



Contents lists available at ScienceDirect

## ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: [www.elsevier.com/locate/isprsjprs](http://www.elsevier.com/locate/isprsjprs)

# The Naïve Overfitting Index Selection (NOIS): A new method to optimize model complexity for hyperspectral data



Alby D. Rocha<sup>a,\*</sup>, Thomas A. Groen<sup>a</sup>, Andrew K. Skidmore<sup>a,b</sup>, Roshanak Darvishzadeh<sup>a</sup>, Louise Willemen<sup>a</sup>

<sup>a</sup> Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Hengelosestraat 99, P.O. Box 217, 7500 AE, Enschede, The Netherlands

<sup>b</sup> Department of Environmental Science, Macquarie University, NSW 2106, Australia

## ARTICLE INFO

## Article history:

Received 9 March 2017

Received in revised form 21 July 2017

Accepted 19 September 2017

## Keywords:

Remote sensing

Model tuning

Cross-validation

Prediction accuracy

Dimensionality

Multicollinearity

## ABSTRACT

The growing number of narrow spectral bands in hyperspectral remote sensing improves the capacity to describe and predict biological processes in ecosystems. But it also poses a challenge to fit empirical models based on such high dimensional data, which often contain correlated and noisy predictors. As sample sizes, to train and validate empirical models, seem not to be increasing at the same rate, overfitting has become a serious concern. Overly complex models lead to overfitting by capturing more than the underlying relationship, and also through fitting random noise in the data. Many regression techniques claim to overcome these problems by using different strategies to constrain complexity, such as limiting the number of terms in the model, by creating latent variables or by shrinking parameter coefficients. This paper is proposing a new method, named Naïve Overfitting Index Selection (NOIS), which makes use of artificially generated spectra, to quantify the relative model overfitting and to select an optimal model complexity supported by the data. The robustness of this new method is assessed by comparing it to a traditional model selection based on cross-validation. The optimal model complexity is determined for seven different regression techniques, such as partial least squares regression, support vector machine, artificial neural network and tree-based regressions using five hyperspectral datasets. The NOIS method selects less complex models, which present accuracies similar to the cross-validation method. The NOIS method reduces the chance of overfitting, thereby avoiding models that present accurate predictions that are only valid for the data used, and too complex to make inferences about the underlying process.

© 2017 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

## 1. Introduction

Data collection using *in situ* measurements is time-consuming and expensive, constraining the availability of information to limited areas and specific periods (Plaza et al., 2009; Muñoz-Huerta et al., 2013; Ramoelo et al., 2012). Remote sensing technologies can mitigate these limitations and provide opportunities to monitor biological processes over wider temporal and spatial scales (Stroppiana et al., 2011; Wilson et al., 2011). The monitoring of biological processes in ecosystems by remote sensing relies mostly on empirical models to predict a variety of biochemical and biophysical properties of vegetation, soil or water (such as nitrogen concentration, organic carbon and biomass stocks), estimated from spectral information (Huber et al., 2008; Kokaly et al., 2009; Nguyen and Lee, 2006; Shepherd et al., 2003; Thiemann and Kaufmann, 2002).

Hyperspectral images present even greater potential, as they consist of many narrow spectral bands that can detect changes in specific regions of the spectrum to which concentrations of such substances or structural characteristics of vegetation can be related (Acevedo et al., 2017; Curran, 1989; Hansen and Schjoerring, 2003; Manolakis et al., 2003; Darvishzadeh et al., 2011). Predictive empirical models face two important challenges when using hyperspectral data, as a result of the high dimensions involved: (1) there is a large number of predictors relative to the number of observations to fit the model (Zhao et al., 2013) and (2) there is strong multicollinearity in the predictors, resulting in highly redundant reflectance values at close spectral distances (Dormann et al., 2013). Multicollinearity is enhanced when the sample originates from a homogeneous land cover type, because similar surfaces result in more similar reflectance values across wavelengths (Cho et al., 2007).

High dimensionality and multicollinearity complicate the identification of relevant spectral bands to predict the response variable and the estimation of their regression coefficients, since

\* Corresponding author.

E-mail address: [a.duarterocha@utwente.nl](mailto:a.duarterocha@utwente.nl) (A.D. Rocha).

several explanatory variables can be written as a linear combination of the others (Gelman and Hill, 2006; James et al., 2013; Kuhn and Johnson, 2013). Also, multicollinearity can falsely increase prediction accuracy when a variable that has no correlation with the response but correlates well with another variable that does correlate with the response is used in the model (Meehl, 1945).

There are two main solutions to process high dimensional and multicollinear hyperspectral data with regression models (Stroppiana et al., 2011). Firstly, the number of predictors (bands) can be reduced before fitting an ordinary least squares (OLS) type of model. This can be achieved by selecting a spectral index based on a priori knowledge, by grouping bands to create latent variables using techniques such as principal components and wavelets (Bioucas-Dias and Nascimento, 2008; Bruce et al., 2002), or by finding an optimal combination of bands using stepwise multiple linear regression or genetic algorithms (Ramoelo et al., 2012; Darvishzadeh et al., 2008; Schlerf et al., 2010). Secondly, models can be fitted using all explanatory variables based on non-ordinary least square techniques (non-OLS). Commonly used non-OLS regressions applied to remote sensing are: dimension reductions such as Partial Least Squares Regression (Carvalho et al., 2013; Martin et al., 2008), tree-based ensembles such as Random Forest or Boosted Regression Trees (Abdel-Rahman et al., 2013; Feilhauer et al., 2015), support vector machine regression (Feilhauer et al., 2015; Mountrakis et al., 2011), and artificial neural networks (Farifteh et al., 2007; Mirzaie et al., 2014; Skidmore et al., 1997).

Regardless of whether or not there is a true relationship between predictors (spectral bands) and the response variable, using a large set of predictors in relation to the number of observations with a supervised method is likely to cause model overfitting (Hastie et al., 2009). A model may fit the training set almost perfectly, but lead to lower accuracy predictions when applied to new samples or a testing set (Gelman et al., 2014; Lee et al., 2004).

Overfitting is the situation where overly complex models capture more than the underlying relationship, and also fit random and systematic errors (noise) in the data (James et al., 2013). This is even more of a concern in non-OLS regression techniques that use the residuals from a model fitted in a previous step as new response in a subsequent step (Hastie et al., 2009). Also, predictors derived from hyperspectral data may present a considerable amount of noise in some regions of the spectra, depending on the capacity to control variations in illumination and atmospheric conditions during the measurements (Manolakis et al., 2003).

Therefore, empirical models need to be constrained regarding the number of predictors or parameters included to avoid overfitting. The type and number of terms per predictor used in a fitted model varies between techniques, including parameter coefficients, interaction, second order terms, nodes, trees, and so on (James et al., 2013). The number of terms used determines the level of model complexity (Hastie et al., 2009). The maximum model complexity to avoid overfitting depends greatly on the number of observations relative to the number of predictors used for fitting the model (Fassnacht et al., 2014; Kuhn and Johnson, 2013).

The procedure to select an optimal model complexity that balances the trade-off between accuracy and overfitting is called the tuning process (James et al., 2013). This process is typically performed by adjusting or “tuning” parameters that control the number of terms in the model, such as the “number of components” in partial least squares regression or “cost” in support vector machine regression (Hastie et al., 2009).

The optimal model complexity cannot be calculated directly from the data but can be defined by fitting models with different complexities and evaluating their prediction accuracy (Krstajic et al., 2014; Verrelst et al., 2012). Some metrics to assess model

accuracy, such as the adjusted coefficient of determination ( $R^2_{adj}$ ), Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC) are inappropriate for selecting the best model complexity from different non-OLS regressions as the degrees of freedom are impossible to determine or compare between regression techniques (James et al., 2013). Often the coefficient of determination ( $R^2$ ) of the simple regression between observed data and model predictions is presented as accuracy metric for non-OLS regressions.

Assessing model performance with the same dataset to which it was fitted, greater complexity automatically means higher accuracy because error declines monotonically as complexity increases (James et al., 2013). Therefore, it is inappropriate to use the same dataset to select model complexity and to report the prediction accuracy, requiring a method that separates the data into training and testing (sub) sets (Esbensen and Geladi, 2010). Whether the most suitable splitting of data will be based on approaches such as cross-validation or bootstrapping or even the collection of an independent validation set, will depend on the sample design and data availability (Fassnacht et al., 2014; Kuhn and Johnson, 2013).

Independent validation can be achieved by splitting the existing data into training and testing sets, keeping the validation set apart to quantify the accuracy of each level of model complexity. In this case, the fitted model will be considered overfitted when the accuracy of an independent validation set is significantly lower than the accuracy of the training set (Dormann et al., 2013). Although non-representative samples or samples from different populations can also lead to lower accuracies, overfitting is related exclusively to the process of modelling (Hawkins, 2004).

Despite being widely employed, splitting a single dataset into a training and a testing set may only have a limited ability to characterize the uncertainty in the predictions (Kuhn and Johnson, 2013). Model performance can be highly variable depending on the size of the testing set and the variability in the population that was sampled (Darvishzadeh et al., 2008; Kuhn and Johnson, 2013). In addition, when the number of observations is limited, most of them need to be allocated to calibrate the model (Hawkins, 2004). In these cases, cross-validation is an alternative approach to evaluate a model as it randomly splits off multiple combinations of training and validation sets (James et al., 2013).

