Resilient Forests Research Programme

SCION FORESTS PRODUCTS INNOVATION

Date: Dec 2020
Reference: RFP-TN

Technical Note

Understanding spatial drivers for Red Needle Cast: An East Coast pilot study

Summary:

This tech note describes what topological factors potentially influence RNC behaviours and how. We use both a machine learning and statistical approach to carry out this analysis to provide confidence in our observed findings. In this preliminary study, we focus on the East Coast, where satellite images of known RNC infection sites were available and data had already been processed. Initial observations derived from this study indicate that the following topological variables associated with each plot play a significant role in detecting groups of RNC infected trees. The variables are distance of a plot from the forestry boundary, gradient of slope, terrain elevation of each plot. For example, observations from the study demonstrated plots with low slope gradient, of around 400m elevation, close to forest boundary and that are east-ward facing tend to be more likely to exhibit the presence of RNC. However, due to the scope and limitations in the dataset used, these observations need to be validated with further, more in-depth studies. This initial study will be used to inform future trial designs for further understanding how topological factors drive RNC infection and development of RNC management plans.

Author/s: Alan Yu Shyang Tan, Chanatda Somchit, Ellen Mae Leonardo, Grant Pearse, Stuart Fraser

Introduction

Red needle cast (RNC), caused by *Phytophthora pluvialis,* is a foliar disease of *Pinus radiata* and *Douglas-fir*. Sporadic field expression has resulted in difficulties in setting up trials for studying and understanding the disease. Developing a good understanding of the disease is crucial for predicting outbreaks and developing control measures.

Past research on understanding and modelling RNC, carried out under the *Healthy Trees Healthy Future* (HTHF) and *Resilient Forests* programmes, have led to initial understanding on the behaviour of RNC and external factors driving the disease. An RNC Susceptible-Infectious (SI) model (Wake, Williams, & Pleasants, 2018; Wake & Zaidi, 2017) has provided a theoretical basis to model and quantify the dynamics of RNC at a single needle level. For example, Gomez-Gallego et al. (Gomez-Gallego, Gommers, Bader, & Williams, 2019) extended the SI model and identified the need to differentiate needle death from pathogen death to help understand reproduction of pathogen and epidemic development. This theoretical basis allowed us to breakdown the behaviour of RNC into various aspects for focused studies, as shown in Figure 1.

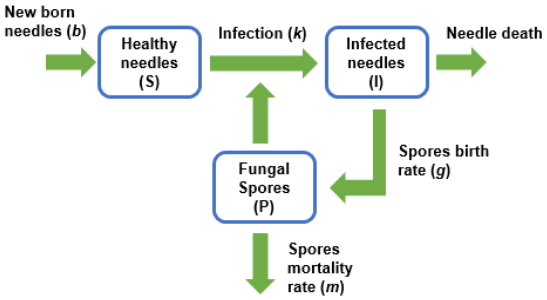


Figure 1. RNC infection process at needle level, as modelled by the RNC SI model.

Observations from studies of RNC in the field carried out in the Pacific Northwest United States and in New Zealand by Gomez-Gallego et al. (Gómez-Gallego et al., 2019) have showed climate factors such as relative humidity in winter played a significant role in explaining the variation in the relative abundance of *Phytophthora pluvialis* in sampled foliage of Douglas fir. Similarly, detection of inoculum of *Phytophthora pluvialis* is greater under cooler and wetter conditions (Fraser et al., 2020). Previous Resilient Forest studies (RA 3.2.3) have also shown climate factors such as rainfall and maximum temperature were significant in areas observed with RNC.

Although we now know climate factors play an important role in influencing RNC expression, little systematic research has been done to understand whether topological factors are important in

FOREST GROWERS RESEARCH

supported by forestgrowers commodity levy

influencing RNC. We had previously attempted to model RNC observations gathered from the Forest Health Database and long-term site monitoring (Hood, Gardner, & Wright, 2017) against site climate and topological factors, using a machine learning approach. Although the machine learning models alluded to climate factors such as rainfall and relative humidity as important factors, topological factors such as elevation were ranked quite highly too. In this study, we will investigate further which topological factors that are associated with RNC expression and identify factors that can inform further studies on how climate and topology drive RNC behaviours.

