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, penalized regression, machine-learning, variable importance, spatial cross-validation

1. Introduction

2. Data and study area

2.1. In-situ data

The four *Pinus radiata* plots 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 largest area size. All plots besides Luiando are located nearby the coast. The data was collected 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.

Additional information is provided in Table 1:

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

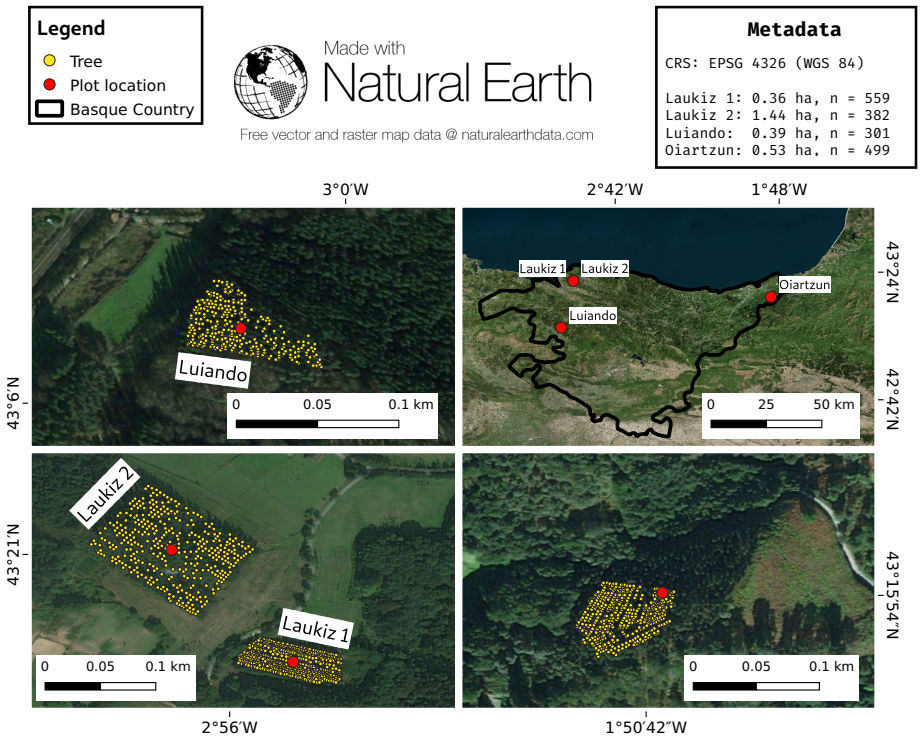

Figure 1: Information about the plot locations, the area of hyperspectral coverage and the number of trees per plot.

3. Methods

3.1. Derivation of indices

All vegetation indices (90 total) suitable for the wavelength range of the hyperspectral data and offered by the `hsdar` package have been calculated.

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

Additionally, all possible Normalized Ratio Index (NRI) were calculated from the data using the formula:

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

where i and j are the respective band numbers.

To account for geometric offsets, we used a buffer of 2 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 ($\frac{125*126}{2}$). Some indices returned `NA` values for some observations and were removed from the dataset, leaving a total of 7471 variables without missing values.

3.2. Exploratory plot characteristics

To get a better impression on the infection state of the four plots, an exploratory data analysis was done. The distribution of the response variable `defoliation` was visualized as well as the spectral signatures of each plot.

3.3. Benchmarking of algorithms

Multiple algorithms were benchmarked on predictive performance to find the best performing one. Besides well-known machine-learning algorithms like Random Forest (RF) and Support Vector Machines (SVM) we also used *xgboost* due to its promising results in recent machine-learning competitions. To expand the algorithm selection even more, we also used penalized L2 (Ridge) regression. This algorithm is able to account for the high correlation between the predictors by penalizing the estimated coefficients of the fitted model.

