Optimization of the Random Forest Algorithm for Multispectral Derived Bathymetry

Manessa, M. D. M.,[1*] Setiawan, K. T.,[2] Haidar, M.,[3] Supriatna, S.,[1] Pataropura, Amesanggeng[4] and Supardjo, A. H.[5]
[1]Geography Department, University of Indonesia, Depok, West Java, Indonesia
E-mail: manessa@ui.ac.id
[2]Remote Sensing Application Center, National Institute of Aeronautics and Space (LAPAN), Jakarta, Indonesia
[3]Geospatial Information Agency, West Java, Indonesia
[4]Departement of Software, Buddhi Dharma University, Tangerang, Banten, Indonesia
[5]Departement of Physics, Indonesia University, Depok, West Java, Indonesia
Correspondence Author

Abstract

The random forest (RF) algorithm is among the most commonly applied machine learning algorithms in remote sensing. In this study we tested a new approach to improving the accuracy of RF algorithm when applied to multispectral derived bathymetry by increasing predictor numbers and improving hyperparameter tuning. This approach goes beyond previous work that only applied an auto-tuning hyperparameter and linearized reflectance. We tested our experimental approach on the Gili Islands of Indonesia by comparing the optimized RF to basic RF algorithms used to determine water depth from multispectral imagery. The findings of this study indicate that the optimized RF approach was particularly advantageous in high-dimension data: errors in water depth prediction accuracy improved by 46% after optimization.

1. Introduction

The Random Forest (RF) algorithm is a machine learning approach that belongs to the multiple decision tree learning family (Breiman, 2001). It has been widely used to solve problems related to remote sensing data since it can deal with regression or classification but is sensitive to the identification of critical variables and proper sampling design (Belgiu and Drăgu, 2016). Recent research has explored new modifications and applications of the algorithm for remote sensing, including modifying it to resolve issues with out-of-bag cross-validation (Cánovas-García et al., 2017) and applying it to land cover identification (Smith, 2010), urban tree space (Chen et al., 2017), Antarctic moss health (Turner et al., 2018), biomass (Mutanga et al., 2012) and bathymetry (Manessa et al., 2016).

The feasibility of deriving bathymetric information that estimates from remote sensing imagery was first demonstrated using aerial photographs over clear shallow water (Lyzenga, 1978). The technique has been expanded to include the use of passive optical multi-spectral satellite imagery. The basic concepts of multi spectral derive bathymetry (MDB) method is based on the simple assumption of a linear relation between water depth and surface reflectance (Lyzenga, 1978 and Stumpf et al., 2003).

Compared with conventional survey methods, MDB is a preferred approach due to the efficiency of extracting depth data from shallow-water areas using multispectral imagery (e.g., Landsat series, SPOT-6, WorldView series, or Quickbird). The empirical MDB approach is preferable due to its simplicity, as it uses a simple linear regression (Lyzenga, 1978 and Stumpf et al., 2003), although this is based on several unrealistic assumptions (homogeneous bottom substrate and water quality). For this reason, advanced statistical models such as generalized adaptive models (Kanno et al., 2011), support vector machines (Mohamed et al., 2017 and Mohamed and Nadaoka, 2019) and RF analyses (Manessa et al., 2016 and Mohamed and Nadaoka, 2019) have been used to address the issue and improve depth prediction accuracy.

We first implemented the RF approach for MDB in a previous study with promising results (Manessa et al., 2016). Subsequently, compared RF with other MDB methods, revealing that a semiparametric regression using depth-independent variables and a spatial coordinates algorithm (Kanno et al., 2011) performed much better than RF but had a longer processing time. However, the RF algorithm can be a useful alternative because of its shorter processing time and higher accuracy when compared with other

MDB methods (Lyzenga, 1978, Mishra et al., 2005, Mohamed et al., 2016, Mohamed and Nadaoka, 2019 and Stumpf et al., 2003).

Despite this improved accuracy, such previous work (Manessa et al., 2016, 2018) has ignored two issues: (1) use of auto-tuning hyperparameters (number of trees, mtry, sampling size, and node size) that are not consistently effective and (2) the fact that multispectral imagery has a limited number of bands that can be used as predictors, while the RF algorithm works well with high-dimension data. Moreover, Stumpt et al., (2002) and Kanno et al., (2011) proposed a modified variable based on images' band values (i.e., reflectance ratio, modified linearized reflectance, and bottom invariance index) that showed a positive result for MDB. This modified variable can be used to increase the number of predictors. With this in mind, in this study we tried to optimize the RF algorithm for MDB by manually tuning the hyperparameters and adding several modified variables, using SPOT-7 imagery with 6 m spatial resolution at nadir.

