DK Forest LiDAR v.0.1.0 (beta)

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Disclaimer: This project is still under development and not yet peer-reviewed.

Data description

- 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)
- Area summar stats for projections (summary_stats.html)

Data outputs

- Gradient Boosting Projections v0.1.0 (23 MB, GeoTiff) (https://dkforestlidar2022.s3.eu-central-1.amazonaws.com/forest_quality_gbm_biowide_cog_epsg3857_v0.1.0.tif)
- Random Forest Projections v0.1.0 (37 MB, GeoTiff) (https://dkforestlidar2022.s3.eu-central-1.amazonaws.com/forest_quality_ranger_biowide_cog_epsg3857_v0.1.0.tif)
- Disturbance map v0.1.0 (36 MB, GeoTiff) (https://dkforestlidar2022.s3.eu-central-1.amazonaws.com/disturbance_since_2015_cog_epsg3857_v0.1.0.tif)
- Training Polygons (44.3 MB, GeoJson) (https://dkforestlidar2022.s3.eu-central-1.amazonaws.com/training_polygons.geojson)

AARHUS UNIVERSITY

SustainScapes
Center for Sustainable Landscapes under Global Change

Aarhus University / VegDyn / SustainScapes

[last update: 2 March 2022]

DK Forest LiDAR - Predictor Data Overview

Jakob J. Assmann

02/03/2022

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.

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 - Gradien Boosting Model Performance

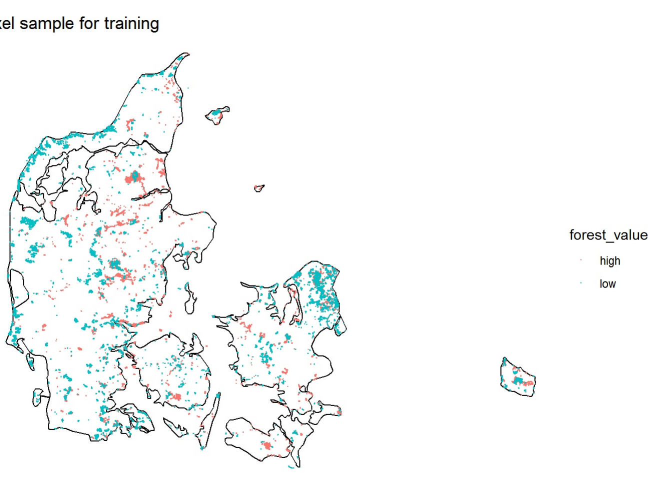
Jakob Assmann

02/03/2022

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.

Pixel sample for training



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.

