A Systematic Approach for Variable Selection With Random Forests: Achieving Stable Variable Importance Values

Amir Behnamian, Koreen Millard, Sarah N. Banks, Lori White, Murray Richardson, and Jon Pasher

***Abstract*— Random Forests variable importance measures are often used to rank variables by their relevance to a classification problem and subsequently reduce the number of model inputs in high-dimensional data sets, thus increasing computational efficiency. However, as a result of the way that training data and predictor variables are randomly selected for use in constructing each tree and splitting each node, it is also well known that if too few trees are generated, variable importance rankings tend to differ between model runs. In this letter, we characterize the effect of the number of trees (ntree) and class separability on the stability of variable importance rankings and develop a systematic approach to define the number of model runs and/or trees required to achieve stability in variable importance measures. Results demonstrate that both a large ntree for a single model run, or averaged values across multiple model runs with fewer trees, are sufficient for achieving stable mean importance values. While the latter is far more computationally efficient, both the methods tend to lead to the same ranking of variables. Moreover, the optimal number of model runs differs depending on the separability of classes. Recommendations are made to users regarding how to determine the number of model runs and/or trees that are required to achieve stable variable importance rankings.**

***Index Terms*— Mean decrease in accuracy (MDA), mean decrease in Gini (MDG) index, random forest, variable reduction.**

I. Introduction

RANDOM Forests, based on the ensembles of classification and regression trees, has become a widely used classification approach in various fields, including remote sensing. It is relatively easy to implement in a variety of software packages (e.g., R Statistics and Python) and is also computationally efficient. The latter is especially relevant today, since high-dimensional data sets from different sources are widely available, and are commonly used for image classification. However, in many cases, not all data sets and predictor variables provide relevant information to the classifier, and thus, it is oftentimes desirable to reduce the model data load to the fewest number of inputs with maximal predictive accuracy. This is especially relevant for large data sets (e.g., Landsat imagery for all of Canada, RADARSAT-2 archive data, and with its four day repeat pass cycle, high-frequency temporal data via the RADARSAT Constellation Mission in the near future [1]–[3]) and/or data acquired from multiple sensors. Reducing model data load can reduce processing times and storage requirements, and can also be used to inform long-term analyses, as attention can focus on just the sensors and variables that provide relevant information to a given classification problem. Furthermore, it has also been demonstrated that with very high dimensional data sets, results can be noisier than models where only the most important variables are used [4]. Both the mean decrease in accuracy (MDA) and the mean decrease in Gini (MDG) are commonly used statistical measures of variable importance for determining which predictor variables are best suited to differentiate the classes of interest and for reducing the dimensionality of large data sets [4]–[7]. MDA quantifies variable importance by measuring the change in prediction accuracy when the values of the variable are randomly permuted. MDG is the sum of all decreases in Gini impurity due to a given variable, normalized by the number of trees (ntree) [8], [9]. However, because of the random way in which training data and variables are selected to determine the split at each node in Random Forests, importance rankings differ from one model run to another, especially when if only a small ntree are generated [4], [7], [10]–[12]. As such, users should not rely on rankings derived from a single model run [13]–[15].

Manuscript received June 13, 2017; revised July 27, 2017; accepted August 19, 2017. This work was supported in part by Environment and Climate Change Canada and in part by Defence Research and Development Canada. *(Corresponding author: Amir Behnamian.)*

A. Behnamian, S. N. Banks, L. White, and J. Pasher are with Environment and Climate Change Canada, National Wildlife Research Centre, Ottawa, ON K1S 5B6, Canada (e-mail: amir.behnamian@canada.ca; sarah.banks@canada.ca; lori.white2@canada.ca; jon.pasher@canada.ca).

K. Millard is with Defence Research and Development Canada, Ottawa, ON K1A 0Z4, Canada (e-mail: koreen.millard@drdc-rddc.ca).

