# **Summary of research**

Increases in the global temperature, as a result of climate change, have led to impacts to the wildfire regime (Collins *et al.* 2013). These impacts are greatest in northern environments were the warming is occurring four times faster than the global average (GNWT 2019). While it is well understood that climate change is altering the wildfire regime, understanding the primary drivers of this change is crucial to predicting future forest compositions. My analysis assesses the impacts of site specific factors that may be drivers of increased fire severity. Fire severity is one of the components of a wildfire regime and can be measured through the depth of soil organic layer consumed during the fire (Walker *et al.* 2018a). Using data from sites that burned in 2014 in the Northwest Territories, Canada I hypothesis that soil moisture, slope, stand age, and dominant tree species of the site will impact the depth of organic soil burned during wildfire. Statistically, my null hypothesis is that soil moisture, slope, stand age, and dominant tree species will all have a coefficients equal to zero, and my alternative hypothesis is that they will all have coefficients greater or less than zero.

Soil moisture was measured using six moisture classes that range from xeric (dry) to sub-hygric (wet) (Johnstone *et al.* 2008). Stand age was calculated using basal tree disks and cores collected from the sites (Walker *et al.* 2018a). Tree dominance was determined by the species that comprised at least 50% of the canopy, if neither jack pine or black spruce dominated the stand then the dominant species was listed as other. Slope was measured at each site using a clinometer (Walker *et al.* 2018b). Burn depth was calculated by using the adventitious root measurements from each site and the appropriate offset based on the tree species measured (Walker *et al.* 2018a).

# **Summary of statistical approaches**

Model selection was used to determine which predictor variables should be included in the model in order to create the best model possible. The minimum adequate model can be determine by several different metrics but I chose to use AIC scores to rank the models. AIC, also known as the Akaike information criterion, is an estimate of the relative amount of information that is lost by a given model. The score accounts for trade-offs between the goodness of fit of the model and the simplicity of the model. Using multiple linear regression, the full model was made using all four of the predictor variables, soil moisture, slope, stand age, and tree dominance, with burn depth as the response variable. I used stepwise backwards selection to remove variables based on the AIC scores of the model. The model with the lowest AIC score was used as the minimum adequate model for this analysis.

K-fold cross validation was conducted in order to determine how accurately the best model can predict burn depth. Cross validation breaks the data into two datasets, the training dataset and the testing dataset. The training dataset is used to create the model, while the model is tested against the testing dataset. By testing the model against data that was not used to create the model, problems like overfitting and selection bias can be avoided. Additionally, cross validation provides insight into how generalized the model will be towards independent datasets. K-fold is a type of cross validation, where the data is randomly partitioned into k equal sized subsamples. One of the subsamples is used as the test dataset, while the other subsamples are grouped as the training dataset. This process is then repeated until each subsample has been used as the testing dataset. The results are then averaged to produce one estimate.

Model selection was conducted before the k-fold cross validation in order to determine the minimum adequate model. While both approaches take overfitting and model simplicity into consideration, they provide different statistical outputs. Model selection is used to determine the best predictor variables to include to explain as much variation in the response variable as possible. While k-fold cross validation assesses how well the selected model can be applied to other independent datasets. The two statistical approaches are therefore complimentary to each other.

# **Results**

Assumptions for a multiple linear regression are linearity, homoscedasticity, independence, and normality. The residual versus fitted values were plotted to check the assumptions of linearity and homoscedasticity (Figure 1a). These assumptions are met based on there being no obvious pattern and the residuals are equally spread around zero. The assumption of homoscedasticity is further met based on there being no obvious pattern in the scale-location plot (Figure 1c). Based on the QQ-plot the assumption of normality is met (Figure 1b), however the results from the Shapiro-Wilks test found that the residuals from the model are significantly different from a normal distribution (p=3.39e-5). The data was transformed using the bestNormalize package, however the Shapiro-Wilks test was still significant (p=3.39e-5) (Peterson 2019). The Cook’s distance versus Leverage plot indicates if there are any influential observations (Figure 1d), while observation 101 has a larger Cook’s distance than the other observations it does not stand out on the other plots and is below a value of two. Therefore, it has been left in the data set. The assumption of independence is met based on the sampling procedure that was used to collect the data. The variance inflation factor detected no correlation between predictor variables, except for tree dominance, where sites dominated by black spruce or jack pine where moderately correlated.

