Available online at www.sciencedirect.com

ScienceDirect

Advances in Space Research 70 (2022) 2443–2457

ADVANCES IN SPACE RESEARCH (a COSPAR publication)

www.elsevier.com/locate/asr

Random Forest for rice yield mapping and prediction using Sentinel-2 data with Google Earth Engine

K. Choudhary [a,b,*], W. Shi [a,*], Y. Dong [a,c], R. Paringer [b]

[a] Department of Land Surveying and Geo-informatics, Smart Cities Research Institute, The Hong Kong Polytechnic University, Hong Kong
[b] Scientific Research Laboratory of Automated Systems of Scientific Research (SRL-35), Samara National Research University, Samara, Russia
[c] Institute of Geophysics & Geomatics, China University of Geoscience, Wuhan, PR China

Received 8 March 2022; received in revised form 13 June 2022; accepted 30 June 2022
Available online 8 July 2022

Abstract

Accurate information on crop yield prediction is essential for farmers, governments, scientists, and agricultural agencies to make well-informed decisions. Majority of yield prediction methods have been based on data assimilation, which incorporates consecutive observation of canopy development from remote sensing data into model simulations of crop growth processes. But this study used high resolution Sentinel-2 data with combination of different types of secondary data in Random Forest (RF) regression model on different phases of the crop growing season for higher accurate rice yield prediction. For that First, computed crop/non-crop and rice/non-rice crops through RF classifiers were applied on seasonal median composites of Sentinel-2 data for each pixel in the region. Thousands of crop/non-crop labels were collected using an in-house google earth engine (GEE) labeler, and several crop type labels were obtained from various sources during the crop growing seasons. Results demonstrate that sentinel-2 imagery is useful to detect crop/non-crop classes from cropland with more than 85% accuracy, thus it can be used for crop prediction. Furthermore, the Sentinel-2 imagery with secondary data such as environmental, soil and topographic data perform higher accuracy for yield prediction. Its show 0.40 to 1.01 t/ha yield production range at a landscape level. Overall, this study illustrates the Sentinel-2 imagery, GEE platform, advanced classification and rice yield mapping algorithms are enhance the understanding of precision agricultural systems.

Keywords: Sentinel-2; Yield prediction; GEE; Environmental data

1. Introduction

Rice (Oryza sativa L.) is a staple agricultural crop and feeding over 50% of the global population (Song et al., 2018). China is the world's largest rice producer (approximately 206 million metric tons) accounting 28% of the world's rice production. Rice yield predictions can assist farmers in diagnosing crop conditions and implement essential steps to improve production as well as valuable for policymakers to develop government development plans for food security (Kailou et al., 2015). However, critical challenges such as climate change, rising temperatures, and severe droughts have posed considerable challenges to the accurate prediction of crop yield (Jeong et al., 2022). Previous studies revealed that yield differences amongst the same rice types under varying environmental conditions (Qiu et al., 2015). For example, Taoyuan, Yunnan Province, China, had a larger grain yield than other rice planting locations (Chen et al., 2011), indicating that prevailing climatic circumstances have a significant impact on rice growth and productivity (Mosleh et al., 2015). The different climatic variables consequence the differences in temperature and precipitation which eventually affected

* Corresponding authors.
E-mail addresses: komal.kumarri@connect.polyu.hk (K. Choudhary), lswzshi@polyu.edu.hk (W. Shi).

https://doi.org/10.1016/j.asr.2022.06.073
0273-1177/

rice yield formation (Gao et al., 2008). Therefore, an effective method for crop yield prediction is necessary.

Crop monitoring has been the major focus of earth observation activity since the first satellites were launched. Sentinel 2 satellite sensors with spatial resolutions of 10 m have opened new possibilities for monitoring crop fields (Choudhary et al., 2021). For example, In a Landsat image with a resolution of 30 m, a 1 ha field consists of only 9 to 10 pixels, making field delineation extremely challenging. In an illustration from the same period, that same field would have about 100 pixels in an image from Sentinel-2 with 10 m resolution. We rely on the Google Earth Engine platform's particular capabilities to analyze this huge amount of data (Gorelick et al., 2017; Gomes et al., 2020).

Though, several aspects of small field farming, despite the small sizes of fields, are major challenging in the context of remote sensing. First are the abundant clouds covers in many small fields which are active growing crops during the rainy season and secondly are the smallest fields are extremely diverse, with families generally growing multiple crops, which are commonly interspersed in their systems (Choudhary et al., 2019). Furthermore, to this within-farm heterogeneity, there are often disparities in the crops grown by farmers in a neighboring district or state (Zhao et al., 2020). Despite these impediments, recent work has demonstrated that Google Earth Engine (GEE) can be applied to yield prediction (Jin et al., 2019; Yang et al., 2021). Google Earth Engine contains petabyte scales of remotely sensed, geophysical datasets, and other ready-to-use products spanning over 40 years (Tu et al., 2020; Jeong et al., 2022). It also assists users to search, analyze, and visualize geospatial data without needing access to specialized coding expertise. Land use and land cover (LULC) classification, crop mapping, yield prediction (Zhao et al., 2020), forest mapping, and other studies based on GEE have been conducted at regional to global scales.

Machine learning algorithms are applied in different fields; it has been used in agriculture for a few years. Crop yield estimation is a challenging task in agriculture, and various methods have been created and tested. As agriculture production is influenced by a variety of elements like weather, fertilizer types, environment, etc., these issues require the use of many datasets. This indicates that estimating crop yield is not a straightforward process. Although yield prediction algorithms can be currently substantially equal to the estimated value, more accurate predictions are required. Most of the Random forest (RF) applications in machine learning have centered on its application as a classification tool, with only a few research exploring its regression capabilities for yield prediction (Vincenzi et al., 2011; van Klompenburg et al., 2020). In this research work Random forest (RF) model was trained and validated for rice yield prediction and for that combined harvester dataset implied over 7000 points were collected in 30 rice fields across Shanwei, Guangdong. The main objective of this study was following:

1. Develop a yield prediction model in machine learning (RF) based on harvester data and compare it with linear regression (LR) or Decision tree (DT).
2. Identify how different combinations of data influence the accuracy of rice yield prediction.
3. Identify and mapping the existence of field crops in general, and rice in the study area, which are important preludes to mapping rice yields in the region.
4. Presents a simplified approach to mapping rice yields, based on combined Sentinel-2 and environmental data and compare estimates to available field data.
5. Finally, discusses the key importance and necessity of this study and implications for future research.

2. Study area

We choose Shanwei (22°47′14′′N, 115°22′32′′E) in eastern Guangdong province as the study area (Fig. 1). This study area is chosen because of its unique role in agriculture and rice production (Qiu et al., 2016). Located on the north shore of the South China Sea, Shanwei retains a total area of about 4,861.79 km^2 with a coastline of 302 km. The average annual rainfall was around 1,891 mm and the average temperature 22.60 °C, respectively. The study area has a humid subtropical monsoon climate (Tu et al., 2020). The landscape of Shanwei slopes from north to south: the north part is mostly mountainous, and the south is covered by plains and hills.

The total cropland areas were 149244 ha and sown rice areas were 68,585 ha with reference to agricultural census data from the National Statistical Bureau of China (NSBC) (https://www.stats.gov.cn/tjsj/ndsj/). Rice has been planted in the area for more than 40 years (Zhang et al., 2019). The rice is cultivated with a rotation system: early rice and late rice. The other major ground categories within the study area are mountain forests, fishponds, sugar cane, bananas, fruits trees, and other kinds of economic crops.

3. Materials and methods

Fig. 2 illustrates the methodology applied in this research work, demonstrating how the combined satellite and other data were analyzed for yield prediction. Broadly-five major types of data were used as satellite, environmental, soil, crop, and topographic data. These data were obtained from different sources in different format thus first converted this heterogeneous data into homogeneous data and remove all errors such as radiometric, geomantic, and atmospheric errors. Then standardized whole data from 0 to 1 range therefore all individual data will get similar weight in the analysis process/model and any individual indicator will not affect the results. Furthermore, all indicators/thematic layer were run in RF, DT and LR models rice yield estimation. Finally,