Data Preparation

Satellite image pre-processing

We focused our study on areas identified with RNC in RA 3.1.1 (Leonardo, Pearse, Fraser, & Estarija, 2020), using satellite images and machine learning. Three sites along the East Coast area of North Island, New Zealand, were selected, namely the Waimata Valley Road, Wharerata and Tauwhareparae sites as shown in Figures 2 and 3. As the regions were obtained from satellite images, no information on the stocking and age class of the forests were known.

High resolution satellite images, obtained from the Worldview-3, Worldview-2 and Geoeye-1 satellite products were pre-processed and annotated. Following the workflow used in RA 3.1.1. (Leonardo et al., 2020), pixel-wise spectral indices Normalised Difference Vegetation Index (NVDI), Green Normalised Difference Vegetation Index (GNVDI), Red-edge Normalised Difference Vegetation Index (RENVDI) and Enhanced Vegetation Index (EVI) were computed using spectral band information for each image. The spectral indices, along with reflectance information and panchromatic band data were used to cluster pixels into regions using an object-based classifier. Regions were then classified into three approximate classes; non-vegetation (i.e. road, plains, mountain terrains, etc), healthy radiata pine trees or unhealthy radiata pine trees (e.g. infected by RNC), using a rule-based classification method. Note that due to the approximation, the radiata pine classification classes may capture some native species and weeds. Non-vegetation classified regions were then removed from the dataset, essentially leaving only healthy and unhealthy forested radiata pine trees regions. Details of the pre-processing, object segmentation and rule-based classification of regions can be found in the technote submitted under the RA 3.1.1. milestone (Leonardo et al., 2020).

The classified regions were then annotated and used to train a random forest model. The aim of the random forest model is to improve the classification of detected forest areas into RNC infected and non-infected classes. The random forest model implementation in python scikit-learn v0.21.1 was used for the classification. Finally, to ensure that we are looking at RNC infected regions, we manually selected and verified 1716 infected regions and 1716 non-infected regions and visually inspected the images and cross checked regions with records of known infected areas obtained from field observation.

Augmenting topological variables

Geographic Information System (GIS) shapefiles for the 3432 healthy and unhealthy regions were generated and used for extracting topological information from various Digital Elevation Models (DEM) of New Zealand. Separate nationwide DEMs, capturing three topological variables, “Aspect”, “Slope” and “Elevation”, each of 25 meters resolution were used. For each region, the minimum, median, maximum and centroid of region for each topological variable were extracted.

Aspect is defined as the direction the terrain feature in question is facing with respect to the compass north, in degrees. In our analysis, this is an issue as the true representation of the directions are not semantically represented in the numerical representation used by the variable. For example, a region having 350 degrees is essentially similar with a region having 10 degrees aspect value – both are north-ward facing. To remedy this issue, we discretise aspect into a categorical variable with the categories definition as shown in Table 1.

