University of Exeter

Advisors

Dr. Johan Wahlström
Dr. David Reynolds

STUDENT

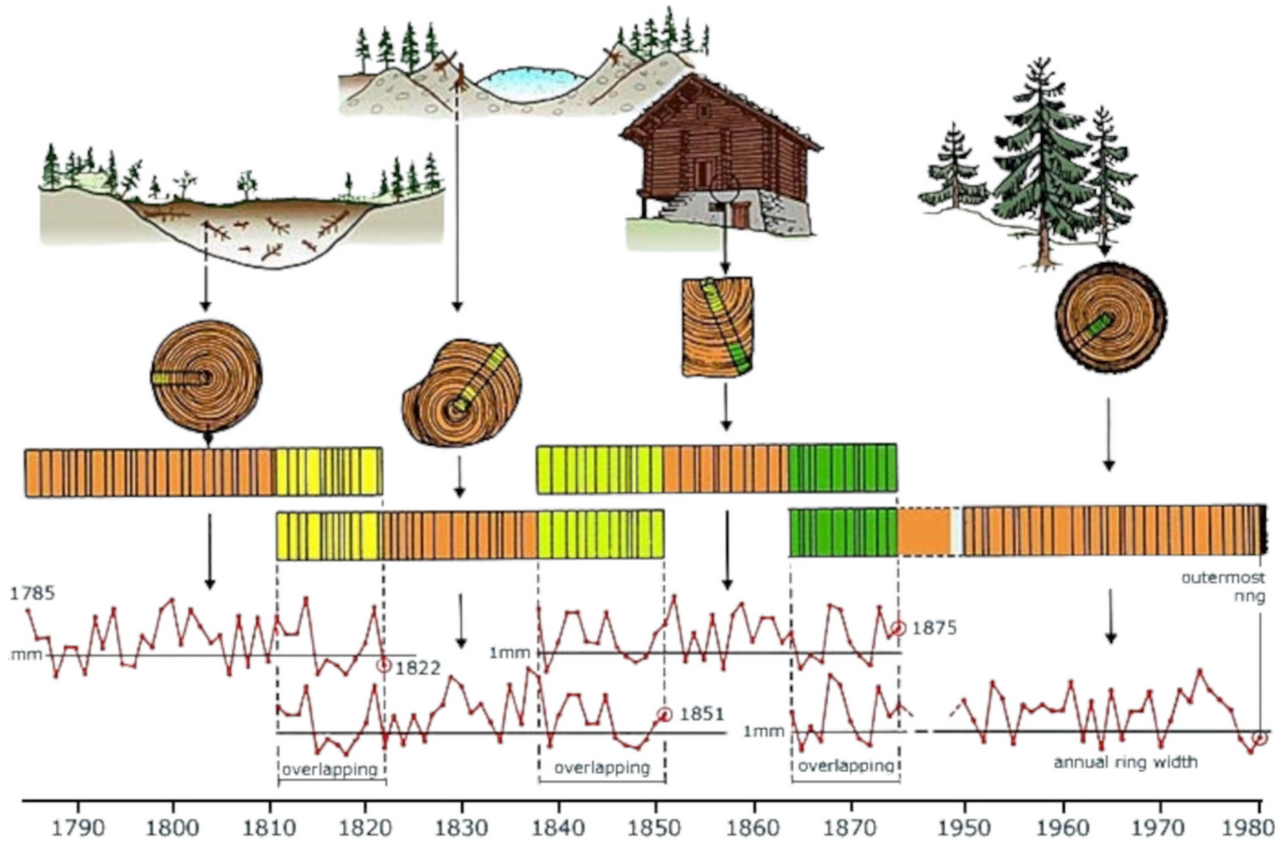
257534

A Comparison of Statistical and Machine Learning Techniques for Crossdating

Table of Contents

Background & Motivation

