ESTIMATION OF NDVI FOR CLOUDY PIXELS USING MACHINE LEARNING

R.Agrawal[1], J.D.Mohite[1], S.A.Sawant[1]*, A.Pandit[1], S. Pappula[2]

[1] TCS Research and Innovation, Tata Consultancy Services, Mumbai, India
[2] TCS Research and Innovation, Tata Consultancy Services, Hyderabad, India
(rishabh.agrawal3, jayant.mohite, suryakant.sawant, ankur.pandit, srinivasu.p)@tcs.com

Commission III, WG 10

KEY WORDS: Cloud Free NDVI, Multi-spectral, SAR, Remote Sensing, Machine Learning

ABSTRACT:

The Normalized Difference Vegetation Index (NDVI) is a useful index for vegetation monitoring. However, due to cloud cover the observations of NDVI are discrete and vary in the intensity. Therefore, there is a need to estimate the NDVI during cloud cover using alternative sources of satellite observations. The main objective of this study is to estimate NDVI during cloudy conditions using moderate resolution multi-spectral and synthetic aperture radar (SAR) observations. Two approaches were identified: 1) pixel replacement and 2) machine learning based regression analysis to estimate cloud free NDVI. Moderate Resolution Imaging Spectroradiometer (MODIS) 8-day NDVI composite, Sentinel-1 SAR and cloud masked Sentinel-2 multi-spectral observations were collected for entire cropping season. The satellite observations were selected only for agricultural areas by applying the agriculture, non-agriculture land use land cover mask. Machine learning algorithms such as Linear Regression (LR), Random Forest Regression (RFR), and Support Vector Regression (SVR) were used for NDVI estimation. Regression analysis was performed using Sentinel-2 NDVI as an independent variable and VV, VH, Cross Ratio (i.e., VV/VH), and MODIS NDVI as dependent variables. NDVI of the cloudy pixel was estimated using the trained regression models over the agriculture areas. A regression model was trained and applied to each Sentinel-2 tile that covers an area of 100 km x 100 km. The RFR and SVR showed the highest R^2 of 0.73 and a RMSE of 0.12. A visual comparison of time series graphs showed good alignment between actual (Sentinel-2) and predicted NDVI and usual crop growth trend.

1. INTRODUCTION

The Normalized Difference Vegetation Index (NDVI) is one of the most widely and frequently used vegetation indexes for vegetation monitoring. Time series of NDVI have been used for a variety of agricultural applications, including crop health monitoring, crop area classification, crop yield prediction, and land cover classification (Zheng et al., 2015, Friedl et al., 2002). These applications require NDVI at high spato-temporal resolution. With the advent of global monitoring satellites like Sentinel 2 (A/B), Landsat 8, and MODIS, massive amounts of multispectral satellite observations are available. However, satellite-based NDVI is often affected by cloud cover and snow. In some regions during the rainy season, due to persistent cloud cover, it is difficult to obtain a continuous time series of NDVI. Therefore, to reduce the effect of clouds, there is a need to remove or replace the cloudy pixel NDVI for spatio-temporal vegetation growth analysis.

Researchers have proposed temporal image compositing and cloud masking techniques to reduce the impact of cloud cover. The most frequently used method is the maximum value composite (MVC) (Liao et al., 2016). This method selects the maximum NDVI over multiple time instances to create a single NDVI output. As the image compositing approach requires at least one cloud-free observation in a selected time window, the approach fails over regions having cloud cover for a long duration, i.e., more than 10 to 15 days. Studies have shown that vegetation dynamics in heterogeneous landscapes often require NDVI datasets with both high spatial and temporal resolution. A single satellite based sensor has technical limitations

and cannot achieve the required spatio-temporal coverage (Holben, 1986). Multi-sensor fusion techniques have been used to address the temporal scalability issues. An improved spatial and temporal data fusion approach (ISTDFA) combines 250-m MODIS NDVI product and 10-m Sentinel-2 NDVI data to generate a synthetic Sentinel-2 NDVI time series for monitoring of a crop pest and disease (Wu et al., 2018). To obtain the cloud-free image, a possible alternative is to use Sentinel-1 synthetic aperture radar (SAR) images that provide continuous all-weather, day-and-night imagery in C-band. NDVI derived using Sentinel-1 can be used for crop monitoring (Filgueiras et al., 2019). Also, studies have shown that SAR data can be used for vegetation index monitoring (Baghdadi et al., 2015).

