Large-Scale Phenotyping Inferences: From Trees to Forest through Machine Learning

Maxime Bombrun, PhD – Data Scientist, Data Analytics Team
Jonathan Dash, Heidi Dungey, David Pont, Michael Watt
February 2019

Context

New Zealand's Top 3 Exports:

- Dairy, eggs, honey: $15 billion (27.6%)
- Meat: $7 billion (12.7%)
- **Wood: $4.9 billion (9%)**

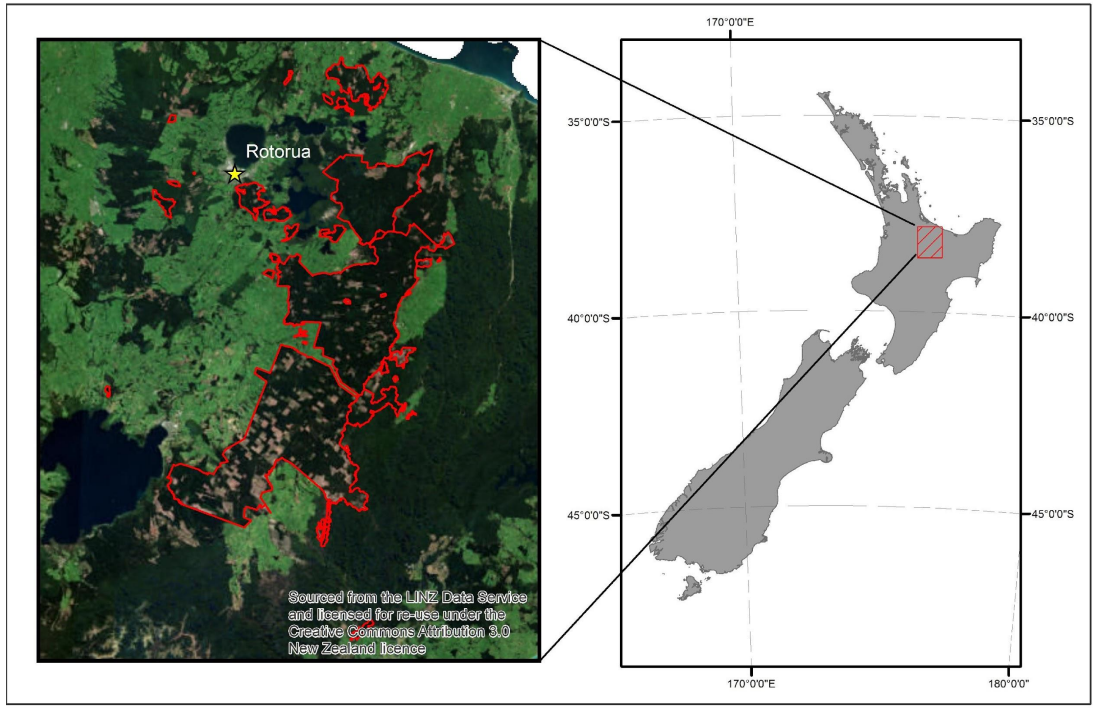
Radiata Pine accounts for **89%** of the total planted resource in New Zealand.

The *Grow Confidence in Future Forestry* programme goals are:

- increase returns from existing forests through mid-rotation interventions
- improve returns from existing forests through better knowledge of wood quality
- increase the productivity and consistency of future forests

SCION

Objectives



The genomics revolution

- Shorten the breeding cycle,
- Increase crop productivity.

However,

- Accurate phenotyping is challenging,
- Phenotyping at a forest scale.

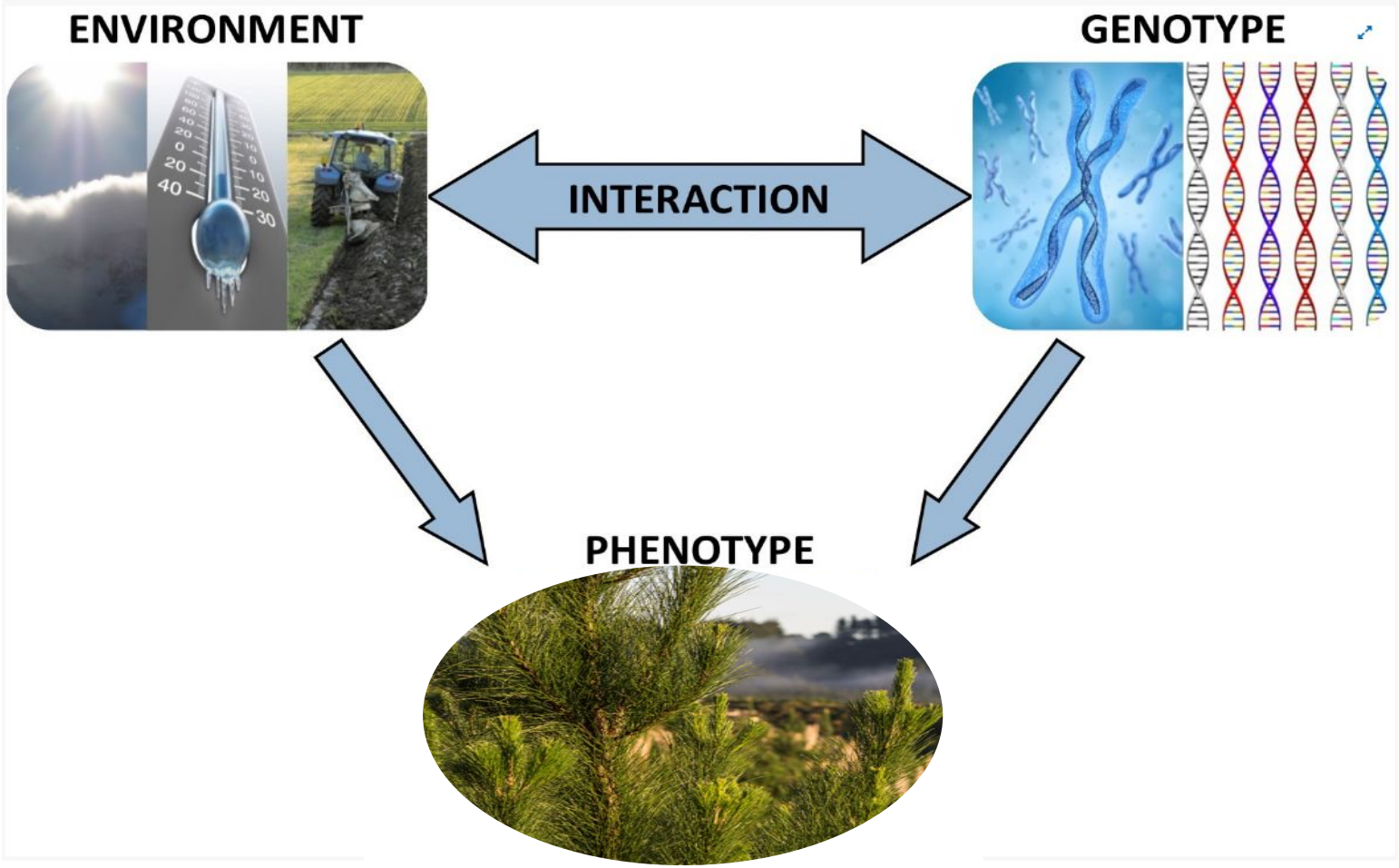
Currently,

- Large dataset for the Kaingaroa forest,
- Stand productivity quantified using the Site Index,
- Machine learning approach to model productivity based on these variables.

