TCC-EFI Review, March 2022

CabinGIS: 0_FAIB-LiDAR_Caret_Predict_to_writeRasterOutput.R

Intro

The following summarizes the main data operations complete in the predictive ecosystem mapping of the Gaspard Operating Area. These operations include data processes undertaken in QGIS and R tools to explore, transform, cross-validate, and predict rasters of Whole Stem Volume using FAIB permanent sample plots and pre-processed 1m resolution LiDAR datasets as key inputs.

Methods: Import, process, and package FAIB sample plot data

Permanent sample plot data was imported into RStudio and subsetted to include predictor variables that matched expected spatial covariates along with target variable of whole stem volume per hectare (wsvha). Two predictor sets were prepared: model 1 was fitted with only seven predictors excluding stems_ha, while model 2 was fitted with all eight predictors. To test linear assumptions and assess predictor influence, predictor variables were fitted with univariate linear functions, and Wilcoxon normality test and Breush-Pagan test of constant variance reported.

n=5264	Mean (SD)	Max	Min	Wcx	Bp
elev	1001.1(166.1)	1750.0	653.0	13.9+6e****	10.572***
slope	10.4(10.5)	70.0	0.00	13.5+6e****	0.183
asp_cos	0.00(0.69)	1.00	-1.00	69.5+6e	125.251****
asp_sin	0.05(0.72)	1.00	-1.00	73.4+6e****	0.3913
lead_htop	19.97(5.86)	39.56	3.93	13.9+6e****	207.74*****
baha	27.87(13.81)	76.53	0.11	13.9+6e****	1389.05****
species	0.79(1.13)	5.00	0.00	—	—
stemsha	1563.1(1749.1)	30271.5	10.0	13.9+6e****	86.38****
wsvha	233.78(150.34)	801.34	0.00	13.7+6e****	—

Six predictors exhibited non-normal distributions with left-leaning skewness (Wcx, $p<0.001$) and five predictors produced non-constant variance against the response variable (Bp, $p<0.001$). Residuals showed increasing trends and funneling that suggested clustering around larger fitted values. In addition, significantly negative influences were observed by slope and stemsha on wsvha, while initial modeling generated negative value estimates confirming need for data normalizations (wsvha $<$ 0.00m3/ha) (Appendix). In response, modeled data was transformed with three caret functions. The 'center' method was used to subtract the mean of the predictors data from the predictor values, the 'scale' method divided them by their standard deviation, and a 'BoxCox' transformation applied an exponential lambda to positive values to coerce a Gaussian distribution.

Methods: Import LiDAR and VRI for terrain, stem/ha, species, and mask rasters

Point cloud processing operations and several algorithms were compared for ground classification, noise removal and height normalization, digital terrain and canopy modelling, individual tree detection and segmentation, and tree-based metrics. Classification of ground points improved with Cloth Simulation Filter

algorithm fitted with slope_smoothing tuning over a cloth resolution of 10cm and rigidness factor of 1. Noise removal was applied using the Statistical Outlier Removal algorithm with a default neighbour if k=10 and a multiple factor of 3. Terrain modelling and height normalization were derived using Inverse Distance Weighting with default paramaters, before tree metrics were derived including 95th percentile returns used in canopy modelling.

