title

Patrick Schratz[a], Jannes Muenchow[a], Eugenia Iturritxa[1], Alexander Brenning[a]

[a] *Department of Geography, GIScience group, Grietgasse 6, 07743, Jena, Germany*

Abstract

Keywords: hyperspectral imagery, forest health, machine-learning, variable importance, model comparison

1. Introduction

Remote sensing data is successfully used in forestry to monitor temporal changes across large areas REFERENCES. The use of SAR techniques enables scientists to estimate Above-Ground Biomass (AGB). Forest health is commonly modelled using optical data from multispectral satellites by applying temporal change detections. With a well-trained model, scientists are able to predict the modelled response to a large area, giving valuable information about the environmental condition of this region. To model forest health, usually vegetation indices are derived from the band values, i.e. band combinations that are proven to be sensitive to health changes of vegetation. However, the pool of possible indices is often limited due to a low spectral resolution of freely available data from multispectral sensors. Also, if the spatial resolution of the data is too coarse (e.g. ¿ 5 m), the value of a pixel contains information of multiple trees and possibly even bare-ground information. These problems introduce bias into the data which is passed onto the fitted model, leading to non-optimal modeling results.

*Corresponding author
Email address: `patrick.schratz@uni-jena.de` (Patrick Schratz)

In this study we will use hyperspectral remote sensing data with a spatial resolution of one meter and 126 spectral bands to model defoliation of *Pinus radiata* trees in northern Spain. These trees suffer from infections of invasive pathogens such as *Diplodia sapinea*, *Armillaria* or *Heterobasidion* leading to a spread of cankers or defoliation before the tree dies. The fungis are assumed to infect the trees through open wounds, possibly caused by hail damage (Iturritxa et al., 2014). The dieback of these trees, which are mainly used as timber, causes high economic damages in the Iberian Peninsula area. Remote sensing data in combination with state-of-the-art machine-learning techniques can help to constantly monitor forest health in this region using proxies such as defoliation of the tree crown. With this information, an increase in defoliation can serve as an early-warning indicator of a possible infection of a plot.

To extract the most information from the available remote sensing data, we not only calculated the most often used vegetation indices (e.g. NDVI) to link against defoliation. The narrow bands of the given data enable the calculation of up to 90 vegetation indices. Additionally, all possible combinations of Normalized Ratio Indeces (NRI) were calculated from the data and supplied to the machine-learning algorithms as predictors. With their ability to find patterns between the response and its predictors in the data, we hope to increase the predictive accuracy of defoliation based on remote sensing indices.

In this study the following objectives are addressed:

- Checking multiple machine-learning algorithms on their ability to model defoliation of *Pinus radiata* trees