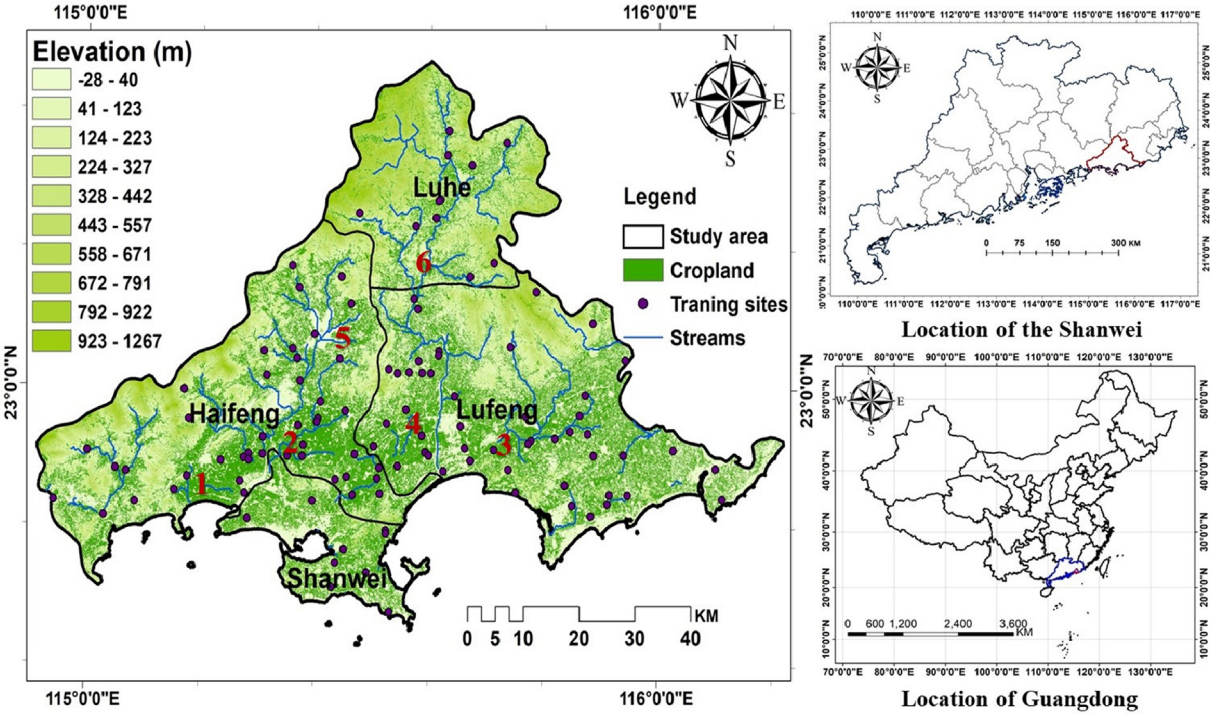


Fig. 1. The location map of the study area by using the digital elevation model (DEM).

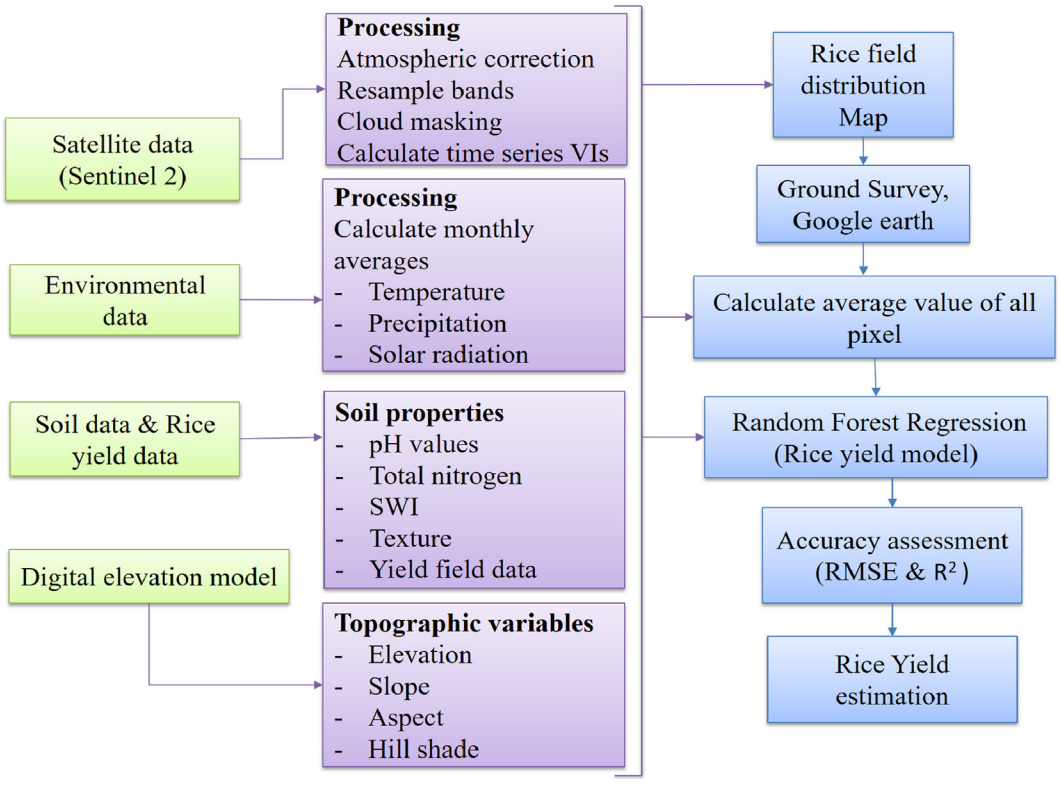


Fig. 2. A flow chart of the methodology applied to yield prediction.

accuracy assessment was check through RMSE and R^2 (Fig. 2).

3.1. Sentinel-2 data processing

Sentinel-2 satellites have an optical sensor called Multi-Spectral Instrument (MSI) that obtains data from 13 spectral bands of different resolutions (Boori et al., 2020) (Table 1). The Sentinel-2 images were downloaded from the Copernicus Open Access Hub. All bands were resampled at 10 m resolution before the use.

3.2. Vegetation indices

We calculated a set of VI that has been utilized in the yield prediction. An overview of the VI and their definitions are described in Table 2. This phenological property of rice was discovered by correlating numerous vegetation indices such as Normalized Difference Vegetation Index (NDVI), Rice Growth Vegetation Index2 (RGVI2), Enhanced Vegetation Index (EVI), and Land Surface Water Index (LSWI) from different seasons of rice areas (Zhou et al., 2017).

The Normalized Difference Vegetation Index (NDVI) has been widely used for evaluating vegetation growth. It represents the reflection differences in green vegetation in the visible and near-infrared (NIR) regions of the spectrum to provide information on the plant condition. The Land Surface Water Index (LSWI) employs the electromagnetic spectrum's shortwave infrared (SWIR) and near-infrared (NIR) regions. There is very strong light absorption through liquid water in the SWIR. The LSWI is recognized to be delicate to the total amount of liquid water in plant as well as its soil background. The Enhanced vegetation index (EVI) was developed to improve the LAI index's indication. The EVI is a blue band absorption in radiant energy that aids to distinguish soil from plants. The rice growth vegetation index 2 (RGVI2) uses the spectral structures of red, near infrared (NIR), and SWIR bands. These VI are beneficial to agronomists because they provide statistical and spatial information on rice crop growth at the field scale.

3.3. Precipitation and temperature

The monthly mean temperature (°C) was downloaded from Earth-Explorer USGS website. While precipitation data was download from NOAA-NCDC. Both data downloaded from entire year. The Soil Water Index (SWI) is a measure of soil moisture. The SWI values were acquired from the Sentinel-1 C-band SAR dataset. The SWI images have a resolution of 1 km (Table 3).

3.4. Digital elevation model

The Digital Elevation Model (DEM) was obtained from the 30 m SRTM Tile downloader (Boori et al., 2021). This data was utilized to construct slope, hill shade, and aspect variables (Table 3). Other data, such as soil (soil ph. values, total nitrogen and soil texture), was downloaded from the Land-Atmosphere Interaction Research Group at Sun Yat-sen University.

3.5. Rice yield data

Field data for the rice classification and yield prediction were acquired from several sources. These data included geographic information in the form of field boundary polygons or sample points obtained from within the field. These spatial geometries were linked to attributes, which included observations of the crop type in a field and any intercropping, as well as yields measured with a series of crop cutting. Sample points were generated in the centre of each raster cell. We collapsed all crop type data into binary classes for rice classification. To generate training points for rice classification, we merged point data with sampled from within field polygons. The sample data was divided into two datasets: training and validation. Other data was obtained from "Shanwei statistical yearbook on agriculture".

3.6. Linear regression (LR)

Linear regression is utilized here. It is an elaboration of simple linear regression when there are numerous indepen-

Table 1
The details of the Sentinel-2 data.

Spectral band	Central wavelength (nm)	Bandwidth (nm)	Spatial resolution (m)
Band 1 Coastal aerosol	443	20	60
Band 2 Blue	490	65	10
Band 3 Green	560	35	10
Band 4 Red	665	30	10
Band 5 Vegetation red edge	705	15	20
Band 6 Vegetation red edge	740	15	20
Band 7 Vegetation red edge	783	20	20
Band 8 NIR	842	115	10
Band 8a Narrow NIR	865	20	20
Band 11 SWIR1	1610	90	20
Band 12 SWIR2	2190	180	20

Table 2
The spectral vegetation indices were computed based on the following equations:

Index	Description	Equation
NDVI	Normalized different vegetation index	$NDVI = \frac{NIR-Red}{NIR+Red}$
EVI	Enhanced vegetation index	$EVI = G * \frac{NIR-Red}{NIR+C1*Red-C2*Blue+L}$
LSWI	Land surface water index	$LSWI = \frac{NIR-SWIR}{NIR+SWIR}$
RGVI2	Rice growth vegetation index2	$RGVI2 = 1.05 - \frac{Blue+Red}{NIR+SWIR1+0.5}$

NIR: near infrared band, Red: red band, Blue: blue band, SWIR: short wave infrared, G: gain factor (2.5), C: coefficients (C1:6, C2:7.5), L: canopy background (1).