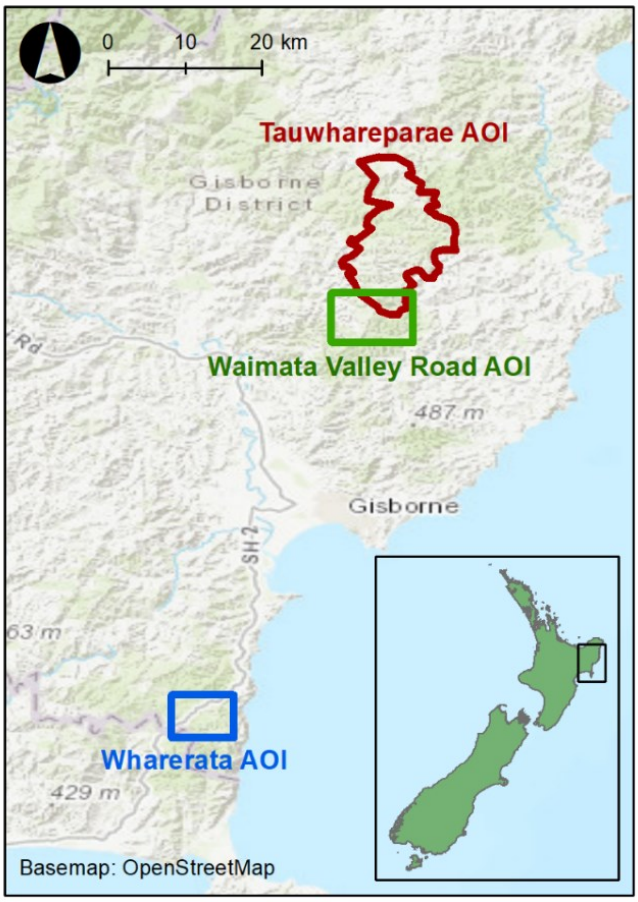
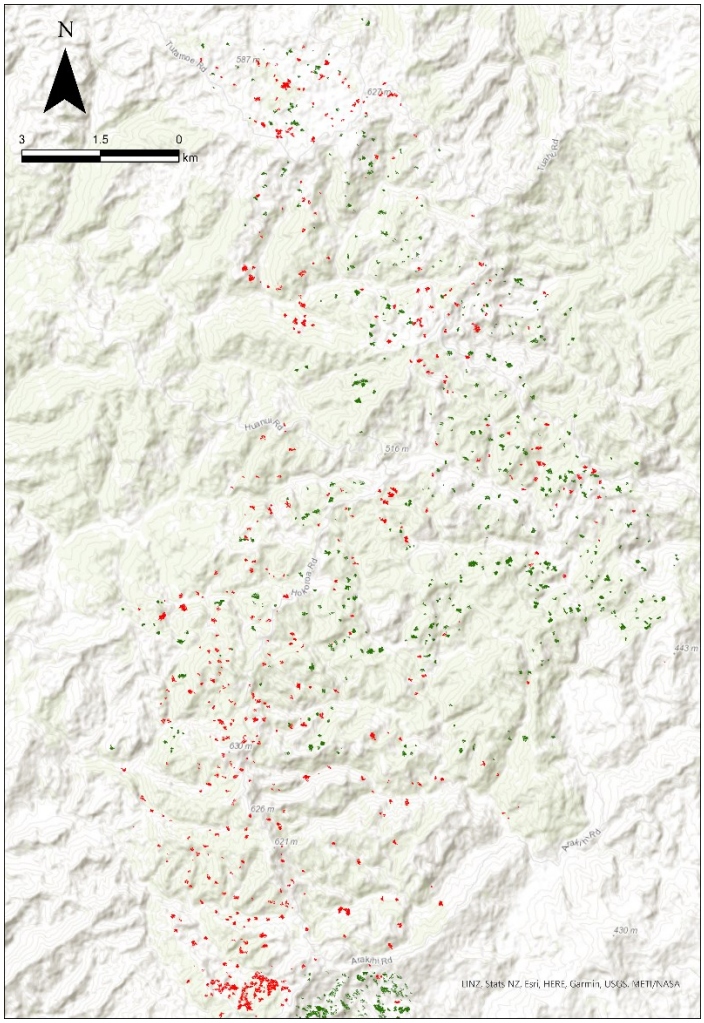
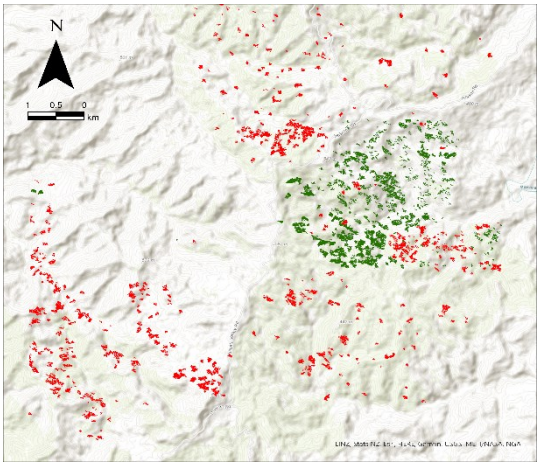


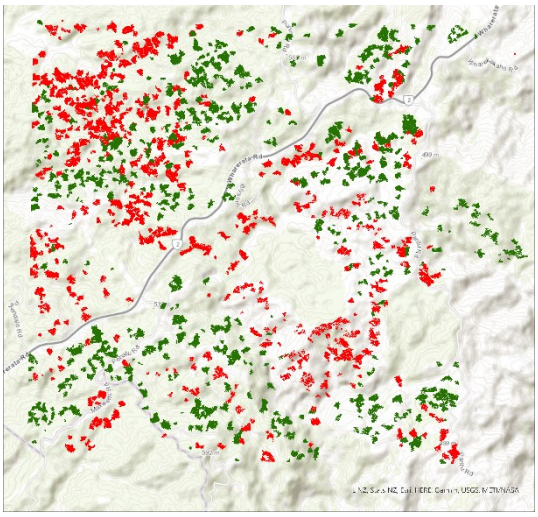
Figure 2. Location of the three selected sites, along the East Coast of North Island, New Zealand, for this study.



c) Tauwhareparae site



a) Waimata Valley Road



b) Whararata site

Figure 3. Predictions made by the random forest model on the satellite images for each of the three sites. Red regions represent unhealthy radiata pine and green regions represent healthy radiata pine.