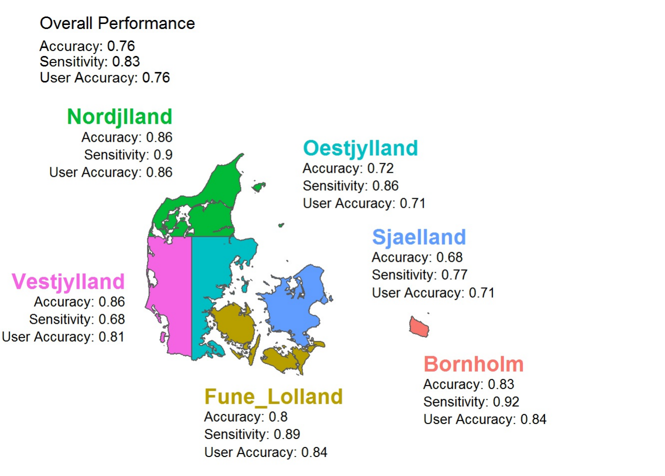
Variable importance

Variable importance for this boosted regression tree model.

```
##                                                           var   rel.inf
## treetype_bjer_dec                           treetype_bjer_dec 10.896051
## Sand_utm32_10m                                 Sand_utm32_10m  9.748158
## dtm_10m                                               dtm_10m  9.554361
## ns_groundwater_summer_utm32_10m ns_groundwater_summer_utm32_10m  6.315921
## dtm_10m_sd_110m                               dtm_10m_sd_110m  6.155180
## Clay_utm32_10m                                 Clay_utm32_10m  5.738498
## amplitude_sd                                     amplitude_sd  5.392395
## Soc_utm32_10m                                   Soc_utm32_10m  5.087160
## solar_radiation                               solar_radiation  4.656150
## canopy_height                                   canopy_height  4.484893
## canopy_height_sd_110m                   canopy_height_sd_110m  4.416347
## vegetation_density                         vegetation_density  4.257838
## normalized_z_sd                               normalized_z_sd  4.075467
## ns_groundwater_summer_sd_110m     ns_groundwater_summer_sd_110m  3.991426
## vegetation_density_sd_110m         vegetation_density_sd_110m  3.963466
## amplitude_mean                                 amplitude_mean  3.932958
## foliage_height_diversity             foliage_height_diversity  3.744206
## slope                                                   slope  2.575710
## openness_difference                       openness_difference  1.013815
```

Performance in BIOWIDE regions:

Performance map based on the independent validation data:



Performance table based on the independent validation data:

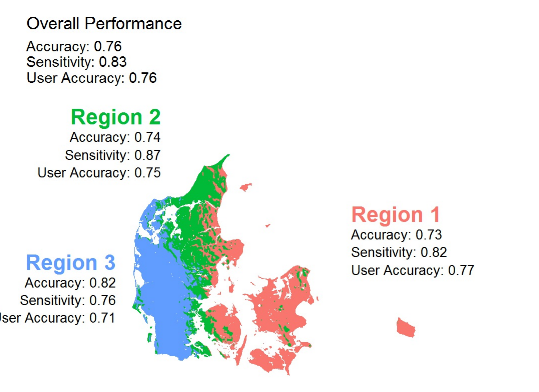
Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	0.76	0.83	0.80	0.86	0.72	0.68	0.86
Error	0.24	0.17	0.20	0.14	0.28	0.32	0.14
Sensitivity (True Positive Rate)	0.83	0.92	0.89	0.90	0.86	0.77	0.68
Specificity (True Negative Rate)	0.67	0.64	0.57	0.80	0.51	0.58	0.93
Fall-out (False Positive Rate)	0.33	0.36	0.43	0.20	0.49	0.42	0.07
Positive predictive value (User Accuracy)	0.76	0.84	0.84	0.86	0.71	0.71	0.81

Performance table based on the dependent training data:

Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	0.77	0.89	0.82	0.87	0.72	0.70	0.88
Error	0.23	0.11	0.18	0.13	0.28	0.30	0.12
Sensitivity (True Positive Rate)	0.84	0.96	0.90	0.91	0.86	0.77	0.72
Specificity (True Negative Rate)	0.69	0.76	0.62	0.82	0.53	0.61	0.94
Fall-out (False Positive Rate)	0.31	0.24	0.38	0.18	0.47	0.39	0.06
Positive predictive value (User Accuracy)	0.77	0.89	0.86	0.87	0.72	0.71	0.82

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	0.76	0.73	0.74	0.82
Error	0.24	0.27	0.26	0.18
Sensitivity (True Positive Rate)	0.83	0.82	0.87	0.76
Specificity (True Negative Rate)	0.67	0.57	0.54	0.85
Fall-out (False Positive Rate)	0.33	0.43	0.46	0.15
Positive predictive value (User Accuracy)	0.76	0.77	0.75	0.71

Performance table based on the dependent training data:

Measure	Overall	Region 1	Region 2	Region 3
Accuracy	0.77	0.75	0.75	0.84
Error	0.23	0.25	0.25	0.16
Sensitivity (True Positive Rate)	0.84	0.84	0.87	0.78
Specificity (True Negative Rate)	0.69	0.61	0.57	0.86
Fall-out (False Positive Rate)	0.31	0.39	0.43	0.14
Positive predictive value (User Accuracy)	0.77	0.79	0.76	0.72

Performance by forest type (boradleaf vs. coniferous)

Performance table based on the independent validation data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	0.76	0.74	0.82
Error	0.24	0.26	0.18
Sensitivity (True Positive Rate)	0.83	0.87	0.55
Specificity (True Negative Rate)	0.67	0.50	0.92
Fall-out (False Positive Rate)	0.33	0.50	0.08
Positive predictive value (User Accuracy)	0.76	0.76	0.71

