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, statistical learning, spatial cross-validation

1. Introduction

2. Data and study area

2.1. Ground 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). Laukiz 1 has the most trees (n = 559) 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)

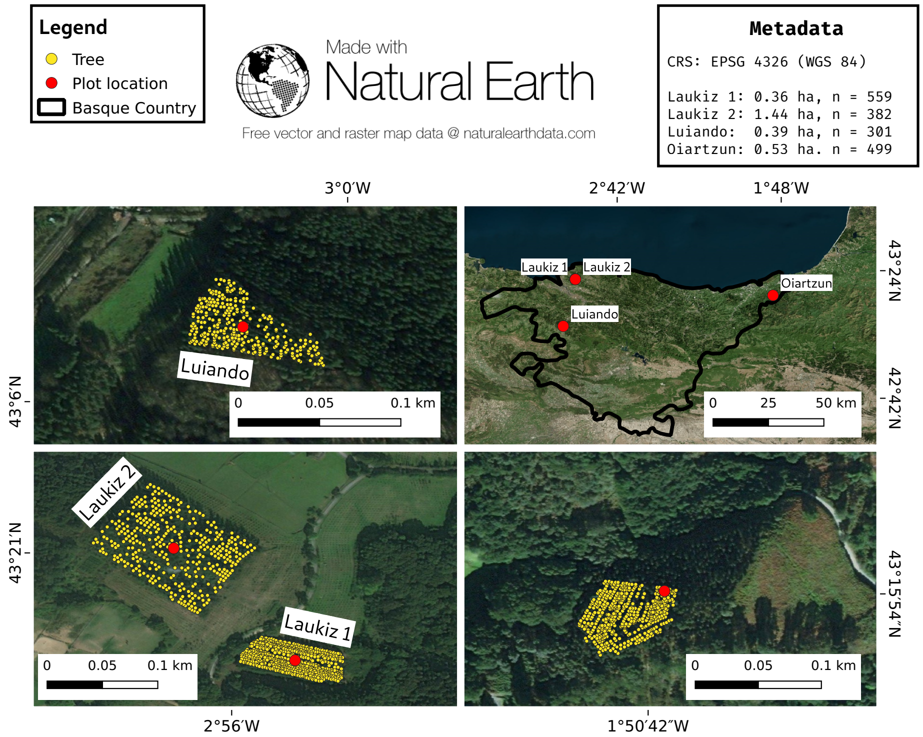

Figure 1: Information about the plot locations, the area of hyperspectral coverage and the number of trees per plot.

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 NRIs 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 indices that were available for all plots without missing values. Note that due to the mass of variables we cannot state which indices in detail have been removed.

3.2. Penalized regression

The aim of this work was to find the indices that best explain defoliation within the plots. We used penalized regression to account for the large amount of highly correlated predictor variables. In a standard Ordinary Least Squares (OLS) regression one of the assumptions is that the predictor variables should be independent when minimizing the Residual Sum of Squares (RSS) (Bare & Hann, 1981; Hastie et al., 2001):

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

where β_0 is the intercept, β_i the coefficient, x_{ij} the predictor variable and y_i the response variable.

If this assumption is violated, regression coefficients can be highly biased. They can even show the wrong sign and are very sensitive to adding new independent variables or data points to the model. These points reduce the robustness and performance of OLS regression when dealing with multicollinearity. One approach to overcome these limitations is to penalize the coefficients. This method leads to a substantial decrease in variance and better predictive performance compared to OLS regression. However, it also sacrifices the assumption of unbiased coefficients. Hence, the resulting coefficients cannot be used for statistical inference but should be interpreted as a measure of variable importance.

3.2.1. 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 RSS, RR adds a penalization term $\lambda \sum_{j=1}^{p} \beta_j^2$ to the equation:

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3}$$

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 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 $\lambda \rightarrow \infty$, 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).

3.2.2. The lasso penalty

While the Ridge penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ shrinks the coefficients, it does not exclude any predictor from the model by setting the coefficient to zero. This is what the Lasso penalty (also called ℓ_1) does: It works similar to a ”best subset selection” approach by only keeping the most important variables in the model and reducing the coefficients of unimportant ones down to zero. This is done by a different *shrinkage penalty* $\lambda \sum_{j=1}^{p} \mid \beta_j \mid$ that is added to the OLS term when minimizing the RSS:

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid \tag{4}$$

Subsequently, Lasso is performing variable selection and results in a sparse model that has the advantage of a simplied interpretability compared to Ridge models. However, coeffcients are still biased due to penalization (Hastie et al., 2001; James et al., 2013).