In-forest phenotyping – a new concept

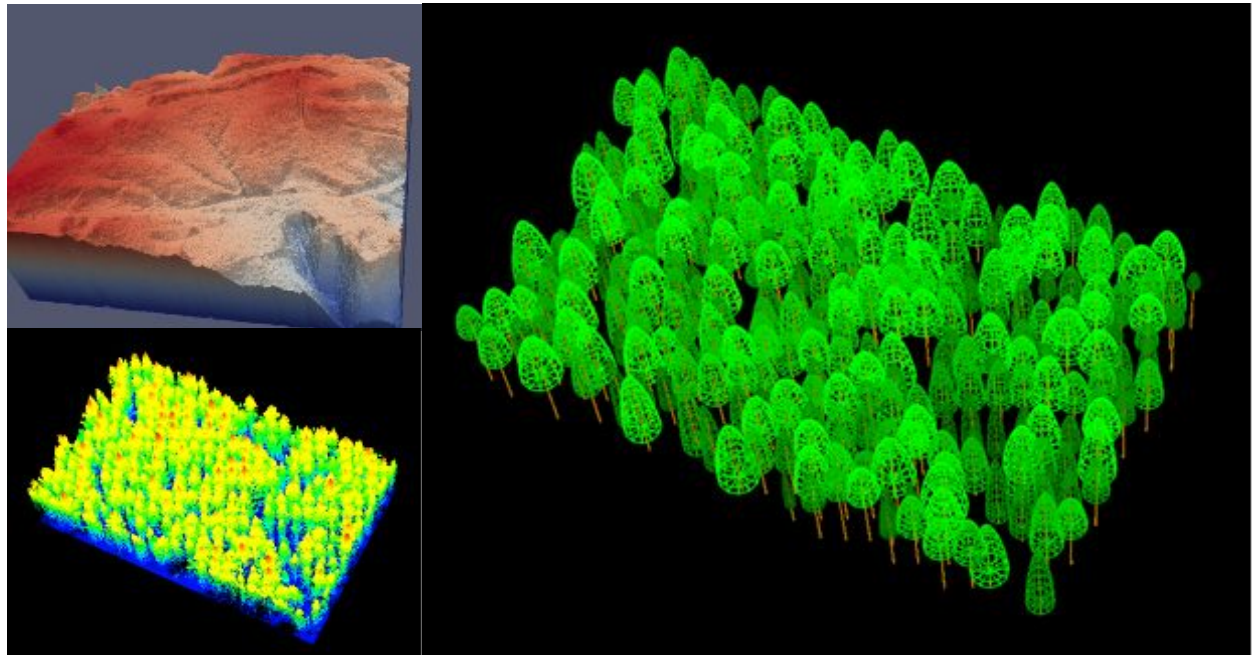
- Enables quantification of **plant traits** and **environmental conditions** to identify and assess the performance on germplasm of interest

- Allows better matching of **seedlot to site**, based on past and current performance



Phenotyping methodology

Many sensors and platforms have been developed for plant phenotyping

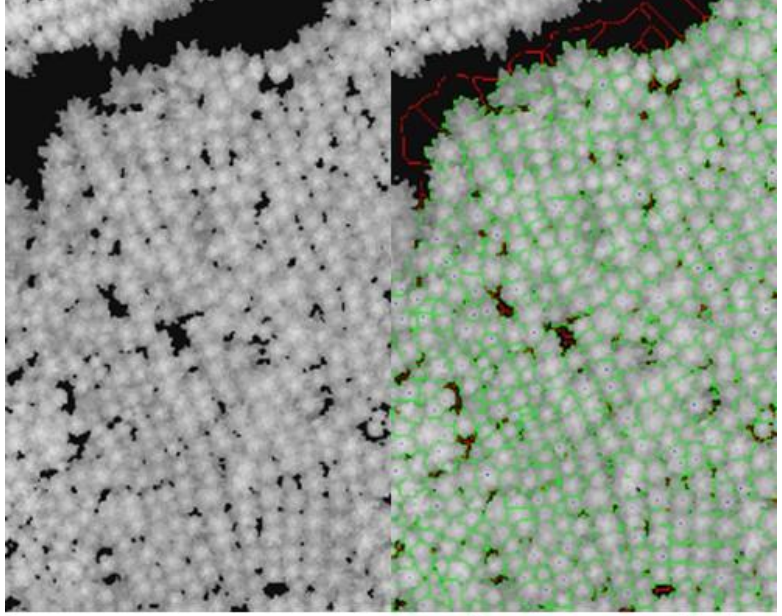


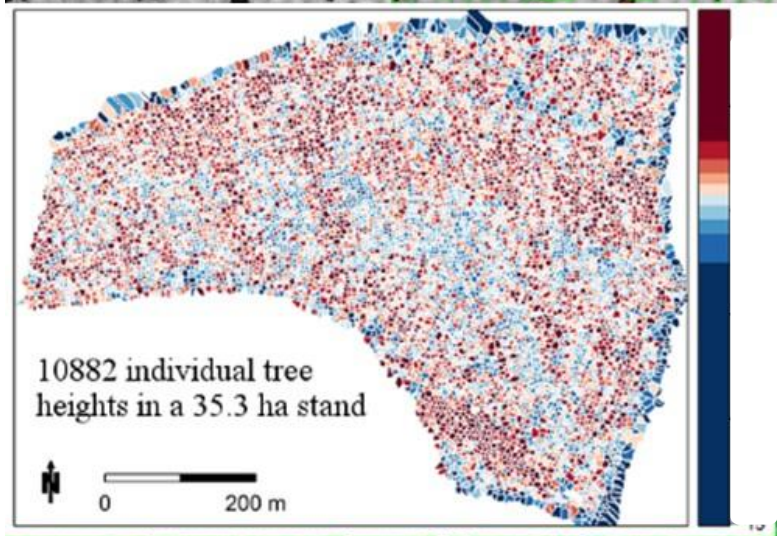
Two levels of phenotyping

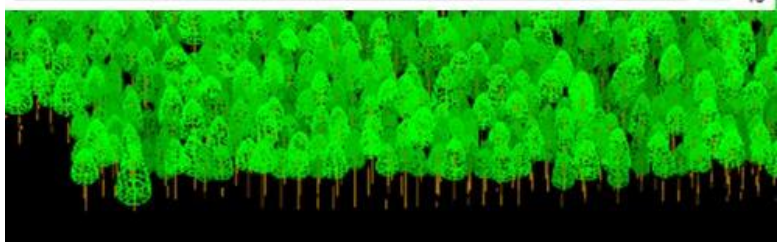
- tree-based, applied to stands and eventually the forest estate
- area-based (25 x 25 m grid), applied to the forest estate

Application of tree-based phenotyping

- Airborne Laser Scanning used to carry out tree detection and crown delineation
- We have identified crown metrics that accurately estimate tree height and volume
- Currently detecting and characterising every tree in operational stands (30-40 ha, 10,000+ trees)
- Developing methods to select individual trees exhibiting exceptional height or volume growth

