Jheronimus
Academy
of Data Science

Professional Doctorate in Engineering

Data Science

Prepared by:

Ankit Majhi

Haripriya Sudhakaran

Loic Nguekam

Prepared for:

UNDP India

Permanently confidential

25 July 2022

Executive Summary

UNDP, has partnered with the Indian State of Telangana for Data in Climate Resilient Agriculture (DiCRA), a digital solution for strengthening food systems and food security by harnessing open-source technologies. For the DiCRA platform, UNDP wants a geospatial crop type map for Telangana that shows different crops growing in different farmlands in Telangana. India lacks open source, georeferenced survey data on farms that contains location information of farms and the crops grown on it. For generating crop type maps, such ground truth data is used.

Lack of such ground-truth, as in the case of Indian states, makes it challenging to not only develop crop type maps but also developing it for multiple crops. Therefore our project goal was to develop a proof of concept focussing on the identification of paddy in Telangana and to provide recommendations to UNDP on approaching this problem from here on.

Paddy's unique growth pattern allows it to be modelled using VH polarization value obtained from satellite data. We implemented a paddy mapping technique and generated a geospatial map for the same. We also performed an unsupervised clustering for identifying paddy fields using scientific indices associated with the growth pattern of paddy.

With our work we concluded that ground truth data is necessary for generating a crop type map. If data is sparse then it can be used for validation of the taken approaches, VH polarization method for paddy in this project. Sparse data can also aid in more accurate clustering of crops for unsupervised learning. If ground truth data is abundant then it may be utilised for building models with much better accuracy than aforementioned appraoches. Be it any approach ground truth data is necessary to atleast validate the obtained results.

In light of the lack of ground truth data and difficulty in obtaining it, we recommend that effort should be focussed on pattern analysis of one crop at a time by building models that utilise the unique growing patterns of each crop. For this a domain expert in agriculture and remote sensing can work with data scientists by sharing information on crop growth patterns and relevant remote sensing datafeatures that can be used to model these patterns. A data scientist can then leverage this domain knowledge for implementing data pipelines and building pattern recognition models that can perform crop identification.

Table of Contents

1. Management Introduction

1.0. Problem background

DiCRA is a digital solution geared towards strengthening food systems and food security by harnessing open-source technologies to attain the Sustainable Development Goals (SDGs). DiCRA can be used to identify farms and lands that are vulnerable to climate change and those that are resilient through remote sensing and pattern detection algorithms. This will help farmers mitigate the effects of climate change on their crops and livestock, boost the resilience of their livelihoods, and enable wider food security. Furthermore, it will also facilitate data analysis and insights on climate resilience, based on empirical inputs crowdsourced from hundreds of data scientists and citizen scientists on best-performing farms.

1.1. Goal specification and added value

UNDP India would like to classify crops across the state of Telangana, using open-source data. The part that we played here is specific to the classification of paddy. This is beneficial to the organization as it moves a step closer towards multiple crop classification. We initially set our goal for crop classification, but based on the literature review, the given time, and the available data, we scoped the goal down to paddy classification.

During this project, multiple phases were tackled, as listed below.

- Literature review: finding out which techniques had been attempted/implemented, and which data was necessary for that.
- Data collection
- Data preprocessing/annotation
- Implementation of unsupervised and supervised learning techniques

We were free to use open-source datasets that we deemed relevant to achieving the goal. We decided to stick to satellite imagery across districts in Telangana. Other datasets containing ground truth labels would be of utmost importance for validation of the results.

1.2. Strategy

We made use of VH polarization to identify paddy fields in districts in Telangana. We also used unsupervised classification for attempting to reach the same goal. We assumed that the transplanting time for paddy was the same across different districts in Telangana.

1.3. Results

We were able to generate, as a proof of concept, a spectral map with farm areas in Telangana where paddy grows.

1.4. Conclusions

Unsupervised learning was not very successful in identifying paddy fields. On the other hand, VH polarization was able to approximate the areas in which rice is grown.

1.5. Recommendations

Domain expertise and ground truth data are extremely valuable for good accuracy of the results.

2. Stakeholder Analysis

2.1. Client

The United Nations Development Programme (UNDP), partners with governments to develop sustainable agricultural programs to boost crop yields, reduce famine and permit farmers to profit from crop cultivation. UNDP has partnered with the Indian State of Telangana for Data in Climate Resilient Agriculture (DiCRA). DiCRA is a multi-stakeholder collaboration for data sharing involving governments, research organizations, citizens and data scientists across the world to strengthen climate resilience in agriculture. In this project the UNDP was represented by Parvathy Krishnan, who served as our stakeholder.

2.2. JADS (TU/e)

The Jheronimus Academy of Data Science (JADS) is a joint initiative of Eindhoven University of Technology, Tilburg University, the municipality of ‘s-Hertogenbosch, and the province of North Brabant. JADS provides a wide variety of educational programs, including the two-year long Professional Doctorate in Engineering (PDEng) program in Data Science. The PDEng program aims to train the current generation of data scientists and data entrepreneurs. This project is part of projects that are carried out during the first year of the program.