2. Materials and Methods

2.1 Study Area and Data Sets

We conducted our MDB analysis around Indonesia's Gili Islands, located in the central portion of the Java Sea approximately 50 km north of Lombok Island (between 8º 21′ 00″ to 8º 22′ 30″ S and 116º 1′ 00″ to 116º 5′ 30″ W, Figure 1). The islands' coastal morphology consists of cliffs, small bays, and a narrow coastline, while the diverse bottom substrate in this coral reef environment (Case 1 water) varies between hard coral, soft coral, dead coral, dead coral with algae, rubble, sand, and seagrass. This diversity makes the area an ideal representative of shallow-water areas for MDB studies.

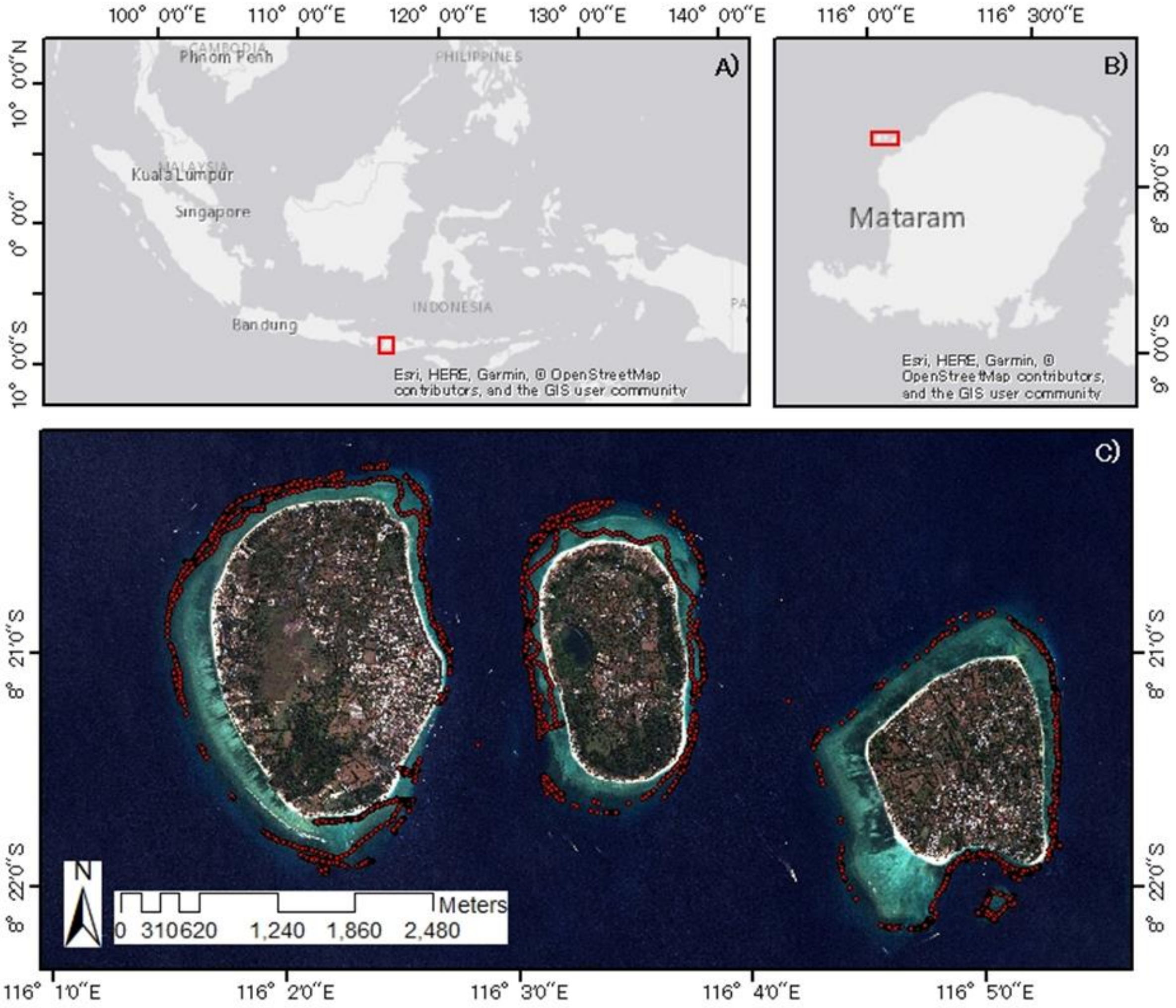

Figure 1: Study setting relative to (A) Indonesia and (B) Lombok Island; (C) True-colour SPOT-7 imagery and depth measurement points (red dots)

We used SPOT-7 multispectral imagery for the study area collected on 28 July, 2018 (Figure 1), using a 1B level that passed geometric correction with four wavelength bands: Blue (0.455–0.525 µm), Green (0.530–0.590 µm), Red (0.625–0.695 µm), and Near-Infrared (0.760–0.890 µm). Prior to further use, the raw SPOT-7 image was passed through radiometric and atmospheric correction pre-processing following methods previously defined in Manessa et al., (2017).

From 22–28 July, 2018, water depth data were collected using the single beam echo-sounder along the coasts of the Gili Islands by the Remote Sensing Application Centre team of the Indonesian National Institute of Aeronautics and Space and the Indonesian Navy's Hydrographic and Oceanographic Centre. The maximum water depth in the survey area reached ~60 m, but this study used about 10,000 water depth points ranging from 0.2–30 m (Figure 1). The depth measurements were referenced to the tidal datum (Lowest Astronomical Tides) and adjusted to tide level during satellite over flight before further use.

2.2. Random Forest Algorithm