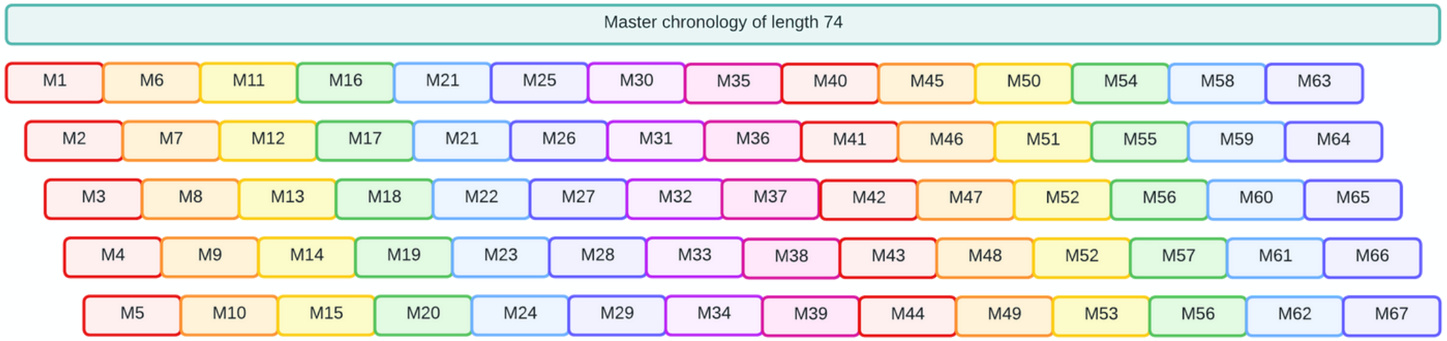
I Research Background & Motivation

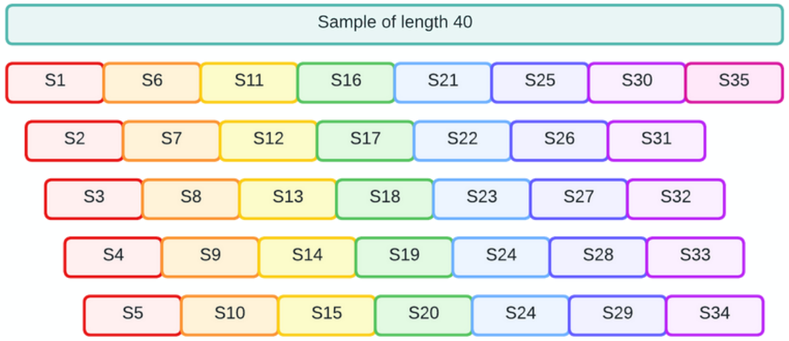


An example of the process of creating a time series, also known as a chronology, using Dendrochronology reprinted from Schweingruber, F. H., 1983. [1]

