

Final Report

Using predictive analytics to decompose site index

CAFS.20.83

Jason Cross, University of Washington
Eric Turnblom, University of Washington

Eric Turnblom, Presenter



Justification

1. Site index is a required input for many growth & yield models. Unless & until a replacement is coded, site index will need to be evaluated. This is reflected by CAFS IAB membership placing high priority on improved parameterization of growth & yield models.
2. The relationship between modified-Weibull rate and shape parameters fit empirically indicate that dominant & co-dominant trees captured in HT40 calculations are on the photosynthetic frontier for a given site.
3. Height/age curves are constructed using historic data where each dimension of growth varied within historic ranges. Observations on future sites will be at odds with predictions where dimensions of growth depart from historic ranges.
4. Planting density confounds on observed site index: Top height is related positively with planting density early in stand development; with the relationship possibly turning negative later in stand development.
5. The utility of site index is highest at the time of planting in order to make decisions that most affect future stand development, such as initial density, whether to PCT and at what timing & intensity, and whether to fertilize.

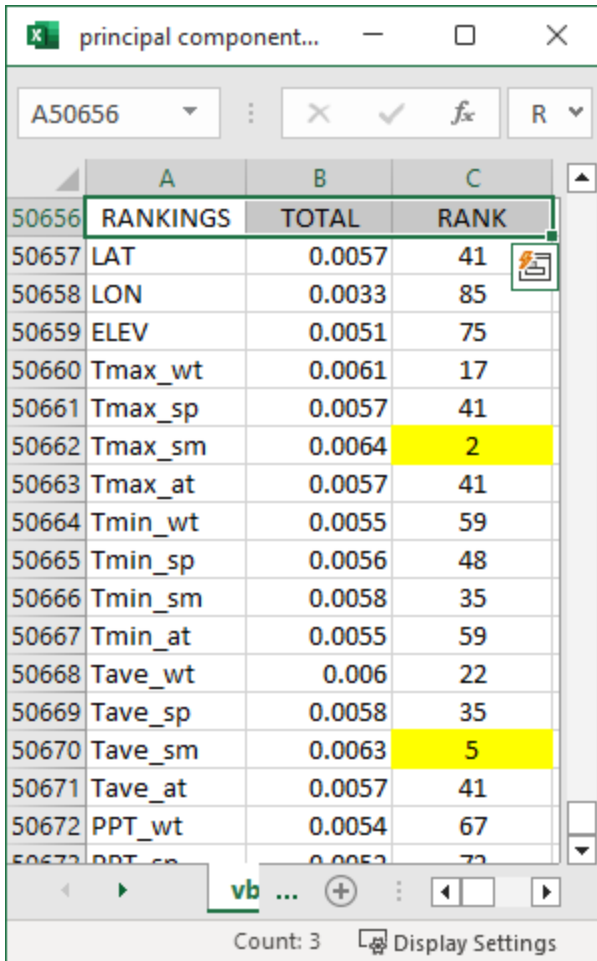


Hypotheses and Objectives

1. We hypothesize that observed site index is a function of both static attributes (e.g. elevation, latitude, soil composition), dynamic stand measurements (e.g. basal area, mean diameter), and regional attributes that measure within historic ranges (e.g. temperature, precipitation, insolation).
2. The objective is to build a direct model of site quality that captures the effect of interactions between multiple independent variables in the form of top-height: a dynamic measure that substitutes directly as site index in applications.



