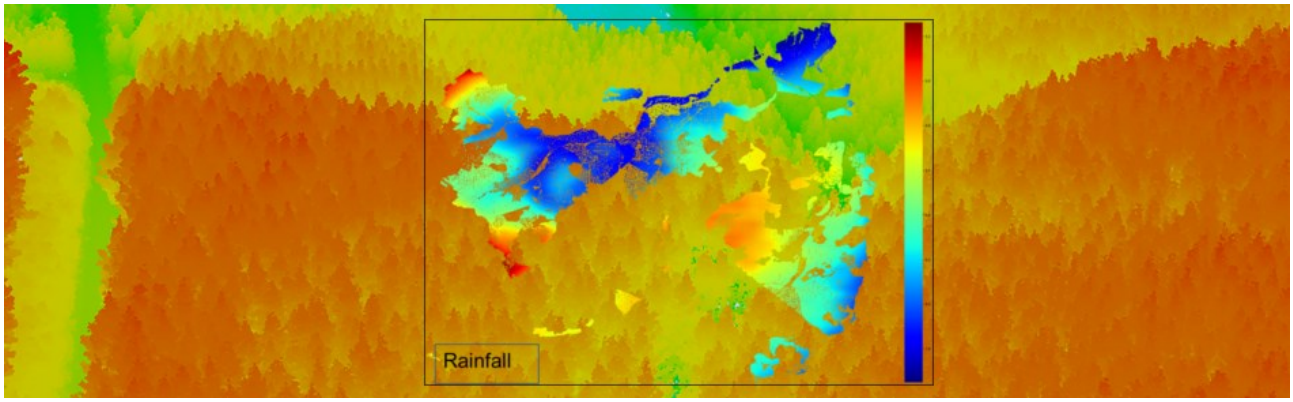
Resilient Forests Research Programme

SCION FORESTS ▪ PRODUCTS ▪ INNOVATION

Predicting Estate Productivity using Area-Based Phenotyping and Machine Learning

David Pont, Sean Husheer, Simon Papps, Maxime Bombrun, Heidi Dungey



Date: **19 June 2020**

Report No: RFP-T007

FOREST GROWERS RESEARCH

TABLE OF CONTENTS

supported by forestgrowers commodity levy

EXECUTIVE SUMMARY

There is real potential for machine learning methods to be used to evaluate and improve forest productivity over forest estates. Forest managers implement these operationally using the data sources available to them and processes described here. The potential benefits include matching genotypes to sites, and guiding management, to improve productivity. The results can also provide insights into environmental and management factors affecting productivity. The important role of stocking is seen as a tangible opportunity to improve productivity by optimally managing site occupancy.

This study successfully developed new methods for predicting and mapping productivity across a forest estate. This required development of area-based phenotyping and machine learning approaches applied to LiDAR data. Prior research into area-based phenotyping research was carried out within the GCFF programme using LiDAR from the Kaingaroa Timberlands estate. Here, we have worked closely with Hancock Forest Management NZ Ltd and used data from several of their forests to further developed the methods have been and apply them to a second forest estate.

This study has met the project objectives and successfully developed practical methods to:

- Compile forest environmental, genetic, management, and productivity data
- Fit machine learning models to these data
- Use trained models to predict forest productivity across a range of management scenarios

The positive outcomes of the study support further evaluation of the methods. The correctness of the processes needs to be verified, and the accuracy and robustness of productivity predictions needs to be quantified. It would be desirable to add spatialised measures of uncertainty to model estimates. To aid adoption by end users, the analytical processes need to be further developed, streamlined, packaged and documented. This further work could be done with the current data sets, but additional data could also be beneficial.

INTRODUCTION

Advances in remote sensing combined with the emergence of sophisticated methods for large-scale data analytics provide new methods to model complex interactions in biological systems. Phenotyping is the observable characterisation of an organism as a consequence of multifactorial genetic traits and environmental influences. The process uses the whole forest as an experiment by drawing inferences from very large amounts of data.

