High-resolution Remote Sensing to Identify Tree Plantations from Natural Forests and Agriculture in Southern India

By

Weiqi Zhou

A thesis submitted in partial fulfillment of the requirements for
the degree of Master of Science
(School for Environment and Sustainability)
at the University of Michigan
April 2019

Faculty advisors:

Assistant Professor Meha Jain, Chair

Associate Research Scientist Kathleen Bergen

Abstract

Tree plantations play an important role in tropical and subtropical countries, including providing economic benefits as well as ecosystem services. Over the past decade, the amount of land under agroforestry and plantations has increased rapidly in smallholder systems. One way to identify the extent of agroforestry and plantations is to use remote sensing, which can map land cover at large spatiotemporal scales. However, remote sensing classifiers often confuse agroforestry and plantations with forest cover; this is because these land cover classes often have similar spectral signatures, particularly high Near-infrared (NIR) reflectance and Normalized Difference Vegetation Index (NDVI) values. In addition, smallholder plantations in tropical areas are especially difficult to identify due to the small size of the plantation plots and the high cloud cover during the monsoon season. However, with the launch of Sentinel-2, which has additional spectral bands in the red edge, it may be possible to better classify these land cover types. Our study objective was to develop a general classification model using high spatial- and temporal-resolution Sentinel-1 and Sentinel-2 imagery to identify smallholder plantations using random forest algorithms. We developed this algorithm in southern India, in the four states with the largest plantation areas - Kerala, Karnataka, Tamil Nadu, and Andhra Pradesh. We find that using only Sentinel-1 imagery has lower classification accuracy (~70%) than using only Sentinel-2 imagery (~90%). Additionally, the combination of Sentinel-1 and Sentinel-2 can be used to map smallholder plantations with high accuracy: 93.91% (Kerala), 93.40% (Karnataka), 88.38% (Tamil Nadu), and 92.91% (Andhra Pradesh). Our results demonstrate the feasibility of systematically identifying tree plantations using high-resolution remote sensing data and machine learning algorithms.

Keywords*:* plantation, remote sensing, classification, random forest

Table of Contents

1 Introduction

The Third International Congress on Planted Forests has noted that the products and services provided by planted forests have become more diverse as the areas of planted forests have continuously increased (Payn et al., 2014). From 1990 to 2015, the annual growth rate of the global planted forest area was 1.84% (FAO, 2015). Among global planted forests, Asia has the largest proportion, accounting for about 50% of global planted forest area (FAO, 2015). In India, which is the 10th largest country in terms of forest area, plantation area and production rapidly increased from 2016 to 2018 (DAC&FW, 2018). The statistical report from the Department of Agriculture Cooperation & Farmers Welfare in India showed that from 2016 to the end of 2018, the area of forest plantation crops in India increased 4.11%, and the production increased 2.70% (DAC&FW, 2018).

Plantations play an important role in the economic landscape of tropical and subtropical countries. The profitable plantation tree crops, including bananas, cocoa, tea, coconuts, coffee, oil palm, and rubber, make plantations more economically valuable when compared with smallholder agricultural farms. In addition, there is an increase in the number of plantations due to climate variability. This is because plantation crops, such as rubber, tend to be more tolerant of extreme weather compared to traditional cash crops, like rice and wheat (Dong et al., 2013). From the socioeconomic perspective, plantations have the ability to increase farmers' income and provide job opportunities in rural areas (Obidzinski et al., 2012).

Though planted forest area is increasing according to census statistics, it is unclear which specific regions and farms are transitioning to growing plantations. One way to better understand the exact location and extent of plantation forests is to use satellite data to map plantations at large spatiotemporal scales. Doing so would provide fine-scale information that cannot be gathered from census datasets. Yet, in existing global land-cover products, plantations are often classified as forests. Yet, it is not always appropriate to include plantation areas in forest statistics. Plantations in the smallholder systems are essentially uniform agricultural systems that typically cannot replace natural ecosystems or provide the benefits of natural forest vegetation. However, vague definitions of forest in remote sensing analyses often conflate plantations with forests

(Tropek et al., 2014). For example, the definition of forest that Hansen et al. (2014) used when they created the global maps of 21st-century forest cover change was “all vegetation taller than 5m in height”. This simplistic definition does not contain the features that effectively distinguishing forests from plantations in the smallholder system, such as species diversity. By this definition, oil palm plantations and rubber plantations are also forests. This directly leads to inaccuracy of global forest coverage statistics. In addition, existing global agricultural products (e.g., GFSAD) group plantation areas with annual agriculture, which is also not appropriate.

It is critical to map tree plantations, non-tree agricultural crops, and natural forests as separate classes. In this study, the natural forest refers to the unplanted forests and the tree plantation plot refers to the plantation areas in the smallholder system that mixed with agriculture plots. To date, it has been challenging to map plantations in smallholder systems given two difficulties. First, plantations in smallholder systems typically grow in small areas that require high spatial-resolution satellites for detection. In smallholder systems, resource-poor farmers typically have an average farm area of smaller than 4 ha (Kouser et al., 2011). However, the spatial resolution of typically used satellite products, such as MODIS and Landsat, are often too coarse spatially and spectrally to map individual smallholder farms. The finest MODIS spatial resolution is 250 m which means that the area of a typical smallholder plantation plot is smaller than a single pixel. Landsat-8 has a spatial resolution of 30 m which allows more than 40 pixels to cover a single plantation plot of average size, but the spectral resolution for Landsat-8 may not be sufficient since it does not include any bands in the red-edge bandwidth, which is important for mapping vegetation types. Thus, in this study, we used Sentinel-2 imagery for classification. Sentinel-2 imagery has a spatial resolution of 10 m and it has 4 extra bands in the wavelength of the red edge.

The second difficulty in mapping forest plantations in smallholder systems is that the cloud cover in tropical areas like southern India is extremely high during the monsoon season (May to August) due to the high evapotranspiration rate. Optical sensors, therefore, cannot always clearly capture the reflectance of vegetation on the ground since these sensors are passive and cannot collect data through clouds. Therefore, Sentinel-1, which is an active radar satellite that can collect information through clouds, was included in the study.

The specific objectives of this study are:

1) Generate site-specific models to identify the amount of annual agriculture area, tree plantation, and natural forest using Sentinel-1 imagery, Sentinel-2 imagery, and both Sentinel-1 and Sentinel-2 imagery,

2) Compare the classification accuracy of using different remote sensing data sources: Sentinel-1 vs Sentinel-2 vs Sentinel-1 and Sentinel-2,

3) Develop a generalized model with the data from all study sites for southern India,

4) Compare the generalized models with the site-specific models based on the classification accuracy to see whether a generalizable model can map land cover as accurately as site-trained models,

5) Determine the most important spectral features for maximizing accuracy when identifying plantation areas, and

6) Create classification maps using the model with the highest classification accuracy.

2 Methods

2.1 Study area

The study region encompassed four states with the largest plantation area and production (DAC&FW, 2018) in India - Kerala, Karnataka, Tamil Nadu, and Andhra Pradesh (Figure 1). The major plantation crop species included areca nut, cashew nut, cocoa, and coconut (DAC&FW, 2018). In each state, an area of 50 x 50 kilometers that contained plantations, agriculture, and natural forests were selected as the study sites (Table 1).

