Growing confidence in forestry's future Research Programme

SCION FORESTS PRODUCTS INNOVATION

Date: May 2017
Reference: GCFF TN-014

Technical Note

Development of a prototype phenotyping platform for plantation forests

Author/s: Jonathan P Dash, Heidi S Dungey, Michael S. Watt, Peter Clinton

Corresponding author: jonathan.dash@scionresearch.com

Summary: This technical note summarises the key concepts and steps followed in the development of a phenotyping platform for plantation forestry in New Zealand.

Introduction

The exponential increase in the global population is causing unprecedented demands for natural resources. In particular the demands for food and wood fibre are causing significant pressure and demands on the planet's land. Plantation forests have a major role to play in meeting these demands. This will be achieved through the provision of fibre from fast growing, fit for purpose, forests. This will meet society's needs for fibre, provide valuable ecosystem services, and also alleviate the demand for wood fibre currently being extracted in a non-sustainable manner from natural and semi-natural forests. The increase in demand for agricultural and urban land uses means that the expansion of forest area in most parts of the world is not feasible. Therefore to meet the increasing demand for fibre those responsible for management and stewardship of plantation forests must increase forest productivity. Forest productivity is commonly expressed as the yield per hectare, or perhaps more appropriately, through the output units per input unit. Increasing productivity by a magnitude large enough to meet the growth in demand will require improvements in all aspects of forest management including improved genetic tree stocks, precise deployment of silviculture. In the agricultural sector productivity has increased at an impressive rate over the past 50 years. For example in the United States corn yields have increased from an average of 2500 kg/ha in 1950 to 9300 kg/ha in 2013 (Boggess et al 2013) and similar increases have been achieved in the dairy sector. Continued productivity increases of this magnitude (327 % in the case of corn yields) have been driven, in recent years by the emergence of the "omics" fields of research; genomics, proteomics and latterly phenomics.

Phenomics is an emerging field of scientific endeavour and in this context relates to the high throughput assessment of an organisms traits. With the emergence of phenomics as a field of applied research phenotyping platforms have been developed by several research groups and commercial companies around the world. Phenotyping platforms can be split into those where the plant is brought to the sensor, typically in a laboratory environment, and those that are field deployed where the sensor is moved to the plant. The latter form is more challenging because the growing conditions of the plant cannot be controlled and so must be carefully described. In this research programme we propose that due to improvements in remote sensing, tree breeding programmes, and careful spatial stand record keeping the plantation forest system can, with appropriate research and development, be thought of as an extension of a sensor to plant phenotyping platform. If we are correct then this may mean that the benefits of the emerging field of phenomics can be applied to industrial plantation forests.

The purpose of this document is to provide a summary of method development for the phenotyping platform prototype. Full documentation of the development stages and extensive reporting of results are beyond the scope of this report.

Methods

The methods used for the development of the phenotyping platform prototype for plantation forestry

supported by forestgrowers commodity levy

FOREST OWNERS ASSOCIATION

This programme is also funded by Scion and the Ministry of Business, Innovation and Employment, New Zealand.

in New Zealand are summarised in the following sections. Figure 1 provides a visual overview of the system and details how the various modules under development interact and how data flow through the prototype.

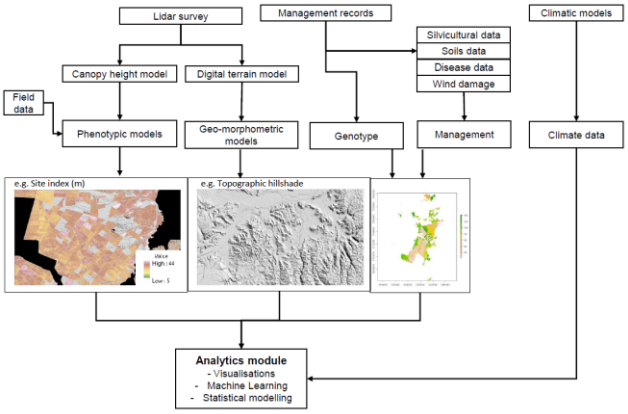

Figure 1. An overview schematic of data flow through the phenotyping platform prototype

