INVESTIGATING THE PERFORMANCE OF RANDOM FOREST AND SUPPORT VECTOR REGRESSION FOR ESTIMATION OF CLOUD-FREE NDVI USING SENTINEL-1 SAR DATA

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
(jayant.mohite, suryakant.sawant, ankur.pandit, srinivasu.p)@tcs.com

Commission TCIII, WG IVb

KEY WORDS: Cloud-Free NDVI, SAR, Random Forest Regression, Sentinel 1, Sentinel 2

ABSTRACT:

The current study focuses on the estimation of cloud-free Normalized Difference Vegetation Index (NDVI) using the Synthetic Aperture Radar (SAR) observations obtained from Sentinel-1 (A and B) sensor. South-West Summer Monsoon over the Indian sub-continent lasts for four months (mid-June to mid-October). During this time, optical remote sensing observations are affected by dense cloud cover. Therefore, there is a need for methodology to estimate state of vegetation during the cloud cover. The crops considered in this study are Paddy (Rice) from Punjab and Haryana, whereas Cotton, Turmeric, and Banana from Andhra Pradesh, India. We have considered, observations of Sentinel-1 and Sentinel-2 sensors with the same overpass day and non-cloudy pixels for each crop. We used Google Earth Engine to extract surface reflectance for the Sentinel-2 and Ground Range Detected (GRD) backscatter for Sentinel-1. The Red and NIR bands of Sentinel 2 were used to estimate NDVI. Sentinel-1 based VV, and VH backscatter was used for estimation of Normalized Ratio Procedure between Bands (NRPB). Regression analysis was performed by using NDVI as an independent variable, and VV, VH, NRPB, and radar incidence angle as dependant variables. We evaluated the performance of Linear regression with tuned Support Vector Regression (SVR) as well as tuned Random Forest Regression (RFR) using the independent data. Results showed that the RFR produced the lowest RMSE for all the crops in the study. The average RMSE using the RFR was 0.08, 0.09, 0.11, and 0.10 for Rice, Cotton, Banana, and Turmeric, respectively. Similarly, we have obtained R^2 values of 0.79, 0.76, 0.69, and 0.71 for the same crops using the RFR. A model with 80 trees produced the best results for Rice and Cotton, whereas the model with 90 trees produced the best results for Banana and Turmeric. Analysis with NDVI threshold of 0.25 showed improved R^2 and RMSE. We found that for grown crop canopy, SAR based NDVI estimates are reasonably matching with the optical NDVI. A good agreement was observed between the actual and estimated NDVI using the tuned RFR model.

1. INTRODUCTION AND STATE OF THE ART

Continuous regional crop mapping and monitoring is essential especially in countries like India to keep a track on spatio-temporal coverage of various crops. This information can be consumed by various stakeholders like the government for the planning of various import-export activities, agri-input companies for facilitation of various fertilizers/chemicals, farmers to get the status of their crop in real-time (Mohite et al. (2018)). Satellite based remote sensing sensors are being effectively used over the years for continuous crop mapping and monitoring. Such methods are always preferred over manual surveys due to efficiency in terms of time, accuracy, spatial coverage, etc. Space exploration agencies such as the Indian Space Research Organization and international agencies such as the National Aeronautics and Space Administration (NASA), European Space Agency (ESA) have launched multiple Optical (IRS, Landsat 5,7,8, MODIS Terra, Aqua, Sentinel 2) as well as Synthetic Aperture Radar (RISAT-1, Sentinel 1) satellites. These satellites are extensively being used for crop mapping and monitoring.

