Hindawi
Journal of Spectroscopy
Volume 2018, Article ID 7960314, 11 pages
https://doi.org/10.1155/2018/7960314

Research Article

A Comparison of Regression Tree Approaches to Modelling the Efficacy of Water Hyacinth Biocontrol Using Multitemporal Spectral Datasets

Na'eem Hoosen Agjee, Onisimo Mutanga, Michael Gebreselasie, and Riyad Ismail

School of Agriculture, Earth and Environmental Sciences, University of KwaZulu-Natal, P/Bag X01 Scottsville, Pietermaritzburg 3209, South Africa

Correspondence should be addressed to Na'eem Hoosen Agjee; agjeen2@gmail.com

Received 25 July 2017; Revised 4 December 2017; Accepted 15 February 2018; Published 14 May 2018

Academic Editor: Pedro D. Vaz

Water hyacinth (*Eichhornia crassipes*) is an exotic plant species that is effectively controlled by *Neochetina* spp. weevils. This study is aimed at determining if spectroscopic data may be utilized to predict insect-induced stress on water hyacinth plants. Single target regression trees (STRTs), multitarget regression trees (MTRTs), and random forest multitarget regression trees (RF-MTRTs) have been used to predict feeding scar damage (FSD) and relative leaf chlorophyll content (RLCC) from hyperspectral canopy reflectance data. Results from this study show that the correlation coefficient of STRTs (training accuracy: 76%–97%; validation accuracy: 47%–86%) performs better than MTRTs (training accuracy: 74%–90%; validation accuracy: 45%–77%) for all infestation levels but are difficult to interpret simultaneously. In contrast, MTRTs (size: 23–35 nodes) are much smaller and more interpretable than STRTs (size: 11–47 nodes) because they predict FSD and RLCC simultaneously. Importantly, RF-MTRTs (training accuracy: 95%–98%; validation accuracy: 55%–88%) yield better predictive performance than STRTs and MTRTs for all infestation levels. It is concluded that MTRTs can be utilized for model interpretation as they are more interpretable; however, RF-MTRTs offer an improved predictive performance.

1. Introduction

Water hyacinth (*Eichhornia crassipes*) is an exotic invasive plant species that occurs as mats on the surface of freshwater bodies [1]. Native to Brazil, water hyacinth has spread to most tropical and subtropical countries suitable for their development [2, 3]. Water hyacinth was first introduced into South African waters around the 1900s [2] and is currently classified as a category 1b invader according to South African legislation, requiring compulsory control. The resilience of this highly invasive exotic weed can be attributed to the prevalence of highly eutrophic waters and the absence of natural enemies [3–5]. Water hyacinth plants have been reported to hinder fishing activities, reduce water quality, impede water usage, and obstruct navigation waterways [6–9] thus placing severe strain on South Africa's limited water resources. In response to this growing ecological concern, biocontrol programs have been initiated in an effort to alleviate the ecological impact imposed on freshwater ecosystems.

The release of biocontrol agents is recognized as an effective solution to sustainably control water hyacinth monocultures. *Neochetina eichhorniae* and *Neochetina bruchi* are two biocontrol agents currently being introduced into freshwater ecosystems throughout South Africa. Their utility is warranted as many studies have demonstrated the efficacy of *N. eichhorniae* and *N. bruchi* weevils in reducing weed density, plant vigor, and reproductive potential [10, 11]. Adult weevils achieve this through feeding by forming rectangular scars on the surface of the leaf [12, 13]. The weevils remove extensive proportions of epidermal tissue at the leaf surface as well as feed on the photosynthetic layers below the leaf surface [13, 14]. Continuous damage negatively affects the functioning of the chloroplasts subsequently reducing relative leaf chlorophyll content (RLCC) and the photosynthetic

capacity of the leaves [15]. Consequently, feeding scar damage (FSD) and RLCC can be used as bioindicators of morphological and physiological damage inflicted to water hyacinth plants. Over time, the combined effects of morphological and physiological stress result in increased leaf mortality, a reduction in plant biomass, and possible plant mortality [16–18]. The ability to quantify the damage inflicted by biocontrol, that is, RLCC and FSD, of variable infestation levels is essential to establishing the efficiency of biocontrol agents to characterize the health status of water hyacinth plants.

Currently, reconnaissance surveys are conducted by manually sampling water hyacinth plants periodically to ascertain the severity of the damage inflicted by biocontrol and the health status of water hyacinth plants. Recently, hyperspectral remote sensing technologies have emerged as a powerful tool to synoptically detect, monitor, and predict vegetation stress [19–22]. Laboratory-based spectroscopic studies can contribute towards exploring the operational potential of predicting different severities of biocontrol damage from remotely sensed data [23]. Hyperspectral data is captured at a high spectral resolution (10 nm) warranting the identification of key spectral regions or diagnostic features that form the leaf optical properties which are related to the biochemical and/or biophysical status of the plants [24]. Importantly, identifying spectral regions that represent responses to key physiological processes (chlorophyll content, chlorophyll fluorescence, carbon, and nitrogen) can be used to detect vegetation stress prior to effects being seen visually [25]. Generally, changes of leaf reflectance in the visible region (350–700 nm) and near-infrared region (700–1000 nm) of the electromagnetic spectrum are indication of vegetation stress [26]. The ability to relate key spectral regions or bands with biocontrol damage reference measurements would allow for the development of calibrated models to monitor and possibly predict previsual and visual biocontrol damage. Consequently, it is imperative to investigate state-of-the-art modelling techniques to determine if these techniques can produce high nowcasting and possibly predictive accuracies when dealing with high dimensional datasets.

Over the last decade, a suite of machine learning algorithms (e.g., artificial neural networks, support vector machines, and fuzzy logic) has emerged as an accurate alternative to conventional parametric linear modelling techniques. One such technique is single target regression trees (STRT) conducting binary recursive partitioning producing a set of rules and a regression model to predict a single response variable [27, 28]. Several studies have successfully demonstrated the utility of STRTs as a powerful tool for data prediction [29–31]. This study attempts to predict biocontrol damage on water hyacinth plants which to the author's knowledge has not been explored before. STRTs offer numerous advantages as a potential operational tool for biocontrol damage monitoring and prediction. STRTs are efficient when dealing with high dimensional datasets and produce a descriptive model [32]. Importantly, STRTs do not rely on data distribution assumptions and the algorithm can map nonlinear relationships between features (i.e., bands) and response variables in complex data spaces [32]. However, a limitation of STRTs is that only one response variable can be predicted per training session. In addition, STRTs can lead to the construction of complex trees that do not generalize well from the training data resulting in overfitting. To ascertain and understand the overall status of water hyacinth plants, environmental managers would have to construct STRTs for each response variable and then try to aggregate the output of the models. This process would be time consuming and inefficient to conduct. Alternatively, a more efficient approach would be to construct a model that simultaneously predicts multiple biocontrol parameters (i.e., responses) with one training session.