Table 1. Study sites

Site	State	District	Location
1	Andhra Pradesh	Chittoor	13°18’N, 78°57’E
2	Karnataka	Chamarajanagar	12°01’N, 77°11’E
3	Kerala	Palakkad and Thrissur	10°34’N, 76°35’E
4	Tamil Nadu	Dindigul	10°18’N, 77°58’E

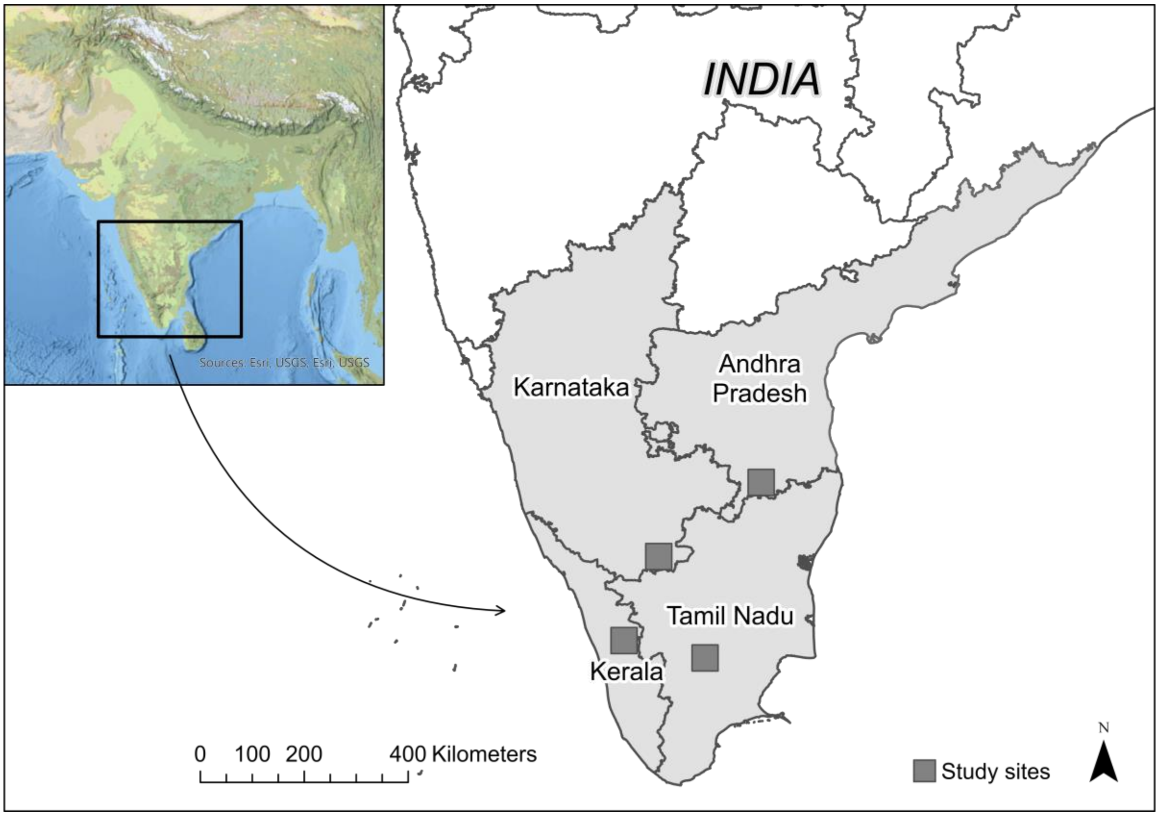

Figure 1. Map of the study sites in India's 4 states – Andhra Pradesh, Karnataka, Kerala, Tamil Nadu. **Figure was updated in February 2021*

2.2 Training and test data

While we did not have field data, we determined that the difference among plantation, forest and agriculture areas is distinct enough for accurate visual interpretation of high-spatial-resolution remote sensing images in Google Earth Pro. In my study, the plantation areas have a clear pattern of rows and columns of trees, agriculture areas have noticeable plot boundaries and the smoothest texture, and forest areas refer to all forests that have the largest area and the roughest texture (Figure 2).

At each study site, 200 agriculture polygons, 200 plantation polygons, and 200 natural forest polygons were digitized randomly using the most recent high-spatial-resolution remote sensing images in Google Earth Pro as the base ground-truth map. The image date of different locations in Google Earth Pro is heterogeneous, however, we constrained the time frame of the images to less than two years, from April 2016 to January 2018. In the classification process, a subset of 70% of the polygons from each class was randomly selected as training data, and the remaining 30% of the polygons were reserved for validation.

Figure 2. Example locations within high-resolution images from Google Earth Pro (Left: Forest polygon; Middle: Agriculture polygon; Right: Plantation polygon).

2.3 Data sources

2.3.1 Sentinel-2

The Sentinel-2 satellite images (Multispectral Instrument, Level-1C) for the four study sites were processed in Google Earth Engine (GEE). The spatial resolution for Sentinel-2 is 10 m and the temporal resolution is 5 days. The advantages of Sentinel-2 are that it provides images over global terrestrial surfaces and its Level-1C product is available on both the Sentinel Online website and GEE. In addition, the spatial resolution requirement of this study could be achieved by using the Sentinel-2 image (10 m). The mean area of training plantation polygons in this study is 1.5 ha which is the same area of 150 pixels on a Sentinel-2 image. This would provide sufficient data for building the classification models. In addition, Sentinel-2 has five spectral bands in the NIR which is the essential band when using remote sensing data map vegetation. Apart from the NIR band (B8), Sentinel-2 has three red edge bands (B5, B6, B7) that have shorter wavelengths than the NIR band and one red edge band (B8A) with a longer wavelength than the NIR band.

The remote sensing images for classification were converted from top of atmosphere reflection to surface reflection using Python 3 and GEE (Murphy, 2018). In this study, a total of 10 spectral bands and the normalized difference vegetation index (NDVI) were extracted in GEE (Table 2) and used as variables for classification.

Table 2. 10 spectral bands and NDVI from Google Earth Engine

Name	Resolution	Wavelength	Description
B2	10 meters	496.6nm (S2A) / 492.1nm (S2B)	Blue
B3	10 meters	560nm (S2A) / 559nm (S2B)	Green
B4	10 meters	664.5nm (S2A) / 665nm (S2B)	Red
B5	20 meters	703.9nm (S2A) / 703.8nm (S2B)	Red Edge 1
B6	20 meters	740.2nm (S2A) / 739.1nm (S2B)	Red Edge 2
B7	20 meters	782.5nm (S2A) / 779.7nm (S2B)	Red Edge 3
B8	10 meters	835.1nm (S2A) / 833nm (S2B)	NIR
B8A	20 meters	864.8nm (S2A) / 864nm (S2B)	Red Edge 4
B11	20 meters	1613.7nm (S2A) / 1610.4nm (S2B)	SWIR 1
B12	20 meters	2202.4nm (S2A) / 2185.7nm (S2B)	SWIR 2
NDVI	10 meters	(B8-B4) / (B8+B4)	normalized difference vegetation index

2.3.2 Sentinel 1

Sentinel-1 is another satellite that was funded by the European Union and carried out by the European Space Agency (ESA). It provides data from a dual-polarization C-band Synthetic Aperture Radar (SAR) instrument. Like Sentinel-2, it has a spatial resolution of 10 m and a temporal resolution of five days. Sentinel-1 is a radar sensor that uses radio waves. Unlike optical waves, radio waves can pass through clouds and detect the surface structure. This could provide more remote sensing data for classification, especially during the monsoon season. Sentinel-1 can transmit a signal in either horizontal (H) or vertical (V) polarization and then receives these signals in both H and V polarizations. In this study, the original VV band and VH band were used as two independent variables. In addition, the cross ratio (CR) index was calculated based on those two bands (Table 3). The monthly average of both VV and VH backscatter are extracted and then CR is calculated as VH/VV (Vreugdenhil et al., 2018).

