RICE YIELD ESTIMATION USING FREE SATELLITE AND FIELD DATA

Tharshini Shanmugam[1] and TL Dammalage

Department of Remote Sensing and GIS, Faculty of Geomatics,
Sabaragamuwa University of Sri Lanka, P.O. Box 02, Belihuloya, 70140, Sri Lanka
[1]*tharsha0012@gmail.com*

Abstract- An effective pre-harvest rice yield estimation method is truly significant for the assessment of seasonal rice production in terms of strategic planning purposes. In Sri Lanka, a conventional method named crop-cut survey is used to estimate seasonal rice production, yet it fails to forecast rice yield before the harvest as it is conducted during the harvest. Therefore, this study is focused on identifying cultivated paddy lands and forecasting rice yield using free satellite data. Landsat 8 OLI/ TIRS images (30m spatial resolution) from Earth explorer and 8-day composite images (250m spatial resolution) from Moderate Resolution Imaging Spectro-radiometer (MODIS) sensor on board NASA EOS Terra/Aqua satellite were used from 2014 to 2017. Paddy cultivated lands were identified by land cover classification by using field training samples and Landsat 8 OLI/ TIRS data. In addition, the temporal change of Normalized Differenced Vegetation Index (NDVI) for paddy and forest was also analyzed to validate the classification. The observed minimum accuracy of the land cover classification out of the tested four (4) seasons was 99.4%, and the minimum kappa coefficient was 0.9916. The correlation coefficient between reference net harvested paddy area and paddy cultivated area identified by Landsat 8 is 0.93. Linear and exponential yield forecasting models proposed by Sirisena, et al. (2014) for Kurunagala district were validated and tested based on NDVI and EVI2 vegetation indices obtained through MODIS (MOD09Q1v006) surface reflectance image of Polonnaruwa District. The comparison of the estimated yield with national statistical records, both NDVI and EVI2 based models, provide more reliable estimations about 80 days after the transplanting of each season, but, EVI2 based model (derived at 80 days) gives more reliable estimations than NDVI based model with 86.37% of average accuracy. Therefore, seasonal rice yield can be successfully forecasted one month prior to the harvest time using EVI2 based model in the Polonnaruwa district.

Keywords- EVI2, Landsat 8 OLI/TIRS, MODIS, NDVI, Rice yield

I. INTRODUCTION

Paddy occupies approximately 37 % (0.77 million ha) of the cultivated land area of Sri Lanka. It is cultivated during two major seasons; Yala and Maha. The periods of the commencement of each season are uncertain as the commencement of Yala season varies from, end of March to mid of May, and Maha season from end of September to mid of December. The majority cultivates in Maha season. Sri Lanka produced nearly 2.9 million tons of rice in the last season of cultivation (2015/2016). There was an overproduction in that season, and therefore farmers in many parts of the country were angered and disappointed over the failure to set up an adequate mechanism to market their paddy. In the 2016/2017 Maha season the production plummeted as low as 1.7 million tons of rice, and there was food shortage.

Also rice yield estimation in Sri Lanka is based on conventional techniques of data collection for crop and yield estimation based on ground-based field visits and reports. It is time-consuming, subjective, and prone to significant discrepancies as a result of insufficient ground observations that cause poor rice production assessment. The outcomes are usually made available to the government and public after several months of the harvesting of the rice, and thus not useful for food security purposes, and It is costly, depending on the survey areas. Mostly, the data become available too late for appropriate actions to be taken to avert food shortage. Therefore, we need to estimate rice before the harvesting of the rice.

According to that, satellite remote sensing was widely applied and it was recognized as a powerful and an effective tool for identifying agriculture crops. The possibility to estimate seasonal crop yield before a specific time period of harvesting is vital to take precautions regarding seasonal crop production. Specially, it is of paramount importance in strategic planning and decision making regarding the food security and facilitation of safe harvest storages. On the other hand, proper import (in shortfall case) or export (in surplus case) policies can be made based on such reliable yield estimations (Noureldin et al., 2013). Since there are such considerable advantages related to a pre-harvest yield estimation method, it is very important to focus on modern science and technology in developing reliable yield forecasting models. Satellite Remote Sensing can be successfully applied to this research area as it is powerful and effective in estimating and forecasting crop yields (Ferencz et al. 2004).