Study site

The prototype was developed using remotely sensed data and stand records data kindly provide by the forest manager for Kaingaroa forest. Kaingaroa is New Zealand's largest plantation occupying approximately 180,000 ha of the pumice plateau south of Rotorua. The forest is predominantly composed of *Pinus radiata* (radiata pine) which constitutes ca. 90 % of the plantation area. Prototype development was limited to areas of the forest planted in radiata pine and that was established after 1999. This subset of the forest was included because we have reasonable information on the seedlot planted for these stands and the family heritage of these seed sources.

Computing resources

Development of the phenotyping platform prototype was completed on a HP desktop with an i5-4670 3.4 GHz CPU and 32GB RAM. The operating system used was a 64 bit version of Windows 7. Due to unacceptably slow computation times when moving to the analytics phase analysis was moved to a 16 core server instance with 64 GB RAM. Even with more resources computation time was still substantial. The computation module took around 24 hours to resolve a single gradient boosting machine even with 16 cores running at 100 % CPU. This is prohibitively slow for some uses and a change to high-performance computing may be advisable for a production implementation.

Prototype development

Unless otherwise stated all development took place in the open source statistical computing environment R (R Core Team 2017). The geographic data abstraction library (GDAL) was also used frequently to provide spatial transforms and geographical projection where required. As the development pathway was not clear at the start of the project an agile development approach was used. This allowed for rapid progress to a minimum viable working application so that the functionality of all aspects of the prototype could be developed and tested simultaneously. The minimum viable prototype was then iteratively improved by the developer based on the results obtained. Wherever possible free and open source software products were used to simplify licencing and ensure that a production implementation could be rolled out to a wider user base efficiently.

Datasets

The following sections detail the datasets used during the development of the prototype.

Field data

Phenotypic data was available from 500 field plots collected in the study forest between January and May 2014. The field plots were collected as part of a forest resource description that comprised an ALS survey. Field data was collected in 0.06 ha slope adjusted bounded field plots. The centre of each field plot was fixed by field crews using a survey grade global navigational satellite system (GNSS) that was post-differentially corrected using a local base station network. This meant that plot locations were known with a high degree of accuracy and could be spatially correlated with remotely sensed data. Within field plots the diameter at breast height (DBH at 1.4 m) was recorded on all trees and total tree height was recorded on a subset of suitable trees selected from across the diameter range present. These measurements were used to produce response variables for phenotypic modelling. Phenotypic variables of interest in this study were Mean Top Height (MTH), calculated as the average height of the 100 largest trees per hectare where largest is defined in terms of DBH, Basal Area (BA), Total Recoverable Volume (TRV), Site Index (SI) and 300 Index (I300).

Airborne laser scanning

An airborne laser scanning (ALS) survey was completed between the 23rd January and 6th March 2014. An Optech Pegasus scanner was used to collect a discrete, dataset with a small footprint. The data were collected with a pulse rate frequency of 100 kHz, a maximum scan angle of 12 off nadir, and a minimum swath overlap of 25%. This produced a dataset with a footprint size of 0.25 m and a pulse density of 11.46 points/m^2. Returns were classified by the supplier into ground and non-ground returns amongst other classifications. Ground returns were classified automatically using the TerraScan module of the TerraSolid software product. Subsequent manual inspection and reclassification, where required, was used to improve the classification accuracy. Within the field plot boundaries LiDAR metrics were extracted including height percentiles (P5, P10, P20, ..., P95, m), the mean (Hmean, m) and maximum height