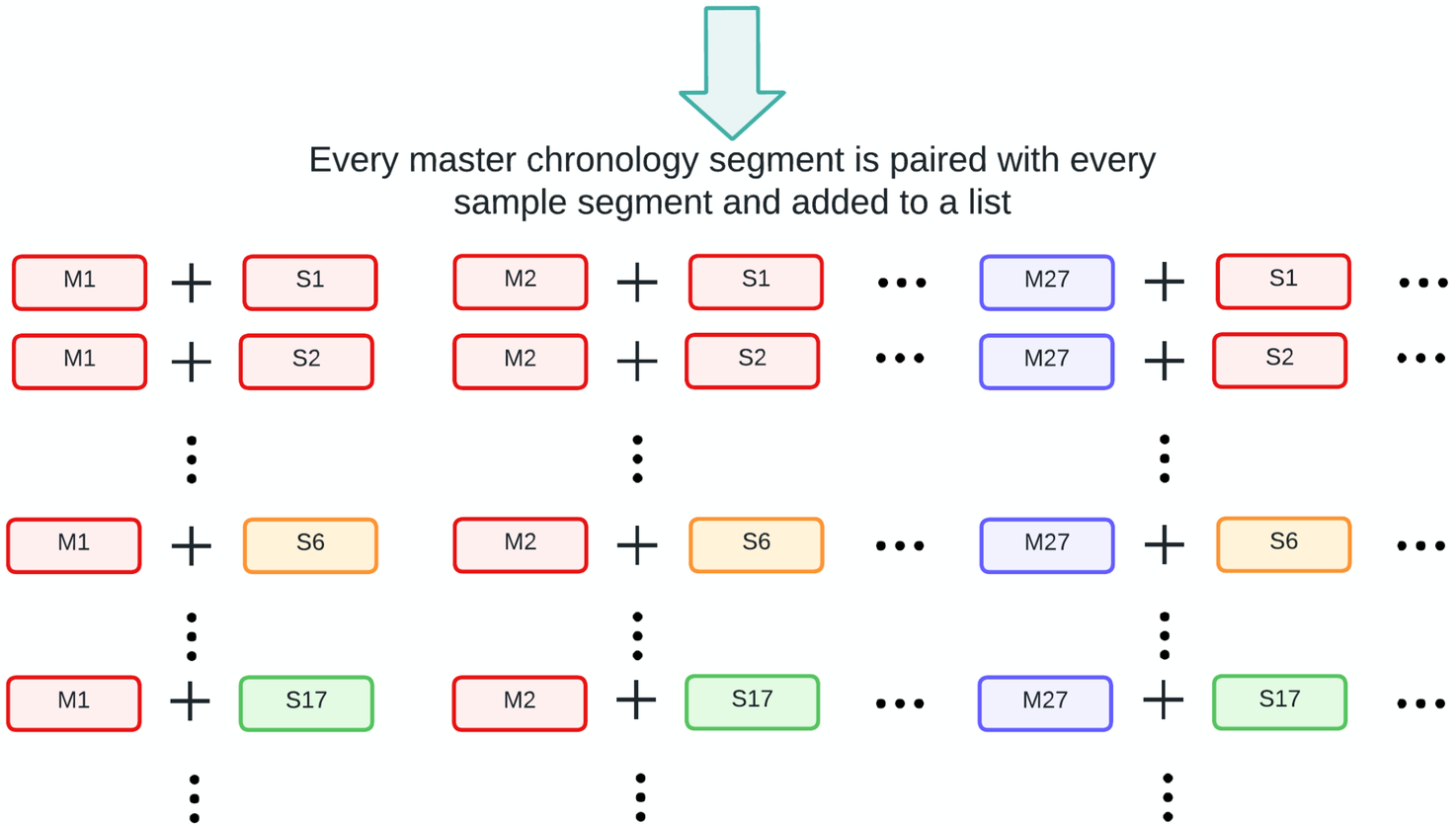
I Research Background & Motivation: Pairwise Lead-Lag Analysis

The master chronology and the sample chronology is split into as many segments of equal length as possible





I Research Background & Motivation: Pairwise Lead-Lag Analysis



Calculate the t values for each of the pairs

Example:

$$t\_val\ (M1, S1) = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

where r is the pearson correlation coefficient between M1 and S2 and n is the length of the segments.

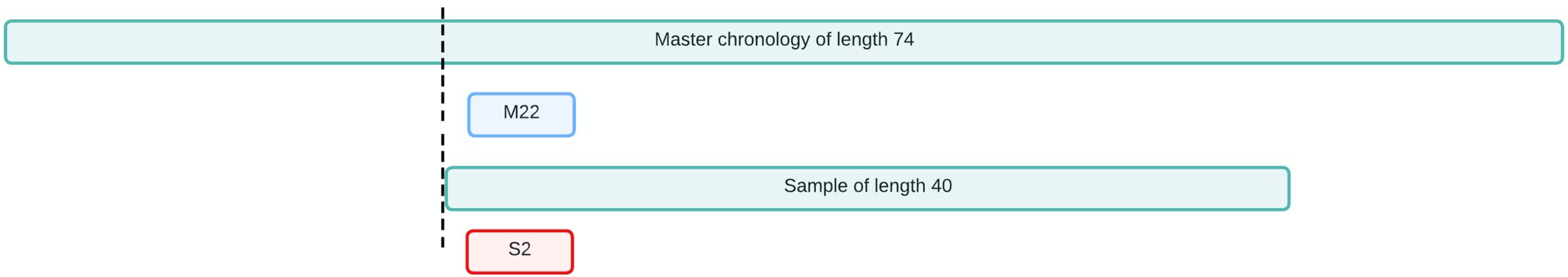
I Research Background & Motivation: Pairwise Lead-Lag Analysis