Table 1. Category definition for Aspect variable

Category	Numerical range
North	315 – 45 degrees
East	46 – 135 degrees
South	136 – 225 degrees
West	226 – 314 degrees

In addition to “Aspect”, “Elevation” and “Slope”, we also computed a derived variable to determine how far the region is from the edge of a forest, “Forest edge distance”. This is calculated by first finding the forest in which the region is contained within using a separate shapefile that captures the boundaries of forest areas in the target sites. The distance of the centroid of the region to the boundary of the forest is then computed. This distance is captured in metres. Table 2 shows the list of variables used for modelling in this study.

Modelling Approach

Machine learning-based modelling

The objective of using machine learning to model the data is to investigate which variables are of interest from a non-parametric approach. Gradient Boosting Machines (GBM) (Friedman, 2002; Mason, Baxter, Bartlett, & Frean, 1999) have been well-known to handle and model structured data well. Given categorical variables are present in our dataset (e.g. aspect, healthy/unhealthy forest), we choose to use CatBoost (Dorogush, Ershov, & Gulin, 2018), an implementation of GBM that can handle categorical variables for prediction. The python implementation of CatBoost was used.

The “Aspect” and its variations and “Disease incidence” were set as categorical variables in CatBoost during training and analysis. A logistic regression function was used as the model loss function due to our predictor, “Disease incidence”, being a categorical variable. One hot encoding, with maximum size set to six, was used to encode categorical variables.

Table 2. List of variables used in study

Variable Name	Variable Type	Variable Definition
Slope	Numerical	Gradient of slope of the centroid of the region, in degrees
Slope min	Numerical	Min gradient of slope for the region, in degrees
Slope median	Numerical	Median gradient of slope for the region, in degrees
Slope max	Numerical	Max gradient of slope for the region, in degrees
Elevation	Numerical	Elevation of the centroid of the region, in metres
Elevation min	Numerical	Min elevation for the region, in metres
Elevation median	Numerical	Median elevation for the region, in metres
Elevation max	Numerical	Max elevation for the region, in metres
Aspect	Categorical	Direction the terrain feature at the centroid of the region is facing
Aspect min	Categorical	Direction the terrain feature, on average, the region is facing
Aspect median	Categorical	Direction the terrain feature, on average, the region is facing
Aspect max	Categorical	Direction the terrain feature, on average, the region is facing
Forest edge distance	Numerical	Distance of centroid of region from forest boundary
Disease incidence	Categorical	Healthy or RNC infected

To fine-tune the model parameters, a grid-search approach, with 5-fold cross-validation was used for evaluation. Four parameters of the CatBoost model, *depth, learning rate, L2 leaf regression* and *iterations* were tuned. The parameters are defined as follow:

- **Depth** – maximum depth allowed for the trees being built. This parameter influences how easily the model can be overfitted.
- **Learning rate** – Parameter used to control the step size to use for internal model parameter tuning. Influence how quickly the model converges.
- **L2 leaf regression** – Parameter for setting the coefficient in the L2 regularisation term of the loss function.
- **Iterations** – Maximum number of iterations to run the model. Terminate too early may result in model not being able to converge.

Using the grid-search approach, the parameter set (depth=6, learning rate=0.15, l2 leaf regression=9, iterations=100) is used.

Statistical modelling

A statistical modelling approach was used to verify observations derived from the machine learning model and to examine the contribution of each variable to positive RNC detection in depth. The statistical modelling was carried out independently from the machine learning model to avoid bias.