(Hmax, m), several metrics describing LiDAR height distribution through the canopy (skewness, coefficient of variation, standard deviation (SD), and kurtosis) and measures of canopy density such as the percentage of returns reaching within 0.5 m of the ground (Pzero, %) and the percentage returns above 0.5 m (Pcover, %). These metrics were used as candidate predictor variables during the development of phenotypic models. It has been widely proven that these descriptive metrics are highly correlated with phenotypic traits in plantation forests. In a separate processing step the same set of descriptive metrics were produced for the entire study area. This formed a target dataset that could be used to apply the phenotypic models and describe the spatial patterns in phenotypic expression across the entire forest. The target dataset was represented by a series of tessellated pixels stored within a raster stack where each layer represented a different canopy metric derived from the ALS survey. A square 25 m pixel was used in this analysis. This pixel size was selected as it is approximately equivalent to the bounded field plot sized used. This plot size is based on sound forest mensuration principles. A 25 m pixel also provides a reasonable trade-off between a good resolution phenotypic description and computational performance.

Data processing and modelling

The data processing and modelling steps are summarised in the following sections.

Phenotypic modelling

The field and ALS data were used to map forest phenotype across the study area. Forest productivity is a critical phenotypic trait that can be used to assess growth rates amongst plantation trees. Using the field plot data and the ALS data surfaces of SI and I300 were developed using the techniques developed by Watt et al. 2015 and 2016 respectively. Other phenotypic traits (BA, MTH, TRV) were modelled using the random forest machine learning algorithm (Breiman 2001). This approach was chosen because it has several favourable properties including the ability to avoid model overfitting when there are many inter correlated predictors. Despite this ability dimensionality in the candidate predictor dataset was reduced by using an initial variable selection algorithm. A separate random forest model was developed for each phenotypic response variable using the default settings available in the randomForest R package. The models were then used to predict phenotypic traits across the entire forest using the wall to wall ALS data available.

Terrain modelling

All ALS tiles were processed to produce a digital terrain model (DTM) based on the ground classified returns using the blast2dem function of the LASTools software. Each tile DTM was buffered and then merged to produce a single, artefact free, elevation DTM for the study forest at a 10 m resolution. This resolution was selected following initial explorations that suggested that it provided a rigorous basis for terrain modelling given the terrain features of interest in the forest whilst ensuring a reasonable computational load given the resources available.

The DTM was exposed to a series of terrain analysis algorithms available through the open source System for Automated Geospatial Analysis (SAGA) 2.1.2 software tool. This approach was selected because of the capacity to automate analysis via a command line interface, reasonable computational performance and access to a large number geospatial algorithms. The terrain variables calculated were selected as it was believed that they contained information relating to the drivers of forest growth and therefore phenotypic expression. These terrain variables were related to the fine tuning of exposure to climatic variables such as incoming solar radiation, wind exposure and likely frosting patterns, and biotic factors such as foliar disease expression.

Output	Algorithm	Source
Slope	Maximum slope	Travis et al 1975
Aspect		Travis et al 1975
Mid-slope position	Relative Heights and Slope Positions	Boehner and Conrad 2008
Sky view factor		Boener and Antonic 2009
Topographic openness		Anders et al 2009
Valley depth	Top hat approach	Rodriquez et al 2002
Vector ruggedness	A Terrain Ruggedness that Quantifies Topographic Heterogeneity	Sappington et al 2007
Visible sky		Boener and Antonic 2009
Wind exposure	Topoclimatic assessment of wind exposure	Boehner and Antoni 2009
Wetness Index	Topographic wetness index	Boehner and Selige 2006
Topographic Position	Topographic Position index	Guisan et al. 1999
Terrain classification	Automated classifications of topography from DEMs	Iwahashi and Pike 2007

Stand record system

A geospatial database containing all stand management records associated with the study forest was made available by the forest manager. Amongst the spatially explicit data contained in the dataset was tree species, initial planted stocking, stand density, pruning and thinning status, fertilisation and disease treatment records, site preparation and records

detailing spraying for foliar disease occurrence in the forest. Conversion procedures were developed to convert these data from the forest mangers storage format within commercial stand management software to a format that could be ingested into the prototype system.