Cross-validation estimation can produce a reasonable indication of overfitting, and has shown, in general, to be efficient in finding optimal model complexity, giving a satisfactory estimation of the predictive performance (Kuhn and Johnson, 2013). A widely used cross-validation method is the K-fold approach, based on the random splitting of observations into  $k$  groups of similar size (James et al., 2013). This procedure can be repeated many times, using a different selection of folds as testing set each time, to increase the robustness (Krstajic et al., 2014).

Being widely accepted as tuning method, cross-validation procedures may still select overly complex model in the case of hyperspectral data. Hawkins (2004) stated that a model overfits when it is more complex than another model that performs equally well. Also, robust cross-validation can be computationally intensive and thus time consuming for high dimensional data such as hyperspectral datasets, depending on the number of parameters to tune (Hastie et al., 2009; Krstajic et al., 2014). Another limitation is that tuning parameters are often not comparable between different modelling methods and the available methods do not evaluate the adequacy of the model complexity selected from different non-OLS regressions (Kuhn and Johnson, 2013). In addition, cross-validation tuning methods do not quantify the amount of overfitting as the (true) maximum model contribution for a given set of predictors is normally unknown, making it difficult to fairly compare the accuracy of different regression techniques.

The novelty of this study is to present a new tuning method for modelling hyperspectral data that overcomes these limitations of existing techniques. The new method is termed Naïve Overfitting Index Selection (NOIS) and it (1) provides an efficient and structured method to tune over a range of parameters, showing a gradual increase in model complexity, for non-OLS regressions; (2) determines the maximum level of model complexity supported by a specific data structure without overfitting; and (3) quantifies the relative amount of overfitting across regression techniques consistently, highlighting the trade-off between prediction accuracy and overfitting.

The performance of models derived from this tuning method is compared to a tuning method based on robust cross-validation, and tested using different hyperspectral datasets and regression techniques.

## 2. Methods

The Naïve Overfitting Index Selection (NOIS) requires three steps. Firstly, a dataset of artificial spectra is generated, having the same data structure as the original spectra, but uncorrelated with the response variable. Secondly, the amount of overfitting at different levels of model complexity is calculated using the generated spectra as predictors. Thirdly, a model complexity is selected based on an overfitting threshold that is compatible with the data structure and comparable between datasets and regression techniques. In this paper, the NOIS method is subsequently compared with a traditional cross-validation tuning method by fitting seven commonly used non-OLS regression techniques to five hyperspectral datasets.

### 2.1. Database

A selection of hyperspectral datasets (Table 1) composed of different surfaces and measured using diverse instruments under singular conditions is used to assess the robustness of the NOIS method. These datasets originate from various scientific contexts, representing plausible combinations of number of observations versus number of predictors. These include a dataset with a number of observations higher than the number of spectral bands (e.g., the soil organic carbon dataset), as well as a dataset where the number of observations is considerably smaller than the number of spectral bands (e.g., the leaf water content dataset).

The second last row of Table 1 indicates the risk of multicollinearity in the model, as in hyperspectral data a large proportion of bands can be considered redundant when a specific surface is measured. For example, if a maximum correlation

threshold of 0.75 between any pair of bands is defined as “not being sufficiently different”, only a few individual bands will be considered non-redundant in all datasets, implying a strong risk of multicollinearity.

### 2.2. Generating artificial spectral data

A new dataset of predictors with the same dimensions as the original dataset (Table 1) is generated from a multivariate normal distribution. This generated dataset preserves the number of bands and has an equivalent mean, variance and covariance to those observed in the original spectra. This procedure intends to create predictors that are completely uncorrelated with the response variable, but maintain the data structure of the original predictors (Fig. 1).

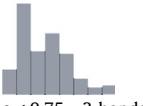

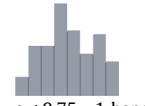
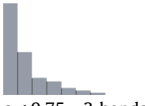
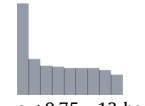
Artificial spectra were generated using the *mvrnorm* function from the MASS package in R version 3.2.5 (Venables and Ripley, 2002; Ripley, 2009; R Core Team, 2016). This function requires a vector of means and a positive-definite symmetric covariance matrix extracted from the original spectra. The generated data was rescaled according to the original spectra, preserving the same reflectance range of each band using the function *rescale* from the package *plotrix* (Lemon, 2006).

The process of generating spectral datasets gives a good indication of the amount of noise present in the predictors (all generated datasets can be found in Appendix B). For instance, the generated spectra for the moisture dataset present all bands as almost completely uncorrelated with the response variable (Appendix B), indicating low noise in the data. Because sand samples allow for well-controlled experiments to be conducted in a laboratory, precise measurements could be made for this dataset. Also, only wavelengths between 350 and 2100 nm are included in the analysis, as wavelengths over 2100 nm are considered by the data provider to have a low signal-to-noise ratio (Nolet et al., 2014). On the other hand, the LWC dataset contains bands between 2500 and 16,700 nm (thermal) and no specific pre-processing in the data has been applied to reduce the noise in the data. A high level of noise in certain regions of the spectra for this dataset can produce generated predictors that may, by chance, still be slightly correlated with the response variable.

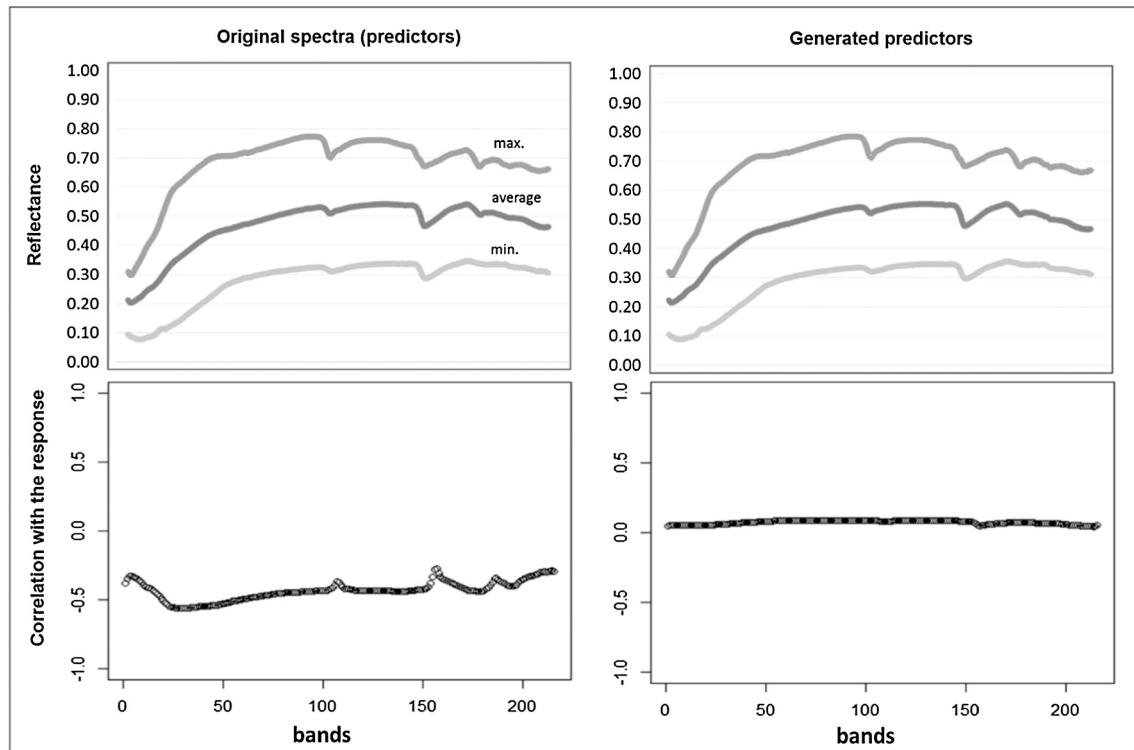
### 2.3. Quantifying overfitting

The generated predictors' dataset ( $X'$ ) preserves the relationship across spectral bands, but makes them uncorrelated (i.e., independent) with the response variable ( $y$ ). Given that  $y$  and  $X'$  are independent, the conditional distribution  $y|X'$  does not depend on the

**Table 1**  
Description and structure of the five selected datasets used for assessing the new tuning method NOIS.

Data structure	Vegetation traits			Soil traits	
	Leaf Area Index (LAI)	Leaf Chlorophyll Content (LCC)	Leaf Water Content (LWC)	Sand Moisture Content (SMC)	Organic Carbon Content (OCC)
Observations (n)	129	111	108	208	292
Predictors (p)	592	126	6612	2150	216
Wavelength	352–2382 nm	436–2485 nm	2500–16,700 nm	350–2500 nm	350–2500 nm
Instrument	GER3700	Hymap	Bruker Vertex 70	ASD Fieldspec	FieldSpec FR
Type	Field	Airborne	Laboratory	Laboratory	Laboratory
Distribution (Y)					
Redundancy (pair of bands)	$\rho < 0.75 = 3$ bands $\rho < 0.90 = 8$ bands	$\rho < 0.75 = 3$ bands $\rho < 0.90 = 5$ bands	$\rho < 0.75 = 1$ band $\rho < 0.90 = 3$ bands	$\rho < 0.75 = 3$ bands $\rho < 0.90 = 3$ bands	$\rho < 0.75 = 13$ bands $\rho < 0.90 = 35$ bands
Published by	Darvishzadeh et al. (2008)	Darvishzadeh et al. (2011)	Buitrago et al. (2016)	Nolet and Roosjen (2014)	ICRAF-ISRIC Spectral Library

Note 1: More detailed information about each dataset can be found in the supplementary material (Appendix A).