Multitarget regression trees (MTRTs) predict several numeric response variables simultaneously [29] and offer several advantages over STRTs. For example, MTRTs are smaller in size than STRTs and are faster to train thus making them more efficient to implement [30]. Furthermore, MTRTs explain dependencies between different variables [30] and are more interpretable than several STRTs [33]. Several studies have explored and successfully demonstrated the utility of MTRTs to predict multiple response variables simultaneously [29, 30, 33–35]. However, to the author's knowledge, only Stojanova et al. [30] have used MTRTs and STRTs to predict vegetation height and canopy cover from remotely sensed data. Results showed that the MTRTs performed significantly better than STRTs when predicting canopy cover. This highlights the operational potential of MTRTs to simultaneously predict RLCC and FSD from hyperspectral data. This study attempts to implement MTRTs to not only predict biocontrol damage but also identify the most influential bands which is important to understand the relationship between influential bands and response variables. Although MTRTs construct easily interpretable models with good predictive performance, they are unstable. Small variations in the data might result in a completely different tree being generated. Unstable predictive models can be combined into an ensemble to improve predictive performance. Random forest multitarget regression trees (RF-MTRTs) are an ensemble of predictive models that when combined increases the predictive performance of their base classifiers [30]. For example, Kocev et al. [34] reported an improvement in the predictive performance when implementing RF-MTRTs compared with MTRTs and attributed the improvement to the ensemble method. However, despite the advantage of improving the predictive performance, RF-MTRTs are not interpretable because hundreds of MTRTs are constructed in an ensemble. Consequently depending on the goal of the application, either an interpretable model can be generated or a model that yields a high predictive performance.

In light of the above, this study is aimed at determining if hyperspectral data can be applied to monitor and predict biocontrol measures of variable infection levels to water hyacinth plants. More specifically, the objectives of this study are to (1) compare the interpretability of STRTs and MTRTs to predict FSD and RLCC of variable infection levels on water hyacinth plants and (2) compare the predictive performance of STRTs, MTRTs, and RF-MTRTs to predict FSD and RLCC of variable infection levels on water hyacinth plants.

2. Materials and Methods

2.1. Experimental Procedure. The experimental procedure implemented in this study was similar to that implemented by Agjee et al. [23]. However, in this study, three *Neochetina* spp. infestation levels, that is, low (two adult male weevils per plant), medium (four adult male weevils per plant), and high (six adult male weevils per plant) were considered to model biocontrol measures from plant spectral reflectance [6]. The three infestation levels were then applied for all subsequent analysis.

2.2. Leaf Variables. Leaf variables sampled included FSD and RLCC. Leaf variables were sampled on the two youngest and two oldest unfurled leaves on each plant [13]. FSD is determined by counting the number of weevil feeding scars on the adaxial leaf laminae on each of the leaves. Subsequently, the chlorophyll content of each of the leaves has been measured using a SPAD-502 chlorophyll meter [36]. The SPAD-502 chlorophyll meter has a measurement area of 0.06 cm^2 and utilizes the 650 nm and 940 nm wavelengths to estimate relative chlorophyll content [37, 38]. Three measurements were recorded on each of the leaves by positioning the leaf over the receptor window and closing the measuring head. FSD and RLCC measurements were averaged for each plant.

2.3. Canopy Reflectance Measurements. Canopy reflectance spectra were captured for low, moderate, and high infestation levels in the same manner as that employed by Agjee et al. [23] over five weeks of infestation. However, in this study, reflectance spectra captured for each week were combined for each infestation level.

2.4. Statistical Analysis

2.4.1. Analysis of Variance. A one-way analysis of variance (ANOVA) was used to ascertain whether differences in FSD and RLCC occur between the variable infestation levels. ANOVAs were performed using TANAGRA version 1.4.50 [40].

2.5. Machine Learning for Biocontrol Modelling

2.5.1. Single Target Regression Trees. Individual STRTs were constructed to predict FSD and RLCC from canopy reflectance spectra for each infection level. A STRT is a hierarchical structure that recursively partitions a set of training observations to produce a model that will predict a single response variable from unseen observations [41]. A STRT is comprised of a root node, branches, internal nodes, and leaves [34]. Initially, the algorithm begins at the root node which contains all the training observations. Subsequently, the dataset is recursively partitioned into subsets at each internal node based on the predictor test. The heuristic function used for selecting the predictor test at each internal node is based on the intracluster variation summed over the subsets induced by the predictor test [29]. The intracluster variation is defined by

$$N \cdot \sum_{t=1}^{T} \mathrm{Var}[y_t], \tag{1}$$

where N is the number of examples in the cluster, T is the number of response variables, and Var $[y_t]$ is the variance of response variable y_t in the cluster.

The goal of the heuristic function is to guide the algorithm towards small trees with good predictive performance [29]. The partitioning process is terminated when a stopping criterion is met [29, 34]. In this study, the F-test stop criterion is used where a node will be split only when a statistical F-test indicates a significant reduction of variance inside the subsets. The F-test value is optimized using the following values: 0.001, 0.005, 0.01, 0.05, 0.1, and 0.125. On termination, the prediction value of the response variable is stored in each leaf. The predicted value is calculated as the mean value of the response variable for the observations that are stored in that leaf [30].

In this study, STRTs were pruned using the M5 pruning method [42–44]. The M5 pruning method builds a multivariate linear model for each node using the observations in the node and the predictors tested in the subtree [42–44]. M5 then calculates the mean absolute deviation of the linear model which is then multiplied by a heuristic penalization factor [42–44]. The resulting error estimate is then compared with the error estimate for the subtree, and if the latter is larger, the subtree is pruned [42–44]. STRTs were constructed using the CLUS software [45].

2.5.2. Multitarget Regression Trees. A MTRT was constructed to simultaneously predict FSD and RLCC using canopy reflectance data for each infection level. A MTRT is a hierarchy of clusters that produces a model to simultaneously predict several response variables from unseen observations [29]. Initially, the algorithm begins at the root node which contains a set of training data. Subsequently, the training dataset is recursively partitioned into smaller subsets using a heuristic function that selects a predictor test at each node [29]. Similar to STRTs, the heuristic function used for selecting the predictor test at each internal node is based on the intracluster variation summed over the subsets induced by the predictor test [29]. The variance function is standardized so that the relative contribution of the different targets to the heuristic score is equal [33]. The partitioning process stops when the F-test stop criterion is met [35]. On termination, the response variables (i.e., FSD and RLCC) are calculated for each leaf. The predicted value for each response variable is calculated as the mean value of the response variable for the observations that are stored in the corresponding leaf [30]. MTRTs are pruned using the M5 pruning method [42–44]. In this study, MTRTs were constructed using CLUS software [45].