The time-series of radar backscatter (VV, VH, and Cross Ratio) and NDVI derived from optical images are proven to be correlated for multiple crops. SAR VV and cross ratio are correlated with NDVI for Winter Wheat, with a R^2 of 0.74 and 0.58, respectively (Veloso et al., 2017). The study also showed that for the Maize crop in the summer season, R^2 of 0.91 and 0.89 was obtained between NDVI-VH and NDVI-Cross Ratio, respectively. The studies have also focused on neural networks for estimation of the NDVI using SAR-based images. Results showed that the SAR data allows a better reconstruction of the NDVI (Mazza et al., 2021). For agriculture areas, VV and VH are mostly used because cover crops such as Wheat present a high VV response and grasslands and seedlings give a higher VH response (Wu and Sader, 1987). Studies have used the SAR backscatter to identify the relationship between different vegetation indices at field level (Macelloni et al., 2001). The shape of a plant has a impact on the relationship between SAR backscatter and biomass, specifically backscatter decreases with the

---

* Corresponding author

increase in biomass in the case of small plants with narrow leaves (Huang et al., 2019). Several studies have compared the NDVI derived from various satellite observations with ground-based sensors to find the best correlation between satellites and ground-based sensors. Sentinel-2 derived metrics have better agreement than MODIS derived metrics (Lange et al., 2017). Due to coarse spatial resolution, MODIS NDVI is higher compared to Sentinel-2 NDVI. Sentinel-2 NDVI represented the attributes of the agriculture area better than MODIS NDVI due to the better discrimination ability within a small area (M.B. Son and Kim, 2021).

The main objective of this study is to estimate NDVI during cloudy conditions using moderate resolution multi-spectral and SAR observations for agricultural monitoring. The proposed method attempts to identify the relationship between non-cloudy pixels from multiple satellite observations using machine learning algorithms like LR, RFR, and SVR. As cloud cover, weather and crop growth conditions are dynamic across geographical extent single machine learning model fails to generalize over large area. We propose a framework for dynamic model training and evaluation to consider local variations in the vegetation and effective cloud free NDVI estimation.

2. MATERIALS AND METHODS

This section covers information about the study area, insights into the satellite data used and detailed approach for estimation of NDVI for cloudy pixel using machine learning.

2.1 Study Area

We selected 59 field boundaries from West Bengal and Telangana state of India. Out of 59 fields, ten fields were planted with the Cotton, while remaining 49 had the Rice crop. Rainy cropping season (locally called as *Kharif*) is considered for estimation of NDVI. During *Kharif* cropping season summer monsoon covers Indian sub-continent and region is mostly covered with the dense clouds. The cropping season begins in June-July and lasts until October-November. All selected fields are covered in six Sentinel-2 tiles from West Bengal and Telangana states of India. Four tiles (i.e., T44QWF, T45QWF, T44QWG and T45QWG) covered the area of West Bengal and the remaining two tiles (T44QLD and T44QLE) covered the study area of Telangana state. Figure 1 shows the selected states and Sentinel-2 tile id's.
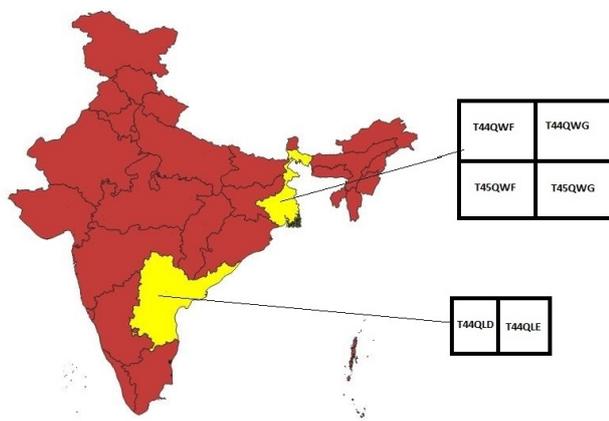

Figure 1. Study Area

2.2 Data Used

In this study, we have used Sentinel-1 SAR, Sentinel-2 multispectral, and MODIS 8-day NDVI composite data for cloud free NDVI estimation. The MODIS has 8-day, 16-day, and one-month composite images with minimal cloud interference. We selected the similar temporal window for both MODIS and Sentinel 2 satellite observation to minimize the effect physical changes on the ground. Therefore, the MODIS 8-day composite was aligned with Sentinel-2 observations. Table 1 shows the selected satellite sensors spatial, temporal resolution and duration of analysis. The details of the satellite pass and data processing approach are covered in the subsequent sections.