Results: Cross-validate, tune, re-model

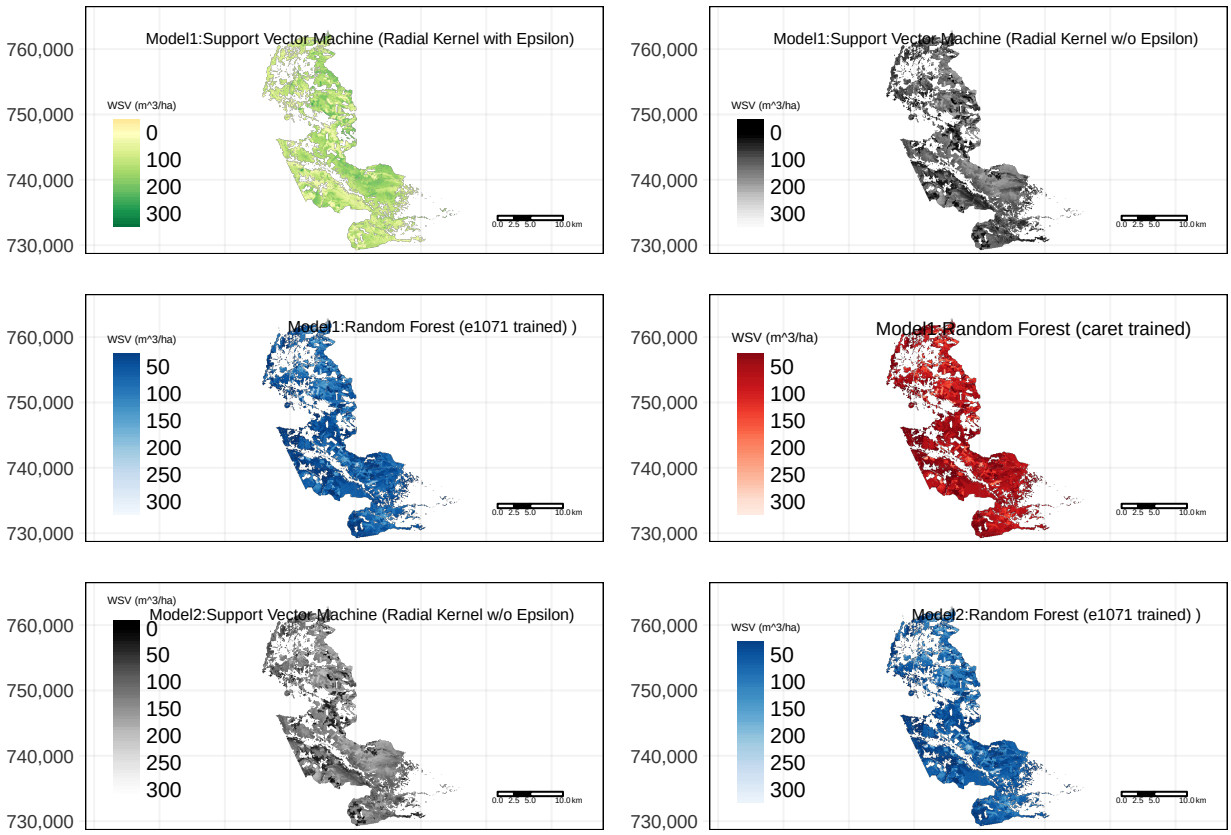
Pre-processed training data was fitted to model 1 and model 2 using nine different algorithms.1 To compare model performances, the sample plot data was split using a 80:20 test index measured against the response variable. Models were fitted with 80% of the data and were used to predict unseen data in the remaining 20% test set. Algorithms were trained using a 10k-fold cross validation technique. Performance metrics were reported for Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Root Mean Squared RMSEratio in order to account for overall accuracy, level of precision, and model bias, respectively.

