Modeling defoliation as a proxy for tree health: A case study using machine-learning and hyperspectral remote sensing data

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

Data retrieved from remote sensing satellites is successfully used in forestry to monitor temporal changes across large areas (Martinez del Castillo et al., 2015; Sexton et al., 2015). The use of Synthetic Aperture Radar (SAR) techniques enables scientists to estimate Above-Ground Biomass (AGB) (Lu et al., 2016; Sinha et al., 2015). Forest health is commonly assesed using optical data from multi-/hyperspectral satellites by applying temporal change detections (Zhang et al., 2016). With the recent success story of machine-learning methods in the field of remote sensing, modeling techniques such as Random Forest (RF) are frequently used to model relationships of possible triggers to forest health (Belgiu & Drăguţ, 2016; Lary et al., 2016; Michez et al., 2016).

With a robust model, predictions to large areas can be conducted, providing valuable information about the health condition of forest stands. One approach to model forest health is to extract information from spectral signatures of affected and unaffected trees (Lelong et al., 2010). Spectral (vegetation-)indices

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

have shown the potential to provide valuable information to increase predictive accuracy of forest pathogens (Jiang et al., 2014; Adamczyk & Osberger, 2015).

However, the amount of possible (vegetation-)indices that can be calculated is often limited due to a low spectral resolution of freely available data from optical multispectral sensors (e.g. Sentinel-2). Also, there is currently no freely available data from hyperspectral sensors that could be used for such studies (after the decommission of the EO-1 Hyperion satellite). If the spatial resolution of the data is too coarse (e.g. $> 5m$), the value of a pixel usually contains information from multiple trees and possibly even bare-ground information. This point makes it impossible to use data from sensors like Sentinel-2 to train a model on a tree level.

In this study we will use hyperspectral data with a spatial resolution of one meter and 126 spectral bands to model the health status of Monterey Pine (*Pinus radiata*) plantations in northern Spain (Figure 1). The trees in the study area suffer from infections of invasive pathogens such as *Diplodia sapinea*, *Fusarium circinatum Armillaria mellea* or *Heterobasidion annosum* leading to a spread of cankers or defoliation (Mesanza et al., 2016; Iturritxa et al., 2017). In-situ measurements of defoliation on a tree level have been collected to serve as the response variable (as a proxy for tree health). The fungi are assumed to infect the trees through open wounds, possibly caused by previous hail damage (Iturritxa et al., 2014). The dieback of these trees, which are mainly used as timber, causes high economic damages (Ganley et al., 2009). Hyperspectral remote sensing data in combination with state-of-the-art machine-learning techniques is used to help monitoring the health status in this region by early detecting affected trees/plots. The aim is to spatially predict the fitted model to other plots/the whole Basque Country to create a forest health map.

To extract the most information from the available remote sensing data, we not only calculated the most common vegetation indices like *NDVI* to link against defoliation but all possible ones within the spectral region of the data (400 nm - 1000 nm) that are implemented in the *hsdar* package in R (Lehnert et al., 2018). Additionally, all possible combinations of Normalized Ratio Indices

(NRI) were calculated from the data and supplied to a selection of different algorithms (*Support Vector Machines (SVM)*, *Ridge Regression (RR)*, *xgboost*) as predictors.

Specifically the following objectives are addressed:

- Comparison of multiple algorithms on their performance to model defoliation of *Pinus radiata* trees using highly-correlated indices
- Exploration of the most important indices of the best performing model
- Spatial prediction of defoliation to *Pinus radiata* plots and the whole Basque Country

2. Data and study area

2.1. In-situ data

The *Pinus radiata* plots of this study, namely *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). In total 1750 observations are available (*Laukiz 1* = 479, *Laukiz 2* = 451, *Luiando* = 291, *Oiartzun* = 529). The data was surveyed 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 using a AISAEAGLE-II sensor. All preprocessing steps (geometric, radiometric, atmospheric) have been conducted by the Institut Cartografic i Geologic de Catalunya (ICGC). The first four bands are corrupted, leaving 122 bands with valid information. Additional metadata information is available in Table 1:

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

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) (*xgboost*), *kernlab* (Karat-

