UNIVERSITY OF EXETER

Computer Science Department

M.Sci. Mathematics and Computer Science
Computer Science Department

Comparing Statistical and Machine Learning Methods to Crossdate in Python

Candidate

Ursula Mennear

Student ID 257534

Supervisor

Dr. Johan Wahlström

University of Exeter

Co-supervisor

Dr David Reynolds

University of Exeter

Academic Year
2022/2023

Abstract

(200 WORDS) Crossdating is an important technique in dendrochronology (the study of tree rings) and involves matching tree ring measurements between samples to create a time series. It has been used to establish multi-millennial histories of past ecosystems and climate dynamics. These records have been fundamental in developing our understanding of modern climate change. This project aims to compare a machine learning-based technique and a statistical-based technique to see which can facilitate crossdating by identifying consistent patterns between trees. Machine learning models, such as Multi-layer Perceptrons, have been successful at pattern matching in other fields, so they could be an effective approach for crossdating. Currently, there have been no published attempts at crossdating using machine learning, so the theory of Multi-layer Perceptrons and possible implementation are discussed. Crossdating processes have, primarily, been performed manually and are, therefore, vastly time-consuming. However, some valuable programs have been developed to automate parts of the crossdating process, and their strengths and limitations are reviewed. A popular method in previous programs is Pairwise Lead-lag, which has been chosen as the statistical approach. Developing fully automated approaches to crossdating would significantly increase the ability to generate environmental histories.

	Yes	No
I certify that all material in this dissertation which is not my own work has been identified.	☑	☐
I give the permission to the Department of Computer Science of the University of Exeter to include this manuscript in the institutional repository, exclusively for academic purposes.	☑	☐

Contents

List of Figures

List of Tables

List of Algorithms

List of Code Snippets

Glossary

Crossdating A procedure which matches the pattern of annual growth rings of an organism to another of the same species alive at the same time.

Sample A set of ring widths measurements from one organism.

Chronology A collection of individual samples which have been correctly dated and overlap with each other.

Dendrochronology The study of tree growth rings to ascertain when each ring was formed.

CSV Comma Separated Values

MLP Multi-layer Perceptron

1 Introduction

Crossdating is a procedure which matches the pattern of growth rings [1]. Organisms such as trees and clams grow annually with a new layer or ring forming every year of the organisms' life [2][3]. Samples of the same species grown in similar locations are compared to form a precise time series, identifying the corroborating years through variations in the ring size [4]. Growth patterns reflect the environment; hence species grown in a similar location will have the same pattern. The objective of crossdating is to create a chronology, which is a collection of correctly crossdated organisms which were alive at the same time and each of their growth rings are labelled with the correct year. The collection of samples forming a chronology can then be averaged using a robust mean to form a master chronology with a stable signal to date more samples [5]. The study of Dendrochronology uses the tree's growth rings to determine characteristics of the past [4]. Crossdating is used to create chronologies from living trees, which can live for hundreds of years [6], and can be extended using preserved dead trees [2]. Dendrochronology is used in multiple fields such as climatology and dating wooden panels in art history [7]. It is particularly important for climatology because modern climate observations only date back to the late nineteenth century, during which industrialisation highly impacted the climate [8][9]. It is crucial to gather information on how the climate operated before human interaction to predict how it may behave in the future [10].

Chronologies have allowed scientists to discern unique insights, such as changes in atmospheric conditions (cold periods), fluctuations in the water table (flash floods and droughts) and forest fires [11][12][13]. In order to do this chronologies should be absolutely dated where it is deemed zero errors exist in the crossdating [14]. This means that every ring must be assigned to the correct year, and all erroneous data called noise, such as missing or added rings, have been found and corrected. An absolutely dated chronology is necessary in order to fully utilise the climate data hidden in the tree rings such as interesting patterns or trends [15]. In order for crossdating to occur the growth rings must be preprocessed and standardised so that correlation methods can be applied. As crossdating techniques are used to determine environmental trends, it is essential to first eliminate growth variability that is not caused by environmental factors in order to examine useful variations in growth ring width [16]. These factors include growth trends which are age-related. Detrending facilitates the removal of the long-term growth slowdown associated with these age-related growth trends while maintaining growth that is driven by the environment [17].

The process of dating detrended samples is labour-intensive, and, as a result, the creation of time series is very expensive and time-consuming. Especially since historically, crossdating was a manual process using a graph called a skeleton plot [18]. However, since the 1980s, many successful statistical-based programs have been developed to help automate the process [19]. Current statistical methods predominantly rely on the same statistical analysis called pairwise

lead-lag analysis. Figure 1.1 depicts a timeline of advancements of popular dendrochronology programs throughout history.

