agriculture. The subplots (a) to (j) display the correlation between field-measured yield values and those predicted by different indices, represented by scatter plots with corresponding trend lines for each index.

In each case, the scatter plots demonstrate a positive correlation between predicted and observed yield values, with the red lines indicating the fitted linear regression models. The trend in the plots shows that most indices yield reasonable predictions of actual field yield, although the strength of the relationship varies across different indices, likely due to differences in sensitivity to vegetation, soil, and environmental factors.

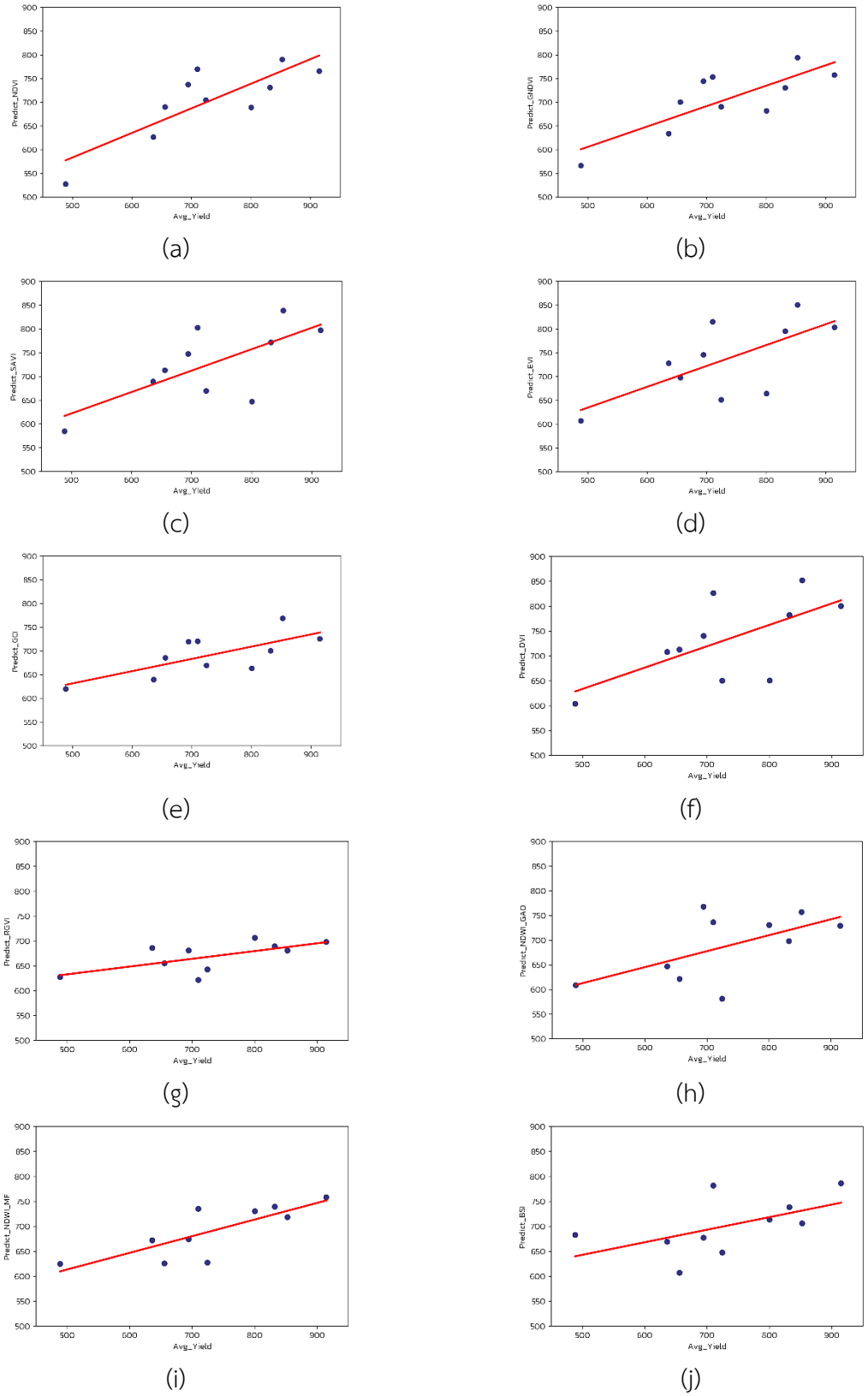


Figure 5: Linear relationships between field yield and predicted yield from each index: (a) NDVI (b) GNDVI (c) SAVI (d) EVI (e) GCVI (f) DVI (g) RGVI (h) NDWI_gao (i) NDWI_mf and (j) BSI.