Methods: Principal Components

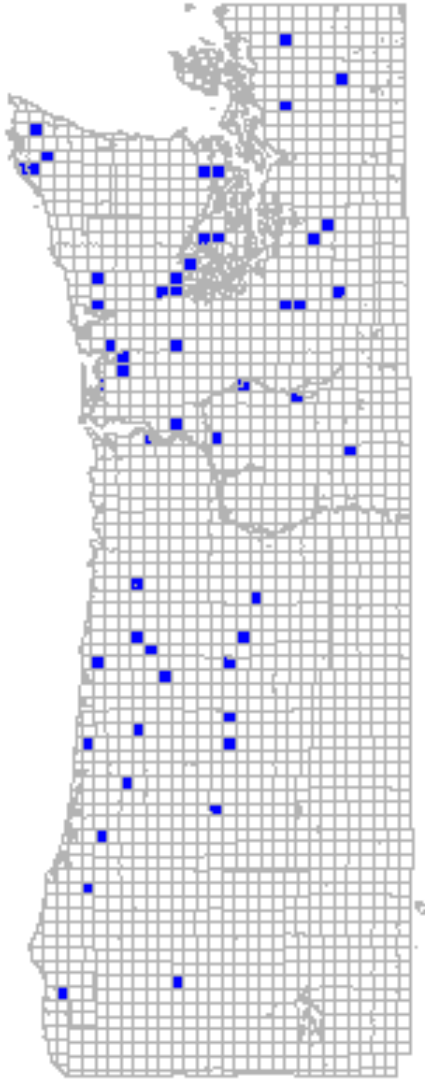


	A	B	C
50656	RANKINGS	TOTAL	RANK
50657	LAT	0.0057	41
50658	LON	0.0033	85
50659	ELEV	0.0051	75
50660	Tmax_wt	0.0061	17
50661	Tmax_sp	0.0057	41
50662	Tmax_sm	0.0064	2
50663	Tmax_at	0.0057	41
50664	Tmin_wt	0.0055	59
50665	Tmin_sp	0.0056	48
50666	Tmin_sm	0.0058	35
50667	Tmin_at	0.0055	59
50668	Tave_wt	0.006	22
50669	Tave_sp	0.0058	35
50670	Tave_sm	0.0063	5
50671	Tave_at	0.0057	41
50672	PPT_wt	0.0054	67
50673	PPT_sp	0.0052	72

1. Raw data (predictors)
2. Standardize values:
3. Covariance matrix
4. Eigenvectors & Eigenvalues
5. Principal Components
6. Vector loadings
7. Variance by predictor
8. Predictor importance



Methods: Model Fitting



1. Selected model: modified Weibull CDF
$$HT = \{b_0 \cdot [1.0 - \exp(-b_1 \cdot BHAGE^{b_2})]\} + 4.5$$
2. Nelder-mead method^{[1] [2]}
3. Initial reduction of categorical variables
4. Secondary addition of continuous variables identified by principal components
5. On completion, resample data with replacement (bootstrap), refit with reduced convergence criteria.
6. Compute predictor p-values after k bootstraps; eliminate least important.

[1] Gao, F. and Han, L. 2010. Implementing the Nelder-Mead simplex algorithm with adaptive parameters. *Comp. Optim Appl.* <https://doi.org/10.1007/s10589-010-9329-3>