Table 3
The details of the used data.

Data name	Attribute	Resolution	Acquisition date	Sources
Sentinel-2	NDVI LSWI EVI RGVI2	05-Day temporal & 10 m spatial resolution	15/07/2020 to 15/06/2021, Monthly data	Copernicus Open Ac cess Hub
MODIS 11A2	Temperature (°C)	8-Day temporal & 1 km spatial resolution	15/07/2020 to 15/06/2021, Monthly data	Earth-Explorer USGS https://earthexplorer.usgs.gov/
NOAA-NCDC	Precipitation (mm)	7-Day temporal & 1 km spatial resolution	15/07/2020 to 15/06/2021, Monthly data	NOAA https://www.ncdc.noaa.gov/cdo-web/
Sentinel-1C	SWI	12-Day temporal & 1 km spatial resolution	15/07/2020 to 15/06/2021, Monthly data	https://land.copernicus.eu/global/products/swi
Soil data	Soil ph / N / Soil texture	shp	–	https://globalchange.bnu.edu.cn/research/soil2
DEM	Elevation / Aspect/ hill shade/ Slope	30 m spatial resolution	–	SRTM https://dwtkns.com/srtm30m/
Field data	Rice yield	–	–	Official website of Shanwei https://www.shanwei.gov.cn/swtjj/xhtml/depny.htm

dent factors occur (Sections, n.d.). It is used to consider the impact of different variables. When there is no rational relationship between factors, a mathematical method can be used to try to connect them.

3.7. Decision tree (DT)

The decision tree is a machine learning algorithm that tests circumstances at each tree level and progresses down the tree to identify different decisions (Matzavela and Alepis, 2021). The scenario is determined by the application and the outcome of the decision-making process. It is created by comparing the instance to the split and then proceeding to the next node. Decision trees can manage large amounts of data.

3.8. Random forest (RF)

Random Forest (RF) is a machine learning (Ml) technique. Its ability to tackle the problem of overfitting trees is among the reasons for its success as an ml algorithm (Bhatnagar and Gohain, 2020; Everingham et al., 2016). The RF regression's trees run concurrently. This algorithm starts by randomly sampling a predetermined number of training sets and then generates a distinct tree. A random selection of independent variables is employed to iteratively partition the data at each node into more units within every tree. The trees are completely grown, and the predicted value of a constant response is determined by the mean fitted reaction from all the different trees. The R package is used for RF analysis. It was trained to predict crop yield using various factors.

4. Results

4.1. Crop classification

Ground reference data and rice maps were collected from "Guangdong Academy of Agricultural Sciences" (GDAAS). The official statistics of rice harvested areas were also collected from the local department of agriculture. A field survey was conducted to collect the data. Once the region is selected, we create a recommended number of random points within that region and sequentially visualize each point on the base map at a high zoom level. We divided the classes into five main groups of cover types, namely 1) crop, 2) forest, 3) water bodies, 4) bare land and grassland, 5 urban areas. These classes were defined by visual assessment and based on a previous study in Shanwei. We obtained total 3990 labels in Shanwei, as shown in Fig. 3. The goal of detecting rice pixels was separated into two sections: one to separate cropland pixels from non-crop pixels, and another to distinguish rice pixels from other crop pixels. Now run the random forest classi-

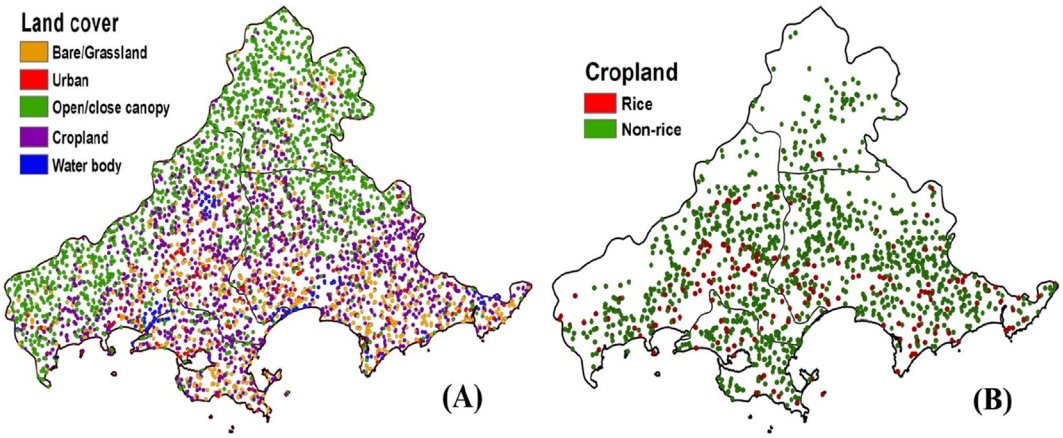


Fig. 3. Location of training fields showing (a) LULC classification and (b) rice/non-rice classification.

fication in machine learning based on selected labels and produce results. As the final goal was to create the rice map, thus generate a cropland map to separate crop pixels as well as reduce the overall interclass variance in the rice classified map.

For a better understanding of the extent of rice fields in the region, a zoomed site of rice parcels demonstration in Fig. 4. The rice crop map and field survey were applied for verification with the mapping results as reference data.

Phenological characteristics have been shown to be a significant factor in deciding different cover types (Fig. 5). The forest regions were detected if enhances vegetation index (EVI) values were greater than 0.5 during at least 365 days. The water bodies and built-up areas were identified if LSWI, RGVI2, and NDVI values were smaller than 0.1 and higher than −0.3 during at least 365 days. Moreover, because rice was grown in plains areas, the areas higher than 150 m, were also identified using the Aster

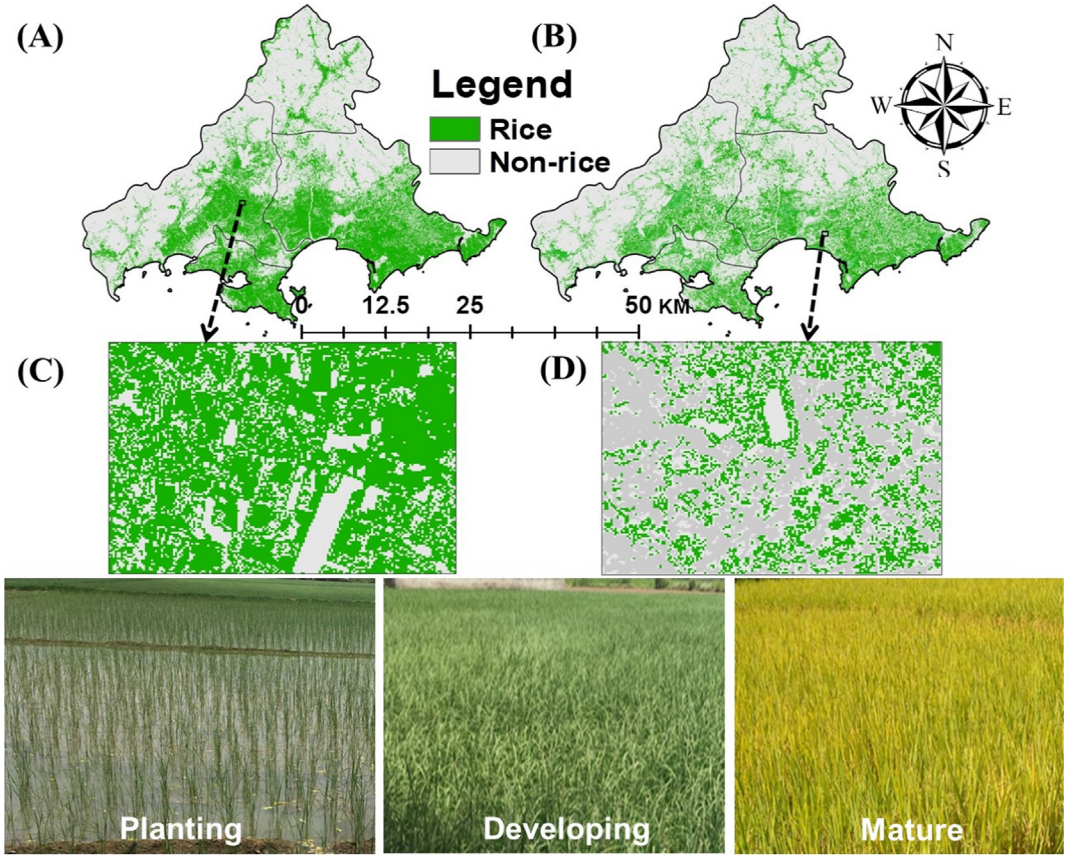


Fig. 4. Maps show: (a) early rice and (b) late rice (c) and (d) zoomed sites as well as some field photographs presented rice growth.