Use of satellite data for crop classification and crop yield estimation has a long history. Primarily, for paddy rice classification, existing studies mainly use two types of sensors: Visible and near-infrared (NIR) sensors: such as Landsat, MODIS, and Sentinel-2, Radar sensors: such as ALOS, Radar Sat, and Sentinel-.1. There is very limited study to use satellite data for paddy rice yield estimation at the field level. The major bottlenecks include: Lack of field-level crop yield data (thus current project is very critical) (Lobell et al., 2014). Lack of continuous satellite data. Due to the lack of data, the methodology improvements are slow. Cloudy issues for the visible/NIR sensors (Guan et al., 2012; Gao et al., 2015). Though methodology progress for paddy rice is relatively slow (Nelson et al., 2014), visible and near-infrared (NIR) sensors are most available data. However, it is hard to use it. For classification: a few usable images during the growing season sometime suffice the needs. However, for yield-estimation, we need both: high spatial-resolution for field-level information. High temporal-resolution for growth condition or phenology and we simple donot have the free data at both high resolutions.

The study aims to improve the MODIS images-based rice yield estimation model with the Landsat 8 OLI/TIRS images and determine the best age of paddy plants for yield forecasting in Polonnaruwa, to identify the paddy cultivated area based on Landsat 8 OLI/TIRS images, and to validate the MODIS images based yield estimation model for the study area and Prediction of the rice yield.

II. STUDY AREA AND MATERIALS

A. Study Area

The study area was Polonnaruwa which is located in the North Central Province, Sri Lanka, at 7° 56 latitude 81°0 longitude (Figure 1). Polonnaruwa is the best area for rice production in Sri Lanka and it has approximately 85,505 acres (34,629.525 ha) of rice fields (Irrigation Department of Sri Lanka 2009). Due to the wideness of the area, this farming system can be easily monitored by remote sensing.

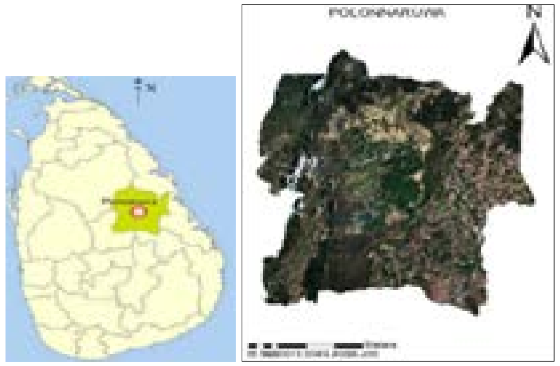


Figure 1. Spatial extent and location of the study area in Sri Lanka with a subset of the Landsat 8 true color bands combination, acquired in 2017.

B. Satellite and Rice Yield Data

Free satellite images Landsat 8 OLI/TIRS level 1 collection 1 were downloaded from the USGS Earth Explore Web site from year 2014 to 2017(6 season) images. This product consists of 30m spatial resolution and there is no proper temporal resolution. MODIS Surface Reflectance 8-Day L3 Global 250m (MOD09Q1) Product was used in this study from 2014 to 2017.MOD09Q1 provides Bands 1 and 2 at 250-meter resolution in an 8-day gridded level-3 product in the Sinusoidal projection. Each MOD09Q1 pixel contains the best possible L2G observation during an 8-day period as selected on the basis of high observation coverage, low view angle, the absence of clouds or cloud shadow, and aerosol loading. Science Data Sets provided for this product include reflectance values for Bands 1 and 2, and a quality assurance rating. Rice yield data and other required paddy statistics from 2014 to 2017 were obtained from Census and Statistics Department, Sri Lanka.

In order to achieve the major objective of the study, a method was suggested to identify paddy cultivated lands in each growing season using Landsat 8 OLI/TIRS data and MODIS satellite data. In Landsat 8 OLI/TIRS images Radiometric and FLAASH atmospheric corrections were conducted. Clouds were removed. There were three science data sets in MODIS satellite images which were 250m surface reflectance band 1 (620-670 nm), 250m surface reflectance band 2 (841-876 nm) and 250m reflectance band quality data sets.

III. METHODOLOGY

A. Vegetation Indices

NDVI: The Normalized Difference Vegetation Index (NDVI) gives a measure of the vegetative cover on the land surface over wide areas. Dense vegetation shows up very strongly in the imagery, and areas with little or without vegetation were also clearly identified. NDVI calculated by using NIR band-5(0.845-0.885μm) and Red band-4 (0.630-0.680μm) In Landsat 8 OLI/TIRS images. NDVI temporal change of the paddy and forest was derived. Temporal pattern of paddy cultivated area was identified through that graph.