Dataset	Spatial Resolution	Temporal Resolution	Duration of analysis	Source
Sentinel-1	10 meter	6 Days	Jun-Oct 2021	ESA
Sentinel-2	10 meter	5 Days	Jun-Oct 2021	ESA
MODIS	250m	8 Days	Jun-Oct 2021	NASA

Table 1. Details of the satellite data

2.2.1 Sentinel-2 Data Collection & Pre-processing
Sentinel-2 is a European wide-swath, high-resolution, multi-spectral imaging mission. The mission is intended to provide high-resolution satellite observations with a 5-day revisit frequency (ESA, 2022b). Each Sentinel-2 multi-spectral observation has 13 spectral bands, of which four, six, and three bands have spatial resolution of 10, 20, and 60 metres, respectively. The data is also available on Amazon Web Services (AWS). The Registry of Open Data on AWS provides the Sentinel-2 observations at Level 2A for the analysis (AWS, 2022). The red and near-infrared bands of Sentinel-2 were used to estimate NDVI. The NDVI is computed as the difference between near-infrared (NIR) and red (RED) reflectance divided by their sum (equation 1). Also, we used the Quality Index (QI) band from processing Level-2A (L2A) for cloud masking. All Sentinel-2 satellite tiles from June to October 2021 were processed for NDVI analysis.

$$NDVI = \frac{NIR - Red}{NIR + Red} \qquad (1)$$

2.2.2 Sentinel-1 Data Collection and Pre-processing
Sentinel-1 imagery is provided by two polar-orbiting satellites, 1-A and 1-B. Sentinel-1 provides continuous all-weather, day-and-night imagery in the C-band (ESA, 2022a). The observations are provided by the satellite on a six-day cycle. The Sentinel 1 Ground Range Detection (GRD) product is provided by AWS (AWS, 2022). We converted VV and VH backscatter of Sentinel-1 from intensity values to decibels and used for analysis. Also, the cross ratio (CR) using VV and VH backscatter was calculated (equation 2). Sentinel-1 observations were resampled to 10 meter spatial resolution, i.e., same as of Sentinel-2 resolution.

$$CR = \frac{\sigma_{VV}}{\sigma_{VH}} \qquad (2)$$

2.2.3 MODIS Data Collection and Pre-processing The 250 meter surface reflectance product MOD09Q1 obtained using Terra satellite was used for the analysis (MOD09Q1, 2022). For each pixel of MOD09Q1 8-day surface reflectance composite was created. The surface reflectance in RED and NIR spectrum was used to calculate NDVI (equation 1). Finally, MODIS 250 meter NDVI was resampled to 10 meter spatial resolution to match Sentinel 2 spatial resolution. The fusion of MODIS and Sentinel-2 NDVI was performed using L2A cloud probability as mask for cloudy pixels. These cloudy pixels were replaced with 8-day MODIS NDVI at same spatial resolution. The detailed methodology for MODIS Sentinel-2 NDVI fusion is skipped for brevity.