2.5.3. Random Forest Ensemble of Multitarget Regression Trees. The random forest algorithm constructs an ensemble of individually grown MTRTs with the prediction of response variables (i.e., FSD and RLCC) based on an average prediction of the response variables for all the regression trees in the forest [46–48]. At the outset bootstrap,

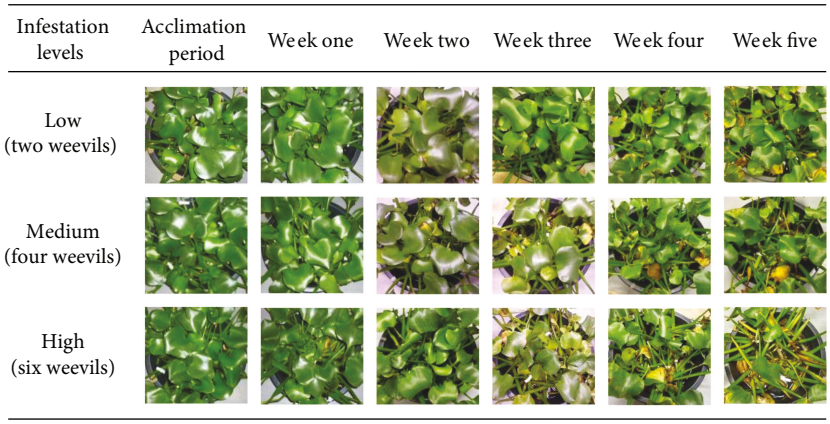

Figure 1: Photographs showing the progression of low, medium, and high weevil damage on water hyacinth plants over five weeks of infestation (adapted from Agjee et al. [23]).

aggregation is employed to create new bootstrap samples [49, 50]. Subsequently, a single MTRT is built for each bootstrap sample and the tree is grown fully without pruning [33]. Since random forest introduces randomness to the regression process, the accuracy of the prediction is improved and the correlation between individual MTRTs is reduced [48, 51]. Random forest introduces randomness through bagging and by choosing a random subset of predictors at each splitting node. The final prediction of each response variable is calculated by averaging the output predictions of the MTRT models in the ensemble [30]. The random forest multitarget analysis was implemented within CLUS software [45]. In this study, the number of trees grown (ntree) per ensemble is 500 trees. The default *mtry* value was used which is given by the function F where $F = \log_2$ (number of predictors + 1).

2.5.4. Evaluating Regression Trees. Model interpretability has been evaluated by determining the size of STRT and MTRT after pruning. The size of the regression trees was calculated as the sum of the nodes (internal nodes and leaves) used to construct the tree [29]. Model size is important to note because the complexer the tree the more bands are used and the more complex the interpretation can be. In addition, each STRT and MTRT model was inspected to determine key spectral regions and identify influential bands used as decision rules to construct the trees.

A 10-fold crossvalidation was performed to validate the regression models constructed. The original dataset was partitioned into ten stratified subsamples, where each subsample was used as a validation dataset while the remaining subsamples were used as training datasets [52]. A regression model was then constructed for the training dataset and the error computed using the test dataset for each fold [52]. The final error is an average of 10 folds to provide a single error estimation.

As recommended by Stojanova et al. [30] and used in other studies [29, 34, 35], the predictive performance of the STRT, MTRT, and RF-MTRT was evaluated by computing the Pearson correlation coefficient and root mean square error (RMSE). The correlation coefficient indicates the direction and strength of a linear relationship between two random variables and has been calculated using

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_{i=1} \sum_{i=1}^{n} Y_{i=1}/n\right)}{\sqrt{\left(\sum_{i=1}^{n} X^2 - \left(\left(\sum_{i=1}^{n} X_i\right)^2/n\right)\right)\left(\sum_{i=1}^{n} Y^2 - \left(\left(\sum_{i=1}^{n} Y_i\right)^2/n\right)\right)}}, \quad (2)$$

where X_i and Y_i are the ith observations of the variables x and y, and n is the total number of pairs of x-y observations.

The RMSE is a measure of the differences between the value predicted by the model and the values actually observed. The RMSE was calculated using formula

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}, \quad (3)$$

where y_i is the observed value and $\hat{y}_i$ is the predicted value for the ith observation.

3. Results

3.1. Biocontrol Damage of Variable Infestation Levels. The extent of biocontrol damage on water hyacinth plants for the three infestation levels over a period of five weeks is shown in Figure 1. It was observed that water hyacinth plants with a low infestation level were healthy after four weeks of infestation. Plants with a medium and high infestation level showed moderate and severe damage after three weeks of infestation. Water hyacinth plants exposed to a high infestation level decreased producing new leaves and decreased in plant size. The base of the petioles were severely eaten with leaves showing signs of desiccation and eventually falling of the petiole.

3.2. Reflectance Spectra of Variable Infestation Levels. Figure 2 shows the spectral reflectance for the three infestation levels over five weeks. It is clear that the low infestation level exhibited higher reflectance spectra than the medium and high infestation levels. In addition, the spectral

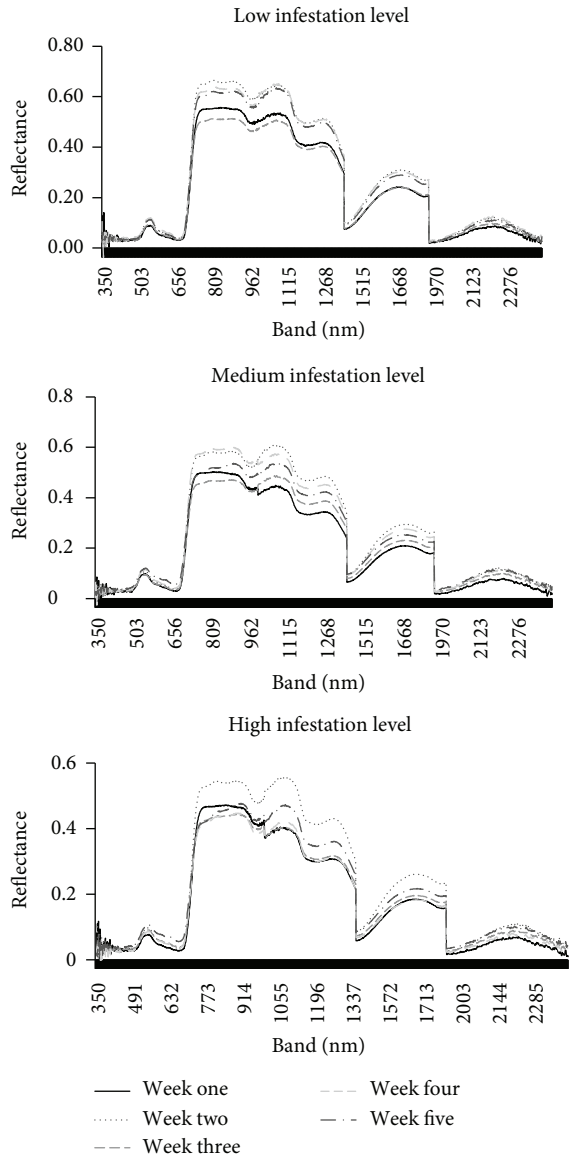

Figure 2: Spectral reflectance for the variable infestation levels (low, medium, and high) under laboratory conditions. Noisy spectral regions were removed (adapted from Agjee et al. [23]).

reflectance for the infestation levels was variable over the duration of the stressor.

3.3. Descriptive Statistics. The descriptive statistics for FSD and RLCC measurements is presented in Table 1. The highest mean RLCC was achieved for the low infestation level (42.95 spad units). The mean RLCC for the high infestation level (41.00 spad units) was lower than the medium infestation level (41.91 spad units). The mean number of feeding scars progressively increased from the low infestation level ($n = 20.47$) to the high infestation level ($n = 69.39$).