Soils data

Soils data was provided by the forest manager in a geospatial file format (.shp). These data have been digitised and maintained by the forest manager as updated versions of the original soil maps produced for the study forest (Rijkse, WC 1988). These maps detailed soil classification for the entire study forest. However the spatial data had variable levels of detail depending on the information available for specific sections of the forest. All spatial data was merged and aligned and transformed as required for inclusion in subsequent processing steps. An extensive survey of soil chemistry throughout the study forest is currently underway and will be incorporated into the phenotyping platform once completed.

Climate data

Climate data for the study forest was extracted from a national dataset developed and maintained by the National Institute for Water and Atmospheric Research (NIWA) https://www.niwa.co.nz/climate/research-projects/national-and-regional-climate-maps. Annual and seasonal means for climatic variables of interest were extracted from the dataset and made available as a raster covering the study forest. The climate variables used in the prototype were temperature, wind speed, rainfall, global radiation, and sunshine hours. Climate data was included in the prototype in order to provide a measure of the climatic growing environment experienced by the study trees during the growth cycle.

Genetics data

The study forest comprises genetic material data that has been the subject of tree breeding programmes for many decades (Dungey et al. 2009). The spatial distribution across the study forest of some aspects of genetic composition of the planted tree stocks was known. The genetic information available for the study forest included the seedlot name, the GF rating in some cases, the seedlot planting stock type and the source nursery for the seed stock. In total there were 1151 *Pinus radiata* seedlots distributed across the study forest. Around 20 % of these seedlots have a GF value recorded in the database that can be used in analysis currently. Other research streams within this research programme are improving our understanding of the genetic composition of the study forest. Once these results become available they can be incorporated into later iterations of the phenotyping platform and used to improve the analysis.

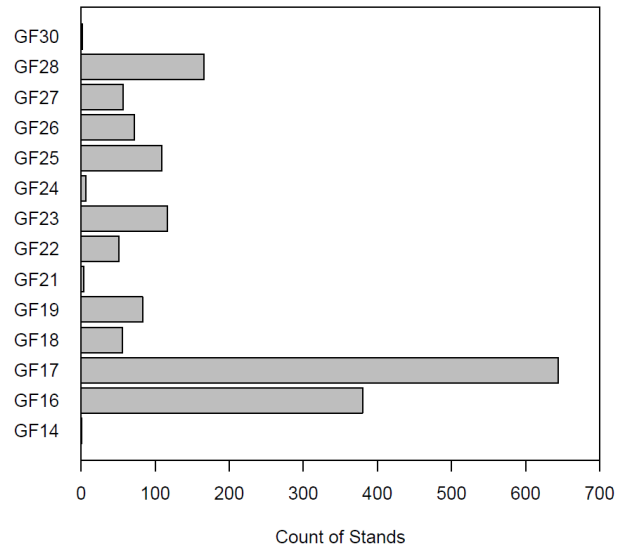

Fig 2. Histogram of the count of GF values present in the study forest

Extraction layer

A series of spatial manipulations, projections and transforms have been developed to align the various data sources and make these available for analysis to an analytics module via an extraction layer. Logic rules ensure that all data in the extraction layer is suitable for analysis and is in an appropriate format. The data modelling and data generation process is compute intensive and time consuming and so it is logical that the extraction layer is stored and used to move the phenotyping platform between computing hardware. The extraction layer consists of a raster stack where each square pixel represents a 25 m patch of the study forest. Each layer in the extraction layer raster stack relates to either a phenotypic variable or a candidate explanatory variable extracted from the processes outlined here. This forms the basis of visualisation and statistical analysis available through the phenotyping platform.

Analytics module

An analytics module was developed to operate on data delivered by the extraction layer. The analytics module was developed in R as this environment has access to many statistical modelling ad learning libraries and a powerful, and well supported, visualisation framework. R is becoming increasing popular, not just within the scientific community, but also amongst forest industry professionals. However, it is recognised that other languages, notably Python's Scikit-learn, offer viable and possibly favourable alternatives to the current approach.

