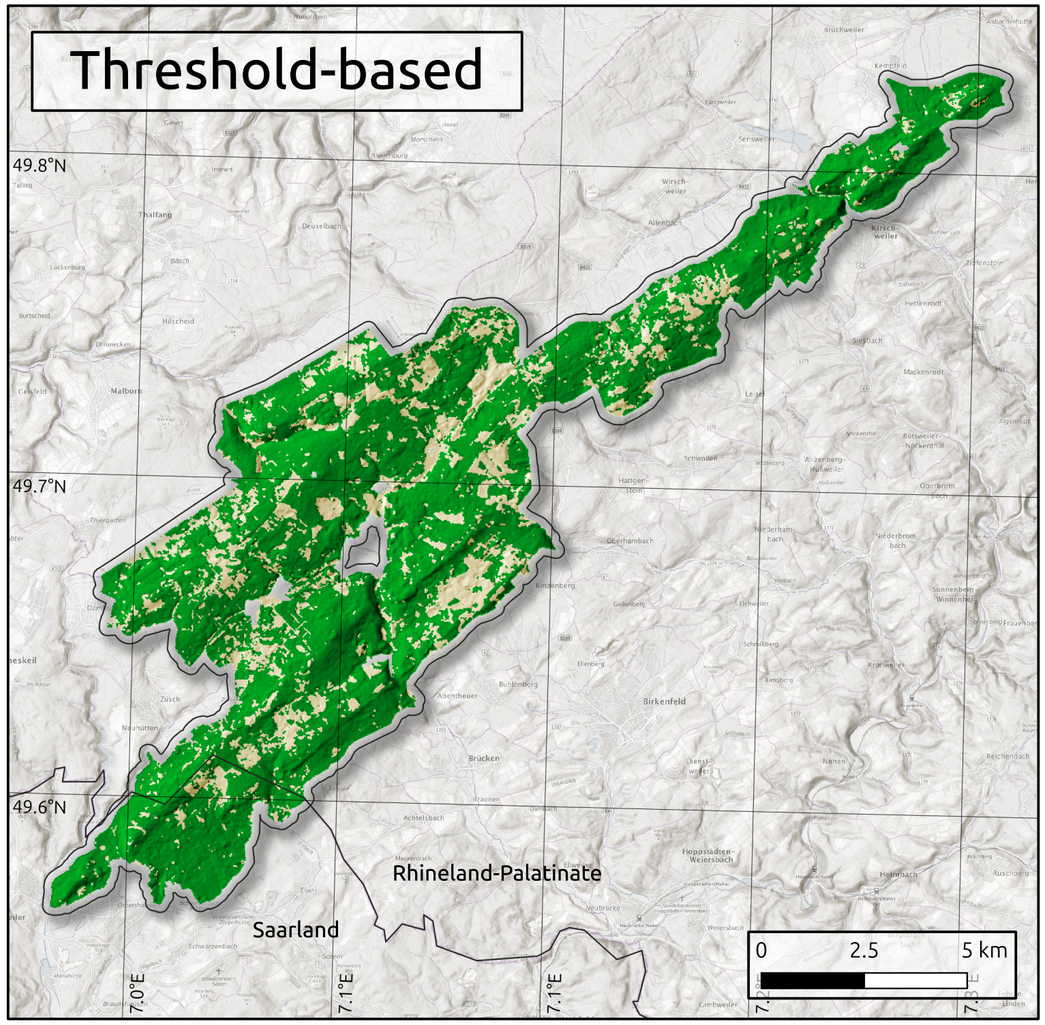
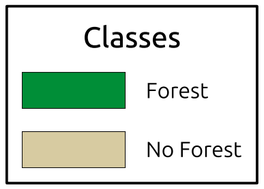
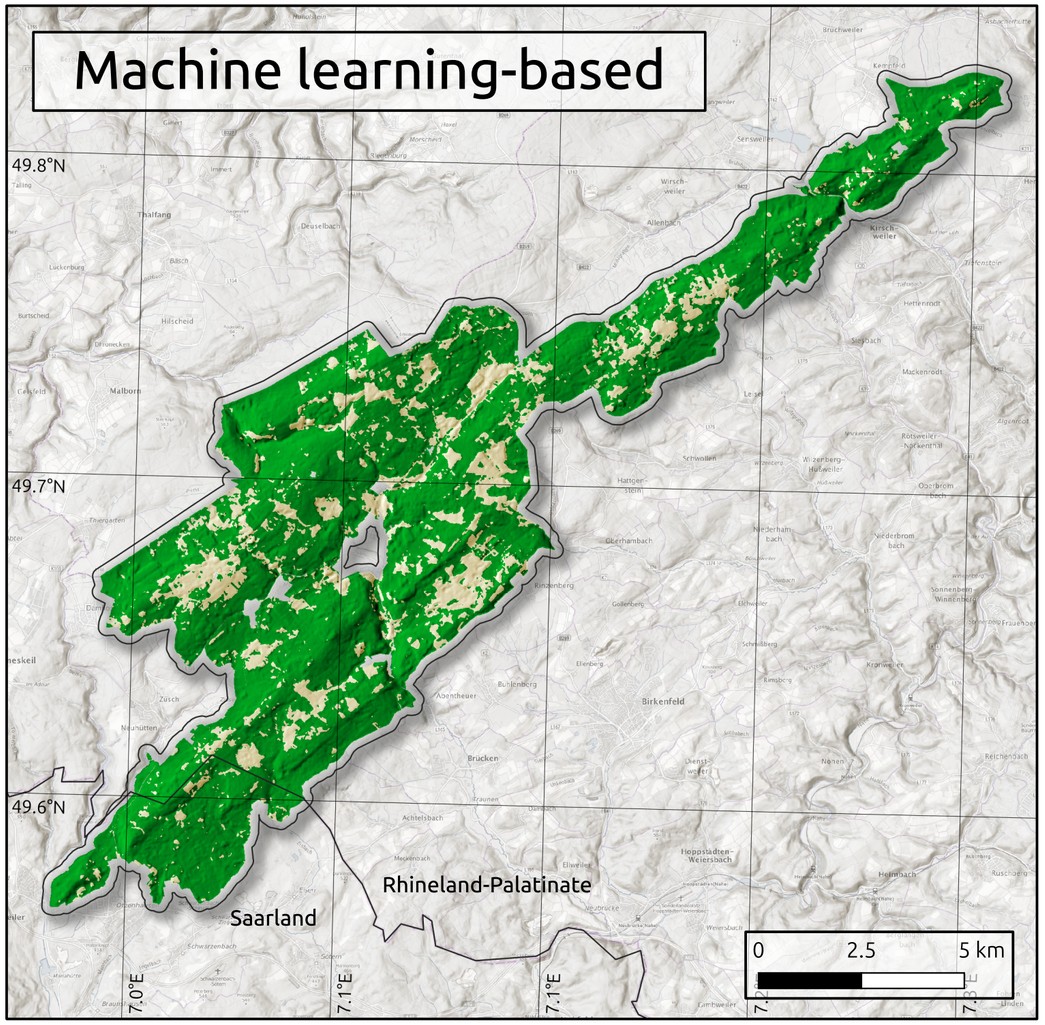
Comparison of a threshold-based and machine-learning based forest mask for the Hunsrück-Hochwald National Park