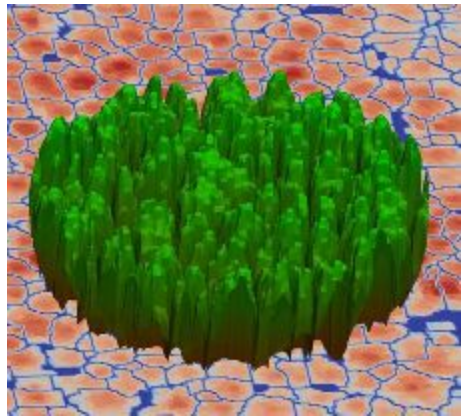


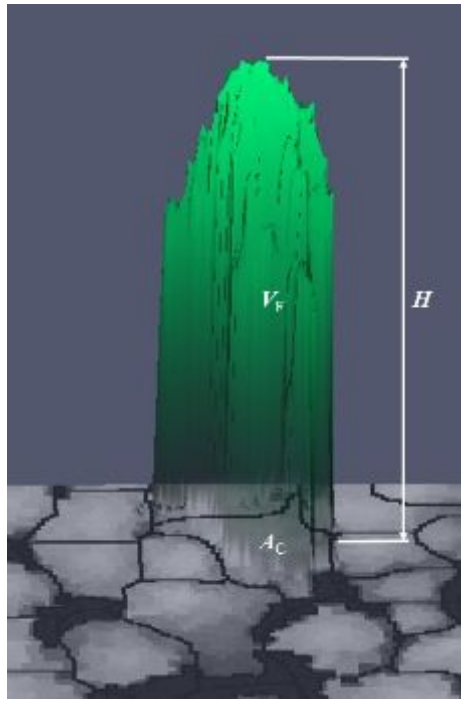


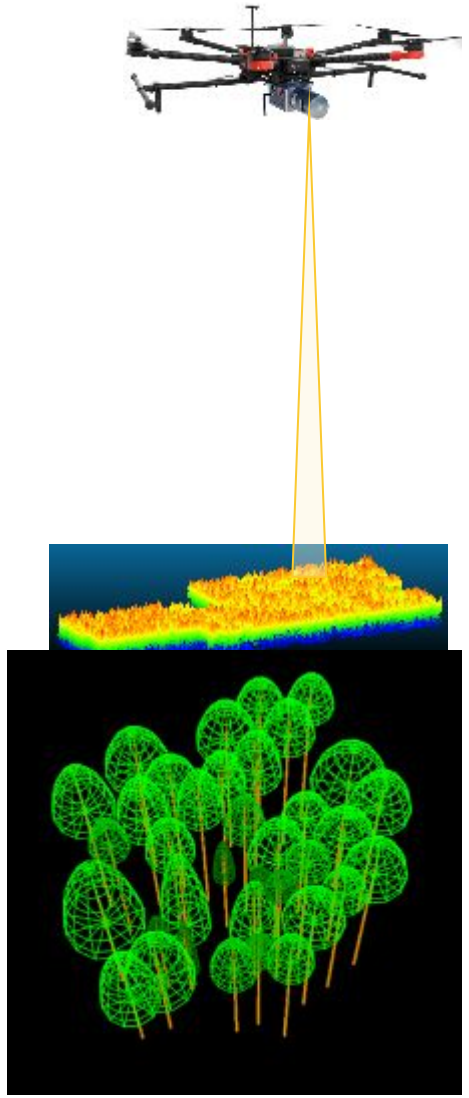


Benefits of tree-based phenotyping

- Quantifying GxExS interaction to determine tree growth
- Identification of superior tree performance
- Accelerated tree breeding
- Optimising the matching of breeds to sites
- Precision stand mapping and management
- Supports higher productivity - higher returns from stands



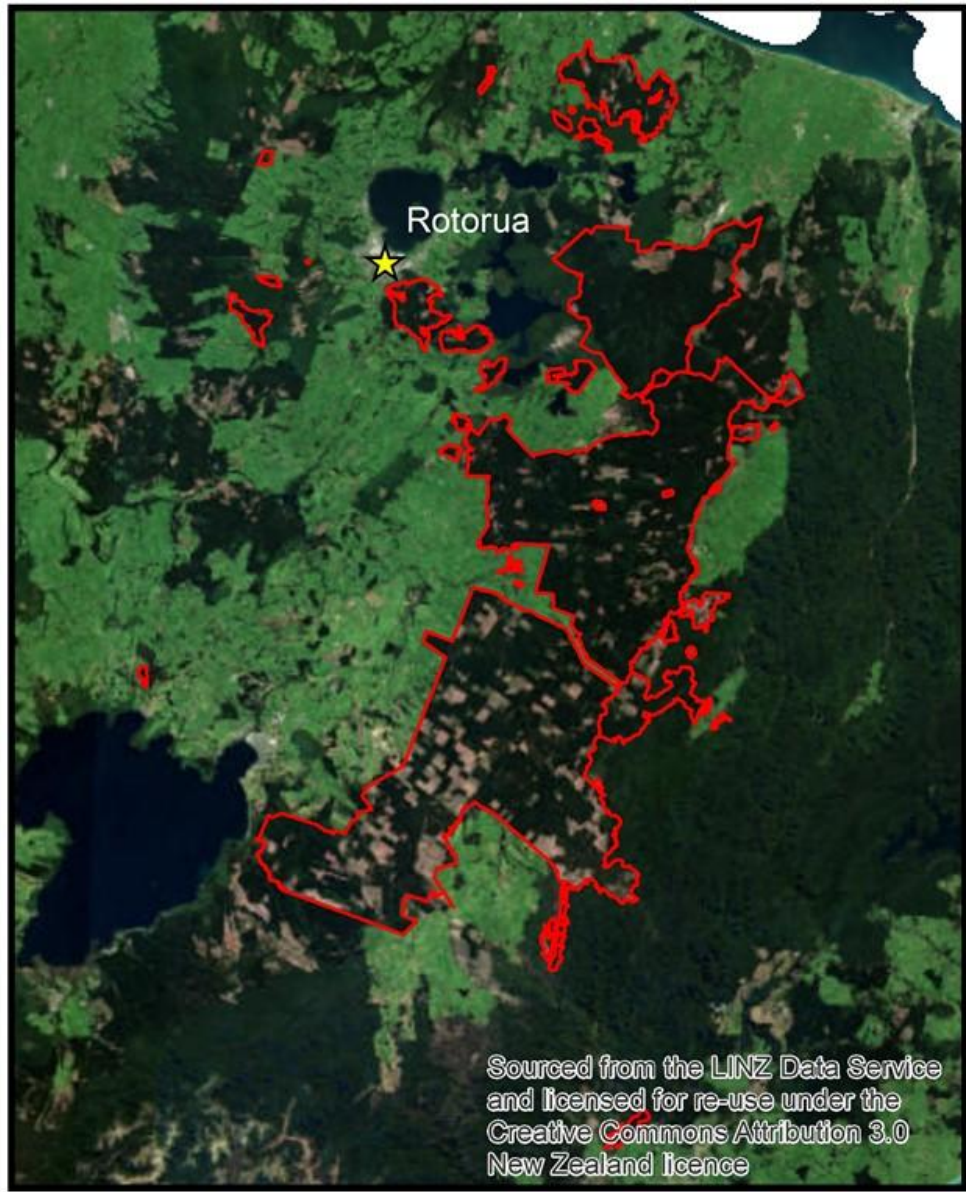




SCION

Application of area-based phenotyping

- Remote sensing to model forest phenotype across landscapes on an area basis in **Kaingaroa**
- Seedlot information provides us with a useful starting point
- Climatic data is integrated to identify key sites properties across the landscape
- The forest phenotyping platform allows us to