Generalized additive models (GAMs) were used to analyse the effect of topological variables on the probability of areas being affected by RNC. *Disease incidence* was assessed based on a binary scale (0 and 1) where areas where RNC were not found were classed as non-infected areas (0), and areas on which RNC found were classed infected areas (1). *Disease incidence* using binary categories was analysed using a GAM with a binomial distribution and logit link function. The explanatory variables considered in the analysis were 'Forest edge distance', 'Slope', 'Slope min', 'Slope median', 'Slope max', 'Elevation', 'Elevation min', Elevation median', Elevation max', 'Aspect', 'Aspect min', 'Aspect median' and 'Aspect max'.

We used an iterative variable selection method to determine which topological variable is of importance to modelling unhealthy regions. GAMs were fitted by penalized likelihood maximization implemented in the *mgcv* package (Wood & Wood, 2015).

To determine which variable to add to the model, a single smooth binomial GAM model was run for each topological variable first. The most significant variable which had the smallest P-value with the lowest Akaike Information Criterion (AIC) obtained from the first step was chosen and added to the model. Then the most significant variable was screened for collinearity using Pearson correlations. The Pearson correlations between the topological and the most significant variables were then calculated. Any pairs of the variables with a Pearson's $r > 0.5$ were considered collinear. The topological variables correlated to the most significant variable were then removed from the model. All steps described above were repeated by adding the most significant variable to the model, screening for collinearity and dropping the correlated variables until reaching the stopping rule (i.e. none of the remaining variables were significant at $\alpha = 0.05$).

For model validation, plots of Pearson residuals against the fitted values and each explanatory variable and plots of ordered deviance residuals against their theoretical quantiles were used. The significant factor terms were followed up by applying a multiple-comparison procedure using Tukey-adjusted contrast. We made predictions using all cases in the data to evaluate the predictive power of the proposed model.

Table 3. Summary of topological variable values for the 3432 samples

Variables	Min	Median	Mean	SD	Max
Forest edge distance	0.00	28.64	39.36	35.04	264.85
Slope	0.09	17.61	18.00	8.17	43.35
Slope min	0.01	9.46	10.11	6.89	34.95
Slope median	0.45	17.62	17.96	7.06	39.96
Slope max	1.44	24.58	24.43	7.06	49.33
Elevation	119.68	404.45	390.11	87.79	620.32
Elevation min	119.52	394.03	375.80	89.04	609.68
Elevation median	120.01	403.98	389.91	87.73	616.86
Elevation max	126.43	419.45	403.94	86.12	620.62

Table 4. GAM results for the probability of areas infected by RNC based on disease incidence data.

Parametric coefficients	***Estimate***	***SE***	***z***	***P***
Intercept	-0.23	0.06	-3.81	<0.001 ***
Aspect min: East	0.72	0.10	7.57	<0.001 ***
Aspect min: South	0.16	0.10	1.67	0.09
Aspect min: West	0.03	0.12	0.24	0.81

Approx. significance of smooth terms	**edf**	***Chi-sq***	***P***
s(Elevation min)	8.06	115.39	<0.001 ***
s(Forest edge distance)	4.75	83.39	<0.001 ***
s(Slope min)	3.52	51.78	<0.001 ***
The predictive accuracy	64%		

s( ) = smooth term for a continuous variable, *SE* = standard error of the estimate, *z* = *z*-statistic, *P* = *P*-value, *edf* = estimated degrees of freedom, *Chi-sq* = *Chi Square*-statistic. Significant values are denoted with P <0.05 = *, P <0.01 = **, P <0.001 = ***. edf values >1 indicate a non-linear effect. 'Aspect min: North' was included as baseline in the model.

Summary of topological variable values for the 3,432 samples is shown in Table 3. Frequency plots of verified infected and non-infected areas, by topological variables are given in Figure 4.

Model Observations

Based on the trained CatBoost model, a variable ranking graph was generated using the SHAP evaluation toolkit. The variable ranking graph is shown in Figure 5. From the figure, we observed that "Slope min", "Forest edge distance", "Aspect min" and "Elevation min" stood out, in order of importance.

In the areas investigated, there was greater detection of RNC in flatter sites, or in regions closer to the forest edge or more elevated areas. We observed that "Slope min" and "Forest edge distance" resulted in positive RNC detections as those variables approached the lower value ranges. Likewise, positive RNC detection associates more with lower "Aspect min" readings, pointing towards North and East orientated slopes tend to more likely present RNC detection. The inverse applies for "Elevation min" with positive RNC detection increasing with elevation.