The results of the one-way ANOVA indicated that there was a significant difference ($p < 0.01$) in the number of feeding scars between the variable infestation levels. In addition, the results of the one-way ANOVA showed that there was a significant difference ($p < 0.01$) in the RLCC between the infestation levels.

Table 1: Descriptive statistics of the measured RLCC and FSD over five weeks.

Infestation level	Parameter	Mean	Minimum	Maximum	Standard deviation
Low	RLCC	42.95	21.45	58.03	7.69
	FSD	20.47	0.00	58.75	14.13
Medium	RLCC	41.91	20.43	57.90	8.14
	FSD	38.14	1.75	110.50	21.69
High	RLCC	41.00	10.87	56.00	10.04
	FSD	69.39	0.00	338	44.90

RLCC: relative leaf chlorophyll content; FSD: feeding scar damage.

3.4. Single Target Regression Trees. The size of STRT models was used as a measure of interpretability because the deeper the model, the more numerous and more complex the decision rules are, thereby decreasing the interpretability of the model [29]. The interpretability of STRT models used to predict FSD decreased as the size of the models increased and the level of infestation also increased (low = 17 nodes, medium = 56 nodes, and high = 65 nodes). However, the size and interpretability of STRT models used to predict RLCC were variable across the three infestation levels (low = 71 nodes, medium = 41 nodes, and high = 53 nodes). The deepest and least interpretable model was the STRT model which was used to predict RLCC for the low infestation level (size = 71 nodes). In contrast, the most interpretable model was the STRT model used to predict FSD model for the low infestation level (size = 17 nodes). Overall, the results indicate that STRT models constructed to predict FSD and RLCC were large and deep, exhibiting numerous complex decision rules that are difficult to interpret.

Even though STRT models were difficult to interpret, each STRT model was inspected to determine key spectral regions and identify influential bands that were used as decision rules to construct the models. Generally, bands from all three spectral regions (visible, near-infrared, and shortwave infrared region) were used as decision rules to construct each FSD and RLCC model (Figure 3). However, more bands from the visible region (350–700 nm) were used as decision rules than bands from the near-infrared (700–1000 nm) and shortwave infrared region (1000–2500 nm), highlighting the importance of the visible region in predicting FSD and RLCC. A closer inspection of STRT models used to predict FSD revealed that the most influential bands at the root node were located at 696 nm, 409 nm, and 384 nm for the low, medium, and high infestation levels, respectively. The most important bands at the root node of STRT models used to predict RLCC were located at 695 nm, 629 nm, and 746 nm for the low, medium, and high infestation levels, respectively.

The STRT models used to predict FSD and RLCC for the high infestation level are illustrated in Figure 4. The models describe the important bands that influence FSD and RLCC prediction for the high infestation level. The

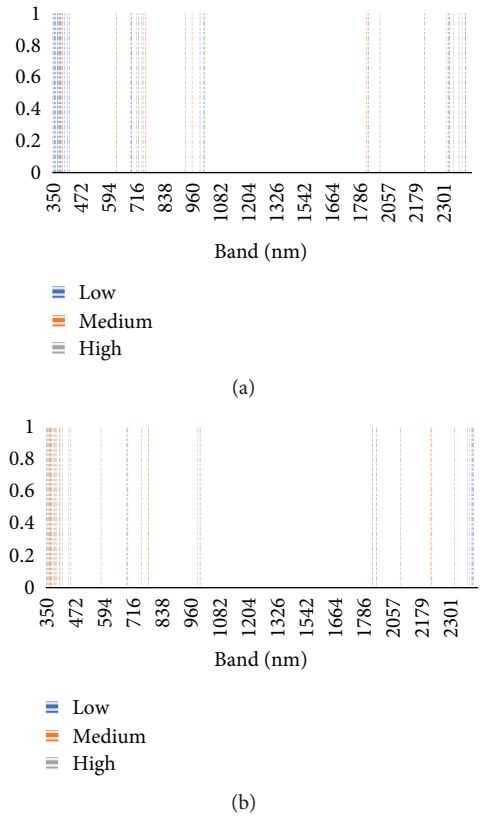

Figure 3: The most influential bands used to construct (a) STRTs to predict RLCC and (b) STRTs to predict FSD for the low, medium, and high infestation levels.

most influential bands defining the FSD model are located at 384 nm, 355 nm, 689 nm, 406 nm, 990 nm, 583 nm, 362 nm, 363 nm, and 2383 nm whereas the most influential bands defining the RLCC model are located at 746 nm, 689 nm, 351 nm, 740 nm, 929 nm, and 350 nm. After inspecting both models, it is evident that both models are deep models exhibiting many different complex decision rules. The complexity of the FSD and RLCC models hinders the simultaneous interpretation of both models.

The correlation coefficient and RMSE were used to evaluate the predictive performance of STRT models used to predict FSD and RLCC for the three infestation levels (Table 2). The results obtained in this study show that the correlation coefficient determined on the validation set increased as the level of infestation increased for both FSD (low: $r = 49\%$; medium: $r = 51\%$; high: $r = 57\%$) and RLCC (low: $r = 64\%$; medium: $r = 67\%$; high: $r = 86\%$). However, the RMSE determined on the training set and validation set for both FSD and RLCC models was variable for the three infestation levels. Overall, it can be observed that the validation set yielded a weaker predictive performance than the training set for both FSD and RLCC models for the three infestation levels (Table 2). Despite the reduction, the FSD and RLCC models constructed for the low infestation level still performed relatively well. Comparatively, STRT models used to predict RLCC yielded a better predictive performance than models predicting FSD for the three infestation levels (Table 2). The best predictive performance for FSD as indicated by the validation set was achieved for the medium infestation level ($r = 51\%$, RMSE = 0.33) while the best predictive performance for RLCC was achieved for the high infestation level ($r = 86\%$, RMSE = 7.21).

3.5. Multitarget Regression Trees and Random Forest Ensemble of Multitarget Regression Trees. MTRT models used to simultaneously predict FSD and RLCC varied in size and interpretability for the three infestation levels (low = 23 nodes, medium = 35 nodes, and high = 29 nodes). The smallest most interpretable MTRT model was constructed for the low infestation level (size = 23 nodes) while the largest least interpretable model was constructed for the medium infestation level (size = 35 nodes). Overall, the MTRT models constructed were small in size with few decision rules resulting in models that are less complex and more interpretable.

Each MTRT model was inspected to determine key spectral regions and identify influential bands that were used as decision rules to construct the models. The most influential bands used to split the observations at the root nodes were located in the visible region (350–700 nm) more specifically at 694 nm, 661 nm, and 689 nm for the low, medium, and high infestation levels, respectively (Figure 5). Generally, bands from the visible region (350–700 nm) were used as decision rules to construct each model for the three infestation levels.

Figure 6 illustrates the MTRT used to simultaneously predict FSD and RLCC for the high infestation level. The most important band at the root node is band 698 nm located in the red edge region. However, the most influential bands for biocontrol damage prediction are bands from both the visible region (350 nm, 363 nm, and 689 nm) and the near-infrared region (743 nm, 757 nm, and 1001 nm).