M. Richardson is with the Department of Geography and Environmental Studies, Carleton University, Ottawa, ON K1S 5B6, Canada (e-mail: murray.richardson@carleton.ca).

Digital Object Identifier 10.1109/LGRS.2017.2745049

II. Background

A conservative approach to dealing with varying importance values is to average outputs from a sufficiently large number of forests and sufficiently large ntree (e.g., 50 forests with more than 1000 trees), followed by a "forward" or "reverse" stepwise approach to reduce model inputs to only the most important predictor variables, until the minimum out of bag error (OOBE) is achieved [12], [16]. It is notable that an iterative variable importance reduction (i.e., recalculating variable importance) is computationally expensive for big data sets (in this context, and throughout this letter, the computational expense refers specifically to the amount of time required to generate importance values and/or predict the classification) [6].

Several automated methods of variable selection with Random Forests exist (implemented in commercial software such as R). For example, ggRandomForests [17], [18] assumes that the variables that are used in the split closest to the root are the most important. However, it appears to base this ranking on a single forest. varSelRF [10], [16] runs a single model, removes the 20% least important variables, and then recalculates errors iteratively with the new set of variables until an unacceptable level of error is reached. While this process is iterative, the rankings are based on a single model run. VSURF runs Random Forests in a two-step-process [11]. First, it ranks the variables based on the average over 50 runs and removes those ranked below a threshold. Then, it sequentially builds Random Forest models and monitors OOBEs by adding variables; starting from the first most important and excluding those variables that do not improve OOBE (based on the average error over 25 runs).

While it is known that the ntree used in a Random Forest model can impact the stability of variable importance [11], a systematic analysis of the convergence of importance values to a stable mean has not been undertaken in previous studies (in this context, and throughout this letter, a stable mean importance value is one that closely approximates the true mean importance value, which is unknown to the user). Running a model multiple times and subsequently averaging importance values will eventually lead to a stable ranking of important variables. However, the optimum number of model runs using an optimum ntree should be determined in order to maximize computational efficiency. Several attempts have been made to determine an optimum value for the latter parameter [19] but have not addressed a link to the former.

III. Objective

Given the inherent random variation of importance values, we hypothesized that average variable importance values will converge (to its true, but unknown mean) after a certain number of model runs. This may occur across relatively few models, thus unnecessary processing can be avoided. The primary objective of this letter is to develop a systematic approach for determining the number of model runs (i.e., forests) required to achieve a stable mean variable importance value. We also address whether the point of convergence varies as a function of the ntree generated per Random Forest model, as well as the separability of the classes (referring to the physical separation of class values within multivariate feature space). With respect to the latter, we hypothesized that the convergence of variable importance will depend on the separability of the classes, as classification accuracies are higher and more stable in the cases where there is good separability. Thus, in this letter, we have analyzed two data sets: one with "poor" and one with "good" separability.

IV. Method

A. Study Areas

Two study sites are considered in this analysis. The first site (hereafter referred to as Coronation Gulf) encompasses the entirety of Coronation Gulf, Bathurst Inlet, and Dease Strait, Nunavut (Table I), where the focus is to classify shoreline types. This site and data set have been previously presented in [6]. The other study site (hereafter referred to as Alfred Bog) centers on a large peatland complex in South Eastern Ontario, and was previously presented in [4]. For this site, the focus is to discriminate peatland types and to differentiate peatland and nonpeatland classes (Table I). For additional site specific details, as well as information on model training and validation data, readers are referred to [2], [4], and [6].