Optical satellites provide rich spectral information in multiple wavelength bands which offer advantages for various agriculture applications such as crop type identification (Mohite et al. (2018)), crop monitoring, crop loss assessment (Sawant et al. (2019)), yield estimation (Mohite et al. (2019)), etc. Various methods based on the vegetation indices have been proposed in the past for agricultural applications. The Normalized Difference Vegetation Index (NDVI) is one of the widely used vegetation index (Rouse et al. (1974)). NDVI is derived using the Red and Near Infrared (NIR) bands of optical satellites such as Sentinel-2, Landsat-8, MODIS Terra and Aqua, etc. However, loss of information due to the presence of clouds in the optical dataset restricts its utilization to its maximum extent. In India, *Kharif* season is the main cropping season which starts in mid-June with the onset of the Indian Summer Monsoon (ISM) and extends upto November. During this season Indian sub-continent is mostly covered with the dense clouds.

Numerous attempts have been made for the cloud removal and cloud induced gap filling in the optical data using the time-series information and information available in the neighborhood pixels (Roerink et al. (2000); Padhee and Dutta (2019); Adam et al. (2018)). Nonetheless, the cloud removal process is useful in the presence of thin clouds and can be performed effectively but such process can not be considered successful in the case of thick clouds. Also, these methods can not be very useful in India during the *Kharif* season (June-October) when there is thick cloud cover over most of the season. Alternatively, the Synthetic Aperture Radar (SAR) sensor can collect continuous data in cloudy conditions as well as during day/night. Hence, synergistic use

*Corresponding author

of optical and SAR sensor observations can generate the continuous stream of NDVI time-series for vegetation monitoring. Studies have attempted to estimate the NDVI using SAR observations (Capodici et al. (2013); Davidse (2015); Filgueiras et al. (2019); Mazza et al. (2018); Navarro et al. (2016); Vreugdenhil et al. (2018)). Capodici et al. (2013) have shown that temporal changes of HV backscatter acquired with off-nadir angle greater than 40 degree best correlates with variations in the vegetation index from optical data. The study has a dependency on historical optical and SAR observations. Frison et al. (2018) showed a strong relationship between Sentinel-1 backscatter and vegetation phenology derived from Landsat-8. Mazza et al. (2018) have developed a CNN based model to derive NDVI from SAR data. Filgueiras et al. (2019) established the regression-based relationship between Sentinel-1 SAR and NDVI from Sentinel-2 to derive the continuous cloudless NDVI for Soybean and Maize (Corn). The study was focused on adjacent fields from a small area. Limitations of the research studies are a) the dependency on data from optical sensors for model development, b) methods are limited to certain incidence angles, c) heterogeneity in the spatial and temporal resolution of the SAR and Optical observations and d) geographical coverage for the model development. The current study focuses on the estimation of cloud-free NDVI using the SAR observations obtained from Sentinel-1 sensor. The proposed method explores the Linear Regression (LR), Support Vector Regression (SVR) and Random Forest Regression (RFR) for estimation of NDVI using SAR observations. The study was conducted during the Kharif season of the year 2019 for two regions of India.

2. MATERIALS AND METHODS

2.1 Study Area

The analysis was performed over two Indian regions namely, Andhra Pradesh and Punjab-Haryana. The study regions are situated in India's southern and northern parts respectively. The crops considered in this study are Paddy from Punjab and Haryana state. Punjab and Haryana are one of the major paddy producing belt in India. Cotton, Turmeric and Banana crops considered from Andhra Pradesh, India. Figure 1 shows the two locations where the geotagged field data has been collected.

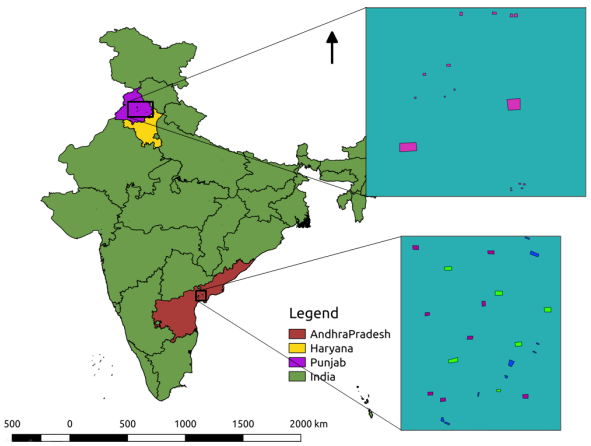

Figure 1. Study Area

2.2 Datasets Used

In this study we have used Sentinel-1 and Sentinel-2 satellite imagery, ground truth data collected from the field visits.

