DK Forest LiDAR v0.9.1

Classifications of Denmark's forest quality using the EcoDes-DK15 dataset (https://github.com/jakobjassmann/ecodes-dk-lidar) and other spatial data.

Disclaimer: This project is under development and not yet peer-reviewed.

Project overview

- Workflow Overview (workflow.html)

Data description

- Forest annotations and training data (to be updated) (training_annotations.html)
- Predictor overview (data_overview.html)
- Focal (window) predictor selection (focal_var_selection.html)

Model performance

- Gradient Boosting performance (gbm_models_performance.html)
- Random Forest performance (ranger_models_performance.html)

Results

- Leaflet web app (map of projections) (data_vis.html)
- Summary stats (area estimates) (summary_stats.html)

Data / Outputs

- Summary report - website snapshot v0.1.0 (2.2 MB, PDF)) (Assmann_et_al-DK_Forest_Quality_Report_v0.1.0.pdf)
- Gradient Boosting Projections v0.9.1 (41.5 MB, GeoTiff) (https://dkforestlidar2022.s3.eu-central-1.amazonaws.com/forest_quality_gbm_biowide_cog_epsg3857_v0.9.1.tif)
- Random Forest Projections v0.9.1 (40.2 MB, GeoTiff) (https://dkforestlidar2022.s3.eu-central-1.amazonaws.com/forest_quality_ranger_biowide_cog_epsg3857_v0.9.1.tif)
- Disturbance map v0.9.1 (17.8 MB, GeoTiff) (https://dkforestlidar2022.s3.eu-central-1.amazonaws.com/disturbance_since_2015_cog_epsg3857_v0.1.0.tif)
- Training Polygons v0.9.0 (115.7 MB, GeoJson) (https://dkforestlidar2022.s3.eu-central-1.amazonaws.com/training_polygons_v0.9.0.geojson)

Auxiliary

- Guide on how to visualise cloud optimised rasters (cog_guide.html)

AARHUS UNIVERSITY

SustainScapes
Center for Sustainable Landscapes under Global Change

[last update: 3 March 2022]

Workflow Overview

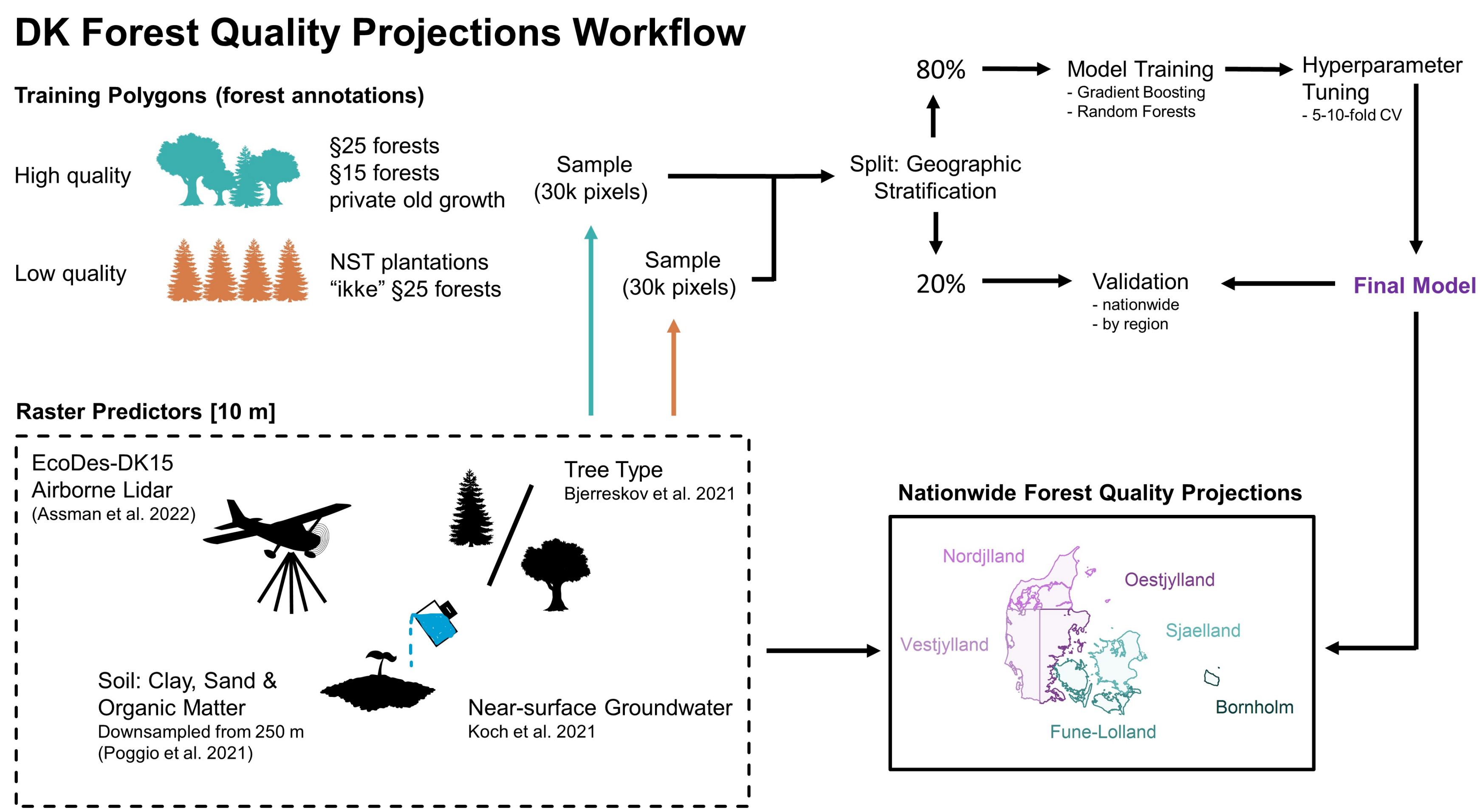
Jakob J. Assmann

09/08/2022

This document provides an overview on the workflow that we used to generate the models for predicting forest quality in Denmark.

In brief:

1. We gathered raster predictors with 10 m res. that we deemed meaningful for predicting the quality of forests in Denmark.
2. We gathered ~20k annotations for forests with high and low quality in Denmark.
3. We generated a training dataset of 60k pixels that fell within the annotated forests.
4. We split the training dataset 80%/20% prior model training using a geographic stratification.
5. We trained Gradient Boosting and Random Forest models.
6. We tuned the model hyperparameters using 5 or 10-fold cross validation based on the training dataset from the 80%/20% split.
7. We tested the final model performance on the validation dataset from the 80%/20% split.
8. We projected the forest quality across the whole of Denmark using the final models and the predictor rasters.



DK Forest LiDAR - Forest Annotations & Training Data

Jakob J. Assmann

09/08/2022

This document provides an overview of the forest annotations used for generating the training dataset that forms the base of our forest quality models for Denmark.

These annotations are vector polygons of forests in Denmark that are of known "high" or "low" quality. We used these polygons to generate a training dataset of 60k pixels based on the 10 m grid of Denmark that is used by our models. The grid is defined by the EcoDes-DK15 dataset. A brief description on how the final pixel training dataset was generated from the forest annotations can be found at the end of this document.

Note: What makes a "high" or "low" quality forest is to some degree arbitrary. Our definitions here are the result of a long discussion and have been developed over multiple years. The aim was to arrive at a workable definition that aligns with the current framework of forest designations in Denmark, while also ensuring that enough training data is available. We appreciate that our chosen definitions of forest quality are a simplification and not without flaws (e.g., we assume that all "plantations" are of low forest quality).

High quality forests

Total number of high quality forest polygons: 9400.

§25 and §15 forest

The core of the high quality forest annotations is made up by the polygons for the designated §25 ("naturmaessigt saerlig vaerdifuld skov") and §15 ("skovnatur") forests. The vector boundaries of these forests were retrieved from "Danmarks Miljøportal" (https://arealinformation.miljoeportal.dk/ (https://arealinformation.miljoeportal.dk/)):

- p25_offentligareal.shp (§25 forests, accessed on on 5 April 2019
- skov_kortlaegning_2016_2018.shp (§15 forests, accessed 24 September 2019).

Number of forests: 9044 (§25 forests: 2906; and §15 forests: 6138).