The primary input in MDB is multiple linearized reflectance created from multispectral visible bands. Lyzenga (1978) showed that bottom reflectance could be assumed as an approximately linear function of bottom reflectance and an exponential function of the water depth. The natural logarithm function of the reflectance value was added to linearize the attenuation effect concerning depth. Thus, a transformed reflectance (*X*) can be built. In this linearization step, the Lyzenga method also engages noise correction using the average value of deep-water pixel radiance (Lyzenga, 1981). The equation for linearizing reflectance can be expressed as:

$$X_i = log\left(\rho_{c_i} - \bar{\rho}_{c_{\infty,i}}\right)$$

Equation 1

where ρ_i is observed spectral reflectance and $\rho_{\infty i}$ represents the water depth-averaged reflectance at band *i*. From this basic linearized reflectance, we propose several new modified variables as follows.

Theoretically, the relationship between depth and linearized surface reflectance should be linear but noise can cause a non-linear condition (Lyzenga, 1978). Manessa et al., (2016) proposed a new approach to determine the nonlinear relation between depth and linearized reflectance (an RF algorithm). RF for nonlinear regression is formed by growing trees dependant on a random vector such that the tree predictor takes on numerical values as opposed to class labels (Breiman, 2001). Then, the depth estimation formula can be written as:

$$\hat{h} = \frac{1}{m}\sum_{j=1}^{m} W_j(X_{blue}, X_{blue}{}') + \frac{1}{m}\sum_{j=1}^{m} W_j\left(X_{green}, X_{green}{}'\right) + \frac{1}{m}\sum_{j=1}^{m} W_j(X_{red}, X_{red}{}') + \varepsilon$$

Equation 2

where $W_j(X_i, X')$ is the non-negative weight of the *i*th training point relative to the new point *x*′ in the same tree, and *m* is a number of the tree.

Table 1: Modified predictors

New Variables	Equation	Source
Band ratio	$X_{ratio,} = X_{blue} / X_{green}$	Stumpt et al., (2003)
Modified linearized reflectance	$\hat{X}_i = log\left(\rho_i - \alpha_0 - \alpha_1\rho_{NIR}\right)$ $\hat{Y}_i \equiv exp(\hat{X}_i)^{-1}$ $\hat{Z}_i \equiv \rho_{c_{NIR}} exp(\hat{X}_i)^{-1}$ where $\alpha_{0,1} = coefficients\ from\ deep\ water$ i $= blue, green, red$	Kanno et al., (2010)
Depth-invariance index	$Y_{ij} = X_i - \frac{k_i}{k_j} X_j$ $where\quad i, j = blue, green, red$ $\frac{k_i}{k_j} = attenuation\ coeficient$	Lyzenga (1978)