**Fig. 1.** Comparison between original and generated reflectance for the soil dataset. The average (dark grey), maximum (lighter grey) and minimum (light grey) from the original spectra (top left) and generated data (top right). And the correlation between the response variable (OCC) and predictors (bands), using original spectra (bottom left) or generated data (bottom right).

value of  $X'$ ,  $E[y|X'] = E[y]$ , and covariance  $y|X'$  should approach zero (Cook and Weisberg, 2009). Consequently, the only information available is the mean of response variable, and any model based on generated spectra as explanatory variables will be referred as a naïve model. It implies that the mean square error of a prediction based on  $X'$  depends only on the variance of the response variable  $\sigma_y^2$ . Therefore, the naïve models, in theory, should not reduce predictor errors (i.e.,  $\hat{y}_i = \bar{y}$  and  $\hat{\sigma}_y^2 \cong \sigma_y^2$ ). Consequently, any reduction in prediction error can be attributed to an increase in the model complexity and thus to overfitting.

The amount of overfitting in a naïve model can be quantified by the difference between the prediction error and the true error (i.e., variance of the response variable), expressed by  $\left(1 - \frac{\hat{\sigma}_y^2}{\sigma_y^2}\right)$ . When values of the predictor error ( $\hat{\sigma}_y^2$ ) are significantly lower than the true error ( $\sigma_y^2$ ) this will indicate model overfitting. The index will achieve a maximum of 1 when the predictor error approaches zero ( $\hat{\sigma}_y^2 \rightarrow 0$ ). In case of no overfitting,  $\sigma_y^2$  and  $\hat{\sigma}_y^2$  should be equal and the overfitting index will approach 0.

### 2.3.1. Naïve overfitting index selection (NOIS)

Since variance or mean square errors depend on the response variable range ( $y$ ), the model accuracy was reported as Root Mean Square Error normalized by the range of the response variable (NRMSE). The naïve overfitting index produced by a specific level of complexity is also calculated based on NRMSE.

naïve overfitting index =  $1 - \left(\frac{\text{NRMSEg}}{\text{NRMSEy}}\right)$ , where :

NRMSEg is the error based on the prediction derived from the naïve model using the generated data ( $X'$ ), and NRMSEy is the error based on the prediction derived from the mean of the response variable ( $y$ ).

For instance, a naïve overfitting index of 0.75 indicates that the true error is falsely reduced 75% by this level of model complexity. In this case, the model complexity should be significantly constrained or the number of observations considerably increased. Negative index values indicate that the model predicts a bigger error than NRMSEy, and the model complexity is constrained excessively (“underfitted”). Because the NRMSEy is only based on the response variable ( $y$ ), and no model contribution is expected from naïve models, the degree of overfitting is directly comparable between regression techniques.

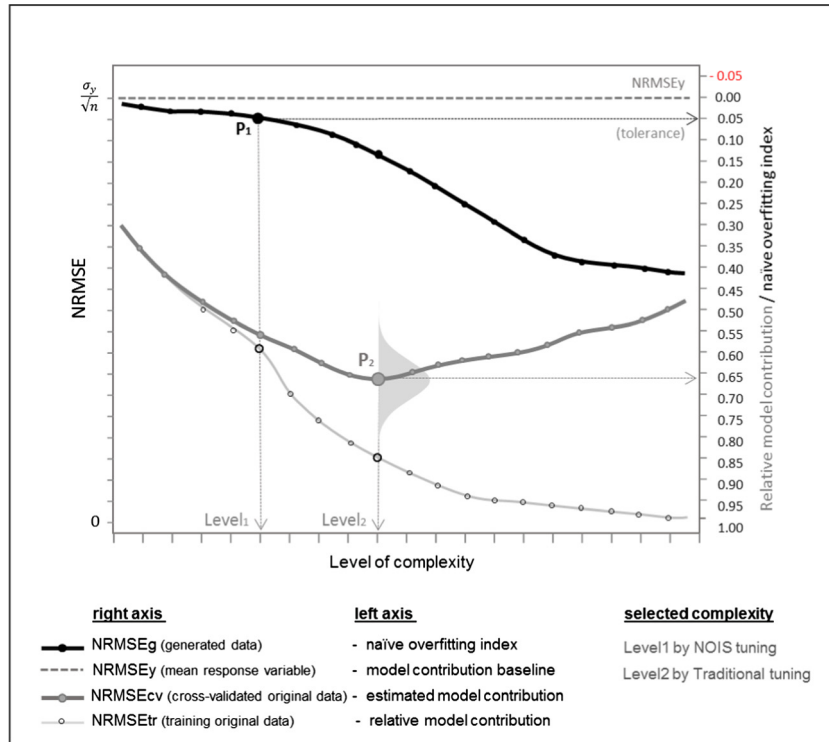
### 2.4. Selecting model complexity

The optimal model complexity supported by the data is selected by increasing tuning parameter values until the naïve overfitting index drops below a pre-defined tolerance (Fig. 2). This tolerance, expressed as a percentage of the NRMSEy, can be adjusted to avoid selecting underfitted models, where the level of complexity is excessively constrained. The tolerance is set at 0.05 in this study, based on the maximum correlation between the response variable ( $y$ ) and the artificially generated spectra (see Appendix B).

In some of the regression techniques there is more than one tuning parameter to define model complexity, requiring a repeat of the procedure for each parameter, whilst keeping other tuning parameters fixed.

Because naïve models are trained and selected using generated artificial spectra, the accuracy can be assessed using the full dataset with original predictors, as opposed to the traditional cross-validation method, which requires multiple splitting of training and validation subsets. Thus, compared to traditional cross validation, the NOIS method avoids uncertainty in the estimation of prediction errors. The naïve overfitting index is defined as the relative model contribution when using generated data (naïve model) and





**Fig. 2.** Process to select the level of model complexity using the NOIS method and the traditional cross-validation tuning. The point P1 represents the level of complexity, where the value of the naïve overfitting index over the NRMSEg curve (generated data) approaches 0.05 (tolerance). The point P2 represents the level of complexity selected by the traditional method (based on original data), where the maximum model contribution is achieved based on minimization of the NRMSE estimate from the cross-validation (NRMSEcv curve).

provides an indication of the amount of overfitting for a given level of model complexity.

2.5. Comparison with a traditional ‘tuning’ method

The NOIS method is compared with a tuning procedure using traditional cross-validation to test its consistency and reliability in the tuning process (Fig. 3). A 10-fold cross-validation is adopted to evaluate the performance of each level of model complexity with the original spectra as predictors. This procedure is randomly repeated ten times, resulting in a combination of 100 subsets of training and validating sets of the original data. The model tuning by means of traditional cross-validation is based on minimization of the cross validated prediction error (NRMSEcv).

The same approach to calculate the naïve overfitting index can be used with the traditional cross validated tuning method to represent the relative model contribution, by replacing the NRMSEg from the naïve model (generated predictors) by the NRMSEcv estimate from the model fitted on the original data. However, as the true model contribution is unknown in this case, the model contribution may be confused with overfitting. Also, the prediction error estimated by cross-validation is based on an average (see Fig. 2 – P2), and the model contribution may vary significantly between sub-models.

2.5.1. Regression techniques tested

Common regression techniques for modelling hyperspectral data are used to compare the NOIS method with a traditional cross-validation method (Table 2). These regression techniques are often selected because they are considered to be reasonably robust regarding highly dimensional data and high multicollinearity (Kuhn and Johnson, 2013; Zhao et al., 2013).

Techniques that contain stochasticity (i.e., Regression Trees and Neural Networks) are initialized with a fixed seed for each level of complexity with the generated (X’) and original (X) predictors. Different initialisations of a Neural Network algorithm may result in differences regarding complexity and relative overfitting (James et al., 2013). When the models are fitted based on the same seed, the results are consistent and comparable. For non-stochastic techniques, the fitted training model will not change, but a fixed seed will guarantee the same selection of k-folds in cross-validation (replicability).

All regression methods are executed in R version 3.2.2 (The R Foundation for Statistical Computing). The package Caret (Classification and Regression Training) is used for fitting models from different regression techniques with cross-validation under the same platform (all the packages are presented in Table 2). The value of all selected tuning parameters for each regression technique and dataset are presented in Appendix C. The response and explanatory variables in each dataset are mean centred and scaled by standard deviation before fitting the models to increase comparability across techniques and datasets (Kuhn, 2008).

3. Results

3.1. Selecting model complexity

The NOIS method, in most cases, identifies lower levels of complexity as suitable than the traditional tuning process using cross-validation does (Fig. 4). Datasets with a higher number of observations (n) in relation to the number of predictors (p) support greater model complexities (Burket, 1943; Hastie et al., 2009). This principle becomes quite clear when the model complexity is selected by the NOIS method, but is less evident when the traditional cross-validation is used.