Find the pair with the highest absolute t value

Max(t value) = M22 + S2

I Research Background & Motivation: Pairwise Lead-Lag Analysis

Use that segment to find the start year for the sample segment and where the chronology aligns



Thus a likely alignment of the sample segment to the master chronology has been found and because the master chronology is dated with years the sample has been crossdated

I Research Background & Motivation

- Crossdating allows scientists to decern insights into the environment before modern climate records began [2].

- It is a time-intensive process so there has been successful development of statistical programs to automate parts of the crossdating process [3].

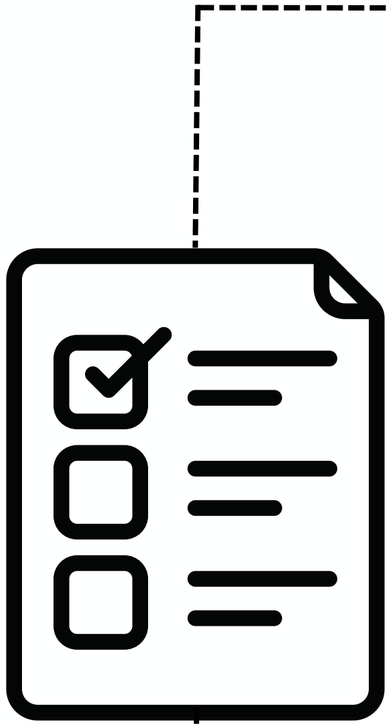
- Machine Learning has been successful at pattern matching in other fields so could be a new approach to crossdating [4].

Project Specification

1. Create 2 novel methods to crossdate a sample chronology to a master chronology.
    - One statistical method using pairwise lead-lag analysis.
    - One machine learning method using Multilayer Perceptrons.
2. Implement a user interface to allow the programs to be used effectively.
3. Compare both methods with each other and a current industry standard crossdating platform.

II Project Specification: Objectives

A successful model will:



- Run in less than 10 minutes
- Work correctly on multiple datasets
- Be explainable
- Be easy to use
- Be customisable
- Return extra helpful information

Methodology

Statistical Method

Pairwise Lead-Lag Analysis

- **Add the dated master chronology indexed by the year and the sample chronology to a data frame.**

- **Divide the sample and the master chronology equal-length segments defined by the user so segments have the maximum possible overlap.**

- **Create pairs of every sample segment with every master chronology segment.**

- **Calculate the t values for all the pairs.**

- **Calculate all the outlying t values above or below a set number of standard deviations**

- **Apply a filter that removes outliers if the next consecutive segment is not an outlier**

- **Calculate the start years of the sample for all the remaining outlying segments from the index of the master chronology segment**

- **Create a list of possible start years and count up the occurrences of each start year**

- **Return the start year which has the highest count and add the sample chronology to the data frame in the correct place**

Machine Learning Method

Multilayer Perceptrons

The Machine learning method uses 3 multilayer perceptron (MLP) classifiers in order to crossdate, the method uses the same counting method used in the statistical method to crossdate.