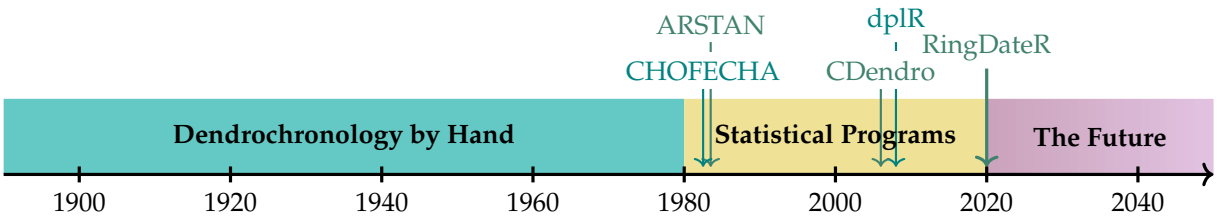

Figure 1.1: Timeline of Development of Popular Crossdating Programs.

More recent programs have implemented methods to pinpoint possible errors and created graphs to support manual error identification.

Pairwise lead-lag analysis is a statistical method for the analysis of two sets of data over a time series; it is currently used in many of the dendrochronology programs available, such as CDendro and RingDateR and has shown to be reliable and robust [20][21].

The data is split up into sections of the same length as the unknown chronology; the sections of the master chronology will overlap for a given number of years. Each section of the master chronology will be paired with the unknown chronology and then a t-value will be calculated. The t-values determine the strength of the correlation whilst taking into account the overlap in the data. This can be done by calculating the following equation [22]:

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{1.1}$$

where n is the number of years over which the correlation is calculated and r is the Pearson correlation coefficient (1.2). The Pearson correlation coefficient is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1.2}$$

where r is the correlation coefficient, x_i is the i^{th} value of the x variable in a sample, $\bar{x}$ is the mean of the values of the x-variable, y_i is the i^{th} value of the y-variable in a sample and $\bar{y}$ is the mean of the values of the y-variable.

If there is a point where the sample and the master chronology correlate, its t-value will be significantly larger compared with the other results. So, in order to find the possible correlations, all the t-values are compared to each other to find an outlier [20]. Once one is found, it is checked using cross-correlation and plotted as a line graph [23]. The lines for the master chronology and the new sample should follow similar trends.

It is important to note that there still exists variations in detrended samples so any match is not perfect and any method must match the slopes of the lines formed by multiple data points rather than the values of each individual data point. Figure 1.2 is an example of a sample against a master chronology to illustrate this point.