The predictive performance (correlation coefficient and RMSE) of the MTRTs and RF-MTRTs for the three infestation levels is presented in Table 3. The MTRT correlation coefficient calculated on the validation set increased as the infestation level increased for both FSD (low: $r = 45\%$; medium: $r = 46\%$; high: $r = 50\%$) and RLCC (low: $r = 57\%$; medium: $r = 59\%$; high: $r = 77\%$). However, the RMSE determined on the training set and validation set was variable for both FSD and RLCC across the three infestation levels. In particular, the correlation coefficient calculated on the validation set was 45% when predicting FSD and 57% when predicting RLCC for the low infestation level (Table 3). Despite the reduction in predictive performance from the training set, the test model still generalizes relatively well to unseen observations. However, the best predictive performance for FSD and RLCC as indicated by the validation set was achieved for the high infestation level (Table 3). Overall, MTRTs predicted RLCC better than FSD for the three infestation levels as indicated by the training and validation accuracies (Table 3).

RF-MTRTs consistently achieved a higher predictive performance than single MTRTs when predicting FSD and RLCC for the three infestation levels (Table 3). The ensemble of MTRTs increased the FSD correlation coefficient between 10% and 20% as compared to single MTRTs for the three infestation levels. Similarly, RF-MTRTs predicting RLCC

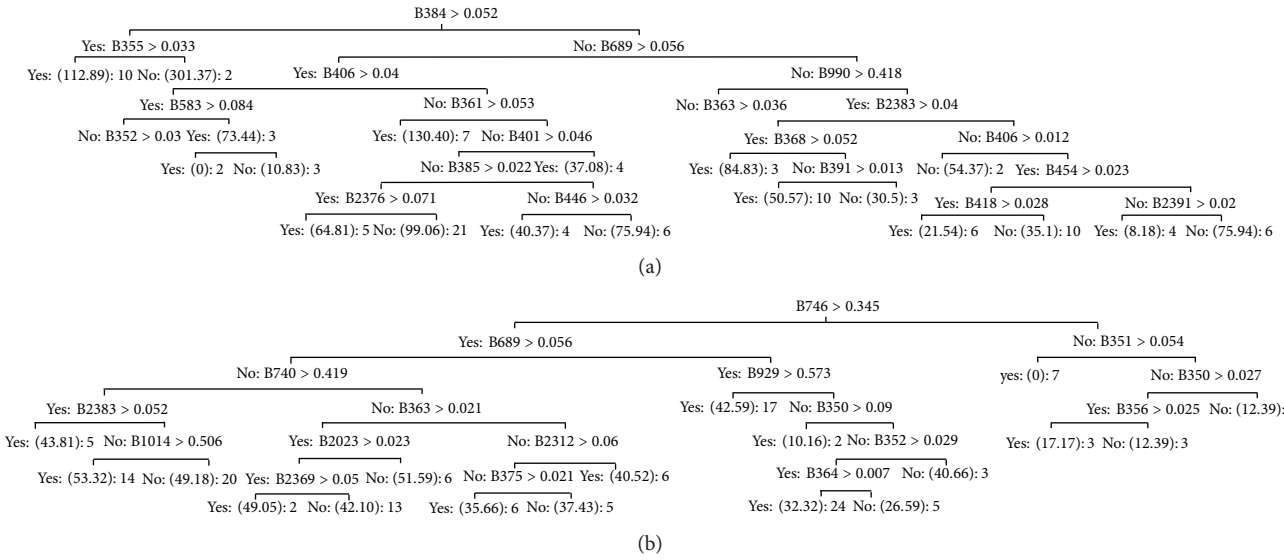

Figure 4: Individual STRTs used to predict (a) FSD and (b) RLCC from canopy spectral reflectance for the high infestation level.

Table 2: Predictive performance of the STRTs for the low, medium, and high infestation levels.

		Correlation coefficient (%)		Root mean square error (FSD—number of feeding scars; RLCC-SPAD units)	
Infestation level	Target	Train	Validation	Train	Validation
Low	FSD	76	49	0.28	0.40
	RLCC	96	64	1.98	6.14
Medium	FSD	93	51	0.13	0.33
	RLCC	92	67	3.14	6.15
High	FSD	90	57	20.63	48.31
	RLCC	97	86	2.70	7.21

RLCC: relative leaf chlorophyll content; FSD: Feeding scar damage.

yielded an improved correlation coefficient of between 6% and 19% as compared to single MTRTs. These results are encouraging highlighting the implementation and utility of an ensemble regression approach to improve the predictive performance of MTRTs. Similar to the results achieved by the MTRTs, the correlation coefficient calculated on the validation set increased as the infestation level increased for both FSD (low: $r = 55\%$; medium: $r = 63\%$; high: $r = 70\%$) and RLCC (low: $r = 63\%$; medium: $r = 78\%$; high: $r = 88\%$). In addition, the RMSE was variable across the three infestation levels (Table 3). The validation set yielded weaker predictive performance that the training set for both FSD and RLCC for the three infestation levels (Table 3). RF-MTRT predicted FSD and RLCC relatively well for the low infestation level (Table 3). However, the best predictive performance for FSD and RLCC as indicated by the validation set was achieved for the high infestation level (Table 3).

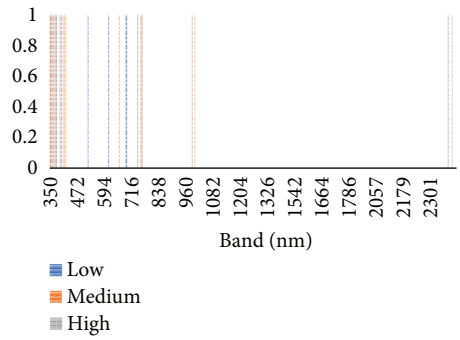

Figure 5: The most influential bands used to construct MTRTs to predict RLCC and FSD for the low, medium, and high infestation levels.

4. Discussion

4.1. Model Size and Interpretability. MTRTs are the best regression tree models to interpret and understand the relationship between reflectance spectra and biocontrol measures. The size of a single MTRT used to simultaneously predict FSD and RLCC is smaller than the sum of the STRTs constructed for each response variable for the three infection levels. MTRTs are less complex in nature because they capture general information about the response variables while considering interactions between response variables [33]. It is more beneficial to interpret a single MTRT that describes all the response variables as compared to interpreting each STRT separately and reconciling the decision rules between trees [34]. Even though the studies did not predict biocontrol measures from hyperspectral data, the results achieved in this study are consistent with the results obtained by Stojanova et al. [30], Kocev et al. [29], and Kocev and Džeroski [33] demonstrating that MTRTs are generally smaller in size and more interpretable than STRTs. Overall, MTRTs are small

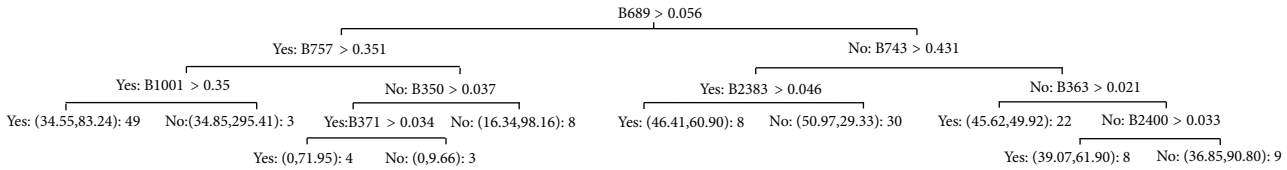

Figure 6: An example of a MTRT to simultaneously predict FSD and RLCC from canopy spectral reflectance for the high infestation level.

