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)

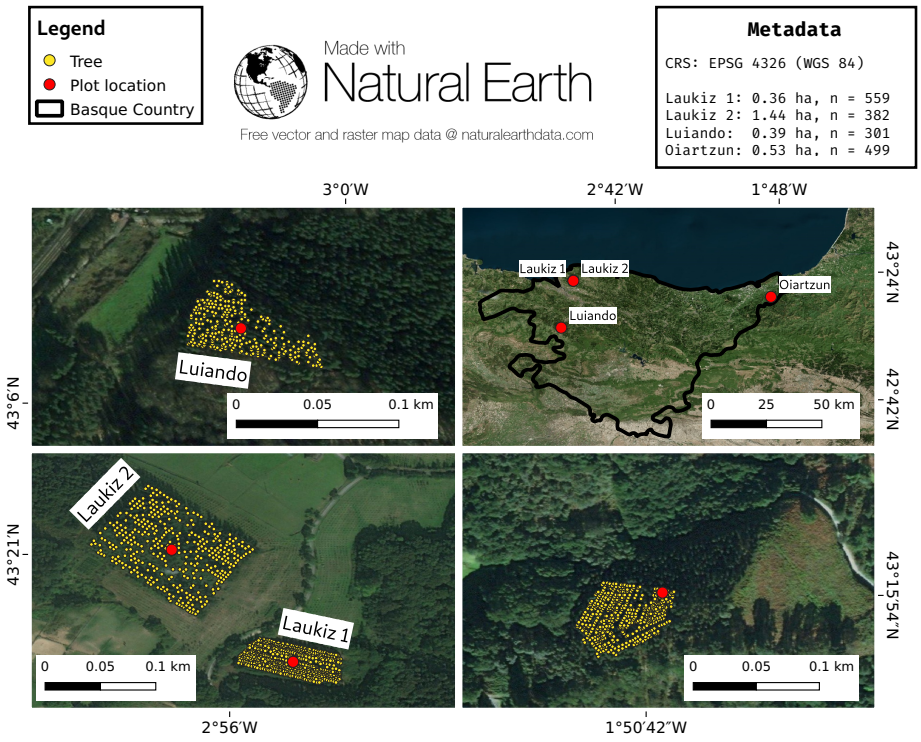

Figure 1: Information about the plot locations, the area of hyperspectral coverage and the number of trees per plot.

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

Penalized regression was used to account for the large amount of highly correlated predictor variables. The aim was to find the indices that best explain defoliation within the plots.

Due to the amount of highly correlated predictor variables in the dataset, the independence assumption of the predictors is violated. To compensate for that, penalized regression penalizes the coefficients. This leads to a substantial decrease in variance and better predictive performance but also to biased coefficients that cannot be used for statistical inference anymore. The resulting coefficients can be seen as a measure of variable importance. The penalization terms that have been used in this work are explained in the following sections.

3.2.1. L1 Penalization

3.2.2. L2 Penalization

3.2.3. Elasticnet

3.3. Modeling

The first step was to decide which penalization term works best for the given data. A nested 10-fold spatial cross-validation (CV) was conducted on every single plot and the merged dataset to explore the best method.

4. Discussion

4.1. 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