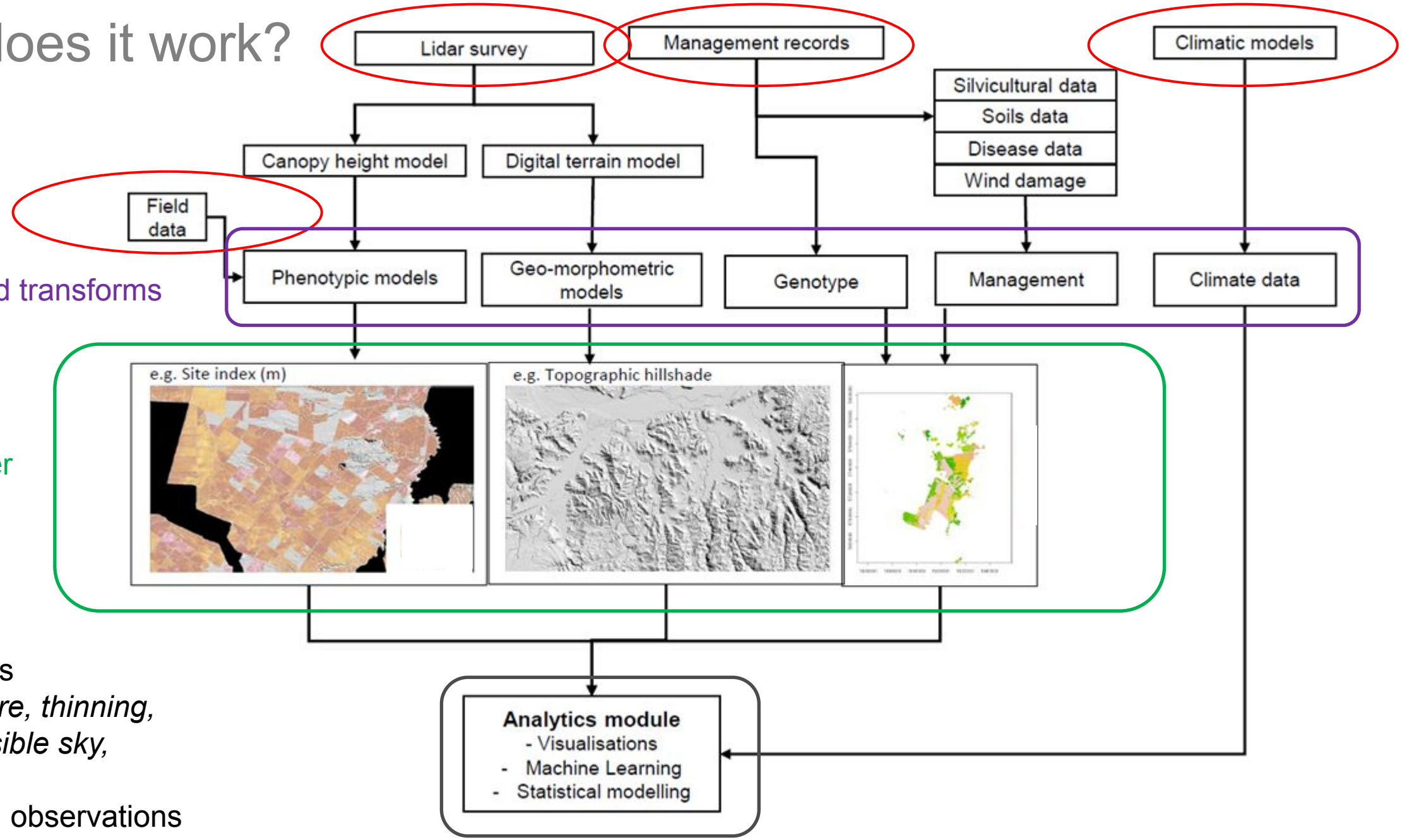


How does it work?

Inputs

Modelling and transforms

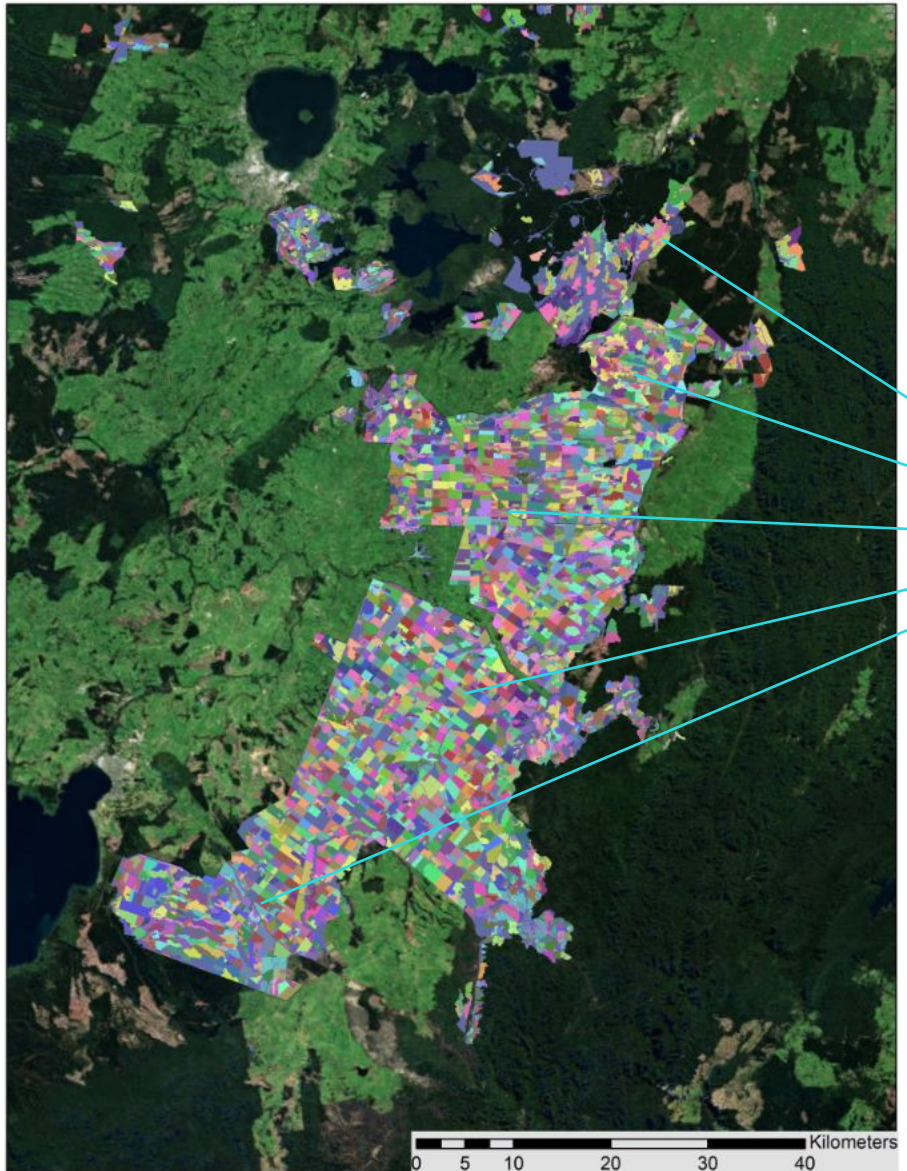
Extraction layer



- 93 variables (*temperature, thinning, seedlot, visible sky, species*)
- 17 781 121 observations

-> Over one billion values!