The linear and multiple regression models (Table 3.) detailed in this research aim to predict rice yield based on vegetation indices derived from Sentinel-2 satellite data. The models consider several indices, including NDVI, GNDVI, SAVI, EVI, GCVI, DVI, RGVI, NDWI, and BSI, to identify which best correlates with and predicts rice yield in specific growth stages. The simple linear regression equations demonstrate varying degrees of success in predicting rice yield based on a single vegetation index. Among these, the model using the maximum NDVI (Eq. 11) yielded the best results, with an R^2 of 0.68, a relatively low RMSE of 75.68, and a MAPE of 8.18%. This suggests that NDVI_max is highly correlated with rice yield, explaining 68% of the variability in the yield data, and is capable of making accurate predictions. Following NDVI, GNDVI_max (Eq. 12) also performed well, with an R^2 of 0.64, an RMSE of 81.28, and a MAPE of 9.22%, making it another strong indicator of rice yield. Both indices reflect the health and density of vegetation, which are critical factors in rice growth. SAVI_max (Eq. 13) and EVI_max (Eq. 14) showed lower predictive power, with R^2 values of 0.49 and 0.46, respectively. These models have slightly higher error rates, suggesting that while they are useful, they are less precise in predicting rice yield compared to NDVI and GNDVI. Other indices, such as GCVI_max (Eq. 15) and DVI_max (Eq. 16), also show moderate correlations, with R^2 values of 0.52 and 0.42, respectively. However, models using indices like RGVI_max (Eq. 17) and BSI_min (Eq. 20) demonstrated weaker performance, with R^2 values as low as 0.30, indicating that these indices are less reliable predictors of rice yield.

The multiple linear regression model combining NDVI_max and BSI_min (Eq. 21) significantly improves predictive accuracy. This model has an R^2 of 0.72, which is the highest among all models, meaning it explains 72% of the variability in rice yield. Furthermore, it has the lowest RMSE (74.92) and a MAPE of 8.47%, indicating improved prediction accuracy compared to models using a single index.

This result highlights the benefit of using a combination of vegetation indices in predicting rice yield. While NDVI_max is a strong predictor on its own, incorporating BSI_min improves the model's ability to capture the variability in rice yield, likely due to the complementary nature of the indices. NDVI primarily reflects vegetation health, while BSI provides information on soil brightness, which could be particularly relevant in agricultural landscapes. The analysis underscores the importance of NDVI and BSI in predicting rice yield, particularly when combined. Using multiple indices allows for a more comprehensive and accurate prediction model, contributing to the advancement of remote sensing applications in agricultural monitoring and yield prediction.

Table 3: Simple linear regression Multiple Linear Regression equation (stepwise selection) and statistical analysis (R^2, $RMSE$, $MAPE$)