Two machine learning algorithms were selected for use in the initial data explorations within the prototype. There were the ensemble statistical learning methods random forest (RF) and gradient boosting machines (GBM).

The ensemble decision tree classifier Random forest (RF) uses bootstrap aggregated sampling (bagging) to construct many individual decision trees, from which a final class assignment is determined (Breiman 2001). RF is now regularly applied to natural resource assessment (Mellor et al 2013) and has previously been used, in combination with remotely sensed data, to successfully model several variables of interest in this forest type (Dash-et al 2015, Dash et al 2016, Watt et al 2015a, Watt et al 2016). Decision trees are constructed using a sample from the available training data, with the remaining assigned as out-of-bag (OOB) samples. At each node, a random subset of predictor variables are tested to partition the observation data into increasingly homogeneous subsets. The node-splitting variable selected from the variable subset is that which resulted in the greatest increase in data purity (variance or Gini) before and after the tree node split (Cutler-2007). This process ends when there are no further gains in purity. Response variables can be continuous, calculated by averaging, or categorical where predictions are derived from a model vote among decision trees. Computational efficiency of the algorithm is enhanced, compared with alternative approaches, as only a sample of variables are used at each node split. This also reduces correlation between trees, improving both predictive power and classification accuracy. The OOB sample data are used to compute accuracies and error rates, averaged over all predictions, and estimate variable importance (Cutler 2007, Mellor et al 2013). RF provides two methods to estimate the importance of each predictor variable in the model. The mean decrease in accuracy (MDA) importance measure is calculated as the normalised difference between OOB accuracy of the original observations to randomly permuted variables (Cutler 2007, Mellor et al 2013). An alternative variable importance measure is calculated by summing all of the decreases in Gini impurity at each tree node split, normalised by the number of trees (Criminisi et al 2012, Mellor et al 2013). RF is a well-regarded machine learning tool that can identify complex and non-linear relationships in fitting datasets and has been shown to offer high classification accuracy (Cutler 2007, Criminisi et al 2012, Dash et al 2017).

Gradient boosting was chosen here for its ability to accommodate nonlinear interactions, resilience to outlier influence, tolerance of collinear predictor variables, and ability to handle categorical and missing data (Elith et al 2008, Balzotti and Asner 2017). In short, GBMs are a decision tree based regression technique that sequentially decreases model bias. Unlike other tree-based models, such as Random Forest, GBMs are computationally more intensive and require proper selection of metaparameters to prevent overfitting (Hastie et al 2009). We followed guidelines, provided by Elith et al (2008) and the R package dismo (Hijmans et al 2016), to choose model tuning parameters during the implementation of the GBM models. The final tuning parameters for the foliar all models included the number of decision trees, tree complexity (number of nodes), contribution of each tree to the growing model (learning rate), and the loss function (Balzotti and Asner 2017).

The performance of both RF and GBM models was evaluated using 10-fold cross validation of the fitting dataset.

Initial Results

The following sections contain detail on provisional results from the prototype phenotyping platform to give an indication of the types of analysis and insights that will be available. The extraction later presents the analytics module with a sizable dataset. With 2.5 M cells the extraction layer for the study forest is still small enough to fit in memory of a normal machine but large enough to make data manipulations and model fitting slow. The extraction layer also contains a moderate number of candidate covariates (74 in the current example) meaning that carefully graphical analysis is necessary. This step will need to be complete by an analyst but several other components of the analytics module can, and have, been automated. The initial results reported here show the output of a GBM model fitted to the phenotyping platform with SI as the response variable. Figure 3 shows the relationship between SI and mean spring temperature across the forest extracted from the phenotyping platform for a subset (around 200,000) of samples where GF values were known during this analysis. A trend between SI and mean spring temperature is evident but there are also significant patterns here associated with the other environmental and genetic factors affecting growth. This highlights the complexity of the scenario and shows why an approach based on machine learning is required to unravel these interactions.