2.2.4 Training Data Creation For cloud-free NDVI estimation, we used Sentinel-1 SAR, MODIS NDVI, and Sentinel-2 multi-spectral observations. The six sentinel-2 tiles were used. Based on the Sentinel-2 overpass date, we have downloaded the Sentinel-1 tiles for the previous eight days. For example, if the tile date for Sentinel-2 is July 11, 2021 then Sentinel-1 observations from July 4, 2021, to July 11, 2021 were downloaded and 8-day maximum value composite of Sentinel-1 observation for VV, VH and CR was created. Figure 2 shows overlay of Sentinel 1 and 2 tiles and extraction of Sentinel 1 data for Sentinel 2 extent. In case of MODIS, 8-days composite of nearest previous date with respect to Sentinel-2 overpass date was considered. Further, stack of Sentinel-2 NDVI, Sentinel-1 VH, VV, CR, and MODIS NDVI was created. Land use land cover mask comprising of two classes (i.e., agriculture and no-agriculture) was applied on the stack to consider pixels only from agricultural areas. Sentinel-2 cloud mask provide the cloud probability of each pixel, we considered pixels with cloud probability of zero. Only cloud-free pixels were considered for model training. During the analysis 30 percent fields were removed for validation. Figure 3 shows the location of some of the validation fields.
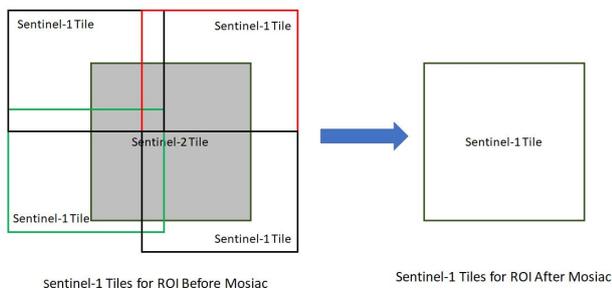

Figure 2. Sentinel 1 and 2 tile overlay

2.3 Machine Learning Model

In the proposed method, we plan to establish the relationship between cloud free pixels of Sentinel-2 NDVI, Sentinel-1 VH, VV backscatter, CR, and MODIS NDVI data at the local (tiles) level using a machine learning algorithm. We used LR, RFR, and SVR to estimate cloud free NDVI on agriculture area. Sentinel-2 NDVI was considered as independent variable and Sentinel-1 VH, VV, CR and MODIS NDVI were used as dependent variables. Hyperparameter tuning was carried out to find the best parameter for the machine learning model. Random Forest was tuned for the maximum depth of tree which was kept at 5, 10, 15, 20, 25 and number of tree at 50, 100, 150, 200, 250, and 300. Similarly, SVR was tuned with the model parameter C ranging between (0.1, 1, 10, 100) and Gamma (1,

Figure 3. Validation fields

0.1, 0.01, 0.001) with Radial Basis Function Kernel. We used the grid search approach with 10-fold cross validation for RFR and SVR. LR model was trained on 80 % data and remaining 20 % was used for model validation. We used 10-fold cross validation for identification of LR model. The cross validation helped to overcome the model overfitting and grid search helped to identify the model parameters for higher accuracy.

Machine learning models were trained separately for each tile and acquisition date. For each tile the data of cloud-free pixels was extracted based on the Sentinel-2 cloud mask. For model training, only cloud-free pixels were considered. Based on the available cloud free pixels, maximum of 10000 pixels were selected for model development. The pixels were further divided into training data (80 %) and testing data (20 %). NDVI available for cloud free pixels was also used as a ground truth for model validation. The LR, RFR and SVR models were evaluated based on R^2 and RMSE values. Model with lowest RMSE value was considered as the best. The best model was used to predict NDVI of cloudy pixels in Sentinel-2. Final cloud free NDVI image was created using mosaic operation between the sentinel-2 NDVI for non-cloudy pixels and the predicted NDVI for cloudy pixels.

3. RESULTS AND DISCUSSION

For analysis, we have extracted the data based on sentinel-2 tile coordinates. A tile-wise model has been developed for cloud free NDVI estimation. We identified set of best LR, RFR and SVR model for each tile and date. Table 2 and 3 shows the performance of the LR, RFR, and SVR models for various tiles on different dates. RFR mostly performed best compared to LR and SVR. The RFR and SVR showed highest R^2 of 0.73 and RMSE of 0.12. For some instances due to limited availability of cloud free observations regression models failed to generalize yielding low R^2 and high RMSE. There is a need for model fine tuning and field level validation of obtained insights.