Untouched forests and "aftaler om natur"

The two other components of the high quality forest annotations are vector boundaries from the untouched forests (private and public), as well as areas with agreements on nature ("aftaler om natur"). The vector boundaries of these areas were retrieved from "Miljøgis - Ansøgning om skovtilskud for private" (https://miljoegis3.mim.dk/spatialmapsecure?profile=privatskovtilskud (https://miljoegis3.mim.dk/spatialmapsecure?profile=privatskovtilskud)):

- tilsagn17_st_uroert_skov_privat_tilskud.shp (untouched forests, accessed on 6 July 2021)
- tilsagn18_st_uroert_skov.shp (untouched forests, accessed on 6 July 2021)
- tilsagn19_st_uroert_skov_privat_tilskud.shp (untouched forests, accessed on 6 July 2021)
- tilsagn20_st_uroert_skov_privat_tilskud.shp (untouched forests, accessed on 6 July 2021)
- aftale_natur_tinglyst.shp (agreements on nature, accessed on 6 July 2021)

Number of forests: 356 ("untouched": 118; "aftaler om natur": 238).

Low quality forests

Total number of low quality forest polygons: 10697.

"Ikke" §25 forests

These forests are forests that were considered for being designated as §25 forests, but did not meet the requirements (e.g., after completion of the field survey). The vector geometries for these forests were shared with us by Bjarne Aabrandt Jensen (Miljøstyrelsen) in a personal communication on 19 November 2019.

- ikkeP25_skov.shp (personal communication, 19 November 2019)

Number of forests: 5848.

NST plantations

These forests are plantations owned by Denmark's environment agency "Naturstyrelsen" (NST). The vector geometries and auxiliary data for these forests were obtained by personal communication from Bjørn Ole Ejlersen at NST to Pil Pedersen on 11 June 2020.

The source dataset includes all forests owned by NST. To subset only forests that are plantations, we filtered the data by excluding all forests that had an "ANV 4" value of 1, were classified as "urørt" or designated as "historical". We then sub-sampled the plantations to ensure a balanced training dataset between high and low quality forests (target :~10k high & ~10k low quality forests). We drew a sample of 5000 plantations. To account for variation within stand ages, we stratified the sample based on the following stand ages classes (years): [0, 10], (10, 25], (25, 50], (50, 75], (75, ∞). For each stand age we drew 1000 forests at random. Not all forests that were drawn in the sample had an associated polygon in the separate vector geometry file (n missing = 151), these forests were not included in the final NST plantation subset.

- NST 2019 08012019 ber 16012020 til bios_au.xlsx (NST forests data table)
- LitraPolygoner_region.shp (NST forests polygons)

Number of forests: 4849.

Cleaning and preparation of geometries

We observed some overlap between the assembled annotations for high and low quality forests, both duplicate mappings within each category, but also some duplicate mapping of parts of forests as high and low quality. These inaccuracies were expected given the extent of the dataset and the fact that it was assembled from multiple sources. To address the issue we carried out a systematic cleaning of the geometries (fully reproducible through our source code).

First, we removed internal overlap within the high quality geometries. For this we iteratively removed all internal overlap starting with the p25 forests (internally sorted in arbitary order^), followed by the private old growth and lastly the p15 forests. We also buffered the geometries, using a negative buffer of 10 cm to avoid line-overlaps due to geometric inaccuracies. Finally, we removed one remaining p15 forest that failed to be filtered out.

Second, we removed all internal overlap within the low quality geometries, taking the same approach as for the high quality geometries except working fully in an arbitrary order. While doing that we also checked for overlap with any high quality geometries and if that was the case removed those. We also buffered the geometries, using a negative buffer of 10 cm to avoid line-overlaps due to geometric inaccuracies. Finally, we reomved one remaining ikke_p25 forest that was dublicated in the dataset.

The resulting dataset consists of two type of forest annotations (8915 high and 9720 low quality polygons) with no internal overlap within or between categories.

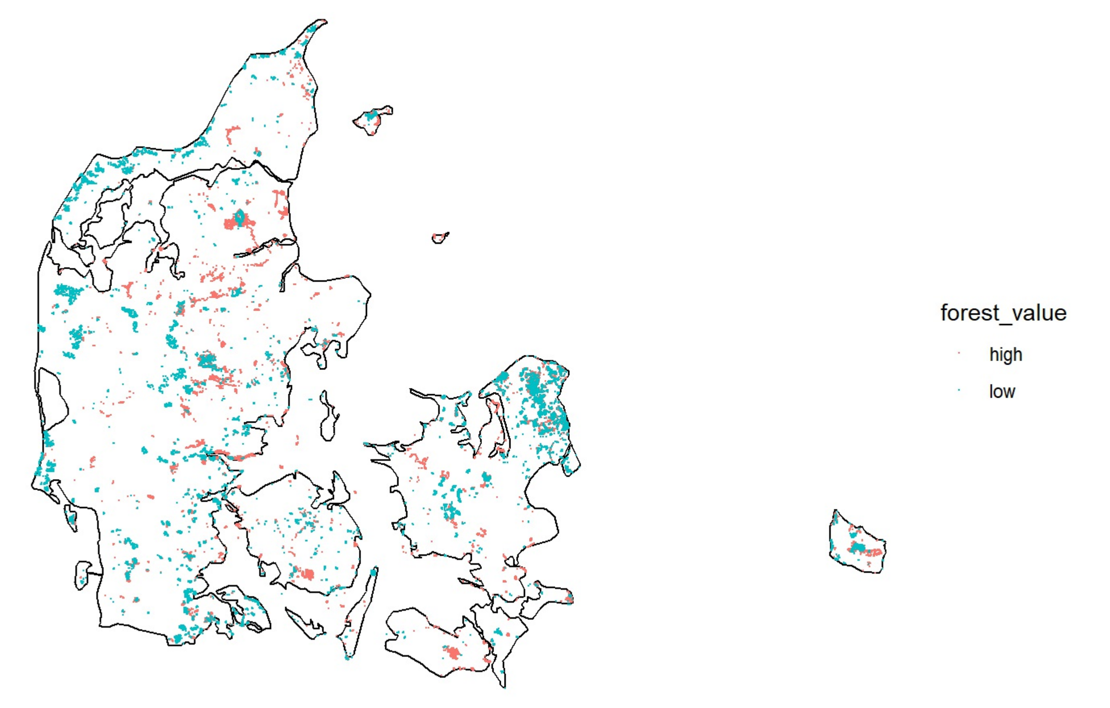
Pixel training dataset

Pixel sampling

We used the forest annotations (8915 high quality forests and 9720 low quality forests) to generate a sample of 60k pixels based on the EcoDes-DK14 grid to train our forest quality models.

The EcoDes-DK15 dataset uses a version of the Danish national grid that divides terrestrial Denmark into 10 m x 10 m cells / pixels (UTM32). For the training dataset we drew a sample of 30k pixels each from within the high quality and the low quality forest polygons. Specifically, the sample was based on the EcoDes-DK15 "dtm10_m" descriptor raster). The sample was drawn in the following fashion: First, we drew a random pixel from within each forest polygon (high or low), we then filled in the missing number of pixels to make up 30k for each forest quality class (high or low). The filing step was done completely at random, drawing from all remaining pixels available per class. The final dataset of pixels therefore contained 30k unique pixels from each class.

Pixel sample for training



Note, that because of the sampling scheme some pixels will be drawn from the individual forest polygons and the sampled pixels are therefore to some degree not statistically independent (as would samples from neighbouring polygons). However, the aim of our project is not to carry out a statistical analysis, but to train a machine learning classifier for predicting forest quality. The reduction in independence of the samples is therefore not an issue, as long as it does not lead to over fitting of the models (which it did not). Instead sub-sampling of the forest polygons is actually desirable as it allows us to capture the variation within the forest polygons themselves and therefore increase the predictive capabilities of our models. We also include focal (window) predictors variables to account for within landscape-scale (100 m x 100 m) variation in our models.

Finally, we chose a sample size of 60k pixels as a compromise between available computing power and model output performance. There are approximately 56.3 million forest pixels in Denmark and the training sample therefore represents ~0.1% of the total forest area in the country.

