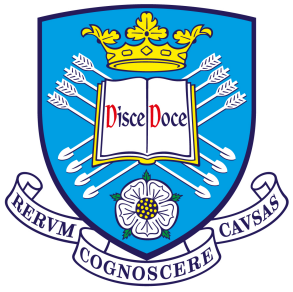
Semi-Automation of Bayesian Chronology Construction Using a Graph-Theoretic Approach



The University Of Sheffield.

Bryony Charlotte Moody

School of Mathematics & Statistics

The University of Sheffield

A thesis submitted in partial fulfilment for the degree of
Doctor of Philosophy

May 2023

Abstract

In the 1990s, Bayesian modelling revolutionised archaeological chronology construction. It allowed for using relative and absolute dating evidence to improve the precision of calendar dates for archaeological events. However, the process of building such chronological models is time-consuming and labour-intensive, which means often only one chronological model is considered. Furthermore, even if multiple plausible chronological models are considered, there is currently a lack of suitable statistical metrics for analysing such sets of models.

Within this research, we use mathematical graph theory to manage both stratigraphic and chronological information during Bayesian chronology construction. Dye and Buck (2015) developed a graph theoretic approach to representing archaeological relative dating evidence, proposing that we might semi-automate chronology construction by managing such archaeological dating evidence using graph theory. The research presented in this thesis builds upon the initial theoretical work by Dye and Buck (2015), showcasing novel software we have produced for the semi-automation of chronology construction.

We present the result of research that sought to address three objectives. The first was to quantify the quality and potential for reuse of digital dating evidence obtained during excavation and then deposited to digital repositories. This review demonstrated a distinct lack of reusable archaeological dating evidence within a prominent digital archaeological archive in the United Kingdom. This absence of reusable data is of particular concern in archaeology due to the non-repeatable nature of excavation. The research presented in this chapter was published in a peer-reviewed journal (Moody et al., 2021), in which we provide recommendations for improving the reusability of digital dating evidence in archaeology.

Following this initial research, we developed novel prototype software (using mathematical graphs) to manage dating evidence that is required for Bayesian chronological modelling. Within the same software, users can visualise and manipulate their dating evidence using point-and-click functionality. In addition, users can automatically obtain calendar age estimates for archaeological events of interest simply by loading in the required dating evidence into the software. Thus semi-automating the process of chronology construction, something which is not possible within existing software for Bayesian chronological modelling. Further, the prototype software improves the process of saving the data and information used, and produced, during the chronological building process, thus improving the potential for reusability of such data and information following future archiving.

Further functionality of this software allows for the rapid semi-automated construction of multiple chronological models, which we demonstrate proof-of-concept results for case studies based on existing dating evidence from archaeological excavations carried out at various sites in Europe. Further, we explore a novel application of an existing statistical methodology that allows us to collectively interpret the results of fitting multiple chronological models, each of which we deem plausible (a priori) for a given site within our analysis.

Finally, we discuss how we plan to develop our prototype software in the future, ensuring that it is functional for the archaeological community. We seek to ensure that all relevant dating evidence obtained during excavation can be managed and interpreted within the software and that it enables efficient and reliable archiving of models, methods, data and results thus facilitating improved repeatability and usability of future chronology construction methods and data.

Acknowledgements

I am very fortunate to have had a wonderful support system throughout this PhD. Therefore I am afraid this list of acknowledgements is going to be a rather long one.

First, I would like to thank my supervisors, Caitlin Buck, Keith May and Gianna Ayala, and our collaborator Tom Dye, for their continued support, feedback and guidance during this PhD. In particular, I thank Caitlin for your initial encouragement, without which I would not have had the confidence to pursue this PhD.