3. Data Value Chain

3.1. Problem background

3.1.1. The business case

It is certain that climate change, as well as the need for nutrition will have an impact on food production therefore sustainable agricultural practices are a necessity. For the government it is important to know what crop is being grown where, and how much of it is being produced. This can be facilitated by collecting crop production statistics, mapping soil productivity, monitoring farm activity and assessing crop damage due to storms and drought. The case at hand is a subpart of the abovementioned objectives, namely, to classify crops in the Telangana state of India. This is essentially done to fill in existing data gaps in where crops are grown.

3.1.2. How the project fits in with the corporate goal

UNDP India has set out multiple goals that they want to achieve by 2030. Some examples of these goals are to double the agricultural productivity and incomes of small-scale food producers, to ensure sustainable food production and implement resilient agricultural practices that increases productivity and production, helps maintain ecosystems and strengthens capacity for adaptation to climate change and other disasters. To move towards more sustainable production, it would help to know where crops are growing and how much of each crop is available in each region.

3.1.3. The company’s questions and objectives

The goal of the DiCRA project is to showcase how digital public goods can feed into policymaking at the subnational level. High resolution crop type maps have remained challenging to create in developing regions due to a lack of ground truth labels for model development and due to the small size of the farmlands in these countries. In this phase of DiCRA platform, UNDP India wants to showcase how digital public good data layers can be generated for the Telangana state in India which can feed into knowledge products and interactive decision-making tools for the state government. UNDP’s main objective is to classify crops in Telangana using open-source data.

3.2. Literature review in brief

Papers on crop classification make use of some amount of ground truth data for building a crop classification model or for validation purposes. There were multiple methods that were suggested and implemented across papers. We consider the following three approaches as the most prominent ones:

- **Transfer learning**: The Cropland Data Layer (CDL) is a crop-specific land cover data layer created annually for the continental United States. This can be used to create a crop classification model. This pre-trained model can be used to classify crops in other regions.

- **Unsupervised learning**: Identifying and using appropriate spectral bands and vegetation indices for crop clustering and using sparse ground truth data for validating the crop clusters.

- **Supervised learning:** Supervised learning approaches require annotated sample data that indicates the location of crops of interest on a spatial map. In the absence of such survey data, manual annotation and data generation is one way to approach it. The quality of the crop maps would depend on the accuracy of the manually annotated data.

Croplands in the USA are mostly large farm holdings while in India they are small holdings. Therefore, the feasibility of the model for India remains to be studied. We were not able to find open-source survey data for crops within the state of Telangana, therefore we decided to take the data generation and data annotation approach for building a supervised learning model for crop classification.

3.3. Project goals and objectives

The initial goal of crop classification seemed to be infeasible given the time limit and the available data sources, therefore in consultation with the stakeholder, the goal was scoped down to paddy classification in the districts of Telangana.

This provided us with two main objectives:

- Paddy classification using data annotation and classification/generalization

- Paddy classification using unsupervised clustering and visual inspection

3.4. Data source(s) and value

3.4.1. Sentinel –2A

The satellite carries a wide swath of high-resolution multispectral imagery with 13 spectral bands. It performs terrestrial observations in support of services such as forest monitoring, land cover changes detection, and natural disaster management. NDVI, NDWI and NIR indices that are calculated from different spectral bands of this satellite are used for the unsupervised clustering of crops. The satellite provides optical images of the area for the different spectral bands. We found an image in the paddy transplanting season for the month of June 2020. With Sentinel-2A data was not always available for required dates. Therefore, we proceeded to work with sample data that was available around the paddy transplanting season.

3.4.2. Sentinel –1

The SAR instrument on the Sentinel 1 satellite provides radar backscatter measurements influenced by the terrain structure and surface roughness. Generally, the more roughness or structure on the ground, the greater the backscatter. Rough surfaces will scatter the energy and return a significant amount back to the antenna resulting in a bright feature. Flat surfaces reflect the signal away resulting in a dark feature. Likewise, more structurally complex targets such as forests will appear brighter as signal interaction with the leaves, branches and trunks will result in a higher proportion of the signal being transmitted back to the sensor. VH polarization values are used from this satellite for classification of paddy. These values are obtained for the paddy transplantation season in Telangana, i.e., every fortnight from mid-June to mid-October.

3.5. Tasks performed

For performing paddy classification, we took two approaches:

3.5.1. Supervised Learning with Data Annotation

This approach needs ground truth data for building a supervised learning model for paddy classification. This approach can be broken down into three main tasks.

- Data annotation for sample pixels
- Generalization for larger area

3.5.2. Unsupervised Clustering with vegetation indices

The other approach is a form of unsupervised learning and consists of the tasks listed below.