- **The first MLP determines the segments of length 100 of the master chronology that the sample lies in**

- **The second MLP determines if segments of length 10 of the master chronology and segments of length 10 of the sample chronology overlap by at least 5 years**

- **The third MLP determines if segments of length 10 of the master chronology and segments of length 10 of the sample chronology align exactly**

- **Segments that align exactly are indexed to find the start year for the sample**

- **Create a list of possible start years and count up the occurrences of each start year**

App Demonstration

IV Code Demonstration

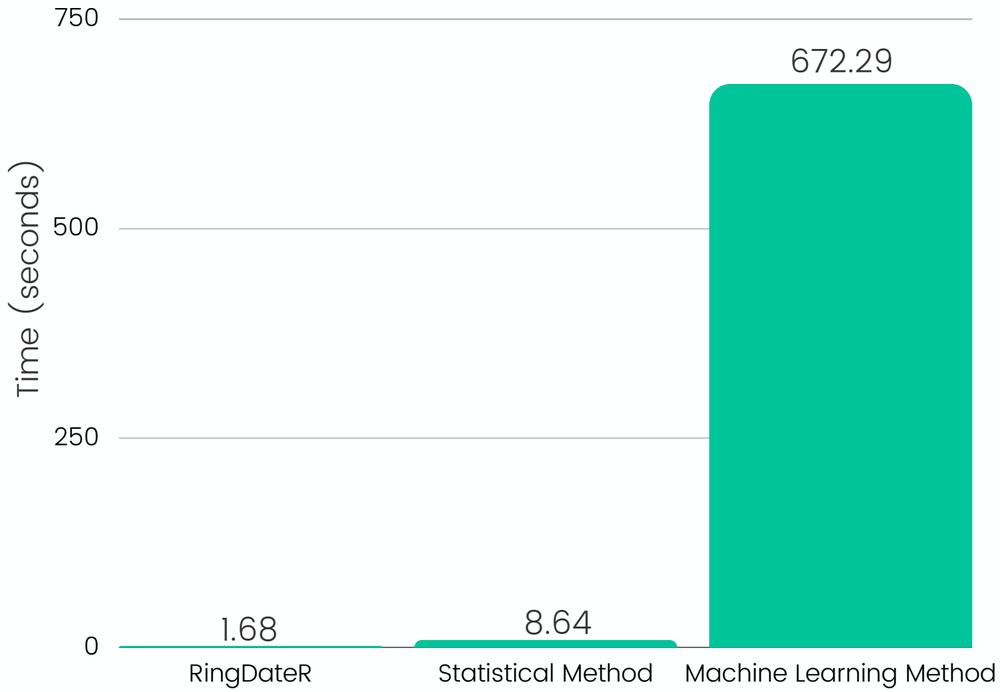
The machine learning method takes 11 minutes to run so it has been preloaded.

Results

Testing was conducted in order to evaluate the success of each model, where possible the models were compared with a state-of-the-art commercially available crossdating application, RindDateR Online Version 1.4 [5].

V Results

Runtime



The runtime of these methods was compared using the same testing data to determine their relative speed.

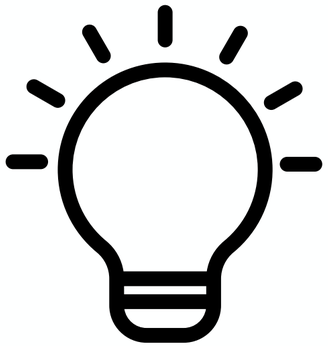
Testing Data

100%
RingDateR

100%
Statistical Method

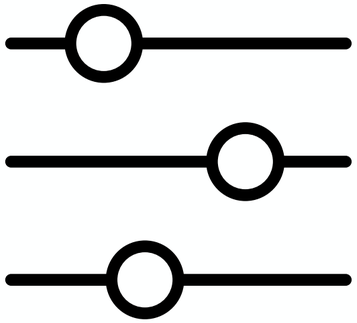
93.9%
Machine Learning Method

Each method was tested using 33 datasets. RingDateR and the Statistical Method passed. The Machine learning method failed 2 tests.



