Technische Universität München

DATE

Project: Tree Species Classification (WS 22/23)

Data Science in Earth Observation

by
Prof. Dr. -Ing. habil. Xiaoxiang Zhu
Prof. Dr.-Ing. Muhammad Shahzad
Mr. Yang Mu

09/02/2023

submitted by:

Jayendra Chorapalli
Vincent Orth
Kirill Volter
Sirui Wang
Manap Shymyr

Contents

1. Abstract

Tree species classification plays an important role in the fields of ecosystem monitoring, biomass prediction and wood fuel production estimation. Machine learning (ML) for Remote Sensing (RS) method provides an efficient way to classify tree species with less cost and high accuracy. In order to get machine learning features from Earth Observation (EO) missions. Sentinel-2 (S2) optical satellite images from European Space Agency (ESA) have been used as the main datasets. After cloud masking, four seasons' scenes from sentinel-2 images have been generated as spectral information. And vegetation indices have been calculated as extra features, which can make great benefits to the training models. 7 species are classified by Random Forest (RF), Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Mobile Neural Network (MobileNet) and Recurrent Neural Network (RNN). The final overall accuracy varies from 76.2% (RNN) to 79.4% (CNN). Remarkably, CNN model by multispectral images archives a great result for tree species classification.

Keywords: tree species classification; Sentinel-2; multi-temporal; deep learning

2. Introduction

Earth observation (EO) has a high potential for biodiversity assessments, mainly for the description of vegetation habitats [Kuenzer, et. al., 2014]. Tree species diversity is very important for the ecosystem.

In the last few years, many related studies have been done by many publications. High spatial resolution of WorldView-2 satellite data classified 10 species with Random Forest method [Immitzer, et. al., 2012]. Besides, hyperspectral images from HySpex used Support Vector Machines (SVM) and Random Forest method to classify three species with the high accuracy [Dalponte, et. al., 2013].

Up to now, the tree species classification still has many challenges. High spatial resolution and high spectral resolution can not be gathered by the same dataset. Different types of trees have small differences in spectral space compared with normal land cover classes. Ground truth references are not perfect and hard to be interpreted by airborne maps. In order to overcome existing changes and get higher accuracy, the workflow in Fig. 1. has been built.