Seedlot distribution



GF Plus™ SEED CERTIFICATE

Seed Producer	PF Olsen Ltd	Date	
		Year of Collection	2012
Orchard	Seddon	Seedlot Number	12/208
Number Of Parents	29	Number of Crosses	27
Pollination Method	CP	Breeding Values Version	2008v1

Relying on the information provided by the seed producer this seedlot is rated as:

	Growth	Straightness	Branching	Dothistroma	Wood Density	Spiral Grain
% Of Seedlot Rated	100 %	100 %	100 %	-	100 %	95 %

Special Comments

SCION SEED CERTIFICATION SERVICE

Manager: C.B. Gould Date: 24-Oct-12

An indication of the estimated ratings for an average unimproved seedlot is:

GROWTH STRAIGHTNESS BRANCHING DOTHISTROMA WOOD DENSITY SPIRAL GRAIN

Ratings allocated are estimates of the seedlot average. They are developed from data that has differing levels of confidence, thus the more parents involved in a seedlot, the higher the confidence level of the rating. An asterisk(*) after the GF rating infers a less than average confidence level. A hyphen (-) means that there was insufficient data available to estimate a rating.

When different seedlots are compared strictly under the same conditions the following will usually apply;

the higher the rating, the:

Better the expected average growth (diameter).
More multinodal branching habit of the seedlot.
Higher average wood density (juvenile wood).
Better the expected average stem straightness.
Greater the resistance to Dothistroma.
Lower the average incidence of spiral grain.

IMPORTANT

The GF PlusTM trade mark, copyright in this GF PlusTM Seed Certificate and all other intellectual property used in the testing and certification of the seedlot described above ("Seed") and in the creation of, or pertaining to, this GF PlusTM Seed Certificate ("Intellectual Property") belong exclusively to the Radiata Pine Breeding Company ("RPBC"). Without limiting the terms of any terms of supply or licence between you and RPBC, you may not use the GF PlusTM trade mark, copy this GF PlusTM Seed Certificate or otherwise use the Intellectual Property to promote, advertise, distribute or sell any plants propagated by or for you from plants grown or derived from any of the Seed ("Plants") unless you have signed a licence to do so from RPBC. For purposes of clarification, propagated means the propagation of any plant by vegetative means (including, without limitation, tissue culture) for the purpose of producing multiple plants from a single plant.

Trait ratings

Growth	Straight	Branch
22	21	23

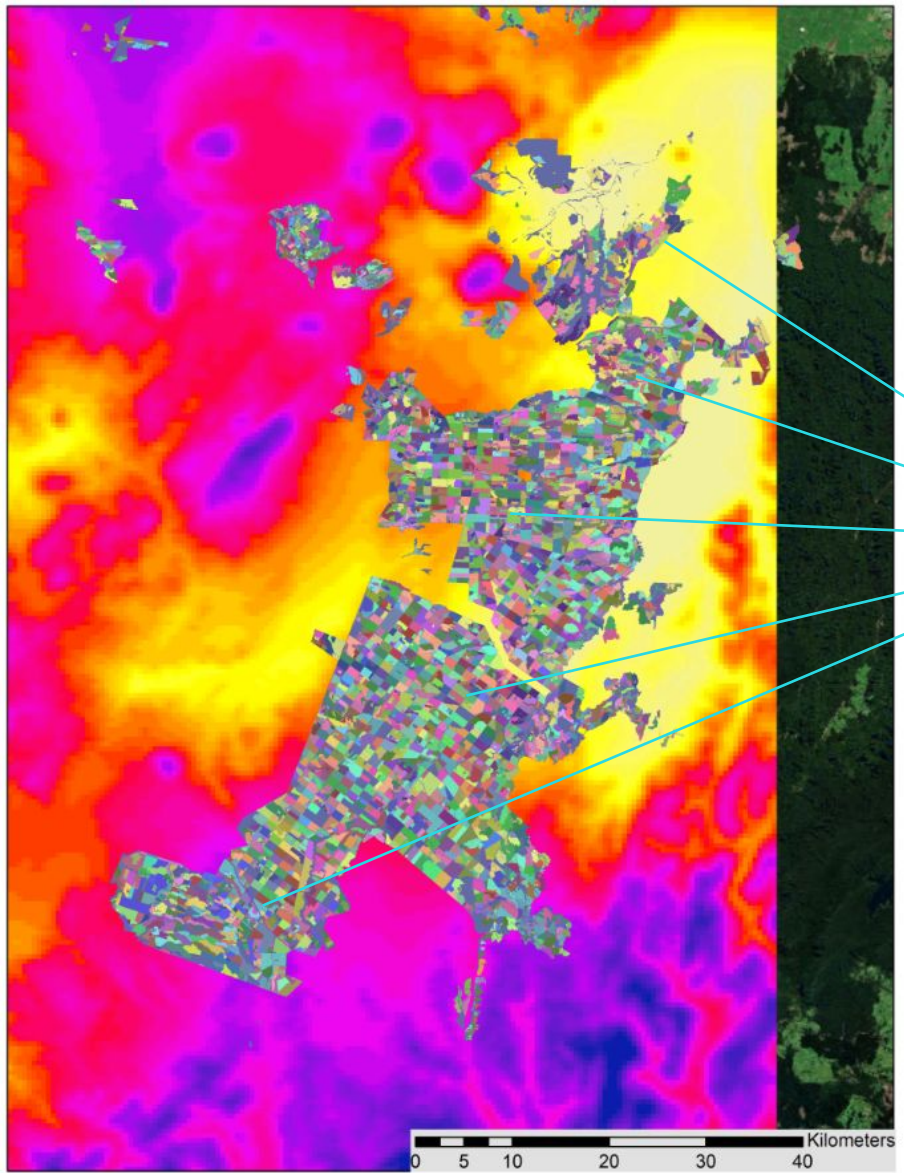
List of parents

APPENDIX I
Seedlot Crosses

No. of Crosses: 40

Parents		% of
Female	Male	seedlot
268.005	268.532	0.8%
268.005	875.220	0.3%
268.007	268.054	4.1%
268.007	268.228	2.1%
268.007	875.220	1.3%
268.054	268.262	1.7%
268.054	268.532	1.9%
268.054	875.066	3.4%
268.123	268.248	1.9%
268.228	268.609	1.3%
268.228	875.076	2.4%

Climate overlay



GF Plus™ SEED CERTIFICATE

Seed Producer	PF Olsen Ltd	Date	
		Year of Collection	2012
Orchard	Seddon	Seedlot Number	12/208
Number Of Parents	29	Number of Crosses	27
Pollination Method	CP	Breeding Values Version	2008v1

Relying on the information provided by the seed producer this seedlot is rated as:

	Growth	Straightness	Branching	Dothistroma	Wood Density	Spiral Grain
% Of Seedlot Rated	100 %	100 %	100 %	-	100 %	95 %

Special Comments

SCION SEED CERTIFICATION SERVICE

Manager: C.B.Gould Date: 24-Oct-12

An indication of the estimated ratings for an average unimproved seedlot is:

GROWTH STRAIGHTNESS BRANCHING DOTHISTROMA WOOD DENSITY SPIRAL GRAIN

Ratings allocated are estimates of the seedlot average. They are developed from data that has differing levels of confidence, thus the more parents involved in a seedlot, the higher the confidence level of the rating. An asterisk(*) after the GF rating infers a less than average confidence level. A hyphen (-) means that there was insufficient data available to estimate a rating.