Explainablitity

- The Statistical Method is easy to explain because it is based on a commonly used statistical technique
- The Machine Learning Method is difficult to understand because the method uses models that employ black-box computing [6]

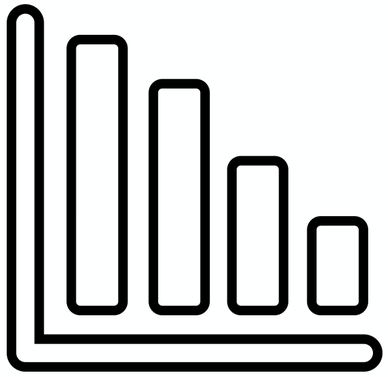


Customisable

- The Statistical Method allows the user to vary the parameters
- The Machine Learning Method allows the user to input their own training dataset

Easy to use

- Both of the programs are integrated into a simple user interface that allows the user to upload CSV files and see the results of the methods in graph form.
- RingDateR can be downloaded or used on a web program and allows the user to upload files of several types of files and see the results of the method in graph form.



Extra Helpful Information

- Both methods output a bar chart of the top 15 start years produced by the method for the user to see.
- RingDateR produces a heatmap of the t values which can help identify errors in the chronologies

V Results: Overview

Method \ Success Criteria	Runs in under 10 minutes on a standard laptop	Works correctly on multiple datasets	Easy to explain	Easy to use	Can be customised	Returns additional helpful information
RingDateR	✓	✓	✓	✓	✓	✓
Statistical Method	✓	✓	✓	✓	✓	✓
Machine Learning Method	✗	~	✗	✓	✓	✓

Discussion & Conclusion

Critical Evaluation

Class Imbalance

In 2 of the 3 MLPs used in the Machine Learning Method, there was an extreme class imbalance which required training data to be heavily resampled

Filter Strength

In a small number of cases, the default filter is too strong to accurately crossdate because the sample was too short to contain the required number of consecutive outliers

Using a Classifier

By using a classifier the method was made more complex because 3 different MLPs had to be implemented to produce correct results

Toolbar on user Interface

The toolbar on the user interface does not render correctly. It still functions but the icons for each of the actions is not visible

Future Directions

Dynamic Default Parameters

Adding functionality to scale the default parameters based off characteristics of the chronology would aid user experience

Default training data

Default training data from a variety of datasets collated allows users to use the Machine Learning Method before they have crossdated several samples

Adding the filter to the Machine Learning Method

Since the Machine Learning method uses the same counting function, implementing a filter may improve the accuracy of the method.

Test statistical method on chronologies with errors

By testing the method with chronologies with errors it can be determined whether it is error resistant

Two methods for crossdating were produced

Each was tested to find their strengths and weaknesses

The statistical method produced better results in line with the success criteria

The project was successful and has possible extension points in the future

VII References

[1] F. H. Schweingruber, Tree Rings: Basics and Applications of Dendrochronology. Springer Netherlands, 1988, pp. 1–4

[2] R. Allan and T. Ansell, "A new globally complete monthly historical gridded mean sea level pressure dataset (hadslp2): 1850–2004," Journal of Climate, vol. 19, pp. 5816–5842, 2006

[3] H. D. Grissino-Mayer, "Evaluating crossdating accuracy: A manual and tutorial for the computer program cofecha," Tree-Ring Research, vol. 57, 2001

[4] S. Dash and A. Dash, "A correlation based multilayer perceptron algorithm for cancer classification with gene-expression dataset," 2014 14th International Conference on Hybrid Intelligent Systems, pp. 158–163, 2014

[5] RingDateR version 1.4, (Accessed: 24th April 2022). [Online]. Available: https://ringdater.shinyapps.io/ringdater_beta_v1_4/

[6] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature machine intelligence, vol. 1, no. 5, pp. 206–215, 2019.

University of Exeter

ADVISORS

Dr. Johan Wahlström
Dr. David Reynolds

STUDENT

257534

Thank you for listening!

University of Exeter

ADVISORS

Dr. Johan Wahlström
Dr. David Reynolds

STUDENT

257534

Link to Presentation:

Presentation.mov

Link to Code Base:

Crossdating