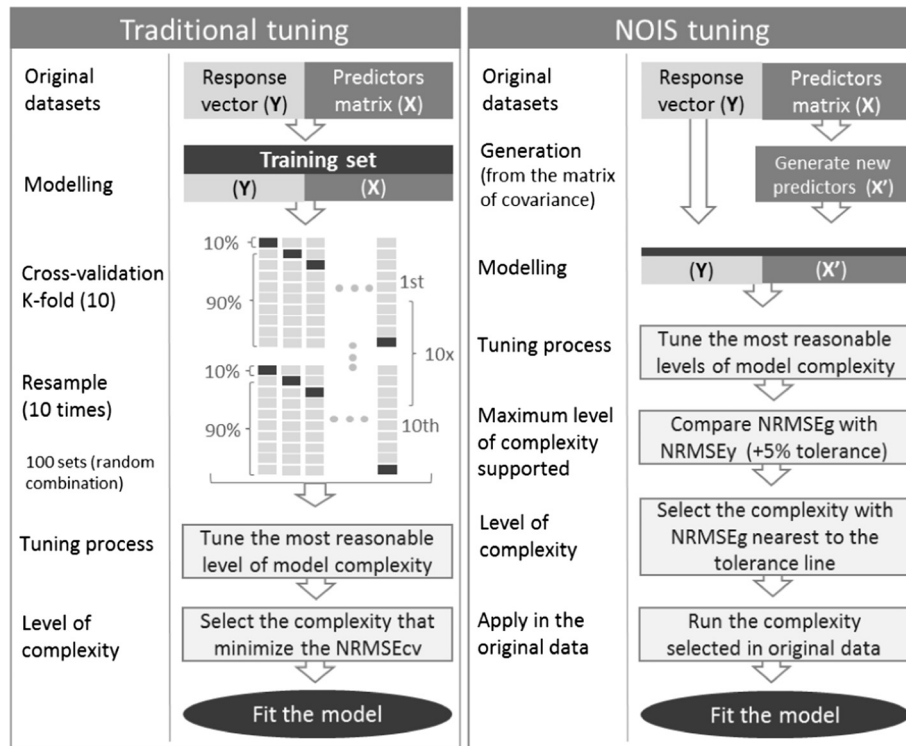


Fig. 3. Comparison between the proposed NOIS method and a traditional approach of cross-validation.

Table 2

List of regression techniques tested, R packages and functions to fit the model, and tuning parameters used for defining model complexity.

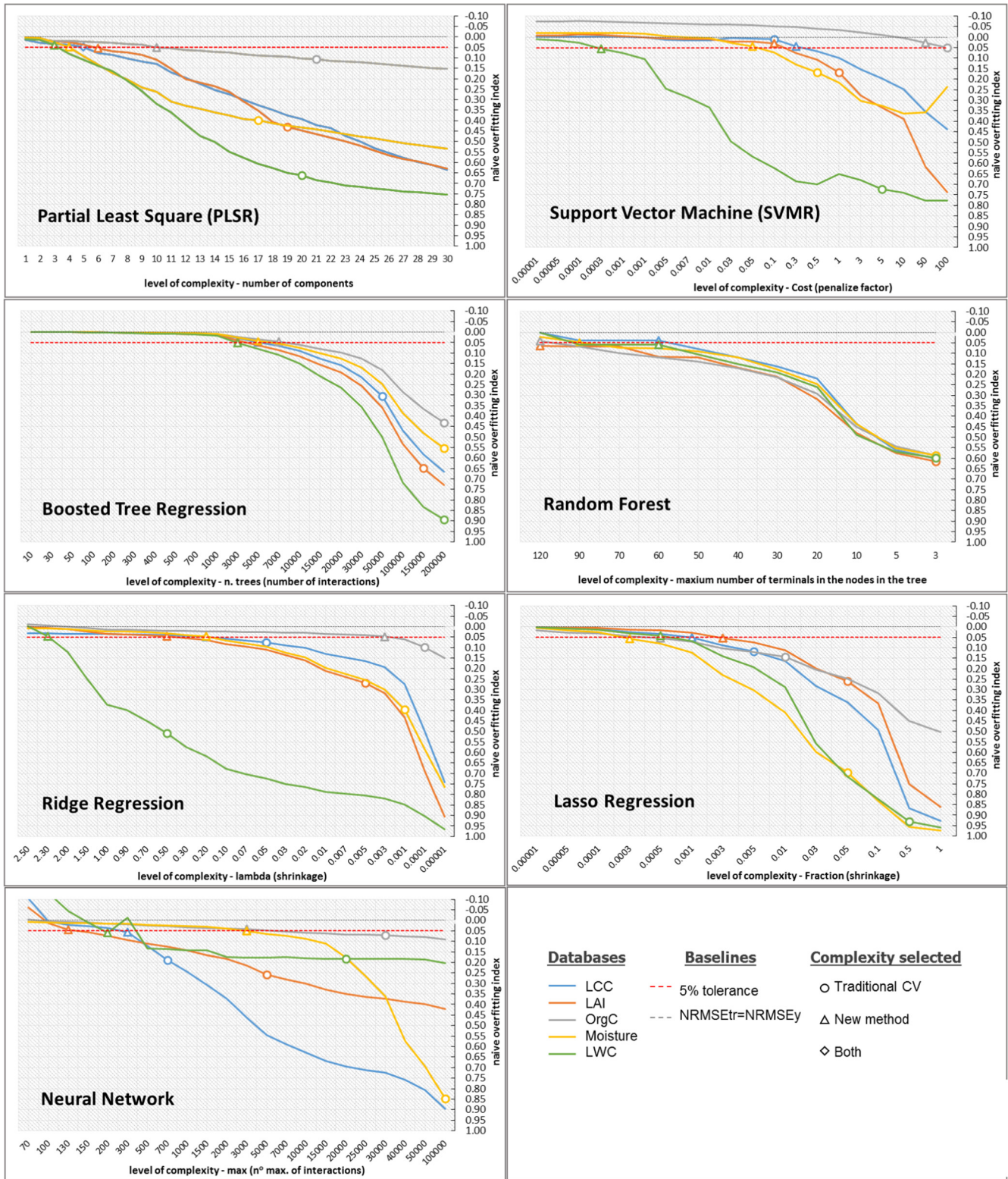
Type	Regression technique	R package and (function)	Tuning parameters to define model complexity
Regression trees (ensemble)	Random forest	randomForest (rf)	mtry (number of randomly selected predictors) maxnode (max number of terminal nodes trees)
	Boosted trees	gbm (gbm)	n.trees (number of interactions) interaction.depth (max. of variable interactions) Shrinkage (shrinkage/learning rate) n.minobsinnode (min. terminal node size)
Artificial neural network	Stuttgart neural network simulator	RSNNS (mlp)	Size (number of hidden units), Max (number max. of interactions), Decay (weight decay/shrinkage/leaning rate)
Dimension reduction	Partial least squares	pls (pls)	ncomp (number of components)
Vector machines	Support vector machines	e1071 (svmLinear)	Cost (cost) Epsilon
Penalized or shrinkage model	The Lasso	Elasticnet (lasso)	Fraction (fraction of full solution – shrinkage)
	Ridge	Elasticnet (ridge)	Lambda (weight decay – shrinkage)

For example, the organic C dataset (grey line in Fig. 4) is the dataset with the highest n/p ratio, namely 292 observations for 216 bands. This dataset also shows the lowest overfitting in almost all the regression techniques. Random Forest models form an exception with similar levels of overfitting occurring regardless of the differences in n/p ratios. In contrast, LWC has the lowest n/p ratio, i.e., 108 observations for 6612 bands, and shows overfitting at relatively low levels of model complexity.

The two tuning methods suggest similar levels of complexity for the LCC dataset for all regression methods. The Support Vector Machine tuning for the LCC dataset generated the only instance where the traditional tuning method selected a lower level of model complexity than the NOIS method did. The reason for this is that the LCC dataset has the smallest number of predictors, which are also the least correlated with the response. As well, the generated spectra present a low level of noise to fit in this dataset (see Appendix B). This leads to select models with low levels of complexity in both methods.

The different tuning approaches result in different levels of overfitting. For example, the PLSR model fitted to the LWC dataset is constrained to a maximum complexity of 3 components (tuning parameter of PLSR) at a naïve overfitting index of 0.04 when the NOIS method is used. On the other hand, the traditional method selects up to 20 components, at a naïve overfitting index of 0.66. Whereas the new method selected a model complexity that, when applied to the original spectra, presents a model contribution of 33%, the traditional method selected a model complexity that presents a model contribution of 48%. The model contribution suggested by cross-validation is only slightly higher than the one indicated by the new method, but has a much higher level of complexity (100,000 more parameter terms in the model). This complexity selected by cross-validation is large enough to present a model contribution of 99% for the training model (NRMSE<sub>tr</sub> = 0.0037).

Some researchers suggest selecting a smaller model complexity if increasing it does not decrease the error by at least 2% (Kooistra



**Fig. 4.** Naive overfitting index selection (NOIS) according to model complexity per regression technique. Note: The range of tuning parameters commonly suggested by software guides or machine learning literature seems unsuitable for the high dimensional hyperspectral data used in this study, and more constrained tuning parameters used to reduce complexity are needed to avoid overfitting. See for example [Kuhn and Johnson \(2013\)](#), [James et al. \(2013\)](#) or <https://cran.r-project.org/> for suggested ranges of tuning parameters.

et al., 2004; Darvishzadeh et al., 2011). In the case of PLSR, selecting a model complexity by the traditional tuning method, with this criterion, 9 rather than 20 components would be selected for optimal complexity. Although less overfitted, this still presents an

NRMSEtr more than twice as small as NRMSEcv, and a level of complexity sufficient to reduce the NRMSEtr one quarter of the NRMSEy in the generated data (with a naïve overfitting index of 0.26). Also, this 2% rule is easily applied in PLSR where there is only



one discrete tuning parameter, but is less applicable in many other regression techniques that present two or more non-discrete tuning parameters for selection.

