Time-series Cross-orbit Sentinel-1 Synthetic-Aperture Radar (SAR) Data for Mapping Paddy Extent: Case Study of Magelang District, Central Java

S Arjasakusuma, S S Kusuma, R Rafif, S Saringatin, and P Wicaksono

Department of Geographic Information Science, Faculty of Geography, Universitas Gadjah Mada, Yogyakarta

*E-mail: Sanjiwana.arjasakusuma@ugm.ac.id

Abstract. Dual-polarized (VV and VH) Sentinel-1 Synthetic-Aperture Radar (SAR) Ground Range Detected (GRD) data are available in 9-m spatial resolution and 12-day repeat orbit. A constellation of two satellites, Sentinel 1A and Sentinel 1B, capture these data with ascending and descending orbits, thus increasing the revisit time at the equator to every six days. Those specifications allow creating dense cross-orbit time-series data with a relatively high spatial resolution, beneficial for identifying land-covers and land-uses with unique temporal dynamics, such as paddies. This study was intended to assess the accuracy of time-series dual-polarized cross-orbit Sentinel 1A and 1B GRD data for mapping paddy extents. The monthly median value of these data was processed in Google Earth Engine and used as inputs in the paddy identification in Magelang District using bagging random forests (RF) and extreme gradient boosting (XGB) algorithms. Variables were ranked based on importance and selected using recursive feature elimination (RFE) and RF model to reduce the data dimensionality and understand the variable importance corresponding to a different month of the year. The resulting variable importance demonstrates better contributions of VV polarization and ascending orbit to the mapping model, and the producer’s and user’s accuracies achieved by RF classifier were 75% and 93.9%. For these reasons, an ascending (ASC) dataset provides better accuracy than its descending (DSC) counterpart and the combination of both (ASC+DSC). The user's accuracy of the paddy identified using the RF model with ascending Sentinel 1-data is 4% and 6% higher than the XGB models built using ASC and cross-orbit (ASC+DSC) datasets, respectively.

1. Introduction

Technological advancement of remote sensing in both the space segments, i.e., sensors and satellites, and the ground segments, i.e., data processing hardware, software, methods, and computation technology, creates the opportunity to perform big data analysis on the archive of remote sensing datasets. Nowadays, active and passive sensors offer many options for mapping applications using multiple datasets. For instance, Landsat-8, Sentinel-2, and Sentinel-1 imageries available in 10–60 m spatial resolutions with a revisit time varying from 3 to 16 days allow a more detailed mapping application, primarily if supported by excellent data computation ability, such as in Google Earth Engine [1].

Among the newer sensors, Sentinel-1 SAR is well-known for its capacity to produce remote sensing data for many mapping applications—including land subsidence identification, change detection, and land-cover mapping classification—utilizing the phase and magnitude of the complex radar signal or
the backscatter intensities of ground range detected (GRD) products with different polarizations, i.e., vertical-vertical (VV) and vertical-horizontal (VH). The versatility of radar in mapping applications in the tropics, such as Indonesia, lies in the radar signal ability to penetrate cloud layers and operate at night. Sentinel-1 data mainly flow in a constellation system (Sentinel 1A and Sentinel 1B) in ascending and descending orbits and have a revisit time of 3 days for each orbit or 6 days combined. Therefore, these data offer an abundant archive of imageries beneficial for characterizing the dynamics of land-use types, such as croplands or paddies, at much higher spatial resolution (± 9 m). More importantly, the images enable clear observation because they are not affected by atmospheric disturbances.

Paddy mapping usually relies upon dense time-series data derived from medium to coarse-spatial resolution optical sensors, such as MODIS [2;3]. However, mixed pixels with 250x250 m² spatial resolution [4;5] and cloud cover [3] affect the accuracy of MODIS in such mapping. Therefore, Sentinel-1 datasets are a perfect alternative for mapping paddies because of the high temporal and spatial resolutions.

Prior scholars, such as Arjasakusuma [6], Bazzi [7], and Lasko [8], have used time-series Sentinel-1 datasets for identifying paddies and their cropping patterns, but the utilization of cross-orbit Sentinel-1 data is still under-researched. A recent study by Sayedain [9] confirms that both ascending and descending orbits produce maps with high classification accuracy. However, whether or not the inclusion of both cross-orbit time-series data can increase the classification accuracy needs to be further studied, particularly because of the added data dimensionality.

Based on the identified problems above, this study sought to address several research questions, namely:

a. Can the cross-orbit Sentinel-1 data (ascending-descending orbit combination) improve the accuracy of paddy mapping?

b. Which variables representing different months or timing are important for mapping paddies using time-series Sentinel-1 data?