- Data preprocessing
- Unsupervised Clustering

3.6. Data Science strategies

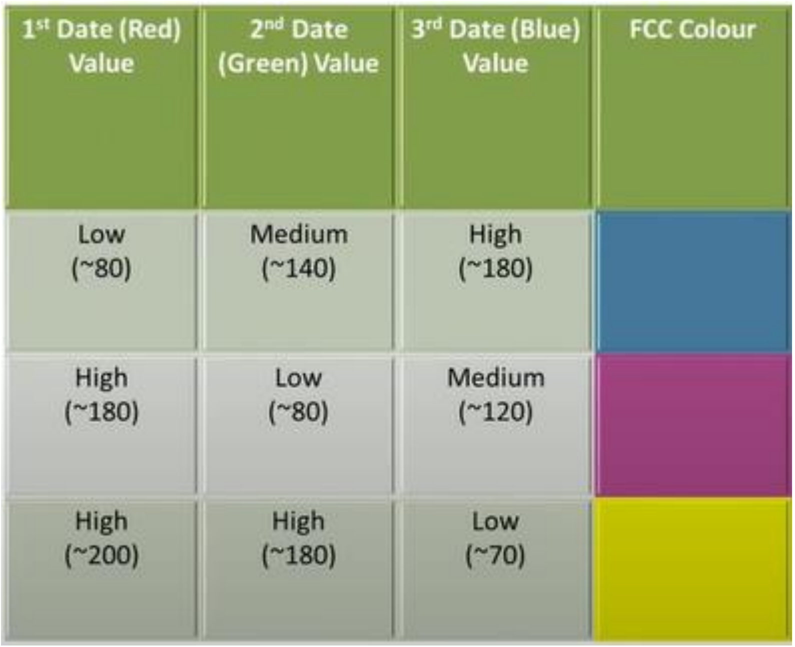
3.6.1. Supervised Learning with Data Annotation

In this approach, we assumed that the transplanting time is the same throughout the state of Telangana. Manual annotation of data was performed for rice growing areas in Telangana using the backscatter value called VH polarization from sentinel 2 satellite images. The idea here is to use the temporal backscatter profiles of the region of Telangana for consecutive fortnights for identifying the change in the amount of water on the land surface. If the backscatter in satellite images changes over time, so does the water level. This analysis is performed during the rice transplanting season from mid-June to mid-August. As rice is essentially flooded vegetation, the assumption is that the change in backscatter values indicated the presence of paddy.

The approach of data annotation and classification in google earth engine is outlined below.

1. Upload shape file for area of interest
2. Filter district of interest
3. Filter Sentinel data based on transplanting time span
4. Create temporal composite of data using VH polarization mean
5. Stack these bands
6. Scale the stacked bands
7. Create classes: rice, urban, water and other
8. Assign polygons/points to each class based on transplanting time
9. Run a Random Forest classifier

This data representation in this approach utilizes the additive RGB color model for representing and comparing the temporal backscatter values on a spatial map. The first row in figure 1, corresponding to blue indicates pixels on a spatial map where the previous fortnight had low backscatter values (therefore, more water) and the current fortnight has relatively higher backscatter (therefore less water). This is assumed to be indicative of flooded pieces of land that have now more vegetation in a span of two weeks. Thus, paddy fields can be identified in this manner, where the crop has been transplanted in the previous fortnight but now the vegetation continues to grow.



1st Date (Red) Value	2nd Date (Green) Value	3rd Date (Blue) Value	FCC Colour
Low (~80)	Medium (~140)	High (~180)	🟦
High (~180)	Low (~80)	Medium (~120)	🟪
High (~200)	High (~180)	Low (~70)	🟨

Figure 1: Additive RGB color model for representation of stacked VH polarization values on a spectral map

In the second row, corresponding to the magenta color, the high and low backscatter values indicate regions on the map where higher vegetation is followed by flooding in a span of two weeks. This indicates newly transplanted pieces of land. Next, we annotated sample pixels for paddy in these colored regions by drawing polygons for choosing the pixels that fall within the colored region.

We kept the Blue component of the RGB color model constant at a high level of backscatter, (VH polarization value for the first fortnight of October) and shifted the first two dates continuously to analyze the differences in water levels and vegetation. We then looked for areas with a very low initial backscatter level which continued to increase. These areas represent places that initially had a lot of water, which decreased over time. In other words, these areas were initially flooded, but as rice continued to grow there, the backscatter in the images increased. As this corresponds with the first row of figure 1, we looked for blue areas in figure 2. Similarly, we shifted the first and second date by two weeks and continued to look for pink areas in the satellite image. This process continued until there was no more pink to be found in the image, indicating that all areas in which rice was grown, had been found.