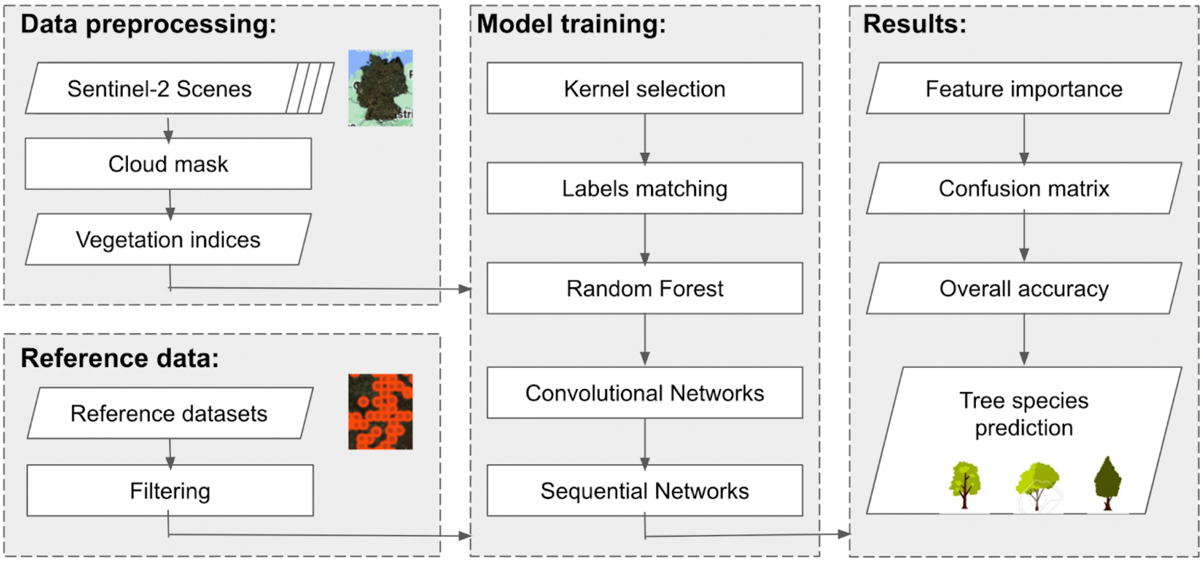

Fig. 1. Workflow of tree species classification

3. Data curation

3.1. Data acquisition

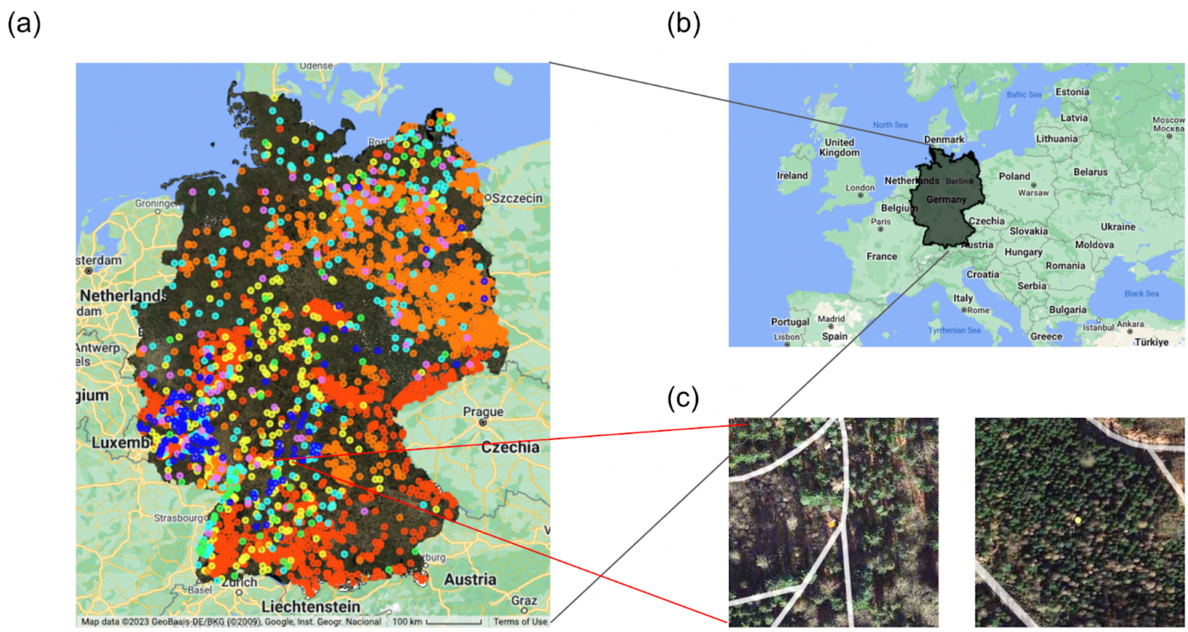

Fig. 2. Overview of the Region of Interest (ROI) and reference data. (a) Regular grid of reference data feature collection and Sentinel-2 RGB scene. (b) Location of ROI in europe. (c) Examples of samples on Google Earth map.

Central Europe has a variety of deciduous and evergreen trees. The vast forests are mainly located in southwestern and northeastern Germany. The 4 datasets in Tab. 1. have been used in our research.

Tab. 1. Overview of datasets

Dataset Name	Time	Resolution (m)	Description
Sentinel-2	2021	10	10 Multispectral bands with 4 seasons
Reference data	2017	1000	Labels of tree species of EU-Forest
CNES/ Airbus	2023	0.5	Very high resolution image on Google Earth from The French Space Agency (CNES) for interpretation
CGLS-LC100	2019	100	Copernicus Global Land Cover Layers Collection 3 for forest mask

3.2. Data preparation

Labeled reference data preparation: The German tree species data of EU-Forest (Mauri et al., 2017) is used as the reference data, which contains a total of 84 tree species and 67,412 sample plots. '*File1.csv*' contains the coordinates (X, Y) and the name (SPECIES NAME) of tree species. (X, Y) indicates coordinates of tree species in a ETRS89-LAEA reference coordinate system, representing the centroid of the INSPIRE-compliant 1 km × 1 km European grid. '*File2.csv*' is the statistical result of File 1, which counts the number of occurrences of each tree species and sorts them. '*File1.csv*' and '*File2.csv*' are downloaded from Moodle and converted into WGS-1984 projection in ArcGIS. '*File2.csv*' is joined to '*File1.csv*' and the shapefile of these 67,412 samples are uploaded to Google Earth Engine (GEE) asserts.