Performance table based on the dependent training data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	0.77	0.75	0.84
Error	0.23	0.25	0.16
Sensitivity (True Positive Rate)	0.84	0.87	0.59
Specificity (True Negative Rate)	0.69	0.53	0.93
Fall-out (False Positive Rate)	0.31	0.47	0.07
Positive predictive value (User Accuracy)	0.77	0.77	0.74

Models trained using Derek’s stratification

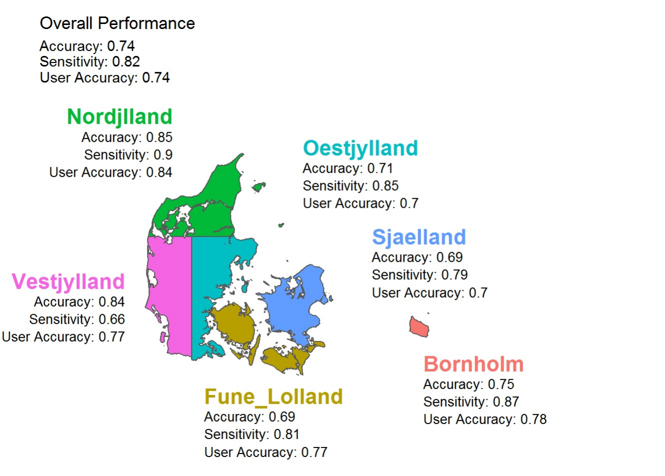
Variable importance

Variable importance for this boosted regression tree model.

```
##                                                           var   rel.inf
## treetype_bjer_dec                           treetype_bjer_dec 10.294828
## Sand_utm32_10m                                 Sand_utm32_10m  8.786217
## dtm_10m                                               dtm_10m  7.556530
## amplitude_sd                                     amplitude_sd  6.734123
## ns_groundwater_summer_utm32_10m ns_groundwater_summer_utm32_10m  6.621882
## canopy_height_sd_110m                   canopy_height_sd_110m  5.397279
## Clay_utm32_10m                                 Clay_utm32_10m  5.263973
## Soc_utm32_10m                                   Soc_utm32_10m  5.081004
## vegetation_density                         vegetation_density  5.046986
## ns_groundwater_summer_sd_110m     ns_groundwater_summer_sd_110m  4.838499
## foliage_height_diversity             foliage_height_diversity  4.823770
## solar_radiation                               solar_radiation  4.578948
## canopy_height                                   canopy_height  4.484568
## amplitude_mean                                 amplitude_mean  4.330737
## dtm_10m_sd_110m                               dtm_10m_sd_110m  4.326786
## vegetation_density_sd_110m         vegetation_density_sd_110m  4.016476
## normalized_z_sd                               normalized_z_sd  3.829831
## slope                                                   slope  2.427349
## openness_difference                       openness_difference  1.560213
```

Performance in BIOWIDE regions:

Performance map based on the independent validation data:



Performance table based on the independent validation data:

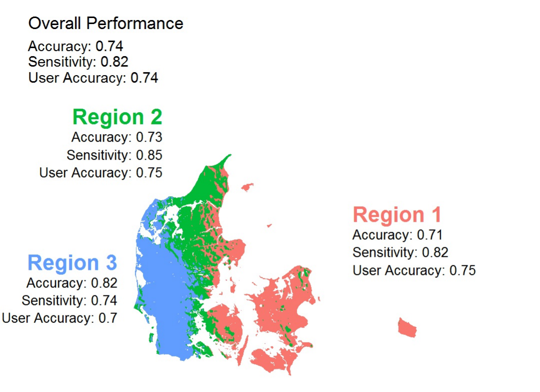
Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	0.74	0.75	0.69	0.85	0.71	0.69	0.84
Error	0.26	0.25	0.31	0.15	0.29	0.31	0.16
Sensitivity (True Positive Rate)	0.82	0.87	0.81	0.90	0.85	0.79	0.66
Specificity (True Negative Rate)	0.64	0.48	0.39	0.77	0.52	0.55	0.92
Fall-out (False Positive Rate)	0.36	0.52	0.61	0.23	0.48	0.45	0.08
Positive predictive value (User Accuracy)	0.74	0.78	0.77	0.84	0.70	0.70	0.77