Other than rice, we considered three other classes, namely urban, water and other. Urban areas could be easily detected by zooming into certain cities in the map view. Water could be detected by the low backscatter values and thus the darker pixels. Finally, the pixels belonging to the class other were generally forested areas, barren land or mountains. For each of these classes, corresponding polygons consisting of multiple pixels were drawn. For instance, in figure 2, a blue polygon was drawn in the water (due to the low backscatter values). Each pixel in the polygon was then assigned the class water, and these pixels then became part of the training set. In other words, the training set consisted solely of pixels within each polygon that was manually drawn per class.

Finally, a random forest classifier was used to build a classification model using the annotated training dataset, and then generalize the classification towards the entire region of interest.

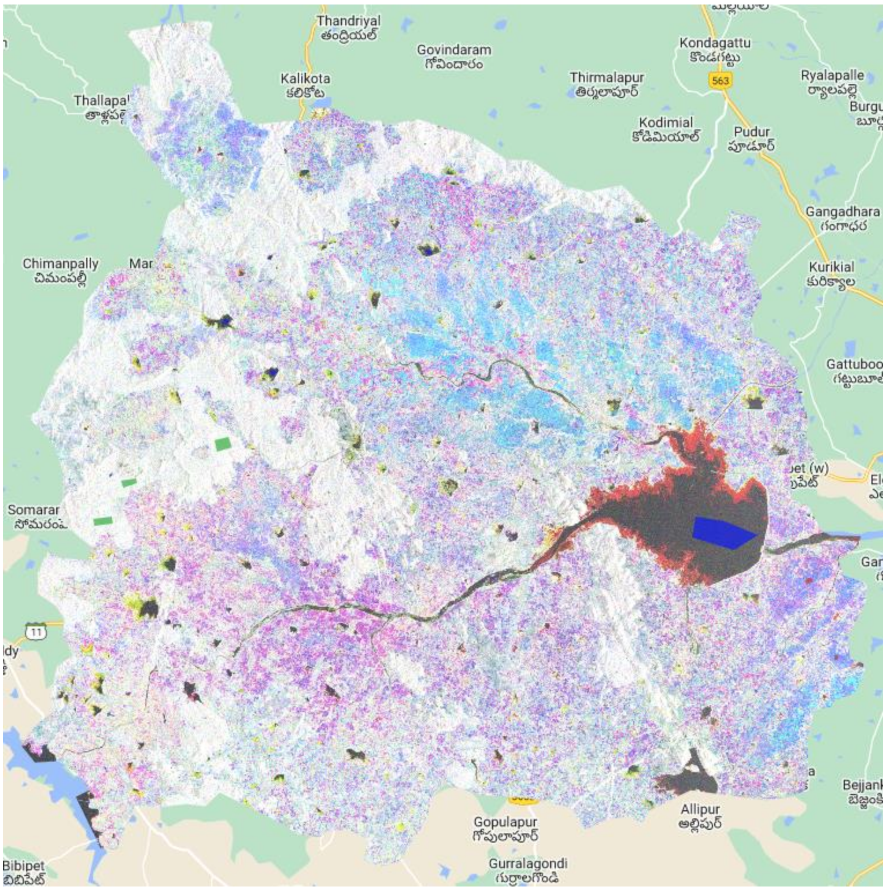


Figure 2: Spectral map showing the stacked VH polarization values for two consecutive fortnights

3.6.2. Unsupervised Clustering with vegetation indices

Normalized Difference Vegetation Index (NDVI) quantifies vegetation by measuring the difference between near-infrared (which vegetation strongly reflects) and red light (which vegetation absorbs). Normalized Difference Water Index (NDWI) is used to monitor changes related to water content in water bodies. As water bodies strongly absorb light in visible to infrared electromagnetic spectrum, NDWI uses green and near infrared bands to highlight water bodies. These spectral indices along with Near Infrared (NIR) band of Sentinel –2 satellite was used to build an unsupervised learning model for identifying areas influenced by these indices, therefore anticipating classification of paddy and non-paddy areas.

4. Results

4.1. Results of task: Supervised Learning with Data Annotation

In figure 3, it becomes clear that backscatter values for some classes are relatively constant while others are quite variable. The pink line for instant is almost always fixed at 80, which is quite a low backscatter value. This in turn indicates high levels of water and thus also represents the class water. The yellow and cyan lines are also relatively constant, but at much higher backscatter values, corresponding to urban and other classes, respectively. The other five classes all represent rice at different times. It is clear to see that the backscatter values initially significantly drop and slowly rise again. This indicates that the water level suddenly increases and proceeds to decrease over time. In other words, these places are flooded and later the vegetation (rice crop) rises above the water level.

