Submitted to Ecosphere

Using machine learning on tree-ring data to determine the geographical provenance of historical construction timbers

Eileen Kuhl, Christian Zang, Jan Esper, Dana F. C. Riechelmann, Ulf Büntgen, Martin Briesch, Frederick Reinig, Philipp Römer, Oliver Konter, Martin Schmidhalter, and Claudia Hartl

Corresponding author: Eileen Kuhl (eikuhl@uni-mainz.de)

Appendix S1:

Table S1 Descriptive statistics of the living tree sites and the historical dataset.

Code	Elevation [m asl]	Exposure	MSL[1]	First year (>5)	Last year	Range[2]	AR1[3] TRW[5]	AR1[3] MXD[6]	Rbar[4] TRW[5]	Rbar[4] MXD[6]
S14	1400	south	73	1884 (1928)	2011	128	0.68	0.5	0.7	0.37
N16	1600	north	134	1860 (1870)	2011	152	0.83	0.44	0.64	0.43
SN17	1700	south & north	133	1679 (1854)	2011	333	0.77	0.44	0.56	0.49
N19	1900	north	205	1582 (1685)	2010	429	0.67	0.54	0.64	0.56
S20	2000	south	213	1672 (1714)	2010	339	0.77	0.61	0.7	0.62
S22	2200	south	218	1542 (1717)	2009	468	0.7	0.52	0.71	0.64
Historical	**unknown**	**unknown**	115	742 (790)	2002	1261	0.75	0.51	0.47	0.31

[1] mean series length

[2] length of chronology

[3] first-order autocorrelation

[4] inter-series correlation

[5] tree-ring width

[6] maximum latewood density

Table S2 Mean values of tree-ring proxies.

Code	Elevation [m asl]	AGR[1]	Mean MXD[2]	Mean EWW[3]	Mean LWW[4]	Mean EWD[5]	Mean LWD[6]
S14	**1400**	1.58	1.02	1.08	0.49	0.35	0.91
N16	**1600**	0.94	1.02	0.66	0.29	0.34	0.91
SN17	**1700**	0.91	0.98	0.62	0.29	0.36	0.87
N19	**1900**	0.97	0.9	0.67	0.3	0.34	0.79
S20	**2000**	0.72	0.86	0.52	0.21	0.34	0.76
S22	**2200**	0.72	0.84	0.51	0.21	0.36	0.73
Historical	**unknown**	1.16	0.99	0.84	0.32	0.37	0.88

[1] average growth rate

[2] maximum latewood density

[3] earlywood width

[4] latewood width

[5] earlywood density

[6] latewood density

Table S3 Hyperparameters of the fine-tuned models.

ML Algorithm	**Hyperparameter**
Ridge Regression Classifier	{'alpha': 0.3, 'solver': 'cholesky'}
Logistic Regression Classifier (Softmax Regression)	{'multi_class': 'multinomial', 'solver': 'newton-cg', 'C': 15, 'penalty': 'l2', 'max_iter': 9000000}
Gaussian Naïve Bayes	{'var_smoothing': 8.111308307896856e-09}
Random Forest	{'criterion': 'entropy','max_depth': 20, 'min_impurity_decrease': 0.0001, 'min_sample_split': 5, 'n_estimators': 75}
Extreme Gradient Boosting (XGBoost)	{'eval_metric': 'mlogloss', 'eta': 0.3, 'max_depth': 5, 'subsample': 1, 'n_estimators': 250, 'colsample_by_tree': 1, 'gamma': 0.1, 'min_child_weight': 1, earlystopping}
k-Nearest Neighbor	{'algorithm': 'kd_tree', 'leaf_size': 1, 'metric': 'euclidean', 'n_neighbors': 1, 'weights': 'uniform'}
Support Vector Machine	{'C': 30, 'gamma': 0.005}
Stochastic Gradient Decent	{'alpha': 0.1, 'learning_rate': 'optimal', 'loss': 'squared_hinge', 'penalty': 'l1'}
Linear Discriminant Analysis	{'shrinkage': 1, 'solver': 'lsqr'}

Table S4 Tested fine-tuned machine learning algorithms with f1 score means and standard deviations (std) of repeated stratified k-fold (k = 10, repeats = 100) cross-validation.

ML Algorithm	**f1 score mean**	**f1 score std**
Ridge Regression Classifier	0.58	0.12
Logistic Regression Classifier (Softmax Regression)	0.57	0.13
Gaussian Naïve Bayes	0.59	0.12
Random Forest	0.68	0.12
Extreme Gradient Boosting (XGBoost)	0.70	0.12
k-Nearest Neighbor	0.68	0.13
Support Vector Machine	0.61	0.14
Stochastic Gradient Decent	0.16	0.08
Linear Discriminant Analysis	0.35	0.08

Table S5 Classification Report: performance metrics of D_{test}[1] on XGBoost in each individual class for DM_{gen}[2] and RWM_{gen}[3] (in brackets).