To my other half, Farhad, thank you for all your support and for always being willing to listen to me claim that "my code is fixed for sure now" despite the fact it never was. My friends, Carys, Alex, Keira and George, thank you for keeping me sane during the write-up stage of this PhD. I would like to thank Carly and Rosie, our lovely coffee dates and chats have been a constant treat during this whole process. To Gemma and Matt, thank you for making what could have been a very stressful start to this PhD so much better. To my family, thank you for always being supportive and proud in my decade of being a student. To Bo, thank you for always being on hand to provide helpful chats about writing and all computer-related issues. Thank you to Kieran for being willing to spend your 10-hour coach journeys around South America reading this thesis and telling me what you think.

I would also like to thank Danny and Cassie for introducing me to the world of Bayes linear statistics and for always being willing to provide your time and such insightful discussions about your research. Finally, would like to thank the staff of the ADS, in particular Tim Evans, Jenny O'Brien, Holly Wright and Julian Richards, for their generosity in giving their time and expertise, and especially to Holly, thank you for always seeming to manage to find funding for me to present my work to the community!

This work was supported by the Arts and Humanities Research Council via funding under their grants (ref AH/S001328/1 and AH/T002093/1), without which this research could not have taken place.

Nomenclature

Radiocarbon dating

t	Time since death of an organic sample
A	Number of ^{14}C atoms at time t
A_0	Number of ^{14}C atoms at time t
λ	Decay constant
$T_{1/2}$	Half life of an organic material
x	Laboratory estimate of the true radiocarbon age of a sample
σ	Laboratory error for radiocarbon estimate x

Bayesian inference

θ	True calendar age of a sample
$r(\theta)$	True radiocarbon age of a sample with true calendar age θ
$r(\cdot)$	True function that coverts calendar age to radiocarbon age

$\mu(\cdot)$	A calibration curve estimate which is an approximation of $r(\cdot)$
$\delta(\cdot)$	Uncertainty on the calibration curve estimate, corresponding to 1 standard deviation from $\mu(\cdot)$
R	interval of calendar ages
J	number of archaeological groups
α_j	age of event that corresponds to start of group j for groups $1, ..., J$
β_j	age of event that corresponds to end of group j for groups $1, ..., J$
$\boldsymbol{\beta}$	Set of true calendar ages of the end event of all groups within a chronological model
$\boldsymbol{\theta}$	Set of true calendar ages of all contexts within a chronological model
$\boldsymbol{\alpha}$	Set of true calendar ages of the start event of all groups within a chronological model
$\boldsymbol{x}$	Set of all laboratory radiocarbon age estimates for all samples dated for a chronological model
$\boldsymbol{\sigma}$	Set of all laboratory errors for radiocarbon age estimates $\boldsymbol{x}$
N_j	Number of contexts in group j
$x_{i,j}$	radiocarbon age estimate of sample found in context i of group j
$\sigma_{i,j}$	laboratory error of radiocarbon determination for sample found in context i of group j

$\theta_{i,j}$ True calendar age of sample found in context i of group j

$X_{i,j}$ Sample found in context i of group j

C Set of possible values for $\boldsymbol{\theta}$ that satisfy the constraints within the stratigraphic sequence

G Set of possible values for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that satisfy the constraints within the stratigraphic sequence

P, Q Labels of archaeological groups

E, F Random events

$s(\boldsymbol{\alpha}, \boldsymbol{\beta})$ Span of calibrated years between the oldest and youngest group boundaries in a chronological model

MCMC methods

M Number of chronological models considered plausible

V_m Set of judgements for model m that encompasses all the prior knowledge, modelling decisions and computational methods that result in model m

$\mathcal{G}$ A vector of estimated marginal posterior expectations parameters of interest, for each of the M models we consider plausible

K Number of coexchangable classes

H_k Number of posterior expectations within k^{th} coexchangable class

$M(C_k)$ Common posterior expectation of the parameters of interest for all models within that class k

$R_h(C_k)$ Residual term attributed to the h^{th} expectation within class k, which accounts for the difference between the common posterior expectation of the class, $M(C_k)$ and the actual posterior expectation of the parameters of interest for model h