3.2.3. The elasticnet penalty

The *elasticnet penalty* combines both penalties Lasso and Ridge by weighting them with a parameter α that needs to be optimized. It was introduced by Zou & Hastie (2005) with the idea to combine the advantages of both penalty terms. The *elasticnet penalty* is written as

$$\lambda \sum_{j=1}^{p} \alpha \beta_j^2 + (1 - \alpha) \mid \beta_j \mid \tag{5}$$

where α weights the contribution of either the Lasso or Ridge penalty. For $\alpha = 0.5$ both are weighted equally. When $\alpha = 0$ only Ridge is used and for $\alpha = 1$ only Lasso applies (Hastie et al., 2001).

3.3. Modeling

3.3.1. Selecting the best penalty

As the introduced penalties behave different depending on the characteristics of the dataset, we first conducted a nested 10-fold spatial cross-validation (CV) to find the best method among those three. The model comparison was applied to all of the four plots using Root Mean Square Error (RMSE) as the error measure. For both the performance evaluation and hyperparameter tuning a spatial sampling using k-means clustering was used (Brenning, 2012). For the tuning, 200 random search iterations were used. The limits of the tuning space for λ were set to the internally calculated (data driven) limits of the R package `glmnet`.

3.3.2. Extraction of the most important variables

Next, we applied the winning method (RR) on all plots. Hyperparameter tuning of λ for the full dataset was again done using a spatial sampling. The ten highest coefficients (both positive and negative) were extracted and reported.

3.3.3. Linking variables to plot characteristics

To interpret the outcomes of the models on a plot level, we linked the winning variables of each to dataset attributes that describe the underlying plot characteristics. For example, a plot which a low tree density might inherit more information from the bare ground in the calculated indices while a plot with a very high tree density might in contrast contain information from multiple trees. Also, the overall level of defoliation for the whole plot might possibly have an effect how well an index is able to describe the situation.

3.3.4. Creation of a super model with all observations

After the plot level analysis, we merged all observations into a single dataset and fitted another RR model. We used a block-level based spatial CV for this

Table 2: 10-fold 20-times repeated spatial CV performances of lasso, ridge and elasticnet penalties on the plot level and the merged dataset using RMSE as the error measure. Values show the overall mean and standard deviation at the repetition level.

Plot/Penalty	Lasso	Ridge	Elasticnet
Laukiz 1	133.15 (4.66)	86.71 (2.96)	121.04 (4.15)
Laukiz 2	91.93 (6.78)	29.73 (0.38)	47.18 (13.70)
Luiando	74.85	76.02	74.77
Oiartzun	327.93	106.65	260.38
Merged dataset	56.88	55.88	56.76

setup. For a total number of m folds (n = number of plots), every plot is once the test set while the training is done on the remaining ones. This ensures that the test set is fully spatially independent from the training set. Tuning was also performned on the block level within the training set on $n - 1$ folds using 200 random search iterations. Due to the fixing of the indices on a plot level, varying those to use multiple repetitions is not possible. The idea behind fitting a supermodel is that this model includes information from all plots rather than just from a single plot. This may possibly reduce the predictive error and create a more robust model.

4. Results

4.1. Plot characteristics

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.

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 Figure 4) while all other plots are more dense (34.35 (Laukiz 1), 33.77 (Luiando), 35.02 (Oiartzun)) (Figure 4).

4.2. 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

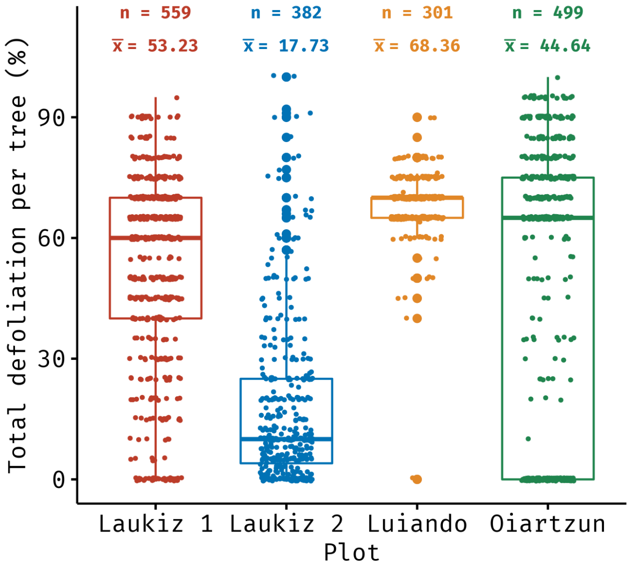

Figure 2: Descriptive statistics of the response variable *defoliation*.