We performed visual analysis of NDVI time series for all fields. Tile level NDVI estimates were used to extract field level estimated NDVI. Finally, the time-series graph was prepared for each field using the mean of original NDVI (with cloud), predicted NDVI (cloud-free), and fusion NDVI (using MODIS and Sentinel-2). Figure 5, 6, 7 and 8 shows the time series graph for fusion approach (Red), proposed machine learning approach (Blue) and original NDVI (Green). In all figures, black circle depicts the drop in NDVI due to cloud cover and not following the standard crop growth trend. Also, blue rectangle shows
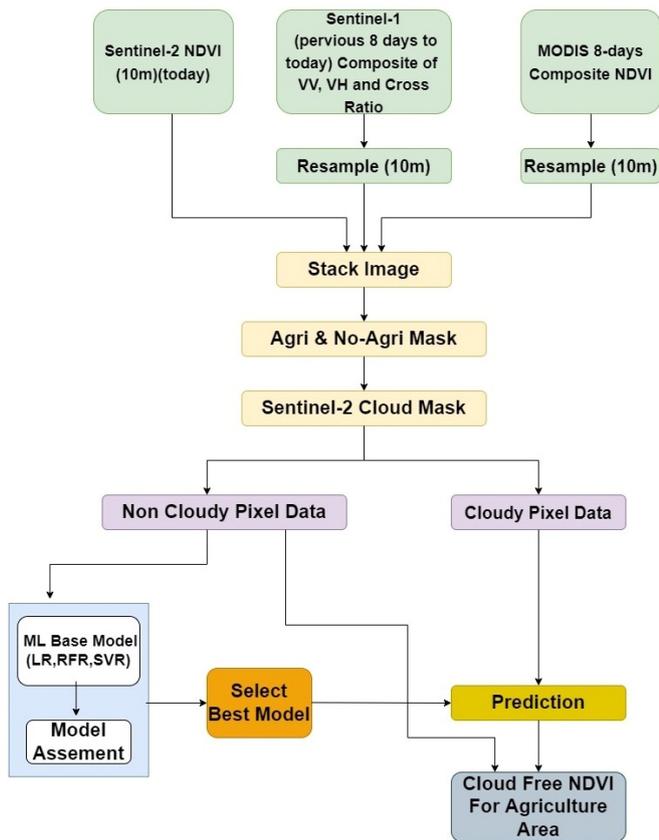
Figure 4. Overall analysis approach

Figure 5. Comparison between actual, estimated, and fused NDVI for field 1

Figure 6. Comparison between actual, estimated, and fused NDVI for field 2

very high value of MODIS-Sentinel-2 fused NDVI. Machine learning approach output at field level showed good alignment with the crop growth trend (blue line in the graph). Time series graph show that impact of cloud is decreased after replacing the NDVI of cloudy pixel with predicted NDVI.

In figure 5, blue rectangle shows that NDVI from fusion approach is very high at (t timestamp) compared to (t + δt) and (t- δt) timestamp. Same trend is also visible in figure 6, 7 and 8. The coarse spatial resolution and value of MODIS NDVI at time t dominates the field level mean NDVI.

Figure 8 shows that from Sep. to 10 Oct., 2021 NDVI is continuously decreasing compared to NDVI of 30 Aug., 2021 due to dense cloud. For same time period fusion NDVI also shows uneven values of NDVI. This shows that for the cloudy pixel high or low MODIS NDVI dominates the fusion NDVI and results in uneven field level mean NDVI. However, for the same time period model based NDVI follows a smooth increasing or decreasing trend. This shows that proposed approach reduces the impact of dense cloud and dominance of coarse resolution MODIS NDVI.

4. SUMMARY AND CONCLUSIONS

The NDVI time series has been widely used for crop monitoring. However, cloud cover results in the noise and obscure further applications of NDVI time series. We proposed use of the machine learning model to reconstruct a high-quality cloud-free NDVI image for the agriculture area. Using the ML model, we established the relationship between Sentinel-2 NDVI and
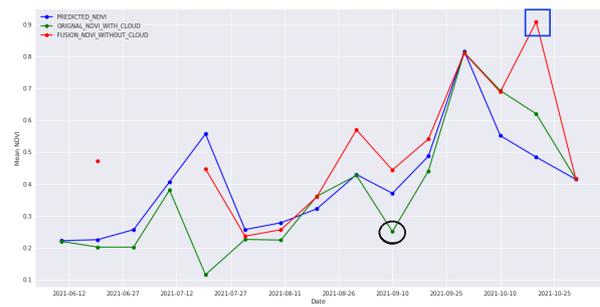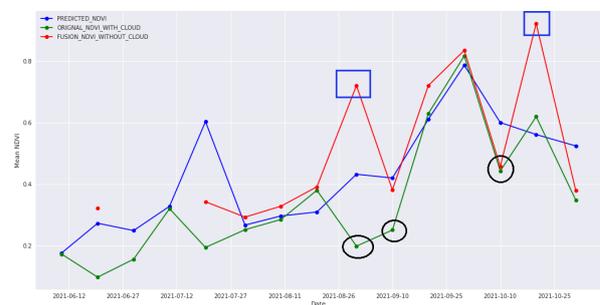