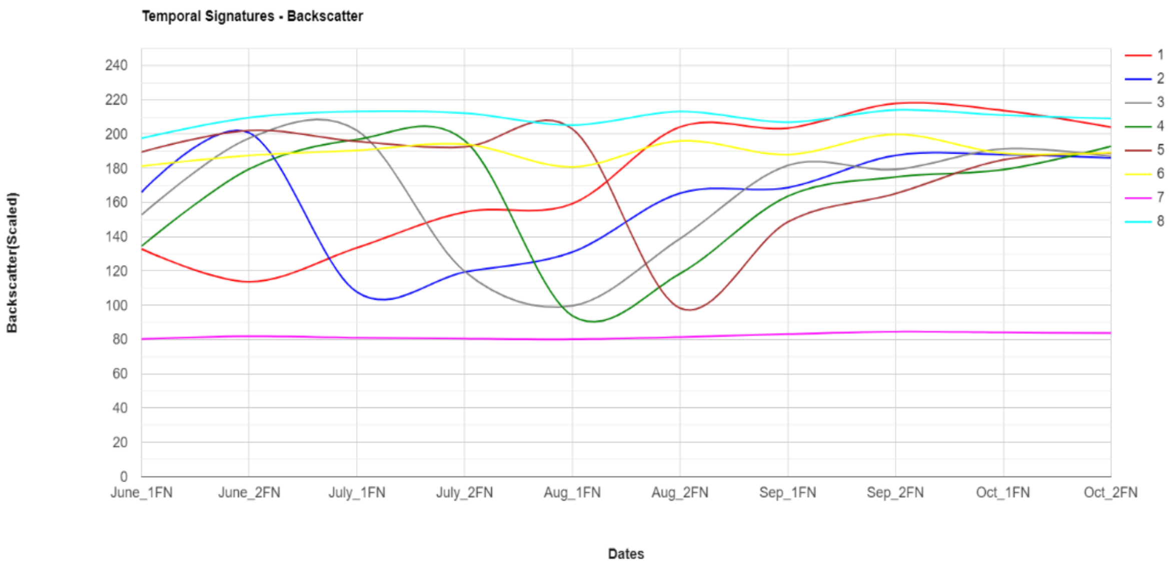


Figure 3: VH polarization values for the different classes for the different fortnights

Figure 3 shows the areas that were classified as rice in yellow, in the district Rajanna Sircilla, in the state of Telangana. As each pixel corresponds to 10m2, the total area of paddy could also be calculated, which was estimated to be 75689 hectares.

The Random Forest classifier ended up with a training accuracy of 99.53%. However, the labels as well as the polygons were created by us, thus accuracy cannot be relied upon as a metric of interest in this case.

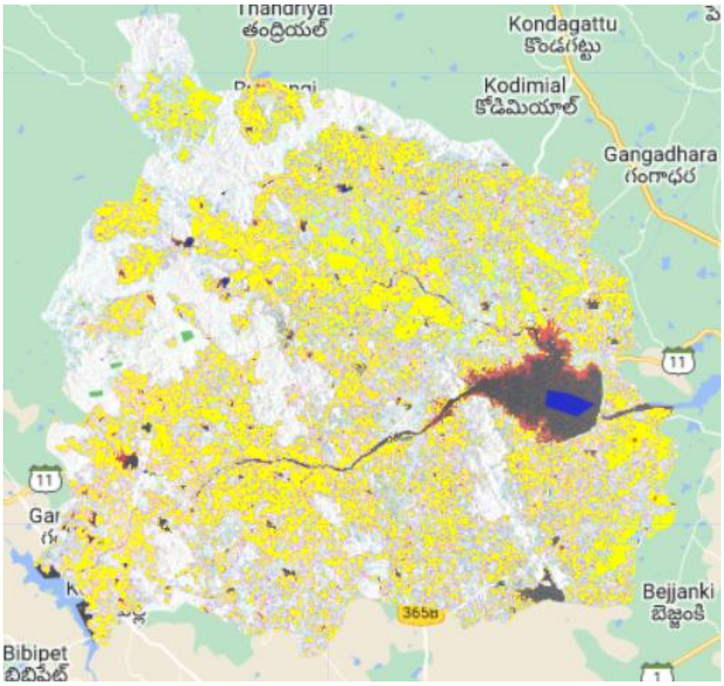
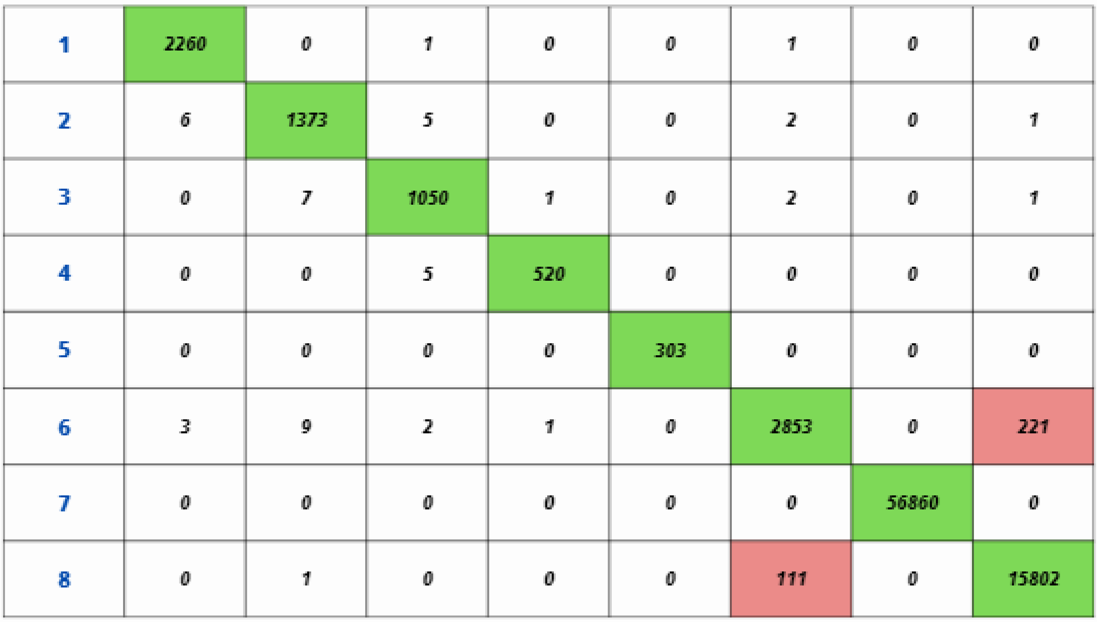