2.4 Time window

Even though the spectral attributes of agriculture areas, plantations, and natural forests are similar, the seasonal phenology of agriculture and plantations may help in identifying a time window when the biggest difference in spectral signatures occurs. Among the agriculture, plantation, and forest, agriculture is the class that has the most obvious seasonal phenology. Thus, we focused our study on the winter (*Rabi*) season, since this is the main agricultural season that is not plagued by cloud cover, unlike the main monsoon (*Kharif*) season. The winter season in India starts after the summer monsoon season in November and ends in the early summer in April of the next year (Krishna Kumar et al., 2004; Figure 3). In this study, only the agriculture plots that were cultivated in winter 2018 are digitized as training agriculture polygons. This is because the uncropped plots could be masked out using a mask based on the maximum NDVI value in January 2018. The mask could remove all pixels that had an NDVI of 0.4 or lower, including non-vegetated areas and fallow agriculture areas (Jain et al., 2017).

Table 3. 2 bands and CR (cross ratio) from Google Earth Engine

Name	X Resolution	Y Resolution	Wavelength	Description
VV	10 meters	10 meters	5.405GHz	single co-polarization, vertical transmit/vertical receive
VH	10 meters	10 meters	5.405GHz	dual-band cross-polarization, vertical transmit/horizontal receive
CR	10 meters	10 meters	VH/VV	cross ratio

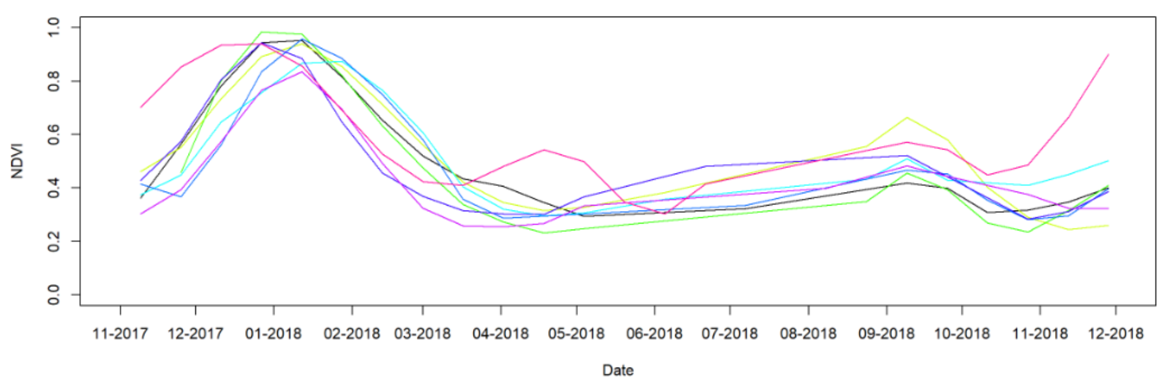

Figure 3. Time Series of ten random cultivated agriculture pixels in the study site in Kerala. The NVDI is splined with a degree of freedom of 10. The growing season is from November to April and the highest NDVI value occurred in January.

To remove the impacts of cloud cover on our spectral signatures, developed a cloud mask based on the 'QA60' band in Sentinel-2, which uses a bitmask to mark the opaque clouds (Bit 10) and cirrus clouds (Bit 11) (ESA, 2015); we masked out both clouds by selecting only the pixels that have a 'QA60' band equals to 0. To further remove the effect of clouds, six quality mosaic images were generated based on the monthly maximum NDVI from November 2017 to April 2018. Cloud pixels have moderate reflectance in every band, so cloudy pixels will not have higher NDVI values than vegetation. However, the cloud cover in November and December is so high that some pixels in the remote sensing image were covered with clouds during the whole month (Figure 4). Taking the phenology of agricultural crops and the effect of clouds into account, the time window used for this study is January 2018 to April 2018. Four quality mosaic images from January 2018 to April 2018 were used in the classification.

Figure 4. Quality mosaic images from November 2017 to April 2018 for the study site in Kerala. The red and white pixels represent cloudy pixels that cannot be used.

2.5 Random forest classification

In this study, the random forest algorithm (Breiman, 2001) was used for classification. This machine-learning algorithm is very popular in the study of remote sensing due to its high classification accuracy (Belgiu et al., 2016; Gislason et al., 2006; Pal, 2005). Compared to the classification and regression tree (CART) classifier, the random forest could randomly select the variables at each split that will minimize the correlation among the bands. In addition, the random forest algorithm is more computationally efficient than the more advanced machine learning algorithms like support vector machine (SVM) and artificial neural network (ANN) and may perform equally as well as these more computationally-intensive algorithms (Pal, 2005).

In this study, three site-specific random forest classification models were generated using the training data from each study site: 1) Using the 4 quality mosaic images from January 2018 to April 2018 with 10 spectral bands and NDVI from Sentinel-2 (Table 2); 2) Using the 4 monthly mean images from January 2018 to April 2018 with 2 bands and CR from Sentinel-1 (Table 3); 3) Using all the extracted data from Sentinel-1 and Sentinel-2 from January 2018 to April 2018. To generate the generalized classification models in southern India, three generalized models were created using the same variable as the 3 site-specific models above (i.e. Sentinel-2 only, Sentinel-1 only, both Sentinel-2 and Sentinel-1), but using all of the training polygons from the 4 study sites.

The output of the random forest classifier assigned each pixel in the image to either agriculture, plantations, or natural forests. In each random forest model, the number of variables considered at each split is the square root of the number of independent variables, and the number of trees in the random forest is 1000. Statistical analysis was performed in R version 3.5.2 (R Development Core Team, 2016) and the random forest analysis was conducted using the ‘randomForest’ package (Liaw et al., 2002).

3 Results

3.1 Classification accuracy assessment

3.1.1 Site-specific models

The 3 site-specific models were generated using the training data from each site and the same algorithm with the same parameters for each study site. Thus, the results from Kerala, the state in India with the largest plantation areas, are shown below as an example. The accuracy assessments for the other three states are in the Appendices.

The confusion matrix for the model using only the data from the Sentinel-1 satellite (Table 4) showed that the overall accuracy is 74.58%, which is not very high.

The producer accuracy is highest for the plantation which is more than 20% higher than that for forests. There is not much difference in the user accuracies across land-cover classes.