	Full Model					Cross Validation		
Model1	Hyperparameter Tuning			MAE	RMSE	MAE	RMSE	$RMSE^{ratio}$
M1.svm.radial$^{\epsilon}$	ε = 0.02	C = 20	γ = 0.5	3.524	5.877	5.751	9.901	0.594
M1.svm.radial	ε = 0.10	C = 20	γ = 0.5	9.424	10.830	10.801	13.961	0.776
M1.svm.linear	ε = 0.10	C = 01	γ = 1.0	30.000	41.173	170.590	213.531	0.193
M1.glm.caret	ε = *1e-08*	C = 25		30.715	40.902	174.112	216.489	0.189
M1.RF.e1071	*Mtry* = 3	*Ntree* = 50		3.662	5.576	8.460	12.841	0.420
M1.RF.caret	*Mtry* = 3	*Ntree* = 500		3.388	5.178	171.073	215.336	0.024
M1.Epanech	*OV* = *4.35*	*Bws* = *aic*		0.226	0.434	176.299	219.720	0.002
M1.LocalConst	*OV* = *4.35*	*Bws* = *leastsq*		0.226	0.454	89.042	159.294	0.003
M1.ensemble	α = *0.10*	$\hat{\lambda}$ = *0.11031*		8.876	13.702			
Model2								
M2.svm.radial$^{\epsilon}$	ε = 0.02	C = 05	γ = 0.5	2.570	3.541	6.321	14.505	0.244
M1.svm.radial	ε = 0.10	C = 20	γ = 0.5	9.146	10.559	11.694	18.327	0.576
M2.svm.linear	ε = 0.10	C = 01	γ = 1.0	27.910	38.471	174.306	216.880	0.177
M2.glm.caret	ε = *1e-08*	C = 25		28.591	37.921	174.485	216.916	0.175
M2.RF.e1071	*Mtry* = 3	*Ntree* = 50		3.857	5.709	8.928	13.587	0.420
M2.RF.caret	*Mtry* = 3	*Ntree* = 500		3.537	5.271	175.490	218.750	0.024
M2.ensemble	α = 0.10	$\hat{\lambda}$ = *0.11025*		9.423	14.463			