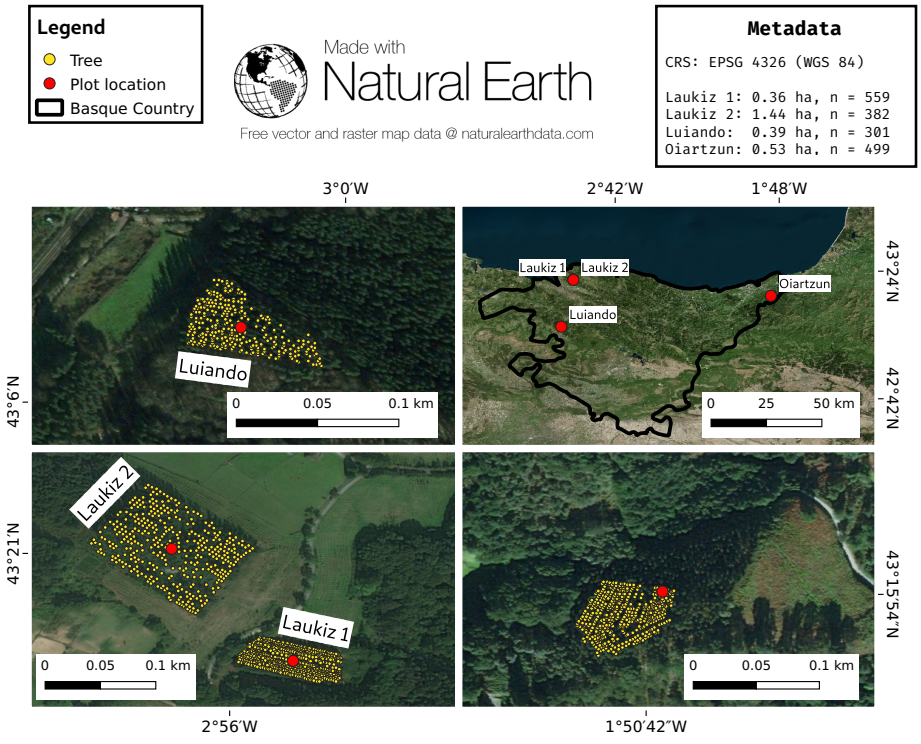

Figure 1: Information about the plot locations, the area of hyperspectral coverage and the number of trees per plot.

zoglou et al., 2004) (Support Vector Machine) 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). We provide the complete code and required data as a Mendeley dataset to make this work fully reproducible ().

3.1. Derivation of indices

To use the full information from the hyperspectral data, we calculated all possible vegetation indices that are available in the *hsdar* package (90 in total) and all possible NRI combinations. We were interested if NRIs of arbitrary band combinations will have a substantial effect on the predictive power of the fitted model. The NRIs were calculated using the following 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 (which were reported with up to 1 m from ICGC), 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 of each index. In total, $\frac{125*126}{2} = 7875$ NRIs were calculated . Due to four corrupted bands and numerical problems for some band combinations, some indices returned `NA` for specific observations. We removed all indices from the dataset that showed one or more `NA` values (across all plots) since we valued a single observation more than having an additional NRI as a predictor variable. In total, 7471 indices had no `NA` values and were subsequently used as predictors.

3.2. Benchmarking of algorithms

Three algorithms (*xgboost*, *SVM*, *RR*) were benchmarked on their predictive performance. Besides the well-known SVM algorithm (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 portfolio due to its ability to handle highly correlated covariates (Hoerl & Kennard, 1970). One of the most popular machine-learning algorithm, Random Forest (Breiman, 2001), was not considered for this study: 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). After the selection of the best model, we checked if the winning algorithm can achieve a similar performance when using only the most important variables compared to using all variables. A successful feature selection simplifies the spatial prediction task because the prediction dataset needs to consist of less variables. Furthermore, model complexity and fitting times are reduced.

3.2.1. Performance estimation

The algorithms were benchmarked in two ways:

1. Using spatial block cross-validation (CV) on the plot level with each plot serving as the test set once. Subsequently, four performance estimates were retrieved, one for each fold.

2. Using five-fold five-time repeated spatial CV within each plot based on the k-means clustering approach of Brenning (2012) (Figure 2).

As we used the best algorithm of 1) for the spatial prediction, we first conducted the model selection on this setup and only applied the winning algorithm on 2).

3.2.2. 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

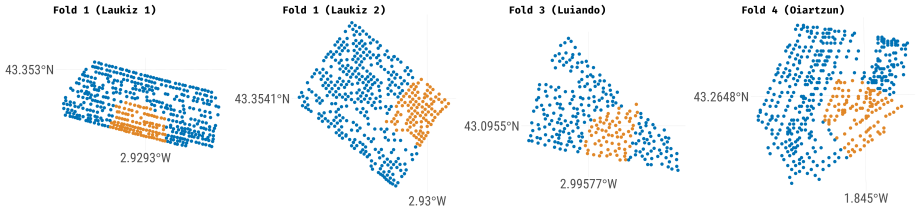

Figure 2: Fold 1 of the spatial partitioning using k-means clustering for *Laukiz 1*, *Laukiz 2*, *Luiando* and *Oiartzun*.

hyperparameter setting, which is going to be evaluated next, is proposed based on a fitted regression model. The regression model estimates the performance of the machine-learning method for unknown hyperparameter settings. Using these estimates, a new promising hyperparameter setting is proposed to be evaluated next. 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 in 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).

3.3. Variable importance

To find indices that contributed most to model performance, we used the internal variable importance measure of the *xgboost* algorithm. The score is calculated by taking the contribution of each feature for each tree in the fitted model. The higher the score of a variable, the more important it is for the fitted model when making predictions (Chen & Guestrin, 2016). The variable importance measure is automatically computed during model fit. In contrast to other approaches such as permutation-based ones, the *xgboost* score is composed out of three parts that contribute to the overall importance (Chen & Guestrin, 2016):