### 3.2. Quantifying overfitting and model contribution

Fig. 5 presents the cross-validated error (NRMSE<sub>cv</sub>) for models fitted to the original spectra with a level of complexity tuned by the NOIS method and the traditional method for all regression techniques. The boxplot shows the variability in NRMSE<sub>cv</sub> among the sub-models' performance by the repeated k-fold cross-validation.

The results, when presented in ascending order of dimensionality (n versus p), indicate that by increasing the number of predictors relative to the observations, the distance between NRMSE<sub>tr</sub> and NRMSE<sub>cv</sub> in the traditional method increases considerably. On the other hand, the new method results in NRMSE<sub>tr</sub> values that are very similar to the NRMSE<sub>cv</sub> values. The amount of overfitting (bars on Fig. 5) for the model complexities selected by the new method are all controlled at a tolerance around 0.05.

The model complexities selected by the traditional method present much higher levels of overfitting for a number of scenarios. In the most extreme case a naïve overfitting index of around 0.90 was found using traditional tuning (Random forest applied to the LWC dataset), suggesting that the error can be reduced to 10% with non-informative predictors by selecting an overly complex model. The results indicate that the new method selects models that are less likely to be overfitted, while in most cases showing similar accuracy.

### 3.3. Comparison between methods

PLSR and SVMR models show only small differences in performance between the levels of complexity selected by the traditional and the new method. These regression techniques also present results that are more consistent across different data structures and different capacities of explaining the response by the predictors. The distribution of NRMSE<sub>cv</sub> between models selected in both methods is mostly similar, yet the level of complexity for the NOIS method is usually significantly smaller than for the traditional method.

While the model selected by the NOIS method presents a single value of prediction error, the cross-validation procedure presents an average of hundred combinations of training and validation sets. The more random noise there is present in the original spectral signal, the more uncertainty is presented in the cross-validation estimation. This is noticeable when comparing the variability of the cross-validation estimates between the Moisture and LWC datasets (Fig. 5). This can also be derived from the higher capacity to generate artificial predictors that are uncorrelated with the response variables in the first step of the NOIS method (see Appendix B).

Taking the LWC original dataset as an example, the relative model contribution of the selected models for different regressions is between 0.66 and 0.99 for the training models, while the model contribution from cross-validation is between 0.16 and 0.49. Such differences may be due to the smallest model contribution coming from an underfitted model (NRMSE<sub>cv</sub> = 0.218) and the highest from a highly overfitted model (NRMSE<sub>tr</sub> = 0.004). Based on these



Fig. 5. Boxplots of the NRMSE distribution from 100 cross-validated models fitted on the original bands with a model complexity selected by the traditional and NOIS method. The bars represent the naïve overfitting index of the model complexity selected. The circles indicate the NRMSE<sub>tr</sub> using all the observations.



results, it is difficult to decide what the most reasonable estimation of accuracy is given the available predictors to explain the response variable LWC. However, concluding that choosing a model with a complexity that minimizes NRMSEcv does not guarantee generalizable non-OLS model predictions. In the proposed NOIS method, the model is selected by the maximum complexity that is supported by the data structure without overfitting, and the accuracy is a single calculated value for that particular model using the original data.

Another limitation of the traditional method is the effect of intensive cross-validation, resulting in a low capacity to indicate overfitting in complex models when the number of observations is insufficient. Fig. 6 presents the difference between cross-validation error estimates (NRMSEcv) and training model errors (NRMSEtr) for the tuning process of PLSR as an example using the traditional method.

As observed in the LCC plot (Fig. 6e) the NRMSEcv decreases to a certain level of complexity, after which it starts to increase, while the NRMSEtr further decreases. The optimal complexity for this dataset occurs at a point where the difference between the NRMSEtr and NRMSEcv is not too great (Hastie et al., 2009; Schlerf and Atzberger, 2006). This, however, is not observed in datasets where the number of observations is much smaller than the number of predictors, such as for LWC and Moisture. After the fourth component, NRMSEtr and NRMSEcv start to bifurcate in the LWC dataset (Fig. 6b). While NRMSEtr reduces to approximately zero when twenty or more components are included, NRMSEcv remains at an almost steady value for nine or more components.

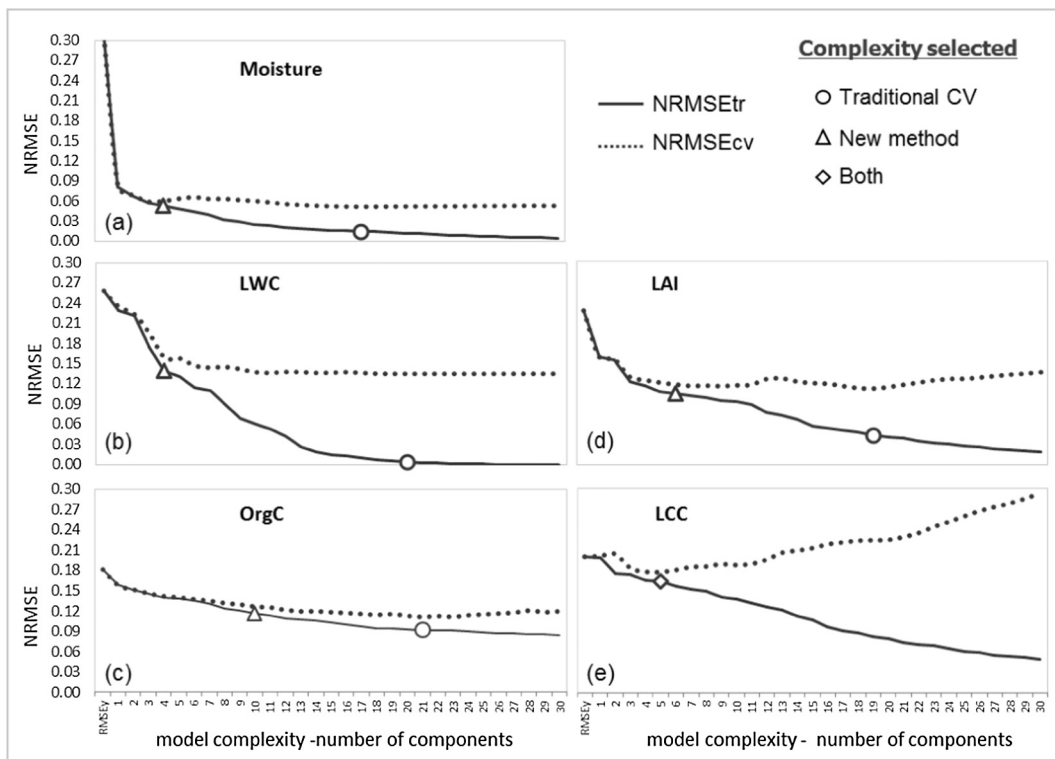
The gap between NRMSEtr and NRMSEcv demonstrates a clearly overfitted model that presents a complexity higher than supported by the data available. This complexity produces an NRMSEtr value of 0.004 for LWC, near to the nominal precision of the instruments used for measuring the spectra (photometric

accuracy of 0.1% T - VERTEX 70 Spectrometer) and is probably more precise than the capacity to determine the true leaf water content values. Nevertheless, this clearly overfitted model would still be selected when using minimization of the cross-validation error (NRMSEcv) as a tuning method.

#### 4. Discussion

The naïve overfitting index selection as presented has a number of advantages. Firstly, it is based on a single error estimation (i.e., NRMSEg). Secondly, it uses all the available observations to calibrate the model ensuring that no degree of freedom is lost in the tuning process. Thirdly, a comparison between different regression techniques is more reliable as the amount of overfitting can be quantified and controlled. This comparison indicates that the maximum level of complexity supported by a model before overfitting depends greatly on the data structure. Especially the number of observations ( $n$ ) versus number of predictors ( $p$ ), the degree of multicollinearity, and the amount of random noise in the data can increase the risk of overfitting considerably. Fourthly, model complexity is hard to standardize across regression techniques, but now the amount of overfitting can be estimated by the naïve overfitting index in a comparable way.

Traditional tuning based on cross-validation does not indicate whether the level of model complexity is appropriate for the data under consideration. The NOIS method allows more control and understanding about the effects of the model complexity in the trade-off between accuracy and overfitting. Finally, robust cross-validation can be time consuming, requiring intensive computing time for high dimensional data such as hyperspectral measurements. The NOIS method is considerably faster, especially for regression techniques that require tuning across a large range of parameters. Cross-validation only needs to be performed for the selected level of complexity to assess the final model accuracy.



**Fig. 6.** Error in model prediction (NRMSE) per level of complexity fitted by PLSR using the traditional tuning method (original data). Solid lines represent training models (NRMSEtr) and dotted lines are cross-validation estimates (NRMSEcv).

#### 4.1. Trade-off accuracy and overfitting

Some machine learning algorithms were initially designed for classification, such as the ones based on regression trees (Random Forest and GBM). Such methods normally produce training models with significantly higher accuracy than the validation models. Thus, these techniques will hardly ever present similar accuracy in the training and validation sets for models using continuous response variables, regardless of the tuning method applied. Also, [Dormann et al. \(2013\)](#) concluded that Random Forests consistently overfit, without there being an obvious solution to correct this.