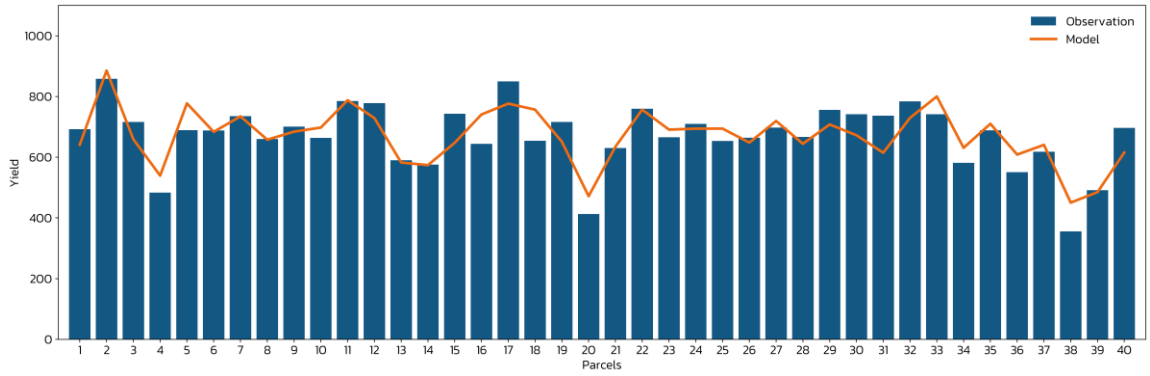
Linear Regression equation	**Eq**	R^2	$RMSE$	$MAPE$
$RY_{NDVI_{\max}} = -35.37 + (978.90 \times NDVI_{max})$	(11)	0.68	75.68	8.18
$RY_{GNDVI_{\max}} = 13.38 + (1058.02 \times GNDVI_{max})$	(12)	0.64	81.28	9.22
$RY_{SAVI_{\max}} = 203.06 + (1039.39 \times SAVI_{max})$	(13)	0.49	84.24	10.61
$RY_{EVI_{\max}} = 229.31 + (836.64 \times EVI_{max})$	(14)	0.46	86.93	11.13
$RY_{GCVI_{\max}} = 503.5 + (455.31 \times GCVI_{max})$	(15)	0.52	100.36	10.81
$RY_{DVI_{\max}} = 311.25 + (1367.92 \times DVI_{max})$	(16)	0.42	89.93	11.42
$RY_{RGVI_{\max}} = 1348.40 + (-6399.86 \times RGVI_{max})$	(17)	0.42	119.20	13.46
$RY_{NDWI\_gao_{\min}} = 667.90 + (444.10 \times NDWI\_gao_{min})$	(18)	0.35	104.13	12.18
$RY_{NDWI\_mf_{\min}} = 237.73 + (-740.61 \times NDWI\_mf_{min})$	(19)	0.63	93.19	11.05
$RY_{BSI_{\min}} = 431.18 + (-1128.24 \times BSI_{min})$	(20)	0.30	103.30	12.88
Multiple Linear Regression equation (stepwise selection)	**Eq**	R^2	$RMSE$	$MAPE$
$RY_{NDVI_{max}\ \&\ BSI_{\min}} = 58.8168 + (721.58 \times NDVI_{max}) + (-430.62 \times BSI_{min})$	(21)	0.72	74.92	8.47

* RY= $Rice\_Yield$

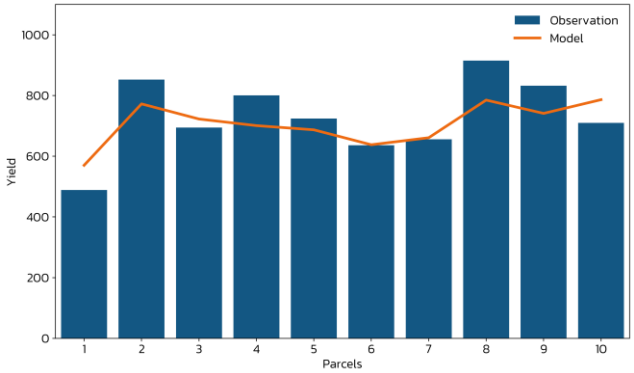
The Figure 6 presents the comparison between observed and predicted yield values based on a multiple linear regression (MLR) model using stepwise selection. The model uses the normalized difference vegetation index (NDVI) and bare soil index (BSI) as key predictors for yield estimation. The Figure 6(a) represents the training data (80% of the total data), where each bar corresponds to the observed yield values for 40 different parcels. The orange line represents the predicted yield values from the MLR model. The close alignment between the observed (bars) and predicted (line) values indicates that the model performs well in predicting the yield within the training dataset.

The Figure 6(b) shows the model's performance on the test data (20% of the total data), consisting of 10 parcels. Here, the difference between the observed and predicted values is

slightly more pronounced than in the training set, as is typical when testing a model on unseen data. Nonetheless, the predicted yield values generally follow the same trend as the observed values, suggesting that the model is effective at generalizing its predictions beyond the training data.



(a)



(b)

Figure 6: presents the comparison between observed and predicted yield values based on a multiple linear regression (a)

This model was developed using stepwise regression, a method that iteratively includes or removes predictors based on their statistical significance, ensuring that only the most relevant variables (in this case, NDVI and BSI) were retained in the final equation. The multiple linear regression equation is:

$$RY_{NDVI_{max}\ \&\ BSI_{\min}} = 58.8168 + (721.58 \times NDVI_{max}) + (-430.62 \times BSI_{min})$$

This model was developed using stepwise regression, a method that iteratively includes or removes predictors based on their statistical significance, ensuring that only the most

relevant variables (in this case, NDVI and BSI) were retained in the final equation. The multiple linear regression equation is:

The model achieved an R^2 value of 0.72, indicating that 72% of the variability in crop yield is explained by the two predictors (NDVI and BSI). The root mean square error ($RMSE$) of 74.92 reflects the average deviation of predicted yield values from the observed values, while the mean absolute percentage error ($MAPE$) of 8.47 shows that the model has a relatively low percentage error in its predictions.