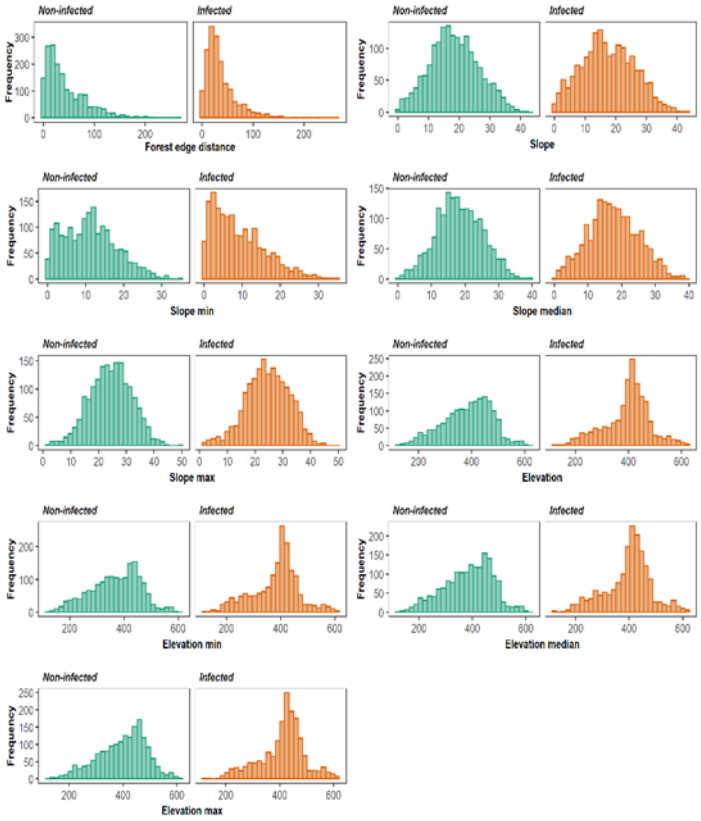


Figure 4. Frequency plots of verified infected and non-infected areas, by topological variables.

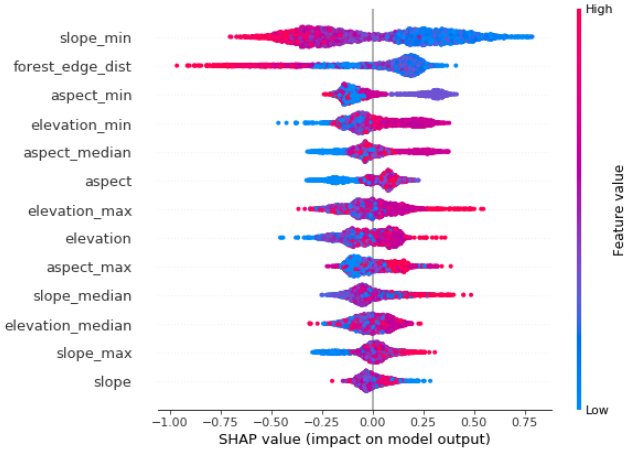


Figure 5. SHAP variable ranking graph. Variables are ranked from most impactful to least impactful on model prediction (based on contribution of variable towards splitting of tree in CatBoost). The x-axis represents the prediction of the model, with negative being absence of RNC and positive being unhealthy and demonstrates the association of value range for each variable to the prediction.

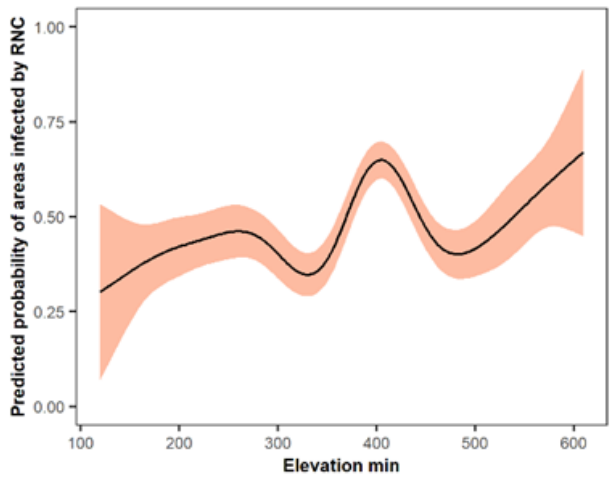


Figure 6. Predicted probability of areas infected by RNC with 95% confidence intervals by 'Elevation min' using the binomial GAMs. Elevation are in metres.