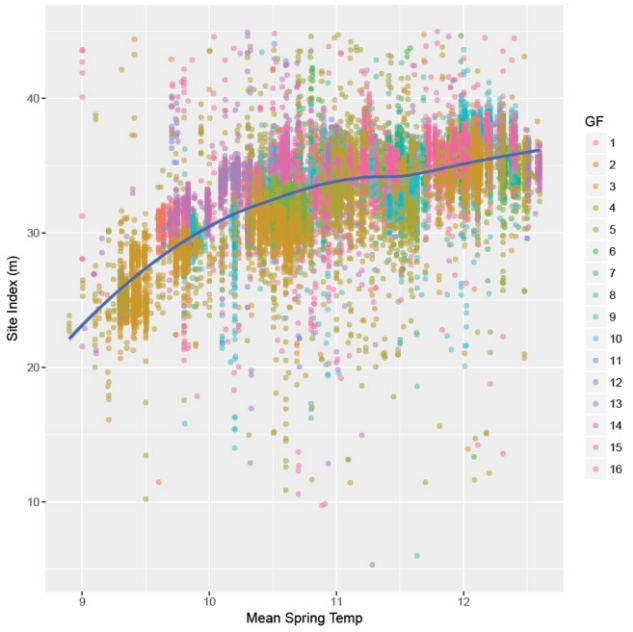

Figure 3. The relationship between SI and spring temperature

The GBM was trained using the methods suggested by Elith et al (2008). At the time of writing specification of all parameters associated with the final GBM model that can be used for prediction are still ongoing. The results reported here refer to the initial GBM model developed using the rules of thumb for comparable datasets set out by Elith et al (2008). Figure 4 shows the relationship between the number of trees fitted in the GBM and the hold out deviance used to assess model performance. The hold out deviance can be seen declining as more and more trees are added to the model indicating the power of statistical learners of this type to improve based on the result of previous weaker models.

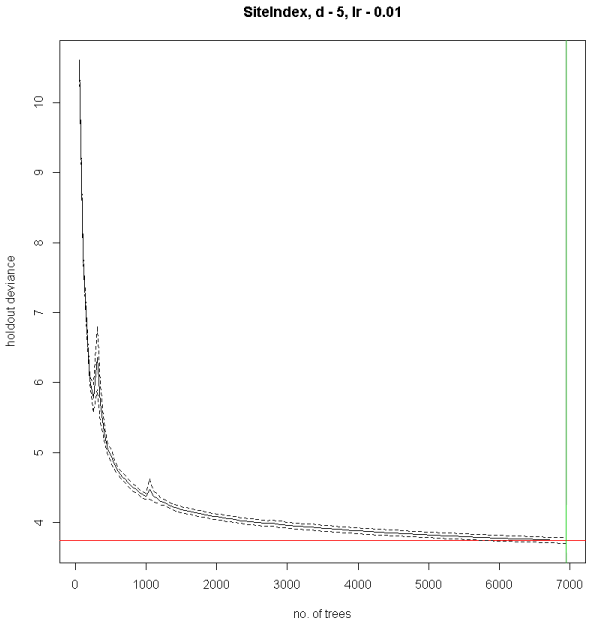

Figure 4. The relationship between holdout deviance and model quality for the initial GBM fitted.

The initial GBM was used to output the relative influence of the candidate predictor variables in the study on SI (Figure 5). This analysis indicated that the genetic variable seedlot code was the most influential predictor within the GBM by a significant margin. This was followed by a variable denoting soil classification and several environmental variables relating to the climatic and terrain characteristics.

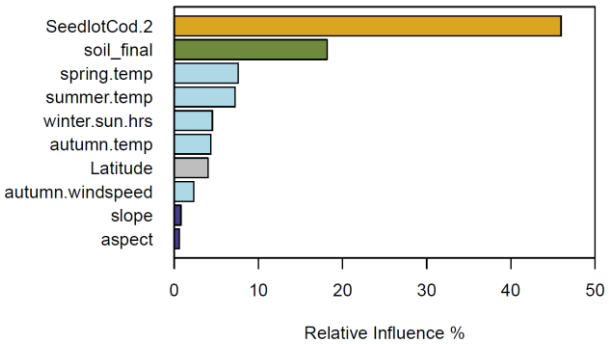

Figure 5. The relative influence (%) of ten most important predictors in the initial GBM.