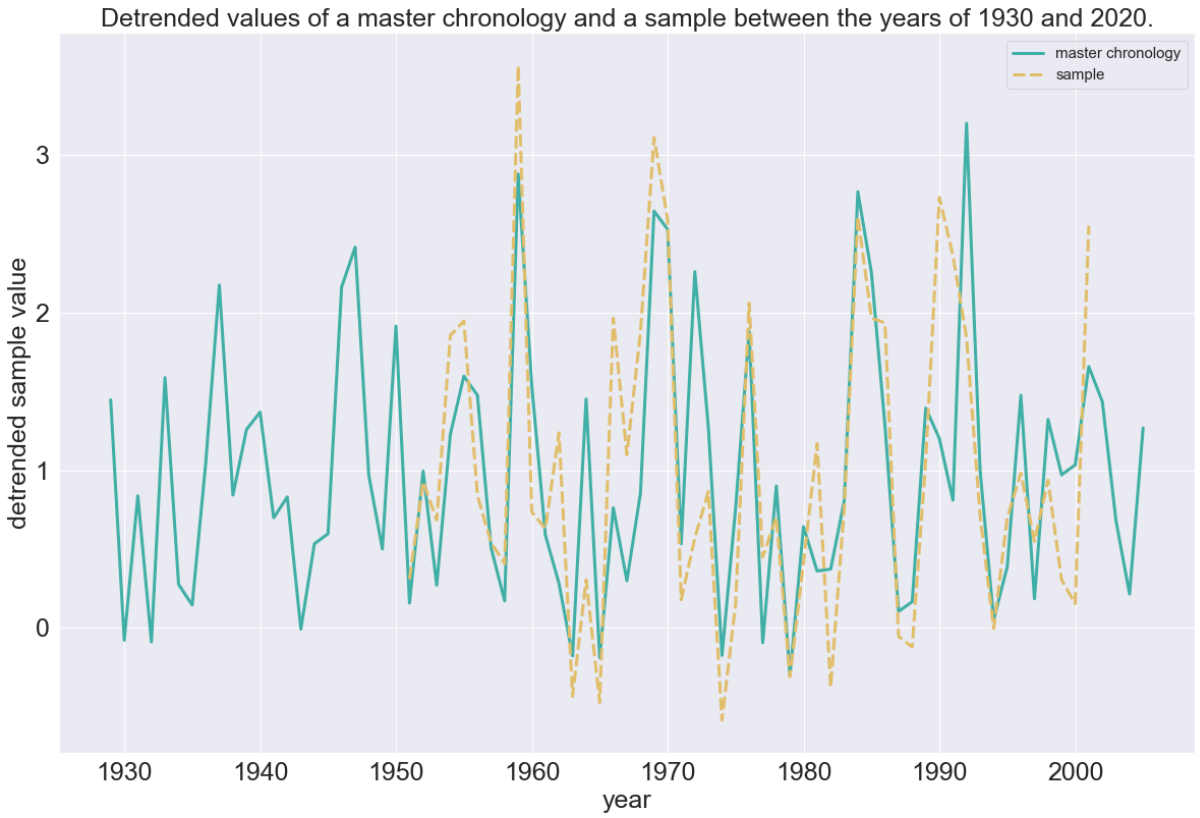

Figure 1.2: Line graph of a detrended master chronology against a sample chronology between the years 1925 and 2020 from the (find the dataset)

This project attempts to use python to experiment and explore both a statistical and machine learning approach to crossdating. Crossdating programs are predominantly written in R as it is a language tailored to statistical methods and graph visualisations. Implementing the tried and tested pairwise lead-lag analysis will allow the method to be experimented with to determine the best variables and suggest helpful insights to the user. Machine learning, in particular multilayer perceptrons (MLPs), have been used in various forms of pattern matching experiments [24]. Errors such as ring insertion and deletion can severely affect the ability of statistical methods to crossdate so new methods to negate their effect on crossdating need to be explored. Thus by designing two different approaches to the same problem and comparing them, their strengths and weaknesses can be properly assessed.

2

Project specification and aims

2.1 Project specification

The main aim of this project is to two methods to crossdate a sample against a master chronology using different methodologies in order to compare them. An essential part of the methodology is experimentation to optimise the accuracy of each of the methods. The methods are first tested with entirely correct master chronologies and samples with the possibility to extend to samples with errors (ring insertions or deletions).

2.2 Objectives

ID	Requirement	Details
Functional Requirements		
1	Input - Take a csv file	The program must take the data collected from the National Centers for Environmental Information Tree Ring database in the form of a csv with a master chronology then a sample chronology.
2	Implement Pairwise lead-lag analysis in python	Create a functioning program that accurately cross-dates an unknown chronology to a master chronology using pairwise lead-leg analysis and experiment with variables to aid user selection.
3	Implement multilayer perceptron in python	Create a functioning program that accurately cross-dates an unknown chronology to a master chronology using Multi-layer Perceptrons.
4	Output - Result in an easy-to-read format	Add a column to the dataframe with the calculated start date and generate figures such as line graphs or bar charts to display the calculated correlation.
Non-Functional Requirements		
5	Programs run in a reasonable time	Each of the completed programs can run in under 10 minutes on a standard laptop.
6	Comprehensive Documentation	Explain all design decisions thoroughly with experimental data to back it up and explain how to run the program clearly.
7	In-depth evaluation of methods	Describe strengths and weaknesses of each program and compare performance between them.

Table 2.1: Project Requirements.

Algorithm 1 An algorithm with caption

Require: $n \geq 0$
Ensure: $y = x^n$
  $y \leftarrow 1$
  $X \leftarrow x$
  $N \leftarrow n$
  while $N \neq 0$ **do**
    if N is even **then**
      $X \leftarrow X \times X$
      $N \leftarrow \frac{N}{2}$ {This is a comment}
    else if N is odd **then**
      $y \leftarrow y \times X$
      $N \leftarrow N - 1$
    end if
  end while

2.3 Objectives

$$e^{j\pi} + 1 = 0 \tag{2.1}$$

3 Methodology

3.1 Statistical Method

3.2 Design

The basis of the method is pairwise lead-lag analysis explained in 1. In order to give dendrochronologists flexibility to optimise the technique for different species experiments have been conducted on possible variables and user input is required with a suggested default. A simple graphical user interface (GUI) has been constructed so that the program can run smoothly without the user requiring technical knowledge. The GUI is basic but does allow figures from the data to be displayed and a progress bar so the user can get the most out of the program.

3.2.1 Methods

3.2.2 Implementation

To begin the program asks for a csv containing 3 columns, the year that lines up with the master chronology and a final column containing the sample chronology to match with padding at the end so all columns are the same length. The csv file is converted into pandas dataframe.