Figure 7. Comparison between actual, estimated, and fused NDVI for field 3

Figure 8. Comparison between actual, estimated, and fused NDVI for field 4

other satellite observations for non-cloudy pixels. The developed models were used to estimate the NDVI on the cloudy

Algorithm	LR	LR	LR	RFR	RFR	RFR	SVR	SVR	SVR
Date/Tile ID	T45QWG	T45QWF	T45QXF	T45QWG	T45QWF	T45QXF	T45QWG	T45QWF	T45QXF
2 Aug	0.06	0.41	0.31	0.17	0.50	0.41	0.16	0.50	0.39
12 Aug	0.08	0.31	0.08	0.08	0.49	0.17	0.01	0.45	0.14
22 Aug	0.04	0.12	0.08	0.09	0.19	0.10	0.07	0.15	0.09
6 Sep	0.04	0.01	0.05	0.03	0.02	0.07	0.02	0.01	0.07
16 Sep	0.05	0.04	0.07	0.10	0.08	0.15	0.09	0.06	0.13
26 Sep	0.06	0.08	0.07	0.10	0.13	0.10	0.08	0.08	0.08
6 Oct	0.11	0.52	0.20	0.19	0.57	0.26	0.14	0.56	0.26
11 Oct	-	0.65	0.36	-	0.73	0.37	-	0.73	0.36
16 Oct	0.19	0.22	0.19	0.20	0.29	0.24	0.18	0.25	0.22

Table 2. R^2 for all models (LR-Linear Regression, RFR-Random Forest Regression, SVR-Support Vector Regression)

Algorithm	LR	LR	LR	RFR	RFR	RFR	SVR	SVR	SVR
Date/Tile ID	T45QWG	T45QWF	T45QXF	T45QWG	T45QWF	T45QXF	T45QWG	T45QWF	T45QXF
2 Aug	0.02	0.18	0.16	0.02	0.17	0.15	0.02	0.17	0.15
12 Aug	0.04	0.14	0.19	0.04	0.12	0.18	0.04	0.12	0.19
22 Aug	0.20	0.23	0.22	0.19	0.22	0.22	0.20	0.23	0.22
6 Sep	0.16	0.16	0.20	0.16	0.16	0.19	0.16	0.16	0.20
16 Sep	0.26	0.17	0.20	0.26	0.16	0.19	0.26	0.17	0.19
26 Sep	0.14	0.16	0.20	0.14	0.16	0.20	0.14	0.16	0.20
6 Oct	0.24	0.23	0.20	0.23	0.22	0.19	0.24	0.22	0.19
11 Oct	-	0.22	0.12	-	0.19	0.12	-	0.19	0.12
16 Oct	0.12	0.13	0.17	0.12	0.12	0.17	0.12	0.13	0.17

Table 3. RMSE for All Models (LR-Linear Regression, RFR-Random Forest Regression, SVR-Support Vector Regression)

pixels. Comparative assessment of various models showed that RFR performed best with lowest RMSE for most of the instances during the cropping season. Further, temporal comparison between estimated NDVI using proposed approach, NDVI generated from fusion and the original Sentinel-2 NDVI (with cloud) was carried. Results showed that, NDVI estimated using the proposed approach was able to capture the trend of NDVI for a typical crop growth season as compared to fusion NDVI and original Sentinel-2 NDVI.

5. FUTURE WORK

In the existing work, we have only considered backscatter from VV, VH and Cross Ration. In future we plan to consider the impact of incident angle and indices such as Normalized Ratio Procedure between Bands (NRPB) along with the backscatter values. We have also planned to collect the ground truth data from different agro-climatic zones of India to evaluate spatial scalability of the proposed approach.