Performance table based on the dependent training data:

Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	0.75	0.74	0.71	0.85	0.72	0.69	0.85
Error	0.25	0.26	0.29	0.15	0.28	0.31	0.15
Sensitivity (True Positive Rate)	0.83	0.88	0.83	0.90	0.85	0.80	0.67
Specificity (True Negative Rate)	0.65	0.49	0.42	0.78	0.54	0.55	0.92
Fall-out (False Positive Rate)	0.35	0.51	0.58	0.22	0.46	0.45	0.08
Positive predictive value (User Accuracy)	0.75	0.77	0.78	0.85	0.72	0.69	0.77

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	0.74	0.71	0.73	0.82
Error	0.26	0.29	0.27	0.18
Sensitivity (True Positive Rate)	0.82	0.82	0.85	0.74
Specificity (True Negative Rate)	0.64	0.50	0.54	0.85
Fall-out (False Positive Rate)	0.36	0.50	0.46	0.15
Positive predictive value (User Accuracy)	0.74	0.75	0.75	0.70

Performance table based on the dependent training data:

Measure	Overall	Region 1	Region 2	Region 3
Accuracy	0.75	0.71	0.74	0.82
Error	0.25	0.29	0.26	0.18
Sensitivity (True Positive Rate)	0.83	0.83	0.85	0.75
Specificity (True Negative Rate)	0.65	0.51	0.56	0.86
Fall-out (False Positive Rate)	0.35	0.49	0.44	0.14
Positive predictive value (User Accuracy)	0.75	0.75	0.76	0.71

Performance by forest type (boradleaf vs. coniferous)

Performance table based on the independent validation data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	0.74	0.75	0.83
Error	0.26	0.25	0.17
Sensitivity (True Positive Rate)	0.82	0.87	0.58
Specificity (True Negative Rate)	0.64	0.53	0.92
Fall-out (False Positive Rate)	0.36	0.47	0.08
Positive predictive value (User Accuracy)	0.74	0.77	0.74

Performance table based on the dependent training data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	0.75	0.75	0.83
Error	0.25	0.25	0.17
Sensitivity (True Positive Rate)	0.83	0.87	0.58
Specificity (True Negative Rate)	0.65	0.53	0.92
Fall-out (False Positive Rate)	0.35	0.47	0.08
Positive predictive value (User Accuracy)	0.75	0.77	0.74

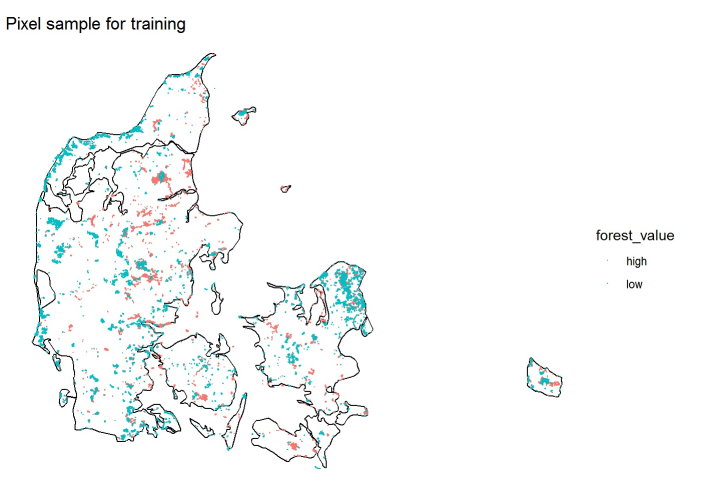
DK Forest LiDAR - Random Forest Model Performance

Jakob Assmann

02/03/2022

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.