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

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

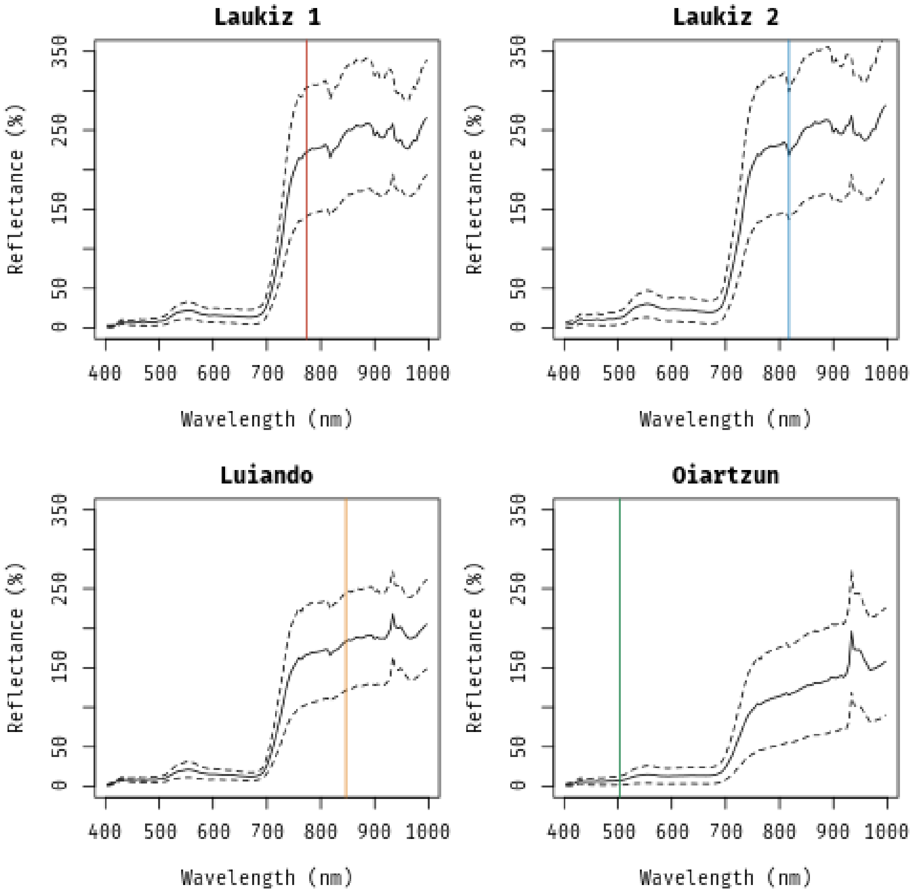

Figure 3: Spectral signatures (mean and standard deviation) of each plot. The colored lines show the most important band for each plot, respectively: Band 80 (773nm, red), band 89 (817nm, blue), band 95 (846nm, orange), band 23 (503nm, green).

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).

5. Discussion

5.1. Index derivation

The exact number of contributing pixels of an index 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, the same buffer (e.g. 3 m) 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 image data, some directions of the buffer may not contain values.

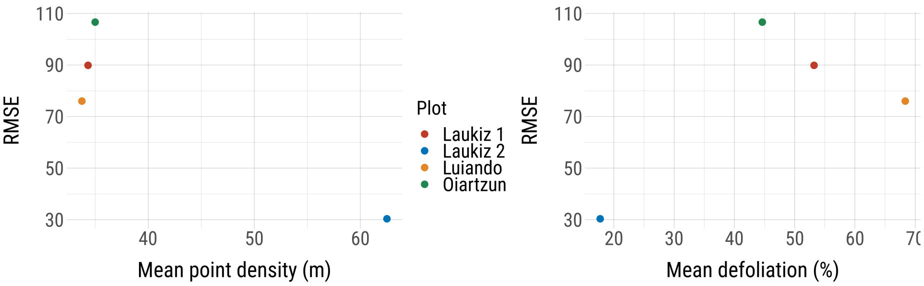

Figure 4: RMSE vs. mean defoliation and point density

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)

5.2. Variable importance

References

Bare, B. B., & Hann, D. (1981). Applications of ridge regression in forestry. *Forest Science*, *27*, 339–348.

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. URL: `https://doi.org/10.1109%2Figarss.2012.6352393`. doi:`10.1109/igarss.2012.6352393` R package version 2.1.0.

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer New York. URL: `https://doi.org/10.1007%2F978-0-387-21606-5`. doi:`10.1007/978-0-387-21606-5`.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer New York. URL: `https://doi.org/10.1007%2F978-1-4614-7138-7`. doi:`10.1007/978-1-4614-7138-7`.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301–320. URL: `https://doi.org/10.1111%2Fj.1467-9868.2005.00503.x`. doi:`10.1111/j.1467-9868.2005.00503.x`.