Site	**Precision**	**Recall**	**f1 score**
S14	0.63 (0.5)	1 (0.8)	0.77 (0.62)
N16	1 (0.67)	0.6 (0.4)	0.75 (0.5)
SN17	0.67 (0.0)	0.4 (0.0)	0.5 (0.0)
N19	0.38 (0.17)	0.6 (0.2)	0.46 (0.18)
S20	1 (0.0)	0.5 (0.0)	0.67 (0.0)
S22	1 (0.2)	1 (0.2)	1 (0.2)
Average	0.78 (0.26)	0.68 (0.27)	0.69 (0.25)

[1] test dataset

[2] density and ring-width parameter model unspecified and general

[3] ring-width parameter model unspecified and general

(a)	Precision	Recall	f1score
(b): Akhmetzyianov et al. 2019/2020 (LWD)	0.4	0.22	0.28
(c): Akhmetzyianov et al. 2019/2020 (TRW)	0.25	0.13	0.16
(d): Wilson, Esper, and Luckman (2004)	0.28	0.24	0.25

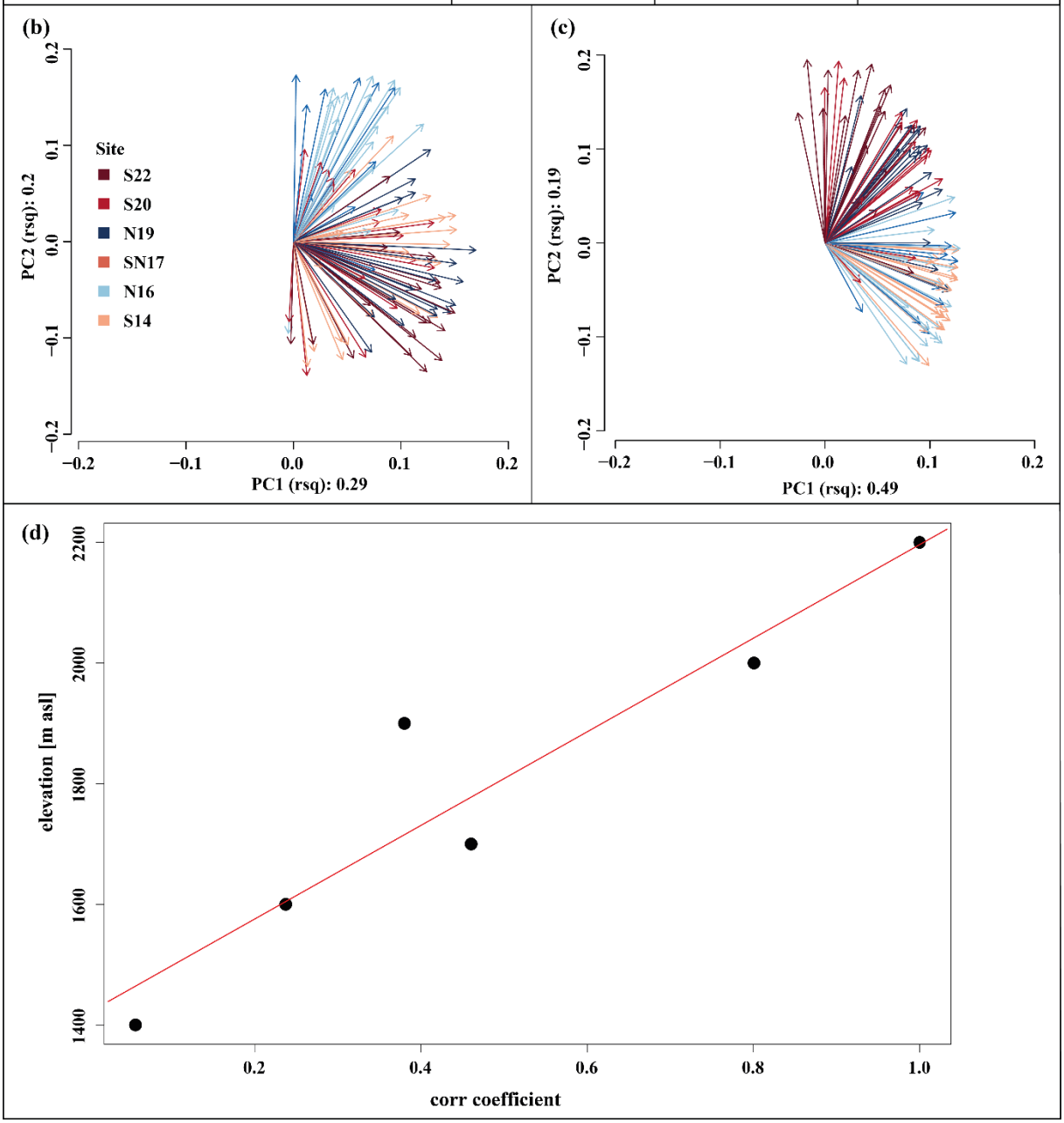

Figure S1 Other tested approaches (a), results from the PCGA (principal component gradient analysis) with latewood density (b) and tree-ring width (c) and a linear regression model approach (d). Although ring width in (c) seems to be better separated, the density PCGA provenancing, (b) works better for provenancing (see (a)), which is in line with findings from Akhmetzyanov et al. (2020).

List of references

Akhmetzyanov, L., A. Buras, U. Sass-Klaassen, J. den Ouden, F. Mohren, P. Groenendijk, and I. García-González. 2019. "Multi-variable Approach Pinpoints Origin of Oak Wood with Higher Precision." Journal of Biogeography 46 (6): 1163–77. https://doi.org/10.1111/jbi.13576.

Akhmetzyanov, L., R. Sánchez-Salguero, I. García-González, A. Buras, Marta Dominguez-Delmás, F. Mohren, J. den Ouden, and U. Sass-Klaassen. 2020. "Towards a New Approach for Dendroprovenancing Pines in the Mediterranean Iberian Peninsula." Dendrochronologia 60: 125688. https://doi.org/10.1016/j.dendro.2020.125688.

Wilson, R. J. S., J. Esper, and B. H. Luckman. 2004. "Utilising Historical Tree-Ring Data for Dendroclimatology: A Case Study from the Bavarian Forest, Germany." Dendrochronologia 21 (2): 53–68. https://doi.org/10.1078/1125-7865-00041.