Contents

List of Figures

List of Tables

Chapter 1

Introduction to chronology construction

A crucial aspect of archaeological research is the construction of archaeological chronologies. These timelines of human activity allow for the historical interpretation of a given geographical location (see Barker, 1993, Chapter 1). The first step of constructing a chronology is excavating the site of interest. As defined by Carver (2014), archaeological excavation is "the procedure by which archaeologists define, retrieve, and record cultural and biological remains found in the ground". During this process, archaeologists remove evidence of human activity from an archaeological site in (approximately) the reverse order in which it was originally deposited, thus obtaining some indication of (reverse) chronology as they do so. Following excavation, archaeologists use their field observations, scientific (e.g. radioacarbon) dates, and expert knowledge (such as a relative ordering of the human evidence that has been excavated) alongside various tools, including a selection of statistical methodologies, to reconstruct a chronological sequence for the site. These methods are used to provide an interpretive framework that enables archaeologists to build a timeline of human activity at a given site. The process from excavation to the final

chronology is referred to within this research as chronology construction.

Chronological modelling is a subset of chronological construction that exclusively utilises statistical methods to synthesise chronologies (see Buck and Millard, 2004) and is the focus of the research within this thesis. A wealth of dating evidence is obtained during excavation which can be used in chronological modelling. Such evidence falls into one of two categories: relative or absolute. Relative dating evidence provides an ordering of archaeological events relative to each other. Absolute dating evidence, which typically derives from some scientific dating method or from expert knowledge, allows archaeologists to anchor their chronology to the calendar scale. However, the precision of most absolute date estimates is limited, leading to associated uncertainty that must also be accounted for during the modelling process.

1.1 Goals of our research

The overarching aim of the PhD project that led to this thesis was to provide prototype software for semi-automating the chronology construction process. The project had three research goals to allow us to achieve this. The first was to characterise the common types of relative and absolute dating evidence used in chronological modelling and then review the reusability and availability of such data in digital archaeological archives. This goal was essential to ensure any software we wrote to semi-automate chronology construction would be suitable for managing and modelling the common types of archaeological dating evidence available in archaeological archives and used in chronological modelling. A further focus of this first goal was to consider which standards for digital archives (of excavation records) would most readily allow for the reuse of dating evidence and thus be suitable for use in this PhD research and the wider research community.

A second research goal was to ensure that we provide software allowing users to manage and interpret large volumes of complex dating evidence in a semi-automated fashion. Due to the complex nature of the dating evidence observed during an excavation, multiple chronologies may be considered plausible for any given site. As a result, the final research outcome was to develop software allowing archaeologists to construct multiple chronologies if they deem them plausible for their site. In addition, we sought to explore options for the interpretation of multiple plausible chronologies, seeking statistical methodologies that could aid in such interpretation.

1.2 Outline of thesis structure

Given the highly multidisciplinary nature of the research presented in this thesis and the somewhat non-standard nature of the PhD, the structure of this thesis is also non-standard. For example, rather than providing a single literature review chapter, each chapter has a corresponding literature review relevant to the research area(s) discussed within it. Furthermore, the archaeological and statistical background relevant to this research is provided in dedicated background chapters, which we now outline.

Chapter 2 introduces the relative dating evidence typically observed during an excavation within the United Kingdom before Chapter 3 introduces radiocarbon dating, a method of scientific dating which provides absolute dating evidence. In Chapter 4, we present research by the author into the availability and reusability of archaeological dating evidence within the UK's digital archives. This data review was an outcome stipulated by my funding body and the resulting formal report was circulated to managers within Historic England. It was also expanded upon and published in a peer-reviewed journal (see Moody et al., 2021).

