Digital Earth Pacific

Fractional Cover Concept Note

[**Title and Objective 2**](#_my7tfdz9gvll)

[Title: 2](#_ua5kykdtxq8g)

[Objective: 2](#_7gf9vqhbyyuu)

[Product Owner: 2](#_dnrq9rwvozvf)

[Supporting Team: 2](#_hpoabfwp2i12)

[**Rationale and Use Cases 2**](#_8a7be7pcyz3r)

[Rationale: 2](#_qp8xpldhy0fa)

[Use Cases: 2](#_6b13bdlekz7i)

[**Data Product Description 2**](#_htnp9as5dq6a)

[Abstract: 2](#_s2htwob356es)

[Data Sources: 2](#_5ncwziw51gxe)

[Methodology: 2](#_plymtla99meu)

[Spatial and Temporal Resolution: 3](#_a6wjrk5qbv1c)

[Validation: 3](#_p8wnkbmby7or)

[**Stakeholders and Collaboration 3**](#_9dkdh3uq31i0)

[Key Stakeholders 3](#_puiblzm38vqt)

[Partners 3](#_3tq4cb5xwezz)

[Funding 3](#_dkb88qipzm6)

[**Implementation Plan 3**](#_qq9j3co5na7x)

[Timeline 3](#_6gh9rtcgrvi)

[Risks 3](#_k3awvqk2f9r5)

[**References 3**](#_mahqb5r0n0ts)

## 

## Title and Objective

### Title: Fractional Cover for Landcover Classification

### Objective:

The goal is to train a Random Forest (RF) Classifier with Fractional Cover (FC) and Spectral Data as input which return Landcover Classification as output.

### Product Owner:

Philippine Laroche and Sachindra Singh

### Supporting Team:

List others that will be in a more supportive function in the development of this product.

## Rationale and Use Cases

### Rationale:

What is the key problem or opportunity the product addresses? Why should this product be prioritized - who is asking for it and why?

The Pacific Islands face a critical lack of locally accurate land cover datasets. Existing global products are designed for continental or global scales and often fail to capture the nuanced land cover dynamics of the Pacific region. This leads to misclassification, underrepresentation of smaller land cover types, and unreliable data for local decision-making. Our Random Forest classifier trained on fractional cover data addresses this gap by providing a regionally tuned, high resolution land cover product specifically trained on Pacific island environments (e.g., Fiji, Cook Islands, Marshall Islands).

Would help James for “Ocean Assests”, also can allow to detect landscape change after a disaster…

### Use Cases:

Who are the users and what use cases can this data product be applied to?

### Product Sponsor or Collaborator:

Is there an individual or organization from whom this product idea originated and can support collaboration or co-development of this product?

## Data Product Description

### Abstract:

Please provide a brief abstract describing this product.

### Data Sources:

This product requires Fractional Cover from Landsat-8 satellites (annually average and monthly average for temporal time series). It also requires spectral data download from Sentinel-2 : RGB bands + NIR from the Sentinel-2 Geomad dataset. From the spectral bands, different indexes were computed: NDVI (vegetation index), NDBI (building index) and NDWI (water index). Then using GLCM (Gray level covariance marix) and more precisely the Haralick Extraction feature function on QGIS 10 bands representing the texture (based on NDVI) were computed. Finally, the median fractional cover over january-april, may-august and september-december is used. As well, the median monthly NDVI (between 2020-2024 from Google Earth Engine.

### Methodology:

Product Development can be split into 3 steps. First, prepare and balance the data. Then train the random forest classifier and achieve tuning of the input data and hyperparameter. Finally, test the model on unseen data to evaluate the generalization/

1. Data Preparation and Balance:

Data from Fiji, Cook Island, Marshall Island and Palaos were download to train a random forest that return land cover classification over the pacific region. We have for each island a shapefile containing point with the associated grountruth land cover label and position (lat, lon). From the Sentinel-2 geomad, the RGB-NIR band were downloaded to compute the NDVI, NDWI, NDBI. From landsat-8 satellite we have .tif file with 3 bands for the annual and daily fractionnal cover over each island. Then, we prepare the data using QGIS to have one dataset (shapefile) for each island containing for each point the following data :

* lat, lon
* Fractional Cover (bs : baresoil ; pv : photosynthetic vegetation ; npv : non-photosynthetic vegetation)
* NDVI
* NDWI
* NDBI (Normalized Difference Build Index)
* RGB-NIR (spectral bands)
* EVI and SAVI
* Label **(class-id : 1. Forest ; 2. Cropland ; 3. Grassland ; 4. Buildup ; 5. Baresoil ; 6. Water ; 7. Mangroves)**
* lulc (land-use, land-cover)
* Texture data (10 bands computed using Haralick Extraction Feature in QGIS --> GLCM)
* Seasonal fractional cover (bs\_jan\_apr, bs\_may\_aug, bs\_sep\_dec, same for pv and npv bands)
* Monthly NDVI (median between 2020 and 2024 with cloud < 10%, download from GEE)

The data are in the folder data, there 4 files (one per island). To build a good classifier that generalizes well across the Pacific, balancing data is important :

- Balancing between classes (so each LULC type is equally learned)

- Balancing between locations (so it doesn't overfit to any specific island)

But a minimum amount of data is needed in order to have a good classifier. Based on this, the objective is to keep aroun 150-300 points per lulc. Finally to preserve spatial generalization, points from each island should be equally reprensented in each class.