- Exploration of the most important indices of the model

- Prediction of defoliation to plots with unknown levels of defoliation

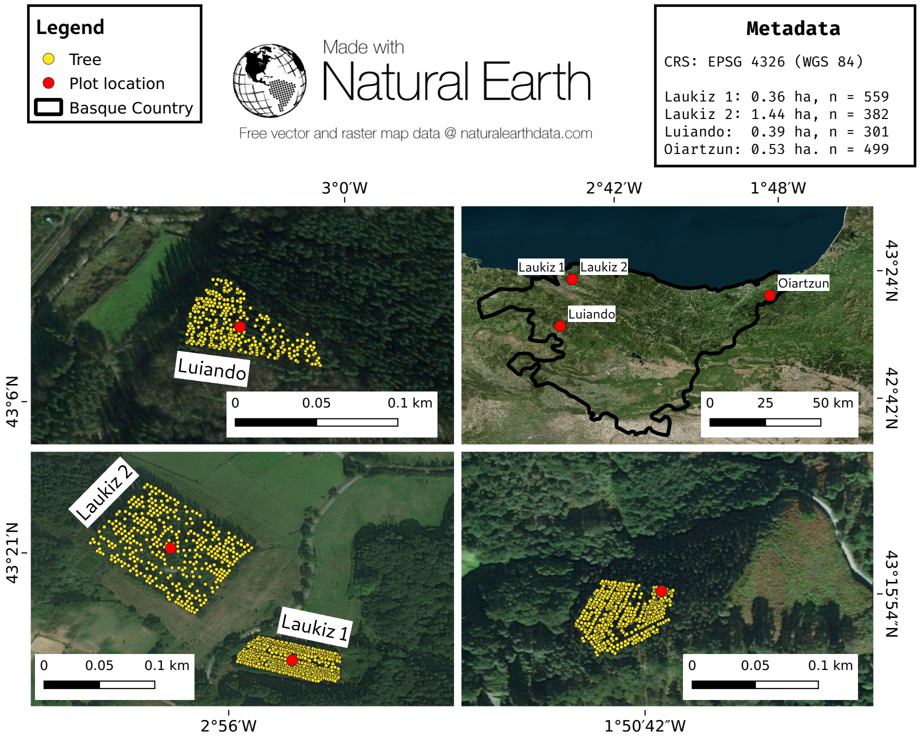

Figure 1: Information about the plot locations, the area of hyperspectral coverage and the number of trees per plot.

2. Data and study area

2.1. In-situ data

The *Pinus radiata* plots of this study, named *Laukiz 1*, *Laukiz 2*, *Luiando* and *Oiartzun*, are located in the northern part of the Basque Country (Figure 1). *Oiartzun* has the most observations (n = 529) while *Laukiz 2* has the largest area size (1.44 ha). All plots besides *Luiando* are located nearby the coast (Figure 1). The data collection took place in September 2016.

2.2. Hyperspectral data

The airborne hyperspectral data was acquired during two flight campaigns on September 28th and October 5th 2016, both around 12 am. The images were taken by an AISAEAGLE-II sensor from the Institut Cartografic i Geologic de Catalunya (ICGC). All preprocessing steps (geometric, radiometric, atmospheric) have been conducted by ICGC. The first four bands were corrupted, leaving 122 bands with valid information.

Additional data characteristics are provided in Table 1:

3. Methods

For all analysis steps we used the open-source statistical programming language R (R Core Team, 2017). The algorithm implementations of the following packages have been used: *xgboost* (Chen & Guestrin, 2016), *kernlab* (Karatzoglou et al., 2004) (Support Vector Machine), Vapnik (1998)) and *glmnet* (Friedman et al., 2010) (Ridge Regression). We used the R package *mlr* for all modeling related steps. It provides a standardized interface for a wide variety of statistical and machine-learning models in R simplifying essential modeling tasks such as hyperparameter tuning, model performance evaluation and parallelization (Bischl et al., 2016).

3.1. Derivation of indices

All vegetation indices (90 total) suitable for the wavelength range of the hyperspectral data that were available in the R package *hsdar* have been calculated. Additionally, all possible NRI were calculated from the data using the formula:

Table 1: Specifications of hyperspectral data.

Characteristic	Value
Geometric resolution	1 m
Radiometric resolution	12 bit
Spectral resolution	126 bands (404.08 nm - 996.31 nm)
Correction:	Radiometric, geometric, atmospheric

$$NRI_{i,j} = \frac{b_i - b_j}{b_i + b_j} \tag{1}$$

where i and j are the respective band numbers.

To account for geometric offsets, we used a buffer of two meters around the centroid of the respective tree. The mean value of all pixels touched by the buffer was assigned as the final value for each index. Missing values were removed from the mean value calculation. In total, 7875 Normalized Ratio Indices NRI have been calculated ($\frac{125*126}{2}$). Due to four corrupted bands and some other numerical problems, few indices returned `NA` values for some observations. These indices were removed from the dataset, leaving a total of 7471 variables without missing values.