With the increased data dimensionality, machine learning algorithms like random forests (RF) and extreme gradient boosting (XGB) were used to build a model from time-series Sentinel-1 data derived by ascending and descending orbits. Before the modeling, the data were selected by deciding the optimal number of variables for mapping paddies and generating variable ranking using Recursive Feature Elimination (RFE) and random forests (RF).

2. Study Area
The study area stretches from 110.03° to 110.47°E and 7.7° to 7.2°S, an administrative part of Magelang District (Figure. 1). It has many land-cover types, including paddies and various croplands. Also, it comprises different landscapes, though the volcanic landforms developed at Merapi, Merbabu, Sindoro, and Sumbing Volcanoes are dominant. While paddies are mostly found on the lower slopes and alluvial plains of those mountains, other croplands and forests occupy the upper slopes.
3. Methods

3.1. Data Processing in Google Earth Engine and Rstudio

The first stage of the research was data pre-processing in Google Earth Engine (GEE) platform. GEE data was selected because it reduces data download time for the entire year and its processing can be conducted in clouds. Sentinel-1 image data used in this research had dual (VH+VV), VH, and VV polarization in ascending and descending orbits (see Figure. 2). The 2019 data is available in Google Earth Engine and has gone through pre-processing steps (i.e., thermal noise removal, radiometric calibration, and terrain correction) and converted to decibels.

The monthly composite data in one year, 2019, was created for each polarization data in each orbit by taking the median value of the images recorded in the same month to avoid any outliers. This data was then exported for further processing in the Rstudio program, to which analysis the VH and VV's backscatter values and an additional variable, Terrain Ruggedness Index (TRI), were inputted. TRI shows surface roughness [10] based on the Digital Elevation Model (DEM) data provided by the Indonesian Geospatial Agency (http://tides.big.go.id/DEMNAS/) in 8.1 m-pixel resolution. TRI is useful for correcting radar mistakes mainly caused by terrain shadow [6]. The DEM data were processed using the TRI module in QGIS 3.12 software. All inputs were then resampled to the same spatial resolution, i.e., the Sentinel-1 pixel’s spatial resolution (10 m), using Rstudio and raster package [11].
3.2. Variable Selection
In this research, variable selection by ranking was conducted in two steps, recursive feature elimination (RFE) and random forests. Each of the four used variables, i.e., VV-ASC, VH-ASC, VV-DSC, and VH-DSC, had 12-month data, thus creating 48 radar variables. Added with a terrain variable (i.e., TRI), the variables analyzed in this research amounted to 49. The RFE method works by iteratively running the random forest wrapper to calculate variable importance, removing variables with lower ranks than the predetermined optimal number, and performing variable ranking again from a reduced dataset [12].

This process was set to cease at the fifteenth variable, meaning that the iteration stops and the ranking returns after 15 important variables are found. The RFE-produced list of variables formed classification accuracy curves with different numbers of variables, and based on these curves, the optimal number of variables used in the classification was determined. The variable selection was performed on three Sentinel-1 datasets: ascending+TRI, descending+TRI, and ascending+descending+TRI.

Figure 2. Examples of paddies, as viewed using Sentinel-1 radar’s color composites (RGB in Jan, Feb, March 2020). (a) Ascending VV, (b) Ascending VH, (c) Descending VV, and (d) Descending VH.
3.3. Classification and Accuracy Assessment

In this study, the classification methods included bagging random forests and boosting strategy (a feature of extreme gradient boosting). The difference between bagging and boosting algorithms lies in the final decision tree construction. Bagging draws on a random subset of data to group tree classifiers, whereas boosting accommodates the loss function of previous classifiers for the ensemble. Hyperparameter tuning was conducted using a random search with ten times iteration and 5-fold cross-validation, thus creating 50 iterations in the search or tuning. The classification used the optimal variables identified from the previous RFE-RF analysis. Based on prior knowledge gained through Google Earth images (satellite view), training sites were identified for nine classes of land covers, namely 1.) Paddies, 2.) Croplands, 3.) Woody Vegetation, 4.) Bareland, 5.) Urban Areas, 6.) Grassland, 7.) Shrubland, 8.) Dryland farming, and 9.) Shadows. Afterward, the classes were merged to differentiate the area observed into paddies and non-paddies for accuracy assessment. The entire process of the variable selection and accuracy assessment was conducted using the caret package (Kuhn, 2012) in R, complemented by random forests [13] and Xgboost package [14].

The validation included the calculations of the producer's, user's, and overall accuracy from the confusion matrix using 2,273 visual inspection points, with Google Earth imagery as the validation source. Figure 3 summarizes the workflow of this study.