When different seedlots are compared strictly under the same conditions the following will usually apply;

the higher the rating, the:

Better the expected average growth (diameter).
More multinodal branching habit of the seedlot.
Higher average wood density (juvenile wood).
Better the expected average stem straightness.
Greater the resistance to Dothistroma.
Lower the average incidence of spiral grain.

IMPORTANT

The GF PlusTM trade mark, copyright in this GF PlusTM Seed Certificate and all other intellectual property used in the testing and certification of the seedlot described above ("Seed") and in the creation of, or pertaining to, this GF PlusTM Seed Certificate ("Intellectual Property") belong exclusively to the Radiata Pine Breeding Company ("RPBC"). Without limiting the terms of any terms of supply or licence between you and RPBC, you may not use the GF PlusTM trade mark, copy this GF PlusTM Seed Certificate or otherwise use the Intellectual Property to promote, advertise, distribute or sell any plants propagated by or for you from plants grown or derived from any of the Seed ("Plants") unless you have signed a licence to do so from RPBC. For purposes of clarification, propagated means the propagation of any plant by vegetative means (including, without limitation, tissue culture) for the purpose of producing multiple plants from a single plant.

Trait ratings

Growth	Straight	Branch
22	21	23

List of parents

APPENDIX I
Seedlot Crosses

No. of Crosses: 40

Parents		% of
Female	Male	seedlot
268.005	268.532	0.8%
268.005	875.220	0.3%
268.007	268.054	4.1%
268.007	268.228	2.1%
268.007	875.220	1.3%
268.054	268.262	1.7%
268.054	268.532	1.9%
268.054	875.066	3.4%
268.123	268.248	1.9%
268.228	268.609	1.3%
268.228	875.076	2.4%

XGBoost model

eXtreme Gradient Boosting (XGBoost):

- Machine learning combining
  - Boosting algorithm
  - Base learner tree method
- High performing solutions in large and complex competitions (e.g., Kaggle)

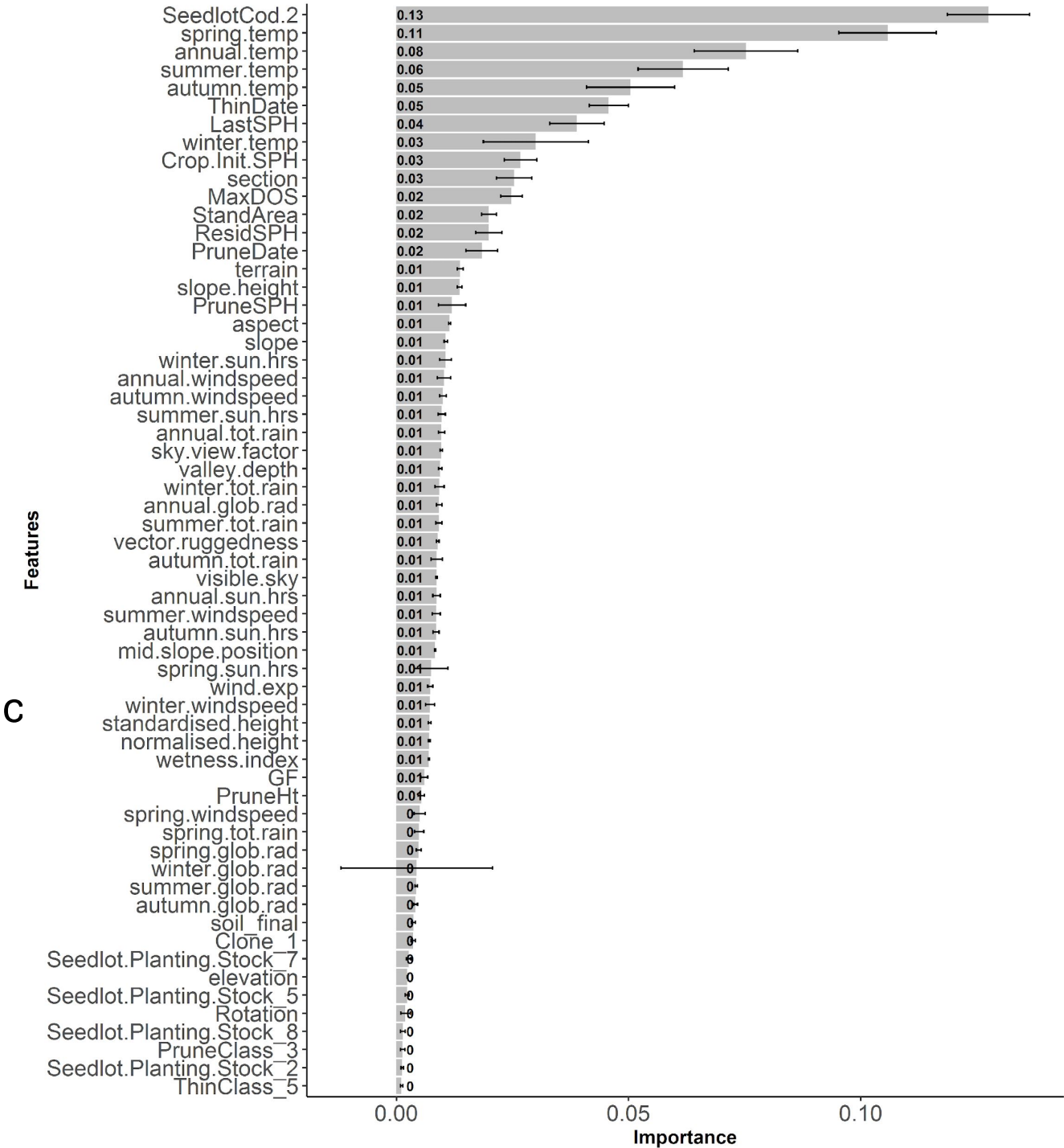
+ Fast (GBM: 24hr, Xgboost: ~6hr) and robust, automatic feature selection, even correlated,

+ Understandable, feature of importance, thresholds of decision, constructed tree(s)