REFERENCES

AWS, 2022. Registry of Open Data on AWS. https://registry.opendata.aws.

Baghdadi, N. N., El Hajj, M., Zribi, M., Fayad, I., 2015. Coupling SAR C-band and optical data for soil moisture and leaf area index retrieval over irrigated grasslands. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(3), 1229–1243.

ESA, 2022a. Sentinel-1. https://sentinel.esa.int/web/sentinel/missions/sentinel-1.

ESA, 2022b. Sentinel-2. https://sentinel.esa.int/web/sentinel/missions/sentinel-2.

Filgueiras, R., Mantovani, E. C., Althoff, D., Fernandes Filho, E. I., Cunha, F. F. d., 2019. Crop NDVI monitoring based on sentinel 1. *Remote Sensing*, 11(12), 1441.

Friedl, M. A., McIver, D. K., Hodges, J. C., Zhang, X. Y., Muchoney, D., Strahler, A. H., Woodcock, C. E., Gopal, S., Schneider, A., Cooper, A. et al., 2002. Global land cover mapping from MODIS: algorithms and early results. *Remote sensing of Environment*, 83(1-2), 287–302.

Holben, B. N., 1986. Characteristics of maximum-value composite images from temporal AVHRR data. *International journal of remote sensing*, 7(11), 1417–1434.

Huang, S., Ding, J., Zou, J., Liu, B., Zhang, J., Chen, W., 2019. Soil Moisture Retrieval Based on Sentinel-1 Imagery under Sparse Vegetation Coverage. *Sensors*, 19(3). https://www.mdpi.com/1424-8220/19/3/589.

Lange, M., Dechant, B., Rebmann, C., Vohland, M., Cuntz, M., Doktor, D., 2017. Validating MODIS and Sentinel-2 NDVI Products at a Temperate Deciduous Forest Site Using Two Independent Ground-Based Sensors. *Sensors*, 17(8). https://www.mdpi.com/1424-8220/17/8/1855.

Liao, L., Song, J., Wang, J., Xiao, Z., Wang, J., 2016. Bayesian method for building frequent Landsat-like NDVI datasets by integrating MODIS and Landsat NDVI. *Remote Sensing*, 8(6), 452.

Macelloni, G., Paloscia, S., Pampaloni, P., Marliani, F., Gai, M., 2001. The relationship between the backscattering coefficient and the biomass of narrow and broad leaf crops. *IEEE Transactions on Geoscience and Remote Sensing*, 39(4), 873-884.

Mazza, A., Sepe, P., Poggi, G., Scarpa, G., 2021. Cloud segmentation of sentinel-2 images using convolutional neural network with domain adaptation. *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*.

M.B. Son, J.H. Chung, Y. L., Kim, S., 2021. A Comparative Analysis of Vegetation and Agricultural Monitoring of Terra MODIS and Sentinel-2 NDVIs. *The Korean Society of Agricultural Engineers*, 63(6). https://www.researchgate.net/publication/357171794.

MOD09Q1, 2022. MODIS/Terra Surface Reflectance 8-Day L3 Global 250 m SIN Grid. https://lpdaac.usgs.gov/products/mod09q1v006/.

Veloso, A., Mermoz, S., Bouvet, A., Le Toan, T., Planells, M., Dejoux, J.-F., Ceschia, E., 2017. Understanding the temporal behavior of crops using sentinel-1 and sentinel-2-like data for Agricultural Applications. *Remote Sensing of Environment*, 199, 415–426.

Wu, M., Yang, C., Song, X., Hoffmann, W. C., Huang, W., Niu, Z., Wang, C., Li, W., Yu, B., 2018. Monitoring cotton root rot by synthetic Sentinel-2 NDVI time series using improved spatial and temporal data fusion. *Scientific reports*, 8(1), 1–12.

Wu, S.-t., Sader, S., 1987. Multipolarization SAR data for surface feature delineation and forest vegetation characterization. *IEEE Transactions on Geoscience and Remote Sensing*, GE-25(1), 67–76.

Zheng, B., Myint, S. W., Thenkabail, P. S., Aggarwal, R. M., 2015. A support vector machine to identify irrigated crop types using time-series Landsat NDVI data. *International Journal of Applied Earth Observation and Geoinformation*, 34, 103–112.