In prediction problems, it is desirable to fit models from a given sample in such a way that the most accurate predictions are produced, also when applied to other samples from the same population ([Burket, 1943](#)). However, building complex models with high dimensional data with techniques that learn from the information in the model residues can reduce the reproducibility of the prediction accuracy considerably for future samples from the same population ([Kuhn and Johnson, 2013](#)). Accuracy metrics used for model selection in non-OLS regression techniques do not take in account the lack of parsimony as common in ordinary least square regressions. The new tuning method overcomes this problem by identifying for each regression technique the maximum model complexity that is supported by the given data structure.

Seeking accurate models by minimizing the prediction error has to be weighed against the risk of overfitting and producing unrealistically small errors. At times, complex models fictitiously perform better than the accuracy of the measuring system used for collecting the set of spectral signals, chemical concentrations or structural components. The random error in measurements or situations when relevant predictors are missing in the model should not be mistaken for lack of fitness (underfitting) and be a reason to increase model complexity.

Predictors derived from hyperspectral data cannot be considered independent because the reflectance is measured by the same instrument, at the same time, and from nearby wavelengths ([Curran, 1989](#)). These characteristics are generally undesirable for modelling, as predictors that are not independent from each other tend to cause serious problems of multicollinearity. However, these characteristics also provide the opportunity to generate artificial spectra using the covariance matrix in such a way that the data structure is replicated, but the result is not correlated with the response variable. So, our proposed method uses these properties of hyperspectral data to present an intuitive tuning process that permits understanding of the trade-off between accuracy and overfitting for the selected model complexity.

#### 4.2. Limitations and precautions

Our proposed method is built on the assumption that the modelling algorithm conducts the same procedure for the original and for the artificially generated predictors. This is not the case for regression techniques that present an internal mechanism of feature selection for explanatory variables. Such techniques (e.g., Lasso and GBM) may actually present different levels of model complexity for the same value of a tuning parameter ([Hastie et al., 2009](#), [James et al., 2013](#); [Kuhn and Johnson, 2013](#)). This can be seen, for example, in the lasso regression of the moisture dataset. This model, when tuned with cross validation retains 341 out of the 2150 predictors available (others have their coefficients shrunk to zero). However, with the NOIS method it retains 571 predictors.

The process to generate artificial predictors may result in a dataset slightly correlated with the response variable when the number of predictors is extremely large and noisy. In this case, when the complexity is constrained to a level that presents no model contribution, the NOIS tuning may select an underfitted

model. This is the case for the LWC dataset (see [Appendix B](#)), and can be seen distinctly in the ridge regression where the model coefficients were shrunken excessively resulting in an error higher than the RMSEy ([Fig. 6](#)). In this case, the remaining correlation from the generated data can overtake the tolerance of 5%, and a higher threshold should be defined to accept more model contribution. A pre-processing filter to smooth the original spectral signal to reduce the noise before generating the artificial predictors could be applied in such cases. As this study aimed to compare the new method for different data structures, no extra pre-processing was applied on the (original) spectra and the tolerance was kept constant, despite the risk of selecting underfitted models for a particular dataset.

## 5. Conclusion

Hyperspectral data provide opportunities to monitor biological processes and structure in a natural environment over wider temporal and spatial scales. However, as demonstrated in this study, empirical models using high dimensional hyperspectral data as predictors are very likely to cause model overfitting. The traditional tuning methods fail to precisely determine the maximum level of complexity that is warranted by the used data. These methods are also unable to estimate the amount of overfitting expected given a selected model complexity. The NOIS method presented here, overcomes these problems by quantifying the relative amount of overfitting and by selecting an optimal model complexity supported by the data. The new tuning method consistently selects a less complex model and is thus less susceptible to overfitting, while the model performance is similar to the ones selected by the traditional tuning method. The NOIS method increases the chances of fitting more generalizable models from hyperspectral data, avoiding models that perform accurately only on the data that they were trained with.

## Acknowledgments

The current research was supported by CNPq (the Brazilian National Council for Scientific and Technological Development). The authors are thankful for the data kindly shared by World Agroforestry Centre (ICRAF) at ISRIC – World Soil Information, and also by Corjan Nolet and Peter Roosjen at 4TU.ResearchData, and Roshanak Darvishzadeh and Maria Fernanda Buitrago from the ITC faculty, University of Twente.

## Appendix A. Datasets

### A.1. LAI and GER3700 canopy spectra

Description extracted from [Darvishzadeh et al. \(2008\)](#).

#### A.1.1. Leaf area index – LAI

Non-destructive measurements of leaf area index were taken using a Plant Canopy Analyzer (LAI-2000), an instrument produced by LICOR Inc. (Lincoln, NE USA). The measurements were taken under clear sky conditions, with a low solar elevation, and without direct sunlight reaching the sensor. Five below-canopy samples and a reference above-canopy radiation were collected to represent the, average, LAI.

#### A.1.2. Canopy spectra

The canopy spectra measurements were captured in the field from June 15 to July 15 in 2005 by the spectroradiometer GER3700 (Geophysical and Environmental Research Corporation, Buffalo, New York). The wavelength range was between 350 nm

and 2500 nm with a spectral resolution of 3–16 nm. Measurements were collected on clear sunny days between 11:30 and 14:00 to reduce atmospheric perturbations and BRDF effects. The sensor captured a base area about 45 cm in diameter. Up to 15 measurements per plot (1 m × 1 m) were recorded, changing the position slightly to represent the plot area. The average of the measures was used in order to reduce noise.

## A.2. LCC and Hymap image

Description extracted from [Darvishzadeh et al. \(2011\)](#).

### A.2.1. Leaf chlorophyll content – LCC

Leaf chlorophyll content (LCC) was measured in the field by the instrument SPAD-502 Leaf Chlorophyll Meter (Minolta, Inc.). SPAD values are unitless measurements based on the transmittance in red (650 nm) and NIR (920 nm) wavelength regions. Many studies have demonstrated that these values are highly correlated with the leaf chlorophyll concentrations derived from chemical processes. A total of 30 leaves of main, dominant species were measured and averaged to represent the LCC in each plot.

### A.2.2. Hymap image

Hyperspectral airborne HyMap sensor data were acquired over the study area on 4 July 2005. The sensor contained 126 spectral channels in a wavelength range of 436–2485 nm with a spectral resolution of between 13 nm and 17 nm and a spatial resolution of 4 m.

## A.3. LWC and FTIR spectrometer

Description extracted from [Buitrago et al. \(2016\)](#).

### A.3.1. Leaf water content (LWC)

LWC was destructively measured at each stage of the experiment using leaves from the same cohort as the marked leaves, which were used for the spectral measurements. The relative gravimetric LWC was calculated using the equation:  $LWC = 100 * (Ww - Wd) / Ww$ , where  $Ww$  is the weight of the fresh leaf, and  $Wd$  is the weight of the dried leaf. Leaves were dried in an oven at 65 °C. Cuticle thickness was measured from a thin transverse section of the marked leaves, using a Leitz Wetzlar microscope, with an amplification of 250×. This trait was measured at least 3 times in each leaf and the measurements averaged and expressed in  $\mu\text{m}$ .

### A.3.2. Leaf spectral

All plants were measured with a Bruker Vertex 70 FTIR spectrometer, adapted with an external integrating sphere. An Infragold plate with known spectral emissivity was used to calibrate each measurement. Spectra were measured in the range 4000–600  $\text{cm}^{-1}$  (2.5–16.7  $\mu\text{m}$ ) with a resolution of 4  $\text{cm}^{-1}$ . Per leaf eight samples, with 520 scans per sample, were taken. These measurements were averaged and the results were calculated per leaf. Five leaves per plant were measured in the same way for a total of 75 leaves per treatment at every stage of the experiment.

## A.4. Moisture and laboratory spectroscopy

Description extracted from [Nolet and Roosjen \(2014\)](#). Data are publicly accessible at doi: 10.4121/uuid:866135c2-2be3-4b74-8f9c-922505285a7b.

### A.4.1. Moisture

A representative sample of beach sand was collected from the ‘Sand Motor’ (GPS location: 52.0520N 4.1840E). Before the experiment, the sample was coarsely sieved (2 mm) to remove shells and constituents other than sand. The sand, composed of quartz with

some feldspar, had a dry bulk density of 1.655  $\text{g cm}^{-3}$ . For each experiment, a sub-sample of the collected beach sand was placed in a matte black petridish (5 cm radius, 1.5 cm height), filling it up to the rim, and oven dried for 24 h at 105 °C. The sample was, after measuring its initial weight, slowly saturated with distilled water. The water was allowed to distribute itself uniformly throughout the sample and excess free water was drained from the surface. The sample was placed on a data-logging weighing scale with milligram precision.

### A.4.2. Laboratory spectroscopy

A laboratory spectroscopy experiment was conducted twice to observe spectral reflectance in the optical domain (350–2500 nm) under different moisture conditions. The spectral reflectance was measured at 1 nm intervals using an ASD Fieldspec Pro spectrometer (Analytical Spectra Devices, Boulder, CO). A 40,640 cm white Spectralon panel (LabSphere, Inc., North Sutton, NH) was used to calibrate the spectrometer. The spectrometer was fitted with a 10 FOV foreoptic which was directed at nadir at 40 cm distance from the sample. As an artificial light source, a 900 W Quartz Tungsten Halogen (QTH) lamp was placed 70 cm from the sample at a 300° zenith angle. The spectrometer was programmed to take a measurement every 5 min. Each time the weight of the sample was also measured and stored.

## A.5. Soil OrgC and VNIR spectral library

Data from: [World Agroforestry Centre \(ICRAF\) and ISRIC – World Soil Information \(2010\)](#). ICRAF-ISRIC Soil VNIR Spectral Library. Nairobi, Kenya: World Agroforestry Centre (ICRAF). Available at <http://africasoils.net/>.