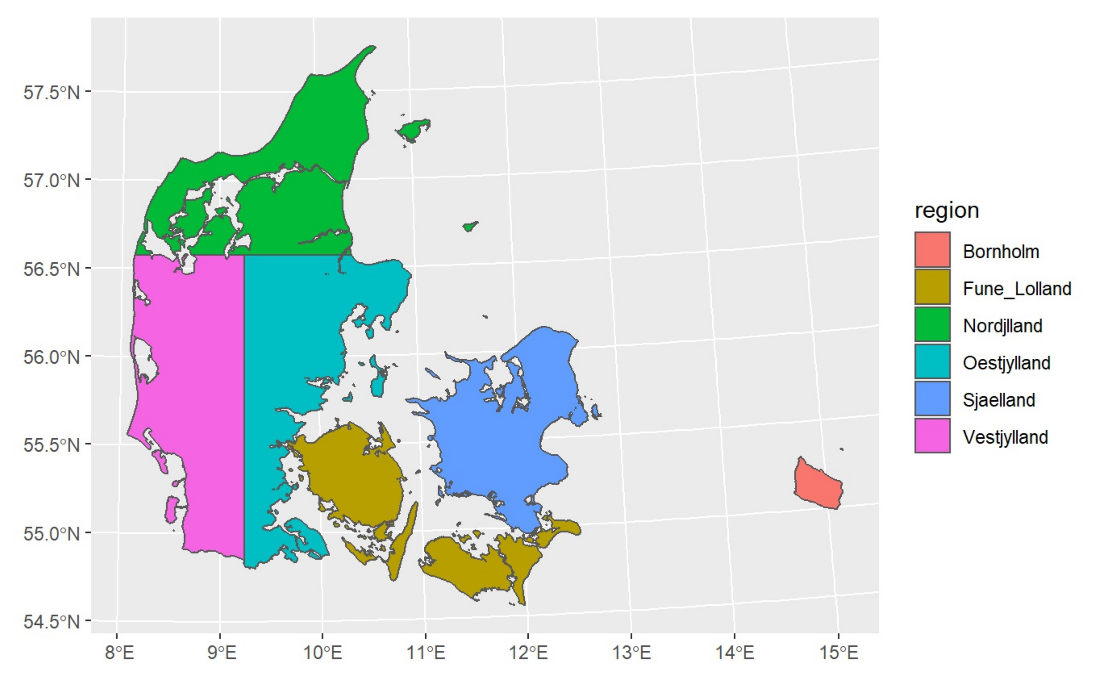
Training / validation split (geographic stratifiaction)

To allow for an independent validation of the model performance, we split the data 80/20 (training / validation) before carrying out the training. To account for potential geographical covariation in the training data we used a geographic stratification when carrying out the split. This means that the split was not conducted at random on the whole dataset, but randomly within regions (80/20 in each region). We used two different stratification schema of Denmark (see below).

Biowide stratification

This stratification was developed for the BIOWIDE project (Brunbjerg et al. 2019). The geometries are not publicly available and were kindly shared with us by Ane Brunbjerg (personal communication on 1 September 2021). Some further clean up of the geometries was required on our end. We had to make sure the boundaries of geometries were flush among neighbouring regions and that the coastlines were buffered.

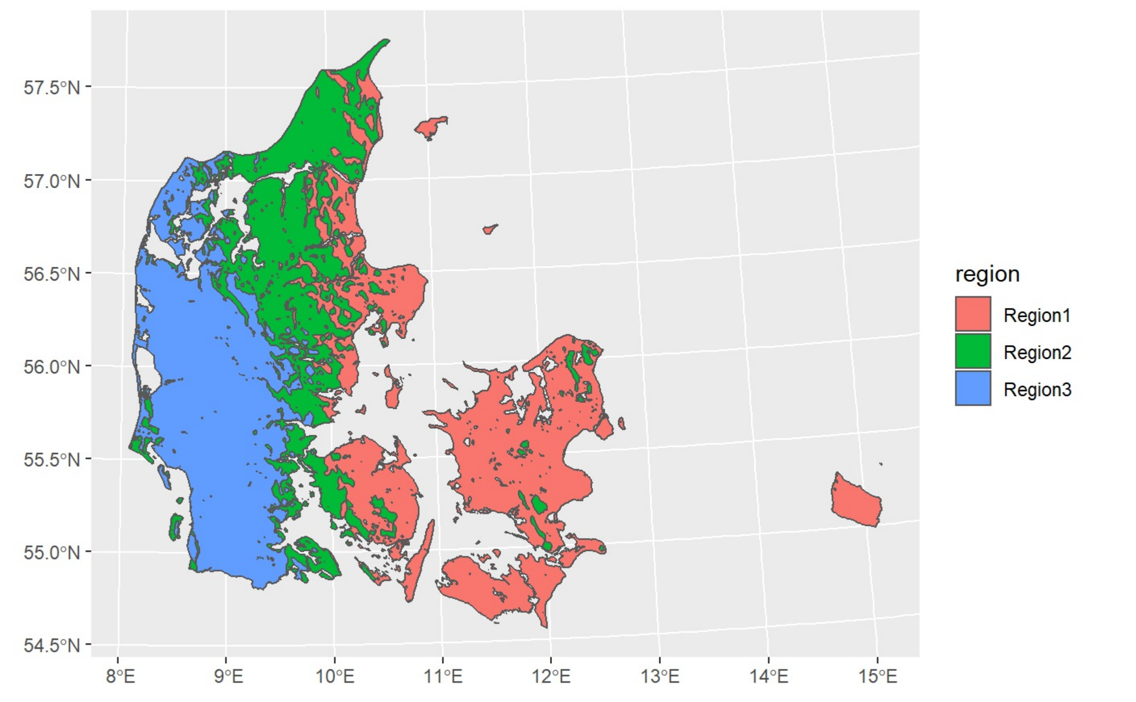
The stratification divides Denmark into six regions.



Derek's stratification

The second stratification was developed by co-author Derek Corcoran based on climatic and other ecological covariates.

It divides Denmark into three regions.



References

- Brunbjerg, A. K., Bruun, H. H., Brøndum, L., Classen, A. T., Dalby, L., Fog, K., Frøslev, T. G., Goldberg, I., Hansen, A. J., Hansen, M. D. D., Høye, T. T., Illum, A. A., Læssøe, T., Newman, G. S., Skipper, L., Søchting, U., & Ejrnæs, R. (2019). A systematic survey of regional multi-taxon biodiversity: Evaluating strategies and coverage. BMC Ecology, 19(1), 43. https://doi.org/10.1186/s12898-019-0260-x (https://doi.org/10.1186/s12898-019-0260-x)

DK Forest LiDAR - Predictor Data Overview

Jakob J. Assmann

09/08/2022

Predictor variables and selection

EcoDes-DK15 descriptors

The core of the predictor variables is formed by the EcoDes-DK15 rasterised lidar descriptors (Assmann et al. 2022) generated from the 2014/15 national airborne laser scanning campaign conducted by the Danish government.

From the 76 available EcoDes-DK15 layers (incl. auxiliary layers), we removed the date_stamp_xxx, point_count_xxx, point_source and building_proportion layers as we deemed those non-informative for the task of predicting forest quality. We kept the sea and water mask layers to try out sub-setting of the training data to make sure only land pixels are included, but discarded the mask layers later in the analysis.

Furthermore, we removed the following descriptors: canopy_openness, point_count, normalized_z_mean, heat_load_index, openness_mean, twi - as the ecological meaning of these was conceptually redundant with other descriptors (vegetation_density, canopy_height, solar_radiation, openness_difference and ground water respectively) and initial model runs indicated that these variables had a low predictive power. We also removed the aspect variable because it was a very weak predictor. This makes sense conceptually as the aspect at 10 m likely has little meaning on whether a forest cell is of high quality or not (all cardinal directions would theoretically be expected to be of high quality).

Finally, we removed all vegetation_proportion variables. These variables demonstrated a low predictive power by themselves. However, to capture the vertical variability in the lidar point cloud we calculated a foliage height diversity variable.

The final set of used EcoDes-DK15 variables is:

- amplitude_mean
- amplitude_sd
- canopy_height
- dtm_10m
- normalized_z_sd
- openness_difference
- slope
- solar_radiation
- vegetation_density

Foliage height diversity

To capture the vertical variation in the forest canopy we calcualted the “foliage height diversity” (MacArthur and MacArthur 1961) from the EcoDes-DK15 point proportion descriptors We followed the height bins used by Wilson (1974): 0 m – 1.5 m, 1.5 m – 9 m, and >9 m.