- Gain: The relative contribution of the feature to the model

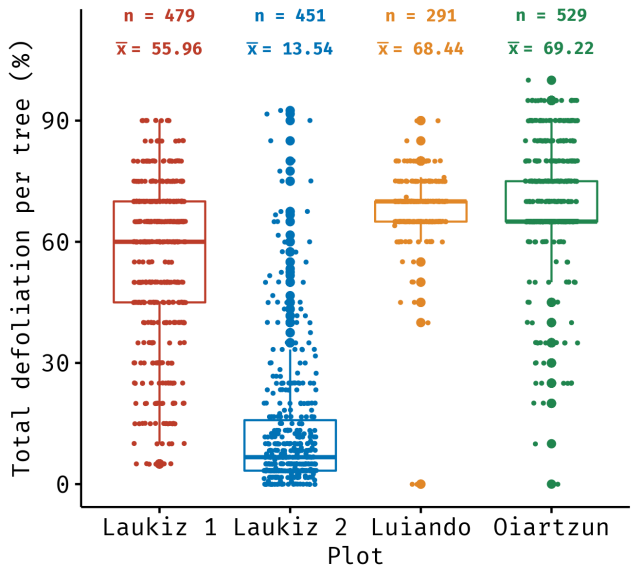

Figure 3: Descriptive statistics of the response variable *defoliation*.

- Cover metric: How often a feature was selected to be the deciding feature in a tree for a specific observation
- Frequency How often a feature occurs in all trees of the model

The *Gain* features is the most important one among the three. All measures sum up to one (Chen & Guestrin, 2016).

4. Results

4.1. Plot characteristics

Oiartzun shows the highest defoliation ($\bar{x} = 69.22\%$) among the plots while *Laukiz 2* is the healthiest ($\bar{x} = 13.54\%$) (Figure 3). All plots besides *Luiando* show an evenly distributed level of defoliation across the entire plot.

The high degree of defoliation for *Luiando* and *Oiartzun* is also visible in the spectral signatures of the plots (Figure A.8). Both plots show lower mean reflectance values around the wavelength range 800 nm - 1000 nm compared to

Table 2: Spatial block CV performances of *RR*, *SVM* and *xgboost* using RMSE as the error measure. Mean and standard deviation are shown.

RR	SVM	xgboost	xgboost (7 variables)
59.10 (22.71)	36.23 (15.73)	33.26 (16.61)	29.59 (16.09)

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 (Figure 4) while all other plots have a higher density (34.64 m (Laukiz 1), 33.01 m (Luiando), 34.96 m (Oiartzun)) (Figure 4).

4.2. Predictive performance

4.2.1. Algorithm benchmarking

The *xgboost* algorithm showed the lowest error (33.26 RMSE) when benchmarking the learners on the complete dataset of all plots (Table 2). While the *SVM* performance was only slightly worse (36.23 RMSE), *RR* showed a substantially

Table 3: Predictive performance of *xgboost* using all observations and all variables (All Observations/all variables), all observations and the seven most important variables only (All Observations/7 variables) and observations from specific plots only (Plot level observation/all variables) with RMSE as the error measure. The performance estimates for "All Observations" correspond to the fold for which the respective plot was serving as the test set (block CV). Column "Plot level observations", shows the mean RMSE estimates at the repetition level of a five-fold five-time repeated spatial CV, scored by using data of the respective plot only.

Plot/Data	All Observations/ all variables (Block CV)	All Observations/ 7 variables (Block CV)	Plot level observations/ all variables (SpCV)
Laukiz 1	22.03	21.47	19.18
Laukiz 2	51.75	49.94	17.24
Luiando	13.20	15.37	8.30
Oiartzun	32.97	17.62	14.40

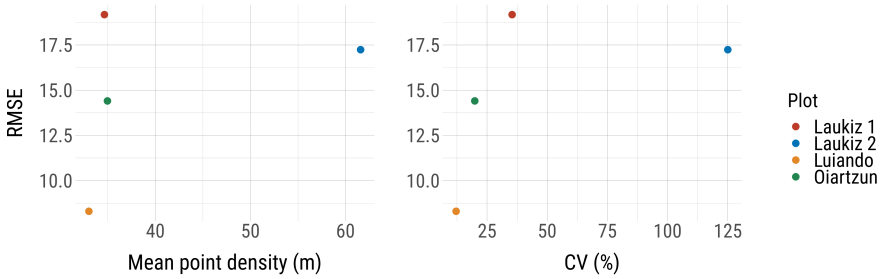

Figure 4: RMSE vs. mean point density and coefficient of variation (defoliation).

worse performance than *xgboost* (59.10 RMSE).

4.2.2. Single models vs. super model

Comparing the mean predictive performance of models fitted at the plot level against the performance of the model that was fitted using all data (super model), the plot-level models showed a better performance in all cases (Table 3). The highest difference between both datasets occurred for plot *Laukiz 2* with a difference of 34.51 RMSE.

Using only the seven most important variables (Figure 5) for the super model showed small increases in performance for *Laukiz 1* and *Laukiz 2*, a small decrease for *Luiando* and almost a reduction of 50% of the error for *Oiartzun* (32.97 vs. 17.62 RMSE) (Table 3).

4.2.3. RMSE vs. plot characteristics

An increase of the error rate was observed with an increase of descriptive plot measures such as mean point density and the coefficient of variation (based on the response variable *defoliation*) (Figure 4).

4.3. Variable importance

The seven most important features of the super model in this study were vegetation indices with the *EVI* (Huete et al., 1997) being the most important one (Figure 5).

Table 4: Formulas of the five most important vegetation indices of the super model. R = Reflectance at wavelength, D = First derivation of reflectance value at wavelength.

