Random forest for large-scale forest resource inventory in northern Canada

C. Mahoney1, C. Hopkinson1, R. Hall2, & M. Filiatrault2

1Department of Geography, University of Lethbridge, Lethbridge AB T1K 3M4, Canada

2 Northern Forestry Centre, Canadian Forest Service, Natural Resources Canada, Edmonton AB T6H 3S5, Canada

# Abstract

Large-scale contemporary forest resource inventories (FRI) are becoming increasingly important in enabling policy makers to establish sustainable forest management strategies, especially as climate change continues to affect our planet’s forest regions with uncertain consequences. The current study follows on from a previous study that investigated the use of the historically favored k-nearest neighbor (kNN) imputation technique for predicting FRI. The current study follows the exact same methods as the kNN study, but substitute’s random forest (RF) imputation in place of kNN to predict stand height and crown closure over a 200,000 km2 region in northern Canada. Predictions from both methods are compared with equivalent values from an independent ALS dataset, indicating that RF (R2=0.45, RMSE=4.44 m) better characterizes stand height relative to kNN (R2=0.39, RMSE=4.66 m), whereas crown closure exhibits a slight decrease in accuracy (kNN: R2=0.41, RMSE=7.00 %, RF: R2=0.41, RMSE=7.09 %). The performance of RF highlights its use as a means to quantify FRI in dynamic landscapes at large geographies and possibly beyond, and even replacing kNN in all future investigations.

# Introduction

Northern boreal forests are of key importance regarding the sustainability of global forests through management strategies, particularly as this region of the planet is warming most rapidly ([Winton 2006](#_ENREF_53), [IPCC 2013](#_ENREF_22), [IPCC 2014](#_ENREF_23)). Approximately 28% of the world’s boreal forest falls within Canada ([Brandt 2009](#_ENREF_5), [NRCan 2015](#_ENREF_42)) where the primary means of understanding forest status is via forest resource inventory (FRI) ([Gillis 2001](#_ENREF_10), [Wulder et al. 2004](#_ENREF_55), [Gillis et al. 2005](#_ENREF_11)). Typically Canadian forest inventory practices have relied on the interpretation of aerial photographs ([Leckie and Gillis 1995](#_ENREF_24), [Hall 2003](#_ENREF_12)). However, over vast, largely inaccessible regions such as the Northwest Territories (NWT), completing seamless coverage by such methods is cost prohibitive, as a result existing inventory is sparse and inconsistent ([Smith 2002](#_ENREF_51)). Aside from FRI data, the most spatially complete vegetation information available in the NWT was completed as part of the Earth Observation for Sustainable Development of Forests (EOSD) project to map land cover for the forested areas of Canada (circa 2000) and was later updated in 2007 ([Wulder et al. 2003](#_ENREF_54)). At present, multi-sensor remote sensing data and models are used to estimate forest structure, stand volume and aboveground biomass as value-added products within the framework of the circa 2007 EOSD land cover maps ([Hall et al. 2012](#_ENREF_14)).

Given the well documented success of Airborne Laser Scanning (ALS) data, and satellite-based Geoscience Laser Altimeter System (GLAS) data, transitioning to utilize such data is key for assessing FRI moving forward. Recent developments have focused on frameworks (Mahoney et al. 2017) that allow a variety of input data (such as field, ALS, and spaceborne) to map key attributes to regional scales. Current within framework methods typically rely on k-nearest neighbor (kNN) algorithms due to their historic capabilities regarding FRI attribute mapping (McRoberts et al. ????). However, a variety of machine learning techniques are available for use in such frameworks the capabilities in such a context have not been quantified; of particular interest here is Random Forest (Breiman 2001) due to its feature selection capabilities. Given these capabilities, the current study will mirror that of Mahoney et al. (2017) but will substitute the use of kNN with RF, and compare results in an attempt to determine which algorithm yields best results for FRI quantification.