Table 4. Accuracy assessment of the site-specific model using only the data from Sentinel-1. The confusion matrix is for the validation pixels in the study site in Kerala.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1310	231	130	78.40%
Forest	124	966	290	70.00%
Agriculture	66	303	1080	74.53%
Producer Accuracy	87.33%	64.40%	72.00%	74.58%

The confusion matrix for the model using only data from the Sentinel-2 satellite (Table 5) showed that the overall accuracy is 93.16% which is much higher than the overall accuracy when using only the images from Sentinel-1. The producer accuracy is still the lowest for forests, but the difference is much smaller than the model using only the data from Sentinel-1 (Table 4).

Table 5. Accuracy assessment of the site-specific model using only the data from Sentinel-2. The confusion matrix is for the validation pixels in the study site in Kerala.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1418	143	72	86.83%
Forest	48	1346	0	96.56%
Agriculture	34	11	1428	96.95%
Producer Accuracy	94.53%	89.73%	95.20%	93.16%

The confusion matrix for the model using both Sentinel-1 and Sentinel-2 imagery (Table 6) showed that the overall accuracy is 93.91%, which is the highest among the three types of models. The producer accuracy is still the lowest for forests. The user accuracy is lowest for agriculture which is 10% lower than the user accuracy for plantation.

From the comparison of the overall accuracy of test pixels for each site-specific random forest model (Table 7), the models created with Sentinel-2 data have much higher overall accuracy than the model using only the data from Sentinel-1. The combination of Sentinel-1 and Sentinel-2 will increase the overall accuracy by about 1% - 3%. The overall accuracy has the highest improvement in Tamil Nadu when adding Sentinel-1 data to the random forest models.

Table 6. Accuracy assessment of the site-specific model using both the data from Sentinel-1 and Sentinel-2. The confusion matrix is for the validation pixels in the study site in Kerala.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1430	3	39	97.15%
Forest	0	1384	49	96.58%
Agriculture	70	113	1412	88.53%
Producer Accuracy	95.33%	92.27%	94.13%	93.91%

Table 7. Overall classification accuracy of each site-specific random forest model

	Sentinel-1	Sentinel-2	Sentinel-1 and Sentinel-2
Andhra Pradesh	68.42%	91.09%	92.91%
Karnataka	72.67%	92.47%	93.40%
Kerala	74.58%	93.16%	93.91%
Tamil Nadu	73.51%	85.71%	88.38%

3.1.2 Generalized models

Three generalized models were created using all the training polygons from all of the 4 study sites. The overall accuracy for these models ranged from 65.87% to 93.60% (Table 8). Generally, the generalized models have lower overall accuracy than the site-specific models by about 1%. However, the generalized model has the highest overall accuracy in Karnataka, even more than the site-specific models.

Table 8. Overall classification accuracy of the 3 generalized models for each study site

	Sentinel-1	Sentinel-2	Sentinel-1 and Sentinel-2
Andhra Pradesh	65.87%	90.51%	92.49%
Karnataka	72.09%	91.89%	93.60%
Kerala	71.82%	92.33%	93.00%
Tamil Nadu	73.09%	86.38%	86.98%

3.2 Important variables

The ‘randomForest’ package in R version 3.5.2 has another output that assesses the most important variables in the model. If not including a variable in the model will significantly decrease the classification accuracy, then this variable is the most important variable in the model. We identified the top 5 most important variables for the three generalized models (Table 9). The top 5 most important variables for each site-specific model created with both Sentinel-1 and Sentinel-2 data are in the Appendices. The results suggested that the VH band and the CR are important when using data from Sentinel-1. In addition, the SWIR 1 (B11), blue (B2), red edge 2 (B6), and red edge 3 (B7) are important variables for the random forest classification when using data from Sentinel-2.

Table 9. The top 5 most important variables in the three generalized models: 1) only the data from Sentinel-1, 2) only the data from Sentinel-2, 3) both the data from Sentinel-1 and Sentinel-2

Data Source	No. 1	No. 2	No. 3	No. 4	No. 5
Sentinel-1	VH_Apr	CR_Mar	CR_Jan	CR_Apr	VH_Mar
Sentinel-2	B11_Feb	B11_Jan	B2_Jan	B6_Jan	B7_Mar
Sentinel-1+Sentinel-2	VH_Feb	B11_Feb	VH_Jan	VH_Mar	B6_Jan

3.3 Classification map

The output of the random forest classifier is a classification map for each site (Figure 5; Appendix Figure A1, Figure A2, Figure A3). The classification in this study is only for vegetated areas, so an urban mask and a water mask were used when creating the final classification maps. The urban mask is the Global Human Built-up And Settlement Extent (HBASE) Dataset from Landsat-8 which has a spatial resolution of 30 m (Wang et al., 2017). The water mask is the Yearly Water Classification History which was developed by the European Commission's Joint Research Center and also has a spatial resolution of 30 m (Pekel et al., 2016).

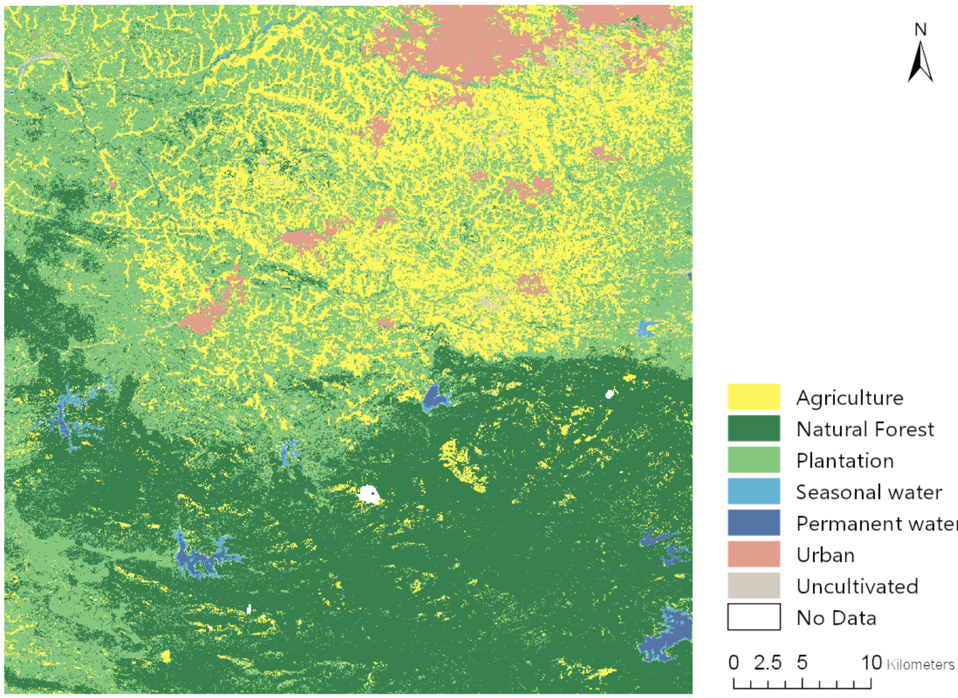

Figure 5. Classification map of the study site in Kerala when using both the data from Sentinel-1 and Sentinel-2.