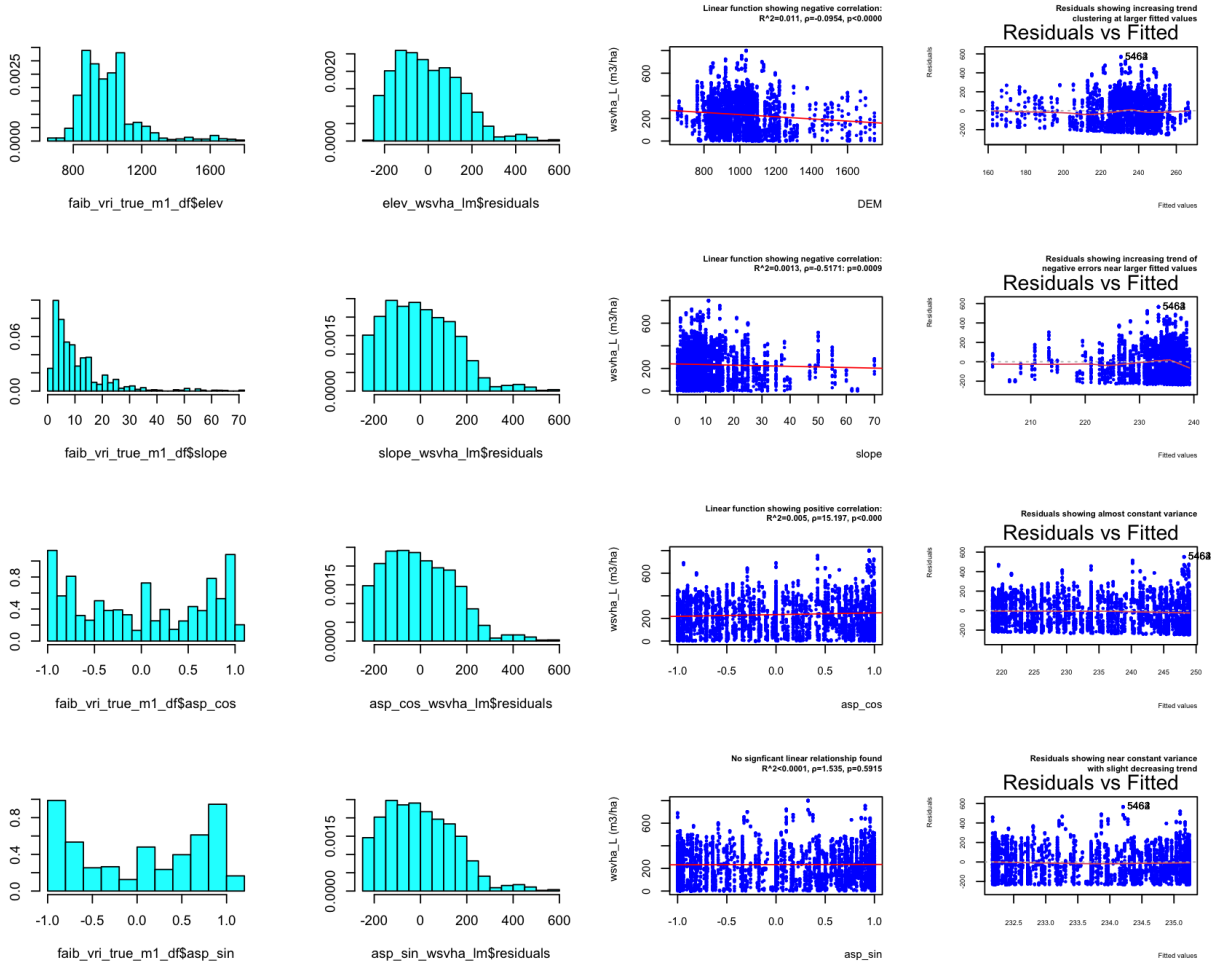
Results: Report and predict

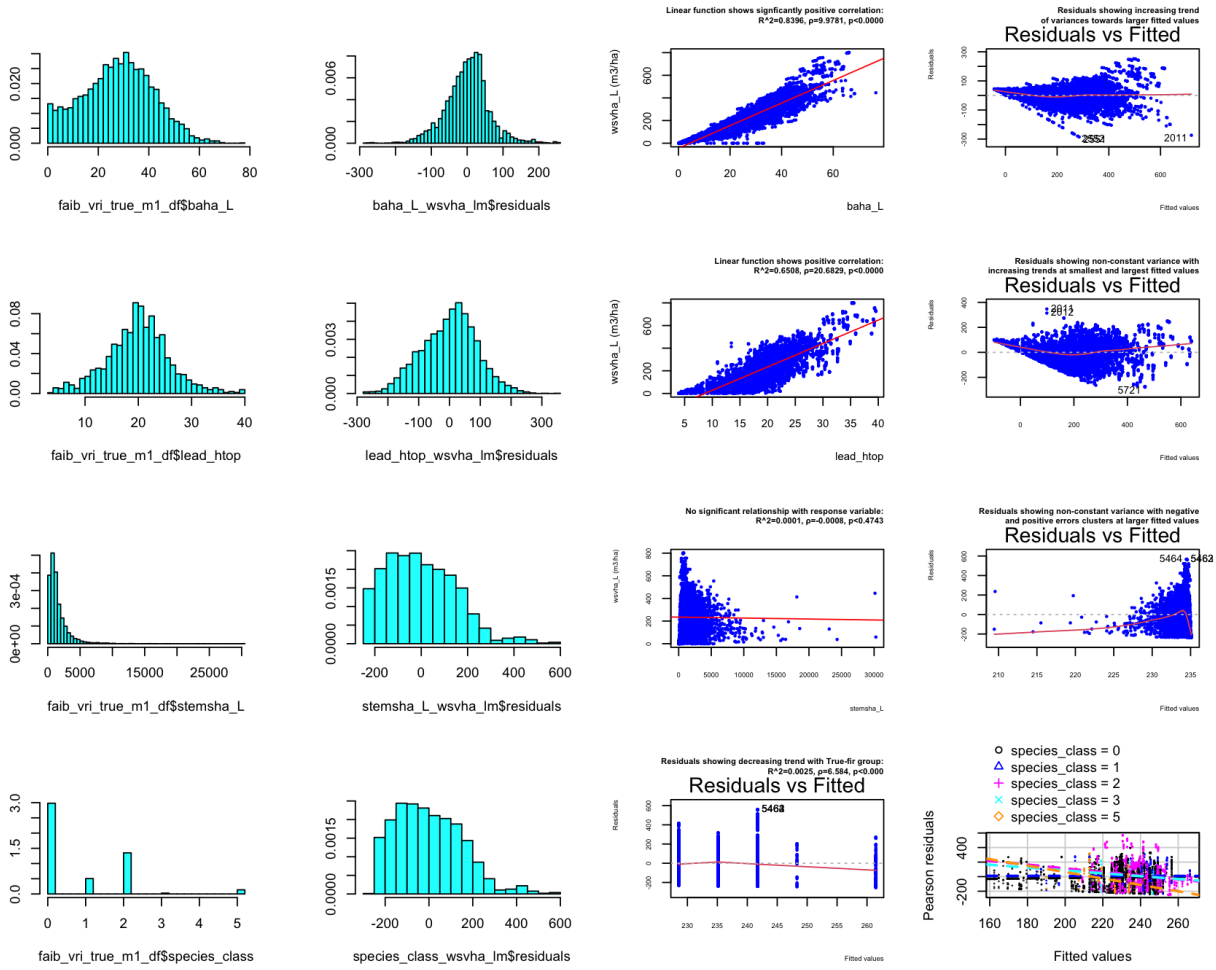
The two Support Vector Machine (SVM) algorithms, which were fitted with a radial and linear kernel, were calibrated using the same tuning grid that tested for optimal cost parameter values in the range of 1 and 20 and scanned for optimal gamma parameter values between -1 and +1. Tuning produced varied parameters between the two SVM algorithms, which may explain why the linear kernel produced such high bias (RMSEratio = 0.193). Performances greatly improved with SVM models when an epsilon buffer was added (0.02), though as model accuracy increased (RMSE=5.877m3/ha) so too did model bias (RMSEratio 0.594). Random Forest (RF) models were calibrated with two hyperparameters using a grid search of Mtry between 2 and 10 variables at each split over two regression trees consisting of 50 and 500 decision branches.