MODIS images consist of two bands to calculate NDVI. Both bands were used as green vegetation displays strong absorption in the red part of the spectrum (reflectance of around 3-5%) and weak absorption in the NIR part (reflectance around 40 - 60%) (Gitelson, 2004). In MOD09Q1 images calculated NDVI by equation 1 using 250m surface reflectance band 1 (620-670 nm), 250m surface reflectance band 2 (841-876 nm).

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{1}$$

EVI2: The Enhanced Vegetation Index (EVI) (Huete et al., 1999) was developed as a standard satellite vegetation product for the Terra and Aqua Moderate Resolution Imaging Spectro radiometers (MODIS). EVI provides improved sensitivity in high biomass regions while minimizing soil and atmospheric influences. The major limitation of EVI is, it utilizes blue band in addition to the red and near-infrared bands. Jiang et al. (2008) developed and evaluated a 2-band EVI (EVI2), without a blue band, which is synonymous with the 3-band EVI, particularly when atmospheric effects are insignificant and data quality is good. EVI2 index of the MOD09Q1 images calculated.

$$EVI2 = 2.5\frac{NIR - RED}{NIR + 2.4RED + 1} \tag{2}$$

B. Data Collection

The field route was selected through Google earth application before commencing the field work. The centre of the study area, which consists of more paddy fields was selected for further purposes. The field work was planned for one day. GPS co-ordinations of the paddy fields and the forests were collected through CT Droid Sri Lanka application in mobile phones. 50 paddy samples were collected. Ground truth data was used for the classification. The identification of paddy area and the collection of 20 samples of remaining places were done by using Google Earth application. The areas in terms of percentage and hectares were also computed. Accuracy assessment was captured.

The paddy cultivated area could be identified through the basic statistics of the maximum likelihood classification, and it was compared with the information of census and statistics department (reference data). As mentioned above, there are two main paddy cultivation seasons in the study area, yet, the commencement of the relevant seasons is uncertain. Therefore, several composites (covering the whole season) were taken for each season in terms of time series analysis.

Then NDVI values for all images were calculated using equation 1. Before calculating correlation coefficients, NDVI temporal profiles were smoothed using "moving average" method. Correlation coefficients between NDVI changing pattern of normal paddy cultivation and smoothed profiles were calculated. In this respect, 10 correlation coefficients were calculated for one smoothed

NDVI temporal profile and 10 correlation coefficients were also calculated. Then an appropriate threshold value for correlation coefficient was selected to extract cultivated paddy pixels by equalizing the total paddy area identified through algorithm and census data. The pixels which included correlation coefficient values higher than threshold values, were classified as paddy cultivated lands. A few pixels (identified as paddy) were randomly selected for verification of the classification result. In this process, randomly selected paddy pixels were compared with Google earth and field data to determine whether those pixels are truly paddy or not. This method was applied for the other years of the seasons for identify cultivated paddy lands in the study area. Paddy cultivated area was identified through Landsat 8 and MODIS images. Both extents were compared with reference to data and also check whether the areas are truly cultivated or not with the training samples and Google earth application. Finally get the cultivated paddy extent from analysing those extends.

C. Validation of Model Equation and Yield Prediction.

Dammalage et al. 2017 has proposed several yield estimation models for Kurunagala district using MODIS data. This study investigates the use of the same models to estimate the rice yield in Polonnaruwa district. For validation correlation coefficient from MODIS images was used as the one of the input quantity, and also paddy area from both Landsat 8 images and MODIS images were used to find out which is most suitable for this analysis. According to that, the cultivated paddy area identified through Landsat 8 OLI/TIRS images was widely used for research purposes. After validating and checking the accuracy of each model, the most accurate one will be used for yield prediction in Polonnaruwa district.

IV. RESULTS AND DISCUSSION

A. Identification of Cultivated Paddy Fields

Figure 2 shows the temporal pattern of the cultivated paddy and forests identified through 2017 yala season images.