4. Results and Discussion

4.1. Monthly patterns of paddies and other land-cover types

Figure 4 presents the temporal patterns of paddies and other land-cover types by month. It shows that the temporal dynamics and the magnitude of the radar backscatter values differed between land-cover classes. For instance, the urban areas had the highest backscatter values among all classes. Meanwhile, croplands and woody vegetation had similar magnitude but different temporal patterns because of varying seasonality. In the study area, these two land-cover types shared similar vegetation structures, i.e., crop types, in which snake fruit plantations (Salacca zalacca) were abundant. Paddies had the lowest backscatter values among the classes because the combination of rice plant structure and water in the background causes this land cover to have lower radar signal attenuation.
4.2. Variable selection results
The variable selection using recursive feature elimination was carried out on the ascending (ASC), descending (DSC), and ascending-descending (ASC+DSC) datasets. Figure 5 shows the resulting accuracies of land-cover classifications with different numbers of variables used, and a flattening curve marks the optimal number of variables. This study selected 6, 8, and 10 variables for the DSC, ASC, and ASC+DSC datasets. As seen from the RFE analysis results, the total number of the inputted variables could be reduced by almost 75 to 80% to achieve similar accuracy to the classification that used the whole variables.
After the optimal number of variables was identified, the random forest model was built to produce variable ranks by which variables were selected for the classification. Variables occupying the highest rank until the lowest one within the predetermined optimal number were extracted, and Figure 6 shows the list of these variables. Also, it can be inferred that the TRI (terrain roughness index) had the highest importance value (highest rank), implying that the terrain configuration controls most of the land cover distribution in the study area. From the selected variables in Figure 6, it is clear that the important variables are related to the start and end of the year, which correspond to the rainy season in the study area. In other words, the data collected during these times of the year can clearly distinguish paddies from other land-covers appearing on satellite imagery because they are mostly at the flooding stage of rice cultivation. Variable ranking based on the combined (ASC+DSC) dataset indicates that ascending-based variables contribute higher to machine learning models than their descending counterparts. Of the nine radar variables, seven were from the ascending dataset. Based on the polarization, the three datasets tested in the variable selection showed that VH had a better contribution to the model, as compared with VV.

**Figure 6.** The selected variables and their ranking, as derived from the random forests model

### 4.3 Classification and accuracy results

The variables selected in Figure 6 were used as inputs to RF and XGB algorithms to generate eight land-cover classes. Afterward, maps with these classes were processed with a majority filter (window of size 3x3) and reclassified into two major categories: paddy and non-paddy (Figure 7). Although the reclassification results looked similar, particularly for the ASC and ASC+DSC datasets, some noises were apparent, and the paddies in certain areas were slightly overestimated (see the orange box in Figure 7). Misclassification into croplands (see the green box in Figure 7) was found in the descending dataset, especially from the random forest data.
Figure 7. XGB and RF classification results presenting the spatial distribution of paddies (yellow)

The classified paddies and non-paddies were validated using 2,273 visual inspection points identified from Google Earth images. Table 1 presents the validation results, i.e., classification accuracy, and it showed that the random forest model built using the ASC dataset had better performance even than other classifications with higher accuracy components. Although the overall accuracies of all models were above 80 %, the producer's accuracies were mostly under 76 %. This finding indicates that although map users can find accurate information (representing the field's condition) based on the map, the map itself has lower accuracy in mapping paddies. Table 1 also suggests that the random forest model using ASC variables produces a more accurate paddy map in the study area than other methods and datasets. ASC variables were proven to be the most reliable data for mapping paddies, which is attributable to their higher sensitivity to moisture change in this land cover, as shown in Figure. 2.

Table 1. The accuracy of paddy classification (user's and producer's accuracy) and overall map (overall accuracy)

| Classifications   | User's Accuracy (%) | Producer's Accuracy (%) | Overall Accuracy (%) |
|-------------------|---------------------|-------------------------|----------------------|
| XGB_DSC           | 81.95               | 68.11                   | 84.73                |
| RF_DSC            | 82.46               | 72.43                   | 86.01                |
| XGB_ASC           | 90.23               | 74.86                   | 89.18                |
| RF_ASC            | 93.94               | 75.41                   | 90.41                |
| XGB_COMBINED      | 88.11               | 75.14                   | 88.61                |
| RF_COMBINED       | 86.79               | 74.59                   | 88.03                |
This study showed that the ASC+DSC (combined) datasets produced maps with low accuracy, thus contrasting [9], which conclude that the cross-orbit dataset has high accuracy. However, it should be noted that the land cover classification in Sayedain [9] used single instead of time-series datasets. In that sense, using more data like dual-polarized Sentinel-1 in ASC+DSC orbits is most likely to give the classification algorithm a better opportunity to derive a more accurate mapping model. As for time-series analysis, the model accuracy tends to saturate at certain points where the inclusion of more temporal data will not give any significant accuracy improvement, as seen in Figure 5.

