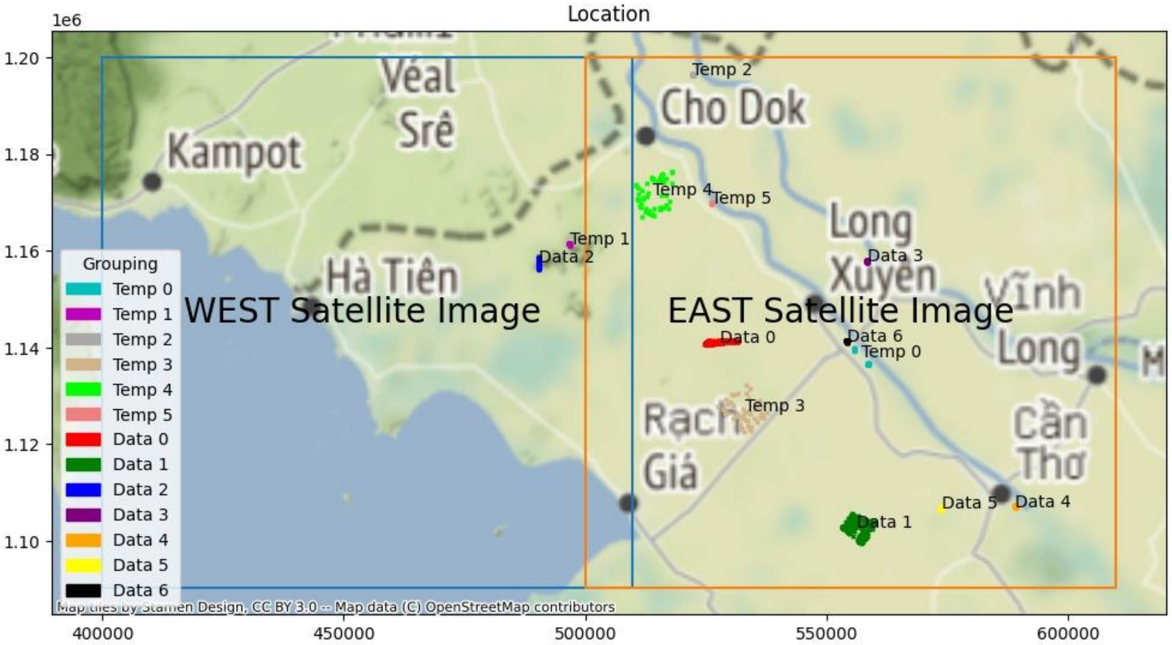
Data Preparation Considerations

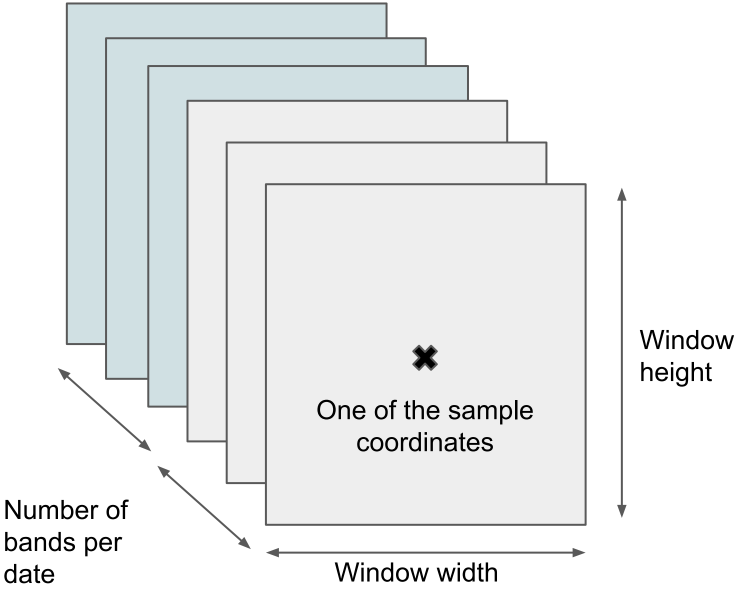


From EDA, we realised two characteristics of data that will affect the data preparation.

1. The training data is highly clustered into several groups. **When querying data, conditions are set based on the clusters instead of the full scene.**

2. Not all data can be retrieved from the same scene. **Data is retrieved from separate East scene and West scene respectively.**

Data Preprocessing and Structure



The diagram above shows the feature window extracted from satellite data for a sample coordinate.