Acronym	Name	Formula	Reference
EVI	Enhanced vegetation index	$2.5 * \frac{R_{800}-R_{670}}{R_{800}-(6*R_{670})-(7.5*R_{475})+1)}$	Huete et al. (1997)
GDVI	Generalized DVI*	$\frac{R^n_{800}-R^n_{680}}{R^n_{800}+R^n_{680}}$	Wu et al. (2008)
D1	Derivative Index	$\frac{D_{730}}{D_{706}}$	Zarco-Tejada et al. (2003)
mNDVI	Normalized DVI*	$\frac{R_{800}-R_{680}}{(R_{800}+R_{680}-2*R_{445}}$	Sims & Gamon (2002)
mSR	Simple Ratio Index	$\frac{R_{800}-R_{445}}{R_{680}-R_{445}}$	Sims & Gamon (2002)

* Difference Vegetation Index

$$EVI = 2.5 * \frac{R_{800} - R_{670}}{R_{800} - (6 * R_{670}) - (7.5 * R_{475}) + 1)} \tag{2}$$

where R = Reflectance at the respective wavelength.

Vegetation index *GDVI* appears three times among the first seven most important features (Figure 5) with different `n` values. This is because it was computed four times, with `n` ranging from 1 - 4 (Wu et al., 2008):

$$GDVI = \frac{R^n_{800} - R^n_{680}}{R^n_{800} + R^n_{680}} \tag{3}$$

The seven most important features (*EVI*, *GDVI_4*, *D1*, *GDVI_3*, *GDVI_2*, *mNDVI* and *mSR*) showed a substantial difference in the importance score compared to all following variables (Figure 5).

The best NRI scored rank eight (band 112 and band 62). All further places up to rank 30 were occupied by NRIs.

4.4. Spatial prediction

The plots with a higher mean defoliation (*Luiando* and *Oiartzun*) showed a good visual separability compared to the healthier plots *Laukiz 1* and *Laukiz 2* (Figure 6). All predicted values ranged between 30 % and 80 % defoliation with *Luiando* showing the smallest variance (Figure 7).

The high error of the super model for *Laukiz 2* (49.94 RMSE) is also visible in the respective histogram as most predictions range between a defoliation of 40 % - 50 % (Figure 7) while in fact most trees of *Laukiz 2* show an actual defoliation of around 20 % (Figure 3).

For the less defoliated plots *Laukiz 1* and *Laukiz 2* a subtle level of separation between trees and bare ground is visible (Figure 6). This is not the case for the other two higher defoliated plots for which bare ground and defoliated predictions are mainly in the same value range (around 60 % - 80 %).

[ADD FIGURE WITH SENTINEL2 PREDICTION]

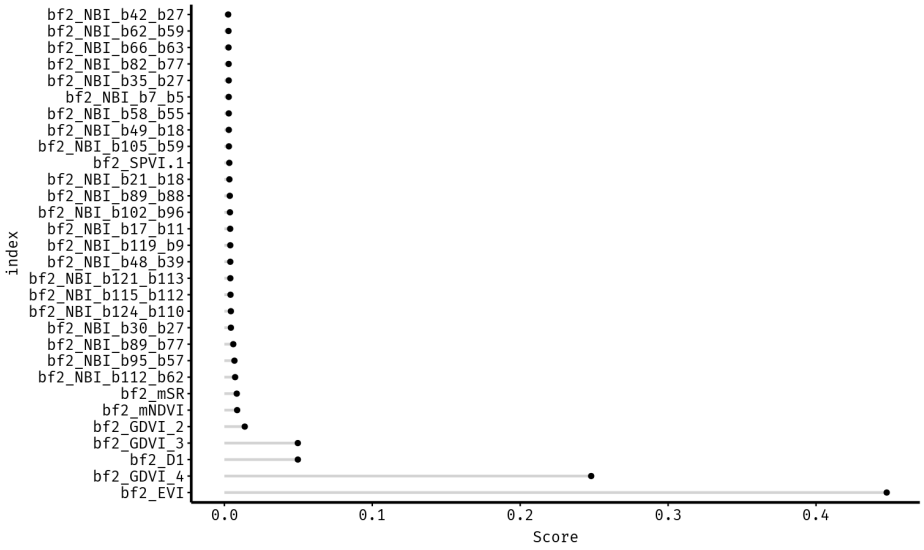

Figure 5: The 30 most important variables as estimated by the internal variable importance measure of the *xgboost* algorithm. The higher the score, the more important the feature. "bf2" notes that a buffer of 2 meter was used to extract the variable information to the tree observation. "NRI" means that a normalized ratio index with the subsequent bands was calculated. Features without "NRI" prefix are vegetation indices, e.g. "bf2_EVI".

.