GEO INFORMATICS

UNIVERSITÄT TRIER







The forest mask was created using a combination of LiDAR-derived canopy height data and vegetation indices. Areas were classified as forest if they met the following criteria:

- **Canopy height ≥ 5 meters** or missing height data (to include potential tree-covered areas with no height values),
- **NDVI > 0.35**, indicating healthy green vegetation,
- **NDRE > 0.25**, capturing chlorophyll content typical of forested regions.

The LiDAR-derived canopy height has two different acquisition dates that both do not match the acquisition date of the Sentinel-2 satellite image that was used for the calculation of the vegetation incdices. Thus this approach does not claim a high accuracy. Nevertheless the approach is validated according to the acquisition date of the Sentinel-2 images, which is the same that was used for the machine-learning-based approach and can therefore be easily compared to it.

Agreement in forested area

Total agreement **85.63 %**

Jaccard simmilarity: **82.86 %**

Confusion Matrix: Threshold-based vs ML-based

		ML-based	
		Forest	**No Forest**
TH-based	**Forest**	186,508	15,182
	No Forest	21,821	33,980

The forest mask is based on Sentinel-2 satellite imagery and a machine learning approach using the Random Forest algorithm. The workflow includes the following steps:

1. **Data Preparation:** A Sentinel-2 image stack was extended by additional derived features (see p. 2, Feature Table) and used as input. Sample polygons representing forest and non-forest areas were loaded and the spectral values were extracted for each polygon. Only a subset of 50% of the data was used, maintaining a balanced distribution across polygons. Missing values were removed.
2. **Model Training:** A spatial 3-fold cross-validation was implemented to ensure spatial separation between training and validation data. A Random Forest classifier was trained, and the most important spectral bands were selected based on variable importance. A second Random Forest model (final model) was trained using only the most relevant predictors (those with importance > 40). This improves model efficiency while maintaining accuracy.
3. **Prediction** The trained model was applied to the NLPHH masked Sentinel-2 image to generate a binary forest/non-forest map.