Figure 4: Yellow regions showing the area under paddy cultivation

The same can be said about the confusion matrix: since this is based on the training data which we have created ourselves, it does not mean it is representative for other areas. Here, the first five classes represent rice, the sixth one represents urban areas, the seventh represents water and the eighth represents other areas. As can be seen from figure 5, the random forest model had the most difficulty in distinguishing between the sixth and eighth class, but the remaining classes were generally well classified.



1	2260	0	1	0	0	1	0	0
2	6	1373	5	0	0	2	0	1
3	0	7	1050	1	0	2	0	1
4	0	0	5	520	0	0	0	0
5	0	0	0	0	303	0	0	0
6	3	9	2	1	0	2853	0	221
7	0	0	0	0	0	0	56860	0
8	0	1	0	0	0	111	0	15802

Figure 5: Confusion matrix

4.2. Results of task: Unsupervised Clustering with vegetation indices

A visual inspection of the results based on different number of clusters was performed and 6 clusters were formed for areas like water bodies, barren land and urban areas. Areas with vegetation were classified into more than one cluster. Further visual inspection of these clusters would throw more light on the homogeneity and accuracy of the clusters. For performing a visual inspection of the satellite images the image quality of the satellite images is important. We worked with the satellite images on Jupyter Notebook and clarity of the images was affected in this environment.

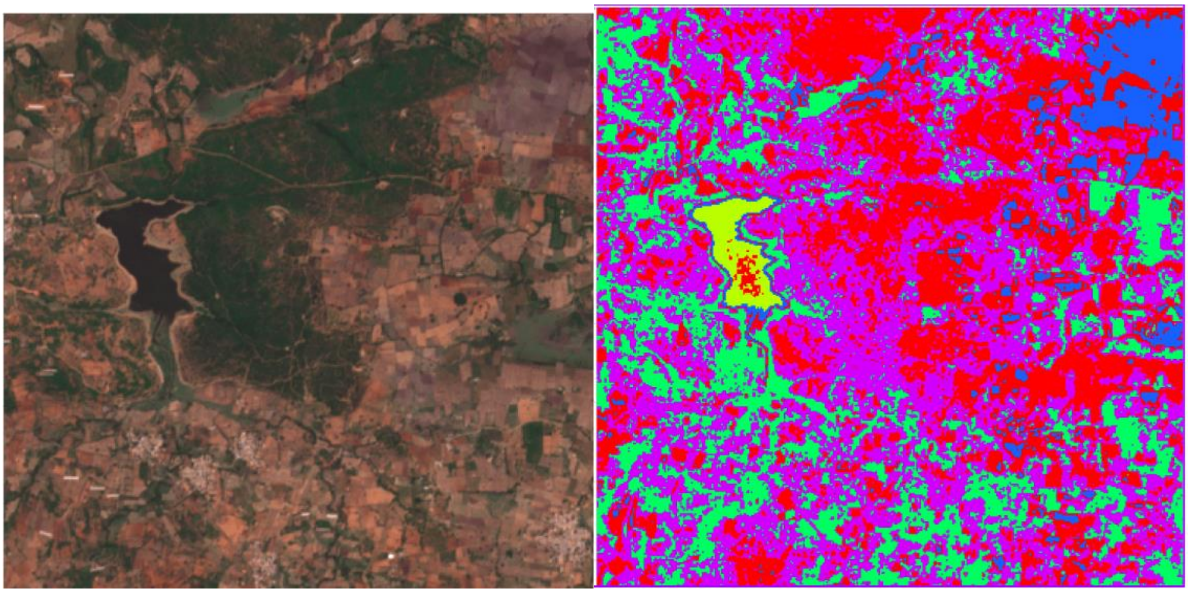


Figure 6: Results of Unsupervised Clustering

4.3. Limitations of the study

Based on figure 3, the classes urban and other were relatively indistinguishable from each other, even in the training data that we had created ourselves. This is turn means that while creating the polygons, some pixels within the polygons most likely belonged to the other class. Domain expertise could help in preventing such misclassifications.

