Analysis of Paddy Productivity Using NDVI and K-means Clustering in Cibarusah Jaya, Bekasi Regency

Liyantono*, Yusup Almadani, Yusuf Adillah, Muhammad Maulana Yusuf, Muhammad Nyak Reza Mahbub, and Alvin Fatikhunnada

Department of Mechanical and Biosystem Engineering, Faculty of Agricultural Engineering and Technology (FATETA), Bogor Agriculture University, Bogor, Indonesia.

*E-mail: liyantono@apps.ipb.ac.id

Abstract. Information about rice productivity is one of the references for government to maintain food availability. With remote sensing technology, rice productivity can be known faster. This research was conducted using UAV (Unmanned Aerial Vehicle) and Sentinel-2 Satellite. Sentinel-2 NDVI which has a low resolution with high resolution UAV images, both variables have similarity values and regression reaches 0.8. NDVI are grouped into 8 classes using k-means clustering based on the similarity of the waveforms of each data retrieval point. Based on characteristic of k-means classes, field which has earlier planting times and the location closer to the water source, allowing a higher paddy productivity. Further analysis was also carried out to get the best period to estimate paddy productivity using Sentinel-2 imagery. Sentinel-2 was chosen because it has a distance between data as far as 5 days, allowing it to be more accurate. The best time is obtained at 63 DAP (Days After Planting), which is when NDVI reaches its maximum state. The estimation model of rice productivity based on UAV has a high coefficient of determination compared to Sentinel-2 so that the relationship between maximum NDVI UAV and rice productivity is better than Sentinel-2.

1. Introduction
Paddy is widely use as a staple food for Indonesian people. The availability of safe rice is very important to achieve stable food security. Information about rice productivity has an important role for government to make national decisions. According to the Central Statistics Agency (2017), rice production in Indonesia is increasing every year. In 2011, Indonesia's rice production amounted to 65.7 million tons and continued to rise to 75.4 million tons in 2015. In estimating rice production in Indonesia, the Agriculture Ministry estimates rice production by considering the parameters of the area of planting / harvesting, the number of seeds spread by farmers and calculation of productivity by utilizing the institutional structure below, namely agricultural extension workers and agricultural staff from the Central Statistics Agency [5]. Most data retrieval is still done manually so it requires more time and effort. Therefore we need a more effective and efficient method of estimating rice productivity, one of which is using remote sensing technology. The development of methods for estimating rice productivity using this technology has been developed in Indonesia.

Remote sensing is science and art in obtaining information about an object area, or phenomenon through analysis of data obtained with a tool without a direct contact [3]. Remote sensing technology has been developed so rapidly and its application is increasingly widespread for various fields including agriculture [4]. Prediction of crop yields can be done by identifying the greenness (vegetation index) of
a plant using a ratio between infrared and near infrared bands [1]. The ratio formula is known as the vegetation index which can provide an overview of the greenness of vegetation based on plant biomass and indicates plant health. The formulation that is commonly used as a parameter in determining the vegetation index with the level of crop production is NDVI (Normalized Difference Vegetation Index) [8]. NDVI has a range of values ranging from -1 to 1 where values -1 to 0 indicate objects or non-vegetation objects while values from 0 to 1 indicate the presence of vegetation objects. The value index listed in NDVI determines how much influence vegetated land cover, or not vegetation, and the potential distribution characteristics of an area. The higher the NDVI value, the more plants are in the ready-to-harvest phase, on the contrary the lower the value, the more unproductive plants [6].

Instruments commonly used in remote sensing are UAV (Unmanned Aerial Vehicle) and Satellite such as Sentinel-2. Each instrument has advantages and disadvantages. The UAV has a spatial resolution that is more detailed than Sentinel-2, however Sentinel-2 images are easier to obtain because its available for free every 5 days. In this research, a comparison was made between UAV and Sentinel-2 to estimating rice productivity based on NDVI waveform characteristics and regression from NDVI and paddy productivities.

2. Methodology
2.1. Data Collection

The data collected in this research consist of 3 types of data including Sentinel-2 image data, UAV image data and field surveys data. Sentinel-2 image data downloaded from the sentinel official website via http: /scihub.copernicus.eu/. The image taken every 5 days from April 10, 2018 to August 13, 2018 according to one cycle of the second planting season in Cibarusah District, Bekasi Regency. Drone image taking is done every 16 days, from April 17, 2018 to August 8, 2018. Images are taken using Parrot Sequoia multispectral camera. Field surveys were conducted to obtain rice productivity data. Data is collected from farmers through direct interviews in the field.