Table 3: Predictive performance of the MTRTs and RF-MTRTs for low, medium, and high infestation levels.

		Multitarget regression trees		RF multitarget regression trees	
Infestation level	Target	Correlation coefficient (%) Train/validation	Root mean square error (FSD—number of feeding scars; RLCC-SPAD units) Train/validation	Correlation coefficient (%) Train/validation	Root mean square error (FSD—number of feeding scars; RLCC-SPAD units) Train/validation
Low	FSD	74/45	0.28/0.41	95/55	0.16/0.36
	RLCC	80/57	4.49/6.36	96/63	2.56/5.87
Medium	FSD	87/46	0.17/0.35	96/63	0.12/0.28
	RLCC	88/59	3.63/6.73	97/78	2.19/4.99
High	FSD	85/50	25.34/44.66	98/70	15.23/35.55
	RLCC	90/77	5.81/9.25	97/88	2.83/6.33

RLCC: relative leaf chlorophyll content; FSD: feeding scar damage.

interpretable models that can be implemented efficiently to predict biocontrol measures from hyperspectral canopy reflectance.

In this study, the most influential bands used to partition the observations at the root nodes of MTRT models are located in the red edge region more specifically at 694 nm, 661 nm, and 689 nm for the low, medium, and high infestation levels, respectively. Red-edge bands are directly correlated with RLCC and are affected by vegetation stress consequently its central role in partitioning the observations is logical [53, 54]. The central role of red-edge bands also suggests that RLCC is the dominant response variable being predicted from spectral reflectance. Generally, bands from the visible region (350–700 nm) are used as decision rules to construct each model for the three infection levels. This is rational because weevil feeding stress negatively affects the plant's ability to perform photosynthesis, causing chlorophyll to deteriorate and absorb less efficiently thus influencing visible reflectance [15]. However, weevil feeding stress also affects the morphology of the leaf, in particular destroying the lamina, epidermis, and mesophyll cells thus influencing reflectance in the near-infrared region [55, 56]. Nonetheless, variables with the most explanatory power were primarily selected from the visible region as compared to the near-infrared region highlighting the importance of the visible region to predict biocontrol measures.

Comparatively, STRTs were larger in size, more complex in nature, and less interpretable because they captured more detailed information about the response variables. Results from this study show that the size and complexity of STRTs are influenced by the level of infestation. The size of STRTs predicting RLCC decreased as the infestation level decreased whereas the size of STRTs predicting FSD decreased as the infestation level increased. This was attributed to more tests being required to partition the observations because as the level of infestation increases for FSD and decreases for RLCC, the variance of the response variable decreases. For example, for the high infestation, FSD damage was very extensive and uniform that recorded values were similar and variance lower. Similarly, for the low infestation level, feeding damage was minimal at this stage resulting in infested plants having a similar RLCC to healthy plants. Generally, STRTs constructed for FSD and RLCC used different bands at the root node and a different sequence of bands to model each response variable for each infestation level. This added complexity to the interpretation of STRTs because it was difficult to reconcile the information contained in both trees to determine variables with the most explanatory power. Similar observations were noted by Kocev et al. [34] who compared single target, multitarget, and ensembles to model a compound index of vegetation condition. Overall, STRTs are larger, less interpretable, and less informative than MTRTs to predict biocontrol measures from hyperspectral canopy reflectance.

4.2. Single Target Regression Trees and Multitarget Regression Trees. The predictive performance of both MTRTs and STRTs for all infection levels is relatively strong. A comparison between MTRTs and STRTs reveals that STRTs have a slightly higher predictive performance than MTRTs for all infestation levels (Table 2; Table 3). The lower predictive performance of MTRTs could be attributed to the complex nature of interactions between biocontrol measures. The results obtained in this study were similar to that obtained by Kocev et al. [29] who observed that STRTs performed equally or slightly better than MTRTs. The lower predictive performance of MTRTs should not be a deterrent to its implementation. Many studies still advocate its utility for

multitarget prediction because of its ease of interpretation [29, 34]. However, one of the key limiting factors of both STRTs and MTRTs is that trees constructed on the validation dataset performed much lower than trees constructed on the training dataset for all infection levels. For example, the correlation coefficient of MTRTs constructed on the validation dataset for the medium infestation level was 29% lower than the training dataset when predicting RLCC. This finding was previously observed by Kocev et al. [29] who also noted that the predictive power of STRT and MTRT models on unseen observations was weak. Despite the fact that STRTs and MTRTs overfit the training dataset, both MTRTs and STRTs predicted RLCC better than FSD for all infection levels. In particular, it is assumed that MTRTs and STRTs predict FSD more accurately at the low infection levels because it is the dominant form of damage and a reduction in RLCC is only achieved after significant damage is done [57]. However, result shows that of the two biocontrol measures, physiological damage is predicted better from spectral data and the dominant biocontrol measure at each infestation level. Overall, the findings show that MTRTs perform well and are favored because they are interpretable; however, they are suboptimal when predicting biocontrol measures.

4.3. Multitarget Regression Trees and Random Forest Multitarget Regression Trees. In this study, RF-MTRTs perform better than both STRTs and MTRTs for all infestation levels (Table 2; Table 3). The predictive performance of RF-MTRTs for all infestation levels was strong. When compared to MTRTs, RF-MTRTs improved the predictive performance by between 7% and 21% on the training set and between 6% and 20% on the validation set. This is because RF-MTRTs combine the predictions of numerous base predictive models thereby increasing the overall predictive performance of the model [33]. The results obtained in this study compare favorably with studies conducted by Stojanova et al. [30] and Kocev et al. [58] who demonstrated that RF-MTRTs provide better predictive performance than MTRTs. This clearly demonstrates the utility of RF-MTRTs as a modelling technique as it is able to simultaneously predict multiple response variables while still achieving high predictive performance.

RF-MTRTs predicted both biocontrol measures for low infection levels with a relatively high accuracy and better than MTRTs. The ability to predict biocontrol measures accurately at the initial stages of the infestation is highly beneficial to biocontrol initiatives in determining if weevil populations are alive and establishing within water hyacinth infestations. This will enable water resource managers to establish the efficacy of the weevil and if supplementary weevil releases are required. The highest predictive performance was achieved for the high infestation level using RF-MTRTs. Generally, this was expected because of the extensive damage of both biocontrol measures. The ability to predict high and medium infestation levels is also beneficial as it provides an indication of areas where further interventions are not required. However, as mentioned in the sections above, RF-MTRTs are limited in their interpretability. Kocev et al. [58] suggested implementing both MTRTs and RF-MTRTs to a prediction problem. MTRT offer trees that are more interpretable for knowledge extraction whereas RF-MTRTs offer improved predictive performances. Utilizing both modelling techniques will be highly beneficial when predicting biocontrol measures on water hyacinth plants. Overall, the results from this study demosntrate the excellent performance of RF-MTRTs in predicting biocontrol measures from spectral information.