3.2. Exploratory analysis of plot characteristics

Plot characteristics like age, stand density and defoliation were analysed to show differences among the plots. Additionally, the spectral signatures of each plot have been visualized.

3.3. Benchmarking of algorithms

Multiple algorithms were benchmarked on predictive performance to find the best performing one. Besides the well-known Support Vector Machines (SVM) (Vapnik, 1998) we also used *xgboost* which is ensemble method relying on the idea of tree boosting that gained a lot of attention in recent years (Chen & Guestrin, 2016). We also added penalized L2 (Ridge) regression to the algorithm collection due to its ability to handle highly correlated covariates. We also want to explain why we did not consider the probably most popular machine-learning algorithm, Random Forest: Due to the high number of variables model fitting times in the range hours for a single model fit were not practicable for this work. These high fitting times are caused by hyperparameter `mtry` which scales with the number of variables (Probst et al., 2018).

3.3.1. Performance estimation

The algorithms were benchmarked in two ways: (1) Using spatial cross-validation (CV) for each plot using on the k-means clustering approach of Brenning (2012). To reduce runtime we used a five-fold five-times repeated CV setup. (2) Using spatial CV on the plot level with each plot being the test set once. This results in four performance estimates, one for each fold. For (1) we only used the best performing algorithm from (2). The reason why the (2) was chosen for algorithm selection is that this model will also be used to spatially predict defoliation in other plots.

3.3.2. Hyperparameter tuning

To tune the hyperparameters of the algorithms, we used Sequential-based Model Optimization (SMBO) via the R package *mlrMBO* (Bischl et al., 2017). This Bayesian approach first composes n randomly chosen hyperparameter settings out of a user defined search space. After these n tries have been evaluated, a new hyperparameter setting to be evaluated next is proposed based on the setting that performed best. This strategy continues until a termination criterion, defined by the user, is reached (Hutter et al., 2011; Jones et al., 1998). In this work we used an initial design of 30 randomly composed hyperparameter settings and a termination criterion of 20 iterations, resulting a total budget of 50 evaluated hyperparameter settings per fold. The advantage of this tuning approach is that it substantially reduces the tuning budget which is needed to find a setting close to the global minimum compared to methods that do not use information from previous runs such as *random search* or *grid search* (Bergstra & Bengio, 2012).

3.4. Variable importance

To find indices that contributed most to model performance, we used permutation-based variable importance on the best performing algorithm.

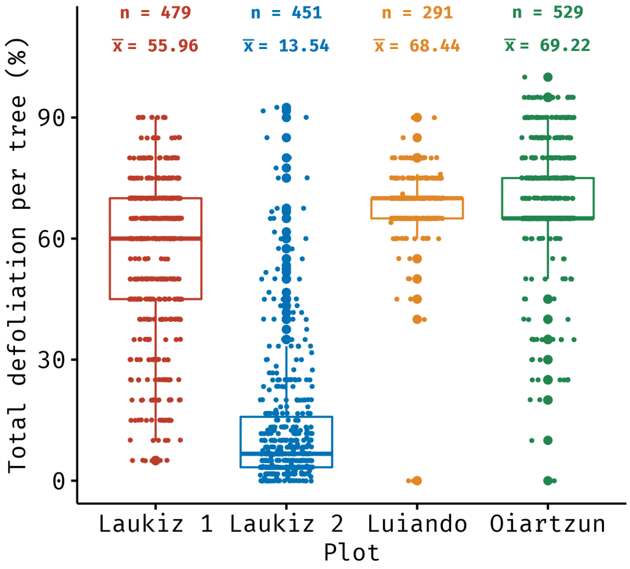

Figure 2: Descriptive statistics of the response variable *defoliation*.

4. Results

4.1. Exploratory data analysis

Oiartzun shows the highest defoliation ($\bar{x} = 69.22\%$) among the plots while *Laukiz 2* is the healthiest ($\bar{x} = 13.54\%$) (Figure 2). All plots besides *Luiando* show an evenly distributed level of defoliation across the entire plot.