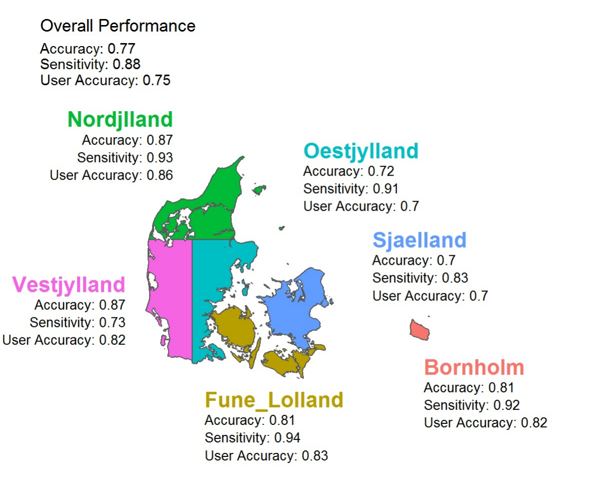
Variable importance

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

	Overall
treetype_bjer_dec	100.000000
Sand_utm32_10m	91.623261
dtm_10m	68.306076
dtm_10m_sd_110m	61.471200
ns_groundwater_summer_utm32_10m	49.774668
Clay_utm32_10m	48.733725
amplitude_sd	31.449933
slope	26.574606
openness_difference	25.213739
normalized_z_sd	23.334754
canopy_height	21.993877
Soc_utm32_10m	19.003812
solar_radiation	12.008270
amplitude_mean	11.700365
ns_groundwater_summer_sd_110m	8.569490
vegetation_density	8.510326
canopy_height_sd_110m	4.557115
vegetation_density_sd_110m	1.524886
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:

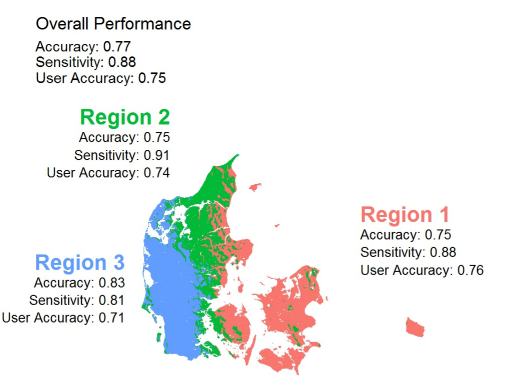
Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	0.77	0.81	0.81	0.87	0.72	0.70	0.87
Error	0.23	0.19	0.19	0.13	0.28	0.30	0.13
Sensitivity (True Positive Rate)	0.88	0.92	0.94	0.93	0.91	0.83	0.73
Specificity (True Negative Rate)	0.63	0.59	0.51	0.79	0.45	0.52	0.93
Fall-out (False Positive Rate)	0.37	0.41	0.49	0.21	0.55	0.48	0.07
Positive predictive value (User Accuracy)	0.75	0.82	0.83	0.86	0.70	0.70	0.82

Performance table based on the dependent training data:

Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	0.78	0.89	0.84	0.88	0.73	0.71	0.88
Error	0.22	0.11	0.16	0.12	0.27	0.29	0.12
Sensitivity (True Positive Rate)	0.89	0.97	0.95	0.93	0.91	0.84	0.75
Specificity (True Negative Rate)	0.65	0.75	0.57	0.80	0.48	0.55	0.94
Fall-out (False Positive Rate)	0.35	0.25	0.43	0.20	0.52	0.45	0.06
Positive predictive value (User Accuracy)	0.76	0.88	0.85	0.86	0.71	0.70	0.82

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	0.77	0.75	0.75	0.83
Error	0.23	0.25	0.25	0.17
Sensitivity (True Positive Rate)	0.88	0.88	0.91	0.81
Specificity (True Negative Rate)	0.63	0.51	0.48	0.84
Fall-out (False Positive Rate)	0.37	0.49	0.52	0.16
Positive predictive value (User Accuracy)	0.75	0.76	0.74	0.71

Performance table based on the dependent training data:

Measure	Overall	Region 1	Region 2	Region 3
Accuracy	0.78	0.77	0.76	0.84
Error	0.22	0.23	0.24	0.16
Sensitivity (True Positive Rate)	0.89	0.89	0.91	0.83
Specificity (True Negative Rate)	0.65	0.56	0.51	0.85
Fall-out (False Positive Rate)	0.35	0.44	0.49	0.15
Positive predictive value (User Accuracy)	0.76	0.78	0.75	0.72