Recent research carried out by Scion has developed methods to combine phenotypic data on site productivity with information on management and genetics from stand records (Bombrun et al., 2020; Dungey et al., 2018). Machine learning methods can be applied to these data to predict productivity for different combinations of genetics, environment and management. The benefit of machine learning methods over traditional modelling methods is the ability to fit models to data sets with very large numbers of variables and very large numbers of observations. This offers the potential to maximise productivity or explore alternative scenarios for large existing, and planned, forest plantings.

This report outlines the application of a forest phenotyping system across several plantation forests managed by Hancock Forest Management (NZ) in the central North Island of New Zealand. Spatial estimates of productivity can be correlated to specific site, genetic, environmental, and silvicultural features and provide the opportunity to improve growth and profitability through better management interventions.

The approach used LiDAR to derive area-based estimates of productivity as continuous raster surfaces at 25 by 25 m resolution across the forest estate. These data were merged with environmental raster data and stand records to create a multi-layered raster data set with variables in four major classes:

- Productivity
- Genetics
- Environment
- Management

Individual productivity variables became target (dependant) variables to be estimated from models fitted using machine learning. The numerous genetic (G), environmental (E), and management (M) variables were the inputs (independent variables, referred to as features in machine learning), to G by E by M models. The input data sets were large, having hundreds of thousands of raster cells in each layer. Genetic and management variables from stand records numbered in the tens, while environmental variables were more numerous, resulting in hundreds of layers. The complete data set can be represented as a three-dimensional cube of data, typically having several millions of individual data points.

METHODS

Forest data sets

Forest data sets were provided by Hancock Forest Management NZ Ltd. Initial analyses were carried out using data from the Tiaki Forest as this contained reliable stand records, useful variation in the key variables, and a compact size. Data from the OTPP forest estate were subsequently added to provide a larger data set with a more extensive range of variation. Initial analyses of Tiaki used Site Index (SI) as the target variable, the addition of OTPP forest also saw the addition of the 300 Index (I300) productivity metric.

Forest LiDAR data and productivity layers

Productivity data was provided in a spreadsheet where each record represented a single 25 x 25 m cell and each column contained variables derived from area-based analyses of LiDAR. Productivity measures of interest were SI, mean annual volume increment (MAI) and I300.

Adding environmental layers

All spatial data was processed in the New Zealand Transverse Mercator 2000 projection and grid reference system to create rasters at a 25 by 25 m resolution matching the LiDAR data. Environmental raster extraction, reprojection, resampling and merging steps were carried out in R version 3.8 (*RStudio Team. RStudio: Integrated Development for R*, 2020). Soil data was extracted from the S-map spatial database ("Manaaki Whenua - Landcare Research. S-map - New Zealand's national digital soil map," 2019). Each soil family is defined as a unique combination of attributes (New Zealand Soil Classification, parent material, rock type, dominant texture and permeability class). Monthly climatic data supplied by NIWA in ArcInfo binary format was converted into R rasters over the extent of the forest areas. Annual, seasonal minimum, and maximum climate summaries for each grid point for multiple data sets were extracted using GIS processing in R. GRASS and SAGA GIS were called from R to calculate topographic predictor variables from elevation. The environmental layers were then merged to the base productivity data in a single R compressed binary data base.

Adding genetic and management layers

Management and genetics information were extracted from forest stand records and provided in a spreadsheet where each record represented a single stand, and columns contained genetic and management values. Management variables included information on silviculture, and genetic variables included parental family codes and breeding values where available. These data were extracted for each raster grid cell and merged with the combined productivity and environmental layers created in the prior step.

The combined Tiaki and OTPP forests occupied an area of 49,426 ha and the raster data set comprised 790820 rows (one per input grid cell) and 236 columns (one per input variable or layer), a total of 186,633,520 observations (data points).