4 Discussion

This study verified that it is feasible to map plantations in smallholder systems using multi-source high-resolution remote sensing imagery. Previously, Landsat-8 and PALSAR-2 were highly used remote sensing products for mapping plantations on the regional scale (Dong, 2013; Gao et al., 2016; Torbick et al., 2016). Considering that the spatial resolution for Landsat-8 and PALSAR-2 is 30 m and 25 m, the use of Sentinel imagery allows for the mapping of smaller field sizes since these data are available at a spatial resolution of 10 m. Especially when using both data from Sentinel-1 and Sentinal-2, the random forest classifier provided high classification accuracies (~90%). Sentinel-1 could avoid the problem of cloud cover and Sentinel-2 has multiple red edge bands that provide more information about vegetation to the classifier.

When separating tree plantations from agriculture and natural forest, the pattern of vegetation in the plantations is an important feature. Though the models with Sentinel-1 imagery had low classification accuracy (~70%), the information provided by Sentinel-1 was important as it increased the accuracy of models that just used Sentinel-2. In addition, when creating the generalized model with both Sentinel-1 data and Sentinel-2 data, one band from Sentinel-1 (VH_Feb) became the most important variable for the random forest classifier. This indicates that the texture, pattern, and roughness of the canopy which can be detected by Sentinel-1 is an essential feature when separating plantation, agriculture, and natural forest. This is likely because plantations in smallholder systems have a clear row and column pattern that is distinct from agriculture and natural forest.

This study also demonstrates that the models with Sentinel-1 and Sentinal-2 data are generalizable and thus likely to be useful over larger regions such as southern India. The generalized models have slightly lower overall accuracies than the site-specific models for each study site by about 2% (Table 7, Table 8). The comparable overall accuracy indicates that the phenology of plantation, agriculture, and forest among the four states in southern India are very similar. This may be because the climate condition and the species of plantation crops are similar across the four states. In all four states, there is a large amount of area planted to areca nut, cashew nut, cocoa, and coconut (DAC&FW, 2018). Thus, using remote sensing images offers an opportunity to scale up the identification of plantation areas without the need to get additional training data.

Future work will focus on how to further reduce the amount of training data needed when building models. In this study, the training polygons from all three classes were created through visual interpretation of very high spatial resolution image data. However, it is very time-consuming to visually interpret and digitize training and test polygons. Furthermore, the objective of this study is to only identify plantation areas. The other land-cover classes, such as agriculture and forests, are not the classes of interest, but these classes have to be collected in order to use the random forest classification algorithm. Thus, one-class classification algorithms could be an optimal method to identify only plantation areas while collecting less training data. The one-class classification algorithm only requires the training data from the class of interest (Deng et al., 2018). In this case, it could create the classifier with only digitized plantation polygons. These classifiers could significantly decrease the time needed for data collection and may achieve a similar classification accuracy as the models developed in this study. Popular one-class classification methods include one-class support vector machine (OCSVM), biased support vector machine (biased SVM), and maximum entropy (MaxEnt) (Mack, 2017).

To further assess the applicability of the generalized models to other study sites, different study sites other than the four considered in this study could be used for validation of the generalized model. Even though the training data set, and test data set in this study are independent datasets separated at the beginning of the study, there may be some spatial autocorrelation between the training data and test data, since they were mapped from the same study site. Therefore, validation that is done in a new study site could evaluate if the models are generalized enough for the four states in southern India.

5 Conclusion

In this study, the use of Sentinel-1 and Sentinel-2 imagery achieved a high classification accuracy when mapping plantations in smallholder systems in southern India. The models created with only the Sentinel-1 data had overall classification accuracies around 70%, and the models created with only the Sentinel-2 data had higher overall accuracies for around 90%. The accuracy comparison of different models indicates that the integration of radar data (Sentinel-1) and multi-spectral data (Sentinel-2) has benefits for mapping plantations on a regional scale. Generally, the models that added the Sentinel-1 data have 1% - 3% higher accuracies than the models using only the Sentinel-2. Specifically, the bands from Sentinel-1, which contain the surface texture attributes are the important predictors for the random forest classifier. This study also suggests that the generalized model for southern India has the same performance as the site-specific models. The generalized models with both the Sentinel-1 and Sentinel-2 data also have an overall accuracy of around 90%. The result of this study suggests that the synergistic use of high-resolution radar data with multi-spectral data has the potential to improve the accuracy of mapping the plantation areas in smallholder systems.

References

Belgiu, M., & Drăgu, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, pp. 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

DAC&FW. (2018). *Area and Production of Horticulture Crops - All India: 2016-17 to 2017-18*. Retrieved from http://agricoop.nic.in/sites/default/files/area_prod20161718.pdf

Deng, X., Li, W., Liu, X., Guo, Q., & Newsam, S. (2018). One-class remote sensing classification: One-class vs. Binary classifiers. *International Journal of Remote Sensing*, *39*(6), 1890–1910. https://doi.org/10.1080/01431161.2017.1416697

Dong, J., Xiao, X., Chen, B., Torbick, N., Jin, C., Zhang, G., & Biradar, C. (2013). Mapping deciduous rubber plantations through integration of PALSAR and multi-temporal Landsat imagery. *Remote Sensing of Environment*, *134*, 392–402. https://doi.org/10.1016/j.rse.2013.03.014

ESA. (2015). Sentinel-2 User Handbook. In *Sentinel-2 User Handbook*. https://doi.org/GMES-S1OP-EOPG-TN-13-0001

FAO. (2015). Global Forest Resources Assessment 2015. In *Forest Ecology and Management* (Vol. 352). Retrieved from http://www.fao.org/forestry/fra2005/en/

Gao, T., Zhu, J., Deng, S., Zheng, X., Zhang, J., Shang, G., & Huang, L. (2016). Timber production assessment of a plantation forest: An integrated framework with field-based inventory, multi-source remote sensing data and forest management history. *International Journal of Applied Earth Observation and Geoinformation*. https://doi.org/10.1016/j.jag.2016.06.004

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random forests for land cover classification. *Pattern Recognition Letters*, 294–300. https://doi.org/10.1016/j.patrec.2005.08.011

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., … Quantification. (2014). *High-Resolution Global Maps of 21st-Century Forest Cover Change*. *850*(2013), 850–854. https://doi.org/10.1126/science.1244693

Jain, M., Singh, B., Srivastava, A. A. K., Malik, R. K., McDonald, A. J., & Lobell, D. B.

(2017). Using satellite data to identify the causes of and potential solutions for yield gaps in India’s Wheat Belt. *Environmental Research Letters*. https://doi.org/10.1088/1748-9326/aa8228

Kouser, S., & Qaim, M. (2011). Impact of Bt cotton on pesticide poisoning in smallholder agriculture: A panel data analysis. *Ecological Economics*. https://doi.org/10.1016/j.ecolecon.2011.06.008

Krishna Kumar, K., Rupa Kumar, K., Ashrit, R. G., Deshpande, N. R., & Hansen, J. W. (2004). Climate impacts on Indian agriculture. *International Journal of Climatology*. https://doi.org/10.1002/joc.1081

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*. https://doi.org/10.1177/154405910408300516

Mack, B. M. (2017). *Applied One-Class Classification of Remote Sensing Data*.