3.3 Machine Learning Method

4 Results

4.1 Testing

4.2 Outcomes

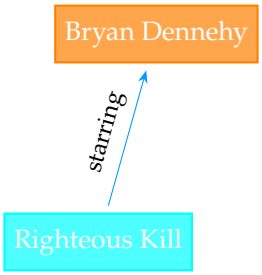

Figure 4.1: Image created with TikZ

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

```
import numpy as np

def incmatrix(genl1,genl2):
    m = len(genl1)
    n = len(genl2)
    M = None #to become the incidence matrix
    VT = np.zeros((n*m,1), int)  #dummy variable

    test = "String"

    #compute the bitwise xor matrix
    M1 = bitxormatrix(genl1)
    M2 = np.triu(bitxormatrix(genl2),1)

    for i in range(m-1):
        for j in range(i+1, m):
            [r,c] = np.where(M2 == M1[i,j])
            for k in range(len(r)):
                VT[(i)*n + r[k]] = 1;
```

```
                VT[(i)*n + c[k]] = 1;
                VT[(j)*n + r[k]] = 1;
                VT[(j)*n + c[k]] = 1;

                if M is None:
                    M = np.copy(VT)
                else:
                    M = np.concatenate((M, VT), 1)

                VT = np.zeros((n*m,1), int)

    return M
```

Code 4.1: Code snippet example

4.3 Evaluation

5 Discussion

5.1 Critical Reflection

5.2 Future Directions

A	**B**
C	D
E	F
G	H

Table 5.1: Table example

6

Conclusion

PLAN:

- Introduce all key terms and motivation.
- Detail project specification, add project specification table and success criteria.
- Methodology -> reference statistical decisions and experiments at each stage, show flexibility and research to back up decisions (Include many tables). Make it into a narrative of how the project progressed, even if it didn't exactly go like that
- Results -> Clear experimentation and helpful figures compare different variables as well as overall methodologies
- Discussion -> Which method works better, decide which is more robust when dealing with errors and whether the loss of explainability is worth the increased performance. The weaknesses of the models and why they remained in the project. Future reflections to expand the project, such as error-checking methods.
- Conclusion -> How the project impacted the field and whether it met objectives and was worthwhile.

A	**B**
C	D
E	F
G	H

Table 6.1: Table example

References

[1] T. Wigley, P. Jones, and K. Briffa, "Cross-dating methods in dendrochronology," *Journal of Archaeological Science*, vol. 14, no. 1, pp. 51–64, 1987, ISSN: 0305-4403. DOI: https://doi.org/10.1016/S0305-4403(87)80005-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0305440387800055.

[2] J. G. A. Lageard, "Dendrochronology," *Encyclopedia of Geoarchaeology*, A. S. Gilbert, Ed., pp. 180–197, 2017. DOI: 10.1007/978-1-4020-4409-0_41. [Online]. Available: https://doi.org/10.1007/978-1-4020-4409-0_41.

[3] D. J. Reynolds, V. R. von Biela, K. H. Dunton, D. C. Douglas, and B. A. Black, "Sclerochronological records of environmental variability and bivalve growth in the pacific arctic," *Progress in Oceanography*, vol. 206, 2022.

[4] M. Kaennel and F. H. Schweingruber, "Multilingual glossary of dendrochronology," *Swiss Federal Institute for Forest, Snow and Landscape Research. Berne, Stuttgart, Vienna, Haupt*, vol. 133, pp. 162–184, 1995.

[5] R. Laxton and C. Litton, "Construction of a kent master dendrochronological sequence for oak, ad 1158 to 1540," *Medieval archaeology*, vol. 33, no. 1, pp. 90–98, 1989.

[6] J. J. Camarero, M. Colangelo, A. Gracia-Balaga, M. A. Ortega-Martínez, and U. Büntgen, "Demystifying the age of old olive trees," *Dendrochronologia*, vol. 65, 2021, ISSN: 1125-7865. DOI: https://doi.org/10.1016/j.dendro.2020.125802. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1125786520301417.

[7] M. Baillie, "Some thoughts on art-historical dendrochronology," *Journal of Archaeological Science*, vol. 11, no. 5, pp. 371–393, 1984, ISSN: 0305-4403. DOI: https://doi.org/10.1016/0305-4403(84)90019-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0305440384900190.

[8] 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, no. 22, pp. 5816–5842, 2006.

[9] Q. He and B. R. Silliman, "Climate change, human impacts, and coastal ecosystems in the anthropocene," *Current Biology*, vol. 29, no. 19, R1021–R1035, 2019, ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2019.08.042. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982219310929.