Ideally, we would have liked to perform the data labeling and classification over the entire state of Telangana, but unfortunately this is not possible due to limited computational power. In fact, when we ran the model on a 3652 km2 district named Kamareddy, we received a memory error from google earth engine.

Other errors also consisted of rice being classified in rivers, which does not seem logical at all. Such an example is shown in figure 7, once again domain expertise could come in handy for checking in which areas rice can definitely (not) grow, which will limit our scope of search.

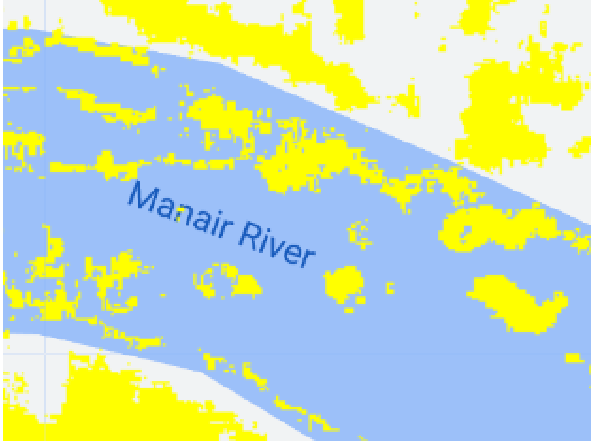


Figure 7: Examples of errors in classification on the map

Furthermore, the accuracy and the confusion matrix are not reliable metrics, as these are both focused on training data that we generated. For the validation to really take place, the model needs to be tested on an area that already consists of ground truth data.

5. Conclusions

We have come to the following conclusion regarding the two approaches taken:

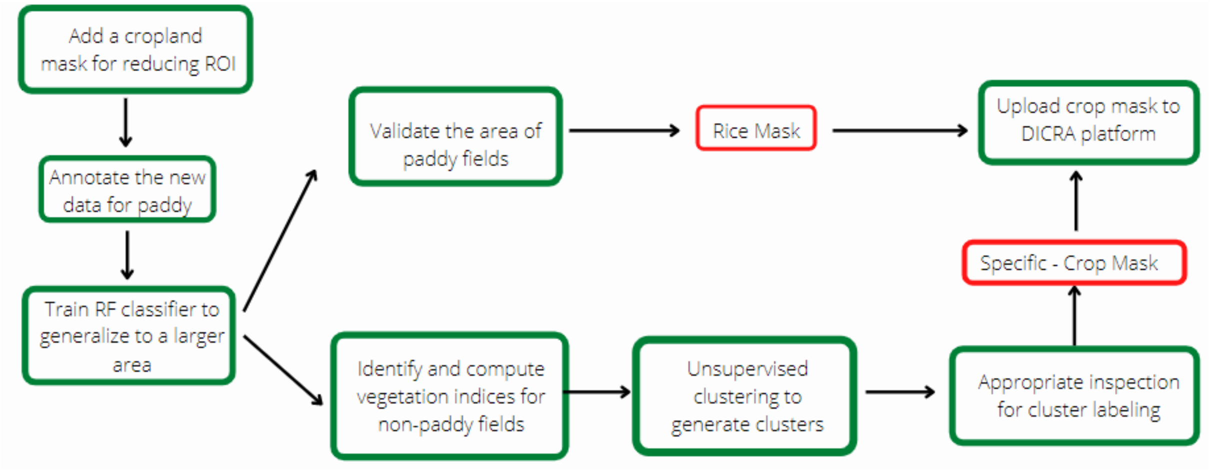
- Regarding paddy classification:
    - It is possible to identify paddy fields up to a certain extent using the VH polarization method.
    - There are quite a few inaccuracies that could be limited with the help of domain experts
- Regarding Unsupervised Clustering with vegetation indices
    - Domain knowledge is the primary requirement for being able to translate real world phenomena in a machine learning model as the inputs to an unsupervised learning algorithm determine the quality of the results
- Ground truth data is extremely valuable even if it is sparse.
    - It is difficult to measure the quality of any solution without ground truth data
- In the absence of ground truth data, building a model or finding patterns in crop growth and other phenomena related to the climate and environment would require domain knowledge in not only agriculture but also in the scope and potential of geospatial data features.