5. Conclusion

This study demonstrates the benefits of implementing MTRTs and RF-MTRTs to predict biocontrol measures from hyperspectral data on water hyacinth plants. MTRTs that predicted multiple biocontrol measures simultaneously were smaller and more interpretable than STRTs for all infection levels. Variables with the most explanatory power were primarily located in the visible region highlighting the importance of the visible region to predict biocontrol measures. While MTRTs achieved acceptable predictive accuracies, RF-MTRTs employing an ensemble approach were more effective in estimating biocontrol measures achieving a higher predictive performance. However, owing to the size of the forest, interpretation of RF-MTRTs was impractical. Consequently, for operational use, water resource managers should seek to implement MTRTs for information generation but implement RF-MTRTs to yield high predictive accuracies. Future research should investigate the implementation of MTRT models and RF-MTRT models using new generation satellite sensors, for example, WorldView-2 or Sentinel-2A satellite sensors. Furthermore, future studies should explore detecting and modelling other previsual physiological indicators of vegetation stress, such as leaf water content and chlorophyll fluorescence, using spectral information. Overall, this study highlights the opportunities available to environmental managers for quantifying the severity of biocontrol damage and assessing the efficacy of biocontrol agents.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank the National Research Foundation (NRF) for the funding provided to undertake this study.

References

[1] W. T. Penfound and T. T. Earle, "The biology of the water hyacinth," *Ecological Monographs*, vol. 18, no. 4, pp. 447–472, 1948.

[2] L. A. Navarro and G. Phiri, *Water Hyacinth in Africa and the Middle East: A Survey of Problems and Solutions*, International Development Research Centre, Ottawa, ON, Canada, 2000.

[3] R. Verma, S. P. Singh, and K. Ganesha Raj, "Assessment of changes in water-hyacinth coverage of water bodies in northern part of Bangalore city using temporal remote sensing data," *Current Science*, vol. 84, no. 6, pp. 795–804, 2003.

[4] M. C. Law, *Willingness to Pay for the Control of Water Hyacinth in an Urban Environment of South Africa [M.S. thesis]*, Rhodes University, Grahamstown, South Africa, 2007.

[5] C. A. Kull and H. Rangan, "Acacia exchanges: wattles, thorn trees, and the study of plant movements," *Geoforum*, vol. 39, no. 3, pp. 1258–1272, 2008.

[6] R. A. Goyer and J. D. Stark, "The impact of *Neochetina eichhorniae* on water hyacinth in Southern Louisiana," *Journal of Aquatic Plant Management*, vol. 22, pp. 57–61, 1984.

[7] A. M. Villamagna and B. R. Murphy, "Ecological and socio-economic impacts of invasive water hyacinth (*Eichhornia crassipes*): a review," *Freshwater Biology*, vol. 55, no. 2, pp. 282–298, 2010.

[8] D. Jianqing, W. Ren, F. Weidong, and Z. Guoliang, "Water hyacinth in China: its distribution, problems and control status," in *Proceedings of the Second Meeting of the Global Working Group for the Biological and Integrated Control of Water Hyacinth*, Beijing, China, 2001.

[9] H. De Groote, O. Ajuonua, S. Attignona, R. Djessoub, and P. Neuenschwander, "Economic impact of biological control of water hyacinth in southern Benin," *Ecological Economics*, vol. 45, no. 1, pp. 105–117, 2003.

[10] K. L. S. Harley, "The role of biological control in the management of water hyacinth Eichhornia crassipes," *Biocontrol News and Information*, vol. 11, no. 1, pp. 11–22, 1990.

[11] T. D. Center, F. A. Dray Jr, G. P. Jubinsky, and M. J. Grodowitz, "Biological control of water hyacinth under conditions of maintenance management: can herbicides and insects be integrated?," *Environmental Management*, vol. 23, no. 2, pp. 241–256, 1999.

[12] R. Van Driesche, B. Blossey, M. Hoddle, S. Lyon, and R. Reardon, *Biological Control of Invasive Plants in the Eastern United States*, USDA Forest Service, Washington, USA, 2002.

[13] P. J. Moran, "Feeding by waterhyacinth weevils (*Neochetina spp.*) (Coleoptera: Curculionidae) in relation to site, plant biomass, and biochemical factors," *Environmental Entomology*, vol. 33, no. 2, pp. 346–355, 2004.

[14] C. J. Deloach and H. A. Cordo, "Life cycle and biology of *Neochetina bruchi*, a weevil attacking waterhyacinth in Argentina, with notes on *N. eichhorniae*," *Annals of the Entomological Society of America*, vol. 69, no. 4, pp. 643–652, 1976.

[15] R. S. Fletcher, "Applying broadband spectra to assess biological control of saltcedar in west Texas," *Geocarto International*, vol. 29, no. 4, pp. 383–399, 2013.

[16] T. D. Center, "Biological control and its effect on production and survival of waterhyacinth leaves," in *Proceedings of the Fifth International Symposium on Biological Control of Weeds*, Brisbane, Australia, 1980.

[17] I. W. Forno, "Effects of *Neochetina eichhorniae* on the growth of water hyacinth," *Journal of Aquatic Plant Management*, vol. 19, pp. 27–31, 1981.

[18] T. D. Center and T. K. Van, "Alteration of water hyacinth (*Eichhornia crassipes* (Mart.) Solms) leaf dynamics and phytochemistry by insect damage and plant density," *Aquatic Botany*, vol. 35, no. 2, pp. 181–195, 1989.

[19] J. H. Everitt, D. Flores, C. Yang, and M. R. Davis, "Assessing biological control damage of giant salvinia with field reflectance measurements and aerial photography," *Journal of Aquatic Plant Management*, vol. 43, pp. 76–80, 2005.

[20] P. E. Dennison, P. L. Nagler, K. R. Hultine, E. P. Glenn, and J. R. Ehleringer, "Remote monitoring of tamarisk defoliation and evapotranspiration following saltcedar leaf beetle attack," *Remote Sensing of Environment*, vol. 113, no. 7, pp. 1462–1472, 2009.

[21] Z. Oumar and O. Mutanga, "The potential of remote sensing technology for the detection and mapping of *Thaumastocoris peregrinus* in plantation forests," *Southern Forests: A Journal of Forest Science*, vol. 73, no. 1, pp. 23–31, 2011.

[22] Z. Oumar, O. Mutanga, and R. Ismail, "Predicting *Thaumastocoris peregrinus* damage using narrow band normalized indices and hyperspectral indices using field spectra resampled to the Hyperion sensor," *International Journal of Applied Earth Observation and Geoinformation*, vol. 21, pp. 113–121, 2013.

[23] N. H. Agjee, R. Ismail, and O. Mutanga, "Identifying relevant hyperspectral bands using Boruta: a temporal analysis of water hyacinth biocontrol," *Journal of Applied Remote Sensing*, vol. 10, no. 4, article 042002, 2016.