Sources:
- ©GeoBasis-DE / LVermGeoRP, December 2020 / May 2023, dl-de/by-2-0, www.lvermgeo.rlp.de - Laser points surface and ground - [Data processed to forest canopy height for the National Park Hunsrück-Hochwald; Accessed: 1st May 2025]>
- Data: European Space Agency (2025). Copernicus Data Space Ecosystem). https://dataspace.copernicus.eu/. S2C_MSIL2A_20250501T104041_N0511_R008_T32ULA_20250501T161558.SAFE. [May 6, 2025].
- Data: University of California, Berkley. Global Administrative Areas Version 4.1 www.gadm.org (2025). [May 9, 2025].
- Basemap: © Esri. "World Hillshade" [basemap]. Scale Not Given. World Hillshade. [May 20, 2025]. Available at: https://www.arcgis.com/home/item.html?id=1b243539f4514b6ba35e7d995890db1d.
- Basemap: © Esri, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap, and the GIS User Community.
- Basemap: © BKG (2025) dl-de/by-2-0 [May 20, 2025]

CRS: ETRS89 UTM Zone 32 (EPSG:25832)
Author: Fabian Schumacher
Affiliation: Trier University, Department for Geoinformatics
Date: 17th June 2025

Accuracy Assessment

Threshold-based

1) Error matrix:

		Validation					
		1 (Forest)	2 (No Forest)	Total	User's accuracy	Total class area (m²)	Wi
Thematic map	1	49	1	50	0.98	81309200.00052	0.78332
	2	11	39	50	0.78	22491600.00014	0.21668
	Total	60	40	100		103800800.00067	
	Producer's accuracy	0.94153	0.91517		0.93666		

2) Accuracy:

Overall:

Overall Accuracy	Standard deviation
0.93666	0.02024

User:

	User's accuracy	Standard deviation
1 (Forest)	0.98	0.02
2 (No Forest)	0.78	0.05918

Producer:

	Producer's accuracy	Standard deviation
1 (Forest)	0.94153	0.01519
2 (No Forest)	0.91517	0.07766

2b) Accuracy matrix of estimated area proportion:

User:

		Validation	
		1 (Forest)	2 (No Forest)
Thematic map	1	0.98	0.02
	2	0.22	0.78

Producer:

		Validation	
		1 (Forest)	2 (No Forest)
Thematic map	1	0.94153	0.08483
	2	0.05847	0.91517

3) Error matrix of estimated area proportion:

		Validation		
		1 (Forest)	2 (No Forest)	Wi
Thematic map	1	0.76765	0.01567	0.78332
	2	0.04767	0.16901	0.21668
	total	0.81532	0.18468	

4) Quadratic error matrix of estimated area proportion:

		Validation	
		1 (Forest)	2 (No Forest)
Thematic map	1	0.00025	0.00025
	2	0.00016	0.00016
	total	0.02024	0.02024

5) Class area adjusted:

	Area adjusted (m²)	Error	Lower limit	Upper limit	Coefficient of variation	Uncertainty
1 (Forest)	84631168.00054	2101442.1336	80512341.41869	88749994.58239	2.48 %	0.04867
2 (No Forest)	19169632.00012	2101442.1336	15050805.41827	23288458.58197	10.96 %	0.21486
total	103800800.00067					

Machine Learning-based

1) Error matrix:

		Validation					
		1 (Forest)	2 (No Forest)	Total	User's accuracy	Total class area (m²)	Wi
Thematic map	1	50	2	52	0.96154	83326000.00053	0.80912
	2	4	46	50	0.92	19657100.00013	0.19088
	Total	54	48	102		102983100.00066	
	Producer's accuracy	0.98075	0.84946		0.95361		

2) Accuracy:

Overall:

Overall Accuracy	Standard deviation
0.95361	0.02301

User:

	User's accuracy	Standard deviation
1 (Forest)	0.96154	0.02693
2 (No Forest)	0.92	0.03876

Producer:

	Producer's accuracy	Standard deviation
1 (Forest)	0.98075	0.00919
2 (No Forest)	0.84946	0.08954

2b) Accuracy matrix of estimated area proportion:

User:

		Validation	
		1 (Forest)	2 (No Forest)
Thematic map	1	0.96154	0.03846
	2	0.08	0.92

Producer:

		Validation	
		1 (Forest)	2 (No Forest)
Thematic map	1	0.98075	0.15054
	2	0.01925	0.84946

3) Error matrix of estimated area proportion:

		Validation		
		1 (Forest)	2 (No Forest)	Wi
Thematic map	1	0.778	0.03112	0.80912
	2	0.01527	0.17561	0.19088
	total	0.79327	0.20673	

4) Quadratic error matrix of estimated area proportion:

		Validation	
		1 (Forest)	2 (No Forest)
Thematic map	1	0.00047	0.00047
	2	5e-05	5e-05
	total	0.02301	0.02301

5) Class area adjusted:

	Area adjusted (m²)	Error	Lower limit	Upper limit	Coefficient of variation	Uncertainty
1 (Forest)	81693721.84668	2369644.26038	77049219.09634	86338224.59702	2.9 %	0.05685
2 (No Forest)	21289378.15398	2369644.26038	16644875.40364	25933880.90432	11.13 %	0.21816
total	102983100.00066					

Accuracy assessment results for the two maps. Assessment was performed according to the methodology of Oloffson et al. (2013, 2014).

Additional information: Preprocessing for machine-learning-based approach

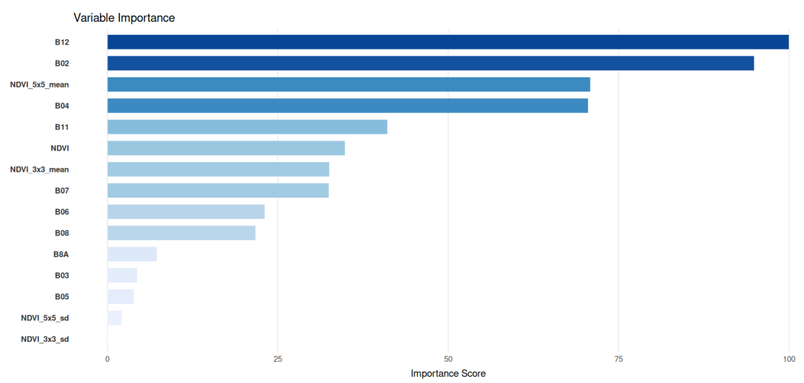
The preprocessing step prepares the multi-band Sentinel-2 image stack used for the forest/non-forest classification. The workflow includes:

1. **Data:** Sentinel-2 bands (B02–B12) at 10m and 20m resolution were loaded from a single acquisition date.
2. **Resampling and Reprojection:** All 20m bands were resampled to 10m resolution to ensure consistency across all layers and the image stack was projected to the UTM coordinate system (EPSG:25832).
3. **NDVI and NDVI Texture Metrics Calculation:** The NDVI was computed from the NIR (B08) and Red (B04) bands and added as an additional feature. Additionally spatial textures of NDVI were calculated using 3×3 and 5×5 moving windows (standard deviation and mean) and added as additional fearures for model training.

Features:

Band / Feature	Short Name	Description
Blue*	**B02**	Visible blue band (490 nm) – useful for water body and atmospheric analysis.
Green	B03	Visible green band (560 nm) – reflects vegetation vigor and soil background.
Red*	**B04**	Visible red band (665 nm) – useful for vegetation discrimination and NDVI calculation.
Red Edge 1	B05	First red edge band (705 nm) – sensitive to chlorophyll content in vegetation.
Red Edge 2	B06	Second red edge band (740 nm) – helps detect vegetation stress.
Red Edge 3	B07	Third red edge band (783 nm) – supports vegetation classification and biophysical retrieval.
Near Infrared	B08	NIR band (842 nm) – highly reflective for healthy vegetation; used in NDVI.
Red Edge 4	B8A	Narrow NIR band (865 nm) – useful for fine vegetation differentiation.
SWIR 1*	**B11**	Short-wave infrared band (1610 nm) – useful for vegetation moisture and soil discrimination.
SWIR 2*	**B12**	Short-wave infrared band (2190 nm) – helpful in land cover and fire mapping.
NDVI	NDVI	Normalized Difference Vegetation Index – indicator of vegetation greenness and biomass.
NDVI Standard Deviation (3×3)	NDVI_3x3_sd	Texture: spatial heterogeneity in NDVI over a 3×3 kernel.
NDVI Standard Deviation (5×5)	NDVI_5x5_sd	Texture: spatial heterogeneity in NDVI over a 5×5 kernel.
NDVI Mean (3×3)	NDVI_3x3_mean	Texture: local mean NDVI over a 3×3 window – captures patch-level vegetation.
NDVI Mean (5×5)*	**NDVI_5x5_mean**	Texture: local mean NDVI over a 5×5 window – useful for forest homogeneity.

*Features used for the final random forest model (after feature selection)



Variable importance of the random-forest model fitted to the whole feature stack.