Modelling approach

All subsequent processing, from data preparation through to model fitting and analyses, were carried out using Jupyter Notebook version 6.0.3 (Kluyver et al., 2016).

Data preparation

Data preparation comprised three phases: cleaning, feature selection, and row filtering. The first phase was cleaning the data set, checking for missing and outlier values, redundant variables, and verifying the different data sources had been merged correctly.

This phase used numerical summaries and graphical methods such as scatter and density plots, as well as geographical maps. These numerical and graphical methods were essential to efficient screening of the large amounts of data.

In machine language terminology, the independent predictor variables input to machine learning models are referred to as “features”. Selection of features is one of the few influences a modeller has over the fitting of a machine learning model. While machine learning methods can handle large numbers of input features, and even correlated features, the process of feature selection is an important part of model development.

In the second phase, feature selection was carried out to create the final set to be used in model training and fitting. We developed four stages of feature selection:

1. Exclude non-target dependant features
2. Exclude constant and empty features
3. Exclude features identified as unwanted using domain knowledge
4. Exclude highly correlated features

In the third phase, row filtering was applied to exclude input rows having missing values for the dependant variable, or for key genetic, environmental or management variables to be used later as inputs to model prediction simulations.

Model training and fitting

The data set columns were split into the dependant variable (Y) and the independent variables (X) and then rows were split randomly to create training and test data sets (70 and 30% respectively). Modelling used the open source machine learning library CatBoost (version 0.22) to carry out gradient boosting on decision trees (Dorogush, Ershov, & Gulin, 2018).

Model parameter tuning was carried out using Bayesian optimisation, which was shown to be slower than an alternative grid search approach, but to provide better parameter estimates (Bombrun et al., 2020). Parameter tuning used the GPU acceleration option, and the target measure for optimisation was RMSE on the estimated values.

RESULTS AND DISCUSSION

Feature selection

Initial modelling runs used the Tiaki data set, with SI as the target variable, where SI is defined as the average height of the 100 largest diameter trees per hectare at age 20 years. This data set was used to develop the feature selection and modelling approach. Subsequently the OTPP data set was added and I300 added as a target variable. I300 Index is an age and stocking independent measure of stand volume productivity, defined as the stem volume mean annual increment in m^3/ha/year at age 30 years for a defined reference regime of 300 stems/ha.

Feature selection options were extensively explored. In particular, the genetic and management feature sets contained several highly correlated variables. These features, obtained from the forest stand records system and analyses of LiDAR data, were well understood, and not so numerous, permitting a reasoned approach and exclusion of features which were clearly illogical to include. Our investigations ultimately led to the development of three initial stages of feature selection, outlined in Table 1. In contrast, the environmental feature set contained numerous soil and climatic variables, for which the potential effects on productivity were unknown. In response, the fourth stage of feature selection was developed to exclude features which were extremely highly correlated (a threshold of Pearson's r>0.90).

Data filtering, feature extraction results

Table 1. shows the numbers of features excluded and remaining after each stage of feature selection. Stages 3 and 4 removed substantial numbers of features, both manually and automatically selected for removal.

Table 1. Numbers of features excluded and remaining at each stage in feature selection for the Tiaki Forest data.

Stage	Excluded	Remaining
1. Exclude non target dependant features	2	240
2. Exclude constant and empty features	6	234
3. Apply domain knowledge to exclude features	68	166
4. Exclude highly correlated features	91	75

Parameter tuning

Parameter tuning was found to be computationally demanding. For example, the parameter tuning run for SI with the smaller Tiaki data set (75 features and 290916 records) took over 16 hours. Optimised parameters for the target variables SI and I300 are presented in Table 2.

Table 2. Optimised CatBoost parameters with the combined data set for SI and I300 target variables.

Target variable	Parameter	Optimised value
SI	eta	0.2996
	iterations	1840
	max_depth	7.93
I300	eta	0.3424
	iterations	1744
	max_depth	11.71

Model fitting for I300 with the combined data set took 1005 seconds (16 minutes 45 seconds), more than 50 times faster than the parameter tuning process. It was noted that model fitting was faster when the fourth stage of feature selection (see Table 1) was not applied. The larger set of features (166 versus 75) apparently reduced computational effort to build the decision tree.