2.2.1 Sentinel-2 Data and Preprocessing ESA launched the constellation of optical satellite Sentinel-2 A and B which provides the earth observation in 10, 20 and 60 meter spatial resolution at five days repeat period (ESA (2020b)). Observations provided by Sentinel 2 are available in the 13 spectral bands mainly visible and NIR at 10 meters, red edge and SWIR at 20 meters, and atmospheric bands at 60 meters spatial resolution, respectively. For research purposes, Google Earth Engine cloud platform (Gorelick et al. (2017)) provides the collection of time-series Sentinel 2 Level-2A orthorectified atmospherically corrected surface reflectance data. In the present study, the data in Red and NIR bands was accessed from GEE to estimate the NDVI. Table 1 shows the location specific availability of Sentinel-2 data overlapping (or 1 day difference) with the Sentinel-1 overpass date. First number in the pair (1) shows the Sentinel-1 overpass date, however second number represents Sentinel-2 overpass date. Pixels with no cloud cover were considered for model development. NDVI threshold is used for obtaining the cloud-free pixels.

2.2.2 Sentinel-1 Data and Pre-processing Sentinel-1 satellite mission launched by ESA also has a constellation of two satellites 1-A and 1-B (ESA (2020a)). Data has been captured in dual-polarization by C-band Synthetic Aperture Radar. Satellite provides the observations at 5 meter in range and 20 meter in azimuth direction with 6 days repeat period. GEE (Gorelick et al. (2017)) has a collection of S1 Ground Range Detected (GRD) scenes, processed using the Sentinel-1 Toolbox to generate a calibrated, ortho-corrected product. The GRD product has been generated by pre-processing the scenes for thermal noise removal, radiometric calibration and terrain correction (Filipponi (2019)). Sentinel-1 C-band SAR has all weather, day-night capability hence all the observations available during the growing season are useful for the analysis. We have accessed backscatter information in VV, VH polarization along with local incidence angle. Normalized Ratio Procedure between Bands (NRPB) was estimated using VV and VH backscatter using equation 1 and used in the analysis as one of the variables.

$$NRPB = \frac{\sigma_{VH} - \sigma_{VV}}{\sigma_{VH} + \sigma_{VV}} \quad (1)$$

2.2.3 Ground truth data from field visits We have developed an android mobile application RuPS (Mohite et al. (2015)) for collection of field geo-coordinates and reporting various agricultural activities and events. For the current research, geo-tagged locations of the fields, crop cultivated on the field, its sowing or planting date and estimated harvest date were collected using the RuPS. Table 2 shows the number of plot boundaries collected for each crop and the total number of pixels associated with those crops.

2.3 Overall Approach

Each crop has a different crop season length therefore based on crop sowing and estimated harvest date concerning the region, we have considered NDVI and SAR data. For each crop and plot, we have identified the same satellite overpass dates and data with 1 day difference for Sentinel-1 and 2 and only that data was considered in the analysis. Data on all other dates were ignored to avoid noise and have the same reference. Plots were scattered all over the region to account for the regional variations of crop growth. The problem was devised as a regression analysis to establish the relationship between NDVI as an independent variable using

SN	Month	Cotton/Banana/Turmeric	Rice
1	June	17-18, 29-30	-
2	July	-	12, 16-17, 23-22, 28-27
3	Aug	4, 28-29	5-6, 12-11, 16, 21, 28
4	Sept	9-8	5, 10, 14-15, 21-20, 26-25
1	Oct	3, 27-28	-

Table 1. Month and crop-wise availability of Sentinel-1 and 2 observations

SN	Crop	No.of Fields	Total Pixels
1	Cotton	56	14988
2	Banana	58	16438
3	Turmeric	37	11352
4	Rice	78	24544

Table 2. Summary of crop-wise field observations

the dependant variables from SAR data. VV, VH backscatter, local incidence angle and NRPB were considered as the dependant variables. We tested the two scenarios (Table 3) using the Linear, Support Vector and Random Forest regression. Support vector and Random Forest regression models were tuned to get the best performance on validation data.