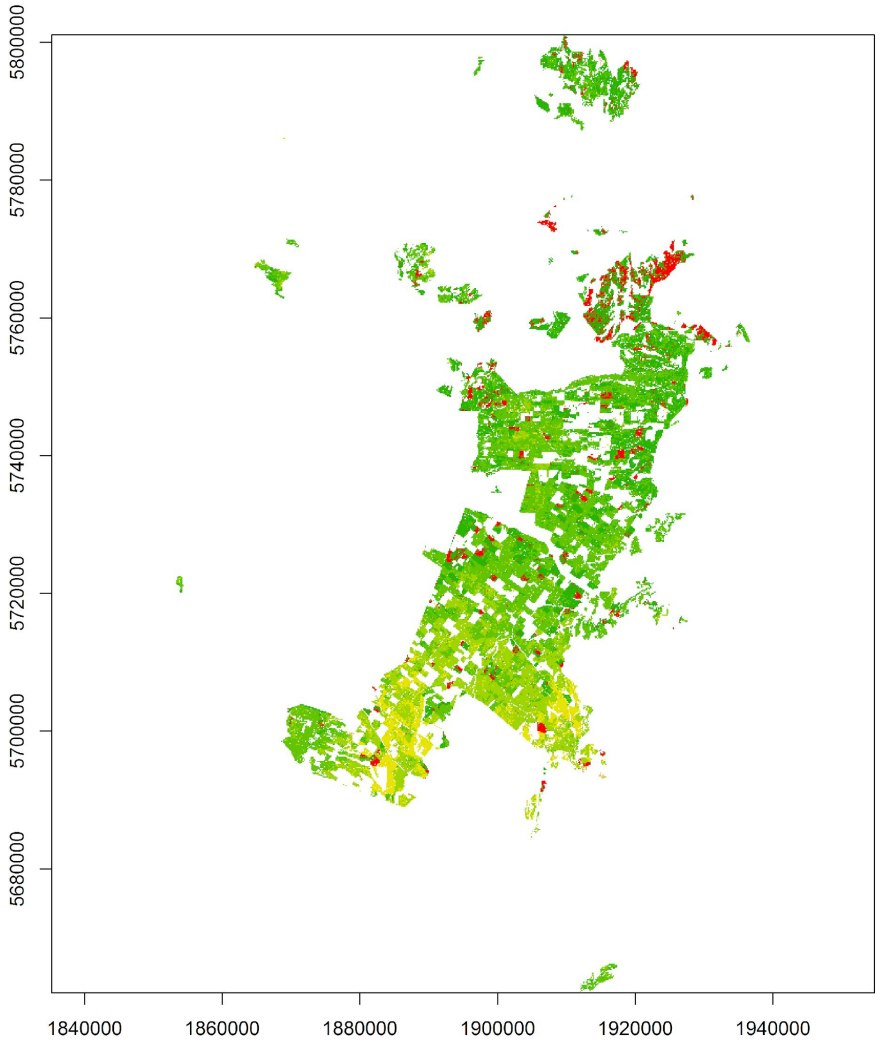
Early results

Initial models account for ~91 % of the variance in forest productivity across the forest

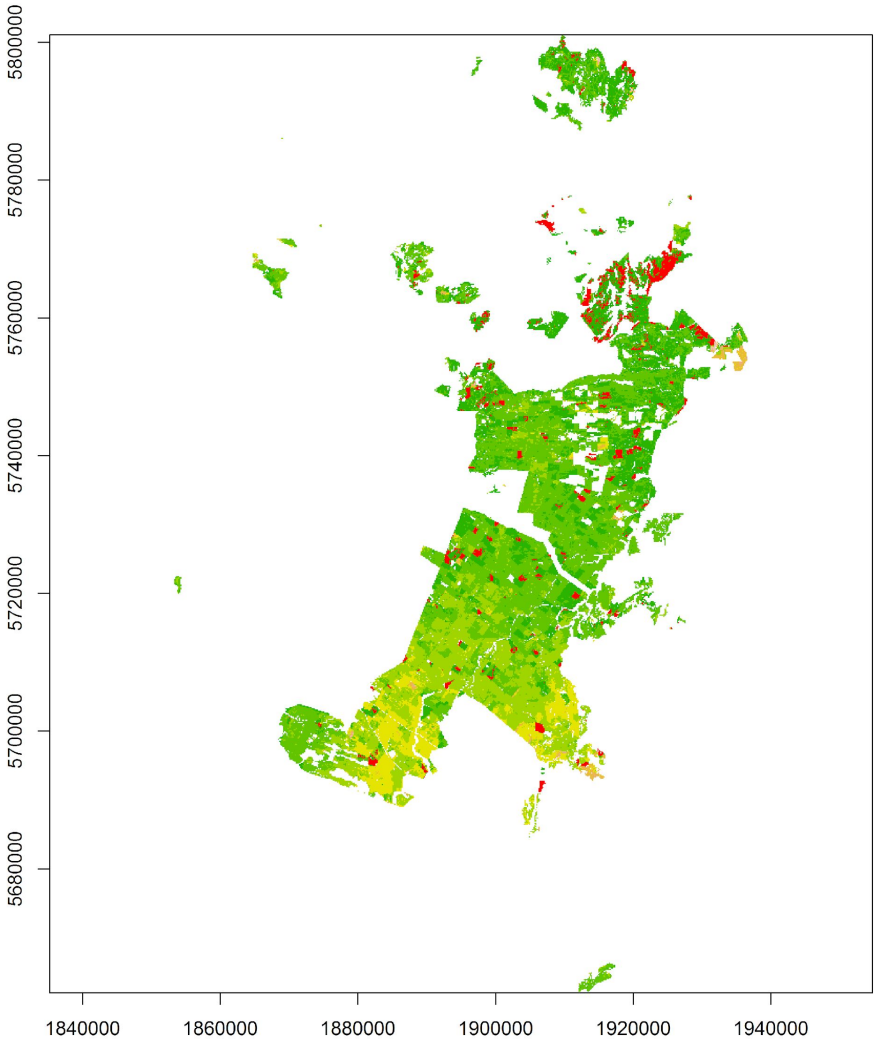
- Seedlot and site attributes are consistently the most important factors, highlighting the importance of matching seedlot to site
- We can use the XGBoost outputs to examine the interaction between genetic and site factors
- Future analysis will indicate where best to locate specific seedlots



