An evaluation of the predictive performance of the Random Forest model for predicting wildfire residuals

Yikalo H Araya[1], and Tarmo K Remmel[2]

1 Geography, York University, Toronto, ON, yikalo@yorku.ca
2 Geography, York University, Toronto, ON, remmelt@yorku.ca

Abstract

We present a method for developing spatially explicit probability maps for the presence of wildfire residuals within a burned landscape. Using the Random Forest method, we learn rules that explain the formation of wildfire residuals based on selected physical predictors. We then implement the rules (akin to inverting the learning algorithm) to build maps of likely residual stand locations. First, satellite derived data from eleven fire events (from the same ecoregion) are partitioned into training and validation using a hold-out approach. The performance of the model is then assessed using an independent and extensive fire event and using threshold-independent measures at 4, 8, 16, 32, and 64 m spatial resolutions. The model has a reasonable or high predictive performance ('marginal' or strong' model outcome) for most of the fire events within the same ecoregion. However, the predictive power of the model is lower for the independent fire event. We further characterize the relative importance of each predictor for the presence of wildfire residuals and identify whether certain land cover types are more likely to escape burning than other types. Our results suggest that the variables interactively affect the residuals occurrence, but natural firebreaks, specifically wetlands and water, are among the most important predictors. The results also indicate that land cover types such as dense conifer and treed wetland appeared to significantly over-represented among the wildfire residuals.

Background and Relevance

Wildfire in boreal forests is usually intense and frequent, and consumes substantial forest cover but does not burn the entire landscape. Owing to the variation in the geo-environmental factors, wildfire creates a complex and heterogeneous spatial mosaic (van Wagtendonk, 2004); resulting in the presence of wildfire residuals. Wildfire residuals are broadly defined as remnants of the pre-fire forest ecosystem that retain their structure and are not completely changed to ash or charcoal (Perera & Buse, 2014). The residuals have been described as insular or peninsular (Perera et al., 2009). Insular patches refer to contiguous undisturbed areas that are entirely contained within a fire perimeter while a peninsular patch is the undisturbed forest patch that are connected physically to the surrounding forested matrix along a narrow interface. We consider only the insular type of residuals due to the ability to define them explicitly, regardless of the age, size, and type of forest species that form them.

Understanding the patterns of wildfire residuals has become a common approach for implementing disturbance-based management practice. Specifically in Ontario, mapping the characteristics of wildfire residuals has become a primary requirement for emulating forest disturbances, emerging as a general forest management goal within disturbance driven landscapes (Perera et al., 2009). The presence of wildfire residuals can be mapped using remotely sensed imagery coupled with field observations. This provides only a snapshot of wildfire residual occurrence, but timely and spatially explicit information on residuals occurrence is required for effective resource management

(Beauvais et al., 2006). This requires the design of a consistent and replicable measurement framework in the study of wildfire residuals. However, the presence of wildfire residuals has not been well recognized in the fire literature (Perera & Buse, 2014); studies on boreal wildfires have focused mainly on the patterns and processes of wildfire and their ecological effects (van Wagtendonk, 2004).

Knowing the variables that explain the wildfire residuals, and the site conditions at which wildfire residuals are likely to occur forms a basic component of natural resource management and ecological research (Beauvais et al., 2006). In this regard, a number of highly computational statistical methods have emerged to unravel the complex interactions among the variables that explain spatial patterns. Random Forests (RF) is one such method (Breiman, 2001). In this context, our broad goal is to develop a replicable approach based on RF for determining the relative importance, assessing the combined effects of the physical variables, and evaluating the predictive performance of RF model. RF is an ensemble-learning method that combines multiple models built using several bootstrap samples (Breiman 2001).