3.3.1. The ridge penalty

In Ridge Regression (RR) (also called ℓ_2 penalization) the assumption of unbiased coefficients is given up in favor of higher predictive accuracy and reduced variance (Hastie et al., 2001). Coefficients are standardized and penalized for

their size. When minimizing the Residual Sum of Squares (RSS), RR adds a penalization term $\lambda \sum_{j=1}^{p} \beta_j^2$ to the equation:

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2}$$

where $\lambda >= 0$ is a tuning parameter responsible for the magnitude of penalization. To make the second term (usually referred to as the *shrinkage penalty*) of this equation small, the coefficients β_j need to become small. Unlike to Lasso however, predictors are not removed from the final model and will always be $\beta_j >= 0$. Hence, λ has the effect of shrinking the coefficients when minimizing the RSS. For $\lambda = 0$, no penalization is done and standard Ordinary Least Squares (OLS) applies (James et al., 2013). The *shrinkage penalty* is only applied to the coefficients and not to the intercept. Also, while OLS generates only one set of coefficient estimates, RR will create multiple sets for every value of λ. In summary, the advantage of RR is based on the bias-variance tradeoff: For $\lambda \to \infty$, the flexibility of the fit is reduced leading to a decrease in variance of the coefficients but also introduces a (substantial) bias. For $\lambda = 0$, the variance is high but coefficients are unbiased (James et al., 2013).

3.3.2. Performance estimation

The algorithms were benchmarked using a spatial cross-validation (CV) on the plot level. The dataset was split into four folds, each fold representing one plot. Algorithms were trained on three out of four plots and evaluated on the remaining plot. Hyperparameter tuning was also performend on the plot level: Each respective training set, consisting of observations from three plots, was again split into three partitions to estimate the best performing hyperparameter set.

3.4. Variable importance

We aimed to find indices that contribute most when predicting defoliation. Permutation-based variable importance was applied on the best performing algorithm.

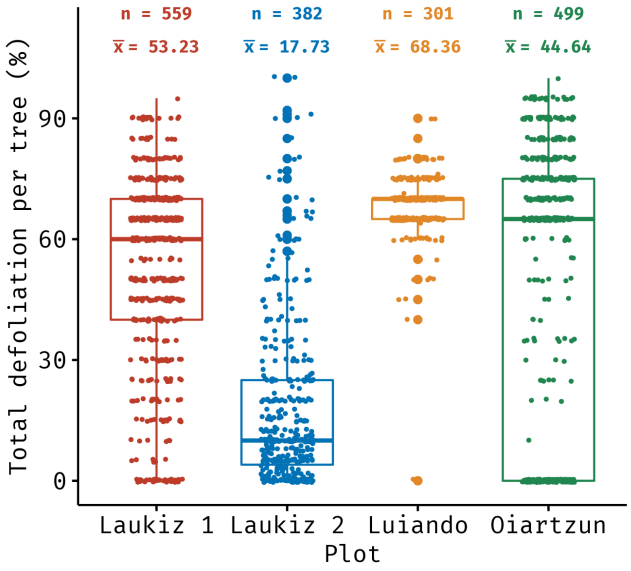

Figure 2: Descriptive statistics of the response variable *defoliation*.

4. Results

5. Results

5.1. Exploratory data analysis

Luiando shows the highest defoliation ($\bar{x} = 68.36\%$) among the plots while Laukiz 2 is the healthiest ($\bar{x} = 17.73\%$) (Figure 2). Laukiz 1 and Oiartzun both show a medium defoliation level around 50 %. Oiartzun consists mainly of trees that show either a very high level of defoliation (70% <=) or none at all. Laukiz 1 shows an evenly distributed level of defoliation across the entire plot.