[24] E. M. I. Adam, O. Mutanga, D. Rugege, and R. Ismail, "Field spectrometry of papyrus vegetation (Cyperus papyrus L.) in swamp wetlands of St Lucia, South Africa," in *2009 IEEE International Geoscience and Remote Sensing Symposium*, Cape Town South Africa, 2009.

[25] P. J. Zarco-Tejada and G. Sepulcre-Cantó, "Remote sensing of vegetation biophysical parameters for detecting stress condition and land cover changes," in *Proceedings of VIII Jornadas de Investigación de la Zona no Saturada del Suelo*, Cordoba, Spain, 2007.

[26] N. Coops, M. Stanford, K. Old, M. Dudzinski, D. Culvenor, and C. Stone, "Assessment of Dothistroma needle blight of *Pinus radiata* using airborne hyperspectral imagery," *Phytopathology*, vol. 93, no. 12, pp. 1524–1532, 2003.

[27] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA, USA, 1984.

[28] G. Xian and M. Crane, "Assessments of urban growth in the Tampa Bay watershed using remote sensing data," *Remote Sensing of Environment*, vol. 97, no. 2, pp. 203–215, 2005.

[29] D. Kocev, A. Naumoski, K. Mitreski, S. Krstić, and S. Džeroski, "Learning habitat models for the diatom community in Lake Prespa," *Ecological Modelling*, vol. 221, no. 2, pp. 330–337, 2010.

[30] D. Stojanova, P. Panov, V. Gjorgjioski, A. Kobler, and S. Džeroski, "Estimating vegetation height and canopy cover from remotely sensed data with machine learning," *Ecological Informatics*, vol. 5, no. 4, pp. 256–266, 2010.

[31] L. Dominguez-Granda, K. Lock, and P. L. M. Goethals, "Using multi-target clustering trees as a tool to predict biological water quality indices based on benthic macroinvertebrates and environmental parameters in the Chaguana watershed (Ecuador)," *Ecological Informatics*, vol. 6, no. 5, pp. 303–308, 2011.

[32] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer classification and regression tree techniques: bagging and random forests for ecological prediction," *Ecosystems*, vol. 9, no. 2, pp. 181–199, 2006.

[33] D. Kocev and S. Džeroski, "Habitat modeling with single- and multi-target trees and ensembles," *Ecological Informatics*, vol. 18, pp. 79–92, 2013.

[34] D. Kocev, S. Džeroski, M. D. White, G. R. Newell, and P. Griffioen, "Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition," *Ecological Modelling*, vol. 220, no. 8, pp. 1159–1168, 2009.

[35] G. Everaert, P. Boets, K. Lock, S. Džeroski, and P. L. M. Goethals, "Using classification trees to analyze the impact of exotic species on the ecological assessment of polder lakes in Flanders, Belgium," *Ecological Modelling*, vol. 222, no. 14, pp. 2202–2212, 2011.

[36] Minolta Camera Co. Ltd., *Chlorophyll Meter SPAD-502 Instructional Manual*, Minolta, Osaka, Japan, 1989.

[37] C. Stone, L. Chisholm, and N. Coops, "Spectral reflectance characteristics of eucalypt foliage damaged by insects," *Australian Journal of Botany*, vol. 49, no. 6, pp. 687–698, 2001.

[38] A. D. Richardson, S. P. Duigan, and G. P. Berlyn, "An evaluation of noninvasive methods to estimate foliar chlorophyll content," *New Phytologist*, vol. 153, no. 1, pp. 185–194, 2002.

[39] ASD, *Handheld Spectroradiometer: User's Guide Version 4.05*, Analytical Spectral Devices Incorporated, Boulder, CO, USA, 2005.

[40] R. Rakotomalala, *TANAGRA: A Free Software for Research and Academic Purposes*, Lyon, France, 2005.

[41] L. Yang, G. Xian, J. M. Klaver, and B. Deal, "Urban land-cover change detection through sub-pixel imperviousness mapping using remotely sensed data," *Photogrammetric Engineering & Remote Sensing*, vol. 69, no. 9, pp. 1003–1010, 2003.

[42] J. R. Quinlan, "Learning with continuous classes," in *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, Singapore, 1992.

[43] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Incorporated, San Francisco, CA, USA, 1993.

[44] L. Torgo, *Inductive Learning of Tree-Based Regression Models [Ph.D. thesis]*, University of Porto, Porto, Portugal, 1999.

[45] H. Blockeel and J. Struyf, "Efficient algorithms for decision tree cross-validation," *Journal of Machine Learning Research*, vol. 3, pp. 621–650, 2002.

[46] D. R. Cutler, T. C. Edwards, K. H. Beard et al., "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007.

[47] S. E. Sesnie, B. Finegan, P. E. Gessler, S. Thessler, Z. R. Bendana, and A. M. S. Smith, "The multispectral separability of Costa Rican rainforest types with support vector machines and random forest decision trees," *International Journal of Remote Sensing*, vol. 31, no. 11, pp. 2885–2909, 2010.

[48] M. Segal and Y. Xiao, "Multivariate random forests," *WIREs Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 80–87, 2011.

[49] J. C. Chan and D. Paelinckx, "Evaluation of random forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery," *Remote Sensing of Environment*, vol. 112, no. 6, pp. 2999–3011, 2008.

[50] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, pp. 93–104, 2012.

[51] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[52] R. Geneur, J. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[53] D. N. H. Horler, J. Barber, and A. R. Baringer, "Effects of heavy metals on the absorbance and reflectance spectra of plants," *International Journal of Remote Sensing*, vol. 1, no. 2, pp. 121–136, 1980.

[54] D. N. H. Horler, M. Dockray, J. Barber, and A. R. Baringer, "Red edge measurements for remotely sensing plant chlorophyll content," *Advances in Space Research*, vol. 3, no. 2, pp. 273–277, 1983.

[55] M. O. Bashir, Z. E. El Abjar, and N. S. Irving, "Observations on the effect of the weevils *Neochetina eichhorniae* Warner and *Neochetina bruchi* Hustache on the growth of water hyacinth," *Hydrobiologia*, vol. 110, no. 1, pp. 95–98, 1984.

[56] A. A. El-Zoghby, F. S. Ali, M. H. A. Abo Bakr, and M. H. Mahgoub, "Effect of feeding by two *Neochetina* species or infestation with *Tetranychus urticae* Koch on histological structure of water hyacinth leaves," *Egyptian Academic Journal of Biological Science*, vol. 2, no. 1, pp. 55–61, 2009.

[57] M. H. Julien, M. W. Griffiths, and A. D. Wright, "Biological control of water hyacinth: the weevils Neochetina bruchi and N. eichhorniae: biologies, host ranges, and rearing, releasing and monitoring techniques for biological control of Eichhornia crassipes," in *Australian Centre for International Agricultural Research Monograph No. 60*, Canberra, Australia, 1999.

[58] D. Kocev, C. Vens, J. Struyf, and S. Džeroski, "Tree ensembles for predicting structured outputs," *Pattern Recognition*, vol. 46, no. 3, pp. 817–833, 2013.