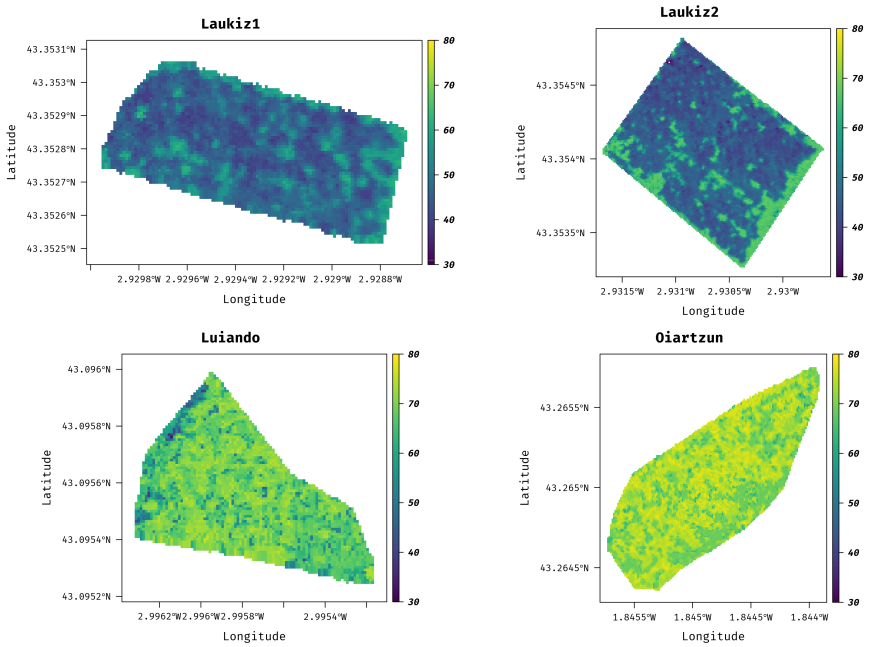

Figure 6: Spatially predicted defoliation (in %) from *xgboost* of *Laukiz 1*, *Laukiz 2*, *Luiando* and *Oiartzun*.

5. Discussion

5.1. Derivation of indices

The buffer of 2 m that we used to generate the index value for each observation can be seen critical. When using no buffer at all, the possibility is high that a pixel value gets assigned to the tree observation that does not spatially match (due to the geometric offset of 1 m in the hyperspectral data). Using a buffer of more than 2 meters would increase the probability of merging information from other trees into the pixel value, blurring the actual value of the tree observation. Thats why in our view using a buffer of 2 m was the best compromise here.

another critical point is that the exact number of contributing pixels to the final index value of an observation cannot be determined as it depends on the location of the tree within the pixel grid. As the buffer is a circle, it depends on the exact location of a tree observation within a pixel how much surrounding

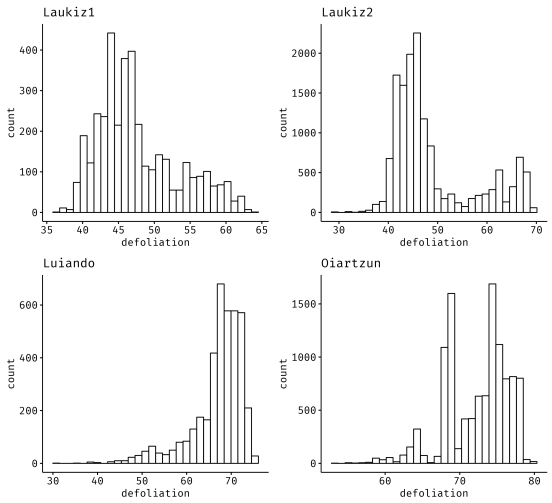

Figure 7: Histograms of predicted defoliation (in %) from *xgboost* of *Laukiz 1*, *Laukiz 2*, *Luiando* and *Oiartzun*.

pixels are touched by the buffer. If a tree observation is located at the border of the plot, some directions of the buffer will contain no values and the subsequent index value will be calculated using less pixels than if the tree observation is located in the middle of the plot.

All these points introduced a bias of an unknown magnitude into the data. This has to be considered when making interpretations about the outcome of this study.

5.2. RMSE vs. plot characteristics

Relating the modeling error to plot characteristics (mean point density, coefficient of variation) did not show a clear picture: For both comparisons, *Laukiz 2* did not follow the pattern that was observed from the other three plots (Figure 4) of having an increase in error with an increase in mean point density and coefficient of variation.

It needs to be considered that we only looked at four plots in this work. To make a robust statement about a possible relationship between modeling error and plot characteristics, a larger sample size of plots is needed.

5.3. Predictive Performance

5.3.1. Algorithm benchmarking

The relatively large difference in performance between *RR* (59.10 RMSE) and the machine-learning models (36.23 and 33.26 RMSE) is remarkable. *RR* has shown promising performance results in other studies when many highly-correlated predictors were involved (REFERECES). However, in this study, *RR* was not able to achieve a sufficient performance score compared to *SVM* and *xgboost* even though its hyperparameter λ was properly tuned using SMBO.

While *xgboost* shows a slightly better performance than *SVM*, the latter has the advantage of only having two hyperparameters that need to be tuned. This results in a shorter runtime. Nevertheless, *xgboost* showed the best performance and was subsequently selected to fit the models on the plot level and for the spatial prediction.

An important point that needs to be considered when interpreting the performance results is that we only related defoliation to indices derived from remote sensing data. Possible other local variables that could help in predicting defoliation were not considered. One example here is tree age: The older a tree the more vulnerable it may be to pathogen infections causing defoliation. However, such predictors would not be available for a spatial prediction scenario and one of the main goals of this study is to relate defoliation to variables that are available on a larger scale (e.g. remote sensing indices).