TABLE I
Land Cover Classes Considered in This Letter

Study Area	Coronation Gulf	Alfred Bog
Class	Water	Fen
	Sand/Mud	Open Bog
	Mixed Sediment	Mixed Sediment
	Pebble/Cobble/Boulder	Treed Bog
	Bedrock	Mixed Forest
	Wetland	Wetland
	Tundra	Agriculture

B. Remote Sensing Data

For both the study areas, a combination of Landsat, RADARSAT-2, and digital elevation model variables were provided as inputs to the model. In total, 49 variables were classified for Coronation Gulf and 50 for Alfred Bog. Banks *et al.* [6] provide all image processing details, which were followed exactly for the Alfred Bog study area, with the exception that: Landsat 8 imagery and Shuttle RADAR Topography Mission data were used in the place of Landsat 5 imagery and Canadian Digital Elevation Data. RADARSAT-2 data for Alfred Bog were also Boxcar filtered instead of Enhanced Lee filtered, and two additional variables (Shannon entropy: phase and intensity) were used for this site (described in [2]), and not for Coronation Gulf. The Julian date of the RADARSAT-2 acquisition was also not included among the set of variables for the Alfred Bog site.

With Users and Producers accuracies for seven land covers $\geq$84%, Coronation Gulf has been selected to reflect the "good separability" case [6]. Alfred Bog represents the "poor separability" case, since Users and Producers accuracies for five classes were much lower ($\geq$63%) [2].

C. Random Forests and Variable Importance

The Random Forests model was implemented using the randomForest [8] package in R. To address the objectives of this letter, four sets of models with a different ntree (50 200 500 and 10 000) were each run for 25 iterations to assess the stability of variable importance rankings, as well as the effect of the ntree on the point at which importance values converge to a stable mean. Each time the model was run using identical training data, and for each run, variable importance was calculated (both MDA and MDG). Mean importance values were then calculated using the following equation:

$$\overline{VI}_p(i) = \left(\sum_{j=1}^{i} VI_p(j) \right) / i \tag{1}$$

where p is the predictor variable of interest listed in Section IV-B, $VI_p(j)$ is the corresponding variable importance value for an individual run j, and $\overline{VI}_p(i)$ is the mean importance variable over i runs.

To further investigate the number of runs required to achieve stable mean importance values, the convergence of the deviation of mean importance values from their true mean at each model run was calculated for all predictor variables using the following equation (assuming that the average of 25 runs provides a good approximation of the true mean importance value):

$$D(i) = \left(\frac{\sum_{p=1}^{P} (\overline{VI}_p(i) - \overline{VI}_p(25))^2}{P} \right)^{0.5} \quad (2)$$

where P is the total number of predictor variables (50 for Coronation Gulf and 49 for Alfred Bog).

Note that the degree of correlation of variables was considered outside the scope of this letter, but it is an important consideration when using Random Forests [4], [6], [11], [20]. For example, Genuer *et al.* [11] showed that the addition of highly correlated replications of a true predictor variable leads to a decrease in the magnitude of the importance of the true variable, and likely results in a decrease in the variability of importance values of the true variables (but not the corresponding correlated ones). Additionally, in this letter, effort was also not made to assess the effect of mtry (i.e., the number of variables tried to determine the optimal split at each node), since the mtry default value (i.e., the square root of the number of predictor variables) has been shown to achieve results that are close to optimal [21]–[24]; increasing this value would greatly decrease the computational efficiency of the algorithm, which is one of the primary benefits associated with Random Forests.

V. Results and Discussion

Fig. 1 shows the plots of mean variable importance rankings for different numbers of trees (50, 200, 500, and 10 000) for both the sites. In the good separability case (Coronation Gulf) with only 50 trees [Fig. 1(A)], the MDA values over the first 25 model runs are highly variable. For instance, the ranking of variable 4 (in bold blue line) has been changed from 4 to 8 after 21 runs. Results are similar in the poor separability case [Alfred Bog; see Fig. 1(B)] though the separation between the top eight variables is not reached even after 21 model runs (e.g., variables represented by blue line and orange line are still crossing). Fig. 1(a) and (b) represents the ranking and the error bars (representing 95% confidence interval) of the mean importance for the first 30 most important variables. For Coronation Gulf, the first 20 variables exhibit a gradual decrease in the value of the mean importance and the root mean square values, but this is only the case for the first eight or nine variables with the Alfred Bog data set. This could be a result of the lesser ranking variables not containing any additional information or that this additional information is not relevant, given the land cover classes of interest.