This spectral library consists of visible near infrared spectra of 785 soil profiles (4437 samples) selected from the Soil Information System (ISIS) of the International Soil Reference and Information Centre (ISRIC). The samples consist of all physically archived samples at ISRIC for which soil attribute data was available in 2004.

### A.5.1. OrgC

Soil samples were air-dried, clods crushed and the resulting sample material sieved through a 2 mm sieve prior to further analysis. Organic carbon content was determined using the Walkley-3.

Black procedure. This involves a wet combustion of the organic matter with a mixture of potassium dichromate and sulfuric acid at about 125 °C. Soil property attributes were provided by ISRIC and had been analysed according to the ISRIC ‘Procedures for soil analysis’ ([Van Reeuwijk, 2002](#)).

### A.5.2. VNIR spectral

Soil diffuse reflectance spectra were recorded for each library sample using a FieldSpec FR spectroradiometer (Analytical Spectral Devices, Boulder, CO) at wavelengths from 0.35 to 2.5  $\mu\text{m}$  with a spectral sampling interval of 1 nm. Samples were illuminated from below using a high-intensity source probe. About 20 g of air-dried soil, ground to pass through a 2-mm sieve, was placed into 7.4 cm diameter Duran glass Petri dishes to give a sample height of about 1 cm. To sample within-dish variation, reflectance spectra were recorded at two positions, successively rotating the sample dish through 90° between readings and an average of 25 spectra was recorded at each position to minimize instrument noise. Before reading each sample 10 white reference spectra were recorded using calibrated spectralon (Labsphere, Sutton, NH, USA) placed in a glass petri dish. Reflectance readings for each wavelength band were expressed relative to the average of the white reference readings. The 1 nm interval spectra were resampled by selecting every tenth-nanometer value from 0.35 to 2.5  $\mu\text{m}$  to give a total of 216 data points for each spectrum.



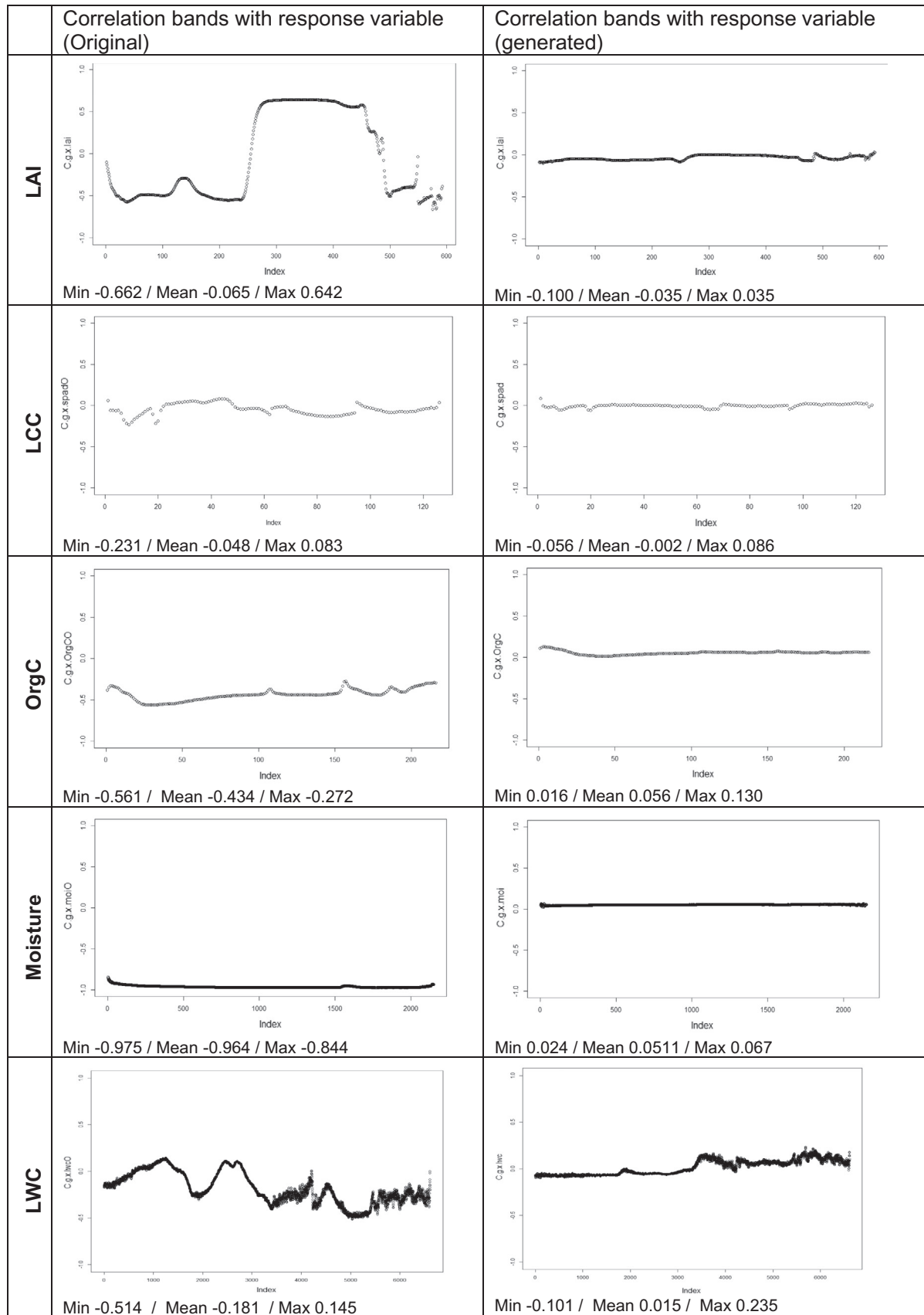


Fig. B.1. Original and generated data for all the used datasets. The average, maximum and minimum correlation between the wavelengths with the response variable.

Regression technique	Tuning parameter	OrgC		LCC		LAI		Moisture		LWC	
		NOIS	CV	NOIS	CV	NOIS	CV	NOIS	CV	NOIS	CV
PLSR	ncomp	10	21	5	5	6	19	4	17	3	20
SVMR	cost	50	100	0.3	0.1	0.1	1	0.05	0.5	0.0003	5
	epsilon	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Rforest	nodesize	120	3	60	3	120	3	90	3	60	3
	mtry	50	50	5	5	10	10	10	10	10	10
	ntree	200	200	200	200	200	200	200	200	200	200
Lasso	fraction	0.0005	0.01	0.001	0.005	0.003	0.05	0.0003	0.03	0.0005	0.5
Ridge	lambda	0.003	0.0001	0.2	0.05	0.5	0.003	0.2	0.001	2.3	0.01
GBM	n.trees	7000	200000	5000	50000	3000	150000	5000	200000	3000	200000
	i.depth	2	2	2	2	2	2	2	2	2	2
	shrinkage	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	n.minnode	10	10	10	10	10	10	10	10	10	10
Nnet	m.interaction	3000	30000	300	700	130	5000	3000	100000	200	20000
	hidden units	4	4	4	4	6	6	3	3	4	4
	learn rate	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.005	0.005	0.0001	0.0001

Fig. C.1. Tuning parameters selected by the NOIS method and the traditional cross-validation per database and regression technique.

## Appendix B. Generated predictors

See Fig. B.1.

## Appendix C. Tuning parameters

See Fig. C.1.