The effect of explanatory variables on model outputs can be outputted from a GBM. These plots can be seen in Figure 6 as an example of the type of analysis that can be extracted from the statistical learners of this type. The effect of seedlot code is quite apparent and the increase in SI along a temperature gradient is also obvious. GBM models also provide functionality for assessing the interactions between predictor variables. At the time of writing these methods were still under development for the prototype but it is expected that these will form a useful component of this analysis.

Further work

Further research will focus on:

- Improving the inputs to the prototype including better incorporation of estimates of disease level.
- Improving seedlot information and contributions to each stand to fully understand the relationship detected.
- Using the improved GFPlus and family information improve the understanding of genotype performance across the estate.
- Better training of the GBM meta-parameters to allow the use of GBM as a predictive model.
- Extension of the GBM modelling process to I300 and then to all other phenotypic traits. Investigation of methods to improve the phenotypic models produced.

Limitations

Whilst we recognise that there are numerous limitations to this research the readers should recognise that this document describes the development of a prototype only.

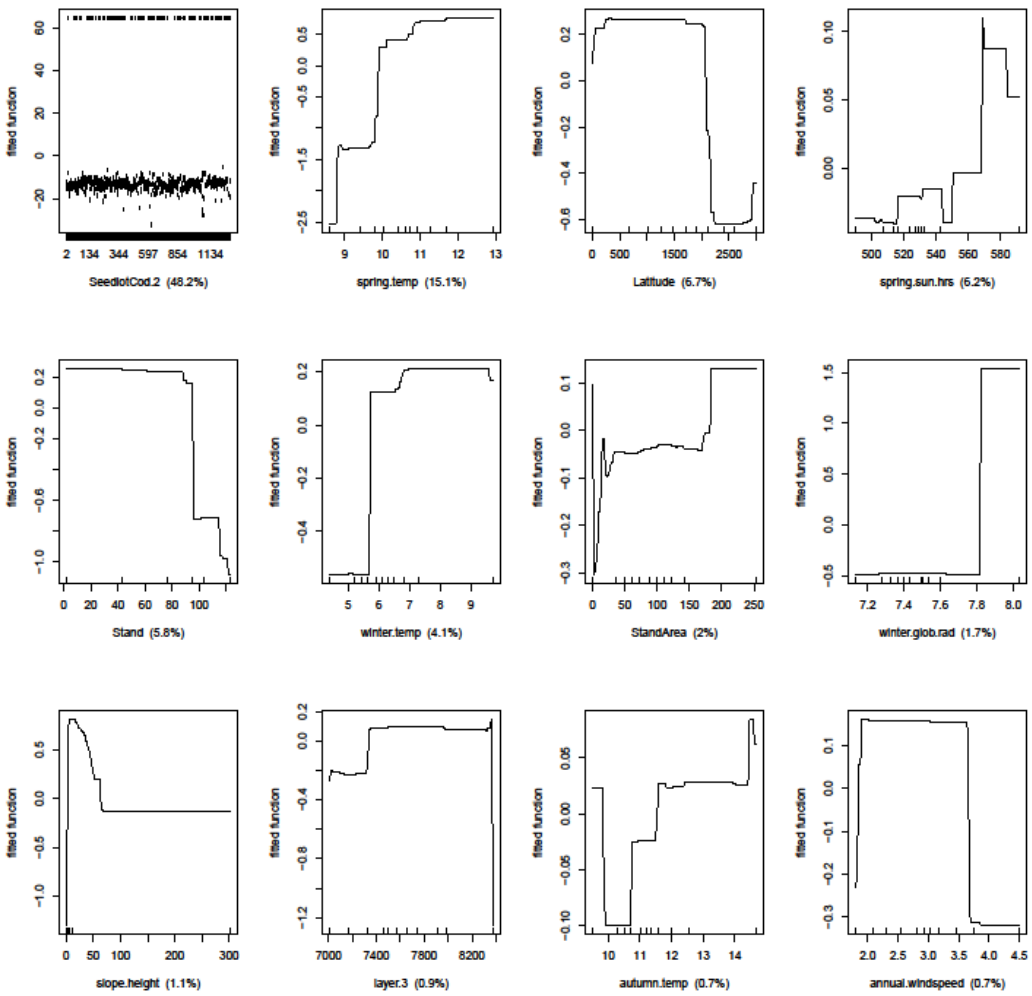

Figure 6. Modelled genetic and environmental factors for the initial GBM fitted.

Acknowledgements

Funding for this research came from the "Growing Confidence in Forestry's Future" research programme (C04X1306), which is jointly funded by the Ministry of Business Information and Employment (MBIE) and the Forest Growers Levy Trust, with the support of the NZ Forest Owners Association (FOA) and the NZ Farm Forestry Association (FFA).

This programme is also funded by Scion and the Ministry of Business, Innovation and Employment, New Zealand.

Bibliography

Balzotti, C.S, and Asner G.P. (2017) Biotic and Abiotic Controls over Canopy Function and Structure in Humid Hawaiian Forests. Ecosystems