2.2.1 Modified variables

The first modified variable consisted of band ratios proposed by Stumpt et al., (2003) to solve the problem of mapping shallow-water areas with significantly lower reflectance than adjacent areas. The second modified variables were based on linearized reflectance (Lyzenga, 1981) but were improved by adding the concept of relaxing uniformity assumption for the water and atmosphere. The third modified variable was a combination of two bands independent of depth and representing an index of bottom type (Lyzenga, 1978) known as the depth-invariance index.

In total, 15 predictors ($3X_i + 3\hat{X}_i + 3\hat{Y}_i + 3\hat{Z}_i + 1X_{ratio} + 2Y_{ij}$) were used (Table 1), then 32,769 different pairs of predictor were tested to evaluate the effective number of predictors. As a further test of each model's efficiency, a cross-validation experiment (70% training, 30% test) was performed over 10 iterations. For each batch, the model prediction accuracy was evaluated using two statistical keys: coefficient of determination (r^2) and root mean square error (RMSE).

2.2.2 Tuning hyperparameters

RF performance strongly depends on hyperparameter settings (Breiman, 2001) including number of trees, mtry, sampling size, and node size. This study used the "tuneRanger" package for r developed by Probst et al., (2018), a simple code for automatic tuning of the RF algorithm. This strategy uses model-based optimization with three parameters: node size (minimum), sampling size, and mtry, all tuned at once. However, this package excluded tree number from the tuning. As Probst et al., 2019 explained, variations in tree number do not need to be tuned; setting this to the highest range will improve the performance.

3. Results and Discussion

Increasing the number of predictors significantly increases the accuracy (high r^2 and low RMSE), with values ranging from 0.3–0.98 and 1–5 m for r^2 and RMSE, respectively (Figure 2). Moreover, a smaller number of predictors resulted in more variable accuracy but stabilized after reaching six predictors.