The results from the model selection found that the minimum adequate model included all four of the predictor variables (Figure 2). Xeric plots had significantly different burn depth than the other moisture categories (p=4.8e-5). Burn depth was found to increase with increases in slope and stand age (p=0.17, p=2.7e-12). Burn depth at stands dominated by jack pine was significantly lower than black spruce or other tree species (p=0.00029). The 10-fold cross validation had a residual mean square error of 43.6, an R2 of 0.537, and a mean absolute error of 35.7.

# **Discussion of results**

The model selection found the minimum adequate model to include all of the predictor variables. That model was then validated using 10-fold cross validation, in order to determine how applicable the model will be to other independent datasets. These two statistical approaches are complimentary to one another. The benefit of doing model selection before cross validation is that it ensures that the minimum adequate model is being validated. The main difference between these two statistical approaches is that model selection uses the whole dataset, while cross validation uses subsets of the dataset. Additionally, model selection uses the AIC scores to compare the different models, while cross validation uses the mean skill scores from all of the models.

# **Conclusion**

In conclusion, these two statistical approaches can be used in conjunction with one another to create a minimum adequate model and ensure that it does not over-fit the data. It is important when making predictive models that they are applicable to other independent datasets. The results from this analysis show that soil moisture, slope, stand age, and dominant tree species can all be used as variables to predict depth of organic soil burned during wildfire.

# **References**

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., *et al.* (2013). *Long-term climate change: Projections, commitments and irreversibility*. *Clim. Chang. 2013 Phys. Sci. Basis Work. Gr. I Contrib. to Fifth Assess. Rep. Intergov. Panel Clim. Chang.* Cambridge, United Kingdom and New York, NY, USA.

GNWT. (2019). *2030 NWT climate change strategic framework*.

Johnstone, J.F., Hollingsworth, T.N. & Chapin III, F.S. (2008). A Key for Predicting Postfire Successional Trajectories in Black Spruce Stands of Interior Alaska. *Gen. Tech. Report- Pacific Northwest Res. Station. USDA For. Serv.*, 1–44.

Ryan A. Peterson (2019). Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. Journal of Applied Statistics, 1-16.

Walker, X.J., Baltzer, J.L., Cumming, S.G., Day, N.J., Johnstone, J.F., Rogers, B.M., *et al.* (2018a). Soil organic layer combustion in boreal black spruce and jack pine stands of the Northwest Territories, Canada. *Int. J. Wildl. Fire*, 27, 125–134.

Walker, X.J., Rogers, B.M., Baltzer, J.L., Cumming, S.G., Day, N.J., Goetz, S.J., *et al.* (2018b). Cross-scale controls on carbon emissions from boreal forest megafires. *Glob. Chang. Biol.*, 24, 4251–4265.