Performance by forest type (boradleaf vs. coniferous)

Performance table based on the independent validation data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	0.77	0.75	0.82
Error	0.23	0.25	0.18
Sensitivity (True Positive Rate)	0.88	0.92	0.55
Specificity (True Negative Rate)	0.63	0.43	0.92
Fall-out (False Positive Rate)	0.37	0.57	0.08
Positive predictive value (User Accuracy)	0.75	0.75	0.72

Performance table based on the dependent training data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	0.78	0.76	0.84
Error	0.22	0.24	0.16
Sensitivity (True Positive Rate)	0.89	0.93	0.60
Specificity (True Negative Rate)	0.65	0.47	0.93
Fall-out (False Positive Rate)	0.35	0.53	0.07
Positive predictive value (User Accuracy)	0.76	0.76	0.75

Models trained using Derek's stratification

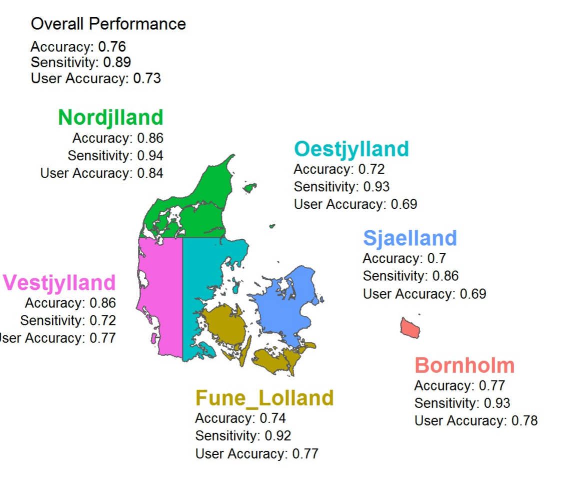
Variable importance

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

	Overall
treetype_bjer_dec	100.000000
Sand_utm32_10m	66.258856
amplitude_sd	48.823849
ns_groundwater_summer_utm32_10m	39.285113
dtm_10m_sd_110m	36.397842
dtm_10m	34.954002
Clay_utm32_10m	34.801037
amplitude_mean	18.526821
canopy_height	16.996106
normalized_z_sd	16.987006
openness_difference	15.870759
slope	12.880257
Soc_utm32_10m	11.485003
vegetation_density	9.652156
ns_groundwater_summer_sd_110m	9.431154
solar_radiation	8.900619
canopy_height_sd_110m	3.482463
foliage_height_diversity	1.739641
vegetation_density_sd_110m	0.000000

Performance in BIOWIDE regions:

Performance map based on the independent validation data:



Performance table based on the independent validation data:

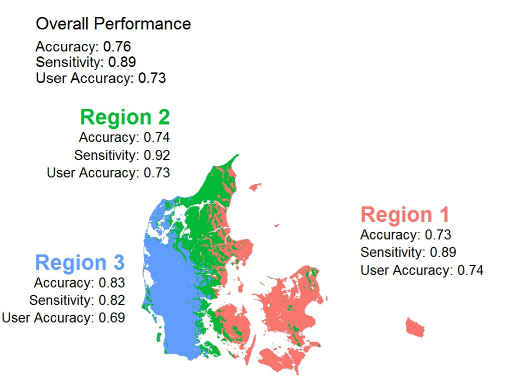
Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	0.76	0.77	0.74	0.86	0.72	0.70	0.86
Error	0.24	0.23	0.26	0.14	0.28	0.30	0.14
Sensitivity (True Positive Rate)	0.89	0.93	0.92	0.94	0.93	0.86	0.72
Specificity (True Negative Rate)	0.59	0.44	0.29	0.75	0.45	0.49	0.91
Fall-out (False Positive Rate)	0.41	0.56	0.71	0.25	0.55	0.51	0.09
Positive predictive value (User Accuracy)	0.73	0.78	0.77	0.84	0.69	0.69	0.77

