

Supplementary Information

Quantification of soil carbon and moisture content in varying-moisture soils with Laser-Induced Breakdown Spectroscopy and machine learning

Paavo Heikkilä¹, Wen Xu¹, Jussi Heinonsalo^{2,3}, and Jan Viljanen^{1,4,*}

¹Cense Analytics Oy, Tampere, 33720, Finland

²Department of Forest Sciences, Faculty of Agriculture and Forestry, University of Helsinki, Helsinki, Finland

³Institute for Atmospheric and Earth System Sciences (INAR), University of Helsinki, Finland

⁴Department of Physics and Technology, UiT The Arctic University of Norway, NO-9037 Tromsø, Norway

*Correspondence: jan.viljanen@uit.no; paavo.kajovirta@cense-analytics.com

Table S1. Total carbon prediction performance by calibration scenario (rows) and model (KR, SVR, PLS, NN, XGB, RF). Metrics are coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE) on the test set (% TC). Each metric is presented for both direct TC prediction (TC) and MC as a feature prediction (MC→TC). Groups **A** and **B**: pooled results over five train–test splits ($n = 5 \times 23$); **C** and **D**: $n = 69$; **E**: $n = 138$.

Total Carbon		R^2		MAE		RMSE	
		TC	MC→TC	TC	MC→TC	TC	MC→TC
A — Ball-milled constant $\approx 15\%$ MC Pooled 5 splits Test $n = 5 \times 23$	KR	0.95	0.95	0.19	0.20	0.27	0.27
	SVR	0.94	0.95	0.22	0.22	0.30	0.29
	PLS	0.95	0.94	0.22	0.23	0.28	0.30
	NN	0.92	0.90	0.26	0.27	0.35	0.40
	XGB	0.93	0.91	0.23	0.25	0.34	0.37
	RF	0.89	0.89	0.28	0.28	0.41	0.41
B — Fresh (freezer) natural humidity Pooled 5 splits Test $n = 5 \times 23$	KR	0.86	0.91	0.36	0.30	0.47	0.39
	SVR	0.81	0.86	0.39	0.34	0.55	0.46
	PLS	0.86	0.83	0.37	0.40	0.47	0.51
	NN	0.83	0.87	0.38	0.34	0.52	0.45
	XGB	0.83	0.82	0.39	0.38	0.52	0.53
	RF	0.73	0.76	0.44	0.41	0.65	0.62
C — Ball-milled lab-controlled varying humidity Test $n = 69$	KR	0.83	0.86	0.34	0.31	0.49	0.44
	SVR	0.85	0.85	0.33	0.33	0.46	0.46
	PLS	0.82	0.83	0.36	0.36	0.50	0.48
	NN	0.67	0.78	0.33	0.36	0.67	0.54
	XGB	0.68	0.75	0.38	0.35	0.66	0.58
	RF	0.68	0.69	0.39	0.38	0.66	0.65
D — Non-milled lab-controlled varying humidity Test $n = 69$	KR	0.86	0.86	0.31	0.31	0.43	0.43
	SVR	0.88	0.86	0.31	0.33	0.41	0.43
	PLS	0.87	0.87	0.33	0.33	0.42	0.42
	NN	0.74	0.80	0.41	0.35	0.60	0.52
	XGB	0.76	0.77	0.40	0.39	0.57	0.56
	RF	0.75	0.74	0.43	0.45	0.58	0.60
E — All humidity states pooled Test $n = 138$	KR	0.92	0.92	0.27	0.26	0.36	0.34
	SVR	0.90	0.91	0.31	0.29	0.40	0.37
	PLS	0.89	0.91	0.32	0.30	0.41	0.38
	NN	0.87	0.89	0.31	0.29	0.45	0.42
	XGB	0.80	0.84	0.39	0.36	0.56	0.50
	RF	0.79	0.77	0.41	0.42	0.57	0.60

Table S2. Moisture content prediction performance by calibration scenario and model. Metrics are R^2 , MAE, and RMSE on the test set (% MC, gravimetric $m_{\text{H}_2\text{O}}/m_{\text{dry}}$). Scenario **B**: mean over five splits ($n = 5 \times 23$); **C** and **D**: $n = 69$; **E**: $n = 138$. Scenario **A** (constant-humidity milled samples) is omitted because MC varies only within the lab-controlled sets **C–E**.