**Tables and Figures**

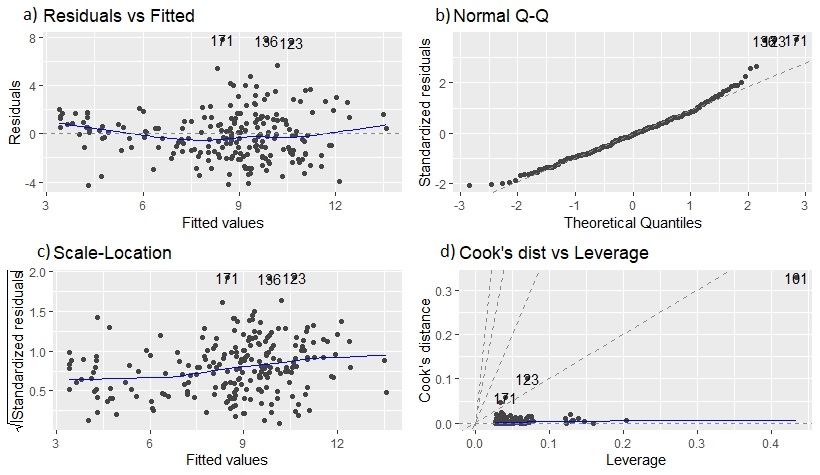


Figure 1: Plot of assumptions for multiple linear regression. A) residuals verses fitted values plot, checking linearity and homoscedasticity assumptions. B) normal QQ-plot, checking assumption of normality. C) Scale-location plot, checking assumption of homoscedasticity. D) Cook’s distance verses leverage plot, checking for outliers in the data.

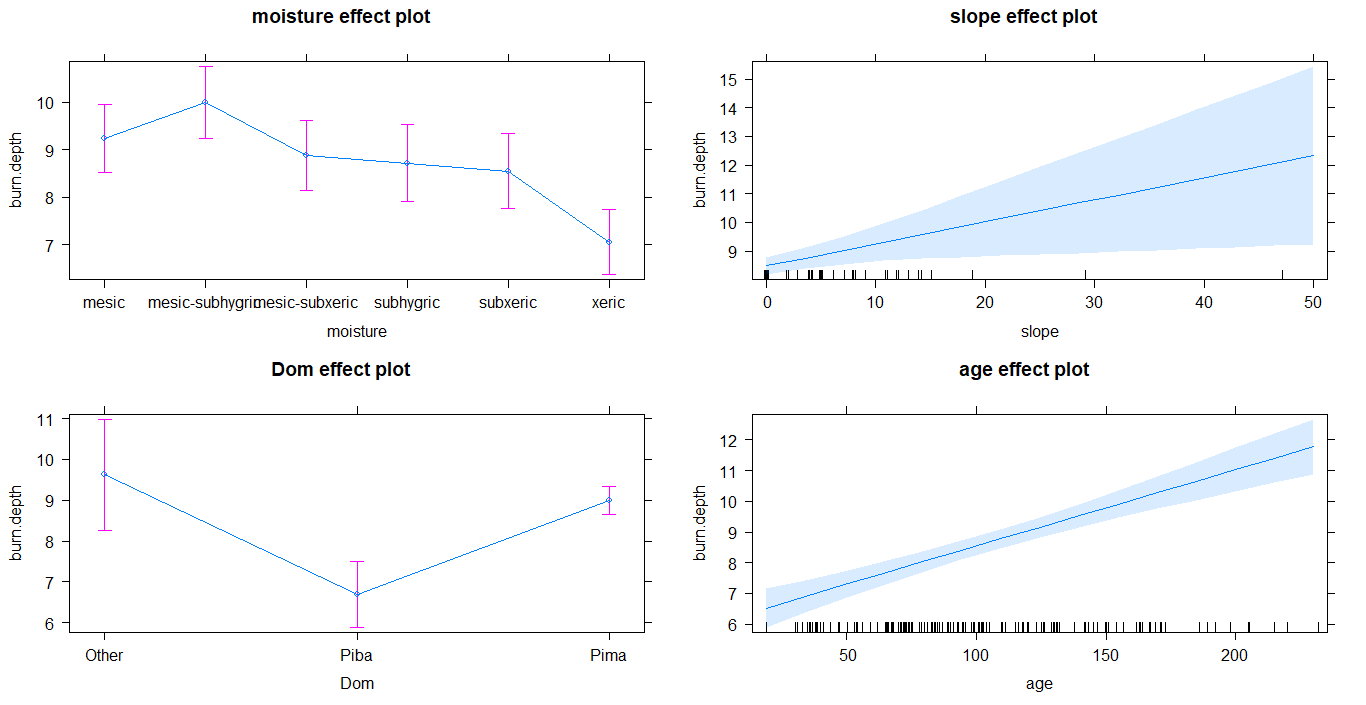


Figure 2: Plot of effects of each predictor variable on burn depth. The four predictor variables are soil moisture, slope, dominant tree species, and stand age. For dominant tree species Piba represents sites dominated by jack pine, and Pima represents sites dominated by black spruce.