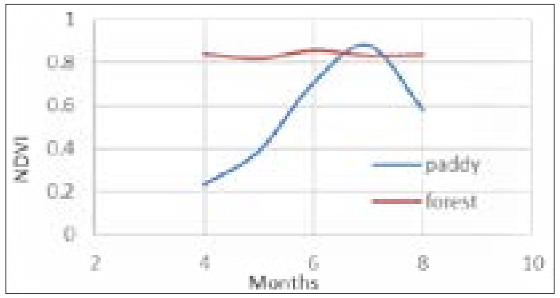
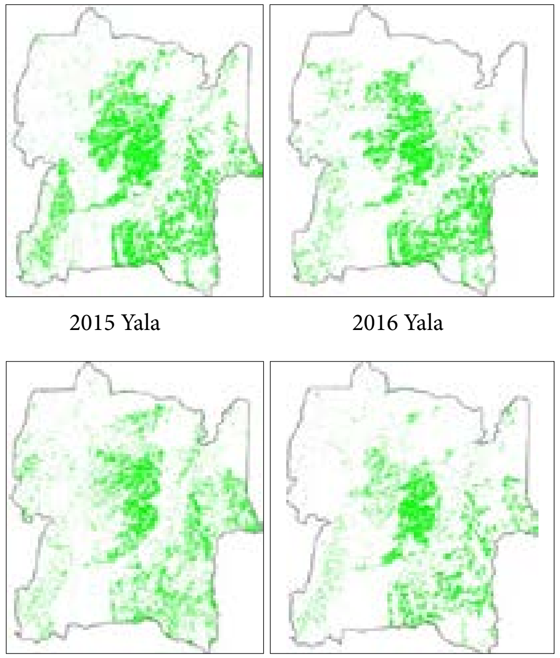


Figure 2.The cultivated paddy lands identification process by temporal change of paddy and forest derived from Landsat 8 images

Generally yala season is from May to August, (not certain). Therefore the images of middle of April - end of August seasons were used. There were paddy fields identified in May. NDVI values of the paddy was increased near NDVI values of forest. In July high NDVI value was occured (after 80 days). However, the utilisation of Landsat 8 images in identifying cultivated paddy fields was unsuccessful due to the low temporal resolution of those images. Then for the land cover classification both (training samples and google earth samples) ground truths were used. Figure 3 shows the classified image over study area.



2015 Yala

2016 Yala

2016/17 Maha

2017 Yala

Figure 3. Landsat 8 classified images. (Green colour indicates paddy cultivatedareaFigure 4. The relationship between rice age and Determinant of Coefficient (R2) values of linear yield forecasting models

Only by using the training samples the identified cultivated paddy extent could not be checked. Still have the problem about the identified paddy extend by Landsat 8 image. But according to the time allocation for the analysis it couldn't find out correctly. After that, compare those images with ground truths (training samples). This method was applied for all the other seasons to identify cultivated paddy lands in the study area. Before calculating correlation coefficients, NDVI temporal profiles were smoothed using "moving average" method. Correlation coefficients between NDVI changing pattern of normal paddy cultivation and smoothed profiles were calculated. Then an appropriate threshold value for correlation coefficient was selected to extract cultivated paddy pixels by equalizing the total paddy area identified by this algorithm and census data. The pixels which included higher correlation coefficient values than threshold values, were classified as paddy cultivated lands. Correlation coefficient (r) has taken a higher value at some points when the specific NDVI time series corresponded to a cultivated paddy land. Required seasonal parameters were obtained based on the number of times that the matching pattern was shifted to right occurring the highest correlation coefficient value. In the above situation, maximum value of the correlation coefficient is 0.90 and it is highly correlated with the matching pattern. Therefore that pixel is classified as paddy cultivated pixel. Using this method, all pixels were classified into two classes as paddy cultivated and uncultivated lands.

Table 1. demonstrates, the interdependency between the amount of area identified as paddy and the threshold value for correlation coefficient.

Season	Threshold for correlation coefficient	Total amount of area identified as paddy (hectares)	Net harvested area (hectares)
2014/15M	0.8116	65,756	65,891
2015 Y	0.8454	54,513	54,668
2015/16M	0.8990	57,650	57,608
2016 Y	0.8920	49,731	49,754
2016/17M	0.9057	42,306	42,268
2017 Y	0.8615	35,625	35, 585

According to the table 1, there is no unique or common threshold value which can be used to distinguish cultivated paddy lands from other Land uses. However, by changing the shape of the matching NDVI pattern and smoothing and filtering methods of time series data, a unique threshold value can be discovered and it provides food for thought for further studies.