Within Chapters 5 and 6, we provide the relevant and established theory of Bayesian chronology modelling, with Chapter 5 motivating the specific statistical methodology that is commonly used for chronological modelling. In addition, we offer an example of a simple statistical model that might be used for chronological modelling. Chapter 6 then outlines the formal structure of the statistical model used in this research and routinely adopted for chronological modelling within the wider research community. In addition, we provide illustrative case studies motivating the value of chronological modelling in archaeology and specifically in considering multiple plausible chronologies for a given site. Further, we introduce dating evidence from archaeological sites used for case studies within the subsequent chapters.

Following this, Chapter 7 returns to the author's research and introduces our novel purpose-built prototype software that semi-automates chronology construction and can rapidly construct multiple plausible models for a single archaeological site. Subsequently, Chapter 8 explores a novel use of an existing statistical methodology, which would allow for the analysis of multiple plausible chronological models. Finally, a summary of the results of this research and a discussion of further work that could follow is provided in Chapter 9.

Chapter 2

Relative dating evidence

The core goal of the research in this thesis is to improve the process of archaeological chronology construction. Two key components required for this are relative and absolute dating evidence. This chapter provides an overview of archaeological excavation methods that produce the former of these. A brief discussion of absolute dating, specifically scientific dating, is provided in the next chapter. For technical details relating to the scientific dating used in this research, see Chapter 5, Section 3.2.

2.1 Archaeological interventions

Broadly speaking, archaeological interventions comprise several stages including pre-excavation investigation, physical excavation of the site, post-excavation analysis, the publication of results and final archiving of data and documentation. We observe relative and absolute dating evidence in the excavation and analysis stages. Thus, we do not discuss pre-excavation strategies but refer the reader to Roskams (2002),

Chapter 3. We focus on the practice of physically excavating a site and the methods that allow us to infer relative dating information, primarily through stratigraphy.

2.1.1 Stratigraphy in archaeology

Geologists refer to layers of rock as strata, inferring the relative age of each layer by utilising the fact that strata deposited deeper down must be older (Harris, 1989, Chapter 1). This theory is also widely used in archaeology since the same principles of stratigraphy hold for deposits (or layers) resulting from human activity as they do for geological activity. In archaeology, stratigraphy can encompass any deposit (such as the fill of a ditch), structure (such as a wall) or an interface between two other deposits/structures. An interface (unit) represents an archaeological event for which no physical depositional evidence can be observed. For example, when digging a ditch, the layers of strata that the ditch was dug into, and the strata that later filled the ditch are all deposits. In addition, the interface created by digging the ditch through the other strata can be an important event in its own right and is commonly called a cut. That cut interface is recognised as a distinct archaeological event within the stratigraphy and should be recorded.

Harris (1989) sought to formalise the process of excavating deposits, structures and interface units by recording their stratigraphic relationships during an archaeological excavation, assuming that stratification occurs to some degree at all archaeological sites. Just as Pitt-Rivers and Wheeler did before him (see Jensen, 2017, for a history of excavation methods), Harris (1989) argued that arbitrarily digging holes as a method of excavation is only sufficient for recovering artefacts and that key archaeological information is lost if one does not observe, excavate and record the stratigraphy. Harris introduced a diagrammatic method for recording and illustrating the stratigraphic relationships systematically. Such diagrams, now

commonly referred to as Harris matrices, help archaeologists to visualise and understand the complexity of the stratigraphic relationships they observe during excavation.

Harris consolidated his archaeological stratigraphic principles in four laws of archaeological stratigraphy, three of which correspond to geological laws of stratigraphy, and the fourth is specific to archaeology (discussed in Section 2.1.3).

1. **The law of original horizontally** tells us that any archaeological strata will have originally been deposited horizontally, with deposits conforming to the shape of the strata directly below.

2. **The law of superposition** tells us that, for any given pairs of strata, the strata below are older than the strata above. This law is fundamental to building stratigraphic sequences since a full site stratification can be constructed based on the pairwise above-below relationships between strata.