Boggess, M.V., Lippolis, J.D., Hurkman, W.J., Fagerquist, C.K., Briggs, S, P., Gomes, A.V. Righetti, P.G. Bala K. (2013) The Need for Agriculture Phenotyping: "Moving from Genotype to Phenotype" Journal of Proteomics. 93 20-39

Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32.

Criminisi, A.; Konukoglu, E.; Shotton, J. Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning; NOW Publishers: New York, NY, USA, 2012.

Cutler, D.R.; Edwards, T.C.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random forests for classification in ecology. Ecology 2007, 88, 2783–2792.

Dash, J.P.; Marshall, H.M.; Rawley, B. Methods for estimating multivariate stand yields and errors using k-NN and aerial laser scanning. Forestry 2015, 88, 237–247

Dash, J.P.; Watt, M.S.; Bhandari, S.; Watt, P. Characterising forest structure using combinations of airborne laser scanning data, RapidEye satellite imagery and environmental variables. Forestry 2016, 89, 159–169.

Dungey, H.S.; Brawner; J.T., Burger, F.; Carson, M.; Henson, M.; Jefferson, P, Matheson, A. C. (2009) A new breeding strategy for *Pinus radiata* in New Zealand and New South Wales. Silvae Genetica, **58**, 28–38.

Elith J., Leathwick, J and Hastie, T. (2008) A working guide to boosted regression trees. Journal of Animal Ecology **77** , 802–813

Hastie T, Tibshirani R, Friedman J. (2009). The elements of statistical learning. Elements 1:337–87.

Hijmans, R.J., Phillips, S.P., Leathwick, J., Elith, J. (2017). Dismo: Species Distribution Modelling. R Package version 1.1-4.

Mellor, A.; Haywood, A.; Stone, C.; Jones, S. The performance of random forests in an operational settingfor large area sclerophyll forest classification. Remote Sens. 2013, 5, 2838–2856

Rijkse, WC (1988). Soils of the Kaingaroa Plateau, North Island, New Zealand. NZ Soil Bureau District Office Report RO 14. New Zealand Department of Scientific and Industrial Research, Rotorua.

Watt, M.S.; Dash, J.P.; Bhandari, S.; Watt, P. Comparing parametric and non-parametric methods of predicting Site Index for radiata pine using combinations of data derived from environmental surfaces, satellite imagery and airborne laser scanning. For. Ecol. Manag. 2015, 357, 1–9

Watt, M.S.; Dash, J.P.; Watt, P.; Bhandari, S. Multi-sensor modelling of a forest productivity index for radiata pine plantations. N. Z. J. For. Sci. 2016, 46, 1–14