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

RF	RR	SVM	xgboost
54.80 (17.58)	53.67 (17.95)	54.66 (17.67)	

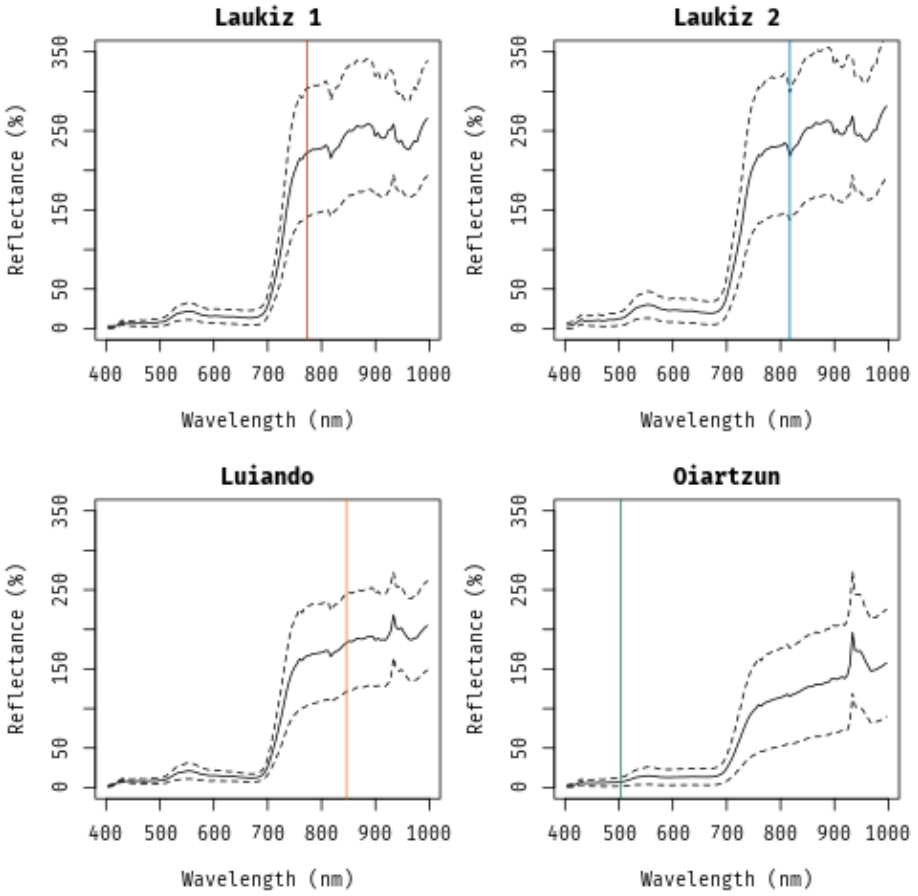

Figure 3: Spectral signatures (mean and standard deviation) of each plot. The colored lines show the most important band for each plot, respectively: Band 80 (773nm, red), band 89 (817nm, blue), band 95 (846nm, orange), band 23 (503nm, green).

The high defoliation level 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 62.51 m **??**) while all other plots are more dense (34.35

Table 3: Predictive performance of RR using the merged dataset (supermodel) and observations on a plot level only (single plot) with RMSE as the error measure. The values for "merged dataset" 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).

Plot/Data	Merged dataset (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

(Laukiz 1), 33.77 (Luiando), 35.02 (Oiartzun)) (**??**).

5.2. Predictive performance

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.

5.2.1. 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 (Table 4). 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.

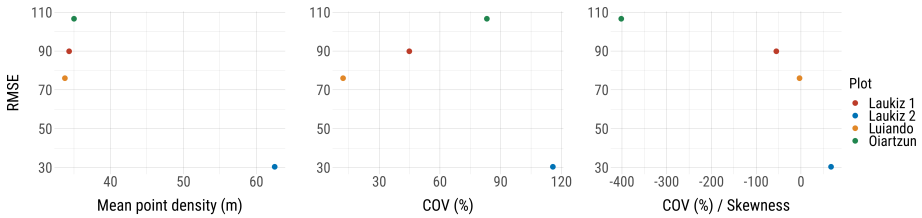

Figure 4: RMSE vs. mean point density, coefficient of variation and coefficient of variation / skewness.

Table 4: The ten highest coefficient estimates for every plot and the merged dataset.