Murphy, S. (2018). Atmospherically Corrected Time Series using Google Earth Engine. *GitHub Repository*. Retrieved from https://github.com/samsammurphy/ee-atmcorr-timeseries

Obidzinski, K., Andriani, R., Komarudin, H., & Andrianto, A. (2012). Environmental and Social Impacts of Oil Palm Plantations and their Implications for Biofuel Production in Indonesia. *Ecology and Society*, *17*(1).

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217–222. https://doi.org/10.1080/01431160412331269698

Payn, T., Carnus, J. M., Freer-Smith, P., Orazio, C., & Nabuurs, G. J. (2014). Third International Congress on planted forests: Planted forests on the globe - renewable resources for the future. *New Zealand Journal of Forestry Science*, *44*(Suppl 1), 1–2. https://doi.org/10.1186/1179-5395-44-S1-S1

Pekel, J.-F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature*. https://doi.org/10.1038/nature20584

R Development Core Team. (2016). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. https://doi.org/10.1007/978-3-540-74686-7

Torbick, N., Ledoux, L., Salas, W., & Zhao, M. (2016). Regional mapping of plantation extent using multisensor imagery. *Remote Sensing*. https://doi.org/10.3390/rs8030236

Tropek, R., Sedla ek, O., Beck, J., Keil, P., Musilova, Z., Imova, I., & Storch, D. (2014). Comment on "High-resolution global maps of 21st-century forest cover change." *Science*, *344*(6187), 981–981. https://doi.org/10.1126/science.1248753

Vreugdenhil, M., Wagner, W., Bauer-Marschallinger, B., Pfeil, I., Teubner, I., Rüdiger, C., & Strauss, P. (2018). Sensitivity of Sentinel-1 Backscatter to Vegetation Dynamics: An Austrian Case Study. *Remote Sensing*, *10*(9), 1396. https://doi.org/10.3390/rs10091396

Wang, P., Huang, C., Brown de Colstoun, E. C., Tilton, J. C., & Tan, B. (2017). *Global Human Built-up And Settlement Extent (HBASE) Dataset From Landsat*. Retrieved from https://doi.org/10.7927/H4DN434S

Appendices

Table A1. Accuracy assessment of the site-specific model using only the data from Sentinel-1. The confusion matrix is for the validation pixels in the study site in Karnataka.

Reference Data / Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1065	158	277	71.00%
Forest	115	1156	229	77.07%
Agriculture	206	245	1049	69.93%
Producer Accuracy	76.84%	74.15%	67.46%	72.67%

Table A2. Accuracy assessment of the site-specific model using only the data from Sentinel-2. The confusion matrix is for the validation pixels in the study site in Karnataka.

Reference Data / Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1355	0	145	90.33%
Forest	29	1441	30	96.07%
Agriculture	130	5	1365	91.00%
Producer Accuracy	89.50%	99.65%	88.64%	92.47%

Table A3. Accuracy assessment of the site-specific model using both the data from Sentinel-1 and Sentinel-2. The confusion matrix is for the validation pixels in the study site in Karnataka.

Reference Data / Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1428	0	72	95.20%
Forest	31	1429	40	95.27%
Agriculture	142	3	1355	90.33%
Producer Accuracy	89.19%	99.79%	92.37%	93.60%

Table A4. Accuracy assessment of the site-specific model using only the data from Sentinel-1. The confusion matrix is for the validation pixels in the study site in Tamil Nadu.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	851	294	355	56.73%
Forest	53	1234	213	82.27%
Agriculture	151	126	1223	81.53%
Producer Accuracy	80.66%	74.61%	68.29%	73.51%

Table A5. Accuracy assessment of the site-specific model using only the data from Sentinel-2. The confusion matrix is for the validation pixels in the study site in Tamil Nadu.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1185	17	298	79.00%
Forest	24	1433	43	95.53%
Agriculture	189	72	1239	82.60%
Producer Accuracy	84.76%	94.15%	78.42%	85.71%

Table A6. Accuracy assessment of the site-specific model using both the data from Sentinel-1 and Sentinel-2. The confusion matrix is for the validation pixels in the study site in Tamil Nadu.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1156	47	297	77.07%
Forest	15	1475	10	98.33%
Agriculture	134	83	1283	85.53%
Producer Accuracy	88.58%	91.90%	80.69%	86.98%

Table A7. Accuracy assessment of the site-specific model using only the data from Sentinel-1. The confusion matrix is for the validation pixels in the study site in Andhra Pradesh.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1133	166	201	75.53%
Forest	136	972	392	64.80%
Agriculture	167	359	974	64.93%
Producer Accuracy	78.90%	64.93%	62.16%	68.42%

Table A8. Accuracy assessment of the site-specific model using only the data from Sentinel-2. The confusion matrix is for the validation pixels in the study site in Andhra Pradesh.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1326	12	162	88.40%
Forest	68	1399	33	93.27%
Agriculture	95	31	1374	91.60%
Producer Accuracy	89.05%	97.02%	87.57%	91.09%

Table A9. Accuracy assessment of the site-specific model using both the data from Sentinel-1 and Sentinel-2. The confusion matrix is for the validation pixels in the study site in Andhra Pradesh.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1331	0	169	88.73%
Forest	53	1385	62	92.33%
Agriculture	31	23	1446	96.40%
Producer Accuracy	94.06%	98.37%	86.23%	92.49%

Table A10. Confusion matrix of the study site in Kerala. The generalized model was created with only the data from Sentinel-1.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1137	182	181	75.80%
Forest	159	1024	317	68.27%
Agriculture	59	370	1071	71.40%
Producer Accuracy	83.91%	64.97%	68.26%	71.82%

Table A11. Confusion matrix of the study site in Karnataka. The generalized model was created with only the data from Sentinel-1.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1105	88	307	73.67%
Forest	160	1054	286	70.27%
Agriculture	229	186	1085	72.33%
Producer Accuracy	73.96%	79.37%	64.66%	72.09%

Table A12. Confusion matrix of the study site in Tamil Nadu. The generalized model was created with only the data from Sentinel-1.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	927	267	306	61.80%
Forest	100	1228	172	81.87%
Agriculture	245	121	1134	75.60%
Producer Accuracy	72.88%	75.99%	70.35%	73.09%

Table A13. Confusion matrix of the study site in Andhra Pradesh. The generalized model was created with only the data from Sentinel-1.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1066	268	166	71.07%
Forest	118	997	385	66.47%
Agriculture	212	387	901	60.07%
Producer Accuracy	76.36%	60.35%	62.05%	65.87%

Table A14. Confusion matrix of the study site in Kerala. The generalized model was created with only the data from Sentinel-2.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1418	31	51	94.53%
Forest	9	1378	113	91.87%
Agriculture	37	104	1359	90.60%
Producer Accuracy	96.86%	91.08%	89.23%	92.33%

Table A15. Confusion matrix of the study site in Karnataka. The generalized model was created with only the data from Sentinel-2.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1398	0	102	93.20%
Forest	42	1410	48	94.00%
Agriculture	159	10	1331	88.73%
Producer Accuracy	87.43%	99.30%	89.87%	91.98%