Moisture content		R ²	MAE	RMSE
B — Fresh (freezer) natural humidity Mean of 5 splits Test n = 5×23	KR	0.70	2.1	3.1
	SVR	0.49	2.8	4.1
	PLS	0.45	3.0	4.2
	NN	0.26	3.6	4.9
	XGB	0.55	2.9	3.8
	RF	0.53	3.0	3.9
C — Ball-milled lab-controlled varying humidity Test n = 69	KR	0.93	1.6	2.0
	SVR	0.88	1.9	2.7
	PLS	0.86	2.1	2.8
	NN	0.78	2.7	3.5
	XGB	0.84	2.2	3.0
	RF	0.83	2.3	3.1
D — Non-milled lab-controlled varying humidity Test n = 69	KR	0.93	1.6	2.4
	SVR	0.89	2.0	2.9
	PLS	0.88	2.0	3.0
	NN	0.82	2.4	3.7
	XGB	0.87	2.0	3.2
	RF	0.85	2.2	3.4
E — All humidity states pooled Test n = 138	KR	0.95	1.4	1.8
	SVR	0.94	1.5	2.1
	PLS	0.91	1.8	2.4
	NN	0.89	2.2	2.7
	XGB	0.88	2.0	2.9
	RF	0.85	2.2	3.2

Table S3. Mean absolute Kernel ridge residual (% TC) by calibration scenario (rows) and soil depth interval (columns). Scenario **A** uses a TC-only model; **B–E** use MC→TC models. Cell entries are MAE (% TC) with n in parentheses. Bold marks the largest residual within each scenario row.

Calibration scenario	0–10 cm	0–20 cm	10–20 cm	20–30 cm	30–40 cm
A — Ball-milled constant ≈ 15 % MC	0.26 (23)	0.19 (20)	0.19 (21)	0.19 (21)	0.16 (30)
B — Fresh (freezer) natural humidity	0.34 (23)	0.25 (20)	0.24 (21)	0.30 (21)	0.34 (30)
C — Ball-milled lab-controlled varying humidity	0.22 (18)	0.44 (13)	0.32 (11)	0.40 (13)	0.22 (14)
D — Non-milled lab-controlled varying humidity	0.24 (18)	0.29 (13)	0.28 (11)	0.44 (13)	0.33 (14)
E — All humidity states pooled	0.34 (33)	0.26 (20)	0.28 (27)	0.23 (23)	0.19 (35)
Weighted mean (n -weighted across A–E)	0.29 (115)	0.27 (86)	0.26 (91)	0.29 (91)	0.24 (123)

Table S4. Optuna hyperparameter search ranges by model. Ranges apply to all calibration scenarios and targets. p denotes the number of input features in a fold; n_k the number of training samples in a fold; p_{total} the full feature count. For the neural network (NN), each additional hidden layer uses half the neuron count of the preceding layer. The number of retained features (n_{features}) was optimized jointly for all models.

Model	Parameter	Search range	Scale	Condition
KR	kernel	linear, poly, rbf, sigmoid	categorical	—
KR	α	$[10^{-4}, 10^2]$	log-uniform	—
KR	γ	$[10^{-6}, 10]$	log-uniform	kernel $\in \{\text{rbf, poly, sigmoid}\}$
KR	degree	$[2, 5]$	integer	kernel = poly
KR	coef0	$[0, 1]$	uniform	kernel $\in \{\text{poly, sigmoid}\}$
SVR	kernel	linear, poly, rbf, sigmoid	categorical	—
SVR	C	$[10^{-5}, 10^2]$	log-uniform	—
SVR	ε	$[0.01, 0.5]$	uniform	—
SVR	γ	$[10^{-6}, 10]$	log-uniform	kernel $\in \{\text{rbf, poly, sigmoid}\}$
SVR	degree	$[2, 5]$	integer	kernel = poly
SVR	coef0	$[0, 1]$	uniform	kernel $\in \{\text{poly, sigmoid}\}$
PLS	n_components	$[2, \min(50, p, n_k - 1)]$	integer	—
PLS	scale	True, False	categorical	—
RF	n_estimators	$[50, 500]$, step 50	integer	—
RF	max_depth	$[3, 30]$	integer	—
RF	min_samples_split	$[2, 20]$	integer	—
RF	min_samples_leaf	$[1, 10]$	integer	—
RF	max_features	$[0.1, 1.0]$	uniform	—
XGB	n_estimators	$[50, 500]$, step 50	integer	—
XGB	max_depth	$[3, 10]$	integer	—
XGB	learning_rate	$[0.01, 0.3]$	log-uniform	—
XGB	subsample	$[0.5, 1.0]$	uniform	—
XGB	colsample_bytree	$[0.3, 1.0]$	uniform	—
XGB	min_child_weight	$[1, 10]$	integer	—
XGB	reg_alpha	$[10^{-8}, 10]$	log-uniform	—
XGB	reg_lambda	$[10^{-8}, 10]$	log-uniform	—
NN	n_hidden_layers	$[2, 6]$	integer	halving layer widths
NN	n_neurons	$[32, 128]$, step 16	integer	—
NN	activation	relu, tanh, elu	categorical	—
NN	learning_rate	$[10^{-5}, 10^{-2}]$	log-uniform	—
NN	batch_size	$[8, 64]$, step 8	integer	—
NN	epochs	$[100, 300]$, step 50	integer	—
All	n_features	$[5, p_{\text{total}}]$	integer	—