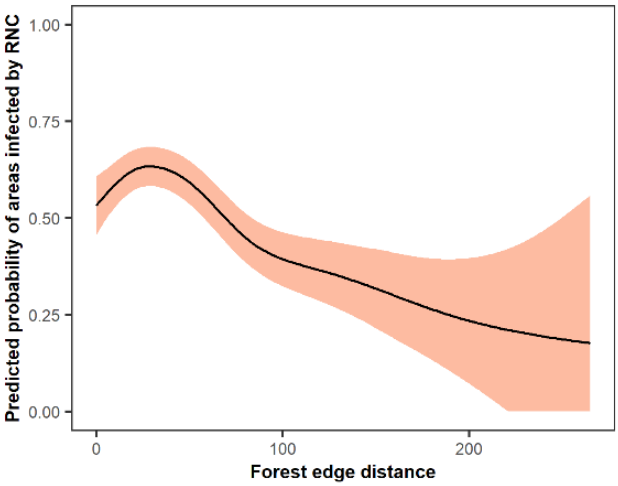


Figure 7. Predicted probability of areas infected by RNC with 95% confidence intervals by 'Forest edge distance' using the binomial GAMs. Distance are in metres

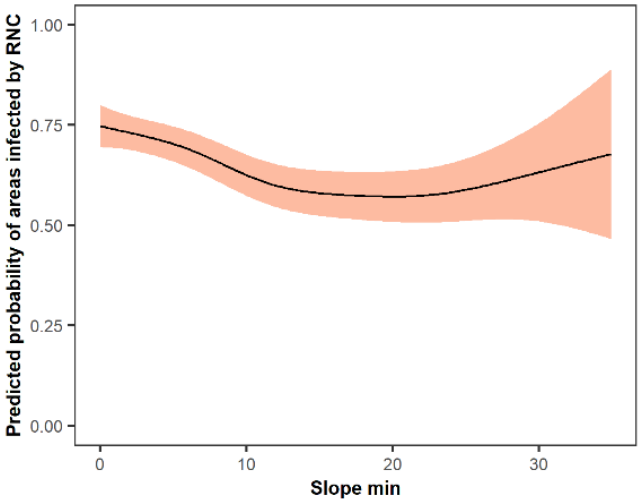


Figure 8. Predicted probability of areas infected by RNC with 95% confidence intervals by 'Slope min' using the binomial GAMs. Slope gradient are in degrees.

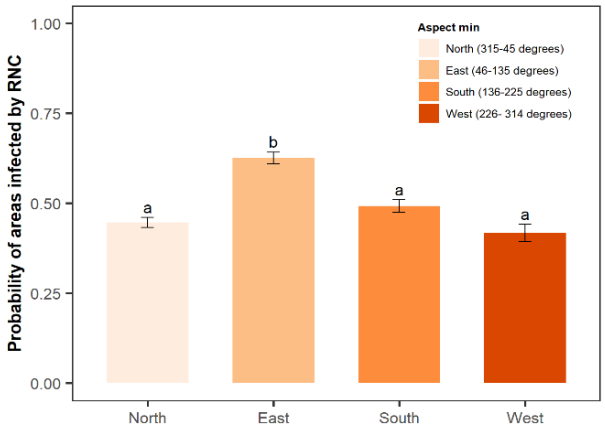


Figure 9. Mean (±SE) areas infected by RNC by 'Aspect min'. Bars followed by different letters indicate statistically significant between 'Aspect min' (multiple comparison tests using Tukey – adjusted contrasts at α = 0.05).

A GAM with a binomial distribution best described the relationship between the probability of areas observed with RNC expression and topological variables. There was a significant effect of the four topological variables: 'Aspect min', 'Elevation min', 'Forest edge distance' and 'Slope min', on the probability of RNC expression (P<0.001; Table 4).

The fitted function for 'Elevation min' indicated that the probability of RNC expression peaked at the 'Elevation min' of about 400m (Figure 6). The probability of RNC expression increased steeply when the sample areas were very close to the forest edge and reached a maximum level when the sample areas was about 30 m away from the forest edge before decreasing sharply (Figure 7). The fitted function indicated that the probability of RNC expression peaked at the lower slope min and decreased with increasing "Slope min", and then was roughly horizontal between 10 – 30 degrees (Figure 8). The probability of RNC expression was significantly greater in areas with an eastern aspect (P<0.05; Figure 9). No statistical differences were found between the North, South and West areas (P>0.05; Figure 9). The predictive accuracy of the proposed model is about 64%.