B. Relationship between Vegetation Indices and Rice Yield

A number of rice yield forecasting models had been built up based on NDVI and EVI2 vegetation indices at different age of paddy starting from a rice age of 32 days.

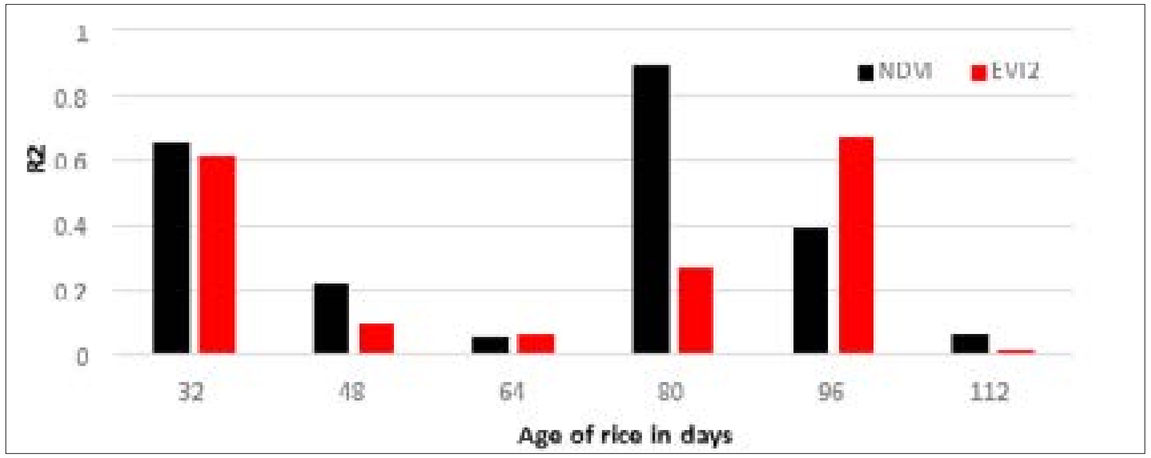


Figure 4. The relationship between rice age and Determinant of Coefficient (R2) values of linear yield forecasting models

Figure 4 shows how determinant of coefficient (R2) values of derived linear models change with the age of the paddy plant. The trend of R2 slightly decreased from the age of 32 days to 64 days, then it has reached its maximum value at the age of 80 days and again it decreased until the harvest time. Based on this statistical analysis, the paddy plant at about 80 days after transplanting has given the best relationship between vegetation indices (NDVI and EVI2) and rice yield. However, in most stages, yield forecasting models which are derived based on EVI2 index have higher R2 values than NDVI based models.

To determine the accuracy and reliability of estimations provided by yield forecasting models, an accuracy assessment has been done for Polannaruwa district over six seasons by using rice yield data provided by the Department of Census and Statistics in Sri Lanka. The accuracy of the estimations is given in table 2 and those results have shown the reliability of yield estimations provided by each model which are based on remote sensing data. Table 2 illustrates the validation of each model over the Polonnaruwa district. There are 5 linear and 5 exponential models. Each and every model consists different accuracy level. According to this validation the correlation between reference data and identified data by using these models were given, NDVI after 80 days gives 0.89 and EVI2 gives 0.27 . Thus, NDVI after 80 days gives an overall accuracy of 77.54% and EVI2 after 80 days model gives an overall accuracy of 83.68%. Therefore, the EVI2 after 80 days model is suitable for rice yield estimation over the study area. Then correlation between the reference yield data and estimated yield data was calculated by using the graphs. X axis determined reference yield data from statistics department and y axis determined estimated rice yield by models. There was good correlation in NDVI after 80 days model, but it has less accurate than EVI2 after 80 days model. Therefore, EVI2 after 80 days based model can used to estimate the rice yield.

B. Languageimprovement needed

Table 3 shows that the comparison of cultivated paddy area. According to that,the area identified by Landsat 8 was in 2 decimal places . And according to the accuracy analysis and other analysis the identified cultivated area by Landsat 8 was more accurate than the other data. But still we have the issue on this data because for the above mentioned reasons.

Table 2. Different type of rice yield forecasting models and a comparison of their forecasting results with national statistical data.