The high degree of defoliation of *Luiando* and *Oiartzun* is also visible in the spectral signatures of the plots (Figure 3). Both plots show lower mean reflectance values around the wavelength range 800 nm - 1000 nm compared to Laukiz 1 and Laukiz 2. Oiartzun is almost completely missing the reflectance drop at around 815 nm that is visible for all other plots but instead shows a higher magnitude for the reflectance increase at around 920 nm. Laukiz 2 shows a mean tree density of 61.59 m **??**) while all other plots are more dense (34.64 (Laukiz 1), 33.01 (Luiando), 34.96 (Oiartzun)) (Figure 4).

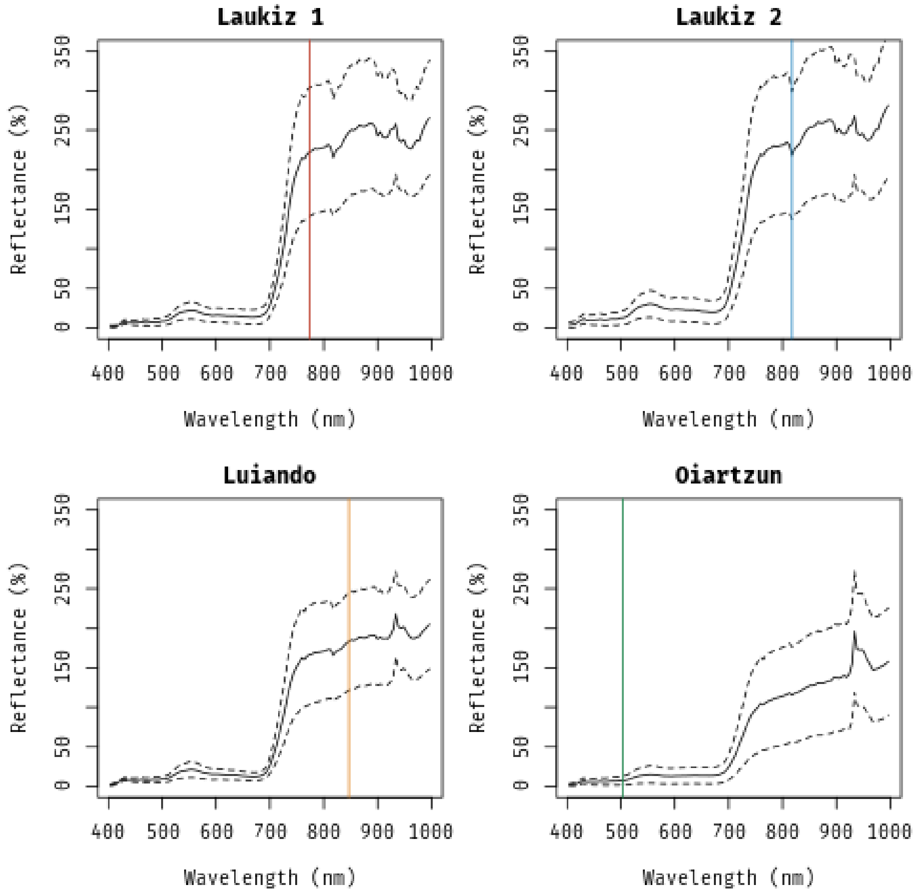

Figure 3: Spectral signatures (mean and standard deviation) of each plot. The colored lines show the most important band for each plot, respectively: Band 80 (773nm, red), band 89 (817nm, blue), band 95 (846nm, orange), band 23 (503nm, green).

Table 2: Four-fold spatial CV performances of RR, SVM and xgboost using RMSE as the error measure. Mean and standard deviation are shown.