The results show that vegetation indices such as NDVI, combined with other spectral indices like BSI, can serve as reliable predictors of crop yield. The integration of these indices in a multiple regression framework enhances the accuracy of yield estimation, which is crucial for agricultural planning and decision-making. The good performance on both the training and test datasets underscores the model's robustness and its potential application in precision agriculture.

Conclusion and Recommendation

This study successfully demonstrated the application of vegetation indices derived from Sentinel-2 satellite data to predict rice yield, providing valuable insights into the efficacy of different indices at various stages of rice growth. Through the use of simple and multiple linear regression models, the research highlighted the significant relationship between satellite-derived vegetation indices and rice yield, underscoring their potential utility in precision agriculture.

The analysis revealed that vegetation indices are powerful tools for estimating rice yield. Among the indices analyzed, NDVI (Normalized Difference Vegetation Index) emerged as the most reliable predictor of rice yield when measured at 70 days after seeding. The high positive correlation (0.83) between NDVI_max and rice yield indicates that NDVI is effective in reflecting the health and productivity of rice plants during their critical growth stages. Similarly, SAVI (Soil-Adjusted Vegetation Index) also demonstrated strong predictive capabilities with a positive correlation of 0.80, suggesting that it is especially useful in areas where soil brightness may interfere with vegetation measurements.

The study also identified key periods where different indices are most effective. For instance, indices such as BSI (Bare Soil Index) and NDWI_MF (Normalized Difference Water Index - Modified), which were evaluated 40 days after tillering, showed strong negative correlations with rice yield. BSI's negative correlation (−0.72) suggests that higher

values, indicative of more exposed or poorer soil conditions, are associated with lower yields. NDWI_MF, with a correlation of −0.63, reflects water stress conditions that negatively impact rice productivity.

Simple linear regression models revealed that NDVI_max provided the most accurate predictions of rice yield, with an R^2 of 0.68. GNDVI_max also performed well with an R^2 of 0.64. In contrast, other indices like SAVI_max and EVI_max had lower predictive power, indicating that while they contribute useful information, they are less effective as standalone predictors.

The integration of multiple indices through a multiple linear regression model significantly enhanced prediction accuracy. The combined model using NDVI_max and BSI_min achieved the highest R^2 of 0.72, the lowest RMSE of 74.92, and a MAPE of 8.47%, demonstrating improved performance over single-index models. This indicates that combining different vegetation indices can provide a more comprehensive assessment of rice yield by capturing various aspects of plant health and environmental conditions.

The findings from this study emphasize the importance of utilizing satellite-derived vegetation indices for agricultural monitoring and yield estimation. By incorporating indices that reflect different growth conditions and stress factors, such as NDVI and BSI, farmers and agricultural planners can gain a more nuanced understanding of crop performance and make better-informed decisions.

The research also highlights the potential for satellite-based monitoring systems to revolutionize precision agriculture. The ability to predict rice yield with high accuracy using remote sensing data can lead to more effective resource management, improved yield forecasting, and enhanced agricultural productivity. This approach not only supports sustainable farming practices but also contributes to food security by enabling more precise and timely interventions.

Future studies could expand on this research by exploring additional vegetation indices, incorporating data from other satellite missions, and applying machine learning techniques to further refine yield prediction models. Additionally, incorporating ground-truth data from various environmental conditions and cropping systems could enhance the generalizability of the models.

In summary, the integration of satellite-derived vegetation indices into yield prediction models represents a significant advancement in agricultural monitoring. This study's results demonstrate the effectiveness of remote sensing technology in providing accurate and actionable insights for rice cultivation, paving the way for its broader application in

precision agriculture.

References

Choudhary, K., Shi, W., & Dong, Y. (2021). Rice growth vegetation index 2 for improving estimation of rice plant phenology in coastal ecosystems. Компьютерная оптика, 45(3), 438-448. https://doi.org/10.18287/2412-6179-CO-827

Department of Foreign Trade. (n.d.). International trade report. Retrieved September 12, 2024, from https://www.ditp.go.th/contents_attach/780244/780244.pdf

European Space Agency. (2012). Sentinel-2A and Sentinel-2B satellites: Technical guide. Retrieved from https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook

Food and Agriculture Organization of the United Nations. (2000). In fao.org dictionary. Retrieved July 25, 2024, from https://www.fao.org/4/x6905e/x6905e00.htm#Contents

Franch, B., Bautista, A. S., Fita, D., Rubio, C., Tarrazó-Serrano, D., Sánchez, A., ... & Uris, A. (2021). Within-field rice yield estimation based on Sentinel-2 satellite data. Remote Sensing, 13(20), 4095. https://doi.org/10.3390/rs13204095

Geo-Informatics and Space Technology Development Agency (GISTDA). (2024, April). Geo-informatics approach for assessing the risk of crops loss and damage associated with drought at farm level phase 2. GISTDA.