5.3.2. Single models vs. super model

It is expected that models that were trained using observations from the respective plot only only achieve a better performance than the super model which was trained on observations from multiple plots (Table 3). The low performance on *Laukiz 2* for the super model is most likely due to the difference of this plot to all others: The fitted model on *Laukiz 1*, *Luiando* and *Oiartzun* is not capable of reaching a good performance when *Laukiz 2* is the evaluation dataset. This is not surprising as *Laukiz 2* shows substantially different plot characteristics

compared to all others plots in terms of the distribution of the response variable *defoliation* (Figure 3) and the mean point density of trees (Figure 4).

The low error for Luiando (8.30 RMSE) for the plot model validates the approach of relating defoliation to vegetation indices and NRIs. The overall error of the super model (33.26 RSME) is expected to decrease if more plots would be available for training. However, even if the fitted model would include at least one instance of every plot showing unique characteristics (e.g. here *Laukiz 2* is substantially different to the others), the overall error would not become smaller than the error achieved when using data from the respective plot only.

An interesting find is that the model with only seven variables shows a better overall performance than the model with all 7471 variables (Table 3). This leads to the conclusion that adding as many variables as possible to a model will not necessarily improve its performance but instead add noise to the model. Too much information can be problematic for a model as it will have a hard time distinguishing between noise and important information in the variables (Li et al., 2017; Guyon & Elisseeff, 2003). However, to find the most important variables in the first place and to check for the performance difference, a model with all variables needs to fitted first. Using a model with only a few predictors does not only simplify prediction tasks but also reduces runtime for hyperparameter tuning and performance estimation (Liu & Motoda, 2007).

5.4. Variable importance

There are some downsides using the internal variable importance approach of *xgboost*: Due to the contribution of three different parts to the overall importance score it is complicated to understand why a specific feature was selected. Furthermore the importance calculation approach is only valid for this algorithm and cannot be compared to others. Nevertheless, as we only relied on the variable importance for this specific algorithm, using the internal *xgboost* approach was sufficient for this work.

It is expected that vegetation indices are most important for the model as these are most sensitive to changes in vegetation health (Croft et al., 2014). Even though we are not directly looking at vegetation health but using the level of defoliation as a proxy for tree health, vegetation indices were most important for the fitted model of this study (Figure 5).
Vegetation indices can help assessing defoliation in two ways:

- Trees that show a high level of defoliation do also reflect their bad health status through the remaining foliation.
- Defoliated trees have more influence of bare ground information in their pixel values and will therefore be classified as defoliated by the model.

Even though no NRI made it among the most important variables in this study (Figure 5) (stating that the first seven of this study are the most important ones), it is notable that all ranks from 8 - 30 are occupied by NRI (Figure 5). However, their relative importance was very small compared to the first seven ranks.

Restricting the important indices of this study to the first seven ranks can be seen critical as we only based the selection on a visual inspection of the variable importance results (Figure 5). The decision to make a cut between rank seven and eight was based on a combination of two facts:

- Using only vegetation indices is easier for large scale predictions using satellites like Sentinel-2 (most NRIs cannot be used with it because the spectral bands do not exist).
- The drop in the importance score of the variable importance results (Figure 5).

However, based on these two points, we could also have made the cut between rank five and 6 but including the two vegetation indices at rank six and seven will eventually improve the model and does not increase runtime.

5.5. Spatial prediction

The spatial predictions showed that the model tries to avoid making extreme predictions since most values ranged between a defoliation of 30 % and 80 %. This behavior is mainly triggered by overfitting on the observations of the *Laukiz 1*, *Luiando* and *Oiartzun* training plots. The overfit then causes a high prediction error for *Laukiz 2* for which a lot of defoliation values actually range between 0 % and 20 %. The fitted model would need more training samples of plots with an unusual defoliation distribution to become more robust and achieve a better overall performance.

5.6. Comparison to other studies

Other studies analyzing defoliation found that reflectance differences between defoliated and 30 % defoliated trees of up to 10% exist in the Near-Infrared (NIR) region (Rengarajan & Schott, 2016). This corresponds with the finding of this study that the most important variables are located in the NIR region.

Goodbody et al. (2018) used NDVI and structural measures in as inputs for a partial least squares analysis to model defoliation caused by the spruce budworm. Results showed that metrics from spectral features were most important. Incorporating spectral metrics could be a possible enhancement for future studies.

Townsend et al. (2012) used Landsat data to model defoliation caused by insect herbivores. They found that the Normalized Difference Infrared Index (NDII) ($\frac{Band4-Band5}{Band4+Band5}$) and the moisture stress index ($\frac{Band5}{Band4}$) gave better results than using NDVI. Overall, they used 10 vegetation indices derived from Landsat data.

MODIS data was used by de Beurs & Townsend (2008) to model defoliation caused by the gypsy moth using vegetation indices such as NDVI, EVI, NDWI and NDII.

All of these examples validate the approach of using vegetation indices to model defoliation. Even though the spatial resolution of the data in these studies varied between hundreds of meters (MODIS) de Beurs & Townsend (2008) and

centimeters Goodbody et al. (2018), high resolution data is preferred to fit accurate models. Also, the importance of certain indices (e.g. NDVI, EVI) will vary based on the data and resolution. The finding of this work that the vegetation indices GDVI and EVI are most important for the fitted model could not be verified by other studies. However, the presented ones did often only use a small subset of the vegetation indices that were used in this study.

We could not find a study that used recent machine-learning techniques in combination with a high amount of variables to model defoliation. This fact highlights the importance of this work and will hopefully encourage scientists to use machine-learning techniques, feature selection methods and a wider selection of vegetation indices when assessing defoliation in the future.