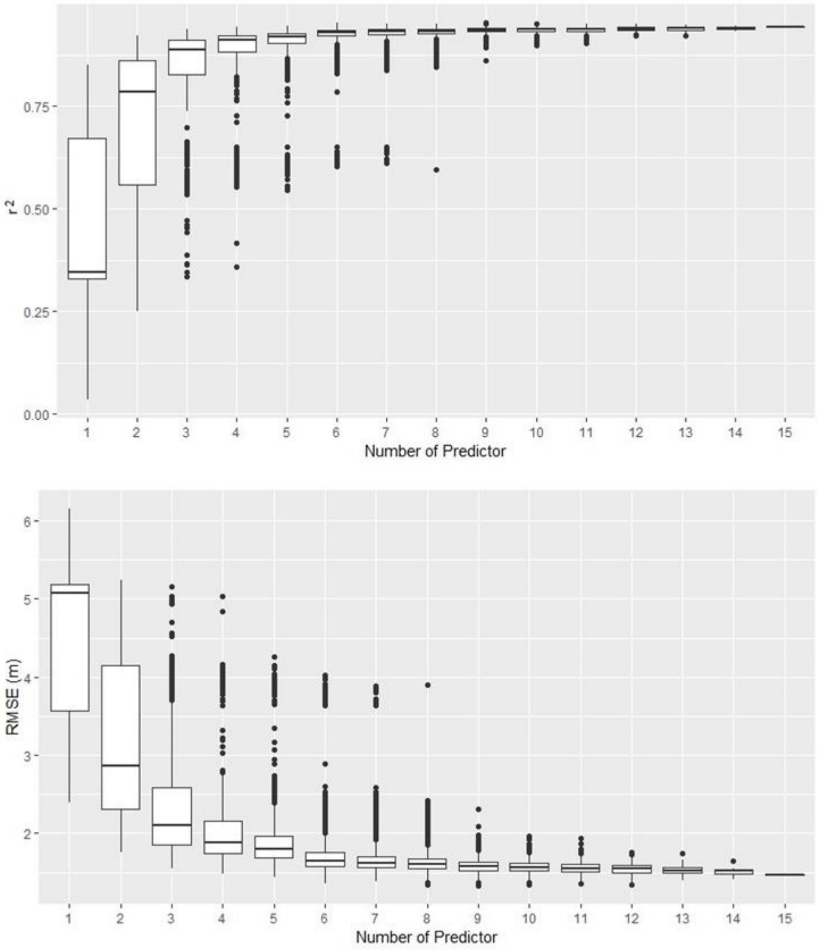

Figure 2: Correlation between accuracy and number of predictors for the RF model

Table 2: Accuracy assessment of depth prediction by different approaches

Case	**Number of predictors**	**Hyperparameter**				**CPU time**	**Accuracy**	
		ntree	**mtry**	**Sampling size**	**Node size**		**RMSE**	**r^2**
Case 1: standard predictor & auto-tuning (**Basic**)	3	500	2	0.632	5	5 s	1.45 m	0.95
Case 2: standard predictor & manual tuning	3	**1,000**	**6**	**0.894**	**2**	4 m 27 s	1.03 m	0.97
Case 3: predictor enhancement & auto-tuning	**15**	500	2	0.632	5	12 s	1.1 m	0.97
Case 4: predictor enhancement & manual tuning (**Optimization**)	**15**	**1,000**	**5**	**0.897**	**2**	10 m 41 s	0.78 m	0.98

The before and after performance of tuning the hyperparameter settings and number of predictors (Table 2) shows that the basic setting with three predictors (linearized reflectance of blue, green, and red bands) and auto-tuning the hyperparameter achieved a prediction accuracy of 1.45 m RMSE and 0.95 r^2. Adding a hyperparameter optimization to the basic case improved the accuracy by 0.42 m RSME and 0.02 r^2, while increasing the number of predictors showed less improvement (by 0.35 RSEM and 0.02 r^2). The best performance was achieved when predictor numbers and hyperparameters were tuned, with greater improvements in accuracy (by 0.67 RMSE and 0.03 r^2).

This study is the first to determine the performance of RF under such optimization conditions. Given that the addition of model tuning has improved prediction accuracy in the past in this location (Manessa et al., 2018), these error reductions were expected. Another reason for the improved performance when compared with the past study was the close time gap between survey and image acquisition. Although the sea floor topography was assumed to be constant under a no-hazard condition, we found that a smaller time gap still produced the best prediction result.

Increasing the number of predictors resulted in 24% less error in predicting water depth and improved the classification result, supporting past study results (Kanno et al., 2010 and Smith, 2010). It should be noted that understanding the imagery's spectral characteristics is crucial to developing a new predictor. While the hyperparameter tuning process created 29% less error, past RF studies showed an identical result and recommended applying the tuning step (Mutanga et al., 2012, Smith, 2010 and Svetnik et al., 2003). This process is easy to add to the processing of an RF model because it is not time-consuming.

4. Conclusions

We demonstrated the potential of an RF optimization algorithm for extracting water depth from multispectral imagery. This was performed in two steps: improvement of predictors and tuning hyperparameters. Applied individually, the latter contributed more than the former, but the combination of both increased prediction accuracy by 0.67 m (RMSE) and 0.3 (r^2) compared to a basic RF algorithm. We therefore suggest applying this optimization approach to other remote sensing applications using RF algorithms to achieve a significant increase in accuracy with a process that is simple and not time-consuming.

Acknowledgments

This work was supported by the University of Indonesia under research grant PUTI 2020. The authors are grateful to two anonymous reviewers whose advice and comments helped to improve this work.

Reference

Belgiu, M. and Drăgu, L. 2016, Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 114, 24-31.

Breiman, L., 2001, Random Forests. *Machine Learning*, Vol. 45(1), 5-32.

Cánovas-García, F., Alonso-Sarría, F., Gomariz-Castillo, F. and Oñate-Valdivieso, F., 2017, Modification of the Random Forest Algorithm to avoid Statistical Dependence Problems when Classifying Remote Sensing Imagery. *Computers & Geosciences*, Vol. 103, 1-11.

Chen, T., Trinder, J., Niu, R., Chen, T., Trinder, J. C. and Niu, R., 2017, Object-Oriented Landslide Mapping Using ZY-3 Satellite Imagery, Random Forest and Mathematical Morphology, for the Three-Gorges Reservoir, China. *Remote Sensing*, Vol. 9(4), 1-14, DOI: 10.3390/rs9040333.