2.2. Research Method

Sentinel-2 image data needs to be processed before further analysis. Data processing consists of several stages, which is summarized in the research methods (Figure 1)
Paddy vegetation index calculated using NDVI (Normalized Difference Vegetation Index) with these following equation:

\[
NDVI = \frac{\rho_{NIR} - \rho_{RED}}{\rho_{NIR} + \rho_{RED}}
\]  

(1)

After NDVI obtained, data processed using k-means clustering. The purpose of using k-means is to generalize NDVI waveform. Main principle of k-means is to compile a prototype or center of mass (centroid) from a dimensionless data set n. k-means requires input of the number of classes as much as k and divides the amount of data n into class k based on the similarity with the class centroid [2]. Determinate of the centroid value in k-means using these equation [7]:

\[
C_i = \frac{1}{M} \sum_{j=1}^{M} x_j
\]  

(2)

Data that has the shortest distance with the centroid will be entered into the class in the centroid. Calculation to determine the distance between centroid points with each point of the object using following equation:

\[
d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}
\]  

(3)

3. Result and Discussion

3.1 Comparison Between NDVI Sentinel-2 and UAV
NDVI UAV is the image that closest to the original conditions in the field. UAV use parrot sequoia camera to capture multispectral images from the field, pixel resolution is capable of reaching 5 cm x 5 cm for RGB and 12 cm x 12 cm for NDVI. Image took using UAV is also carried out at an altitude of 110 m so that the resulting image has minimal atmospheric interference. Comparison of NDVI waves from Sentinel-2 and drone (UAV) as in Figure 2.

![Figure 2. Comparison between NDVI Sentinel-2 and UAV](image-url)
Figure 2 shows that the two images have different characteristics. In the vegetative phase, NDVI Sentinel-2 is lower than the NDVI UAV. Different case in generative phase, NDVI Sentinel-2 is higher than NDVI UAV. In this phase, NDVI UAV is abnormal. NDVI decrease significantly and in the last data number, average NDVI is below 0.1. This value generally occurs in the early phase of rice planting when there is still a pool of water or mud on the land. Validation carried out using RGB drones shows that there were no water body during data collection. This is most likely due to corruptness of the multispectral instrument in parrot sequoia.

| No | Equation of Regression | $R^2$ | No | Equation of Regression | $R^2$ |
|----|------------------------|-------|----|------------------------|-------|
| 1  | $y = 0.448x + 0.0631$  | 0.5412| 5  | $y = 0.6118x + 0.2591$ | 0.4737 |
| 2  | $y = 0.6342x + 0.0681$ | 0.7312| 6  | $y = 0.3767x + 0.3773$ | 0.8025 |
| 3  | $y = 0.6959x + 0.1256$ | 0.7909| 7  | $y = 0.3953x + 0.2928$ | 0.7376 |
| 4  | $y = 0.8015x + 0.0601$ | 0.6127| 8  | $y = 0.5086x + 0.2672$ | 0.7967 |

From Table 1, the slope value for each take is quite diverse. The highest slope is 0.8015 which is at the 4th take on June 2, 2018. The slope value close to 1 indicates that the value of $x$ approaches the $y$ value. $R^2$ of the equation model has the smallest value on June 22, 2018, which is 0.4737. While the biggest value is on July 7, 2018, which is 0.8025. This shows that the relationship between NDVI drone and NDVI Sentinel-2 has a large percentage of up to 80.25%

### 3.2 Paddy Productivity Based on Characteristics of K-means Classes

NDVI data in one season would be grouping into 8 classes using k-mean clustering. The results of the classification as presented in the figure 3.

![Figure 3. Results of NDVI wave grouping based on k-means clustering](image)

Each 8 k-means class has varied characteristics. Class 1 has high NDVI value and does not have a significant change, this class identified as a weed and plots that are disturbed by other vegetation. Class 3 and 4 has the characteristics of NDVI which immediately increase in the first time, this indicates a change from the plot that has been done by processing the land to the plots that have been planted with rice. Class 3 and 4 peaks also reach the peak of NDVI faster because rice growth is better. For other classes, NDVI will go down first after it has increased. There is a change from the plot that has not been carried out till the soil is cultivated, becoming a plot that has been treated by the soil, and then becomes a plot that has been planted with rice. Map those who have not undergone tillage have a higher NDVI because they contain weeds or former planting. NDVI will go down after soil lubrication, higher water content causes NDVI to decrease. After rice is planted, NDVI will increase because there are green plants that grow in the plot.
From the survey result that has been conducted, productivity data are grouped according to the k-means class. Average rice productivity for each class as presented in Table 2.