3. **The law of continuity** relates to strata that are not physically observed together during excavation but where other evidence implies that they were once physically one single strata. For example, digging a ditch may completely remove a portion of a layer of strata, thus cutting it into two apparently separate strata. To a trained archaeological eye, however, since a ditch has been dug, the two now-separate strata must have once been whole (i.e. they are once-whole strata)

2.1.2 Archaeological stratigraphic relationships

Three different relationships between archaeological strata can be used to summarise the laws of archaeological stratigraphy: no direct stratigraphic relationship, strata in superposition, and strata that were once whole or part of the same feature. These

are demonstrated in Figure 2.1.

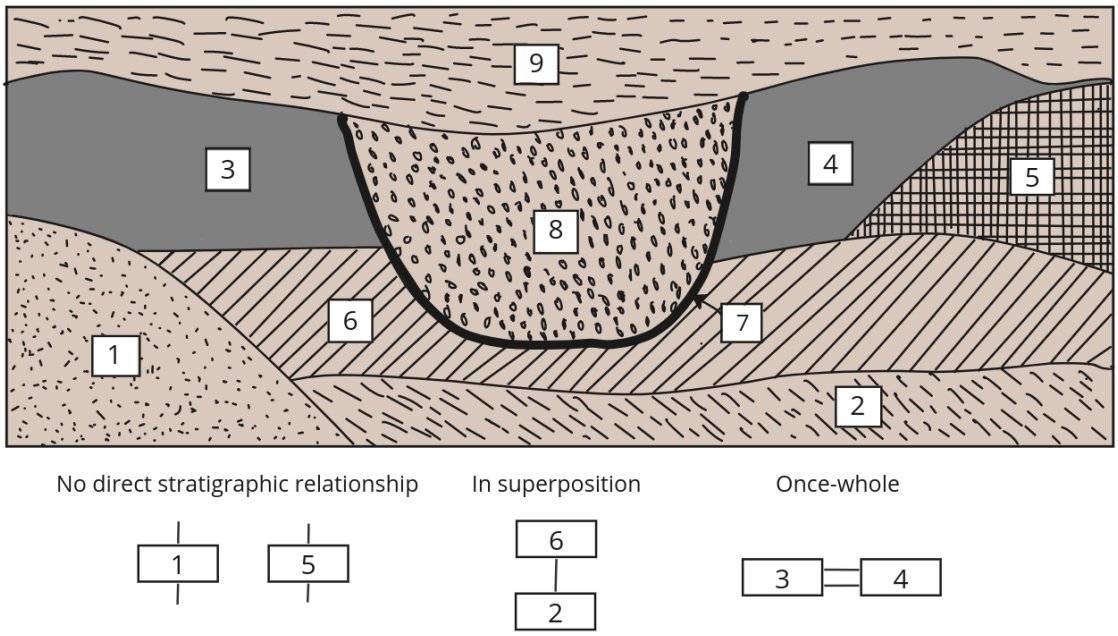


Figure 2.1: Section of an imagined archaeological site showing distinct archaeological strata, the strata labelled [7] is an interface unit (the cut edge of a ditch, and all other strata are deposits. Three stratigraphic relationships are highlighted below the section to demonstrate the three types of stratigraphic relationships.

By following these laws of stratigraphy and recording the relationships between strata on site, a stratigraphic sequence can be constructed, which is a sequence of the deposition or creation (for features) of strata through time (Harris, 1989, Chapter 10).

It is imperative to note here that the archaeological stratification of the site, that is, the order of the strata as observed in the ground, does not always directly correspond to the stratigraphic depth as it does for most geological strata deposited, which are entirely by natural processes. This is because, due to human activity, strata can often be removed completely and redeposited elsewhere or truncated by actions

such as later building or construction. For example, when strata are dug out to make a ditch, they may be cut through and/or redeposited on top of earlier layers of strata. When excavating, archaeologists may recognise signs of human activity and use them in conjunction with their other observations to build a record of the overall stratigraphic sequence.