Concluding Remarks

Observations derived from the independent machine learning and statistical modelling agreed. Both methods identified the factors of gradient of the slope, distance to edge of forest, elevation and aspect which all influenced the occurrence of RNC. More specifically, forested regions which have the following properties:

- Elevation of approximately 400m
- Slope gradient of less than 10 degrees
- Being less than 30m away from edge of a forest

While the 'Aspect min' variable had shown that regions 'East' facing seem to witness more RNC infected regions, this observation should be taken interpreted

with care. This is so because the region is bias to the East Coast. As such, it is not certain whether the significance is caused by being east facing or because of the slope being seaward facing.

Similarly, summary statistics listed in Table 3 shown that Elevation and its derivatives (min, max and median) have similar distributions. While this is unusual, we attribute this to the approach used for clustering pixels into regions prior to classifying regions into non-vegetation, healthy and unhealthy regions. Due to the use of reflectance and spectral indices for clustering, there is a large chance the algorithm inherently clustered regions with similar elevation together. In saying that, this still does not discount the observation that forested regions at 400m tends to exhibit RNC infections.

The observations from both the machine learning and statistical models are in agreement and provide us with the confidence to design future experiments to investigate the influence those four variables. We recommend that the spatial analysis be expanded to more diverse regions of New Zealand and increase the number of random samples of both healthy and unhealthy radiata pine forested regions from a range of stand ages and silvicultural regimes.

Acknowledgement

Funding for this research came from Scion and the Forest Growers Levy Trust, with the support of the NZ Forest Owners Association (FOA) and the NZ Farm Forestry Association (FFA).

References

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. Retrieved from https://arxiv.org/abs/1810.11363 doi:1810.11363

Fraser, S., Gomez-Gallego, M., Gardner, J., Bulman, L. S., Denman, S., & Williams, N. M. (2020). Impact of weather variables and season on sporulation of Phytophthora pluvialis and Phytophthora kernoviae. *Forest Pathology, 50*(2), e12588. doi:10.1111/efp.12588

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367–378. doi:10.1016/s0167-9473(01)00065-2

Gomez-Gallego, M., Gommers, R., Bader, M. K.-F., & Williams, N. M. (2019). Modelling the key drivers of an aerial Phytophthora foliar disease epidemic, from the needles to the whole plant. *PLOS ONE, 14*(5), e0216161. doi:10.1371/journal.pone.0216161

Gómez-Gallego, M., LeBoldus, J. M., Bader, M. K.-F., Hansen, E., Donaldson, L., & Williams, N. M. (2019). Contrasting the Pathogen Loads in Co-Existing Populations of Phytophthora pluvialis and Nothophaeocryptopus gaeumannii in Douglas Fir Plantations in New Zealand and the Pacific Northwest United States. *Phytopathology®, 109*(11), 1908-1921. doi:10.1094/phyto-12-18-0479-r

Hood, I., Gardner, J., & Wright, L. (2017). *Longer term monitoring of red needle cast: Plot establishment and initial data collection* (59277). Retrieved from https://scion.elsevierpure.com/admin/files/17625437/59277_RNC_long_term_monitoring_initial_report_NXPowerLite_.pdf

Leonardo, E. M., Pearse, G., Fraser, S., & Estarija, H. J. (2020). *Red Needle Cast monitoring framework using high resolution satellite imagery*. Retrieved from

Mason, L., Baxter, J., Bartlett, P., & Frean, M. (1999). *Boosting algorithms as gradient descent*. Paper presented at the Proceedings of the 12th International Conference on Neural Information Processing Systems, Denver, CO.

Wake, G., Williams, N., & Pleasants, T. (2018). A dynamical systems model for poly-cyclic foliar forest pathogens. *2018, 59*, 14. doi:10.21914/anziamj.v59i0.12625

Wake, G., & Zaidi, F. (2017). *A proposed systems biology model to determine the spread of red needle cast in pine trees and its control: Formulation and preliminary analysis* (Contract Reference J02006). Retrieved from

Wood, S., & Wood, M. S. (2015). Package 'mgcv'. *R package version, 1*, 29.