Methods and Data

In our previous study, we developed a predictive model using data records from eleven fire events (from the same ecoregion), where data records from a single fire event was used for testing while data records from the remaining fire events were considered for training (Araya & Remmel, 2013). The model was developed using the RF algorithm as implemented in R. The algorithm begins with generating multiple bootstrapped samples, and builds a number of unpruned classifications for each bootstrapped sample set (Breiman 2001). In a typical bootstrap sample, two-thirds of the data are used for constructing any particular tree. Observations in the original dataset do not occur in a bootstrap sample (i.e., one-third of the data that are not used in the construction of a tree) are called out-of-bag (OOB) observations. The trees are fully grown and each is used to predict the OOB observations; the predicted class of an observation is calculated by majority vote. In our study, we evaluated the predictive performance of the RF model using data records from an independent and extensive fire event (RED084). The RED084, ignited by lighting, was occurred in northwestern Ontario in 2011 and burned a total area of 54,828 ha. Unlike the eleven fire events, the RED084 is located within a different ecoregion and within the area of undertaking where forest management practice is permitted. A supervised classification approach was used to map the residuals and the extent of the fire footprint. We resampled the classified image into 4, 8, 16, 32, and 64 m spatial resolutions, hereafter described as R_4, R_8, R_{16}, R_{32}, and $R_{64,}$ based on a non-overlapping block-majority filter for multi-scale analyses. The use of RF for predictive model requires a response and explanatory variables. The response variable often incorporates the presence-absence data; hereafter described as residual and null-residual patches. However, the vast majority of ecological data that are available today are consisting of presence-only datasets; yet, presence-only data are the most difficult element to integrate into statistical modeling (Zaniewski et al., 2002). Additionally, models based on presence-only data do not provide a better performance (Pearce & Ferrier, 2000).

We developed a model based on presence-absence data where the existing residuals were considered as presence-data, but information about the absence data is not readily available. Therefore, a computer simulation approach has been suggested to algorithmically generate null-residual patches. Yet, models designed based on presence-absence data can be affected by class imbalance (Evans & Cushman, 2009). In order to develop a model based on presence-absence data, a simulation algorithm was developed to extract null-residual patches. The algorithm was designed to randomly generate null-residual patches in which the size, shape and orientation of the null-residual patches mimic the residual patches and hence class imbalance would be avoided. The explanatory variables used for the prediction are topographic variables (slope - SL, ruggedness index – RI, and elevation -EL), vegetation cover type (LC), and firebreak features (water - WA, wetland -WL, and non-vegetated areas -BV). The variables were obtained from different sources (digital elevation models and existing land cover maps), and were selected based on a prior ecological studies.

Given the physical variables, we implemented the RF model to determine their relative importance and evaluate the model's performance using an independent dataset. A model based on RF was applied because RF: 1) is a nonparametric and adds an additional layer of randomness; 2) does not over-fit, 3) has high predictive power, and 4) provides additional pieces of information (e.g., importance of variables) (Breiman, 2001). RF also provides error statistics, which is indicative of model fit, but not necessarily the predictive performance of a model. While determining the relative importance of the variables using a mean decrease in accuracy, we specifically examined whether some land cover types were more likely to be observed in residuals than expected under a randomness assumption using a partial Chi-square ($X^2{}_p$). The proportion of land cover observed in the footprint was considered as 'expected' while the land cover proportion within residuals was considered as observed values.

The RF model was calibrated using data records from the 11 fire events while data from independent event (RED084) was used for evaluating the model using threshold-independent measure – receiver operating characteristics curves (ROC). The ROC, which is used to assess the accuracy of the model, provides a graphical depiction of model's discrimination ability over a range of threshold values (Pearce & Ferrier, 2000). However, comparing ROC curves directly from the plot has never been easy; a single index that describes the discrimination ability of a model is required (Zweig & Campbell, 1993). The area under the resulting ROC curve, the AUC, is then considered as an indicator of model's performance. The AUC provides a single measure of model's ability to distinguish between residual and null-residual patches, independent of a specific threshold value. The AUC value was computed for each of the ROC plots to evaluate the model's performance.

Results

The importance values across five spatial resolutions are shown in Figure 1. The results indicate an interactive effect of the variables; yet firebreaks (e.g., wetland) remained the most important predictor. The marginal effect of this important variable was examined, and the results indicated that majority of the residuals are concentrated within 100 m from the wetlands. The results of the land cover composition of residuals suggested that land cover types such as dense conifer, open and treed wetlands appeared to significantly over-represented among the wildfire residuals in RED084 (Table 1).

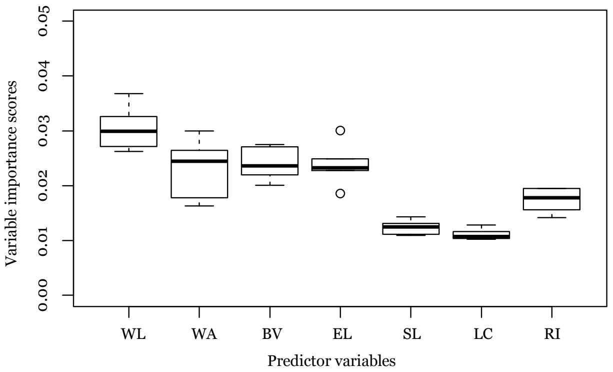

Figure 1. The variability of the relative importance of the predictors.