6. Recommendations

In light of the lack of ground truth data and difficulty in obtaining it we have a few recommendations on working on the crop classification problem. We recommend that

- efforts should be focussed on pattern analysis of one crop at a time by building models that utilise the unique growing patterns of each crop. Trying to classify all crops of interest in one model might not generate accurate results as the lack of ground truth data doesn't allow a model to learn from sample data. Thus, when udertaking pattern analysis one crop at a time should yield better results.
- domain experts also be a part of the problem solving. A domain expert in agriculture and remote sensing can work with data scientists by sharing information on crop growth patterns and relevant remote sensing data features that can be used to model these patterns. A data scientist can then leverage this domain knowledge for implementing data pipelines and building pattern recognition models that can perform crop identification. Therefore, a data scientist learns the tricks of the trade from the domain expert and then makes a machine perform the required task for getting results.
- geo referenced ground truth be gathered and curated for building resources like a crop type map. Ground truth data needs to be georefernced, i.e. the coordinates of the farms and which crop grows in the farm. Other details like the time of collection of the information is important.
- when using optical satellite images, as opposed to SAR data, for building models (depending on the type of solution that is required) one needs to be careful about optical satellite imagery suffering from cloud cover problem for some areas. This along with the availability of optical images for the period that one requires might create a satellite data availability problem. Therefore, we suggest this data availability problem be explored further before identifying the region from where ground truth data is intended to be collected.

Suggestions for future work



Add a cropland mask for reducing ROI
Annotate the new data for paddy
Train RF classifier to generalize to a larger area
Validate the area of paddy fields
Rice Mask
Upload crop mask to DICRA platform
Specific - Crop Mask
Identify and compute vegetation indices for non-paddy fields
Unsupervised clustering to generate clusters
Appropriate inspection for cluster labeling

Figure 8: Proposed process flow for future work

In figure 8, we present our recommendation for approaching crop classification in general. It makes sense to initially start by adding a cropland mask to reduce the region of interest, as we are only interested in distinguishing crops from each other. Then, paddy is annotated using the same approach as described in Section 3.6.1. Similarly, the rice classification is generalized to the larger region of interest. Once this is done, external ground truth labels should be used to validate the area of paddy fields. This could be done by collecting survey data, which would ideally be sampled from districts that are all over Telangana. This eventually results in a rice mask that can be uploaded to the DICRA platform. Simultaneously, the spectral bands and vegetation indices can be calculated for the non-rice fields. Unsupervised learning should be implemented to generate clusters, after which visual inspection can help in cluster labeling for the non-rice fields. Once again, this can be added to the DICRA platform.

References

OpenGeo Lab. (2022, April 17). Google Earth Engine - Rice/Paddy Crop Classification using Sentinel-1 SAR data. YouTube. https://www.youtube.com/watch?v=Ex544uYJRYw

*Kpienbaareh, D.; Sun, X.; Wang, J.; Luginaah, I.; Bezner Kerr, R.; Lupafya, E.; Dakishoni, L. Crop Type and Land Cover Mapping in Northern Malawi Using the Integration of Sentinel-1, Sentinel-2, and PlanetScope Satellite Data. Remote Sens. **2021**, 13, 700. https://doi.org/10.3390/rs13040700*

*Rao, P.; Zhou, W.; Bhattarai, N.; Srivastava, A.K.; Singh, B.; Poonia, S.; Lobell, D.B.; Jain, M. Using Sentinel-1, Sentinel-2, and Planet Imagery to Map Crop Type of Smallholder Farms. Remote Sens. **2021**, 13, 1870. https://doi.org/10.3390/rs13101870*

*Ma, Z.; Liu, Z.; Zhao, Y.; Zhang, L.; Liu, D.; Ren, T.; Zhang, X.; Li, S. An Unsupervised Crop Classification Method Based on Principal Components Isometric Binning. ISPRS Int. J. Geo-Inf. **2020**, 9, 648.*

Wang, S., Azzari, G., Lobell, D.B., 2019b. Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques. Remote Sens. Environ. 222, 303–317.

Mansaray, Lamin & Zhang, Dongdong & Zhou, Zhen & Huang, Jing-feng. (2017). Evaluating the potential of temporal Sentinel-1A data for paddy rice discrimination at local scales. Remote Sensing Letters. 8. 967-976. 10.1080/2150704X.2017.1331472.

*Wang, S.; Di Tommaso, S.; Faulkner, J.; Friedel, T.; Kennepohl, A.; Strey, R.; Lobell, D.B. Mapping Crop Types in Southeast India with Smartphone Crowdsourcing and Deep Learning. Remote Sens. **2020**, 12, 2957. https://doi.org/10.3390/rs12182957*

Singh, R.K., Rizvi, J., Behera, M.D., Biradar, C., 2021. Automated crop type mapping using time-weighted dynamic time warping-a basis to derive inputs for enhanced food and nutritional security. Curr. Res. Environ. Sustain. 3, 100032. https://doi.org/10. 1016/J.CRSUST.2021.100032