Model fitting

All subsequent reported model fitting and analysis was carried out using the larger feature set (166 features) created by not excluding highly correlated features, with I300 as the target feature. The model fitted to the training data set was evaluated with the remaining 30% as test data, had an R-squared of 0.7331 and RMSE of 1.6624. The final model to be used for simulations was then fitted to the full data set, having a R-squared of 0.9170 and RMSE of 0.9273.

Features in the decision tree

An example force plot is presented in Figure 1 (at end of report) to illustrate the way input features influence model estimates. Input features are used in the decision tree to provide a positive or negative influence (force) on the output value. In the force plot arrows are coloured red for positive forces and blue for negative, the size of the arrow represents the relative size of the force. Names and force values are given for some of the top features. This plot represents a single row (grid cell) in the data set, having an expected I300 value of 25.5 and a model estimate of 26.11 (shown in bold). We can see SPH, Understory, SoilType were the features with the top three forces for this grid cell. This illustrates how features have positive or negative influences and differing importance values in the decision tree.

Feature importance

A post-process analysis of the fitted CatBoost decision tree was made to rank features according to their mean SHAP value (Molnar, 2019). This is a measure of how influential each input feature is in the overall decision tree (model). This plot therefore provided a useful insight into the relative importance of different features on forest productivity. The plot in Figure 2 shows feature names and SHAP importance values for the top 20 ranked features from the fitted model. Figure 3 presents additional information showing the distributions, and positive or negative influences, for each of these features.

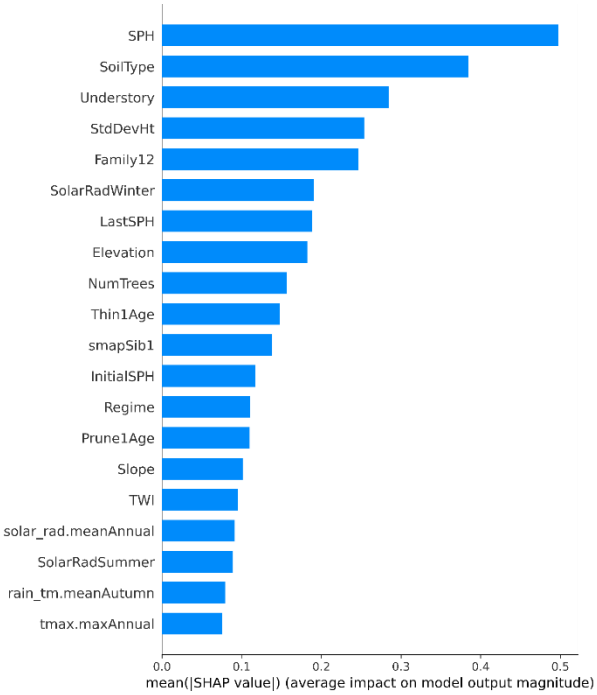


Figure 2. Feature importance values (SHAP) for the top 20 ranked features.

The most significant feature was SPH, a state variable indicating the current stocking in stems/ha based upon the LiDAR imputation. High stocking was associated with both higher and lower I300, and the influence on I300 was mixed for intermediate values. This likely reflected the independence of I300 from stocking, but also the fact that important features can have a complex relationship with I300, because of interactions with other factors. This highlighted the merit of machine learning methods for modelling complex systems that would be intractable with conventional modelling techniques.