Critical p value	*0.2441*	
	$X^2{}_p$	p
Sparse conifer	-0.6773	-
Deciduous	0.0495	*
Dense conifer	0.9298	+
Open wetland	1.6623	+
Treed wetland	4.4442	+
Other	-0.0543	*

Table 1. land cover composition of residuals: the positive (+) and negative (-) sign indicates an over and under-representation of the specified land cover type within wildfire residual patches compared with the proportion of the same land cover category within the fire footprint prior to burning respectively. * indicates no significant difference.

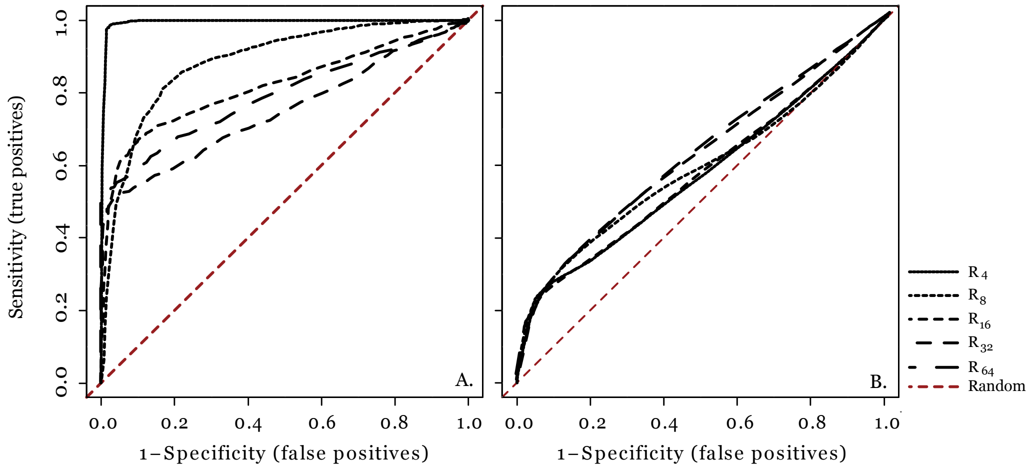

Figure 2. ROC Curves: A) the largest fire event from the training data while B) is the ROC curve for the independent event.

The predictive accuracy of the model for one of 11 fire events (F01) and for RED084 is graphically summarized in (Figure 2). A model that perfectly predicts the residuals generates an ROC curve that follows the left axis and top of the plot, whilst a model with random predictions produces a curve that follows a 45° diagonal from the lower left corner to the upper right corner. The curve for F01 at R_4 appeared closer to perfect discrimination, but it was important to compare the results with the AUC values.

The AUC provides a summary measure of model's performance; the ROC curve with the larger area is, on average, more accurate (Pearce & Ferrier, 2000). As a general rule, the AUC value includes: random guess (AUC = 0.5), low accuracy (0.5 ≥ AUC ≤ 0.7), reasonable accuracy (0.7 ≥ AUC ≤ 0.9), and high accuracy (AUC > 0.9) (Swets, 1988).

Table 2. AUC values for selected fire events.

	Spatial resolutions				
	R4	**R8**	**R16**	**R32**	**R64**
F01	0.995	0.886	0.816	0.749	0.793
RED084	0.571	0.583	0.572	0.615	0.617

The model has the highest discrimination accuracy with AUC value of 0.995 for F01 (Table 2). Based on the rule of thumb set by Swets (1988), the RF model was evaluated as having reasonable to excellent discrimination ability for F01 across the gradient of scales. The results for F01 suggested that the occurrence of wildfire residuals appeared to be explained by the predictors incorporated in the model. However, the model had low predictive performance for the independent fire event, with AUC values that lies within the range of 0.5 and 0.7 across all grain sizes (Table 2). One possible explanation for the low predictive accuracy is that there is an inter-landscape difference within the boreal forest as a function of various physical variables (Burton et al., 2008). It is argued that depending on fuel availability and source of ignition, every fire represents a unique fire combination of fire skips that affect forest species and the subsequent wildfire residuals. Despite the low prediction accuracy, the results of our predictive probability maps (Figure 3) showed that the model was able to identify potential areas (unburnable areas, specifically wetlands) for wildfire residual occurrence. This supports our variable importance assessment where firebreak features (e.g., wetlands) were among the most informative predictors.