Table S5. Cross-validation R^2 during hyperparameter tuning. Format: mean $\pm$ std_{splits} (std_{folds}) for scenarios **A** and **B**, and mean (std_{folds}) for **C–E**, which use a single train–test split. Each trial was scored by 5-fold cross-validation repeated five times. std_{splits} is the standard deviation across five random train–test splits; std_{folds} is the mean within-trial fold variability. Columns give R^2_{CV} for TC-only and MC→TC targets.

Calibration scenario		R^2_{CV}	
		TC	MC→TC
A — Ball-milled constant $\approx 15\%$ MC Pooled 5 splits Test $n = 5 \times 23$	KR	0.929 $\pm$ 0.011 (0.040)	0.931 $\pm$ 0.011 (0.040)
	SVR	0.926 $\pm$ 0.007 (0.039)	0.927 $\pm$ 0.007 (0.039)
	PLS	0.917 $\pm$ 0.009 (0.041)	0.917 $\pm$ 0.009 (0.044)
	NN	0.861 $\pm$ 0.024 (0.076)	0.858 $\pm$ 0.032 (0.091)
	XGB	0.898 $\pm$ 0.012 (0.056)	0.898 $\pm$ 0.013 (0.058)
	RF	0.875 $\pm$ 0.009 (0.065)	0.876 $\pm$ 0.009 (0.065)
B — Fresh (freezer) natural humidity Pooled 5 splits Test $n = 5 \times 23$	KR	0.840 $\pm$ 0.026 (0.074)	0.873 $\pm$ 0.031 (0.059)
	SVR	0.838 $\pm$ 0.027 (0.076)	0.867 $\pm$ 0.025 (0.055)
	PLS	0.822 $\pm$ 0.024 (0.086)	0.833 $\pm$ 0.027 (0.077)
	NN	0.789 $\pm$ 0.027 (0.120)	0.791 $\pm$ 0.032 (0.127)
	XGB	0.765 $\pm$ 0.023 (0.102)	0.788 $\pm$ 0.024 (0.098)
	RF	0.697 $\pm$ 0.018 (0.129)	0.712 $\pm$ 0.019 (0.123)
C — Ball-milled lab-controlled varying humidity Test $n = 69$	KR	0.884 (0.034)	0.903 (0.026)
	SVR	0.887 (0.037)	0.898 (0.031)
	PLS	0.889 (0.031)	0.897 (0.032)
	NN	0.811 (0.057)	0.823 (0.061)
	XGB	0.782 (0.073)	0.798 (0.062)
	RF	0.735 (0.075)	0.737 (0.074)
D — Non-milled lab-controlled varying humidity Test $n = 69$	KR	0.853 (0.037)	0.866 (0.031)
	SVR	0.842 (0.036)	0.853 (0.037)
	PLS	0.839 (0.036)	0.848 (0.031)
	NN	0.803 (0.065)	0.845 (0.043)
	XGB	0.789 (0.064)	0.804 (0.061)
	RF	0.733 (0.079)	0.735 (0.078)
E — All humidity states pooled Test $n = 138$	KR	0.879 (0.027)	0.889 (0.024)
	SVR	0.866 (0.034)	0.871 (0.031)
	PLS	0.858 (0.030)	0.867 (0.033)
	NN	0.845 (0.036)	0.864 (0.034)
	XGB	0.795 (0.048)	0.821 (0.039)
	RF	0.761 (0.054)	0.758 (0.053)