RR	SVM	xgboost
59.10 (22.71)	36.23 (15.73)	

Table 3: Predictive performance of xgboost using the merged dataset (supermodel) and observations on a plot level only (single plot) with RMSE as the error measure. The performance estimates for "All Plots" correspond to the fold for which the respective plot was serving as the test set. For "single plot", the values correspond to the mean value of the SpCV at the repetition level (10 folds, 20 repetitions), scored by models fitted on the data of the respective plot only.

Plot/Data	All Plots (Block CV)	Single plot (SpCV)
Laukiz 1	58.95	89.89
Laukiz 2	27.94	30.37
Luiando	69.72	76.02
Oiartzun	58.09	106.65

4.2. Predictive performance

Ridge Regression (RR) shows the lowest error for three out of four plots (for Luiando *elasticnet* shows a slightly better performance) (Table 2). The magnitude of difference for RR compared to the other penalties for the plots in which RR showed the best performance ranges between XX and XX percent. For the merged dataset, all penalties show a similar mean predictive performance that outperform all single plot models besides the Laukiz 2 model.

When comparing the mean predictive performance of the plot level model against the performance of the super model at the plot level (when the respective plot served as the test set), the supermodel also outperforms the Laukiz 2 model (27.94 vs 30.37 RMSE) (Table 3).

The worst performance of the supermodel on the fold level is reported for Luiando (69.72 RMSE) while for the single plot models Oiartzun shows the highest error (106.65 RMSE).

Laukiz 2 showed contrary results compared to all other plots when linking RMSE against coefficient of variation and mean point density (**??**). Comparing RMSE against $CV/skewness$ shows a $log_2(-x)$ relationship.

4.2.1. Variable importance

NRIs using bands in the wavelength range of 770 nm - 820 nm (band 80 - band 89), which belongs to the infrared region, appear most often among the ten highest coefficient estimates across all plots (**??**). Only one vegeation index (Datt3) showed up among the most important predictors (Laukiz 1). Luiando and Oiartzun also prefered bands with longer (938.39 nm (band 114) - 996.31 nm (band 126)) and shorter wavelengths (480.30 nm (band 18) - 503.26 (band 23)). The first range again belongs to the infrared region while the second is within the region of the visible light, transitioning from blue to green.

5. Discussion

5.1. Index derivation

The exact number of contributing pixels to the final index value of an observations cannot be determined as it depends on the location of the tree within the pixel grid. If a tree is located at the border of a pixel, a buffer of e.g. three meters will include more pixels than if the point is located at the center of a pixel. Also, if a tree is located at the border of the plot, some directions of the buffer will not contain image values.

5.2. Plot characteristics

For Laukiz1, Luiando and Oiartzun RMSE seems to increase with a higher point density at a first glance. However, the point densities of these plots are

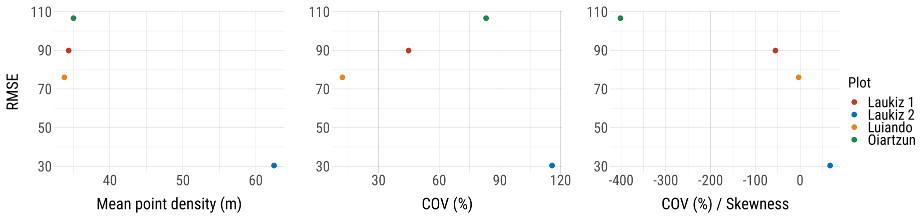

Figure 4: RMSE vs. mean point density, coefficient of variation and coefficient of variation / skewness.

very similar (33.7 m - 35.01 m) and should be interpreted as a group instead of single values. With Laukiz2 being completely off from the other plots in terms of mean point density, no pattern can be extracted from this result. Linking RMSE vs coefficient of variation shows the same relationship as linking against mean point density. The interesting $log_2(-x)$ relationship for RMSE vs. coefficient of variation / skewness should be interpreted with caution: The sample size of four plots is not representative to make general statements here. This finding should be verified with more observations in future studies.