Satellite data of different bands / wavelengths from Jan 2021 to Dec 2021 are queried based on various conditions such as NDVI and cloud coverage to provide comprehensive information to recognise paddy field. A full year of data is used to cover the features of the paddy fields at different seasons.

Windows of various sizes around the coordinate are extracted in the form of arrays. Each array contains the pixel values of a particular band and date of a sample coordinate as shown in the figure on the left. The data queried is then stored as dataframe, with each window stored in a dataframe cell.

Data sources and bands:

Sentinel-1 (radar data): VH, VV (can penetrate cloud but less informative visually)
Sentinel-2 (optical data) data: B01, B02, B03, B04, B05, B06, B07, B08, B8A, B09, B11, B12, SCL, WVP, visual. (more informative but unable to penetrate cloud)

Model Development Experiments

Multiple experiments are carried out to inform us about the development of classification model. By repeating the scientific process of hypothesise-experiment-validate, multiple important principles to develop rice paddy field classification model are extracted. At this preliminary stage, cross-validation is used to test and validate the principles learned.

Principle 1: The use of window return better results than a single pixel.

Random samples 20210301 and 20210311 are chosen from the downloaded Sentinel-2 data and stacked together as features and tested against multiple machine learning models. In this case, windows with 5 pixels around each coordinate is averaged to get the aggregated window values representing the whole window. The performance of models are evaluated with by using the aggregated window values and the single pixel values as data input respectively. The models show improved accuracy with the use of aggregated mean value.

Note: the window size of 5 is selected based on the analysis of distances between sampling points, of which we realise that in between points in the same cluster, most have distance of around 50m between points, corresponding to 5-pixels for the the highest spatial resolution of 10m for Sentinel-2 data.

Model Development Experiments

Principle 2: Adding more dates as features will improve model performance (until they overfit)

Using the same set of machine learning models, adding more dates as features improve the performance of features. This fits our common sense as crops go through multiple cycles throughout the year and more sampled dates result in more useful features informing the prediction model.

In our case, we study this by setting a set of number of dates sampled to be used as our predictive features, i.e. 1 day, 2 days, 1 month, 3 months, 4 months. The accuracy of models increase as we added more dates as features. However, we do not use all available dates throughout the year to avoid overfitting as we only have 600 training samples.

Principle 3: For the same input data, different model can show very different prediction.

Performance of each model fluctuates and one do not show consistent advantage over the other. It shows that model stacking might be able to help to make the model prediction to become more robust.

Model Development Experiments

Limitations of experiments

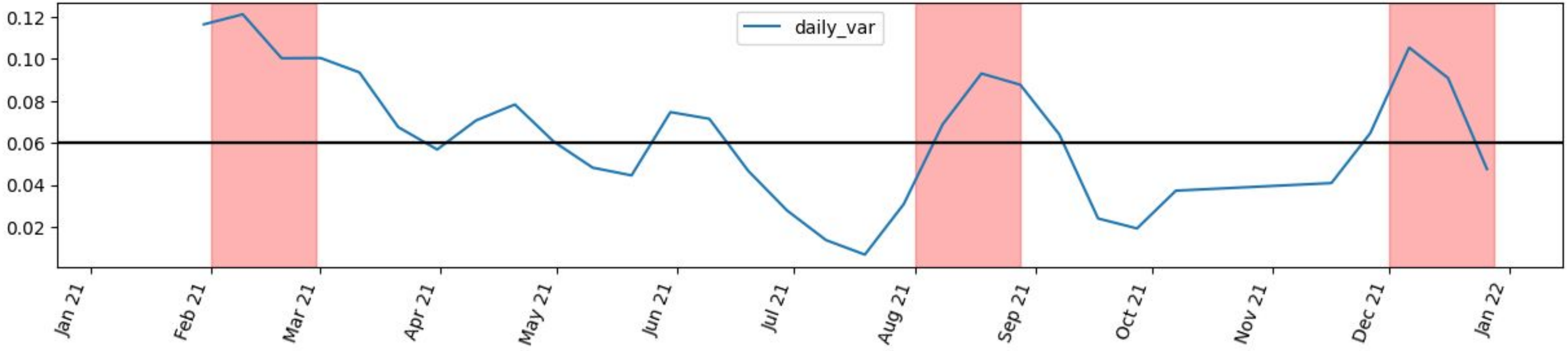
Due to the training samples being highly similar with one another and are not random enough, most of the experiments return F1-score very close to 1.0 when evaluated via cross-validation. This makes us being unable to see significant improvements or drop in f1-score with cross-validated evaluation. Thus, there are multiple explorations done but do not provide any valuable lessons for model development and many more experiments are needed to be validated against unseen test data, e.g.