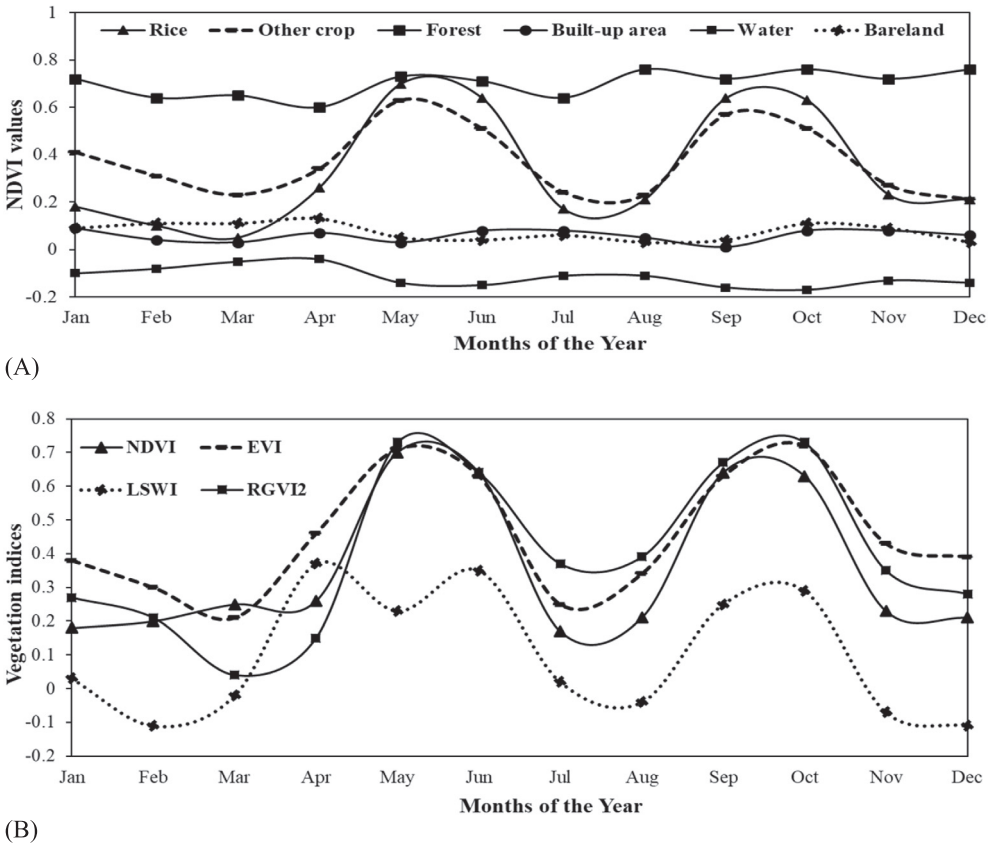


Fig. 5. Phenological profiles of (A) rice or LULC and (B) NDVI, EVI, LSWI, RGVI2 for rice crops.

DEM. Furthermore, two classes of rice and other crop areas were categorized for the first and second crops (Fig. 4).

4.2. Rice yield prediction

In this study, Random Forest (RF) model used for rice yield prediction and compared with Decision Tree (DT), and Linear Regression (LR) methods were using training sites, sample data, and validated data. This research work indicates how different sources of data affect yield prediction accuracy. Results indicate that the RF model expresses the best results and shows higher accuracy in comparison of DTR and LR for yield prediction (Table 4). The results of all analyses were written in the following Table 4. Table 4 shows the best results presented by the RF model with 0.93 R^2 and 0.51 RMSE values. The lowest R^2 is expressed by LR (0.59 & 1.2, R^2 & RMSE respectively) indicates the lowest accuracy in all three models. DT model is in between with 0.89 R^2 and 0.61 RMSE values, respectively.

Table 5 shows RMSE values of specific selected six training sites to validate the results. We take many samples filed but in last finalize 6 filed due to too much variation in terms of crop condition, irrigation system, fertile land, and farming practices. Field 1 and 2 have very good crop condition, while filed 4 and 5 average crop condition and field 5 and 6 have worst crop condition due to above mentioned features, therefor the resulted RMSE values also have variations. It indicates that in all sites RF illustrates the lowest RMSE values means the lowest error, while LR is highest and DT in between, which indicates the highest errors were present in the LR model than DT.

The Sentinel-2 data were resampled at 10 m resolution for this analysis work as higher resolution satellite data provides better results. Results indicate that with the help of secondary data such as environmental, meteorological, topographical, soil moisture data with satellite data, yield estimation and prediction results can improve. Therefore, this research work used satellite data as well as secondary data for better results. In the comparison of all data, RF

Table 4
R-square and RMSE values of DT, LR and RF regressions.

	Decision tree regression	Linear regression	Random forest regression
R-square	0.89	0.59	0.93
RMSE	0.61	1.2	0.51

Table 5
RMSE values of selected fields in the study area.

	Field ID					
	1	2	3	4	5	6
Random Forest	0.58	0.6	0.56	0.65	0.63	0.45
Decision Tree	0.66	0.81	0.7	0.81	0.88	0.53
Linear Regression	1.69	2.5	1.3	1.2	2.41	0.93

shows the best results, later, DT and in the last LR model (Fig. 6).

This suggests that a combination of primary and secondary data improves the yield estimation and prediction. But the yield estimation accuracy based on single date images throughout the years is an important question. In the comparison of all available sentinel-2 images, accuracy estimation was increased from March to June, and the highest accuracy was present for the month of April. Here the addition of secondary data such as environmental data with Sentinel-2 data always improves accuracy for all dates. The inclusion of additional data also improved accuracy estimation. There are many types of data such as temperature, precipitation, topographic, etc. Here topographic data put more effect on estimation accuracy in comparison to other environmental data. However, the single topographic data cannot achieve higher accuracy in comparison

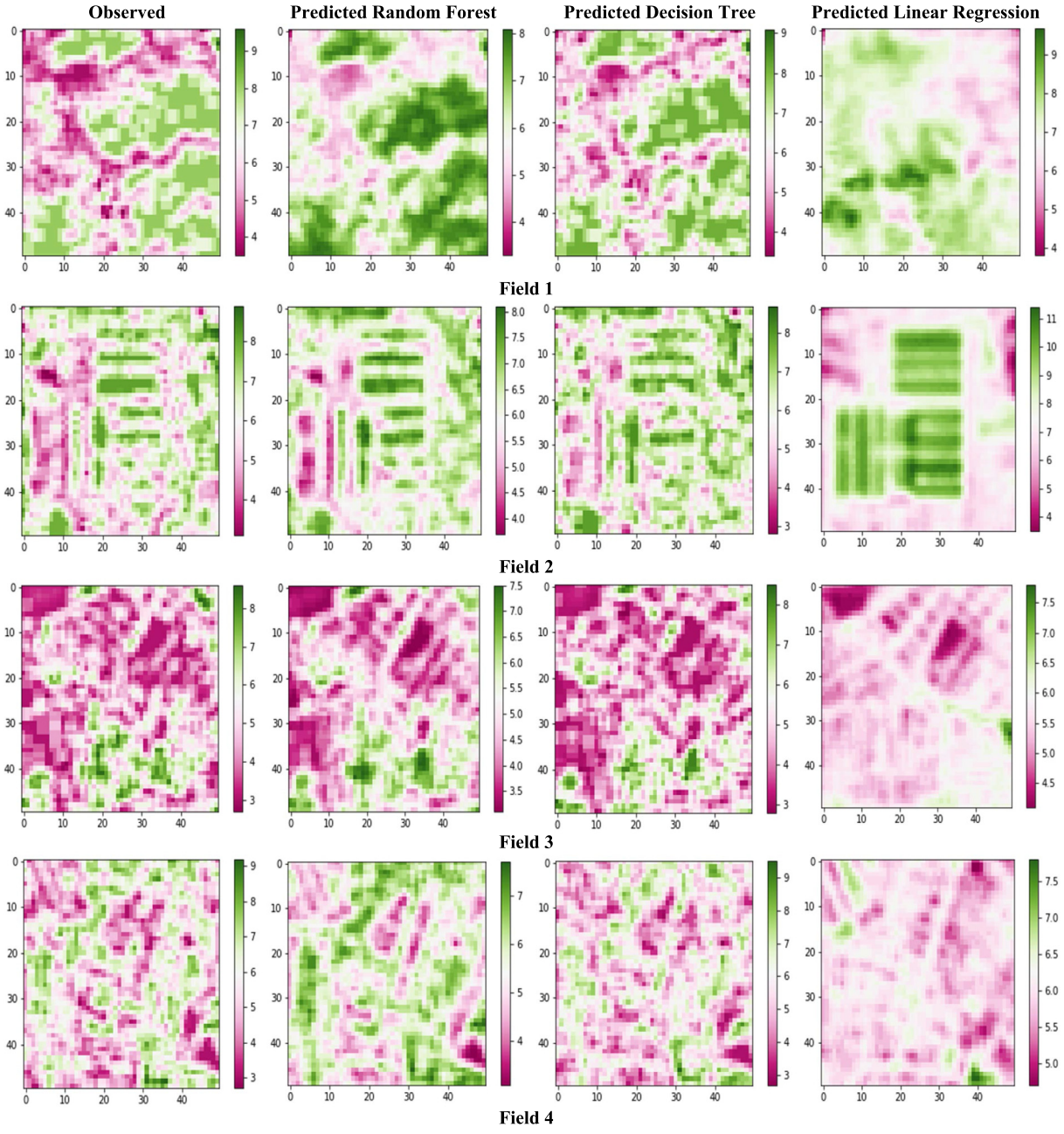


Fig. 6. Observed yield from combined data and predicted yield from the selected sites based on RF, DT and LR.