6. Conclusion

In this work we used various indices derived from hyperspectral remote sensing data to estimate defoliationas a proxy for tree health in northern Spain. In the algorithm comparison *xgboost* showed the best performance among the tested ones. Even though RR is able to handle highly correlated data, it was far from achieving an acceptable performance in this work.

The fitted models on the plot level showed a promising performance with RMSE values between 8 and 20 which validated the approach of relating defoliation to remote sensing indices. The performance of the model containing all data (four plots) was acceptable (RMSE 29.59) but can be improved by adding more observations from other plots in future studies.

The spatial prediction showed a satisfying result making it possible to distinguish highly defoliated plots from plots with low defoliation easily. The most important indices for the fitted model were widely known vegetations indices such as EVI, GDVI or NDVI. In future studies it would be interesting to link the indices of this work to other indicators of forest/vegetation health and analyze their importance and the model performances.

7. Appendix

Appendix A. Spectral signatures of each plot

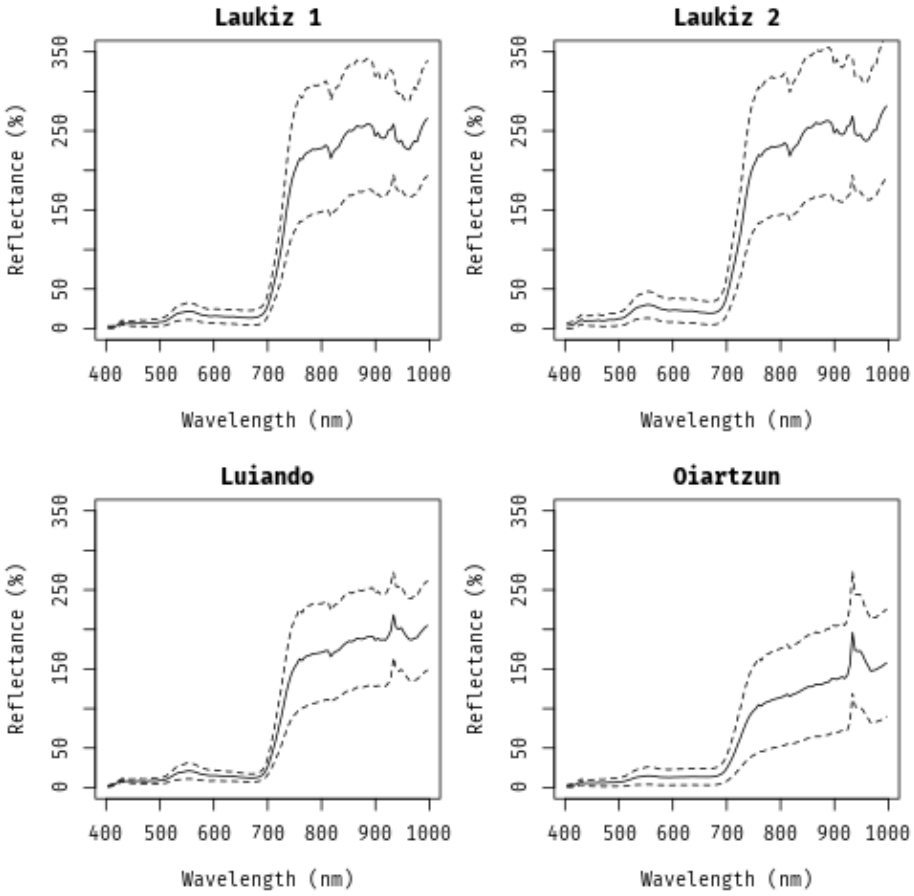

Figure A.8: Spectral signatures (mean and standard deviation) of each plot.

References

Adamczyk, J., & Osberger, A. (2015). Red-edge vegetation indices for detecting and assessing disturbances in Norway spruce dominated mountain forests. *International Journal of Applied Earth Observation and Geoinformation*, *37*, 90–99. doi:10/f64b6c. 00009.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. doi:10/f8ndk8. 00281.

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

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.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. doi:10/d8zjwq.

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 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 01130.

Croft, H., Chen, J. M., & Zhang, Y. (2014). The applicability of empirical vegetation indices for determining leaf chlorophyll content over different leaf and canopy structures. *Ecological Complexity*, *17*, 119–130. doi:10/gdxvd7.

de Beurs, K. M., & Townsend, P. A. (2008). Estimating the effect of gypsy moth defoliation using MODIS. *Remote Sensing of Environment*, *112*, 3983–3990. doi:10/fpqhrc.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*, 1–22. 05097.

Ganley, R. J., Watt, M. S., Manning, L., & Iturritxa, E. (2009). A global climatic risk assessment of pitch canker disease. *Canadian Journal of Forest Research*, *39*, 2246–2256. doi:10/bmj3nk. 00053.

Goodbody, T. R. H., Coops, N. C., Hermosilla, T., Tompalski, P., McCartney, G., & MacLean, D. A. (2018). Digital aerial photogrammetry for assessing cumulative spruce budworm defoliation and enhancing forest inventories at a landscape-level. *ISPRS Journal of Photogrammetry and Remote Sensing*, *142*, 1–11. doi:10/gdxvfk.