# Data & Methods

This study was conducted over a 200,000 km2 region of interest in northern Canada within the Taiga Plains ecozone; coniferous vegetation dominates, followed by mixedwood and deciduous vegetation types as assessed by the EOSD (circa 2007). The region exhibits strategically sampled field data acquired near Fort Simpson in support of an ALS survey (2007), which in turn intersects with multiple GLAS footprints which are also spread throughout the entire region. A total of 38 filed plots were utilized to generate ALS models of stand height and crown closure, which in turn were used to develop GLAS models based on 43 overlapping footprints. GLAS models were applied to laser 2A (fall 2003) and 3A (fall 2004) campaigns only, due to their acquisition times being at similar times of the year, minimizing the effect of varying phenology and snow presence. Both campaigns were subject to a sequential 4 level quality control ruleset which eliminated spurious points from an uncontrolled dataset (level 1) based on height anomalies (level 2), slope (level 3), and snow cover (level 4). Each level of quality control and corresponding predictor variables were submitted to the RF algorithm for training, where regionally available predictor variables were employed to infer regional predictions of stand height, crown closure, and associated uncertainties. Predictor variables were the EOSD land cover map (circa 2007), Landsat bands 3, 4, and 5, elevation, climate moisture index, compound topographic index, and soil moisture index.

An independent ALS dataset (2010) was employed to assess regional predictions using 1% of all available ALS data (12020 points) by the compilations of the following statistics: coefficient of determination (R2), root mean squared error (RMSE), and mean absolute difference (MAD). The best combination of resulting statistics (greatest R2 and lowest RMSE and MAD) from the kNN results of Mahoney et al. (2017) and the current RF analysis were used to determine which level of GLAS quality control and imputation method produced best regional characterizations of the desired forest attributes.

All model development and assessment methods employed in this study mirror those followed by Mahoney et al. (2017), where RF replaces kNN for regional predictions; for any analytical clarifications see the literature documenting the original study.

# Results

Random forest predictions based on level 3 quality controlled GLAS data (Table 1) produced best regional estimates of stand height (Figure 1a) and crown closure (Figure 1a) with respect to independent ALS assessment. Stand height ranges reflected those noted in the field, ALS, and GLAS data, whereas as crown closure estimates appear to saturate at approximately 65%. As a result, summary statistics are poor, but more so for crown closure. Regional uncertainties, scale as a function of attribute value, where stand height (Figure 1c) uncertainties exhibit a more linear relationship than crown closure equivalents (Figure 1d) which appear more sporadic, in both cases, RF modelled uncertainties are lower than kNN equivalents. Maximum regional uncertainties are 13.73 m and 18.61 % for stand height and crown closure, respectively. A comparison between kNN (Mahoney et al. 2017) and random forest performance statistics (Table 1) indicate the latter yields more accurate stand height estimates at the regional scale, but exhibit a slight degradation in crown closure accuracies.

*Table 1.*

*Figure 1.*

Land cover analysis indicates that coniferous vegetation was best characterized across the region for both stand height and crown closure (Table 2), as determined by ALS assessment. This is an expected outcome due to the prevalence of coniferous vegetation throughout the training dataset, and the region of interest. Mixedwood stands were second best characterized, followed by deciduous stands.

*Table 2.*

# Discussion & Conclusion

Random forest predictions of FRI attributes performed better than those reported for kNN (Mahoney et al. 2017) over the exact same landscape. Greatest improvements were noted for stand height in all model performance statistics (R2, RMSE and MAD), where the converse was noted for crown closure with the exclusion of R2 which remained constant. Greatest improvements are expected in stand height as all models, from field plots thru to GLAS data exhibit greater transferability than crown closure. The noted performance degradation in RF crown closure with respect to kNN equivalents is expected to be related to the relatively poor models that characterize crown closure at smaller scales, alternatively, or simultaneously, RF may be more sensitive to more noisy data.

Differences in resulting regional FRI maps are expected to be rooted in the differences between the technical make-up of each algorithm, more specifically, the feature selection and reliance on Shannon entropy embedded within RF. The ability to feature select (i.e. discard predictors that are poor classifiers of the response variable) allows a somewhat autonomous pursuit of maximum overall model accuracy. In addition, the role of Shannon entropy to build an ensemble decision tree enables RF to operate without an assumed model form implied allowing for a purely statistically based model to be extracted, the outputs from which are heavily dependent on the exact data that are used as inputs. kNN model estimates are derived from proximity weighted inputs from a training dataset, therefore outputs are expected to be more closely related to their physical (spatial) surroundings, which is not always the case for RF predictions.