I set manually :

* Forest : Fiji : 80 ; Cook : 75 ; Palaos : 70 ; Marshall : 75 = 300
* Grassland : Fiji : 150 ; Cook : 59 ; Palaos : 77 ; Marshall : 22 = 308
* Mangroves : Fiji : 180 ; Cook : 0 ; Palaos : 18 ; Marshall : 6 = 204
* Cropland : Fiji : 150 ; Cook : 80 ; Palaos : 63 ; Marshall : 4 = 300
* Buildup : Fiji : 110 ; Cook : 110 ; Palaos : 40 ; Marshall : 40 = 300
* Water : Fiji : 39 ; Cook : 55 ; Palaos : 36 ; Marshall : 25 = 155
* Baresoil : Fiji : 15 ; Cook : 115 ; Palaos : 115 ; Marshall : 55 = 300

Following this distribution, the RF would be trained on 1867 data points (80% for training and the remaining 20% for validation).

Secondly, in order to improve performance, we wanted to increase the size of the training data. We therefore decided to maintain a dataset as balanced as possible between classes, but chose to ignore the balance between locations. Following this distribution, the RF is now trained on 3347 data points. Increasing the number of training points allows us to increase the number of features used during training without increasing the risk of suffering from the “curse of dimensionality.”

1. Training RF Classifier :

Initial training was performed using only Fractional Cover (FC) features. While the model learned general patterns, it failed to capture class-specific characteristics, resulting in low accuracy (~55%). To improve performance, spectral features (RGB bands) and vegetation/water indices (NDVI, NDWI) were added. This enriched the feature space with reflectance-based and moisture-related information, increasing accuracy to ~65%. However, the model still struggled to distinguish cropland, grassland, and forest—likely due to their spectral similarity.

To better capture spatial context, NDVI texture features were introduced using GLCM (Haralick features computed in QGIS). A [study](https://www.tandfonline.com/doi/full/10.1080/01431161.2016.1278314) suggests GLCM mean is most effective for classification, with contrast and entropy adding value in edge-rich or detailed contexts. Since band 1 appears to represent the mean, it was retained as texture1. A PCA-based reduction of all texture bands was tested but performed poorly, likely due to loss of spatial nuance.

Vegetation indices such as EVI and SAVI were also tested to better discriminate dense versus sparse vegetation. However, their inclusion led to misclassification, especially overprediction of cropland, suggesting feature redundancy or class confusion. Thus, we only keep one of them for the training to avoid redundancy. NDBI, intended to improve separation of bare soil, lead to more forest pixel classified as baresoil or cropland but increase overall performance.

Alternative classifiers (XGBoost, SVM) and ensemble stacking were tested but underperformed compared to Random Forest, possibly due to the small or imbalanced training dataset.

The NDVI class-wise mean per island (ndvi\_mean) was added. While this greatly boosted accuracy (~96%), it dominated feature importance, causing overfitting by allowing the model to learn class identity rather than general patterns.

Finally, the last attempt to increase overall performance was to add temporality to the data in order to account for yearly changes (wet season and dry season). The fractional cover temporal data were computed (see part 5 of the code) for three seasons: January to April (wet season), May to August, and September to December (dry season). Moreover, we also decided to add the monthly NDVI, as it could help provide more information about seasonal changes over forests and also take into account the harvesting season (drop in cropland NDVI).

Model performance was evaluated using accuracy, F1-score, and the Kappa coefficient to capture complementary aspects of classification quality. Accuracy provides an overall measure of correctly classified samples, while the F1-score (weighted) accounts for class imbalance by combining precision and recall. The Kappa coefficient assesses agreement between predictions and reference data beyond chance, offering a more robust evaluation when classes are unevenly represented or easily confused.

### Spatial and Temporal Resolution:

Data are all from Landsat-8 sensors, the spatial resolution is 30meters. For the temporal resolution, fractional cover is the annual mean of 2024. RGB-NIR and annual indexes NDWI, NDVI, NDBI, SAVI and AVI are annual means of 2023 landsat-8 imagery. Finally, for the temporal data, NDVI monthly median are downloaded using GEE, the median is assessed monthly between 2020/2024.

### Validation:

Provide brief description on how the product will be validated if known.

## Stakeholders and Collaboration

### Key Stakeholders

Identify the primary users and beneficiaries of this product. What ministries would this data product be most useful for?

### Partners

List any partners that are involved in the development, sharing or use of this product.

### Funding

List the core funding source being used for the development of this product.

## Implementation Plan

### Timeline

Provide an overview of the work plan with associated timeline for the following phases: Research, Alpha Product, Product Improvement, Beta Product, Quantitative Validation, Provisional Product, Qualitative Assessment, Operational Product.

### Risks

Summarize any risks or challenges in the development of this product (data, finance, resources, validation, etc.).

## References

Andrade, João, et al. “Evaluating Single and Multi-Date Landsat Classifications of Land-Cover in a Seasonally Dry Tropical Forest.” *Remote Sensing Applications: Society and Environment*, vol. 22, Apr. 2021, p. 100515, <https://doi.org/10.1016/j.rsase.2021.100515>.

Masiliūnas, Dainius, et al. “Global Land Characterisation Using Land Cover Fractions at 100 m Resolution.” *Remote Sensing of Environment*, vol. 259, June 2021, p. 112409, <https://doi.org/10.1016/j.rse.2021.112409>.

Waldner, François, et al. “Land Cover and Crop Type Classification along the Season Based on Biophysical Variables Retrieved from Multi-Sensor High-Resolution Time Series.” *Remote Sensing*, vol. 7, no. 8, Aug. 2015, pp. 10400–24, <https://doi.org/10.3390/rs70810400>.

Hall-Beyer, Mryka. “Practical Guidelines for Choosing GLCM Textures to Use in Landscape Classification Tasks over a Range of Moderate Spatial Scales.” *International Journal of Remote Sensing*, vol. 38, no. 5, Mar. 2017, pp. 1312–38, <https://doi.org/10.1080/01431161.2016.1278314>.