Results: Spatial Predictions of target raster: WSVA_L



Appendix


faib_vri_true_m1_df$elev
elev_wsvha_lm$residuals
Linear function showing negative correlation:
R^2=0.011, p=-0.0954, p<0.0000
wsvha_L (m3/ha)
DEM
Residuals showing increasing trend
clustering at larger fitted values
Residuals vs Fitted
Residuals
Fitted values
faib_vri_true_m1_df$slope
slope_wsvha_lm$residuals
Linear function showing negative correlation:
R^2=0.0013, p=-0.5171: p=0.0009
slope
Residuals showing increasing trend of
negative errors near larger fitted values
faib_vri_true_m1_df$asp_cos
asp_cos_wsvha_lm$residuals
Linear function showing positive correlation:
R^2=0.005, p=15.197, p<0.000
asp_cos
Residuals showing almost constant variance
faib_vri_true_m1_df$asp_sin
asp_sin_wsvha_lm$residuals
No signficant linear relationship found
R^2<0.0001, p=1.535, p=0.5915
asp_sin
Residuals showing near constant variance
with slight decreasing trend


faib_vri_true_m1_df$baha_L
baha_L_wsvha_lm$residuals
Linear function shows signficantly positive correlation:
R^2=0.8396, ρ=9.7781, p<0.0000
wsvha_L (m3/ha)
baha_L
Residuals showing increasing trend
of variances towards larger fitted values
Residuals vs Fitted
Residuals
Fitted values
faib_vri_true_m1_df$lead_htop
lead_htop_wsvha_lm$residuals
Linear function shows positive correlation:
R^2=0.6508, ρ=20.6829, p<0.0000
lead_htop
Residuals showing non-constant variance with
increasing trends at smallest and largest fitted values
faib_vri_true_m1_df$stemsha_L
stemsha_L_wsvha_lm$residuals
No significant relationship with response variable:
R^2=0.0001, ρ=-0.0008, p<0.4743
stemsha_L
Residuals showing non-constant variance with negative
and positive errors clusters at larger fitted values
faib_vri_true_m1_df$species_class
species_class_wsvha_lm$residuals
Residuals showing decreasing trend with True-fir group:
R^2=0.0025, ρ=6.584, p<0.000
species_class = 0
species_class = 1
species_class = 2
species_class = 3
species_class = 5
Pearson residuals