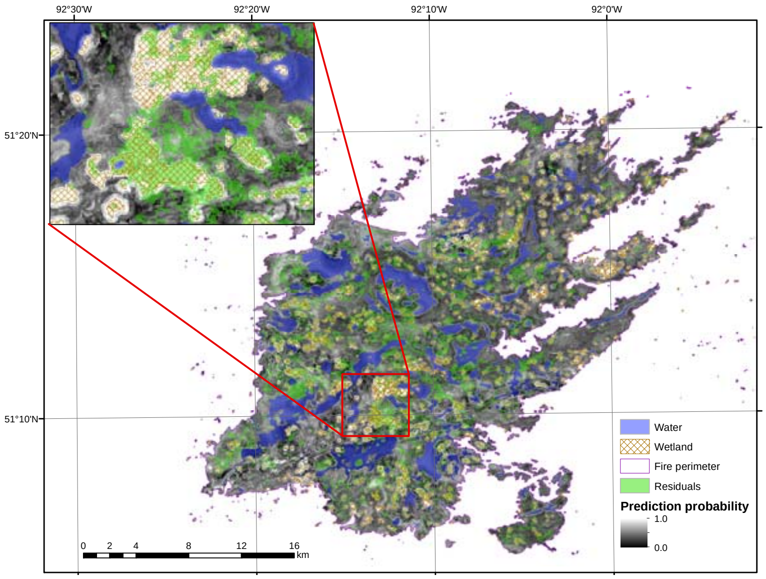

Figure 3. Predicted probability map of residual patch occurrence in the RED084 at R32; light-shading (greater chance of occurrence), green areas (existing wildfire residuals), and cross-hatched areas (distribution of wetland).

Conclusions

We sought to determine the relative importance of the predictors to explain wildfire residuals. Our results revealed that although the variables interactively affect the residuals, firebreaks (wetlands) are the most important predictors to explain the wildfire residuals. Our study also showed that certain land cover types such as dense conifer and treed wetland are likely to escape burning than other land cover types. We evaluated the predictive power of RF model in relation to the combined effects of the variables; a model with good discrimination ability is the one that correctly discriminate between presence and absence in the evaluation dataset, irrespective of the reliability of the predicted probabilities. Our results showed that the predictive power of the model was relatively poor when it is applied in an independent fire event; yet the model was able to identify potential areas (e.g., wetlands) where residual patches are likely to occur. This reflects the potential of the variables (and the model itself) to explain residuals. For all the merits of RF in prediction, its interpretability is limited; it is a black-box and does not provide set of rules that are often obtained from standard classifications (Evans & Cushman, 2009). However, RF excels at identifying predictor variables and visually characterizing the relationship between predictor variables and predicted classes. Therefore, the approach implemented in this study was determined as consistent and replicable for learning complex and non-linear ecological relationship, and predicting wildfire residuals.

References

Araya, Y.H., Remmel, T.K. (2013). Spatially explicit model predicting residual vegetation patch existence within boreal wildfires. Spatial knowledge and Information (SKI), Banff, AB.

Beauvais, G.P., Keinath, D.A., Hernandez, P., Master, L., & Thurston, R. (2006). *Element distribution modeling: a primer.* Retrieved from http://www.natureserve.org/prodServices/pdf/EDM_white_paper_2.0.pdf

Breiman, L. (2001). Random Forests. *Journal of Machine Learning, 45*(1), 5-31.

Burton, P.J., Parisien, M., Hicke, J., & Freeburn, J. (2008). Large fires as agents of ecological diversity in the North American boreal forest. *Int. Journal of Wildfire, 17*, 754-767

Evans, J. E., & Cushman, S.A. (2009). Gradient modeling of conifer species using Random Forests. *Landscape Ecology, 24,* 673-683.

Perera, A.H., Buse, L.J. (2014). Ecology of wildfire residuals in boreal forests. Wiley Blackwell.

Perera, A.H. Remmel, T.K., Buse, L.J, & Ouellete, M.R. (2009). *An assessment of residual patches in boreal fires, in relation to Ontario's policy directions for emulating natural forest disturbance*. OFRI, Sault Ste. Marie, Ontario. Forest Research Report No. 169.

Pearce, J., & Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling, 133*, 225-245

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285-1293.

Van Wagtendonk, J.W. (2004). Fire and landscapes: patterns and processes. USDA Forest service General Technical Report. PSW-GRT-193.

Zaniewski, A.E., Lehmann, A., & Overton, M.C. (2002). Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modeling, 157*, 261-280.

Zweig, M.H., & Campbell, G. (1993). Receiver-Operating Characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39*(4), 561-577.