The variability, and as a result, the stability of the mean importance value for each variable, improves further as the ntree is increased, and this is true for both the poor separability case [see Fig. 1(C) and (D)] and the good separability case [see Fig. 1(E) and (F)]. For example, when the ntree is increased to 200 and 500, the MDA importance ranking of all variables, including the top eight most important variables, stabilizes after fewer runs. This lower variation is also reflected in the magnitude of error bars [see Fig. 1(c)–(f)].

TABLE II
Required Number of Runs for Variable Importance Stability

Site	Separability	ntree	MDA	MDG
Coronation Gulf	good	50	21	21
		200	18	17
		500	10	9
		10000	1	1
Alfred Bog	poor	50	23	23
		200	21	20
		500	16	16
		10000	1	1

These results clearly demonstrate that the convergence of mean importance values to a close approximation of their true mean requires more runs for models built with fewer trees (see Fig. 1). With 10 000 trees, the low variation of mean importance values based on sequential averaging [Fig. 1(g) and (h)], in addition to the fact that there is almost no cross-ranking among the variables [Fig. 1(G) and (H)], indicates values closely approximate the real mean. As such, the maximum deviation of the sequential mean importance values from the mean with 10 000 trees [Fig. 2 (green line)] can be used as a threshold to specify the point of convergence of mean importance values for both Coronation Gulf and Alfred Bog. This threshold is drawn in Fig. 2 using a dashed horizontal line. These values are also listed in Table II, and shown in Fig. 1 (vertical dotted lines). As can be observed, the convergence point is consistently lower for the good separability case (Coronation Gulf) than the poor separability case (Alfred Bog; see Fig. 2). A similar analysis was also performed using MDG (listed in Table II; results not provided in detail here for brevity). Results were similar and indicated that the predicted point of stability is similar to those calculated from the MDA. However, with 200 or more trees, convergence occurred after fewer iterations with MDG, indicating that MDG may be slightly more stable. This observation is consistent with the findings of Liaw and Wiener [7]. We also found larger differences in importance values between the most and least important variables with MDG compared with MDA, meaning that a visual threshold between the important and nonimportant variables was easier to identify.

It is worth noting that the OOBE reached a minimal value with as few as 50 trees for Coronation Gulf and slightly higher for Alfred Bog (OOBE ∼14% and ∼18%, respectively), and remained the same regardless of the ntree that were generated for each model, as shown in Fig. 2(c) (which also suggests the approximate lowest limit of the ntree value