2.1.3 Single context recording and excavation

The technique now known by archaeologists as single-context excavation is accredited to a range of practitioners working on various sites in Winchester and London over the period 1975-1989 to implement the theoretical methods first published by Harris (Harris et al., 1993). Its purpose is to remove the strata sequentially in the reverse order of their deposition (Roskams, 2002, Chapter 6). Each layer of archaeological strata is removed and recorded as a single stratigraphic unit called a context. Each context is allocated a unique context number (or label) and is recorded, along with its direct stratigraphic relationships observed on site according to the stratigraphic principles defined by Harris and refined by many other archaeological practitioners (e.g. Spence, 1990).

In addition to recording stratigraphic sequence, other recording methods are used, including but not limited to plans and section drawings, which are drawn records of part of an excavation in the horizontal and vertical plane, respectively. For single-context excavations, drawings comprise individual stratigraphic units. Section drawings often record the vertical sides of trenches or vertical cross-sections of specific features and clearly show the stratigraphic relationships and the shape and scale of strata observed on site. Such drawings can be useful when checking a stratigraphic sequence. Post-excavation research is the continued analysis of evidence obtained during an excavation after the excavation has finished.

2.1.4 The Harris matrix

The Harris matrix diagram records each stratigraphic unit recorded during excavation and the stratigraphic relationships between stratigraphic units (contexts). These diagrams consist of boxes representing contexts and edges to connect these boxes, producing a diagram representing the stratigraphic relationships and showing the collective stratigraphic sequence for the whole site. Contexts lower down in the diagrams typically represent those that are older since they are lower down in the stratigraphic sequence. However, the Harris matrix itself is only intended to show physical stratigraphic relationships, which are later converted to temporal relationships during post-excavation analysis.

Harris' **law of stratigraphic succession** results from his observation that when combining multiple partial stratigraphic sequences into a single record for a whole site, the Harris matrix may initially include superfluous edges that should be removed. We demonstrate the law of stratigraphic succession (and the removal of a superfluous edge) for a small example in Figure 2.2.

Following excavation, the resulting stratigraphic sequence (recorded as a Harris matrix diagram or a table of context pairs detailing their stratigraphic relationships) gives us a record of relative dating evidence for the chronology we wish to estimate. The process of building the stratigraphic sequence is more detailed and nuanced than outlined in this section, and we refer the reader to Harris (1989), in particular Chapters 10 and 11, for a detailed account. Further discussion is omitted since we do not construct stratigraphic sequences ourselves.

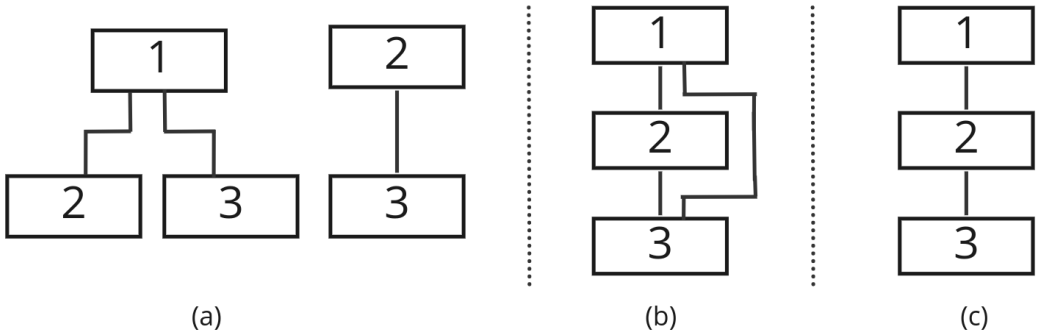


Figure 2.2: Toy example showing (a) two stratigraphic sequences constructed according to the law of superposition, (b) how they are combined without taking into account the law of stratigraphic succession, and (c) the final Harris matrix obtained from applying the law of stratigraphic succession to remove the superfluous edge between 1 and 3.