Forecaster	Model	Accuracy (model/crop cutting) ×100%					
		14/15M	15 Y	15/16M	16 Y	16/17M	17Y
NDVI after 80 days	$y = 2542x + 2087$	73.2	77.7	78.9	79.3	78.1	80.3
	$y = 2391e^{0.6862x}$	73.2	77.8	79.0	79.4	78.2	80.3
EVI2 after 80 days	$y = 8402x + 174.3$	72.9	88.5	81.2	91.8	78.7	96.5
	$y = 1410e^{2.296x}$	72.9	90.6	81.6	94.4	78.9	99.8

Table 3. Paddy cultivated area comparison

Season	Identified paddy area by Landsat 8 (Hectares)	Identified paddy area by MODIS (Hectares)	Net extent (Reference data) (Hectares)
14/15 M	No image	65,756	65,891
15 Y	67,223.61	54,513	54,668
15/16 M	No image	57,650	57,608
16 Y	54,847.53	49,731	49,754
16/17 M	47,185.47	42,268	42,268
17 Y	36,739.80	35,625	35,585

Table 4: Accuracy analysis of total rice yield production by models

Season	Reference total yield (MT)	NDVI after 80 days model (MT)	EVI2 after 80 days model (MT)	Accuracy (model/crop cutting) ×100%	
				NDVI After 80 days	EVI2 after 80 days
15 Y	280,476	268,470.05	312,556.34	95.70	Over production
16 Y	251,131	219,826.87	261,353.47	87.53	Over production
16/17 M	214,722	187,489.52	189,029.05	87.32	88.03
17 Y	173,595	143,935.52	178,805.26	82.90	Over production

Table 4 shows the accuracy analysis of estimated total rice yield by different models. NDVI after 80 days model and EVI2 after 80 days model is considered to estimate total production. Therefore, the total paddy cultivated extent taken from Landsat 8 images and estimate total yield. So according to that, table using NDVI model will give an accuracy closer to that of crop cutting survey method. But using EVI2 model gives a higher accuracy than the crop cutting survey method. Accordingly, those two models can be used for estimation if the reference yield data is assumed to be correct, and if need to the accuracy as crop cutting survey can replace this NDVI after 80 days model because one month before harvesting can estimate rice yield near to crop cutting survey accuracy. Therefore, no need to wait until crop cutting survey. Another thing is that the EVI2 based model consists total estimation more than crop cutting survey output. Therefore, it also can use, but have to analyse about whether the total yield production is correct or not. Couldn't say that the reference data was correct, anyhow have to analyse and then can use that model also for further analysis. According to time allocation it couldn't find out.

V. CONCLUSION AND RECOMMENDATION

The algorithm for identifying cultivated paddy fields was developed based on high spatial resolution images of paddy cultivation. The developed algorithm retains the capacity to find the commencing period of paddy cultivation though it alters from one location to another. The impossibility of using Landsat 8 images successfully was a major limitation of the study. Therefore, as this limitation highly influenced the utilization of Landsat 8 images in the identification of cultivated paddy fields, the relevant areas should be visited and observed to collect paddy statistics and field verifications. Nevertheless, the accuracy of this algorithm can be improved by using high temporal resolution images.

Finally, the researchers arrive at a conclusion that the model based on EVI2 values is the best model for rice yield forecasting using satellite imagery (This study was only based on NDVI and EVI2 indices). Therefore, rice yield can be accurately forecasted approximately for one month (4-month paddy) before harvesting.

REFERENCES

Cheng, Q., & Wu, X. (2011). Mapping paddy rice yield in Zhejiang Province using MODIS spectral index. Turkish Journal of Agriculture and Forestry, 35(6), 579-589.

Dammalage T. L., Sirisena P.M.T.S. and Junichi Susaki (2017). MODIS Satellite Data Based Rice Yield Forecasting Model for Sri Lanka: A Pilot Study in Kurunegala District. Asian Journal of Geoinformatics, Volume 17, Issue 3, 2017, pp. 24-33. ISSN: 1513-6728

Huang, J., Wang, X., Li, X., Tian, H., & Pan, Z. (2013). Remotely sensed rice yield prediction using multi-temporal NDVI data derived from NOAA's-AVHRR. PloS one, 8(8), e70816.

Jiang, Z., Huete, A. R., Didan, K., & Miura, T. (2008). Development of a two-band enhanced vegetation index without a blue band. Remote sensing of Environment,112 (10), 3833-3845.