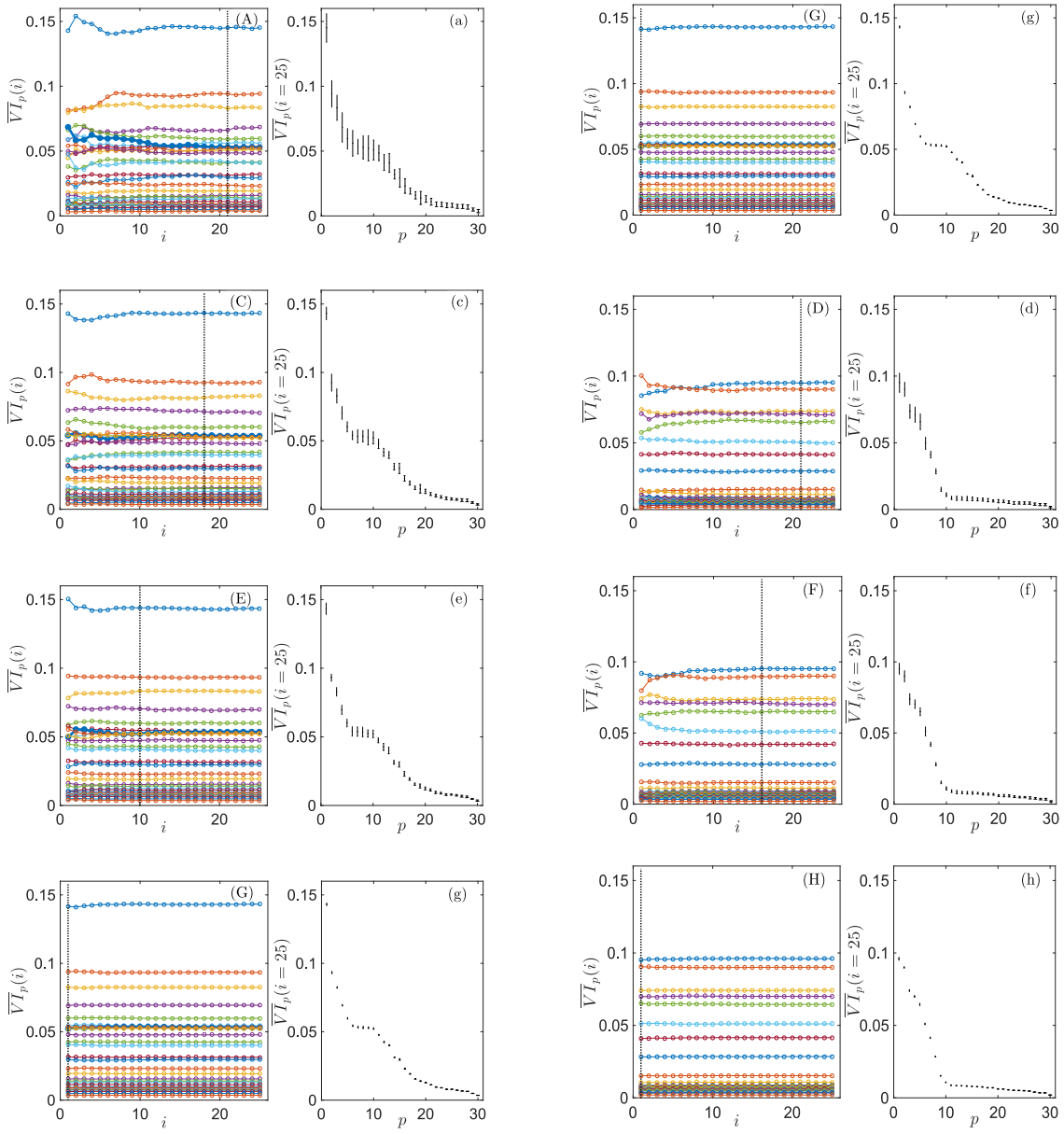

Fig. 1. Sequential averages of the variable importance based on the MDA, plots with capital letters, and the variable importance ranking over 25 runs, plots with small letters. (Left) Coronation data set. (Right) Alfred Bog data set. (A), (a), (B), and (b) 50 trees. (C), (c), (D), and (d) 200 trees. (E), (e), (F), and (f) 500 trees. (G), (g), (H), and (h) 10 000 trees. Dotted lines: number of runs at which the convergence achieved. Only the first 30 important variables are illustrated here.