Scenario	Features Used	NDVI
Sc1	VV, VH, Incidence Angle, NRPB	All NDVI
Sc2	VV, VH, Incidence Angle, NRPB	NDVI greater than 0.25

Table 3. Various Scenarios Considered for Regression Analysis

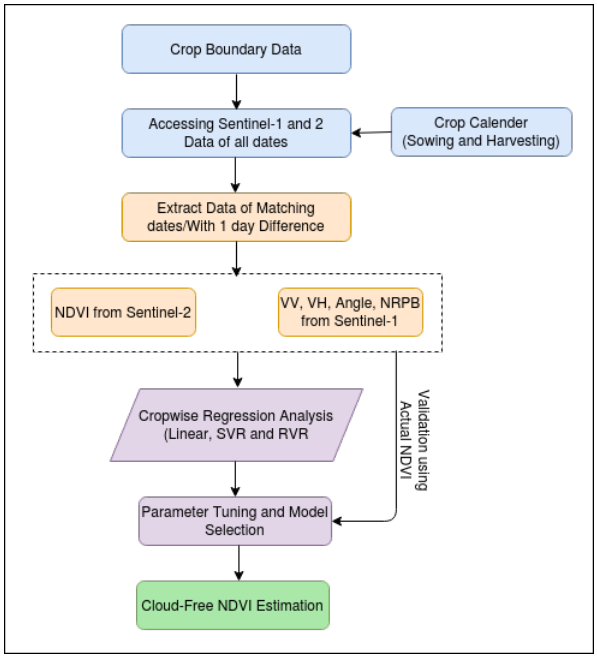

Figure 2. Overall Analysis Approach

3. RESULTS AND DISCUSSION

To carry out the regression analysis, we have extracted the data of NDVI, VV, VH, incidence angle and NRPB for all the pixels associated with individual crops. Crop-wise models are developed for NDVI estimation. For each crop, data was divided into 80% data for model training and 20% data for independent validation of the developed model. We evaluate the performance of Linear Regression (LR), Support Vector Regression (SVR) and Random Forest Regression (RFR). For models such as SVR, RFR there are hyperparameters which could be tuned to obtain the optimum performance. Hence we carried out 3 fold cross-validation on the training data to obtain the best parameters for SVR and RFR.

SVR is tuned for C at 0.1,1,10,100, Sigma at 1, 0.1, 0.01, 0.001 and type of kernel tried were Linear and Radial Basis Function. The model with best parameters (out of 32 models) was determined using 3 fold cross validation. Performance of the best model was evaluated using a 20% validation dataset. RMSE was used as a performance measure to decide the best model. Model with the lowest RMSE was considered as the best model. The same strategy was applied for RFR by tuning the parameters such as number of Trees. The number of trees were varied from 10 to 100 with an interval of 10. A total of 10 models were evaluated to find out the model with optimum trees. In the case of LR, we simply train the model on a random 80% dataset and tested of remaining 20% dataset. To avoid the bias in the random selection of dataset and noise, we ran the LR model 10 times and averaged the RMSE. Table 4 shows the performance of various models for all crops. The LR model shows the average RMSE and RSQ, however the performance of best models was shown for RFR and SVR. The obtained results shows that the RFR produced the lowest RMSE for all the crops in the study. The RMSE using the RFR was about 0.08, 0.09, 0.11, 0.12 for Rice, Cotton, Banana and Turmeric respectively. Similarly, we have obtained R^2 values of about 0.79, 0.76, 0.69 and 0.71 for the same crops using the RFR. The model with 80 trees produced best results for Rice and Cotton whereas, it is observed that the model with 90 trees produced best results for Banana and Turmeric. Also, we have observed that, SVR with RBF kernel was good for all crops when comparing the Linear vs RBF kernel of SVM. The performance of linear regression was poor among all models. Non-linear models such as SVR with RBF kernel and RFR performed well.