## References

- Abdel-Rahman, E.M., Ahmed, F.B., Ismail, R., 2013. Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. *Int. J. Rem. Sens.* 34 (2), 712–728.
- Acevedo, M.F.B., Groen, T.A., Hecker, C.A., Skidmore, A.K., 2017. Identifying leaf traits that signal stress in TIR spectra. *ISPRS J. Photogramm. Rem. Sens.* 125, 132–145.
- Bioucas-Dias, J.M., Nascimento, J.M., 2008. Hyperspectral subspace identification. *IEEE Trans. Geosci. Rem. Sens.* 46 (8), 2435–2445.
- Bruce, L.M., Koger, C.H., Li, J., 2002. Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction. *IEEE Trans. Geosci. Rem. Sens.* 40 (10), 2331–2338.
- Buitrago, M.F., Groen, T.A., Hecker, C.A., Skidmore, A.K., 2016. Changes in thermal infrared spectra of plants caused by temperature and water stress. *ISPRS J. Photogramm. Rem. Sens.* 111, 22–31.
- Burket, G.R., 1943. *A Study of Reduced Rank Models for Multiple Prediction*. Washington University of Seattle.
- Carvalho, S., Maciel, M., Schlerf, M., Moghaddam, F.E., Mulder, P.P., Skidmore, A.K., van der Putten, W.H., 2013. Changes in plant defense chemistry (pyrrolizidine alkaloids) revealed through high-resolution spectroscopy. *ISPRS J. Photogramm. Rem. Sens.* 80, 51–60.
- Cho, M.A., Skidmore, A., Corsi, F., Van Wieren, S.E., Sobhan, I., 2007. Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression. *Int. J. Appl. Earth Obs. Geoinf.* 9 (4), 414–424.
- Cook, R.D., Weisberg, S., 2009. *Applied Regression Including Computing and Graphics*, vol. 488. John Wiley & Sons.
- Curran, P.J., 1989. Remote sensing of foliar chemistry. *Rem. Sens. Environ.* 30 (3), 271–278.
- Darvishzadeh, R., Skidmore, A., Schlerf, M., Atzberger, C., Corsi, F., Cho, M., 2008. LAI and chlorophyll estimation for a heterogeneous grassland using hyperspectral measurements. *ISPRS J. Photogramm. Rem. Sens.* 63 (4), 409–426.
- Darvishzadeh, R., Atzberger, C., Skidmore, A., Schlerf, M., 2011. Mapping grassland leaf area index with airborne hyperspectral imagery: a comparison study of statistical approaches and inversion of radiative transfer models. *ISPRS J. Photogramm. Rem. Sens.* 66 (6), 894–906.
- Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36 (1), 27–46.
- Esbensen, K.H., Geladi, P., 2010. Principles of proper validation: use and abuse of resampling for validation. *J. Chemom.* 24 (3–4), 168–187.
- Farifteh, J., Van der Meer, F., Atzberger, C., Carranza, E.J.M., 2007. Quantitative analysis of salt-affected soil reflectance spectra: a comparison of two adaptive methods (PLSR and ANN). *Rem. Sens. Environ.* 110 (1), 59–78.
- Fassnacht, F.E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P., Koch, B., 2014. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Rem. Sens. Environ.* 154, 102–114.
- Feilhauer, H., Asner, G.P., Martin, R.E., 2015. Multi-method ensemble selection of spectral bands related to leaf biochemistry. *Rem. Sens. Environ.* 164, 57–65.
- Gelman, A., Hill, J., 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
- Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2014. *Bayesian Data Analysis*, vol. 2. Chapman & Hall/CRC.
- Hansen, P.M., Schjoerring, J.K., 2003. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. *Rem. Sens. Environ.* 86 (4), 542–553.
- Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning*. Springer-Verlag, New York, 763 p.
- Hawkins, D.M., 2004. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* 44 (1), 1–12.
- Huber, S., Kneubühler, M., Psoimas, A., Itten, K., Zimmermann, N.E., 2008. Estimating foliar biochemistry from hyperspectral data in mixed forest canopy. *For. Ecol. Manage.* 256 (3), 491–501.
- James, Gareth, Witten, Daniela, Hastie, Trevor, Tibshirani, Robert, 2013. *An Introduction to Statistical Learning*. Springer, New York.
- Kokaly, R.F., Asner, G.P., Ollinger, S.V., Martin, M.E., Wessman, C.A., 2009. Characterizing canopy biochemistry from imaging spectroscopy and its application to ecosystem studies. *Rem. Sens. Environ.* 113, S78–S91.
- Kooistra, L., Salas, E.A.L., Clevers, J.G.P.W., Wehrens, R., Leuven, R.S.E.W., Nienhuis, P. H., Buydens, L.M.C., 2004. Exploring field vegetation reflectance as an indicator of soil contamination in river floodplains. *Environ. Pollut.* 127 (2), 281–290.
- Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminform.* 6 (1), 1–15.
- Kuhn, M., 2008. Caret package. *J. Stat. Softw.* 28 (5).
- Kuhn, M., Johnson, K., 2013. *Applied Predictive Modeling*. Springer, New York.
- Lee, K.S., Cohen, W.B., Kennedy, R.E., Mausersperger, T.K., Gower, S.T., 2004. Hyperspectral versus multispectral data for estimating leaf area index in four different biomes. *Rem. Sens. Environ.* 91 (3), 508–520.
- Lemon, J., 2006. Plotrix: a package in the red light district of R. *R-News* 6 (4), 8–12.
- Manolakis, D., Marden, D., Shaw, G.A., 2003. Hyperspectral image processing for automatic target detection applications. *Lincoln Lab. J.* 14 (1), 79–116.
- Martin, M.E., Plourde, L.C., Ollinger, S.V., Smith, M.L., McNeil, B.E., 2008. A generalizable method for remote sensing of canopy nitrogen across a wide range of forest ecosystems. *Rem. Sens. Environ.* 112 (9), 3511–3519.
- Meehl, P.E., 1945. A simple algebraic development of Horst's suppressor variables. *Am. J. Psychol.*, 550–554.
- Mirzaie, M., Darvishzadeh, R., Shakiba, A., Matkan, A.A., Atzberger, C., Skidmore, A., 2014. Comparative analysis of different uni- and multi-variate methods for estimation of vegetation water content using hyper-spectral measurements. *Int. J. Appl. Earth Obs. Geoinf.* 26, 1–11.

- Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: a review. *ISPRS J. Photogramm. Rem. Sens.* 66 (3), 247–259.
- Muñoz-Huerta, R.F., Guevara-Gonzalez, R.G., Contreras-Medina, L.M., Torres-Pacheco, I., Prado-Olivarez, J., Ocampo-Velazquez, R.V., 2013. A review of methods for sensing the nitrogen status in plants: advantages, disadvantages and recent advances. *Sensors* 13 (8), 10823–10843.
- Nguyen, H.T., Lee, B.W., 2006. Assessment of rice leaf growth and nitrogen status by hyperspectral canopy reflectance and partial least square regression. *Eur. J. Agron.* 24 (4), 349–356.
- Nolet, C., Poortinga, A., Roosjen, P., Bartholomeus, H., Ruessink, G., 2014. Measuring and modeling the effect of surface moisture on the spectral reflectance of coastal beach sand. *PLoS ONE* 9 (11), e112151.
- Nolet, C., Roosjen, P., 2014. Laboratory Spectroscopy Experiment: Spectral Reflectance of Beach Sand as Function of Surface Moisture Content With Sample of Beach Sand Collected From the Sand Motor, the Netherlands. Wageningen University. <https://doi.org/10.4121/uuid:866135c2-2be3-4b74-8f9c-922505285a7b>. Dataset.
- Plaza, A., Benediktsson, J.A., Boardman, J.W., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., Marconcini, M., 2009. Recent advances in techniques for hyperspectral image processing. *Rem. Sens. Environ.* 113, S110–S122.
- R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL <<https://www.R-project.org/>>.
- Ramoelo, A., Skidmore, A.K., Cho, M.A., Schlerf, M., Mathieu, R., Heitkönig, I.M., 2012. Regional estimation of savanna grass nitrogen using the red-edge band of the spaceborne RapidEye sensor. *Int. J. Appl. Earth Obs. Geoinf.* 19, 151–162.
- Ripley, B.D., 2009. *Stochastic Simulation*, vol. 316. John Wiley & Sons.
- Schlerf, M., Atzberger, C., 2006. Inversion of a forest reflectance model to estimate structural canopy variables from hyperspectral remote sensing data. *Rem. Sens. Environ.* 100 (3), 281–294.
- Schlerf, M., Atzberger, C., Hill, J., Buddenbaum, H., Werner, W., Schüler, G., 2010. Retrieval of chlorophyll and nitrogen in Norway spruce (*Picea abies* L. Karst.) using imaging spectroscopy. *Int. J. Appl. Earth Obs. Geoinf.* 12 (1), 17–26.
- Shepherd, K.D., Palm, C.A., Gachengo, C.N., Vanlauwe, B., 2003. Rapid characterization of organic resource quality for soil and livestock management in tropical agroecosystems using near infrared spectroscopy. *Agron. J.* 95, 1314–1322.
- Skidmore, A.K., Turner, B.J., Brinkhof, W., Knowles, E., 1997. Performance of a neural network: mapping forests using GIS and remotely sensed data. *Photogramm. Eng. Rem. Sens.* 63 (5), 501–514.
- Stroppiana, D., Fava, F., Boschetti, M., Brivio, P.A., 2011. 11 Estimation of Nitrogen Content in Crops and Pastures Using Hyperspectral Vegetation Indices. *Hyperspectral Remote Sensing Of Vegetation* 245.
- Thiemann, S., Kaufmann, H., 2002. Lake water quality monitoring using hyperspectral airborne data—a semiempirical multisensor and multitemporal approach for the Mecklenburg Lake District, Germany. *Rem. Sens. Environ.* 81 (2), 228–237.
- Van Reeuwijk, L.P. (Ed.), 2002. *Procedures for Soil Analysis*. sixth ed. Wageningen, ISRIC. Tech. Pap. 9. Available at: <[http://www.isric.org/Isric/Webdocs/Docs/ISRIC\\_TechPap09\\_2002.pdf](http://www.isric.org/Isric/Webdocs/Docs/ISRIC_TechPap09_2002.pdf)>.
- Venables, W.N., Ripley, B.D., 2002. *Modern Applied Statistics With S*. Springer, New York, ISBN 0-387-95457-0.
- Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J.P., Camps-Valls, G., Moreno, J., 2012. Machine learning regression algorithms for biophysical parameter retrieval: opportunities for Sentinel-2 and-3. *Rem. Sens. Environ.* 118, 127–139.
- Zhao, K., Valle, D., Popescu, S., Zhang, X., Mallick, B., 2013. Hyperspectral remote sensing of plant biochemistry using Bayesian model averaging with variable and band selection. *Rem. Sens. Environ.* 132, 102–119.
- Wilson, A.M., Silander, J.A., Gelfand, A., Glenn, J.H., 2011. Scaling up: linking field data and remote sensing with a hierarchical model. *Int. J. Geograp. Inform. Sci.* 25 (3), 509–521.
- World Agroforestry Centre (ICRAF) and ISRIC – World Soil Information, 2010. ICRAF-ISRIC Soil VNIR Spectral Library. World Agroforestry Centre (ICRAF), Nairobi, Kenya. Available at <<http://africasoils.net/>>.