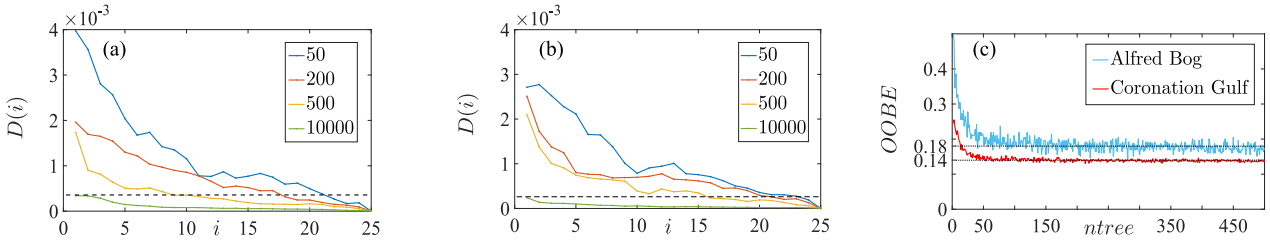

Fig. 2. Deviation of the all predictor variables from their true mean at each model run. (a) Coronation. (b) Alfred Bog. Dashed lines: convergence threshold. (c) OOBE against the ntree for both the Coronation Gulf and Alfred Bog data sets. Dotted lines: minimum value of OOBE.

to users). Furthermore, running one Random Forest model with a large ntree required considerably more time than running multiple models with fewer trees. For example, based on the training data from Alfred Bog (number of training data points $= 500$ and $P = 50$) using a desktop computer (Intel i7 6700HQ at 2.6 GHz and 16 GB of DDR4 RAM

at 2400 MHz), one Random Forest model with 50 trees required 0.114 s and one random forest model with 10 000 trees required 21.54 s (both averaged over 1000 replicates). Thus, the minimum required time to achieve stable importance rankings with the latter is 2.62 s, which is one order of magnitude less than the time required for one Random Forest with 10 000 trees. This difference has important implications for the operational uses of Random Forests with much larger data sets. Specifically, these results show that in obtaining stable mean importance values, it is more computationally efficient to run many iterations of the model with a small ntree than to run a single stable forest of 10 000 trees. Note that, in this case, both the approaches led to approximately the same ranking of variables (i.e., the top ten most important tended to remain constant, while the ranking of lesser important variables varied slightly and the OOBE was not significantly different). The methods used here to determine the optimum number of model runs based on the ntree in each forest can be fully automated by the user. This requires a two-step process, including: 1) defining threshold for the deviation of mean importance values from their true mean by calculating $D(i = 2)$ with a large ntree (for example 10 000), and 2) comparing the convergence plots such as those in Fig. 2(a) or (b) with the calculated threshold value (e.g., for a given ntree determined using [19]).

VI. Conclusion

Importance rankings of MDA and MDG can be variable between runs of Random Forests, even if the same settings are used (e.g., the ntree). Therefore, it is recommended that in order to select variables based on their importance ranking, Random Forests should be run more than once and the variability of values must be assessed. We have demonstrated that variable importance rankings based on the average of sequential models eventually stabilize, but that the minimum number of runs required to achieve stability depends on both the ntree used to build the models and the separability of the classes in the input data. We have demonstrated that convergence to a stable mean can be achieved either by using very large ntree (10 000 or more) or by taking the average variable importance over an optimal number of runs. While both the approaches tend to lead to the same ranking of variables (especially for the top most important), the latter has also been found to be more computationally efficient. A systematic approach to determine the optimum number of runs to achieve a stable mean variable importance has been demonstrated, and recommendations have been made to the user on how to repeat this process.

References

[1] A. Mellor, A. Haywood, C. Stone, and S. Jones, "The performance of random forests in an operational setting for large area sclerophyll forest classification," *Remote Sens.*, vol. 5, no. 6, pp. 2838–2856, 2013.

[2] L. White, K. Millard, S. Banks, M. Richardson, J. Pasher, and J. Duffe, "Moving to the RADARSAT constellation mission: Comparing synthesized compact polarimetry and dual polarimetry data with fully polarimetric RADARSAT-2 data for image classification of peatlands," *Remote Sens.*, vol. 9, no. 6, p. 573, 2017.

[3] A. A. Thompson, "Overview of the RADARSAT constellation mission," *Can. J. Remote Sens.*, vol. 41, no. 5, pp. 401–407, 2015.

[4] K. Millard and M. Richardson, "On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping," *Remote Sens.*, vol. 7, no. 7, pp. 8489–8515, 2015.