Overall modeling was repeated considering NDVI values greater than 0.25. This is to verify whether there is any influence of soil background on the overall model performance. Table 5 shows the performance of various models for the data with NDVI greater 0.25.

We can clearly see the improvements across all the models (both

SN	Crop	LR_rmse	LR_rsq	SVM_rmse	SVM_rsq	RF_rmse	RF_rsq
1	Rice	0.11	0.59	0.08	0.69	0.08	0.79
2	Cotton	0.17	0.51	0.14	0.64	0.09	0.76
3	Banana	0.24	0.37	0.15	0.58	0.11	0.69
4	Turmeric	0.16	0.5	0.11	0.67	0.10	0.71

Table 4. Performance of Various Models for Scenario 1 (Sc1)

SN	Crop	LR_rmse	LR_rsq	SVM_rmse	SVM_rsq	RF_rmse	RF_rsq
1	Rice	0.12	0.66	0.09	0.8	0.05	0.83
2	Cotton	0.17	0.49	0.12	0.67	0.06	0.78
3	Banana	0.21	0.54	0.15	0.62	0.10	0.71
4	Turmeric	0.18	0.52	0.1	0.72	0.09	0.77

Table 5. Performance of Various Models for Scenario 2 (Sc2)

linear as well non-linear) when considering the NDVI greater than 0.25. We observed decrease in the RMSE and improvement in R^2 values for all the crops using the RFR models. Such results show that, the soil background available during initial crop growth period was responsible for poor relationship between NDVI and SAR data.

3.1 Temporal analysis of few pixels

For continuous monitoring of vegetation, it is important to get the temporal and continuous data of NDVI. To check the temporal feasibility of the developed models, we applied the best models (Linear, SVR, RFR) on unknown fields for each crop. We did not consider this field for model development as well as for validation. For each crop, we have chosen one field and plotted the time-series of NDVI estimated using the best model and actual time-series of median NDVI for that field.

Figure 3 shows the time-series pattern for cotton where we can see the RFR model predicts the NDVI which closely matches the actual NDVI for almost all the dates. Also, there was a cloud during the month of July, August and September so there was a drop in the actual NDVI but RFR model predicted NDVI which closely follows the actual temporal NDVI pattern. In the case of Banana crop time-series (Figure 4), although the crop is present throughout the year, we have plotted the time-series between July to Dec 2019. The banana field was mostly affected by clouds during July- September. RFR model predicted NDVI which is closely following the pattern of actual NDVI wherever the actual cloud-free NDVI values are available. Figure 5 shows the time-series pattern for Turmeric. The field is covered by clouds towards the end of August and September. However, RFR predicted NDVI was in good agreement with actual NDVI and predicted the values at cloudy dates which followed the pattern of actual NDVI. Figure 6 shows the time-series of actual and predicted NDVI for rice. All the models were good to follow the actual NDVI however, RFR followed the actual NDVI pattern more accurately among all.