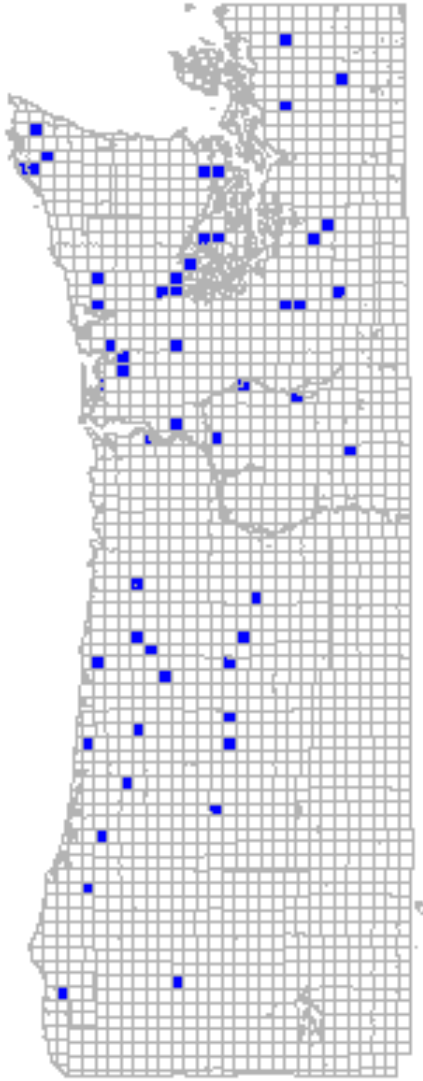
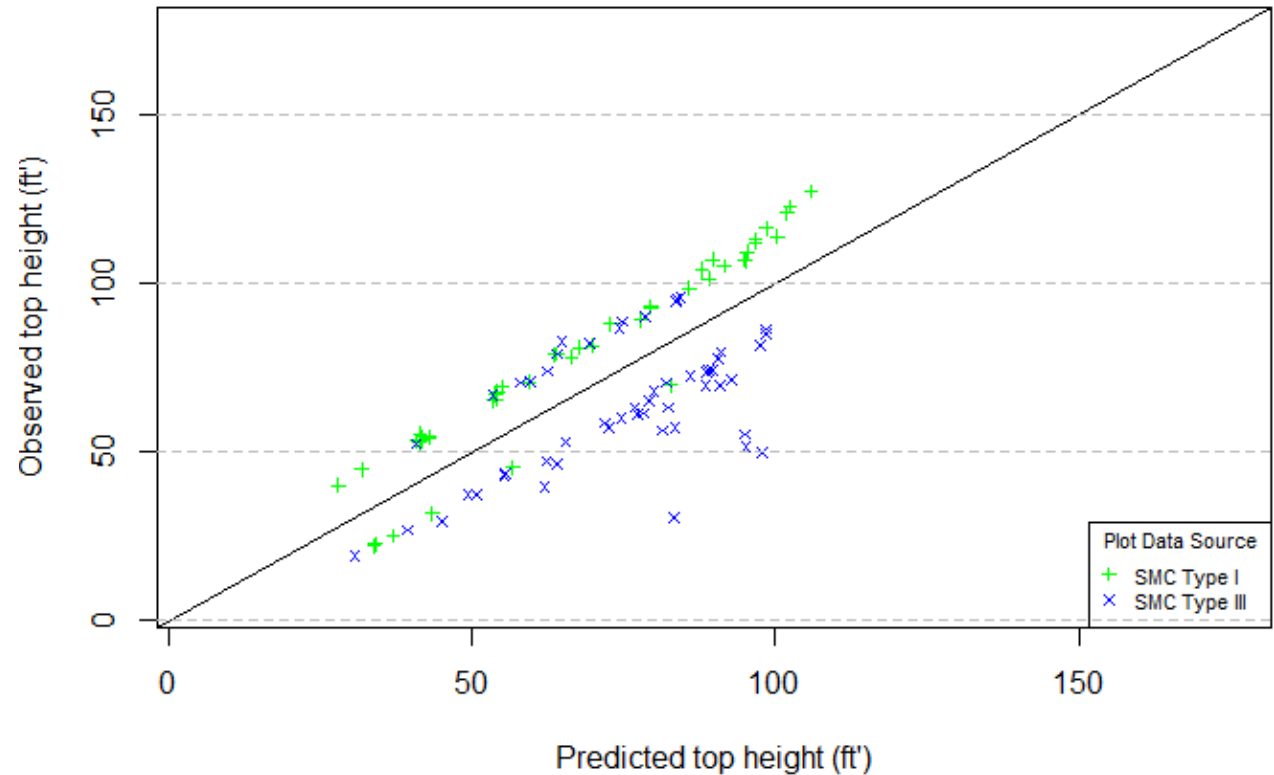
[2] Wessing, S. 2019. Proper initialization is crucial for the Nelder-Mead simplex search. *Optim Lett.* 13:847-856. <https://doi.org/10.1007/s11590-018-1284-4>



Major Findings

Predicted top outliers (residuals > 95%) by source

PS.R2: 95.3% | RMSE: 5.7' | MAE: 4.2' | MAPE: 10.7%



$$b_0 = f(\text{PLDEN2})$$

$$b_1 = f(\text{DD18_sm}, \text{RH_sp}, \text{CMI_wt}, \text{CMI_at}, \text{ELEV}, \text{PPT_sp}, \text{R4}, \text{R5}, \text{R6}, \text{OTHR}, \text{TSHE})$$

$$b_2 = f(\text{DD18_sm}, \text{CMI_at}, \text{ELEV}, \text{PPT_sp}, \text{R9}, \text{OTHR}, \text{TSHE}, \text{PLDEN})$$