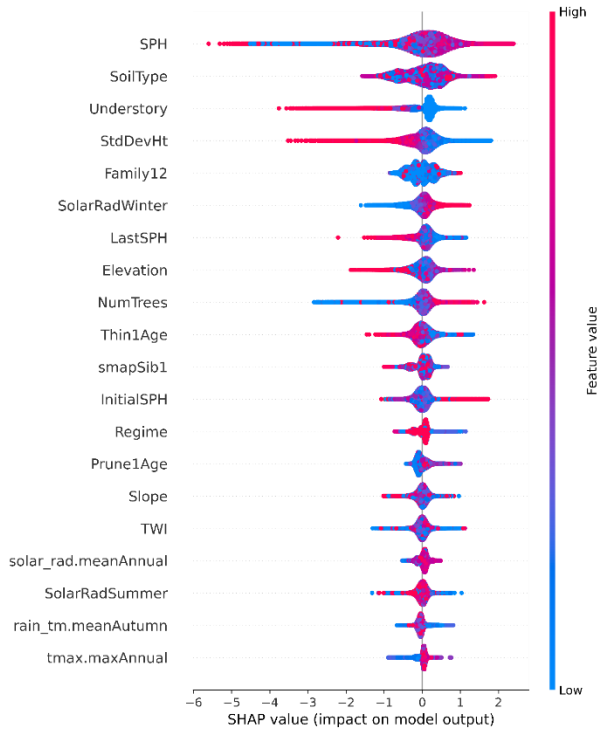


Figure 3. Summary plot (SHAP) for the top 20 ranked features. Colours indicate low (blue) and high (red) feature values. Note that the colours are not meaningful for categorical features such as SoilType and Family12.

Other stocking features were also prominent, LastSPH is the last measured stocking, NumTrees is the LiDAR tree segmentation estimate of stocking, and InitialSPH is the initial stocking at time of planting. One implication is that stocking and competition affect height and volume increment more than is recognised by the 300 Index growth model. Stocking also featured prominently when the target variable was SI, which indicated that height growth is influenced by stocking, whereas the site index metric implicitly assumes that this is not the case.

The second implication is that I300 (volume) is affected by the interaction of multiple features in conjunction with stocking, such as those determined by a principal components analysis. These interrelationships can be teased out from fitting of the CatBoost model and some examples are presented later.

SoilType, a high level categorical classification feature defined in S-map, was the most significant environmental variable. The features Understory and StdDevHt are phenotypic metrics derived from LiDAR, representing the number of distinct identified non-crop vegetation plants and crop height variation respectively. Understory and StdDevHt were found to be inversely related to I300. The Understory feature was consistently ranked highly across several exploratory model runs, with the Tiaki and combined data sets, and for SI and I300 target variables. High values of Understory were found to have a negative influence on model output (Figure 3 with correspondingly low I300). Understory is likely a proxy for unstocked area such as gaps, roads, skids, clear-felled and wind-thrown areas. The CatBoost model results have clearly highlighted the important negative contribution of unstocked areas on forest productivity.

The Family12 feature, representing the families of both parents, ranked fifth. It is interesting to note there are genetic, environmental, and management features in the top five ranked features and that categorical variables dominate rather than soil or genetic variables that take continuous numerical values. This in turn emphasises the merit of using the CatBoost method, specifically designed to permit efficient use of categorical variables in machine learning models.

Feature interrelationships can be explored through partial dependence plots, which show the marginal effect one or two features have on the predicted outcome of the model. For example, Figure 4 shows the effect on SHAP value of slope (degrees on the *x*-axis) and stocking (blue represents low stocking, red is high).