Performance table based on the dependent training data:

Measure	Overall	Bornholm	Fune_Lolland	Nordjlland	Oestjylland	Sjaelland	Vestjylland
Accuracy	0.76	0.75	0.76	0.86	0.73	0.71	0.86
Error	0.24	0.25	0.24	0.14	0.27	0.29	0.14
Sensitivity (True Positive Rate)	0.90	0.93	0.92	0.94	0.92	0.87	0.74
Specificity (True Negative Rate)	0.60	0.42	0.35	0.76	0.46	0.50	0.91
Fall-out (False Positive Rate)	0.40	0.58	0.65	0.24	0.54	0.50	0.09
Positive predictive value (User Accuracy)	0.73	0.75	0.78	0.84	0.71	0.69	0.77

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	0.76	0.73	0.74	0.83
Error	0.24	0.27	0.26	0.17
Sensitivity (True Positive Rate)	0.89	0.89	0.92	0.82
Specificity (True Negative Rate)	0.59	0.44	0.46	0.83
Fall-out (False Positive Rate)	0.41	0.56	0.54	0.17
Positive predictive value (User Accuracy)	0.73	0.74	0.73	0.69

Performance table based on the dependent training data:

Measure	Overall	Region 1	Region 2	Region 3
Accuracy	0.76	0.74	0.75	0.83
Error	0.24	0.26	0.25	0.17
Sensitivity (True Positive Rate)	0.90	0.90	0.92	0.83
Specificity (True Negative Rate)	0.60	0.45	0.49	0.83
Fall-out (False Positive Rate)	0.40	0.55	0.51	0.17
Positive predictive value (User Accuracy)	0.73	0.74	0.74	0.70

Performance by forest type (boradleaf vs. coniferous)

Performance table based on the independent validation data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	0.76	0.74	0.82
Error	0.24	0.26	0.18
Sensitivity (True Positive Rate)	0.89	0.93	0.58
Specificity (True Negative Rate)	0.59	0.37	0.91
Fall-out (False Positive Rate)	0.41	0.63	0.09
Positive predictive value (User Accuracy)	0.73	0.73	0.70

Performance table based on the dependent training data:

Measure	Overall	Broadleaf	Coniferous
Accuracy	0.76	0.74	0.83
Error	0.24	0.26	0.17
Sensitivity (True Positive Rate)	0.90	0.94	0.60
Specificity (True Negative Rate)	0.60	0.39	0.91
Fall-out (False Positive Rate)	0.40	0.61	0.09
Positive predictive value (User Accuracy)	0.73	0.74	0.71

DK Forest LiDAR Summary Stats for Projections

Jakob Johann Assmann

02/03/2022

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 et al 2017) (https://linkinghub.elsevier.com/retrieve/pii/S0924271617302721) 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.

Disclaimer: The current disturbance layer requires an update with the new forest masks and applying a filtering step (MMU = 2). These will likely have a noticeable effect on the final estimates of the total area disturbed.

Name	Area [km²]	Proportion [%]
disturbed forest	38.73	0.70
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	2167.79	38.50
low quality forest	3871.18	68.70
total forest	5633.89	100.00

Disturbance statistics:

Type	Area [km²]	Proportion [%]
disturbed high quality forest	9.24	0.40
total high quality forest	2167.79	100.00

Type	Area [km²]	Proportion [%]
disturbed low quality forest	29.50	0.80
total low quality forest	3871.18	100.00

Type	Area [km²]	Proportion [%]
disturbed high quality forest	9.24	23.80
disturbed low quality forest	29.50	76.20
total disturbed forest	38.73	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	2332.19	41.40
low quality forest	3706.78	65.80
total forest	5633.89	100.00

Disturbance statistics:

Type	Area [km²]	Proportion [%]
disturbed high quality forest	9.80	0.40
total high quality forest	2332.19	100.00

Type	Area [km²]	Proportion [%]
disturbed low quality forest	28.94	0.80
total low quality forest	3706.78	100.00

Type	Area [km²]	Proportion [%]
disturbed high quality forest	9.80	25.30
disturbed low quality forest	28.94	74.70
total disturbed forest	38.73	100.00