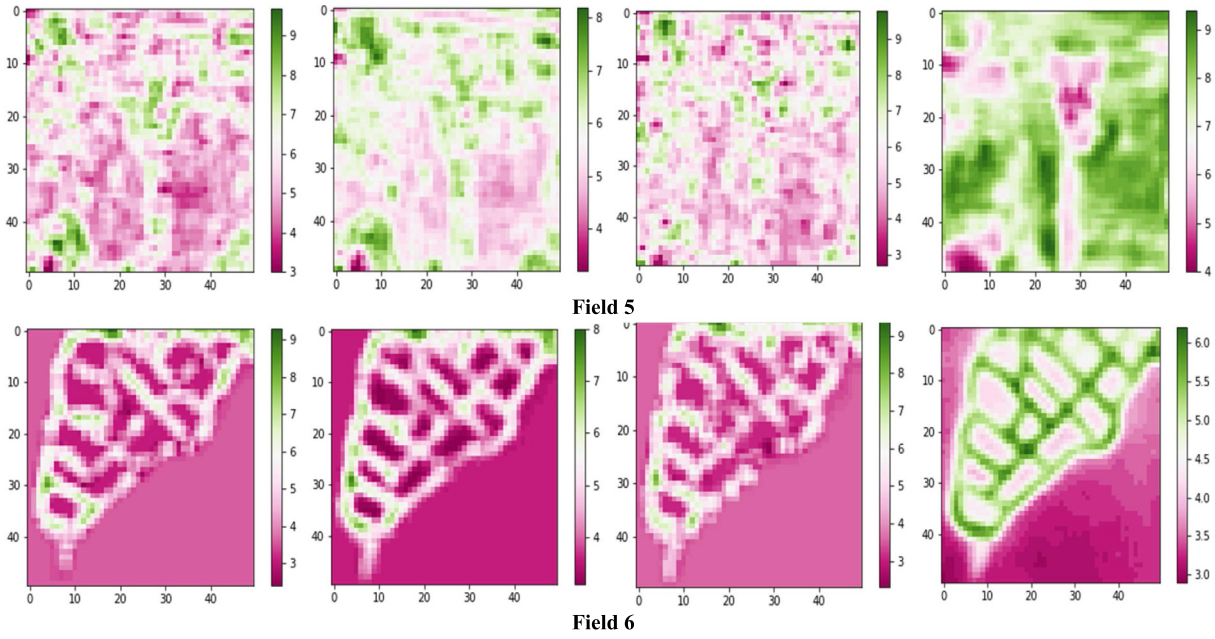


Fig 6. *(continued)*

to the two types of combined environmental data. Fig. 6 indicates that the RF model has the highest yield prediction accuracy from observed data and later DT and the last one LR.

Fig. 6 illustrates the results of observed and predicted yield from RF, DT and LR models for all six sample fields. During ground observation field 1 and field 2 show a healthy and dense rice crop and when compare these results in RF, DT and LR models, RF model predict accurately, DT was little bit less while LR was not accurate. Here green color shows higher production while pinkish color represents very less or no production and this difference and comparison can easily see in Fig. 6 (field 1 & 2). Field 3 and field 4 crop conditions were average due to less fertile land and poor irrigation system. Therefore, in observed values from both fields pinkish color is more dominant in compare of filed 1 and 2. The predicted results from RF, DT, and LR for field 3 and 4 were express similar response as in filed 1 and 2. The fields 5 and 6 have worst crop condition in compare of field 1 to 4 due to unfertile land and worst farming practice. This can also see in observed values in filed 5 and 6 as its show maximum pinkish color. When run regression model, RF model show less production as a higher pinkish color and DT model show average production (average green and pinkish color), while LR model show very high production or higher green color. This is not true as both fields (5 & 6) have very less production. Furthermore, we can say that RF model have highest prediction accuracy, and LR lowest accuracy while DT is in between.

Results indicate that a yield variability can accurately be estimated within the RMSE range of 0.40 to 1.01 t/ha, and it's based on data combination. Generally estimated yield reflects yield variability patterns within an individual field (Fig. 7) and the accuracy is reflected when evaluating the observed and estimated yield. The yield's frequency distribution comparison of separately field demonstrates that the RF model's ability is variate according to fields. Fig. 7 represents the RF model's frequency distribution, which indicates the best results in comparison to DT and LR. The yield distribution shapes also variate in fields, with roughly showing simple unimodal distributions. When we were compared two field distributions, we found that the model values were overstates**.** Normally the model provides accurate field yield estimation variability with the 0.45 to 2.41 t/ha ranges of RMSE in an individual field (Fig. 7). Fig. 8 shows these tendencies.

Fig. 8 shows the accuracy comparison between RF, DT, and LR models. The RF model demonstrates the highest accuracy with 0.93 R^2, while DT has 0.89 R^2 and the lowest one in LR with 0.59 R^2 Values. Therefore, RF model is the best for yield estimation and prediction.

4.3. Rice yield difference

This is one of the best uses of satellite data to accurately monitor and yield estimation mapping in a wide range of landscape areas with field dates. Here RF model was used for yield estimation and prediction from Sentinel-2 images, and it's also compared with DT and LR models. The rice fields were identified by land use land cover maps. To avoid mixed pixels in crop fields especially on boundaries of the crop fields, 10 m buffet techniques were used. This research work covers all individual crop fields of the entire study area and accurately estimates whole areas yield prediction and the whole data range falls within the training data range. The extrapolation outside of the entire input data range was less reliable. Here high resolution of the satellite

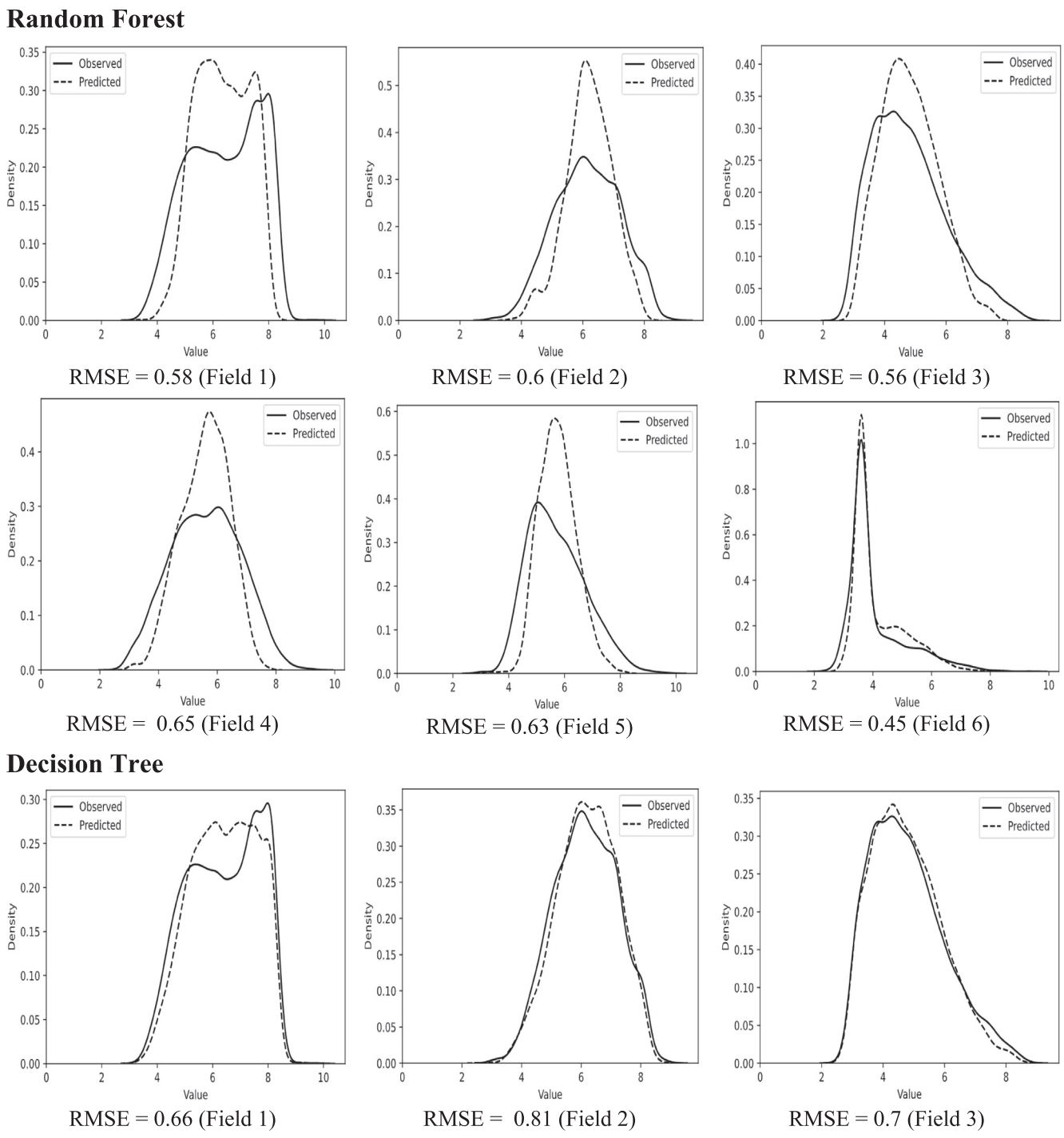
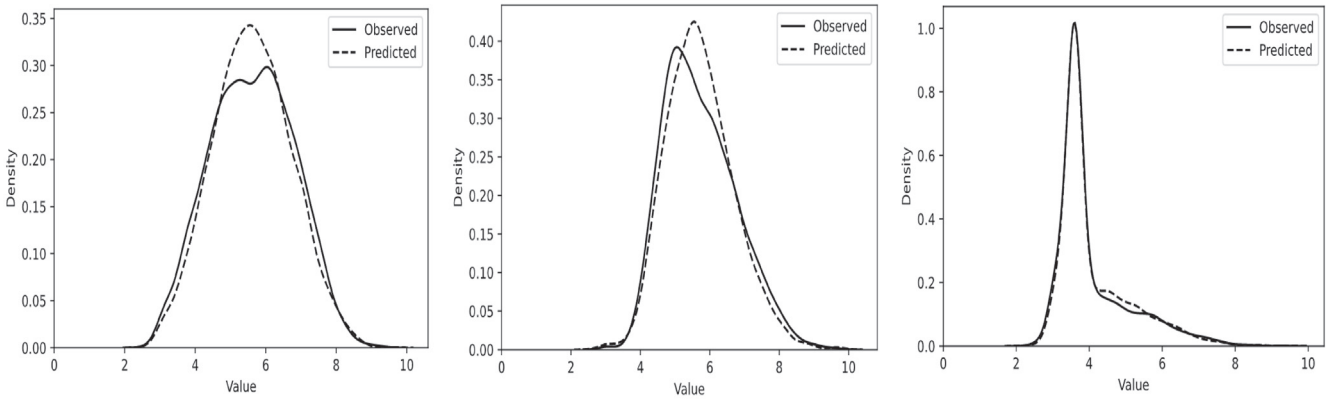