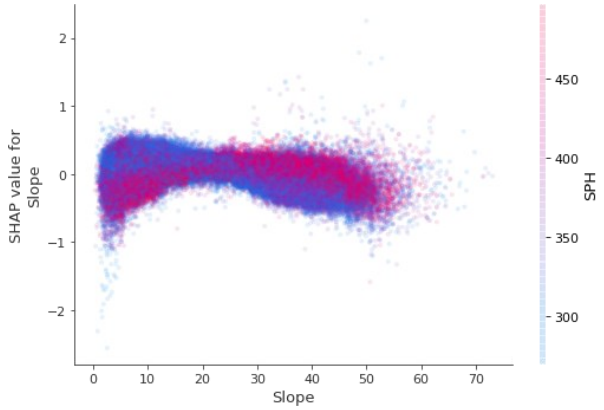


Figure 4. Partial Dependence on SHAP Value of Slope and Stocking

The plot indicated a slight slope was beneficial to growth but decreases thereafter. This is consistent with experience; flat ground is more prone to frosting and foggy conditions conducive to foliar diseases. There was an apparent interrelationship between slope and stocking, where higher stocking was more desirable on steeper slopes, presumably due to greater opportunities for light interception. Steeper slopes tend to be more exposed and subject to wind-throw, which is another reason for maintaining higher stockings.

Model prediction

Figure 5 shows model estimates for I300 plotted against the observed values input to the model. The 1:1 line makes it clear that the model tends to over predict low values and under predict high values of I300. This tendency was also observed for models fitted to SI and to the smaller Tiaki data set (results not shown). This indicated the machine learning model had difficulty accurately representing the full range of observed I300 values and is tending to slightly generalise (predict the mean).

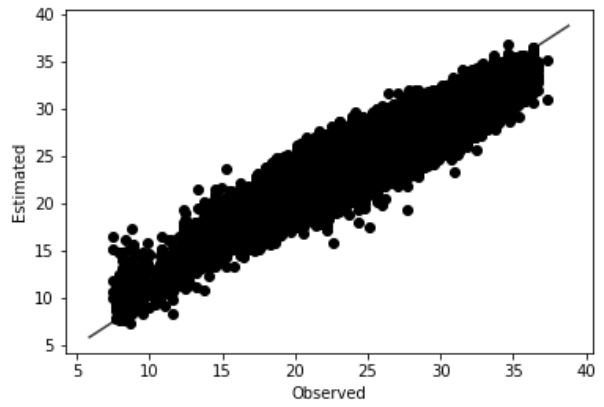


Figure 5. Predicted productivity (I300) plotted against observed values input to the model, with the 1:1 line for reference.

Scenario modelling: estimating productivity under different GxExM

The fitted model was used to predict productivity (I300) under a range of genetic (G), environmental (E), and management (M) options. The top ranked G, E and M features from the model fitting results were Family12, SoilType, and SPH respectively. The data set included pruned and structural regimes, the latter were selected for scenario modelling because they have a single thinning (and pruned regimes do not reflect current management practice). The feature LastSPH represented the final crop stocking after thinning. In exploring management options it was also useful to include the age of thinning, represented by the feature Thin1Age. A set of values were devised to span the observed ranges for Thin1Age and LastSPH for a structural regime (ages 10, 12, 14, 16 and stocking 200, 400, 600, 800 respectively).

The data set included a total of 112 distinct values for SoilType, although many of these occupied very small areas. Initially it was decided to explore the effect of environment using stands selected across the range of productivity. Stands were ranked by observed mean productivity (I300), the range divided into four classes, and a large stand (>100 ha) nearest the middle of each class selected. This provided four stands representative of a range of observed productivity levels. Table 3 (at end of report) presents the total areas and mean I300 for the four selected stands, and for their component soil types. It was evident each stand had a predominant soil class, and sub-type within that, and that the soil groups in the eastern Bay of Plenty were more productive than those in the Waikato group.

The feature used to represent genetics in the model simulations was Family12, having 10 distinct values in the data set. Model predictions were made for the full matrix of 640 G by E by M combinations (10x4x4x4), taking 81 seconds. The highest and lowest improvements in predicted I300 are presented in Table 4 (at end of report). Those results indicate changes in productivity of more than plus or minus 10% can occur depending on the choices of genetic material and management treatments (thinning age and final stocking).

A final set of predictive runs were made across the full set of soil types (SoilType), genotypes (Family12), and management options (Thin1Age, Thin1SPH). The data was filtered to restrict to the Framing regime and to exclude stands less than 10 ha in area. Any instances of SoilType comprising less than 10 ha were also excluded, reducing the number of SoilType from 112 to 80. The simulation covered a matrix of 80x10x4x4=12800 model predict runs.