The 9 by 9 kernel from S2 image around the reference point will be obtained as training data. Since the resolution of the grid is 1 km × 1 km, the reference points need to be filtered by several properties to make sure the kernel can be interpreted as each forest. There are 2 filters have been used in our research:

1. Using Copernicus Global Land Cover Layers (CGLS) as the forest mask, filtered out each broad and needle leaf forest with the canopy density ratio of 70%. And use the very high resolution image as reference, make sure the centroid points are located in the condense forest borders. Then each kernel will cover the high density forest area, and other land covers will be filtered out.
2. Using the coordinate distinguishing method on Google Earth Engine. Remove all the points which contain multi species in one location. This filter can make sure the reference data that we used only contains one specific tree in one location.

Then the final samples which are used in our research are shown in Tab. 2. The first 7 common species in Germany are taken into account.

Satellite image data preparation: All images are grouped per season (Spring: March 1st - May 31; Summer: June 1st - August 31; Autumn: September 1st - November 30); The 10 spectral bands providing valuable information for vegetation are selected (visible bands 2, 3 and 4; red-edge bands 5, 6 and 7; NIR bands 8 and 8A; as well as the SWIR bands 11 and

12); The satellite images around samples in Tab. 2. are obtained from Google Earth Engine. Seasonal composites are stacked, which conta

ining 40 bands (4 seasons times 10 bands) are exported to Google drive and can be downloaded. And the file format is GeoJson. Each tree species has a different file.

Tab. 2. Samples of different tree species

Specie Name	Acronym	Index	Samples	Leaf type
Picea abies	PA	1	1095	Evergreen needle leaf
Fagus sylvatica	FS	2	293	Deciduous broad leaf
Pinus sylvestris	PS	3	1290	Evergreen needle leaf
Quercus robur	QR	4	235	Deciduous broad leaf
Betula pendula	BP	5	88	Deciduous broad leaf
Quercus petraea	QP	6	124	Deciduous broad leaf
Fraxinus excelsior	FE	7	66	Deciduous broad leaf

Training and testing file preparation: In order to improve the speed of Neura

+l network training. All the GeoJson files are loaded by python script '*tree_classification_datapre.ipynb*' and written into '*.npy*' files. The indexes of each tree species are also matched with satellite datasets as labels at the same time. Training and testing samples are separated by the ratio of 80% and 20% from python script '*tree_classification_NN.ipynb*'.

3.3. Data preprocessing & augmentation

The final training dataset has the dimension of (2547, 9, 9, 80). And the final testing dataset has the dimension of (637, 9, 9, 80). The average spectral lines for 7 tree species for 4 seasons indicate the main differences. Different types of trees have different reflectance of visual bands. The red-edge in near infrared red bands represent different biology components and values. Four seasons can provide the information of deciduous or evergreen forest and also can provide information if the tree leaves turn to yellow or red in autumn (Fig. 3. - Fig. 6.).