[5] J. M. Corcoran, J. F. Knight, and A. L. Gallant, "Influence of multi-source and multi-temporal remotely sensed and ancillary data on the accuracy of random forest classification of wetlands in Northern Minnesota," *Remote Sens.*, vol. 5, no. 7, pp. 3212–3238, 2013.

[6] S. Banks, K. Millard, J. Pasher, M. Richardson, H. Wang, and J. Duffe, "Assessing the potential to operationalize shoreline sensitivity mapping: Classifying multiple Wide Fine Quadrature Polarized RADARSAT-2 and Landsat 5 scenes with a single Random Forest model," *Remote Sens.*, vol. 7, no. 10, pp. 13528–13563, 2015.

[7] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[8] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.

[9] L. Breiman. (2003). *Manual on Setting Up, Using, and Understanding Random Forests V4. 0*. [Online]. Available: http://oz.berkeley.edu/users/breiman.Using_random_forests_v4.0.pdf

[10] R. Diaz-Uriarte and S. A. de Andrés. (2005). "Variable selection from random forests: Application to gene expression data," Spanish Nat. Cancer Center, Tech. Rep. [Online]. Available: https://arxiv.org/abs/q-bio/0503025

[11] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognit. Lett.*, vol. 31, no. 14, pp. 2225–2236, 2010.

[12] M. L. Calle and V. Urrea, "Letter to the editor: Stability of random forest importance measures," *Briefings Bioinform.*, vol. 12, no. 1, pp. 86–89, 2011.

[13] M. Immitzer, C. Atzberger, and T. Koukal, "Tree species classification with random forest using very high spatial resolution 8-band WorldView-2 satellite data," *Remote Sens.*, vol. 4, no. 9, pp. 2661–2693, 2012.

[14] K. Millard and M. Richardson, "Wetland mapping with LiDAR derivatives, SAR polarimetric decompositions, and LiDAR–SAR fusion using a random forest classifier," *Can. J. Remote Sens.*, vol. 39, no. 4, pp. 290–307, 2013.

[15] S. V. Beijma, A. Comber, and A. Lamb, "Random forest classification of salt marsh vegetation habitats using quad-polarimetric airborne SAR, elevation and optical RS data," *Remote Sens. Environ.*, vol. 149, pp. 118–129, Jun. 2014.

[16] R. Díaz-Uriarte and S. A. de Andres, "Gene selection and classification of microarray data using random forest," *BMC Bioinform.*, vol. 7, no. 1, p. 3, 2006.

[17] J. Ehrlinger. (Dec. 2016). "ggRandomForests: Exploring random forest survival." [Online]. Available: https://arxiv.org/abs/1612.08974

[18] H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer, "High-dimensional variable selection for survival data," *J. Amer. Stat. Assoc.*, vol. 105, no. 489, pp. 205–217, 2010.

[19] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?" in *Proc. Int. Workshop Mach. Learn. Data Mining Pattern Recognit. (MLDM)*, Jul. 2012, pp. 154–168.

[20] C. Strobl, A.-L. Bouleseix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinform.*, vol. 9, pp. 307–318, Dec. 2008.

[21] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[22] Ö. Akar and O. Güngör, "Integrating multiple texture methods and NDVI to the Random Forest classification algorithm to detect tea and hazelnut plantation areas in northeast Turkey," *Int. J. Remote Sens.*, vol. 36, no. 2, pp. 442–464, 2015.

[23] R. Sonobe, H. Tani, X. Wang, N. Kobayashi, and H. Shimamura, "Random forest classification of crop type using multi-temporal TerraSAR-X dual-polarimetric data," *Remote Sens. Lett.*, vol. 5, no. 2, pp. 157–164, 2014.

[24] V. Svetnik, A. Liaw, C. Tong, and T. Wang, "Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules," in *Proc. Int. Workshop Multiple Classifier Syst.*, 2004, pp. 334–343.