Table A16. Confusion matrix of the study site in Tamil Nadu. The generalized model was created with only the data from Sentinel-2.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1165	33	302	77.67%
Forest	34	1445	21	96.33%
Agriculture	132	91	1277	85.13%
Producer Accuracy	87.53%	92.10%	79.81%	86.38%

Table A17. Confusion matrix of the study site in Andhra Pradesh. The generalized model was created with only the data from Sentinel-2.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1297	8	195	86.47%
Forest	85	1347	68	89.80%
Agriculture	50	21	1429	95.27%
Producer Accuracy	90.57%	97.89%	84.46%	90.51%

Table A18. Confusion matrix of the study site in Kerala. The generalized model was created with both the data from Sentinel-1 and Sentinel-2.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1428	31	41	95.20%
Forest	9	1395	96	93.00%
Agriculture	37	101	1362	90.80%
Producer Accuracy	96.88%	91.36%	90.86%	93.00%

Table A19. Confusion matrix of the study site in Karnataka. The generalized model was created with both the data from Sentinel-1 and Sentinel-2.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1428	0	72	95.20%
Forest	31	1429	40	95.27%
Agriculture	142	3	1355	90.33%
Producer Accuracy	89.19%	99.79%	92.37%	93.60%

Table A20. Confusion matrix of the study site in Tamil Nadu. The generalized model was created with both the data from Sentinel-1 and Sentinel-2.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1156	47	297	77.07%
Forest	15	1475	10	98.33%
Agriculture	134	83	1283	85.53%
Producer Accuracy	88.58%	91.90%	80.69%	86.98%

Table A21. Confusion matrix of the study site in Andhra Pradesh. The generalized model was created with both the data from Sentinel-1 and Sentinel-2.

Reference Data Classified Image	Plantation	Forest	Agriculture	User Accuracy
Plantation	1331	0	169	88.73%
Forest	53	1385	62	92.33%
Agriculture	31	23	1446	96.40%
Producer Accuracy	94.06%	98.37%	86.23%	92.49%

Table A22. The top 5 most important variables in the site-specific models with both data from Sentinel-1 and Sentinel-2

	1	2	3	4	5
Andhra Pradesh	VH_Mar	VV_Mar	B5_Mar	VV_Apr	VH_Apr
Karnataka	VH_Mar	B6_Jan	VH_Jan	B7_Mar	B7_Jan
Kerala	B2_Jan	B7_Feb	NDVI_Apr	B3_Jan	B11_Mar
Tamil Nadu	VH_Apr	B6_Jan	VH_Mar	VH_Jan	VV_Jan

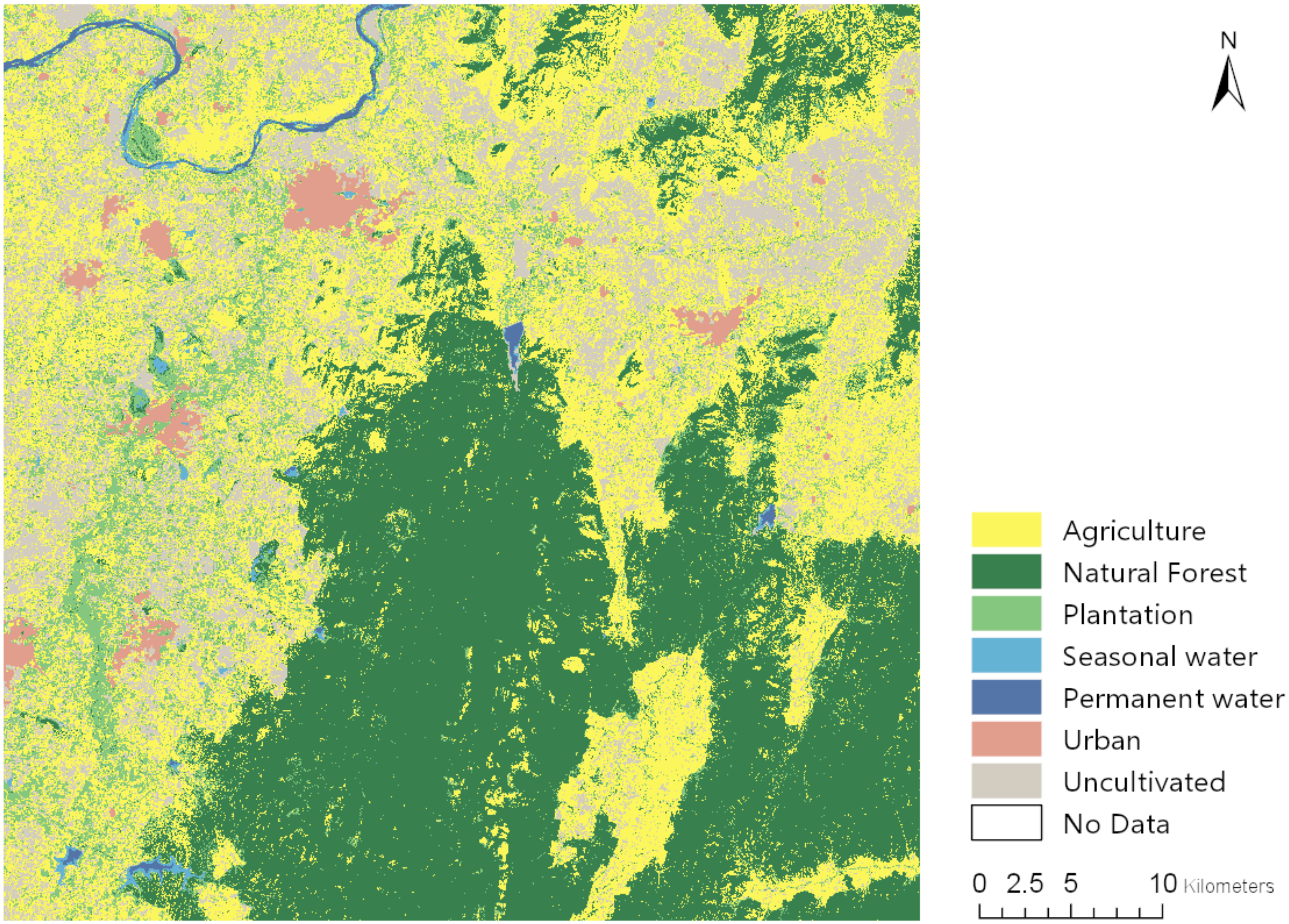

Figure A1. Classification map of the study site in Karnataka when using both the data from Sentinel-1 and Sentinel-2.