Laukiz 1	Laukiz 2	Luiando	Oiartzun	All Plots
b80-b77 (0.0089)	b89-b84 (0.0093)	b95-b93 (1.5e-36)	b23-b18 (-0.0090)	b78-b77 (0.058)
b81-b77 (0.0079)	b89-b87 (0.0086)	b109-b106 (1.4e-36)	b23-b19 (-0.0074)	b115-b113 (0.054)
b78-b77 (0.0079)	b89-b88 (0.0086)	b92-b95 (1.4e-36)	b99-b98 (0.0073)	b82-b77 (0.052)
b77-b76 (-0.0078)	b108-b104 (0.0084)	b114-b6 (-1.2e-36)	b23-b20 (-0.0070)	b79-b77 (0.049)
b79-b77 (0.0077)	b89-b85 (0.0083)	b96-b93 (1.2e-36)	b10-b8 (-0.0063)	b80-b77 (0.049)
Datt3 (-0.71)	b92-b84 (0.0083)	b116-b6 (-1.2e-36)	b102-b98 (0.0062)	b81-b77 (0.049)
b41-b25 (0.0070)	b89-b83 (0.0081)	b114-b5 (-1.2e-36)	b124-b115 (0.0062)	b81-b78 (-0.048)
b116-b113 (0.0068)	b89-b86 (0.0080)	b115-b114 (1.2e-36)	b23-b15 (-0.0060)	b124-b115 (-0.047)
b82-b77 (0.0068)	b92-b88 (0.0080)	b95-b91 (1.1e-36)	b126-b115 (-0.0060)	b23-b20 (-0.047)
b77-b75 (-0.0068)	b108-b96 (0.0079)	b114-b8 (-1.1e-36)	b118-b115 (-0.0059)	b80-b78 (-0.046)

6. Discussion

6.1. 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.

6.2. 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 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.

6.3. Variable importance

References

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer New York. doi:10.1007/978-0-387-21606-5 02665.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer New York. doi:10.1007/978-1-4614-7138-7.

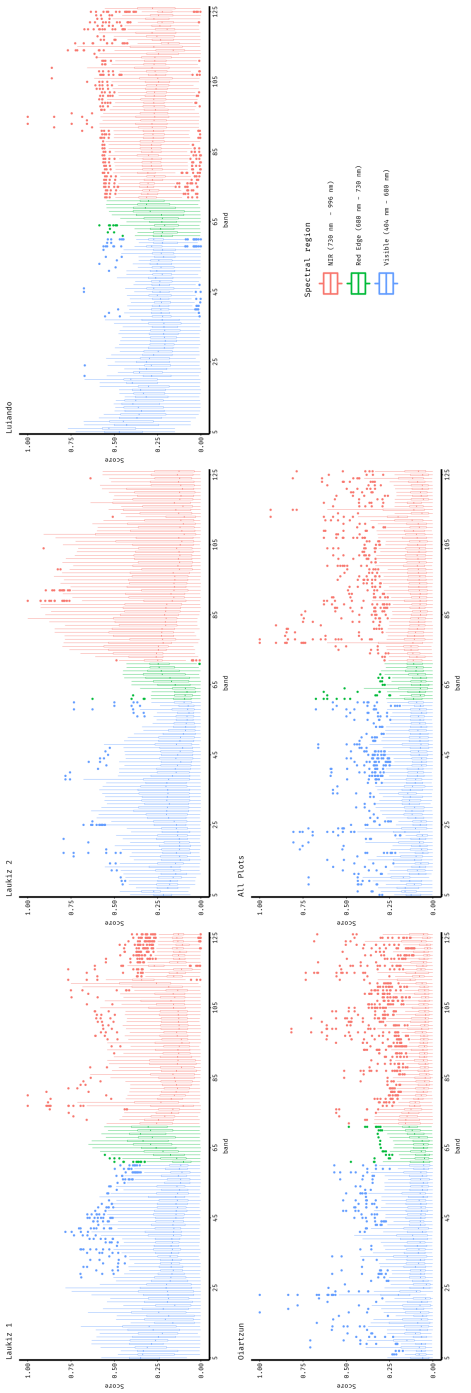

Figure 5: test