Predictions of Site Index values

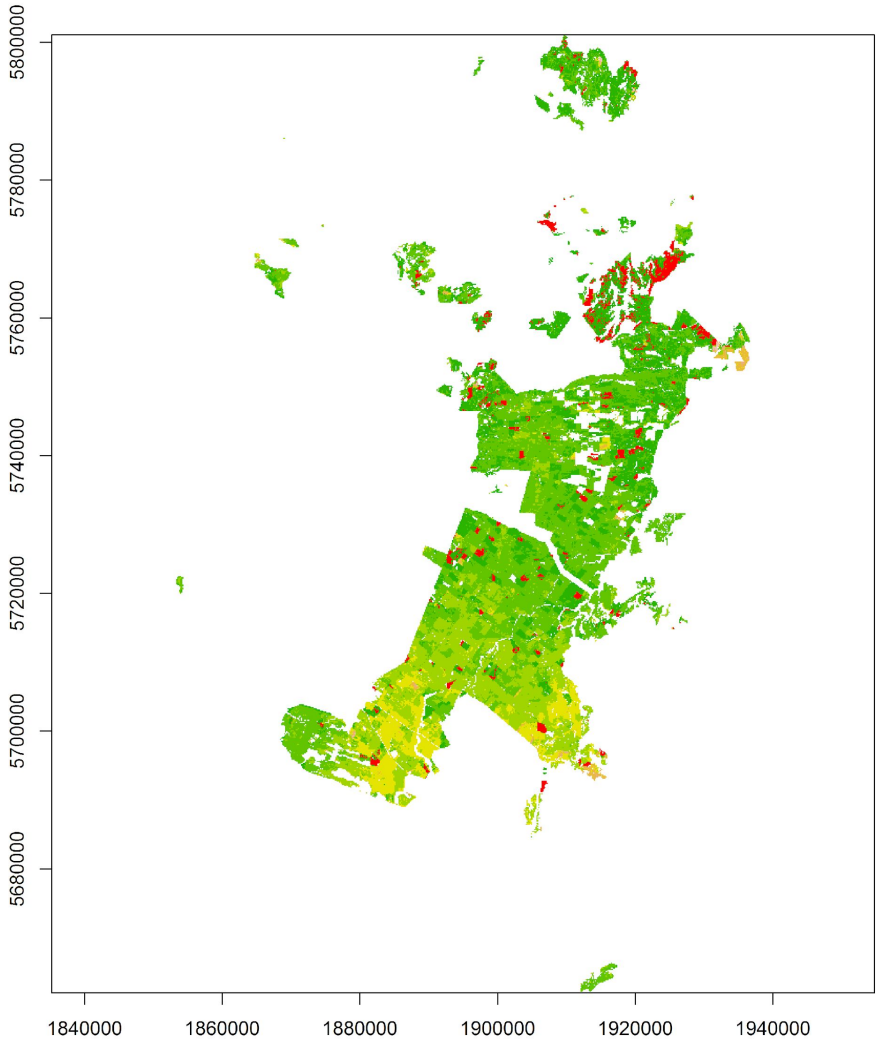


Raw observations

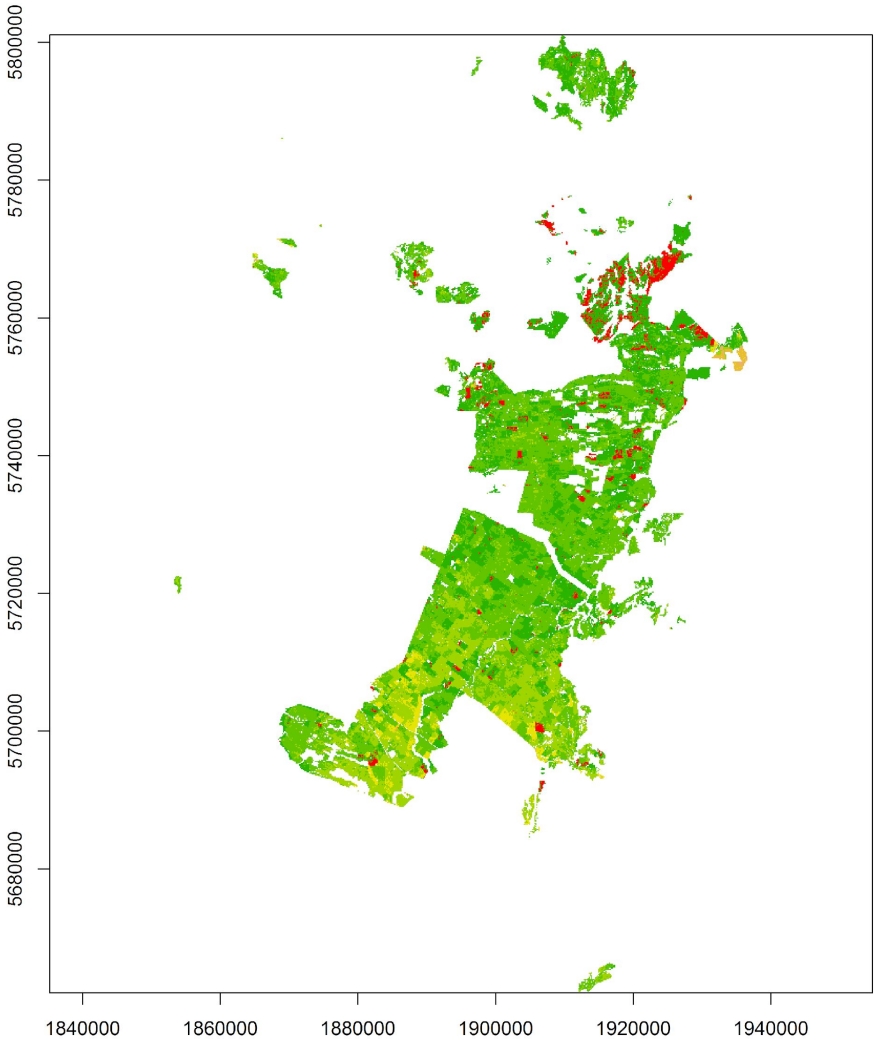


Predicted observation

Predictions of Site Index values for Clone 1



Predicted observations



Predictions Clone 1

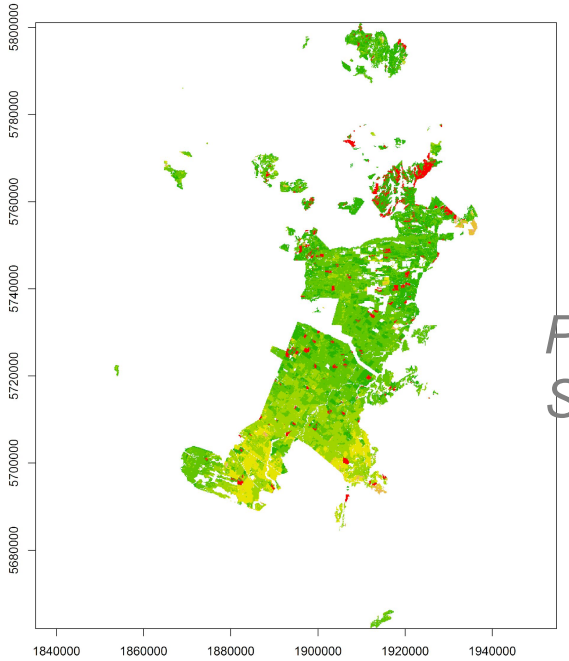
Productivity from genetics

Investigation on seedlot with the highest productivity:

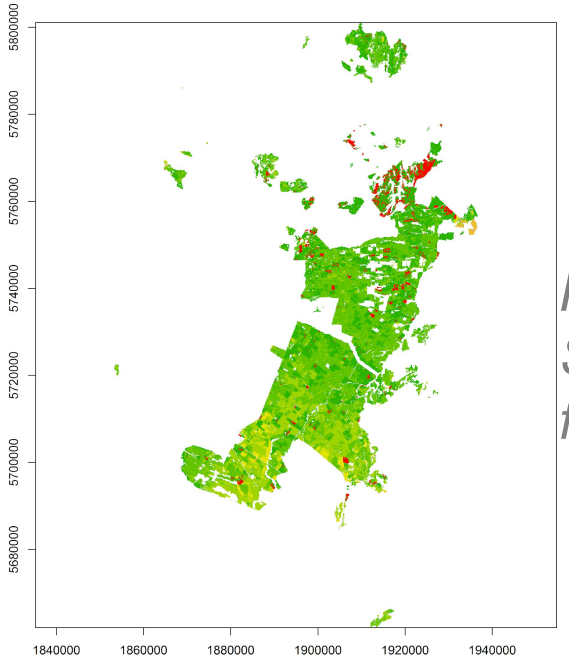
- Selective tree stocks based on environmental conditions,
- Predictions of productivity based on the XGBoost model, thus, enhance the genetic potential,
- Estimations in missing area from known parameters,
- Superior genotype across the forest.

Indicators of interplay between site and genetic factors, possibilities of higher productivity

ID	Percentage_Tot	Percentage_noZero	Improvement
Pos_Clone1	60.77	61.45	2 813 899
Pos_Optimised	16.75	57.8	184 712



Predicted Site Index



Predicted Site Index for clone1

Conclusions

- The Kaingaroa dataset
  - Vast, informative and revealing of the productivity at a forest-scale level.
  - Phenotyping at broad spatial scale was achieved through combining the analytical power of advanced machine learning methods with spatial layers acquired from remotely sensed data, management records and climatic surfaces.

- This methodology can be used to map variation in productivity between seedlots at fine spatial scale under varying environments to identify superior genotypes across the forest and thus, the continual optimisation of deployed genetically improved tree stock.

Maxime Bombrun
Data Scientist
Maxime.Bombrun@scionresearch.com

www.scionresearch.com

www.gcff.nz

www.fgr.nz

Date: 18/02/2019

supported by forestgrowers commodity levy

www.scionresearch.com

Prosperity from trees *Mai i te ngahere oranga*