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)

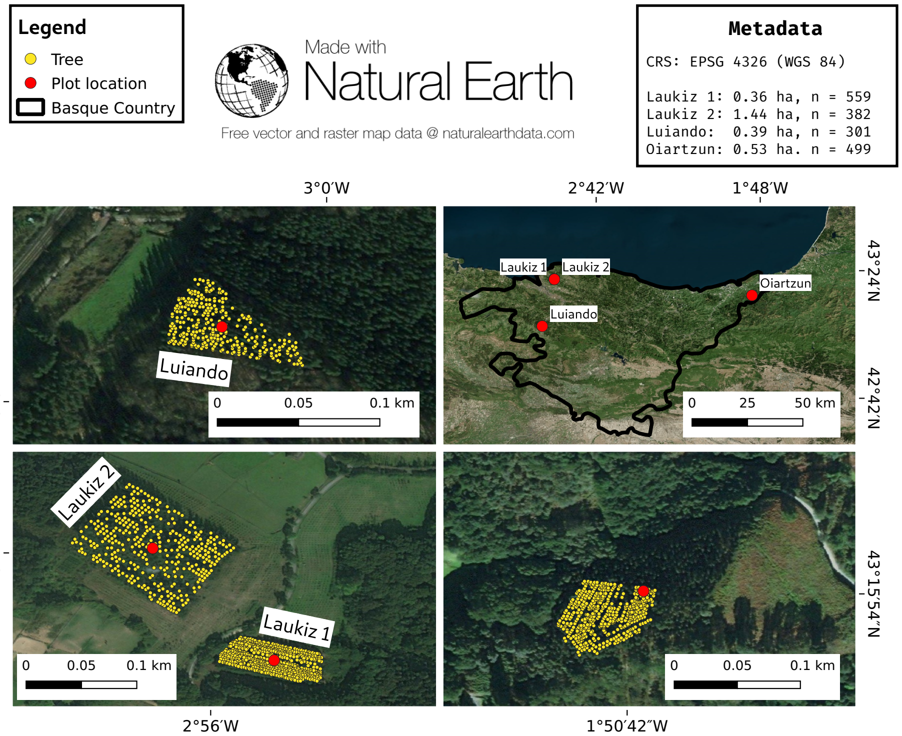

Figure 1: Information about the plot locations, the area of hyperspectral coverage and the number of trees per plot.

3. Methods

3.1. Derivation of indices

All vegetation indices (90 total) suitable for the wavelength range of the hyperspectral data that are 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 sacrifies 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 $\lambda \to \infty$, 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).

3.2.2. 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 idea of the *elasticnet penalty* is to combine both penalties Lasso and Ridge and weight these by a parameter α that 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 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).

3.3.2. Extracting the most important variables

Next, we applied the winning method (RR) on all plots and additionally to the merged dataset of all plots to find the most important coefficients of every setup. Hyperparameter tuning of λ was done using again a spatial sampling again for the single plots. For the merged dataset we used a plot-based sampling approach: One plot serves as the test partition while the remaining plots form the training set. This approach has the advantage of a spatially independent test partition but also comes with the limitation of estimating the performance on a single repetition only as the restriction to the plot-level leaves no opportunity to create multiple resampling instances.

3.3.3. Linking the variables to plot characteristics

To interpret the outcomes of the models on a plot level, we linked the winning variables of each to attributes of the dataset 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 calculdated indices while a plot with a very high tree density might in contrast contain information from multiple trees.

4. Results

4.1. Comparison of penalties

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.

References

Bare, B. B., & Hann, D. (1981). Applications of ridge regression in forestry, . *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.

Table 2: Predictive performances of lasso, ridge and elasticnet penalties on the single plots and the merged dataset.

Plot/Penalty	Lasso	Ridge	Elasticnet
Laukiz 1	146.33	87.76	135.58
Laukiz 2	146.33	87.76	135.58
Luiando			
Oiartzun		106.65	
All Plots	56.88	55.88	56.76