- model stacking
- hyperparameter tuning
- feature selection by removing highly correlated values
- generate NDVI for each date
- adding Sentinel-1 data ('vv', 'vh', 'vv/vh')

Assumption 1: NDVI Variance in Cropping Season

On top of the principles extracted from model experimentations, we used unseen test data for validation of models development. In this case, we assume that **the color of rice during the cropping season is the greatest differentiator between rice field and the other land cover.** So, we selected data from the months with peaks in NDVI variance between rice fields and the other landcover, i.e. Februaru, August and December. As as baseline model, only Sentinel-2 data is used.

.



Evolution of Models + Test F1-scores

Based on the dataset queried based on the NDVI variance assumption, satellite data for February, August and December is used to develop classification models to predict the class of given coordinates

Iteration 1

test F1-score = 0.89

- Random Forest
- window size of 5*5, aggregated by mean

Baseline model.

Iteration 2

test F1-score= 0.86

- all February, August, December data + NDVI
- window size of 5*5, aggregated by mean
- remove highly correlated features
- stack models

From previous experimentation, we assumed that stacking models might lead to improved performance increasing robustness. However, model stacking reduces F1-score when tested against unseen test data. .

Iteration 3

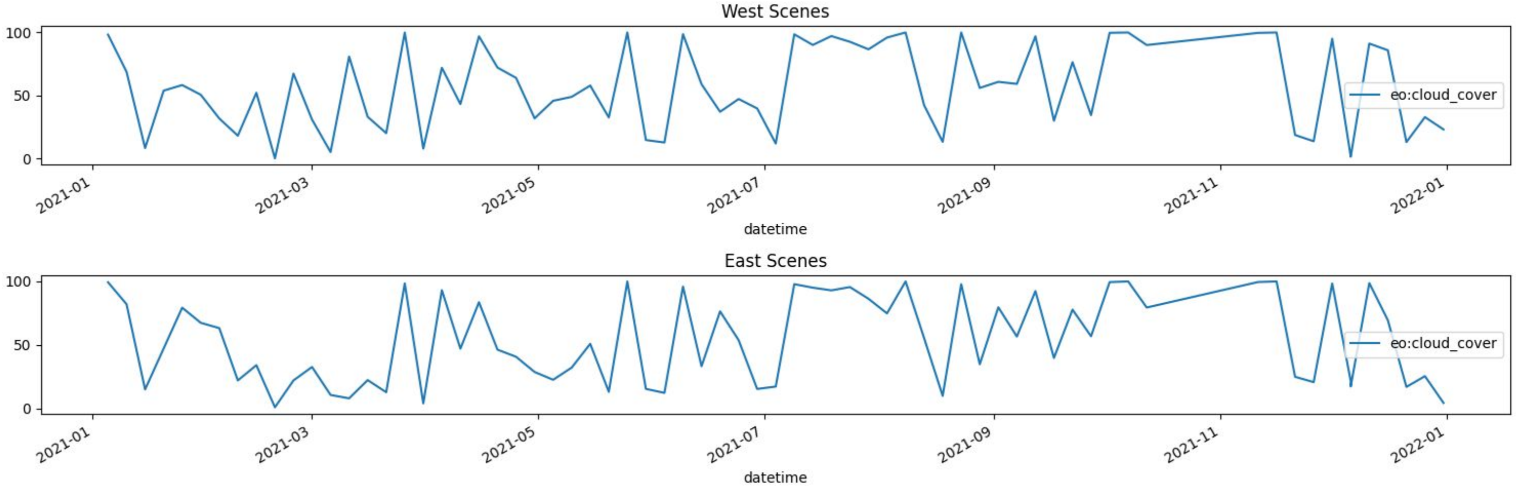
test F1-score= 0.95

- Sentinel-2: Feb, Aug, Dec raw data + NDVI
- Sentinel-1: Sentinel-1 raw data data + vh/vv
- drop highly correlated features
- normalization and stack models

Adding Sentinel-1 data to the configurations used in iteration 2 improved model performance significantly. *New assumption: ability to remove the influence of cloud is the most important factor to increase model performance.*

Assumption 2: Cloud Removal

As we observe that the inclusion of Sentinel-1 data in `iteration 3` significantly improves our prediction accuracy, we assume that **cloud removal is the most important factor in this classification task**. We conduct cloud analysis and find out the dates with the least amount of cloud hanging over the training sample coordinates throughout the year. Based on the dates, we extract their corresponding Sentinel-1 and Sentinel-2 data. We increased window size to 9*9 to reduce the impact of possible cloud covering the window.