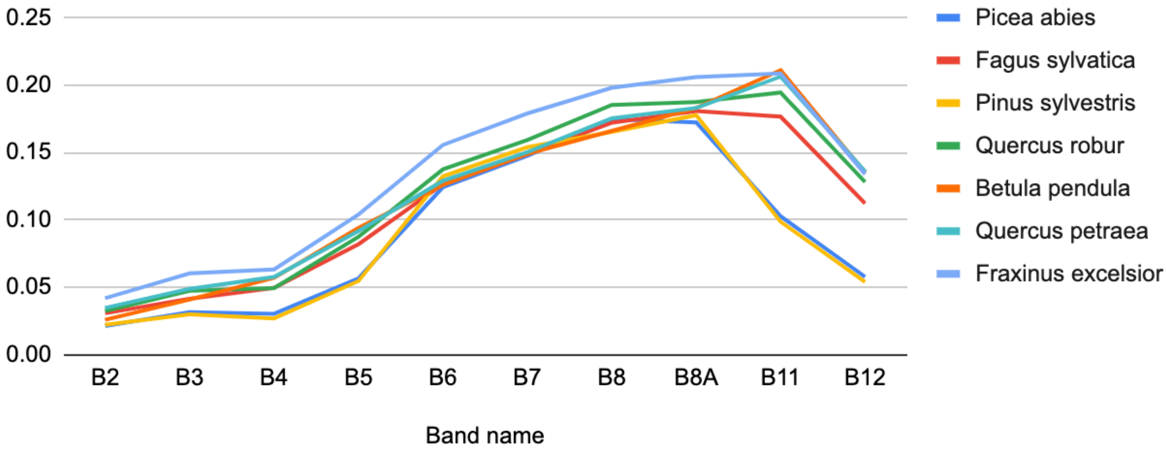

Fig. 3. Spectral lines of tree species in spring

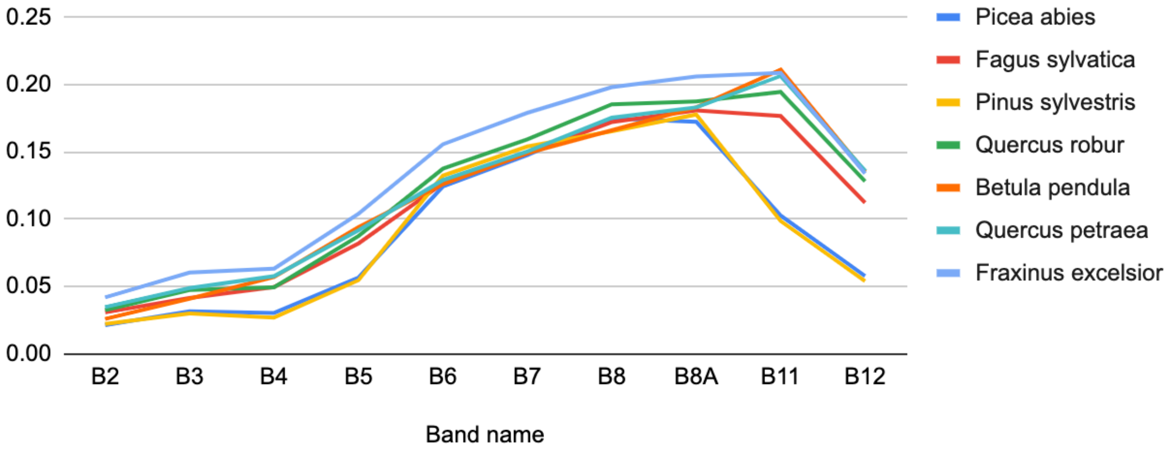

Fig. 3. Spectral lines of tree species in summer

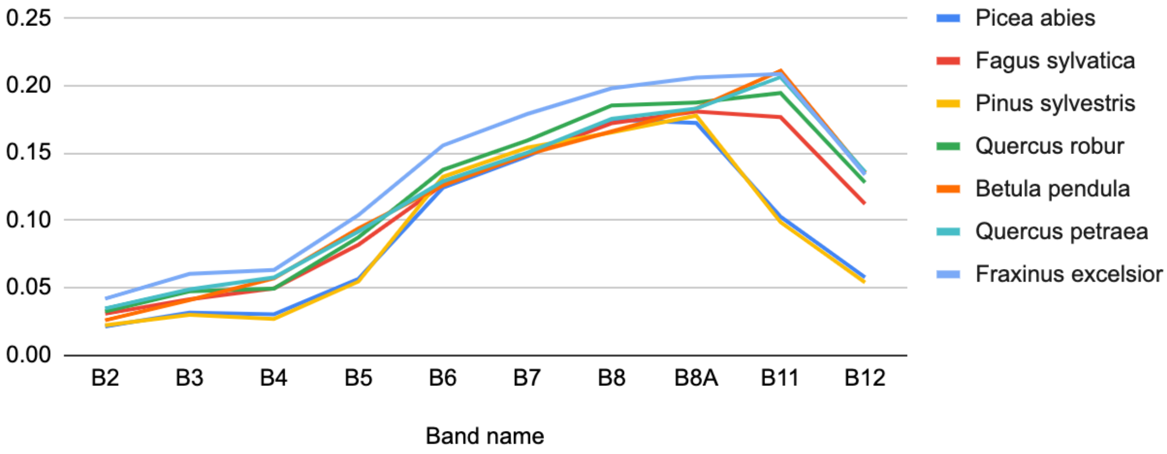

Fig. 3. Spectral lines of tree species in autumn

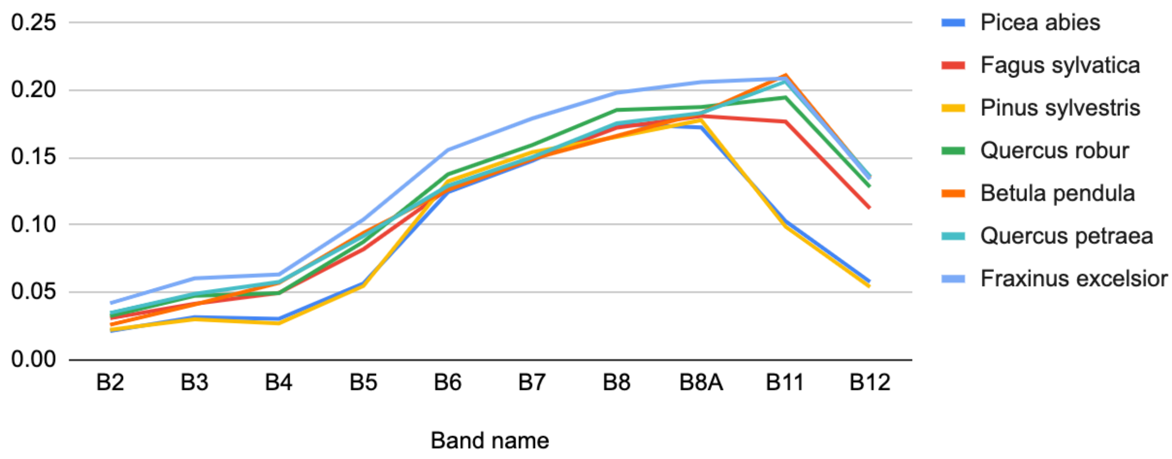

Fig. 3. Spectral lines of tree species in winter

Data augmentation: As we see from spectral lines, there are some species that have very similar spectral information. Especially the trees in the same class, for example, Quercus robur and Quercus petraea. Vegetation indices in Tab. 3. are chosen to decrease the feature similarities. 40 spectral bands and 40 indices bands are merged to 80 bands by data fusion.

Tab. 3. Vegetation Indices Table

Vegetation Indices	Formulary	Reference
NDVI	(NIR-Red)/(NIR+Red)	[]
NDWI	(NIR-SWIR1)/(NIR+SWIR1)	[]
NIR and RE1 ratio (SRNIRRE1)	NIR/RE1	[]
Normalized Difference Red-Edge and SWIR2(NDRESWIR)	(RE2-SWIR2)/(RE2+SWIR2)	[]
Greenness Index (GI)	Green/Red	[]
NDVI	Red/NIR	[]
NDVI	Red/NIR	[]
NDVI	Red/NIR	[]
NDVI	Red/NIR	[]
NDVI	Red/NIR	[]

4. Model selection

5. Model training

6. Network tuning

7. Experimental results

8. Discussion

9. Conclusions

Project package description:

1.Python scripts:
‘tree_classification_datapre.ipynb’
‘tree_classification_NN.ipynb’

2. Reference data:
‘File1.csv’
‘File2.csv’

3. Training and testing data:
‘Training.npy’
‘Testing.npy’

10. References

[1] Tree species introduction: https://www.tree-guide.com/
[2] Kuenzer, C.; Ottinger, M.; Wegmann, M.; Guo, H.; Wang, C.; Zhang, J.; Dech, S.; Wikelski, M. Earth observation satellite sensors for biodiversity monitoring: Potentials and bottlenecks.Int. J. Remote Sens.2014,35, 6599–6647