Li, H., Di, L., Zhang, C., Lin, L., & Guo, L. (2022). Improvement of in-season crop mapping for Illinois cropland using multiple machine learning classifiers. In Proceedings of the 2022 10th International Conference on Agro-Geoinformatics (Agro-Geoinformatics) (pp. 1–6). Quebec City, QC, Canada, July 11–14, 2022. https://doi.org/10.1109/Agro-Geoinformatics55370.2022.9876103

Liu, S., Liu, X., Liu, M., Wu, L., Ding, C., & Huang, Z. (2017). Extraction of rice phenological differences under heavy metal stress using EVI time-series from HJ-1A/B data. Sensors (Switzerland), 17(1), 1–17. https://doi.org/10.3390/s17061243

Martínez-Eixarch, M., Soriano González, J., & Alcaraz, C. (2022). Monitoring rice crop and yield estimation with Sentinel-2 data. Field Crops Research, 281, 108556. https://doi.org/10.1016/j.fcr.2022.108507

Nazir, A., Ullah, S., Saqib, Z. A., Abbas, A., Ali, A., Iqbal, M. S., & Butt, M. U. (2021). Estimation and forecasting of rice yield using phenology-based algorithm and linear regression model on Sentinel-II satellite data. Agriculture, 11(10), 1026. https://doi.org/10.3390/agriculture11101026

Office of Agricultural Economics. (2018). Report on rice production and export in Thailand. Retrieved July 25, 2024, from https://www.oae.go.th

Paul, G. C., Saha, S., & Hembram, T. K. (2020). Application of phenology-based algorithm and linear regression model for estimating rice cultivated areas and yield using remote sensing data in Bansloi River Basin, Eastern India. Remote Sensing Applications: Society and Environment, 19, 100367. https://doi.org/10.1016/j.rsase.2020.100367

Soriano-González, J., Angelats, E., Martínez-Eixarch, M., & Alcaraz, C. (2022). Monitoring rice crop and yield estimation with Sentinel-2 data. Field Crops Research, 281, 108507. https://doi.org/10.1016/j.fcr.2022.108507

Suksiri, T. (2015). Evaluation of NDVI, NDWI and NDDI for drought monitoring (Master's thesis, Department of Water Resources Engineering, Kasetsart University).

Yu, E., Di, L., Meyer, D., Zhao, P., Lin, L., Zhang, C., & Cvejovic, S. (2022). ICroplandNet: An open distributed training dataset for irrigated cropland detection. In Proceedings of the 2022 10th International Conference on Agro-Geoinformatics (Agro-Geoinformatics) (p. 6). Quebec City, QC, Canada, July 11–14, 2022. https://doi.org/10.1109/Agro-Geoinformatics55370.2022.9876103

Zhang, C., Di, L., Lin, L., Li, H., Guo, L., Yang, Z., Yu, E. G., Di, Y., & Yang, A. (2022). Towards automation of in-season crop type mapping using spatiotemporal crop information and remote sensing data. Agricultural Systems, 201, 103462. https://doi.org/10.1016/j.agsy.2022.103462

Zhang, C., Yang, Z., Zhao, H., Sun, Z., Di, L., Bindlish, R., Liu, P.-W., Colliander, A., Mueller, R., Crow, W., & others. (2022). Crop-CASMA: A web geoprocessing and map service-based architecture and implementation for serving soil moisture and crop vegetation condition data over U.S. cropland. International Journal of Applied Earth Observation and Geoinformation, 112, 102902. https://doi.org/10.1016/j.jag.2022.102902

Zheng, H., Cheng, T., Yao, X., Deng, X., Tian, Y., Cao, W., & Zhu, Y. (2016). Detection of rice phenology through time series analysis of ground-based spectral index data. Field Crops Research, 198, 131–139. https://doi.org/10.1016/j.fcr.2016.08.027