Random forest algorithms exhibit more layers of complexity than kNN, and hence may explain improved estimates of FRI attributes. However, even with increased complexity, RF algorithms do not require more computational resources, nor takes longer to execute than kNN equivalents. Furthermore, RF calculations (for the current study) are expected to exhibit greater confidence in predictions (narrower confidence intervals) because final RF predictions (per map cell) were based on 500 individual predictions for both stand height and crown closure, whereas kNN equivalents were based on 6 and 25 nearest neighbors, respectively. These optimal number of neighbors were determined by a well-documented kNN specific optimization process (McRoberts et al. ????, etc.), but systematically restrict model confidence due to the small sample of points from which outputs are inferred.

The current study offers insight to the usefulness of RF as an alternative to kNN for FRI, however, the limitations of use with respect to RF has not been fully tested, unlike kNN which has a greater wealth of historic studies to draw from (Citations). Despite this, the current study demonstrates that RF is a superior classifier of stand height and crown closure over a large, dynamic region in the Canadian north that represents a cross-section of boreal forest characteristics. The suitability, and performance of RF methods with respect to kNN methods has not been tested for significantly different landscapes, therefore conclusions cannot be drawn with respect to the expected performance of RF in other non-boreal regions. However, it is worth noting that RF has been employed for forest attribute quantification elsewhere (Mahoney et al. 2016, etc.), no direct comparisons with other algorithms have been demonstrated.

Based on presented results, we advocate the use of RF as an alternative to kNN for FRI inferences over large geographies. Although kNN exhibits a historic track record in such mapping, RF has demonstrated that improvements can be found with minor modification to current workflows. However, caution should be expressed were noisy datasets are concerned as the exact capabilities of RF under such conditions requires further analysis.

# References

Table 1 Summary statistics for regional predictions of stand height and crown closure from RF and kNN (Mahoney et al. 2017) models for filtered input data from levels 1 to 4; statistics are based on assessment by independent ALS data. Best results are bold.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Stand height [m]** | | | |  | **Crown closure [%]** | | | | |  |
| **Filter** | **N** | **Adj R2** | **RMSE** | **MAD** | **Min** | **Max** | **Adj R2** | | **RMSE** | **MAD** | **Min** | **Max** |
| **Random Forest** | | | | | | | | | | | | |
| 1 | 12020 | 0.27 | 5.11 | 3.77 | 4.70 | 64.04 | 0.37 | | 7.30 | 7.12 | 24.45 | 65.79 |
| 2 | 0.37 | 4.74 | 3.55 | 4.70 | 26.40 | 0.38 | | 7.25 | 7.20 | 24.80 | 65.17 |
| 3 | **0.45** | **4.44** | **3.35** | **4.75** | **25.80** | **0.41** | | **7.09** | **7.52** | **25.01** | **64.83** |
| 4 | 0.40 | 4.65 | 3.48 | 4.85 | 26.83 | 0.38 | | 7.27 | 7.22 | 25.61 | 64.31 |
| **kNN** | | | | | | | | | | | | |
| 1 | 12020 | 0.28 | 5.08 | 3.82 | 4.79 | 49.03 | 0.37 | | 7.25 | 7.53 | 25.96 | 62.25 |
| 2 | 0.32 | 4.92 | 3.71 | 4.63 | 26.65 | 0.38 | | 7.20 | 7.53 | 26.72 | 61.85 |
| 3 | **0.39** | **4.66** | **3.50** | **4.65** | **25.38** | **0.41** | | **7.00** | **7.86** | **26.96** | **60.98** |
| 4 | 0.35 | 4.84 | 3.63 | 4.65 | 27.07 | 0.38 | | 7.19 | 7.47 | 28.91 | 63.28 |
| RMSE = root mean squared error; MAD = mean absolute difference | | | | | | | |  | | | |  |

Table 2 Summary statistics associated with best RF predictions (level 3 filtered GLAS) of stand height and crown closure as a function of dominant EOSD land cover; statistics are based on assessment by independent ALS data.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Stand height [m]** | | | |  | **Crown closure [%]** | | | | |  |
| **Land cover** | **N** | **Adj R2** | **RMSE** | **MAD** | **Min** | **Max** | **Adj R2** | | **RMSE** | **MAD** | **Min** | **Max** |
| Coniferous | 1000 | 0.40 | 2.52 | 2.91 | 4.76 | 24.84 | 0.40 | | 4.87 | 7.50 | 25.39 | 55.61 |
| Deciduous | 0.18 | 3.22 | 4.64 | 6.23 | 24.98 | 0.03 | | 5.80 | 5.61 | 32.56 | 64.83 |
| Mixedwood | 0.24 | 2.87 | 4.69 | 5.19 | 23.94 | 0.22 | | 4.13 | 9.72 | 28.52 | 56.55 |
| RMSE = root mean squared error; MAD = mean absolute difference | | | | | | | |  | | | |  |