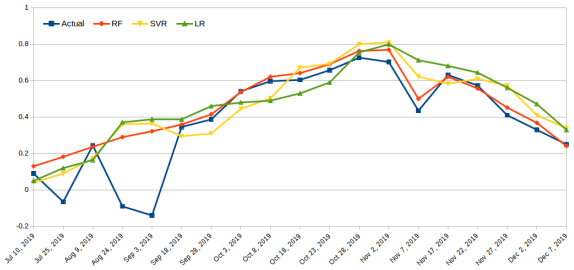

Figure 3. Actual vs Temporal NDVI predicted by Various Models for Cotton

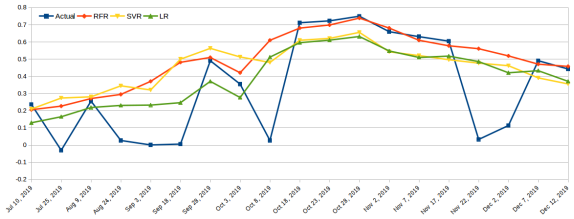

Figure 4. Actual vs Temporal NDVI predicted by Various Models for Banana

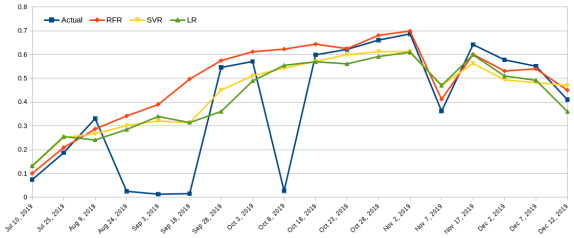

Figure 5. Actual vs Temporal NDVI predicted by Various Models for Turmeric

4. SUMMARY AND CONCLUSIONS

We have attempted to establish a relationship between NDVI derived from Sentinel-2 and Sentinel-1 based VV, VH backscatter, Local Incidence Angle and NRPB. The study has been carried out at two different locations considering the variety of crops during *Kharif* 2019. The crops considered in this study are Paddy from Punjab and Haryana whereas Cotton, Turmeric and Banana from

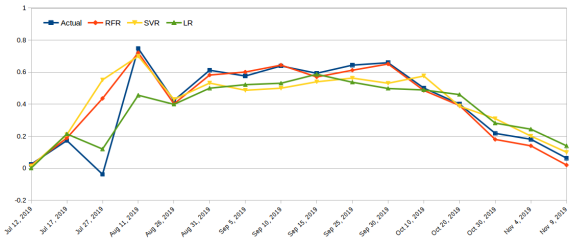

Figure 6. Actual vs Temporal NDVI predicted by Various Models for Rice

Andhra Pradesh, India. Regression analysis was carried out for NDVI estimation using SAR derived variables. We evaluated the performance of various linear (Linear Regression and SVR with Linear Kernel) as well as Non-Linear (SVR with RBF Kernel and RFR) models. Parameter tuning was done for SVR and RFR to get the best results. The Root Mean Square Error (RMSE) and R-Square (R2) were used as the performance indicators. The obtained results shows that the RFR produced the lowest RMSE for all the crops in the study. The average RMSE using the RFR was about 0.08 0.09, 0.11, 0.10 for Rice, Cotton, Banana, Turmeric, respectively. Similarly, we have obtained R^2 values of about 0.79, 0.76, 0.69 and 0.71 for the same crops using the RFR. The model with 80 trees produced best results for Rice and Cotton whereas, it is observed that the model with 90 trees produced best results for Banana and Turmeric. Further, we have considered data with NDVI greater than 0.25 and carried out a similar analysis. We observed a decrease in the RMSE and improvement in R^2 values for all the crops using the RFR models. We found that, RMSE was decreased to 0.05, 0.06, 0.10 and 0.09 for Rice, Cotton, Banana and Turmeric respectively. Moreover, R^2 was increased to 0.83, 0.78, 0.71 and 0.77 respectively for these crops. We found that the estimation of NDVI was good for high canopy density compared to crop in the early stages with soil background. Further, we have also plotted the time-series of actual vs estimated NDVI using all the models for various crops. NDVI predictions made by the RFR model were closely matching with actual NDVI for almost all temporal instances. This was followed by SVR and LR.

5. FUTURE WORK

As a part of future work, we plan to implement the method on every cloudy pixel with respective crop and generate the cloudless NDVI images. This will basically help us to carry out the comparison between the actual and generated NDVI images on a spatial level. In addition to this, we plan to collect more data on other crops cultivated during *Kharif* season and develop models for those crops.