Evolution of Models + Test F1-scores

Iteration 4

Test F1-score = 1.0 !!!

- take sentinel-1 and sentinel-2 data with <15% cloud coverage directly over our training data
- window size: 9*9, aggregated by mean
- normalization (MinMaxScaler)
- apply random forest classification

Based on the conditions established, we realise that we achieved a perfect F1-score of 1.0.

Iteration 4.1

Test F1-score = <1.0

- take sentinel-1 and sentinel-2 data with <15% cloud coverage directly over our training data
- window size: 9*9, aggregated by mean
- normalization (MinMaxScaler)
- remove highly correlated features
- apply random forest classification

Trying to simplify the classification model, removal of some highly correlated features results in reduction in test F1-score.

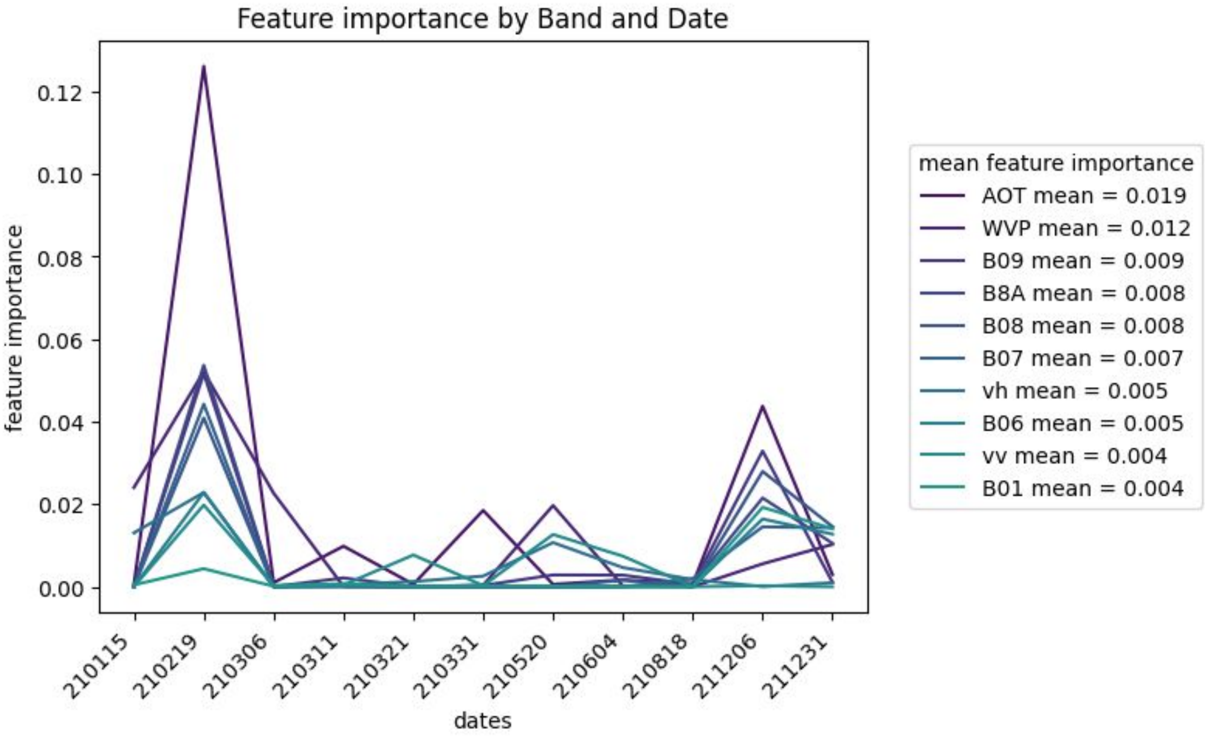
Iteration 4.2

Test F1-score = 1.0

- take sentinel-1 and sentinel-2 data with <15% cloud coverage directly over our training data
- window size: 9*9, aggregated by mean
- add NDVI and vh/vv features
- normalization (MinMaxScaler)
- remove highly correlated features
- apply random forest classification

By adding NDVI and vh/vv features to **iteration 4.1**, we achieve perfect F1-score of 1.0 again.

Highest Performing Features



Due to the large number of features and variation by time is involved, feature importance of the model is not easily interpretable.

In this case, we investigate the feature importance of each band for the dates chosen over the year. The plot shows the feature importances of the highest performing band over time.

Overall, ‘AOT’ and ‘WVP’ band have the highest average feature importance, but we notice that all bands have pretty small feature importance.