Fig. 7. Graphical representation of frequency distribution of observed and predicated yield from individual fields based on validation data sets.

data helps a lot to identify landscape patterns within the field and in between the fields. For example, the yield ranges from 3 to 8 t/ha in the study area show 168673 t yield production. Therefore, these maps are useful for yield prediction with supporting data under this type of climate region. Therefore, this research helps to identification of yield limiting elements to enhance the efficiency of farming techniques in various areas.

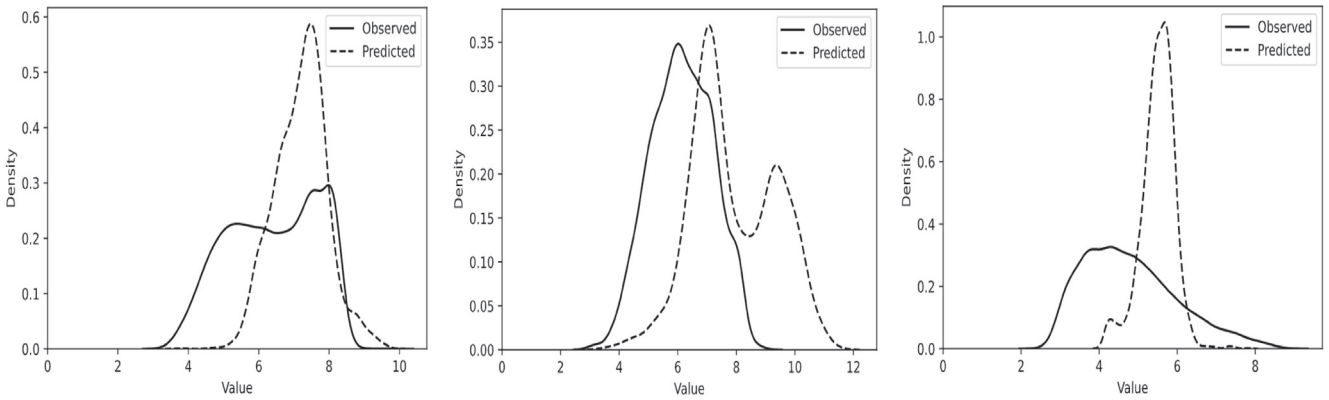
4.4. Accuracy

Yield estimation and prediction accuracy from Sentinel-2 images during growing season normally provide good accuracy. The best results were obtained during April month images little bit less in May month images due to sensor properties and climate effect such as signal–noise ratio during the growing season, sun angle, incoming radi-

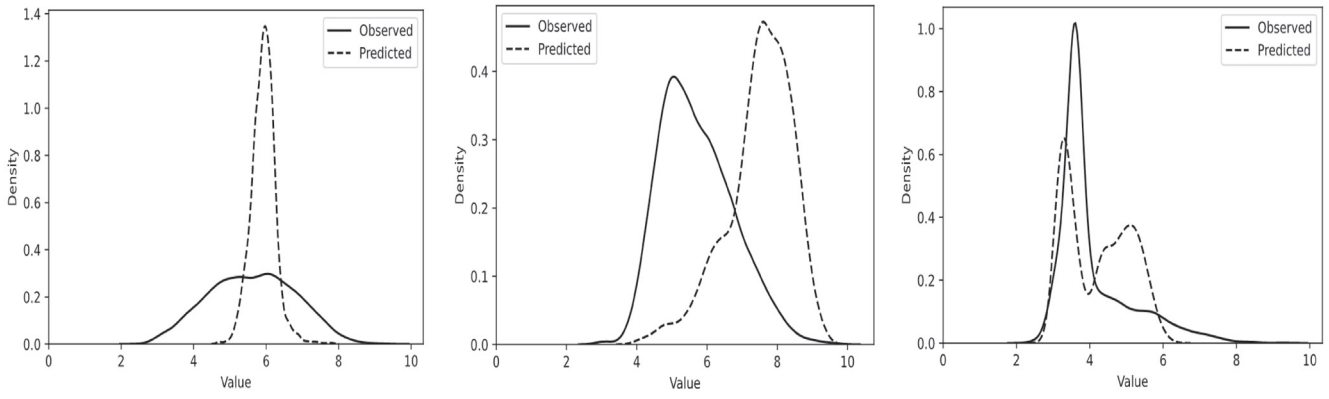


RMSE = 0.81 (Field 4) RMSE = 0.88 (Field 5) RMSE = 0.53 (Field 6)

Linear Regression



RMSE = 1.69 (Field 1) RMSE = 2.5 (Field 2) RMSE = 1.3 (Field 3)



RMSE = 1.2 (Field 4) RMSE = 2.41 (Field 5) RMSE = 0.93 (Field 6)

Yield (t/ha)

Fig 7. *(continued)*

ation intensity, etc. Other than this in the growing season when plant canopy was not developed enough cannot give sufficient spectral reflectance to capture in satellite image but later, when plant canopy developed in later growing stage provide sufficient radiation reflectance for satellite detection so can be measure more accurately. Fig. 9 shows that visual interpretation of Sentinel-2 images was not easy or not possible in June to July due to crop just ripening or beginning to ripen and this effect on accuracy. While these things are overcome in April-May, and we can easily measure and interpret satellite image and accuracy is automatically increase.

5. Discussion

RF regression model was developed in this study based on multi-variable single date VI data; also compare DT and LR with RF. Results indicate RF model shows the best