ACKNOWLEDGEMENTS

We are very thankful to team members of Digital Farming Initiatives (DFI), Tata Consultancy Services Ltd. (TCS) for helping us with field observations. Also, we thank our organization, TCS for funding our research.

References

Adam, F., Mönks, M., Esch, T. and Datcu, M., 2018. Cloud removal in high resolution multispectral satellite imagery: Comparing three approaches. Proceedings of The 2nd International Electronic Conference on Remote Sensing.

Capodici, F., D'Urso, G. and Maltese, A., 2013. Investigating the relationship between x-band sar data from cosmo-skymed satellite and ndvi for lai detection. *Remote Sensing* 5(3), pp. 1389–1404.

Davidse, J., 2015. The relation between the ndvi and backscatter of sentinel-1 for sugarcane monitoring. Technical report, Internship Report GRS-70424, Wageningen University.

ESA, 2020a. Sentinel-1. https://sentinel.esa.int/web/sentinel/missions/sentinel-1.

ESA, 2020b. Sentinel-2. https://sentinel.esa.int/web/sentinel/missions/sentinel-2.

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

Filipponi, F., 2019. Sentinel-1 grd preprocessing workflow. In: *Multidisciplinary Digital Publishing Institute Proceedings*, Vol. 18number 1, p. 11.

Frison, P.-L., Fruneau, B., Kmiha, S., Soudani, K., Dufrêne, E., Le Toan, T., Koleck, T., Villard, L., Mougin, E. and Rudant, J.-P., 2018. Potential of sentinel-1 data for monitoring temperate mixed forest phenology. *Remote Sensing* 10(12), pp. 2049.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. and Moore, R., 2017. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*.

Mazza, A., Gargiulo, M., Scarpa, G. and Gaetano, R., 2018. Estimating the ndvi from sar by convolutional neural networks. IGARSS IEEE International Geoscience and Remote Sensing Symposium.

Mohite, J., Karale, Y., Gupta, P., Kulkarni, S., Jagyasi, B. and Zape, A., 2015. Rups: Rural participatory sensing with rewarding mechanisms for crop monitoring. Pervasive Computing and Communication Workshops (PerCom Workshops), 2015 IEEE International Conference on.

Mohite, J., Sawant, S., Kumar, A., Prajapati, M., Pusapati, S., Singh, D. and Pappula, S., 2018. Operational near real time rice area mapping using multi-temporal sentinel-1 sar observations. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*.

Mohite, J., Sawant, S., Sakkan, M., Shivalli, P., Kodimela, K. and Pappula, S., 2019. Spatialization of rice crop yield using sentinel-1 sar and oryza crop growth simulation model. 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics).

Navarro, A., Rolim, J., Miguel, I., Catalão, J., Silva, J., Painho, M. and Vekerdy, Z., 2016. Crop monitoring based on spot-5 take-5 and sentinel-1a data for the estimation of crop water requirements. *Remote Sensing* 8(6), pp. 525.

Padhee, S. K. and Dutta, S., 2019. Spatio-temporal reconstruction of modis ndvi by regional land surface phenology and harmonic analysis of time-series. GIScience & Remote Sensing.

Roerink, G., Menenti, M. and Verhoef, W., 2000. Reconstructing cloudfree ndvi composites using fourier analysis of time series. *International journal of remote sensing* 21(9), pp. 1911–1917.

Rouse, J., Haas, R., Schell, J. and Deering, D., 1974. Monitoring vegetation systems in the great plains with erts. NASA Special Publication 351.

Sawant, S., Mohite, J., Sakkan, M. and Pappula, S., 2019. Near real time crop loss estimation using remote sensing observations. 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics).

Vreugdenhil, M., Wagner, W., Bauer-Marschallinger, B., Pfeil, I., Teubner, I., Rüdiger, C. and Strauss, P., 2018. Sensitivity of sentinel-1 backscatter to vegetation dynamics: An austrian case study. *Remote Sensing* 10(9), pp. 1396.