- foliage_height_diversity

Tree type predictor

As we expected that most common tree type (broadleaf vs. coniferous) would play an important role in determining if and why a forest is of high or low quality, we included the tree type projections generated by Bjerreskov et al. (2021).

The authors used a multi-temporal Sentinel 1/2 data fusion (SAR and optical) approach to assign forest types in a binary classification (broadleaf vs. coniferous).

As both types are mutually exclusive we discarded the “is confierous” variable after one-hot encoding of the source data. The source data is currently not publcily avialable, but was kindly shared with us by Thomas Nord-Larsen (senior author on Bjerreskov et al. 2021).

- treetype_bjer_dec

Soil predictors

Clay, sand and organic carbon content of soil

Soil type and composition are an important indicator in the key for the paragraph 25 forests. Here we used the following three predictors to account for differences in the soils across Denmark:

- Clay_utm32_10m
- Sand_utm32_10m
- Soc_utm32_10m

These data were obtained from the Soilgrids 2.0 dataset (Poggio et al. 2021). The original data layers were queried using the geodata package (Hijmans, Ghosh, and Mandel 2021) and subset to the extent of Denmark. The original data have a grain size of 250 m and are in a “Interrupted_Goode_Homolosine” projection. We projected them to the EcoDes-DK grid with 10 m grain size (UTM32N) using nearest neighbor resampling.

Note that the nearest neighbour resampling strategy is conservative and makes no assumption about the spatial distribution of the variables during the downsampling of the 250 m dataset. However, the downsampling may give the wrong impression that we have used higher-resolution predictor data than we actually have. Finally, the resampling will inevitably introduce some uncertainties where the downsampled grid and the orignal grid not align.

As a water mask had originally been applied to this data, we had not predictor data were a 250 m x 250 m pixel overlapped with a water body. This became a problem when extrapolating the models to the nationwide extent, as the finer grain size of our maps introduced more detailed shore lines. We therefore had 10 m x 10 m land pixels for which no soil data was available. To address this problem we gap-filled the original 250 m x 250 m soil data. All pixels that were NA and had at least one neighbouring cell that was not NA were filled with the mean of all neighbouring cells that were not NA. The raster was then projected to the EcoDEs-DK grid with 10 m grain size and only used for generating the nationwide forest quality maps from the trained models, but not for training of the models themselves. Forest Quality predictions close to some shores may therefore contain some error, but we are confident that this error is very small due to the inherently high autocorrelation of the soil variables.

Water availability

To account for the wetness of the forest ground and the water availability to the plants we use the summer near-surface ground water estimates by Koch et al. 2021.

- ns_groundwater_summer

Focal variables

To capture the spacial context around a pixel beyond the 10 m grid, we selected four key predictor variables and calculated their mean and variation (sd) for two window sizes of 110 m and 250 m around each pixel. We selected these window sizes as the best candidates based on variograms generated for all variables.

We conducted a collinearity analysis on the focal variables and reduced the v ariables in a step-wise selection process to the following final four focal variables included in the models:

- dtm_10m_sd_110m
- canopy_height_sd_110m
- vegetation_density_sd_110m
- ns_groundwater_summer_sd_110m

Additional documentation of the selection process can be found in the focal variable selection (focal_var_selection.html) document.

Overview table final predictor data sources

Here is an overview table of the final predictor data sources.

Predictor	Source Dataset	Ecological Meaning
amplitude_mean	EcoDes-DK15	Quality of lidar signal reflected (proxy of biomass).
amplitude_sd	EcoDes-DK15	Variation in quality of lidar signal reflected within 10 m pixel (proxy of variation in biomass).
canopy_height	EcoDes-DK15	Lidar estimator of canopy height (95-percentile of height distribution of all vegetation points in 10 m pixel).
canopy_height_sd_110m	EcoDes-DK15	Variation in lidar estimator of canopy height within 110 m focal window (11 x 11 pixels).
Clay_utm32_10m	Poggio et al. 2021	Estimated percentage clay content of soil (250 m resolution downscaled to 10 m).
dtm_10m	EcoDes-DK15	Terrain height above sea level.
dtm_10m_sd_110m	EcoDes-DK15	Variation in terrain height above sea level within 110 m focal window (11 x 11 pixels).
foliage_height_diversity	EcoDes-DK15	Foliage height diversity MacArthur and MacArthur (1979) based on height bins by Wilson (1974)
normalized_z_sd	EcoDes-DK15	Estimated variation in canopy height within 10 m pixel.
ns_groundwater_summer_sd_110m	Koch et al. 2021	Estimate of depth of near-surface groundwater during an average summer.
ns_groundwater_summer_utm32_10m	Koch et al. 2021	Variation in the estimate of depth of near-surface groundwater during an average summer within a 110 m focal window (11 x 11 pixels).
openness_difference	EcoDes-DK15	Presence of linear features in the terrain (valleys, ridges etc.) based on a 50 m search radius.
Sand_utm32_10m	Poggio et al. 2021	Estimated percentage sand content of soil (250 m resolution downscaled to 10 m).
slope	EcoDes-DK15	Terrain slope at 10 m
Soc_utm32_10m	Poggio et al. 2021	Estimated percentage soil organic carbon content of soil (250 m resolution downscaled to 10 m).
solar_radiation	EcoDes-DK15	Annual incident solar radiation based on terrain model (aspect and slope).
treetype_bjer_dec	Bjerreskov et al. 2021	Decidous or coniferous forest.
vegetation_density	EcoDes-DK15	Denisty of vegetation points in 10 m lidar pixel.
vegetation_density_sd_110m	EcoDes-DK15	Variation of density of vegeation points amongst pixels within 110 m window (11 x 11 pixels).

References

- Assmann, Jakob J., Jesper E. Moeslund, Urs A. Treier, and Signe Normand. “EcoDes-DK15: High-resolution ecological descriptors of vegetation and terrain derived from Denmark’s national airborne laser scanning data set.” Earth System Science Data Discussions (2021): 1-32. -Bjerreskov, K. S., Nord-Larsen, T., and Fensholt, R.: Classification of Nemoral Forests with Fusion of Multi-Temporal Sentinel-1 and 2 Data, 13, 950, https://doi.org/10.3390/rs13050950 (https://doi.org/10.3390/rs13050950), 2021.
- Hijmans, Robert J., Aniruddha Ghosh, and Alex Mandel. 2021. Geodata: Download Geographic Data. https://CRAN.R-project.org/package=geodata (https://CRAN.R-project.org/package=geodata). -Koch, J., Gotfredsen, J., Schneider, R., Troldborg, L., Stisen, S., and Henriksen, H. J.: High Resolution Water Table Modeling of the Shallow Groundwater Using a Knowledge-Guided Gradient Boosting Decision Tree Model, 3, 2021.
- MacArthur, R. H., & MacArthur, J. W. (1961). On Bird Species Diversity. Ecology, 42(3), 594–598. https://doi.org/10.2307/1932254 (https://doi.org/10.2307/1932254)
- Poggio, Laura, Luis M De Sousa, Niels H Batjes, Gerard Heuvelink, Bas Kempen, Eloi Ribeiro, and David Rossiter. 2021. “SoilGrids 2.0: Producing Soil Information for the Globe with Quantified Spatial Uncertainty.” Soil 7 (1): 217–40.
- Willson, M. F. (1974). Avian Community Organization and Habitat Structure. Ecology, 55(5), 1017–1029.