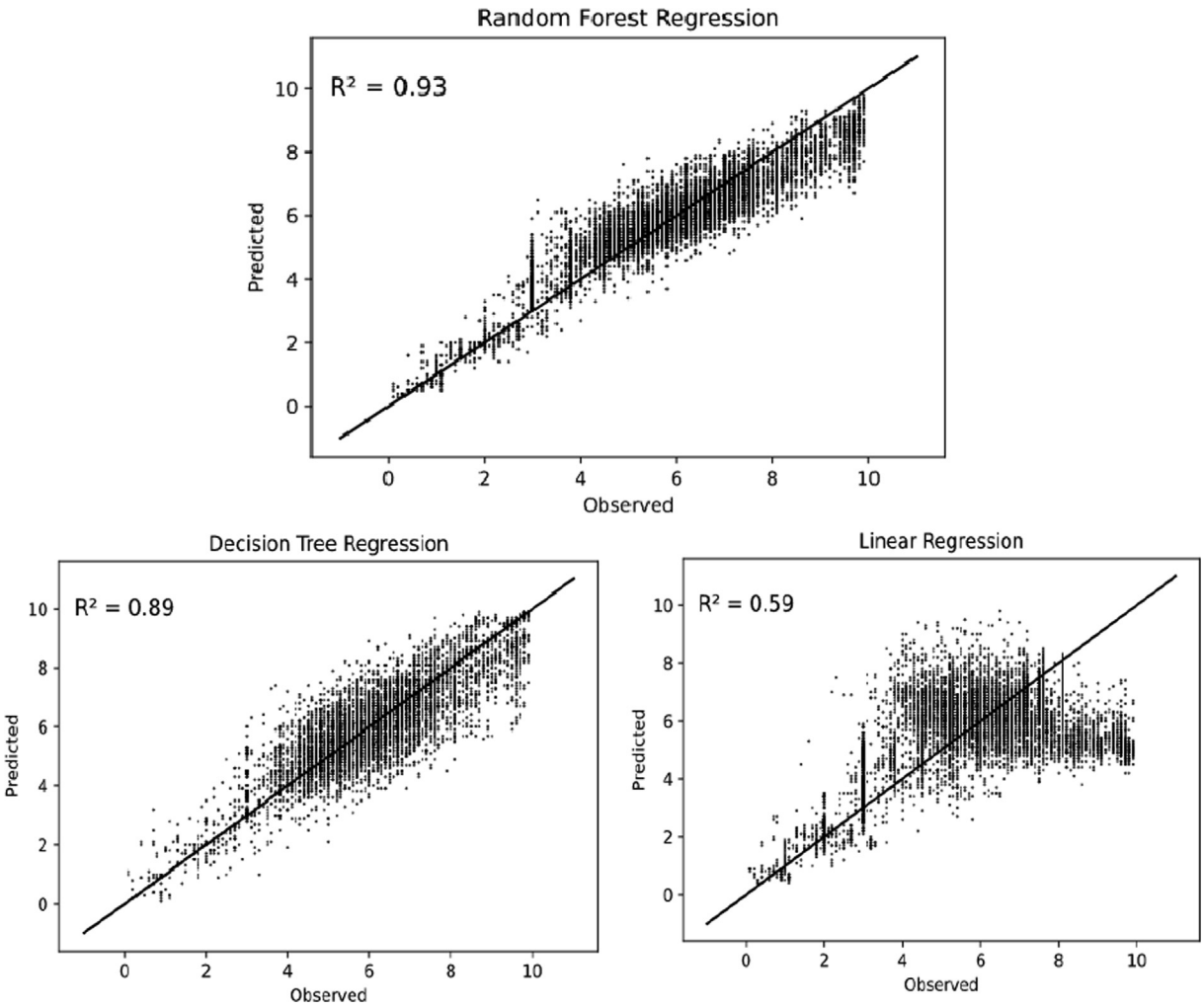


Fig. 8. Accuracy graphs of random forest, decision tree, and linear regression between observed and predicted yield.

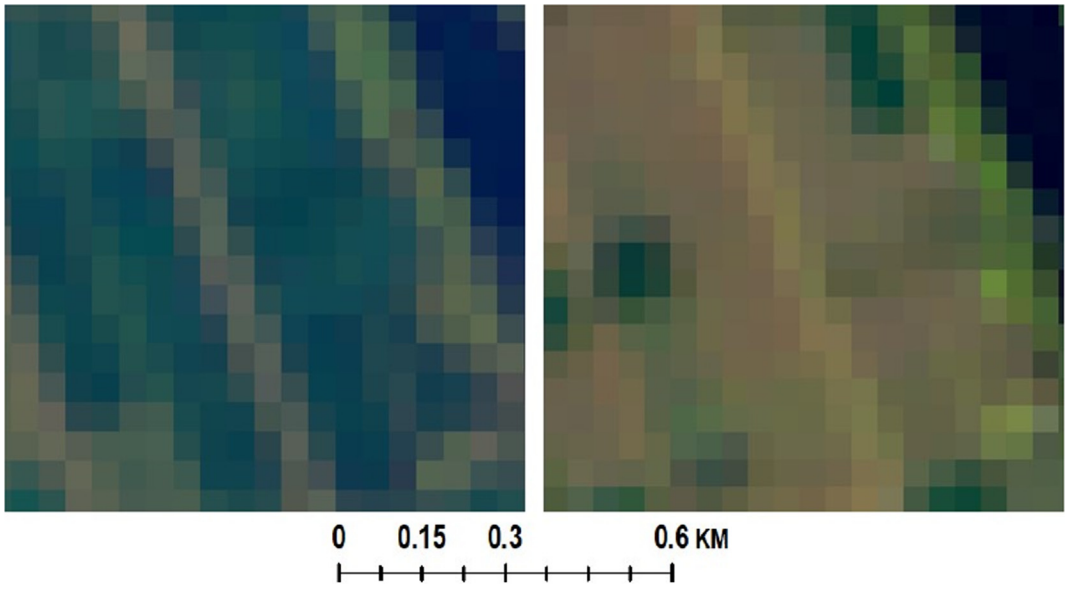


Fig. 9. Optical clarification of consecutive satellite images for (1) June; and (2) July.

performance for yield mapping and prediction based on multi-variable data. RF has been used for image classification and yield prediction. However, the RF model has a lot of advantages in comparison to old and traditional regression modes for yield perdition, which is also demonstrated in the results such as the RF model using many data present in training. RF chooses a subset of the calibration dataset at random to use for model accuracy. Few other things such as splitting of the data into training, outside data validation means checking that the model is overfitting or not.

RF model also make relationships between multi-variable data and control confusing aspects. This research uses different data sets which produce accurate yield prediction results. As a result, data acquisition is a significant benefit in the RF model, such as using soil water index in RF; however, in LR we cannot find any correlation. Thus, the accuracy of RF is always higher due to adding secondary data as well as getting relationships in between different sources of data such as SWI, yield, and spectral reflectance.

Therefore, this is a key advantage of the RF model to use multi-variate relationships from types of heterogeneous data and make the RF model as advanced in comparison of DT and LR. Earlier research used different vegetation indices for yield prediction (Chandra et al., 2020; Li et al., 2022; Bian et al., 2022). This research work finds that there was not any improvement in accuracy by using VI and Sentinel-2 data. But RF makes relationships between VI and Sentinel-2 bands and provides higher accurate results (Rahmati et al., 2022). RF is also simple in processing and less time-consuming (Saha et al., 2022; Li et al., 2022). The results indicate that Sentinel-2 data have the capacity to provide accurate prediction with field–variables. We also observe that higher resolution provides better results in comparison to lower resolution (Camalan et al., 2022; Fernández-Habas et al., 2021).

Normally Sentinel-2 images have higher location accuracy in comparison to other satellite images such as SPOT. Such things affect the accuracy and are sometimes very important for a specific climate region. Accuracy also depends on types of data, sources of data, timings, the season of the data, types of crops, the topography of the data and local climate, framing practice, and their combinations (Schulz et al., 2021; Bian et al., 2022). Yield mapping, monitoring, and prediction also affected some statistical analyses such as calibration errors, time delays, and combine operational errors. Thus, in this research work, all these errors and advantages compile for final yield estimation and prediction accuracy (Xin et al., 2021; Camalan et al., 2022). Earlier studies were used different data correction processes and still there is no universal process. We can assume that correction in data or applying thresholds values affects final accuracy. The image resolution is also an important parameter that affects accuracy as 10 m Sentinal-2 image provides better results than 20 m resolution may be due to sensor quality, design, data processing with harvester data, etc. Thus, best feting or matching in different primary and secondary data is an important factor in results accuracy (Chandra et al., 2020; Chowdhuri et al., 2021).

Shanwei is a coastal area, so satellite data are mostly cloudy and crop prediction was very difficult. As a research, MODIS is an excellent alternative for these areas. A cloud-free image can easily analyze crop growth dynamics and provide good and accurate results. But for the small size of field (approximately 3 ha) low-resolution images are not suitable due to many mixed pixels. Therefore, before study need to analyze research work requirements, objectives, possibilities, and data availability, etc.

As this research work was done on Shanwei, therefore, getting a higher resolution cloud-free data was a necessary and challenging task. While this study used Sentinel-2 images but commercial higher resolution data such as RapidEye, Planet Labs was also a good option. This provides a more detailed assessment due to higher resolution and cloud-free data with secondary data such as temperature and precipitation information. The secondary data such as environmental, metrological, soil moisture, temperature, precipitation increase accuracy due to more accurate site-specific information.

6. Conclusion

This research shows that Sentinel-2 data is ideal for rice yield prediction and advanced classification. The Sentinel-2 data integrated with other data produces additional accurate yields prediction. Random forest based algorithms can successfully detect cropland and rice pixels by using a high number of training labels. Furthermore, RF regression shows to provide higher yield prediction accuracy, than the typically used VI-based decision tree and linear regression. Completing such a mapping in Shanwei is extremely challenging, because of the various smallholder farming landscapes, cloud coverage, and sparse ground data. This study can also apply to different crop yield predictions and estimations on different landscape levels. Further work should be tested on the methodology utilized in this research can be applied to other areas, and crop varieties. Advance research is required with more spectral variables, statistics, and procedures to advance the estimate accurateness which is essential for precision agriculture.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the Hong Kong Ph.D. scholarship from PolyU and research grants from the Research Grants Council of (HKSAR) grant project codes B-Q49D

and 1-ZVE8. The authors would also like to acknowledge the support drawn from the Agriculture department of Guangdong, China.

References

Bhatnagar, R., Gohain, G.B., 2020. Crop yield estimation using decision trees and random forest machine learning algorithms on data from Terra (EOS AM-1) & Aqua (EOS PM-1) satellite data. Stud. Comput. Intell. 835. https://doi.org/10.1007/978-3-030-20212-5_6.

