MD-Dating

The statistical theory behind

Bernhard SPANGL, BOKU University, Institute of Statistics, Vienna, Austria
Johannes TINTNER, BOKU University, Institute of Physics and Material Sciences, Vienna, Austria
Michael MELCHER, BOKU University, Institute of Statistics, Vienna, Austria

Keywords: *Statistical learning — Random forest — nonlinear calibration*

CHNT Reference: B. Spangl, J. Tintner, M. Melcher. 2020. MD-Dating – The statistical theory behind. W. Börner and CHNT Organization Committee. Proceedings of the 25th International Conference on Cultural Heritage and New Technologies. DOI:xxxxxxx.

Motivation and Background

The objective of this talk is to establish a dating tool for wooden artefacts under specified storage conditions based on the relation between chemical characteristics of the decay and time. The combined biotic and abiotic chemical decay is revealed by means of infrared spectroscopy. From an analytical point of view, infrared spectroscopy emerged as a powerful tool.

The new method of MD-dating is based upon the molecular decay. This decay follows in many cases a monotonous function. Therefore, this decay function can be calibrated with a set of artefacts of known age and prediction models can be established. There are several options that can be used for the establishment of such prediction models. While PLS (Partial Least Squares) regression models are state-of-the-art to analyze such data, current publications have chosen an alternative modeling approach for data of this kind, namely the random forest method (Tintner et al., 2020a,b). Dendrochronology serves as the reference method.

The talk will present the models for wood, especially for Scots pine wood, with a keen focus on the statistical modelling.

Statistical Methodology

Using spectral data and dendrochronology as a reference, a random forest model for the prediction of age was created. In the field of machine learning, tree-based methods for regression are well established (Hastie et al., 2017). Although tree-based methods are simple and useful for interpretation they lack prediction accuracy, i.e., they produce good predictions on the training set, but are likely to overfit the data, leading to poor test set performance. The random forest method was introduced by Breiman (2001) to overcome these drawbacks. Like bagging, this approach generates multiple trees which are then combined to yield a single consensus prediction. Combining a large number of trees reduces the variance and will often result in dramatic improvements in prediction accuracy.

As the predicted values y_i of the random forest model always underestimated the true year x_i for all species, especially for very old probes, the final models were subjected to a further calibration step

to improve the prediction quality. As calibration model we use the following extended exponential model:

$$y = h(x;\theta) = h(x;b_0,b_1,\delta) \tag{1}$$

$$\ldots = b_0 + b_1 \cdot 2^{(-(a-x)/\delta)}, \tag{2}$$

where b_0, b_1, and δ are the parameters to be estimated and a is set to 2017, i.e., a-x corresponds to the age of the tree ring. Hence, the dimension of the parameter vector θ is $p = 3$.

Analogous to linear regression and with asymptotic approximation we can specify confidence intervals for the function values $h(x_0; \hat{\theta})$ for the nonlinear function h at x_0. The confidence interval for the function value $\hat{\eta}_0 := h(x_0; \hat{\theta})$ is then approximately given by

$$\hat{\eta}_0 \pm q_{1-\alpha/2}^{t_{n-p}} \cdot s.e.(\hat{\eta}_0) \ , \tag{3}$$

with

$$s.e.(\hat{\eta}_0) = \hat{\sigma}\sqrt{\hat{a}_0^{\top}\left(A(\hat{\theta})^{\top} A(\hat{\theta})\right)^{-1}\hat{a}_0}, \tag{4}$$

and $q_{1-\alpha/2}^{t_{n-p}}$ is the $(1 - \alpha/2)$ quantile of the t-distribution with $n - p$ degrees of freedom. The residual standard deviation $\hat{\sigma}$ is given as the square root of

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \eta_i(\hat{\theta}))^2 \ , \tag{5}$$

where $\eta_i(\hat{\theta}) := h(x_i; \hat{\theta})$ and $i = 1, \ldots, n$, with n the number of observations, and $j = 1, \ldots, p$.

Here, A is the $n \times p$ matrix of partial derivatives with elements

$$A_i^{(j)}(\theta) := \frac{\partial \eta_i(\theta)}{\partial \theta_j} \ . \tag{6}$$

And, a_0 is the linear approximation at x_0, i.e.,

$$a_0 = \frac{\partial h(x_0;\theta)}{\partial \theta} \ . \tag{7}$$

The unknown values are replaced by their estimates to obtain $\hat{a}_0$.