As an example, the results for the most common soil type (SoilType = EBoP\MbH, area = 8884.38 ha) are presented in Figure 6 (at end of report). The plots showed a trend for increasing I300 with decreasing Thin1SPH, maximal I300 at Thin1Age = 14, and strong variation across genetics (Family12).

RECOMMENDATIONS AND CONCLUSIONS

The full set of results from this simulation comprised predicted productivity (I300) for all combinations across the range of soil types, genetics and structural regime management options for the estate area. This effectively provided a prediction of the best combination of genetics and management for any given area in the estate and could also be used to quantify productivity losses if alternative choices were made for genetics or management. This information could be used as an aid to prescribe management for existing stands and selecting planting stock for new sites in the area – based on their soil type and desired regime. Inferences are necessarily limited to sites, genetics, and management well within the confines of the state space, with predictions less certain where there is a paucity of data. Genetics, climate, and pathogen or pest incursions are continually changing which makes future inference more difficult.

The model gives useful insights into the key drivers of productivity, which interact in ways that are poorly understood and intractable to model using conventional methods. The model can also provide a useful guide to the potential trade-offs amongst modelled features such as site, planting stock and management options.

Using management independent target variables such as I300 or SI explores G by E interactions. Exploring trade-offs of management interventions also requires analysis of financial considerations. For instance, a pruned regime might generate more valuable yields at harvest age but requires more expensive interventions during the rotation. Useful target variables would be Internal Rate of Return (IRR) or Land Expectation Value (LEV) calculated at the grid level assuming constant production, overhead, and land costs.

The simulations carried out in this study are a demonstration of what might be possible using machine learning methods with state-of-the-art forestry data sets. The data set and methods presented could be used to carry out other simulations, for example adding climate and more detailed genetic information to the simulation matrix. The methods could also be applied to expanded feature sets, for example including new data on disease and economics.

This project has demonstrated the ability to assemble a large feature set, to train and fit CatBoost models, and to use those models to predict forest productivity for different scenarios. Further research is now required to evaluate the accuracy and robustness of such predictions in a range of scenarios, and to develop operational methods.

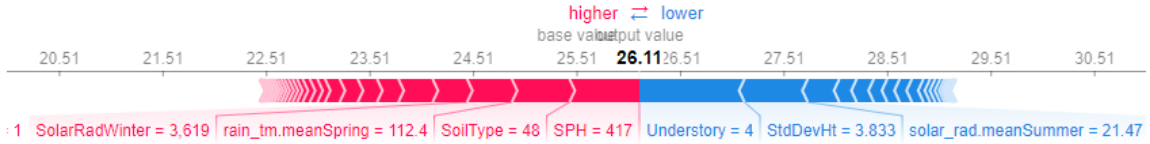


Figure 1. An example force plot (for input grid cell 1 of 290916) from the fitted decision tree.

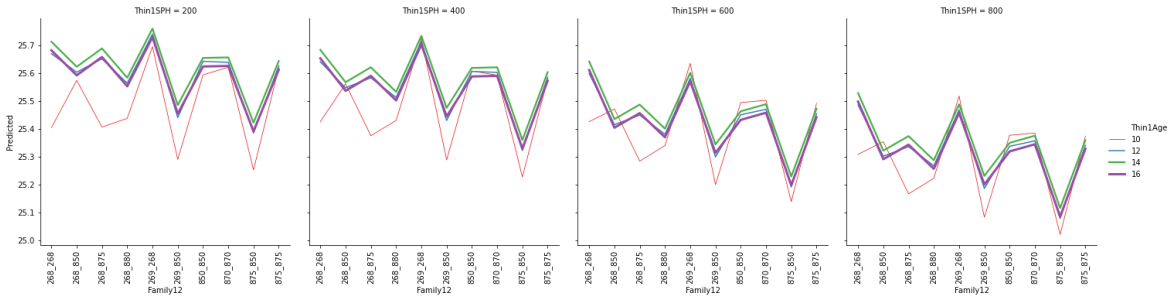


Figure 7. Predicted productivity (I300 y-axis) for the most common SoilType (EBoP\MbH) is shown by genetics (Family12 x-axis) and management (Thin1SPH columns, Thin1Age colour and line thickness).