5.3. Variable importance

References

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-parameter Optimization. *J. Mach. Learn. Res.*, *13*, 281–305. 01590.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., & Jones, Z. M. (2016). mlr: Machine learning in R. *Journal of Machine Learning Research*, *17*, 1–5.

Bischl, B., Richter, J., Bossek, J., Horn, D., Thomas, J., & Lang, M. (2017). mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions. *ArXiv e-prints*, . arXiv:1703.03373.

Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In *2012 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. doi:10.1109/igarss.2012.6352393 00052 R package version 2.1.0.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '16 (pp. 785–794). New York, NY, USA: ACM. doi:10.1145/2939672.2939785.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*, 1–22. doi:10/bb3d.

Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2011). Sequential Model-Based Optimization for General Algorithm Configuration. In *Lecture Notes in Computer Science* (pp. 507–523). Springer Berlin Heidelberg. doi:10.1007/978-3-642-25566-3_40 00686.

Iturritxa, E., Mesanza, N., & Brenning, A. (2014). Spatial analysis of the risk of major forest diseases in Monterey pine plantations. *Plant Pathology*, *64*, 880–889. doi:10/gdq9pb. 00006.

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, *13*, 455–492. doi:10/fg68nc.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). Kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, *11*, 1–20. doi:10/gdq9pc. R package version 0.9-25.

Probst, P., Wright, M., & Boulesteix, A.-L. (2018). Hyperparameters and Tuning Strategies for Random Forest. *ArXiv e-prints*, . arXiv:1804.03515. 00000.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. 00000 R version 3.3.3.

Vapnik, V. (1998). The Support Vector Method of Function Estimation. In *Nonlinear Modeling* (pp. 55–85). Springer US. doi:10.1007/978-1-4615-5703-6_3.

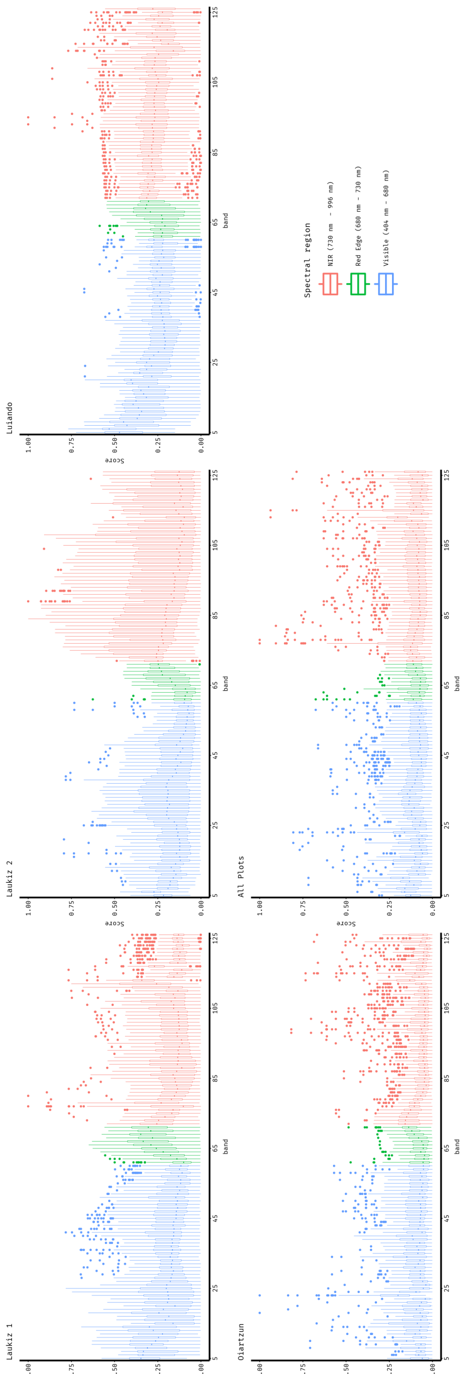

Figure 5: test