5. Conclusion
This study demonstrates the applicability of time-series radar data for mapping paddies using a random forest model. Based on the variable selection and ranking, the important variables correspond to the start and end of the year, i.e., during the rainy season in the study area. It indicates that the rice plant phenology during different planting stages (from flooding to post-harvesting) is significant in mapping paddies. This study has also found that combining ascending and descending orbit datasets may not necessarily increase the map accuracy. As evidence, compared with others, the ascending orbit data produce a map with higher accuracy and contributes more to the model. This is possible because the selected variables in the combined dataset are mostly from the ascending orbit data.

References
[1] 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 202 18-27
[2] Xiao X, Boles S, Liu J, Zhuang D, Frolking S, Li C, Salas W and Moore III B 2005 Mapping paddy rice agriculture in southern China using multi-temporal MODIS images Remote sensing of Environment 95 480-92
[3] Xiao X, Boles S, Frolking S, Li C, Babu J Y, Salas W and Moore III B 2006 Mapping paddy rice agriculture in South and Southeast Asia using multi-temporal MODIS images Remote sensing of Environment 100 95-113
[4] Wang J, Xiao X, Qin Y, Dong J, Zhang G, Kou W, Jin C, Zhou Y and Zhang Y 2015 Mapping paddy rice planting area in wheat-rice double-cropped areas through integration of Landsat-8 OLI, MODIS, and PALSAR images Scientific Reports 5 10088
[5] Sun H-s, Huang J-f, Huete A R, Peng D-l and Zhang F 2009 Mapping paddy rice with multi-date moderate-resolution imaging spectroradiometer (MODIS) data in China Journal of Zhejiang University 10 1509-22
[6] Arjasakusuma S, Swahyu Kusuma S, Astriviany N and Mahendra Krista W Accepted manuscript Mapping Paddy Field Extent and Temporal Pattern Variation in a Complex Terrain Area using Sentinel 1-Time Series Data: Case Study of Magelang District, Indonesia International Journal of Geo-informatics
[7] Bazzi H, Baghdadi N, El Hajj M, Zribi M, Courault D and Belhouchette H 2019 Mapping paddy rice using Sentinel-1 SAR time series in Camargue, France Remote Sensing 11 887
[8] Lasko K, Vadrevu K P, Tran V T and Justice C 2018 Mapping double and single crop paddy rice with Sentinel-1A at varying spatial scales and polarizations in Hanoi, Vietnam IEEE journal of selected topics in applied earth observation and remote sensing 11 498-512
[9] Sayedain S A, Maghsoudi Y and Eini-Zinab S 2020 Assessing the use of cross-orbit Sentinel-1 images in land cover classification International Journal of Remote Sensing 41 7801-19
[10] Riley S J, DeGloria S D and Elliot R 1999 Index that quantifies topographic heterogeneity Intermountain journal of sciences 5 23-7
[11] Hijmans R J, Van Etten J, Cheng J, Mattiuzzi M, Sumner M, Greenberg J A, Lamigueiro O P, Bevan A, Racine E B and Shortridge A 2015 Package ‘raster’.
[12] Darst B F, Malecki K C and Engelman C D 2018 Using recursive feature elimination in random
forest to account for correlated variables in high dimensional data *BMC Genetics* **19** 65

[13] Liaw A and Wiener M 2002 Classification and regression by randomForest *R news* **2** 18-22

[14] Chen T, He T, Benesty M, Khotilovich V and Tang Y J R p v- 2015 Xgboost: extreme gradient boosting.

[15] Kuhn M 2012 The caret package *R foundation for Statistical Computing*

**Acknowledgement**

This study was supported by the PDUPT-DIKTI research, titled “Multi-temporal Mapping Model of Crop Rotation Pattern and Paddy Productivity using the Integration of Big Remote Sensing Data and Machine Learning to Promote National Food Security (*Pengembangan Model Pemetaan Multi-waktu Pola Rotasi Tanam dan Produktifitas Lahan Sawah menggunakan Integrasi Big Data Citra Penginderaan Jauh dan Machine Learning untuk Mendukung Ketahanan Pangan Nasional*)”, with the contract number: 1678/UN1/DITLIT/DIT-LIT/PT/2020