[10] J. Adams and F. Woodward, "The past as a key to the future: The use of palaeoenvironmental understanding to predict the effects of man on the biosphere," in *The Ecological Consequences of Global Climate Change*, ser. Advances in Ecological Research, M. Begon, A. Fitter, and A. Macfadyen, Eds., vol. 22, Academic Press, 1992, pp. 257–314. DOI: https://doi.org/10.1016/S0065-2504(08)60138-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0065250408601385.

[11] F. H. Schweingruber, "Tree rings: Basics and applications of dendrochronology," in Springer Netherlands, 1988, pp. 1–4, ISBN: 978-94-009-1273-1. DOI: 10.1007/978-94-009-1273-1_1. [Online]. Available: https://doi.org/10.1007/978-94-009-1273-1_1.

[12] M. Baillie, "Seven thousand years of alternative history: The tree-ring story.," *Irish Forestry*, 1999.

[13] E.-B. Choi, Y.-J. Kim, J.-H. Park, C.-R. Park, and J.-W. Seo, "Reconstruction of resin collection history of pine forests in korea from tree-ring dating," *Sustainability*, vol. 12, Nov. 2020. DOI: 10.3390/su12219118.

[14] B. Gmińska-Nowak, A. D'Agostino, Y. Özarslan, *et al.*, "Dendrochronological analysis and radiocarbon dating of charcoal remains from the multi-period site of uşaklı höyük, yozgat, turkey," *Journal of Archaeological Science: Reports*, vol. 38, p. 103 078, 2021, ISSN: 2352-409X. DOI: https://doi.org/10.1016/j.jasrep.2021.103078. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352409X2100290X.

[15] E. Cook and N. Pederson, "Uncertainty, emergence, and statistics in dendrochronology," *Dendroclimatology: Progress and Prospects*, vol. 11, Sep. 2011. DOI: 10.1007/978-1-4020-5725-0_4.

[16] S. A. Johnson, S. I. Gass, and M. C. Fu, "Splines," in *Encyclopedia of Operations Research and Management Science*, Springer US, 2013, pp. 1443–1446, ISBN: 978-1-4419-1153-7. DOI: 10.1007/978-1-4419-1153-7_982. [Online]. Available: https://doi.org/10.1007/978-1-4419-1153-7_982.

[17] D. Frank, K. Fang, and P. Fonti, "Dendrochronology: Fundamentals and innovations," in *Stable Isotopes in Tree Rings: Inferring Physiological, Climatic and Environmental Responses*, R. T. W. Siegwolf, J. R. Brooks, J. Roden, and M. Saurer, Eds. Springer International Publishing, 2022, pp. 21–59. DOI: 10.1007/978-3-030-92698-4_2. [Online]. Available: https://doi.org/10.1007/978-3-030-92698-4_2.

[18] P. R. Sheppard, *Making skeleton plots*, (Accessed: 14th of September 2022). [Online]. Available: https://www.ltrr.arizona.edu/skeletonplot/plotting.htm.

[19] H. D. Grissino-Mayer, "Evaluating crossdating accuracy: A manual and tutorial for the computer program cofecha," *Tree-Ring Research*, vol. 57, no. 2, 2001.

[20] D. J. Reynolds, D. C. Edge, and B. A. Black, "Ringdater: A statistical and graphical tool for crossdating," *Dendrochronologia*, vol. 65, 2021, ISSN: 1125-7865. DOI: https://doi.org/10.1016/j.dendro.2020.125797. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1125786520301363.

[21] P. O. Larsson and L.-Å. Larsson, "Match a sample towards a reference curve," *Cybis.se: Technical writing, software development, Dendrochronology*, Jun. 2015. [Online]. Available: https://cdendro.se/forfun/dendro/index.htm.

[22] S. Bennett, M. Cucuringu, and G. Reinert, "Lead-lag detection and network clustering for multivariate time series with an application to the us equity market," *arXiv preprint arXiv:2201.08283*, 2022. [Online]. Available: https://arxiv.org/abs/2201.08283.

[23] T. Derrick and J. Thomas, "Time series analysis: The cross-correlation function," *Innovative analyses of human movement*, pp. 189–205, 2004. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-47849111834%5C&partnerID=40%5C&md5=79f1d9a6506ee071f7edf9ec30e8d825.

[24] 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. DOI: 10.1109/HIS.2014.7086190.

Appendix

1.1 Some Appendix

The contents...

Acknowledgments

Thanks to my supervisors, Johan Wahlström and David Reynolds, and Dr George De Ath for all of the support.