Furthermore, an at least approximate $(1 - \alpha/2)$ forecast interval is given by

$$\hat{\eta}_0 \pm q_{1-\alpha/2}^{t_{n-p}} \cdot \sqrt{\hat{\sigma}^2 + (s.e.(\hat{\eta}_0))^2} \ . \tag{8}$$

The calibration forecast interval for a certain year x_0 can be now defined as the set

$$\{x : \|y_0 - h(x;\hat{\theta})\| \leq q_{1-\alpha/2}^{t_{n-p}} \cdot \sqrt{\hat{\sigma}^2 + (s.e.(h(x;\hat{\theta})))^2}\} \ , \tag{9}$$

where y_0 is the fitted value of the calibration model at x_0. A schematic representation of how a calibration interval is determined is given in Fig. 1. At the points y_0 = 1500 and y_0 = 1800 the resulting intervals are [1367, 1536] and [1759, 1884] respectively.

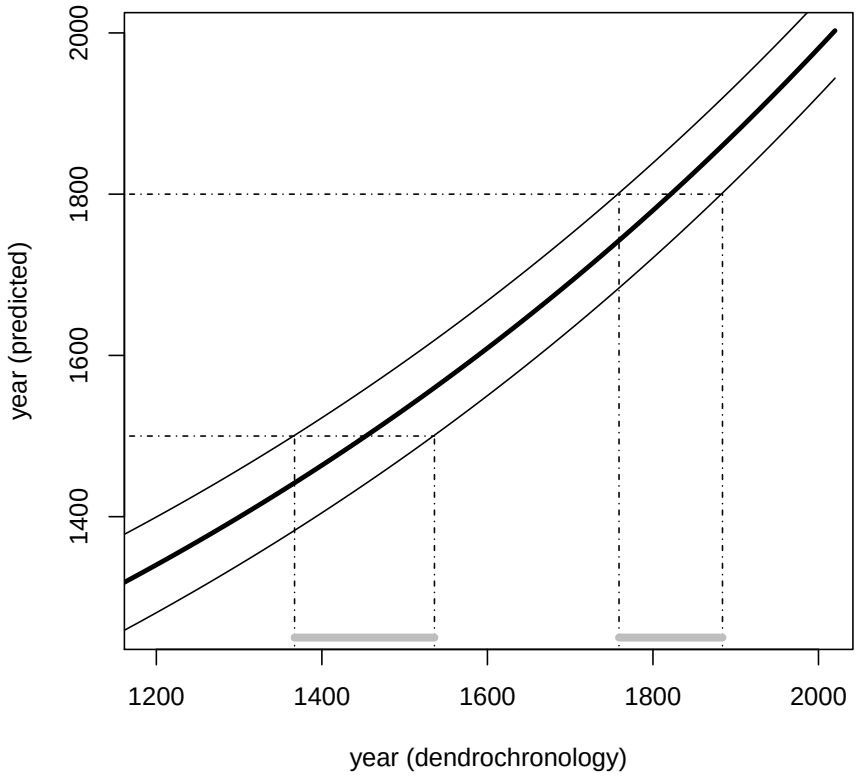

Fig. 1. Schematic representation of calibration intervals.

A calibration confidence interval can be defined in a similar way. Further details may be found in Bates and Watts (1988).

This whole procedure is referred to as MD-Dating model.

Discussion and Outlook

The current models for wooden artefacts demonstrate the potential of molecular decay for dating purposes. Moreover, random forest models applied to infrared spectroscopic measurements result in proper prediction tools. The quality of these models was checked by a cross-validated root mean squared error of prediction (RMSEP). However, as several measurements were usually available for a single wooden artefact the standard procedure of cross-validation has to be adapted accordingly. Hence, all MD-dating models were validated by tree-wise cross-validation. Additionally, it has to be stated explicitly that models are valid only for the preservation conditions covered by the sample set. Future work will test additional statistical learning algorithms like artifical neural networks (ANN) or generalized additive models (GAM). The combination of infrared

spectroscopic measurements and statistical prediction models is a promising approach to stimulate and support the work of building historians, archaeologists, and even environmental scientists.

References

Bates, D. M., Watts, D. G. (1988), 'Nonlinear Regression Analysis and Its Applications', Hoboken, NJ, USA: John Wiley & Sons.

Breiman, L. (2001), 'Random forests', Machine Learning, 45, pp. 5-32, DOI: 10.1023/A:1010933404324

Hastie, T., Tibshirani, R., Friedman, J. H. (2017), 'The elements of statistical learning: Data mining, inference, and prediction', Second edition, New York, NY, USA: Springer.

Tintner, J., Spangl, B., Reiter, F., Smidt, E., Grabner, M. (2020a), 'Infrared spectral characterization of the molecular wood decay in terms of age', Wood Science and Technolgy, 54, pp. 313-327, DOI: 10.1007/s00226-020-01160-x

Tintner, J., Spangl, B., Grabner, M., Helama, S., Timonen, M., Kirchhefer, A. J., Reinig, F., Nievergelt, D., Krąpiec, M., Smidt, E. (2020b), 'MD dating: molecular decay in pinewood as a dating method', Scientific Reports, 10, 11255, DOI: 10.1038/s41598-020-68194-w