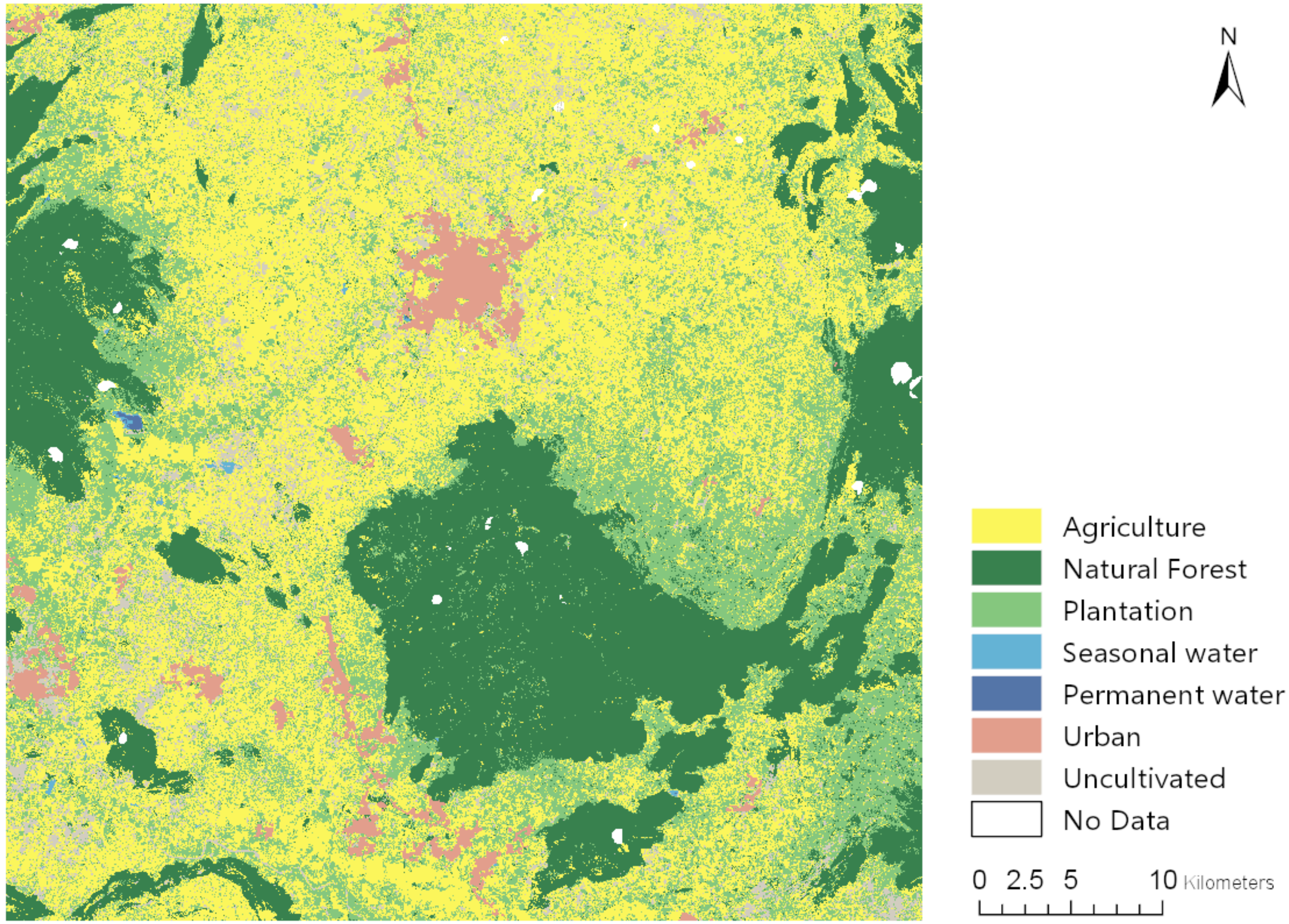

Figure A2. Classification map of the study site in Tamil Nadu when using both the data from Sentinel-1 and Sentinel-2.

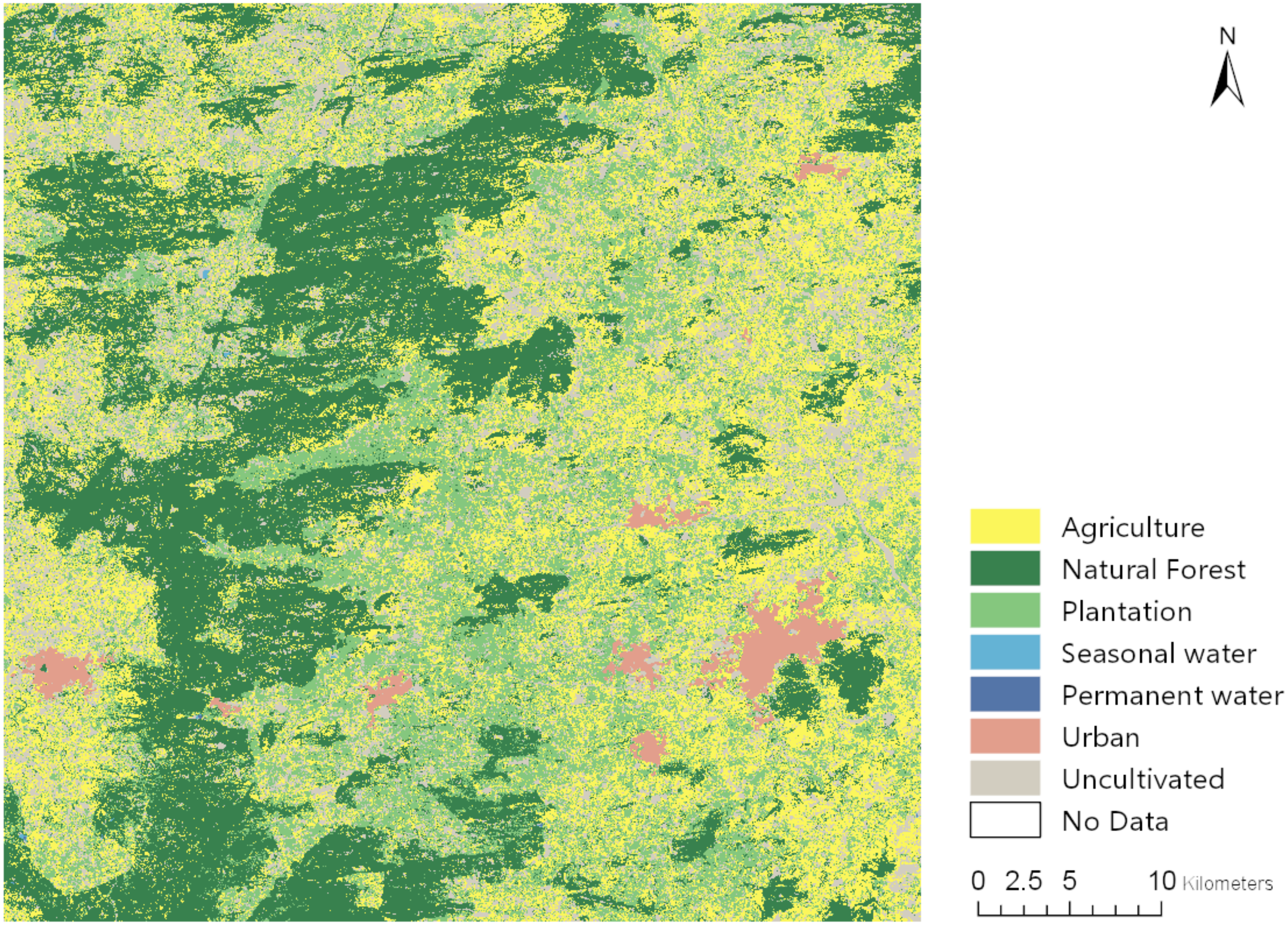

Figure A3. Classification map of the study site in Andhra Pradesh when using both the data from Sentinel-1 and Sentinel-2.