DK Forest LiDAR - Focal predictor selection

Jakob J. Assmann

02/03/2022

Content

We calculated the mean and sd in 110 m and 250 m windows for the following variables:

- dtm_10m
- canopy_height
- vegetation_density
- ns_ground_water

Here is how those measures are correlated with their focal variables:

canopy_height

	cell_10m	mean_110m	mean_250m	sd_110m	sd_250m
cell_10m	+1.00	+0.88	+0.80	+0.40	+0.56
mean_110m		+1.00	+0.95	+0.30	+0.53
mean_250m			+1.00	+0.28	+0.42
sd_110m				+1.00	+0.81
sd_250m					+1.00

dtm_10m

	cell_10m	mean_110m	mean_250m	sd_110m	sd_250m
cell_10m	+1.00	+1.00	+1.00	+0.30	+0.32
mean_110m		+1.00	+1.00	+0.30	+0.32
mean_250m			+1.00	+0.30	+0.33
sd_110m				+1.00	+0.92
sd_250m					+1.00

ns_groundwater_summer_mean_110m

	cell_10m	ns_groundwater_summer_mean_250m	ns_groundwater_summer_sd_110m	ns_groundwater_summer_sd_250m	ns_groundwater_summer_utm32_10m
cell_10m	+1.00	+0.97	+0.33	+0.38	+0.96
ns_groundwater_summer_mean_250m		+1.00	+0.36	+0.41	+0.91
ns_groundwater_summer_sd_110m			+1.00	+0.87	+0.31
ns_groundwater_summer_sd_250m				+1.00	+0.35
ns_groundwater_summer_utm32_10m					+1.00

vegetation_density

	cell_10m	mean_110m	mean_250m	sd_110m	sd_250m
cell_10m	+1.00	+0.82	+0.71	+0.17	+0.29
mean_110m		+1.00	+0.93	-0.03	+0.16
mean_250m			+1.00	-0.07	-0.00
sd_110m				+1.00	+0.76
sd_250m					+1.00

Variation Inflation Factors

To reduce the number of features systematically, we calculate variance inflation factors (vIFs). A VIF above 5 indicates that the variable introduces multicolliniearity in the dataset. A conservative rule is to only keep variables with VIFs below 2.5.

Here we carry out a step-wise selection based on the VIFs and the correlation tables above. VIFs exceeding 5 are highlighted in red.

1) All variables

Variables	VIF
canopy_height	6.66
canopy_height_mean_110m	33.43
canopy_height_mean_250m	25.65
canopy_height_sd_110m	6.12
canopy_height_sd_250m	8.93
dtm_10m	618.45
dtm_10m_mean_110m	1688.75
dtm_10m_mean_250m	511.75
dtm_10m_sd_110m	8.71
dtm_10m_sd_250m	8.92
ns_groundwater_summer_mean_110m	61.96
ns_groundwater_summer_mean_250m	28.76
ns_groundwater_summer_sd_110m	5.14
ns_groundwater_summer_sd_250m	5.27
ns_groundwater_summer_utm32_10m	19.16
vegetation_density	4.54
vegetation_density_mean_110m	23.77
vegetation_density_mean_250m	19.87
vegetation_density_sd_110m	5.33
vegetation_density_sd_250m	6.82

The mean variables seem to introduce a lot of collinearity (very high VIFs, and see correlation tables above). We drop them first.

2) Drop mean variables

Variables	VIF
canopy_height	2.76
canopy_height_sd_110m	5.84
canopy_height_sd_250m	7.42
dtm_10m	1.26
dtm_10m_sd_110m	7.76
dtm_10m_sd_250m	7.76
ns_groundwater_summer_sd_110m	5.09
ns_groundwater_summer_sd_250m	5.27
ns_groundwater_summer_utm32_10m	1.38
vegetation_density	1.72
vegetation_density_sd_110m	4.79
vegetation_density_sd_250m	5.06

The focal variables of different window sizes are highly correlated with each other. The correlation tables (above) suggest the 110 m windows are less correlated with the 10 m cell values, so we drop the 250 m windows next.

3) Drop 250 m variables

Variables	VIF
canopy_height	2.14
canopy_height_sd_110m	2.54
dtm_10m	1.25
dtm_10m_sd_110m	1.77
ns_groundwater_summer_sd_110m	1.5
ns_groundwater_summer_utm32_10m	1.39
vegetation_density	1.7
vegetation_density_sd_110m	2.19

The final set of variables includes only the 10 m cell values and the sd calculated for the 110 m windows.

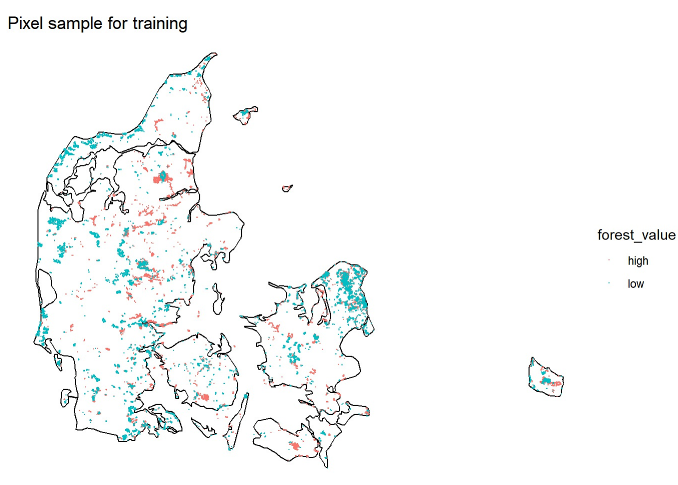
DK Forest LiDAR - Gradient Boosting Model Performance

Jakob Assmann

24/06/2022

Training data overview

I generated a training dataset consisting of 200k pixel samples from the EcoDes- DK15 grid. 100k samples each come from one of the two forest polygon sets ("low" and "high" forest value). I then extracted the predictor data for the pixel centres for those data.



For the training I then split the data set into a training (80%) and validation partition (20%). This split was done based on two stratification. The BIOWIDE stratification and Derek's stratification. The samples were split so that they were distributed evenly in each strata (80/20).

I then trained GBM models on this (currently minimal hyperparameter tuning) using the training dataset. This document evaluates the performance of each stratification based on the validation dataset in both stratification.

Models trained using BIOWIDE stratification

For these models the training data was split according to the BIOWIDE stratification.

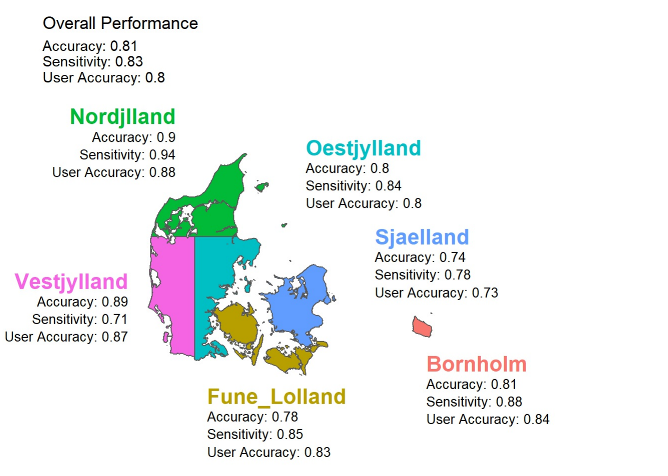
Variable importance

Variable importance for this boosted regression tree model.

```
##                                                                var  rel.inf
## dtm_10m                                                    dtm_10m 9.436428
## Sand_utm32_10m                                      Sand_utm32_10m 9.410692
## treetype_bjer_dec                                treetype_bjer_dec 7.586392
## Clay_utm32_10m                                      Clay_utm32_10m 6.717277
## Soc_utm32_10m                                        Soc_utm32_10m 6.288340
## ns_groundwater_summer_utm32_10m ns_groundwater_summer_utm32_10m 5.873385
## dtm_10m_sd_110m                                    dtm_10m_sd_110m 5.429778
## amplitude_sd                                          amplitude_sd 5.258751
## canopy_height                                        canopy_height 5.022070
## vegetation_density                              vegetation_density 4.780821
## canopy_height_sd_110m                        canopy_height_sd_110m 4.767433
## ns_groundwater_summer_sd_110m        ns_groundwater_summer_sd_110m 4.674164
## solar_radiation                                    solar_radiation 4.625719
## vegetation_density_sd_110m              vegetation_density_sd_110m 4.329235
## normalized_z_sd                                    normalized_z_sd 4.155538
## foliage_height_diversity                  foliage_height_diversity 3.908227
## amplitude_mean                                      amplitude_mean 3.760337
## slope                                                        slope 2.515989
## openness_difference                            openness_difference 1.459423
```

Performance in BIOWIDE regions:

Performance map based on the independent validation data:



Performance table based on the independent validation data:

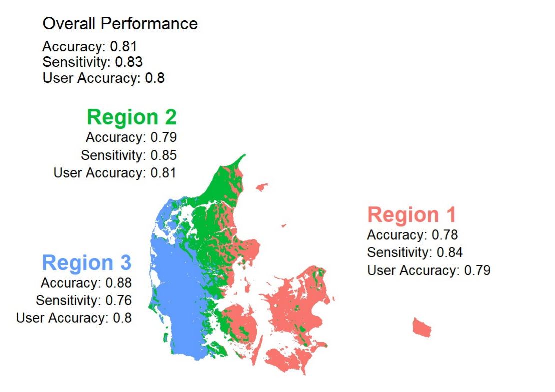
Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	0.81	0.81	0.78	0.90	0.80	0.74	0.89
Error	0.19	0.19	0.22	0.10	0.20	0.26	0.11
Sensitivity (True Positive Rate)	0.83	0.88	0.85	0.94	0.84	0.78	0.71
Specificity (True Negative Rate)	0.78	0.70	0.63	0.86	0.76	0.70	0.96
Fall-out (False Positive Rate)	0.22	0.30	0.37	0.14	0.24	0.30	0.04
Positive predictive value (User Accuracy)	0.80	0.84	0.83	0.88	0.80	0.73	0.87

Performance table based on the dependent training data:

Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	0.99	1.00	0.99	1.00	1.00	0.99	0.99
Error	0.01	0.00	0.01	0.00	0.00	0.01	0.01
Sensitivity (True Positive Rate)	0.99	1.00	1.00	0.99	0.99	0.99	0.97
Specificity (True Negative Rate)	0.99	0.99	0.99	1.00	1.00	0.99	1.00
Fall-out (False Positive Rate)	0.01	0.01	0.01	0.00	0.00	0.01	0.00
Positive predictive value (User Accuracy)	0.99	1.00	1.00	1.00	1.00	0.99	1.00

Perfromance in Derek's regions:

Performance map based on the independent validation data:



Performance table based on the independent validation data:

Measure	Overall	Region 1	Region 2	Region 3
Accuracy	0.81	0.78	0.79	0.88
Error	0.19	0.22	0.21	0.12
Sensitivity (True Positive Rate)	0.83	0.84	0.85	0.76
Specificity (True Negative Rate)	0.78	0.69	0.73	0.92
Fall-out (False Positive Rate)	0.22	0.31	0.27	0.08
Positive predictive value (User Accuracy)	0.80	0.79	0.81	0.80

Performance table based on the dependent training data:

Measure	Overall	Region 1	Region 2	Region 3
Accuracy	0.99	0.99	0.99	0.99
Error	0.01	0.01	0.01	0.01
Sensitivity (True Positive Rate)	0.99	0.99	0.99	0.98
Specificity (True Negative Rate)	0.99	0.99	0.99	1.00
Fall-out (False Positive Rate)	0.01	0.01	0.01	0.00
Positive predictive value (User Accuracy)	0.99	0.99	1.00	1.00

Performance by forest type (boradleaf vs. coniferous)

Performance table based on the independent validation data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	0.81	0.78	0.88
Error	0.19	0.22	0.12
Sensitivity (True Positive Rate)	0.83	0.86	0.63
Specificity (True Negative Rate)	0.78	0.68	0.95
Fall-out (False Positive Rate)	0.22	0.32	0.05
Positive predictive value (User Accuracy)	0.80	0.80	0.80

Performance table based on the dependent training data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	0.99	0.99	0.99
Error	0.01	0.01	0.01
Sensitivity (True Positive Rate)	0.99	1.00	0.97
Specificity (True Negative Rate)	0.99	0.99	1.00
Fall-out (False Positive Rate)	0.01	0.01	0.00
Positive predictive value (User Accuracy)	0.99	0.99	0.99

Models trained using Derek's stratification

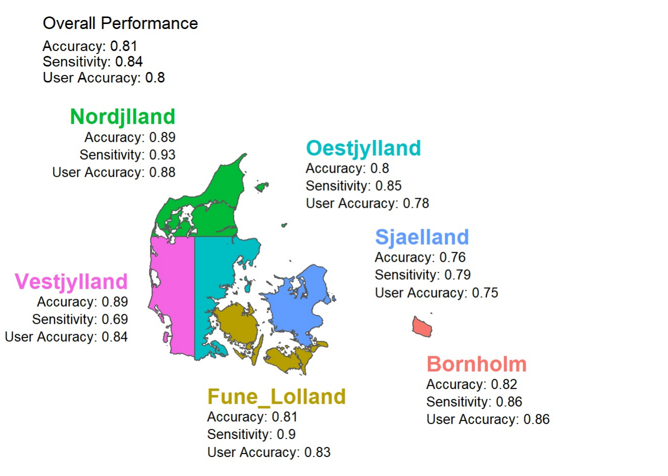
Variable importance

Variable importance for this boosted regression tree model.

```
##                                                                var  rel.inf
## Sand_utm32_10m                                      Sand_utm32_10m 9.292837
## dtm_10m                                                    dtm_10m 9.063782
## treetype_bjer_dec                                treetype_bjer_dec 7.359120
## Clay_utm32_10m                                      Clay_utm32_10m 6.703277
## ns_groundwater_summer_utm32_10m ns_groundwater_summer_utm32_10m 6.048256
## Soc_utm32_10m                                        Soc_utm32_10m 6.036391
## amplitude_sd                                          amplitude_sd 5.627812
## dtm_10m_sd_110m                                    dtm_10m_sd_110m 5.429703
## canopy_height                                        canopy_height 5.251818
## canopy_height_sd_110m                        canopy_height_sd_110m 4.772935
## ns_groundwater_summer_sd_110m        ns_groundwater_summer_sd_110m 4.552738
## vegetation_density_sd_110m              vegetation_density_sd_110m 4.520161
## solar_radiation                                    solar_radiation 4.495919
## vegetation_density                              vegetation_density 4.397915
## normalized_z_sd                                    normalized_z_sd 4.318571
## foliage_height_diversity                  foliage_height_diversity 4.116388
## amplitude_mean                                      amplitude_mean 3.919426
## slope                                                        slope 2.403913
## openness_difference                            openness_difference 1.689039
```

Performance in BIOWIDE regions:

Performance map based on the independent validation data:



Performance table based on the independent validation data:

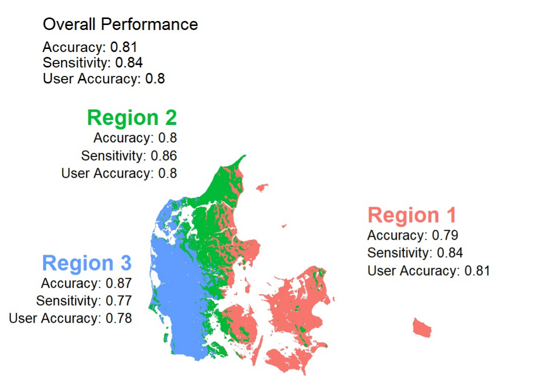
Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	0.81	0.82	0.81	0.89	0.80	0.76	0.89
Error	0.19	0.18	0.19	0.11	0.20	0.24	0.11
Sensitivity (True Positive Rate)	0.84	0.86	0.90	0.93	0.85	0.79	0.69
Specificity (True Negative Rate)	0.79	0.75	0.63	0.86	0.75	0.72	0.95
Fall-out (False Positive Rate)	0.21	0.25	0.37	0.14	0.25	0.28	0.05
Positive predictive value (User Accuracy)	0.80	0.86	0.83	0.88	0.78	0.75	0.84

Performance table based on the dependent training data:

Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	1	1	1	1	1	0.99	0.99
Error	0	0	0	0	0	0.01	0.01
Sensitivity (True Positive Rate)	1	1	1	1	1	0.99	0.98
Specificity (True Negative Rate)	1	1	1	1	1	0.99	1.00
Fall-out (False Positive Rate)	0	0	0	0	0	0.01	0.00
Positive predictive value (User Accuracy)	1	1	1	1	1	0.99	1.00

Perfromance in Derek's regions:

Performance map based on the independent validation data:



Performance table based on the independent validation data:

Measure	Overall	Region 1	Region 2	Region 3
Accuracy	0.81	0.79	0.80	0.87
Error	0.19	0.21	0.20	0.13
Sensitivity (True Positive Rate)	0.84	0.84	0.86	0.77
Specificity (True Negative Rate)	0.79	0.70	0.72	0.92
Fall-out (False Positive Rate)	0.21	0.30	0.28	0.08
Positive predictive value (User Accuracy)	0.80	0.81	0.80	0.78

Performance table based on the dependent training data:

Measure	Overall	Region 1	Region 2	Region 3
Accuracy	1	1.00	1	1.00
Error	0	0.00	0	0.00
Sensitivity (True Positive Rate)	1	1.00	1	0.99
Specificity (True Negative Rate)	1	0.99	1	1.00
Fall-out (False Positive Rate)	0	0.01	0	0.00
Positive predictive value (User Accuracy)	1	1.00	1	1.00

Performance by forest type (boradleaf vs. coniferous)

Performance table based on the independent validation data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	0.81	0.95	0.97
Error	0.19	0.05	0.03
Sensitivity (True Positive Rate)	0.84	0.97	0.90
Specificity (True Negative Rate)	0.79	0.93	0.99
Fall-out (False Positive Rate)	0.21	0.07	0.01
Positive predictive value (User Accuracy)	0.80	0.96	0.96

Performance table based on the dependent training data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	1	0.95	0.97
Error	0	0.05	0.03
Sensitivity (True Positive Rate)	1	0.97	0.90
Specificity (True Negative Rate)	1	0.93	0.99
Fall-out (False Positive Rate)	0	0.07	0.01
Positive predictive value (User Accuracy)	1	0.95	0.96

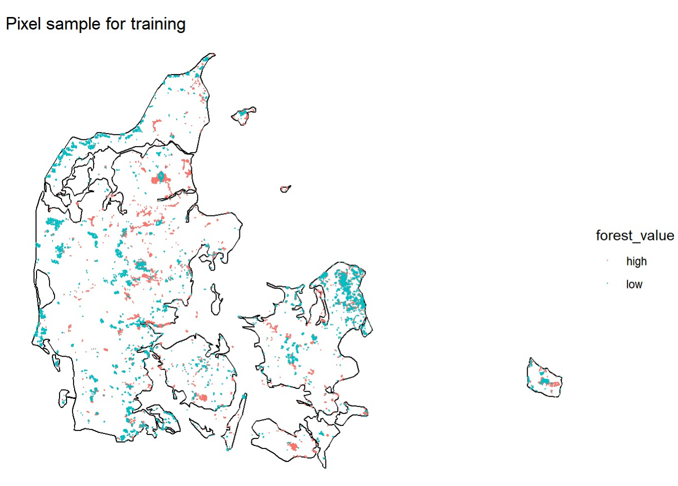
DK Forest LiDAR - Random Forest Model Performance

Jakob Assmann

24/06/2022

Training data overview

I generated a training dataset consisting of 200k pixel samples from the EcoDes- DK15 grid. 100k samples each come from one of the two forest polygon sets ("low" and "high" forest value). I then extracted the predictor data for the pixel centres for those data.



For the training I then split the data set into a training (80%) and validation partition (20%). This split was done based on two stratification. The BIOWIDE stratification and Derek's stratification. The samples were split so that they were distributed evenly in each strata (80/20).

I then trained GBM models on this (currently minimal hyperparameter tuning) using the training dataset. This document evaluates the performance of each stratification based on the validation dataset in both stratification.

Models trained using BIOWIDE stratification

For these models the training data was split according to the BIOWIDE stratification.

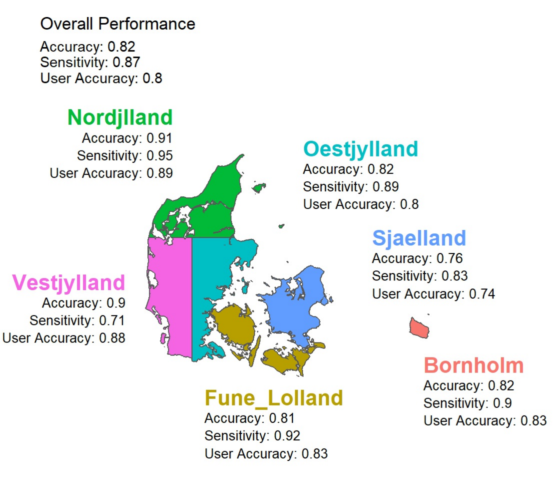
Variable importance

Variable importance for this random forest model, determined using the "permutation" option in ranger.

	Overall
Sand_utm32_10m	100.000000
treetype_bjer_dec	82.402128
dtm_10m	77.419004
Clay_utm32_10m	70.345552
dtm_10m_sd_110m	57.163281
ns_groundwater_summer_utm32_10m	56.787379
Soc_utm32_10m	38.041657
amplitude_sd	37.983167
canopy_height	36.898430
normalized_z_sd	35.367769
openness_difference	30.409109
slope	20.910563
ns_groundwater_summer_sd_110m	16.561303
vegetation_density	15.630589
amplitude_mean	13.397601
solar_radiation	12.448666
canopy_height_sd_110m	4.933190
vegetation_density_sd_110m	3.406289
foliage_height_diversity	0.000000

Performance in BIOWIDE regions:

Performance map based on the independent validation data:



Performance table based on the independent validation data:

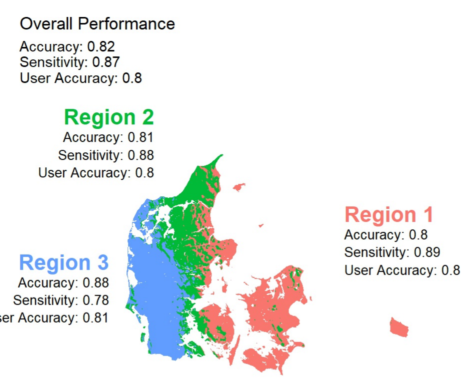
Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	0.82	0.82	0.81	0.91	0.82	0.76	0.90
Error	0.18	0.18	0.19	0.09	0.18	0.24	0.10
Sensitivity (True Positive Rate)	0.87	0.90	0.92	0.95	0.89	0.83	0.71
Specificity (True Negative Rate)	0.78	0.69	0.60	0.86	0.74	0.69	0.96
Fall-out (False Positive Rate)	0.22	0.31	0.40	0.14	0.26	0.31	0.04
Positive predictive value (User Accuracy)	0.80	0.83	0.83	0.89	0.80	0.74	0.88

Performance table based on the dependent training data:

Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	1	1	1	1	1	1	1
Error	0	0	0	0	0	0	0
Sensitivity (True Positive Rate)	1	1	1	1	1	1	1
Specificity (True Negative Rate)	1	1	1	1	1	1	1
Fall-out (False Positive Rate)	0	0	0	0	0	0	0
Positive predictive value (User Accuracy)	1	1	1	1	1	1	1

Performance in Derek's regions:

Performance map based on the independent validation data:



Performance table based on the independent validation data:

Measure	Overall	Region 1	Region 2	Region 3
Accuracy	0.82	0.80	0.81	0.88
Error	0.18	0.20	0.19	0.12
Sensitivity (True Positive Rate)	0.87	0.89	0.88	0.78
Specificity (True Negative Rate)	0.78	0.68	0.71	0.92
Fall-out (False Positive Rate)	0.22	0.32	0.29	0.08
Positive predictive value (User Accuracy)	0.80	0.80	0.80	0.81

Performance table based on the dependent training data:

Measure	Overall	Region 1	Region 2	Region 3
Accuracy	1	1	1	1
Error	0	0	0	0
Sensitivity (True Positive Rate)	1	1	1	1
Specificity (True Negative Rate)	1	1	1	1
Fall-out (False Positive Rate)	0	0	0	0
Positive predictive value (User Accuracy)	1	1	1	1

Performance by forest type (boradleaf vs. coniferous)

Performance table based on the independent validation data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	0.82	0.81	0.87
Error	0.18	0.19	0.13
Sensitivity (True Positive Rate)	0.87	0.91	0.57
Specificity (True Negative Rate)	0.78	0.66	0.96
Fall-out (False Positive Rate)	0.22	0.34	0.04
Positive predictive value (User Accuracy)	0.80	0.80	0.83

Performance table based on the dependent training data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	1	1	1
Error	0	0	0
Sensitivity (True Positive Rate)	1	1	1
Specificity (True Negative Rate)	1	1	1
Fall-out (False Positive Rate)	0	0	0
Positive predictive value (User Accuracy)	1	1	1

Models trained using Derek's stratification

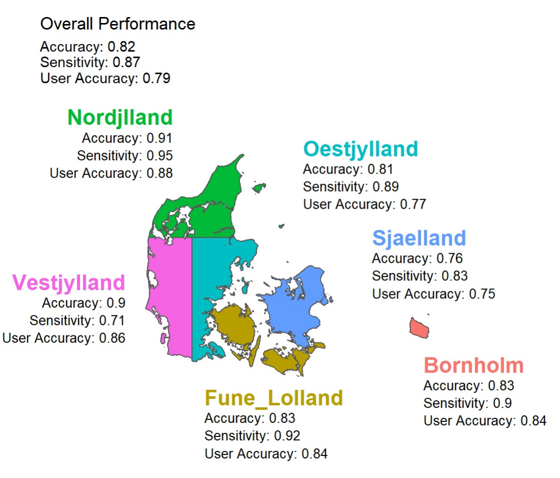
Variable importance

Variable importance for this random forest model, determined using the "permutation" option in ranger.