Kanno, A., Koibuchi, Y. and Isobe, M., 2010, Statistical Combination of Spatial Interpolation and Multispectral Remote Sensing for Shallow Water Bathymetry. *IEEE Geoscience and Remote Sensing Letters*, Vol. 8(1), 64-67.

Kanno, A., Koibuchi, Y. and Isobe, M., 2011, Shallow Water Bathymetry from Multispectral Satellite Images: Extensions of Lyzenga's Method for Improving Accuracy. *Coastal Engineering Journal*, Vol. 53(4), 431-50.

Lyzenga, D. R., 1978, Passive Remote Sensing Techniques for Mapping Water Depth and Bottom Features. *Applied Optics*, Vol. 17(3), 379-383.

Lyzenga, D. R., 1981, Remote Sensing of Bottom Reflectance and Water Attenuation Parameters in Shallow Water Using Aircraft and Landsat Data. *International Journal of Remote Sensing*, Vol. 2(1), 71-82.

Manessa, M. D. M., Haidar, M., Hartuti, M. and Kresnawati, D. K., 2018, Determination of the Best Methodology for Bathymetry Mapping Using Spot 6 Imagery: a Study of 12 Empirical Algorithms. *International Journal of Remote Sensing and Earth Sciences (IJReSES)*, Vol. 14(2), 127-136.

Manessa, M. D. M., Kanno, A., Haidar, M., Sekine, M., Higuchi, T., Yamamoto, K. and Imai, T., 2016, Satellite-Derived Bathymetry Using Random Forest Algorithm and Worldview-2 Imagery. *Geoplanning: Journal of Geomatics and Planning*, Vol. 3(2), 117–126.

Mishra, D. R., Narumalani, S., Rundquist, D., Lawson, M. and Island, R., 2005, High-Resolution Ocean Color Remote Sensing of Benthic Habitats: A Case Study at the Roatan Island, Honduras. *IEEE Geoscience and Remote Sensing Letters*, Vol. 43(7), 1592-1604.

Mohamed, H. and Nadaoka, K., 2019, Assessment of a Hybrid-Based Approach with a Random Forest Ensemble for Determination of Shallow Water Depths from Multispectral Satellite Images. International Journal of Geoinformatics, Vol. 15(1), 47-58.

Mohamed, H., Negm, A., Nadaoka, K., and Elsahabi, M., 2016, Comparative Study of Approaches to Bathymetry Detection in Nasser/Nubia Lake Using Multispectral SPOT-6 Satellite Imagery. *Hydrological Research Letters*, Vol. 10(1), 45-50.

Mohamed, H., Negm, A., Salah, M., Nadaoka, K. and Zahran, M., 2017, Assessment of Proposed Approaches for Bathymetry Calculations Using Multispectral Satellite Images in Shallow Coastal/Lake Areas:A Comparison of Five Models. *Arabian Journal of Geosciences*, Vol. 10(2), DOI: https://doi.org/10.1007/s12517-016-2803-1

Mutanga, O., Adam, E. and Cho, M. A., 2012, High Density Biomass Estimation for Wetland Vegetation Using WorldView-2 Imagery and Random Forest Regression Algorithm. *International Journal of Applied Earth Observation and Geoinformation*, Vol. 18, 399-406.

Probst, P., Wright, M. N. and Boulesteix, A., 2019, Hyperparameters and Tuning Strategies for Random Forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 9(3), 1-18. https://doi.org/10.1002/widm.1301.

Smith, A., 2010, Image Segmentation Scale Parameter Optimization and Land Cover Classification Using the Random Forest Algorithm. *Journal of Spatial Science*, Vol. 55(1), 69–79.

Stumpf, R. P., Holderied, K. and Sinclair, M., 2003, Determination of Water Depth With High-Resolution Satellite Imagery Over Variable Bottom Types. *Limnology and Oceanography*, Vol. 48(1), 547-556.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P. and Feuston, B. P., 2003, Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, Vol. 43(6), 1947-1958.

Turner, D., Lucieer, A., Malenovský, Z., King, D. and Robinson, S. A., 2018, Assessment of Antarctic Moss Health from Multi-Sensor UAS Imagery with Random Forest Modelling. *International Journal of Applied Earth Observation and Geoinformation*, Vol. 68, 168-179.