2.1.5 Representing stratigraphic sequences

Thus far, we have introduced the Harris matrix as a diagram comprising boxes representing contexts, with edges connecting these boxes representing the relationships between the contexts. To a mathematically trained reader, this sounds identical to a mathematical graph. Such graphs comprise nodes (which can represent contexts, just as the boxes on a Harris matrix) and edges connecting such nodes to represent a specific relationship between them. To represent stratigraphic sequences in a mathematical graph, each context would be represented as a node, with edges between nodes representing that one node is above another in the stratigraphic sequence. Therefore, the direction of the edge matters, so a specific type of mathematical graph called a directed graph is required. Such a graph should not contain any cycles since that would indicate that a context is older or younger than itself, which is clearly nonsense. Thus, the type of mathematical graph required for representing stratigraphic sequences is a directed acyclic graph (DAG).

The benefit of representing stratigraphic sequences using DAGs (as opposed to Harris matrices) is that they are widely used for representing other sequential relationships, and so a wealth of algorithms and computer software also exist to allow manipulation and analysis of such graphs. Indeed, using mathematical graphs to display a stratigraphic sequence is not new (see Herzog and Scollar, 1990; Herzog, 1993). However, it was Dye and Buck (2015) who first proposed using DAGs to aid in the automation of chronology construction. We discuss their approach in detail in Chapter 7. For the remainder of this section, we focus on why a Harris matrix is not a DAG and how we can represent stratigraphic sequences as mathematical DAG.

Certain conventions in the drawing of Harris matrices prevent them from being directly translated into mathematical graphs. For example, as first seen in Figure 2.1 and shown again on the left of Figure 2.3, the convention for representing once-whole relationships in a Harris matrix is two parallel edges. Suppose we wanted treat the Harris matrix on the left of Figure 2.3 as a DAG. All edges in the DAG indicate that one context is older than the context it is connected to. Therefore, in a DAG, the two parallel edges between contexts 1 and 3 would create one of two possible relationships. Either, both edges are going in the same direction, which tells us that one context is older than the other rather than once-whole. Alternatively, if both edges go in opposite directions, this would imply a cyclic relationship which is stratigraphically nonsensical and not permitted within a DAG.

To address this, Dye and Buck (2015) propose a simplification of the graphical structure which we adopt in what follows. Nodes representing once-whole contexts are combined into a single node since although they were not identified as a single context during excavation, they are part of a once-whole context; we illustrate this on the right of Figure 2.3. Further discussion of using mathematical graphs to manage

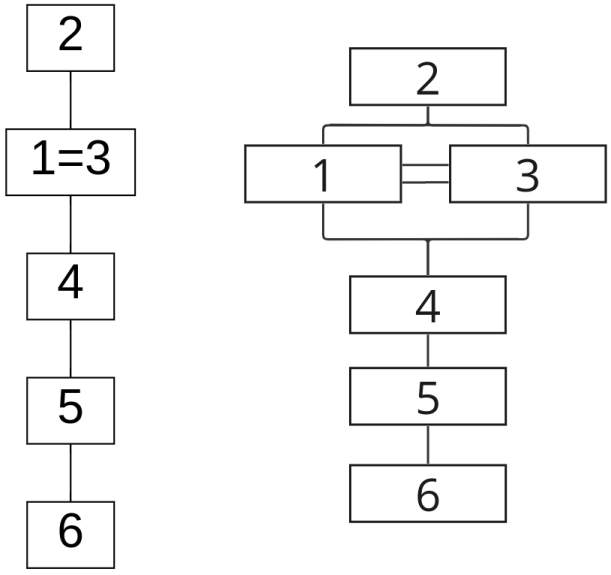


Figure 2.3: *left*: Directed acyclic graph corresponding to the same stratigraphic sequence as the Harris matrix on the left, but nodes representing 1 and 3 have been merged to represent one single archaeological event. Note that in both panels older contexts are below younger ones. *right*: Harris matrix representing a hypothetical stratigraphic sequence such that contexts 1 and 3 are once-whole contexts and, as such, represent events that happened at the same time.

relative dating evidence is provided in Chapter 6. However, to explore the potential for using such mathematical graphs to automate chronology construction, additional background and theory are required.