	Overall
Sand_utm32_10m	100.000000
treetype_bjer_dec	81.570818
dtm_10m	73.198738
Clay_utm32_10m	69.218368
dtm_10m_sd_110m	58.622528
ns_groundwater_summer_utm32_10m	56.875166
amplitude_sd	39.117492
Soc_utm32_10m	37.597694
canopy_height	36.925847
normalized_z_sd	34.734356
openness_difference	30.391050
slope	19.174338
ns_groundwater_summer_sd_110m	16.546973
amplitude_mean	14.016066
vegetation_density	13.985474
solar_radiation	11.460056
canopy_height_sd_110m	4.549356
vegetation_density_sd_110m	3.270389
foliage_height_diversity	0.000000

Performance in BIOWIDE regions:

Performance map based on the independent validation data:



Performance table based on the independent validation data:

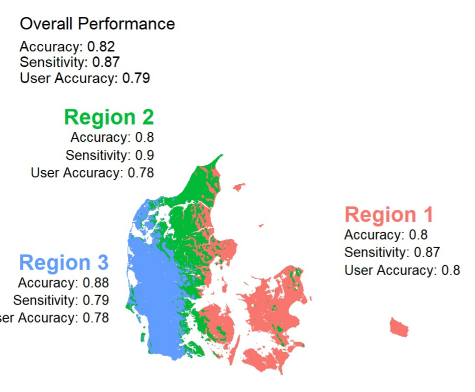
Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	0.82	0.83	0.83	0.91	0.81	0.76	0.90
Error	0.18	0.17	0.17	0.09	0.19	0.24	0.10
Sensitivity (True Positive Rate)	0.87	0.90	0.92	0.95	0.89	0.83	0.71
Specificity (True Negative Rate)	0.77	0.70	0.64	0.85	0.71	0.70	0.96
Fall-out (False Positive Rate)	0.23	0.30	0.36	0.15	0.29	0.30	0.04
Positive predictive value (User Accuracy)	0.79	0.84	0.84	0.88	0.77	0.75	0.86

Performance table based on the dependent training data:

Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	1	1	1	1	1	1	1
Error	0	0	0	0	0	0	0
Sensitivity (True Positive Rate)	1	1	1	1	1	1	1
Specificity (True Negative Rate)	1	1	1	1	1	1	1
Fall-out (False Positive Rate)	0	0	0	0	0	0	0
Positive predictive value (User Accuracy)	1	1	1	1	1	1	1

Perfromance in Derek's regions:

Performance map based on the independent validation data:



Performance table based on the independent validation data:

Measure	Overall	Region 1	Region 2	Region 3
Accuracy	0.82	0.80	0.80	0.88
Error	0.18	0.20	0.20	0.12
Sensitivity (True Positive Rate)	0.87	0.87	0.90	0.79
Specificity (True Negative Rate)	0.77	0.69	0.68	0.92
Fall-out (False Positive Rate)	0.23	0.31	0.32	0.08
Positive predictive value (User Accuracy)	0.79	0.80	0.78	0.78

Performance table based on the dependent training data:

Measure	Overall	Region 1	Region 2	Region 3
Accuracy	1	1	1	1
Error	0	0	0	0
Sensitivity (True Positive Rate)	1	1	1	1
Specificity (True Negative Rate)	1	1	1	1
Fall-out (False Positive Rate)	0	0	0	0
Positive predictive value (User Accuracy)	1	1	1	1

Performance by forest type (boradleaf vs. coniferous)

Performance table based on the independent validation data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	0.82	0.80	0.88
Error	0.18	0.20	0.12
Sensitivity (True Positive Rate)	0.87	0.91	0.60
Specificity (True Negative Rate)	0.77	0.65	0.96
Fall-out (False Positive Rate)	0.23	0.35	0.04
Positive predictive value (User Accuracy)	0.79	0.79	0.80

Performance table based on the dependent training data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	1	1	1
Error	0	0	0
Sensitivity (True Positive Rate)	1	1	1
Specificity (True Negative Rate)	1	1	1
Fall-out (False Positive Rate)	0	0	0
Positive predictive value (User Accuracy)	1	1	1

DK Forest LiDAR Summary Stats for Projections

Jakob Johann Assmann

09/08/2022

This document provides summary stats for the forest quality projections. Here, we concentrate on the models (gradient boosting and random forests) trained with the “BIOWIDE” stratification as these performed best overall.

Content:

1. Forest area in Denmark according to Basemap 03.
2. Disturbance detected in forests overall.
3. Gradient Boosting model summary tats.
4. Random Forest model summary stats.

Forests in Denmark according to Basemap 03

The forest mask used for our projections is based on the DCE Basemap 03 sub-layer “tree cover” for 2016 (Levin 2019) (https://dce2.au.dk/pub/TR159.pdf).

The sub-layer contains five “object types”: 1) tree cover, 2) forest / afforestation, 3) Christmas trees / cut greenery, 4) nursery / plantation, and 5) energy forest.

The table below shows how much area each of the classes cover in the layer (see also Table 4.3, Levin 2019):

Code	Name	Area [km²]	Proportion [%]
1	tree cover	928.3	13.5
2	forest / afforestation	5633.9	81.9
3	Christmas trees / cut greenery	176.2	2.6
4	nursery / plantation	46.8	0.7
5	energy forest	91.4	1.3
•	total	6876.5	100.0

For our projections **we only use the “forest / afforestation” layer (2)**.

To match the grid of the EcoDes-DK15 rasters we had to project the forest mask. For this we used a nearest neighbour algorithm. Here we simply confirm that the forest area (code 2) in the final mask “forest_mask.tif” matches the area noted in the table above.

Layer	Area [km²]
forest mask	5633.89

Disturbance overall

We used a disturbance layer generated by Cornelius (Senf and Seidl 2021) (https://zenodo.org/record/4746129) to estimate the disturbance in Denmark’s forests since the lidar data for EcoDes-DK15 was collected.

Please note that this disturbance mask was projected and down-sampled from a 30 m Landsat grid to the 10 m EcoDes-DK15 grid (nearest neighbour algorithm), potentially adding small uncertainties to the area estimates. Currently, we also only account for disturbances from 2016 till 2020.

Name	Area [km²]	Proportion [%]
disturbed forest	74.48	1.30
total forest	5633.89	100.00

Gradient Boosting projections summary stats

This gradient boosting model was trained based on the “BIOWIDE” stratification.

Type	Area [km²]	Proportion [%]
high quality forest	1409.32	25.00
low quality forest	3238.56	57.50
total forest	5633.89	100.00

Disturbance statistics:

Type	Area [km²]	Proportion [%]
disturbed high quality forest	15.74	1.10
total high quality forest	1409.32	100.00

Type	Area [km²]	Proportion [%]
disturbed low quality forest	58.74	1.80
total low quality forest	3238.56	100.00

Type	Area [km²]	Proportion [%]
disturbed high quality forest	15.74	21.10
disturbed low quality forest	58.74	78.90
total disturbed forest	74.48	100.00

Random Forest projections summary stats

This random forest model was trained based on the “BIOWIDE” stratification.

Type	Area [km²]	Proportion [%]
high quality forest	1434.90	25.50
low quality forest	3212.98	57.00
total forest	5633.89	100.00

Disturbance statistics:

Type	Area [km²]	Proportion [%]
disturbed high quality forest	15.32	1.10
total high quality forest	1434.90	100.00

Type	Area [km²]	Proportion [%]
disturbed low quality forest	59.16	1.80
total low quality forest	3212.98	100.00

Type	Area [km²]	Proportion [%]
disturbed high quality forest	15.32	20.60
disturbed low quality forest	59.16	79.40
total disturbed forest	74.48	100.00