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*, 55–67. doi:10/gdxvdp.

Huete, A. R., Liu, H. Q., Batchily, K., & van Leeuwen, W. (1997). A comparison of vegetation indices over a global set of TM images for EOS-MODIS. *Remote Sensing of Environment*, *59*, 440–451. doi:10/bgtpgv. 01474.

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 00678.

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.

Iturritxa, E., Trask, T., Mesanza, N., Raposo, R., Elvira-Recuenco, M., & Patten, C. L. (2017). Biocontrol of Fusarium circinatum Infection of Young Pinus radiata Trees. *Forests*, *8*, 32. doi:10/f9t3d8. 00000.

Jiang, Y., Wang, T., de Bie, C. A. J. M., Skidmore, A. K., Liu, X., Song, S., Zhang, L., Wang, J., & Shao, X. (2014). Satellite-derived vegetation indices contribute significantly to the prediction of epiphyllous liverworts. *Ecological Indicators*, *38*, 72–80. doi:10/f5q4b4. 00018.

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.

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*, 3–10. doi:10/f79ddn. 00069.

Lehnert, L. W., Meyer, H., & Bendix, J. (2018). *Hsdar: Manage, Analyse and Simulate Hyperspectral Data in R*. 00012 R package version 0.7.1.

Lelong, C. C. D., Roger, J.-M., Brégand, S., Dubertret, F., Lanore, M., Sitorus, N. A., Raharjo, D. A., & Caliman, J.-P. (2010). Evaluation of Oil-Palm Fungal Disease Infestation with Canopy Hyperspectral Reflectance Data. *Sensors*, *10*, 734–747. doi:10/bb8wm6. 00045.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature Selection: A Data Perspective. *ACM Comput. Surv.*, *50*, 94:1–94:45. doi:10/gcvjw3.

Liu, H., & Motoda, H. (2007). *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC.

Lu, D., Chen, Q., Wang, G., Liu, L., Li, G., & Moran, E. (2016). A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems. *International Journal of Digital Earth*, *9*, 63–105. doi:10/gdthzv. 00111.

Martinez del Castillo, E., García-Martin, A., Longares Aladrén, L. A., & de Luis, M. (2015). Evaluation of forest cover change using remote sensing techniques and landscape metrics in Moncayo Natural Park (Spain). *Applied Geography*, *62*, 247–255. doi:10/gdthzt. 00029.

Mesanza, N., Iturritxa, E., & Patten, C. L. (2016). Native rhizobacteria as biocontrol agents of Heterobasidion annosum s.s. and Armillaria mellea infection of Pinus radiata. *Biological Control*, *101*, 8–16. doi:10/f8xnp3. 00004.

Michez, A., Piégay, H., Lisein, J., Claessens, H., & Lejeune, P. (2016). Classification of riparian forest species and health condition using multi-temporal and hyperspatial imagery from unmanned aerial system. *Environmental Monitoring and Assessment*, *188*, 146. doi:10/f8q9wp. 00037.

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. 88058 R version 3.4.4.

Rengarajan, R., & Schott, J. R. (2016). Modeling forest defoliation using simulated BRDF and assessing its effect on reflectance and sensor reaching radiance. In *Remote Sensing and Modeling of Ecosystems for Sustainability XIII* (p. 997503). International Society for Optics and Photonics volume 9975. doi:10.1117/12.2235391.

Sexton, J. O., Noojipady, P., Anand, A., Song, X.-P., McMahon, S., Huang, C., Feng, M., Channan, S., & Townshend, J. R. (2015). A model for the propagation of uncertainty from continuous estimates of tree cover to categorical forest cover and change. *Remote Sensing of Environment*, *156*, 418–425. doi:10/f6v7zc. 00038.

Sims, D. A., & Gamon, J. A. (2002). Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. *Remote Sensing of Environment*, *81*, 337–354. doi:10/fb9nnj. 01985.

Sinha, S., Jeganathan, C., Sharma, L. K., & Nathawat, M. S. (2015). A review of radar remote sensing for biomass estimation. *International Journal of Environmental Science and Technology*, *12*, 1779–1792. doi:10/gdthzw. 00043.

Townsend, P. A., Singh, A., Foster, J. R., Rehberg, N. J., Kingdon, C. C., Eshleman, K. N., & Seagle, S. W. (2012). A general Landsat model to predict canopy defoliation in broadleaf deciduous forests. *Remote Sensing of Environment*, *119*, 255–265. doi:10/fzwbdw.

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.

Wu, C., Niu, Z., Tang, Q., & Huang, W. (2008). Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation. *Agricultural and Forest Meteorology*, *148*, 1230–1241. doi:10/dhcp6r.

Zarco-Tejada, P. J., Pushnik, J. C., Dobrowski, S., & Ustin, S. L. (2003). Steady-state chlorophyll a fluorescence detection from canopy derivative reflectance and double-peak red-edge effects. *Remote Sensing of Environment*, *84*, 283–294. doi:10/c8gjtt. 00238.

Zhang, K., Thapa, B., Ross, M., & Gann, D. (2016). Remote sensing of seasonal changes and disturbances in mangrove forest: A case study from South

Florida. *Ecosphere*, (p. e01366). doi:10.1002/ecs2.1366@10.1002/(ISSN)2150-8925.ExtremeColdSpells.