Top Height & Density Prediction

Primary Species:

- Douglas-fir (202)
- Western hemlock (263)
- Hem-fir Mix

Planting Density:

Get HT40 & TPA



1. Automated tool for assessing variance across dataset.
2. Regional height model implemented in an online map covering western WA, OR:
 - Input: primary species; planting density
 - Output: polymorphic height/age curve; estimated density by age.



Company Benefits: Machine Learning

1. Initial variables that account for regional variation in height growth have been identified here. New variables can easily be compared to existing to determine their significance. This process can be semi- or fully-automated with the products developed here.
2. As new candidate predictors are identified, they can be dropped into the Nelder-Mead fitting process, resulting in one of three outcomes: insignificance; significance with improved performance; significance and displacement of another predictor. This process can be semi- or fully-automated with the products developed here.
3. This process is extensible: treatment flags for PCT, CT, fertilization
4. This process can be replicated for nearly any region
5. This process is being replicated to estimate volumetric yields



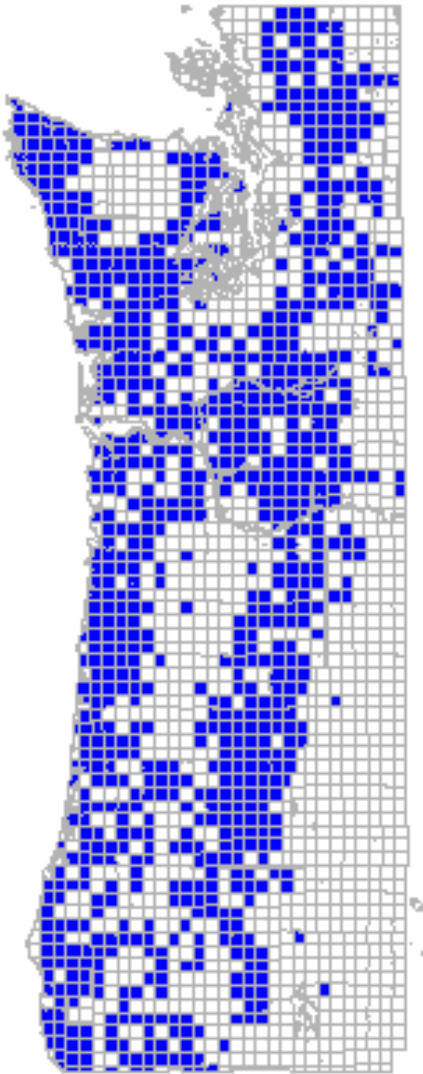
Recommendations

Narrow & deep data

1. Nelder-mead results dependent on initial simplex; continually refit with different simplex initialization tolerances to generate distribution of parameter estimates.
2. Deep time series on plots may allow for TPA to substitute for PLDEN, allowing for additional data (RFNRP).
3. Discrete physiographic regions remaining significant implies there are continuous predictors unaccounted for that would provide differentiation.

Shallow & wide data

1. Principal components analysis likely to identify more predictors that explain a similar amount of variation; and may require more bootstrap iterations to differentiate.
2. Lack of time series at same location may increase the importance of PLDEN in final model.
3. At risk for missing variables that weakly account for variation, but strongly predict.



Summary

1. Automated machine-learning methods for variable selection worked – most remained significant in the model; supplanted by other variables. No swings and misses.
2. Nelder-mead fitting procedure coded in FORTRAN was reliable, showed performance improvements over R-implementation. Still a computing-intensive process (continuous computer time since last CAFS meeting).
3. Resulting model performs well: prediction error $< 4.5'$; 95% of variation explained. Predictions on new data underpredict, would benefit from additional data if metrics can be resolved. This model would predict site at index age 30, which have much utility for early stand decisions – planting density, PCT, whether to CT.
4. Online map-based interface coming live to membership, working paper detailing methods.