Table 3. Characteristics of selected stands and their soil types.

Stand	Area (ha)	Mean I300	SoilType	Area (ha)	Mean I300
WANI/409/2	111.00	19.52	Waikato\KC/Kgg	102.75	19.39
			Waikato\KC/Omg+Kgg	7.44	22.88
			Waikato\KC/OiH	0.82	21.00
TIRA/3/3	104.06	22.56	Waikato\KC/Wng	68.19	22.23
			Waikato\KC/OiH	19.44	22.33
			Waikato\KC/Omg	16.44	22.71
ROTO/214/5	98.56	25.56	EBoP\Tr	84.56	26.39
			EBoP\Mb	14.00	25.42
TAWE/54/7	103.19	28.67	EBoP\HS	101.69	27.49
			EBoP\TrH	1.50	28.68

Table 4. Highest and lowest predicted productivity (I300) for example stands, with associated model input features: G (Family12), E (Stand), and M (Thin1Age and Thin1SPH).

Stand	Observed I300	Predicted I300	Improvement	Range	Family12	Thin1Age	Thin1SPH
WANI/409/2	19.52	23.09	3.57	2.44 (12.5%)	850_850	16	200
		20.65	1.13		875_850	12	800
TIRA/3/3	22.56	22.70	0.14	1.23 (5.45%)	268_875	10	600
		21.47	-1.09		875_850	12	400
ROTO/214/5	25.56	25.47	-0.09	2.02 (7.90%)	269_850	16	400
		23.45	-2.11		850_850	10	200
TAWE/54/7	28.67	27.63	-1.04	2.27 (7.92%)	268_880	14	400
		25.35	-3.31		875_850	10	800

ACKNOWLEDGEMENTS

This research was supported by the ‘Resilient Forests’ programme which is jointly funded by the New Zealand Ministry of Business, Innovation and Employment and the New Zealand Forest Growers Levy Trust. We are grateful to Hancock Forest Management NZ Ltd. for supplying the LiDAR productivity rasters, soils and stand records datasets and to the Radiata Pine Breeding Company Ltd. for making available breeding values.

REFERENCES

Bombrun, M., Dash, J. P., Pont, D., Watt, M. S., Pearse, G. D., & Dungey, H. S. (2020). Forest-Scale Phenotyping: Productivity Characterisation Through Machine Learning. *Frontiers in Plant Science, 11*. doi:10.3389/fpls.2020.00099

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). *CatBoost: gradient boosting with categorical features support*. Paper presented at the Workshop on ML Systems at NIPS 2017.

Dungey, H. S., Dash, J. P., Pont, D., Clinton, P. W., Watt, M. S., & Telfer, E. J. (2018). Phenotyping Whole Forests Will Help to Track Genetic Performance. *Trends in Plant Science, 23*(10), 854-864. doi:10.1016/j.tplants.2018.08.005

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., . . . Team, J. D. (2016). *Positioning and Power in Academic Publishing: Players, Agents and Agendas* F. Loizides & B. Schmidt (Eds.), *Jupyter Notebooks – a publishing format for reproducible computational workflows* doi:10.3233/978-1-61499-649-1-87

Manaaki Whenua - Landcare Research. S-map - New Zealand's national digital soil map. (2019). doi:10.7931/L1WC7

Molnar, C. (2019). *Interpretable machine learning. A Guide for Making Black Box Models Explainable*.

RStudio Team. RStudio: Integrated Development for R. (2020). Boston, MA: RStudio, PBC.