**Table 2.** Rice productivity for each class

| Classes | Average rice productivity (ton/ha) | Classes | Average rice productivity (ton/ha) |
|---------|----------------------------------|---------|----------------------------------|
| Class 2 | 0.841                            | Class 5 | 1.576                            |
| Class 6 | 0.919                            | Class 3 | 2.122                            |
| Class 8 | 0.938                            | Class 4 | 3.280                            |
| Class 7 | 1.324                            | Class 2 | 0.841                            |

**Figure 4.** Spatial distribution of k-means classes

Figure 4 shows the class distribution in the rice field. Each class is given a different color indicator, the color indicator is adjusted to the level of productivity of k-means class. Class 4 which has the highest productivity is given a green color, class 1 which has the lowest productivity is given a red color, and for other classes the color indicator adjusts. The majority of Class 3 and 4 are in group A which is adjacent to water sources, so that the water needs for rice can be fulfilled. This explains why classes 3 and 4 experience an increase in NDVI faster than other classes which far from water sources.

### 3.3 Paddy Productivity Estimation

The best period to estimate paddy productivity is sought based on the highest regression relationship NDVI with rice productivity. NDVI used from Sentinel-2 images, the data has a range of 5 days so that it has smaller error if interpolated to NDVI each day or DAP (Days After Planting). The regression value between NDVI for each DAP and rice productivity is presented in Figure 5.
The best period in estimating rice productivity is 63 DAP with a regression value is 0.27. In this period, NDVI Sentinel-2 reaches its maximum state. This is strengthened by the pattern of average NDVI UAV value of paddy with high productivity and paddy with low productivity which is shown in Figure 6. Based on Figure 6, NDVI value between paddy with high productivity and paddy with low productivity are significantly different when it reach maximum state.

![Figure 5. Regression paddy productivity and NDVI each DAP](image)

**Figure 5.** Regression paddy productivity and NDVI each DAP

**Figure 6.** Pattern comparison of ndvi value on paddy with high productivity and paddy with low productivity

![Figure 7. Relation between NDVI and productivity, A) low productivity, B) high productivity](image)

**Figure 7.** Relation between NDVI and productivity, A) low productivity, B) high productivity
Then if the maximum values of NDVI on every K-means classes connected with the productivity, it will make a linear line which is shown at Figure 7. Rice with high productivity has coefficient of determination $R^2 = 0.5$ and $R = 0.71$. Meanwhile, rice with low productivity has coefficient of determination and correlation $R^2 = 0.53$, $R = 0.73$. Based on Figure 6 the productivity for rice with low productivity can be predicted using equation $y = 4.1233x - 1.9758$ and rice with high productivity can be predicted using equation $y = 6.2123x - 1.6139$.

4. Conclusion
Sentinel-2 images have NDVI values that approach high-resolution NDVI images such as UAV images. The regression formed between the two variables reaches 0.8 and the linear equation slope value is 0.8. In the analysis of rice productivity, NDVI waveforms are grouped using K-means clustering. Class 3 and 4 have the highest productivity. Based on the waveform and the spatial distribution of K-means class, it is known factors that affect of high paddy productivity are the earlier planting times and the adequacy of water requirements for rice growth. The best period to estimate paddy productivity is 63 DAP (Days After Planting) that NDVI reach its maximum state. The regression formed between productivity and maximum value of NDVI Sentinel-2 image is 0.27, lower than NDVI UAV which has regression 0.53 with estimation model is $y = 6.2123x - 1.6139$.

References
[1] Affan MJ. 2002. Assessment of Forest Fire Hazard Levels Based on the Vegetation Index and KBDI (Bogor : Institut Pertanian Bogor)
[2] Giyanto H. 2008. Application of the k-means, k-medoid, gath geva clustering algorithm (Yogyakarta: Thesis in Universitas Gadjah Mada)
[3] Lillesand TM, Kiefer RW, Chipman J 2008 Remote Sensing and Image Interpretation, 6th ed (New York: J Willey)
[4] Martono DN. 2008. Application of Remote Sensing Technology and Validation Test for Detection of Spread of Rice Fields and Land Use / Cover (Jakarta : PUSDATA LAPAN)
[5] Napitupulu TEM. 1998. Estimation System for Production Forecast Results in the Context of Safeguarding National Food Production. Workshop on Rice Monitoring and Prediction Systems in Indonesia (Jakarta : Sari Project – BPPT)
[6] Pradipta D. 2012. Data Analysis of Time Series NDVI-Spot Vegetation for Rice Plants (case study: Karawang district) (Bogor : Institut Pertanian Bogor)
[7] Wakhidah N. 2010. Clustering using k-means algorithm Jurnal Transformatika 8 (1) 33-39
[8] Wahyunto W, Heryanto B. 2006. Estimating the Productivity of Paddy Crops through Analysis of Satellite Images. Warta Informatika Pertanian 15 (2) 853-859