[3] Immitzer, M., Atzberger, C., & Koukal, T. (2012). Tree species classification with Random forest using very high spatial resolution 8-band worldView-2 satellite data. *Remote Sensing*, *4*(9), 2661–2693. https://doi.org/10.3390/rs4092661
[4] Dalponte, M., Ørka, H. O., Gobakken, T., Gianelle, D., & Næsset, E. (2013). Tree species classification in boreal forests with hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(5), 2632–2645. https://doi.org/10.1109/TGRS.2012.2216272
[5] Mauri, A., Strona, G., & San-Miguel-Ayanz, J. (2017). EU-Forest, a high-resolution tree occurrence dataset for Europe. Scientific data, 4(1), 1-8.

Tasks (deleted at the end)

1. Cover page with the your group member names, student numbers, and date.
2. Abstract
(Here you summarise your work and results)
3. Introduction
(Here you could provide background and problem to be addressed and reference any literature which you used)
4. Data curation
4.1 Data acquisition
(Here you provide how did you acquire your data)
4.2 Data preparation
(Here you provide how you prepared the data labels)
4.3 Data preprocessing & augmentation
(Here you provide information related to any preprocessing you carry out)
5. Model Selection
(Why you choose this model? Justify your choice of model)
6. Model Training
(Here provide implemnetation details)
7. Network Tuning
(Here proide information related to hyperparameter tuning)
8. Experimental Results
(Here you should include your model performance on your own train/val/test splits as well on the test data which we shall povide)
9. Discussion
(Here you could provide comments on the results which you get)
10. Conclusions
(Here draw your conclusions)
11. References