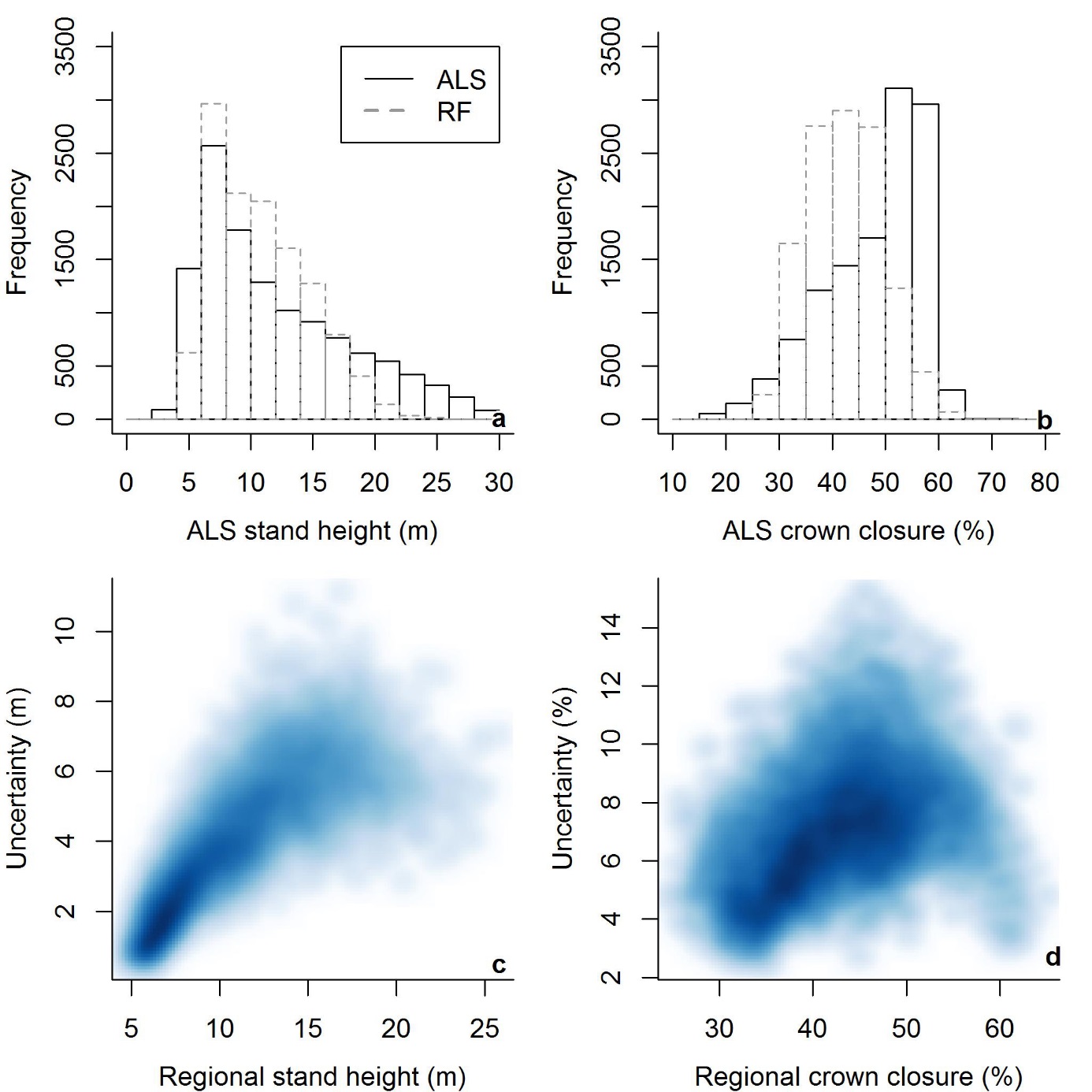


Figure 1 Comparisons of the ALS validation, and regionally predicted RF distributions of a) stand height, and b) crown closure, and the relationship between regional predictions and their associated uncertainties for c) stand height, and d) crown closure for 1% of intersecting ALS data. Darker shades indicate higher point density.