Bian, C., Shi, H., Wu, S., Zhang, K., Wei, M., Zhao, Y., Sun, Y., Zhuang, H., Zhang, X., Chen, S., 2022. Prediction of Field-Scale Wheat Yield Using Machine Learning Method and Multi-Spectral UAV Data. Remote Sens. 14 (6), 1474. https://doi.org/10.3390/rs14061474.

Boori, M.S., Choudhary, K., Kupriyanov, A.V., 2020. Crop growth monitoring through sentinel and landsat data based ndvi time-series. Comput. Opt. 44 (3), 409–419. https://doi.org/10.18287/2412-6179-CO-635.

Boori, M.S., Choudhary, K., Paringer, R., et al., 2021. Spatiotemporal ecological vulnerability analysis with statistical correlation based on satellite remote sensing in Samara, Russia. J. Environ. Manage. 285. https://doi.org/10.1016/j.jenvman.2021.112138.

Camalan, S., Cui, K., Pauca, V.P., Alqahtani, S., Silman, M., Chan, R., Plemmons, R.J., Dethier, E.N., Fernandez, L.E., Lutz, D.A., 2022. Change Detection of Amazonian Alluvial Gold Mining Using Deep Learning and Sentinel-2 Imagery. Remote Sens. 14 (7), 1746. https://doi.org/10.3390/rs14071746.

Chandra, P., 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 Sens. Appl.: Soc. Environ. 19, 100367. https://doi.org/10.1016/j.rsase.2020.100367.

Chen, J., Huang, J., Hu, J., 2011. Mapping rice planting areas in southern China using the China Environment Satellite data. Math. Comput. Modell. 54 (3–4), 1037–1043. https://doi.org/10.1016/j.mcm.2010.11.033.

Choudhary, K., Shi, W., Boori, M.S., et al., 2019. Agriculture Phenology Monitoring Using NDVI Time Series Based on Remote Sensing Satellites : A Case Study of Guangdong, China. Optical Memory Neural Networks 28 (3), 204–214. https://doi.org/10.3103/S1060992X19030093.

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

Chowdhuri, I., Pal, S.C., Chakrabortty, R., et al., 2021. Torrential rainfall-induced landslide susceptibility assessment using machine learning and statistical methods of eastern Himalaya. Nat Hazards 107, 697–722. https://doi.org/10.1007/s11069-021-04601-3.

Everingham, Y., Sexton, J., Skocaj, D., et al., 2016. Accurate prediction of sugarcane yield using a random forest algorithm. Agron. Sustainable Dev. 36 (2). https://doi.org/10.1007/s13593-016-0364-z.

Fernández-Habas, J., Moreno, A.M.G., Hidalgo-Fernández, M.T., Leal-Murillo, J.R., Oar, B.A., Gómez-Giráldez, P.J., González-Dugo, M.P., Fernández-Rebollo, P., 2021. Investigating the potential of Sentinel-2 configuration to predict the quality of Mediterranean permanent grasslands in open woodlands. Sci. Total Environ. 791. https://doi.org/10.1016/j.scitotenv.2021.148101, ISSN 0048-9697.

Gao, M., Qin, Z., Zhang, H., et al., 2008. Remote sensing of agro-droughts in Guangdong Province of China using MODIS satellite data. Sensors 8 (8), 4687–4708. https://doi.org/10.3390/s8084687.

Gomes, F., Queiroz, G.R., Ferreira, R., 2020. An overview of platforms for big earth observation data management and analysis. Remote Sens. 12 (8), 1–25. https://doi.org/10.3390/RS12081253.

Gorelick, N., Hancher, M., Dixon, M., et al., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031.

Jeong, S., Ko, J., Yeom, J.M., 2022. Predicting rice yield at pixel scale through synthetic use of crop and deep learning models with satellite data in South and North Korea. Sci. Total Environ. 802, 149726. https://doi.org/10.1016/j.scitotenv.2021.149726.

Jin, Z., Azzari, G., You, C., et al., 2019. Smallholder maize area and yield mapping at national scales with Google Earth Engine. Remote Sens. Environ. 228, 115–128. https://doi.org/10.1016/j.rse.2019.04.016.

Kailou, L., Huiwen, H., Lijun, Z., et al., 2015. Estimating Rice Yield Based on Normalized Difference Vegetation Index at Heading Stage of Different Nitrogen Application Rates in Southeast of China. J. Environ. Agric. Sci. 2, 13.

Li, Z., Ding, L., Dawei, X.u., 2022. Exploring the potential role of environmental and multi-source satellite data in crop yield prediction across Northeast China. Sci. Total Environ. 815, 152880. https://doi.org/10.1016/j.scitotenv.2021.152880.

Matzavela, V., Alepis, E., 2021. Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments. Comput. Educ.: Artif. Intell. 2, 100035. https://doi.org/10.1016/j.caeai.2021.100035.

Mosleh, M.K., Hassan, Q.K., Chowdhury, E.H., et al., 2015. Application of remote sensors in mapping rice area and forecasting its production: A review. Sensors (Switzerland) 15 (1), 769–791. https://doi.org/10.3390/s150100769.

Qiu, B., Li, W., Tang, Z., et al., 2015. Mapping paddy rice areas based on vegetation phenology and surface moisture conditions. Ecol. Ind. 56, 79–86. https://doi.org/10.1016/j.ecolind.2015.03.039.

Qiu, B., Qi, W., Tang, Z., et al., 2016. Rice cropping density and intensity lessened in southeast China during the twenty-first century. Environ. Monit. Assess. 188 (1), 1–12. https://doi.org/10.1007/s10661-015-5004-6.

Rahmati, A., Zoej, M.J.V., Dehkordi, A.T., 2022. Early identification of crop types using Sentinel-2 satellite images and an incremental multi-feature ensemble method (Case study: Shahriar, Iran). Adv. Space Res. https://doi.org/10.1016/j.asr.2022.05.038.

Saha, S., Mallik, S., Mishra, U., 2022. Groundwater Depth Forecasting Using Machine Learning and Artificial Intelligence Techniques: A Survey of the Literature. In: Das, B.B., Hettiarachchi, H., Sahu, P.K., Nanda, S. (Eds.), Recent Developments in Sustainable Infrastructure (ICRDSI-2020)—GEO-TRA-ENV-WRM. Lecture Notes in Civil Engineering, vol. 207. Springer, Singapore. https://doi.org/10.1007/978-981-16-7509-6_13.

Schulz, D., Yin, H., Tischbein, B., Verleysdonk, S., Adamou, R., Kumar, N., 2021. Land use mapping using Sentinel-1 and Sentinel-2 time series in a heterogeneous landscape in Niger, Sahel. ISPRS J. Photogramm. Remote Sens. 178, 97–111. https://doi.org/10.1016/j.isprsjprs.2021.06.005, ISSN 0924-2716.

Sections, T. (n.d.). I'llSimple Linear Regression I-Least Squares Estimation. 100(10).

Song, P., Mansaray, L.R., Huang, J., et al., 2018. Mapping paddy rice agriculture over China using AMSR-E time series data. ISPRS J. Photogramm. Remote Sens. 144, 469–482. https://doi.org/10.1016/j.isprsjprs.2018.08.015.

Tu, Y., Lang, W., Yu, L., et al., 2020. Improved Mapping Results of 10 m Resolution Land Cover Classification in Guangdong, China Using Multisource Remote Sensing Data with Google Earth Engine. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 5384–5397. https://doi.org/10.1109/JSTARS.2020.3022210.

van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. Comput. Electron. Agric. 177, 105709. https://doi.org/10.1016/j.compag.2020.105709.

Vincenzi, S., Zucchetta, M., Franzoi, P., et al., 2011. Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. Ecol. Modell. 222 (8), 1471–1478. https://doi.org/10.1016/j.ecolmodel.2011.02.007.

Wang, X.S., Ryoo, J.H.J., Bendle, N., Kopalle, P.K., 2021. The role of machine learning analytics and metrics in retailing research. J.

Retailing 97 (4), 658–675. https://doi.org/10.1016/j.jretai.2020.12.001, ISSN 0022-4359.

Yang, G., Yu, W., Yao, X., et al., 2021. AGTOC: A novel approach to winter wheat mapping by automatic generation of training samples and one-class classification on Google Earth Engine. Int. J. Appl. Earth Obs. Geoinf. 102, 102446. https://doi.org/10.1016/j.jag.2021.102446.

Zhang, K., Yu, Y., Dong, J., et al., 2019. Adapting & testing use of USLE K factor for agricultural soils in China. Agric. Ecosyst. Environ. 269, 148–155. https://doi.org/10.1016/j.agee.2018.09.033.

Zhao, H., Mo, Z., Lin, Q., et al., 2020. Relationships between grain yield and agronomic traits of rice in southern China. Chilean J. Agric. Res. 80 (1), 72–79. https://doi.org/10.4067/s0718-58392020000100072.

Zhou, X., Zheng, H.B., Xu, X.Q., et al., 2017. Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery. ISPRS J. Photogramm. Remote Sens. 130, 246–255. https://doi.org/10.1016/j.isprsjprs.2017.05.003.