2.2 Grouping of contexts

Thus far, we have discussed relative dating evidence that derives from physical relationships between archaeological strata. The physical relationships are then converted into temporal relationships during post-excavation, we demonstrate this conversion in Chapter 7 Section 6.3. In addition to recording stratigraphic sequences, archaeologists infer additional relative dating information in the form of temporal

relationships by grouping contexts based on the archaeological evidence observed on site. These groups of contexts will be associated with distinct events in the past. As such, a group will have an associated timespan corresponding to the duration of the collection of events represented by that group. This grouping may happen in a hierarchical nature, with subgroups being grouped inside other groups and so on. Harris describes groups containing contexts as phases, which are grouped into periods. However, as May (2020) describes, archaeologists do not consistently label these groupings. Within this research, we only consider one grouping level, i.e. groups can only contain contexts and not other groups. As such, we refer to any collection of contexts as groups within this research and omit the use of the terms phase and period.

Harris defines two stages to the grouping of contexts within the stratigraphic sequence. The first is building the stratigraphic sequence, as discussed in the previous section, and the next is grouping the contexts into distinct timespans. First, contexts are grouped to represent a period of archaeological activity (Bayliss, 2009). One such example of a period of archaeological activity is the occupation of a particular human settlement.

Grouping of contexts may happen on-site, for example, grouping contexts that represent holes dug in a specific pattern as part of building construction. However, further grouping or revisions of previous groups will most often happen in post-excavation research. Post-excavation research comprises all of the work that follows the end of physically excavating a site through to the publication or archiving of all data and excavation reports. During this part of the process scientific dating occurs in addition to further refining the stratigraphic sequence and grouping based on all evidence observed on-site. Once all contexts have been grouped, then relationships

between these groups should be defined.

2.2.1 Allen operators

Whereas stratigraphic contexts essentially have two relationships, above/below and once-whole, the relationships between any given pair of groups could be one of thirteen. These thirteen relationships correspond to the 13 operators as defined by Allen (1983), which correspond to six temporal relationships, their corresponding inverses and then an operator that defines two groups as equal. These encompass all possible relationships between two temporal intervals (such as groups). These 13 relationships are provided in Table 2.1, alongside their abbreviations.

Abbreviation	**Relation**	**Inverse abbreviation**	**Inverse relation**
>	before	<	after
m	meets	mi	is met by
o	overlaps	oi	overlapped by
f	finishes	fi	is finished by
d	during	di	contains
s	starts	si	started by
=	equals		

Table 2.1: Table of Allen operators that describe relationships between two blocks of time as defined in Allen (1983). Their corresponding abbreviations are also provided.

These 'Allen algebras' are adopted by some archaeologists as a foundation for modelling archaeological groups (e.g. May, 2020; Lucas, 2015). In addition to pictorial representations for each operator, we outline these relationships in Figure 2.4. Typically, archaeologists will define the relationship between groups as one of three: abutting (meets), gap and overlap. These correspond to six of the Allen algebras (including their inverses), and we refer to these as between-group relationships. The remaining seven Allen algebras correspond to relationships between groups and

subgroups contained within them, and as such, we define the remaining seven relationships as between-group relationships. In this research, we only utilise relative dating evidence that uses groups and omit the use of subgroups. Thus, we only utilise the between-group relationships when carrying out chronology construction.

Thus far, we have introduced the relevant methods of obtaining the relative dating evidence that is used in archaeological chronology construction. Further, we introduced the concept of using mathematical graphs, specifically DAGs, to represent stratigraphic sequences. This idea is developed in greater detail in Chapters 6 and 7. The other key component we use in chronology construction is absolute dating evidence which is the focus of the next chapter.

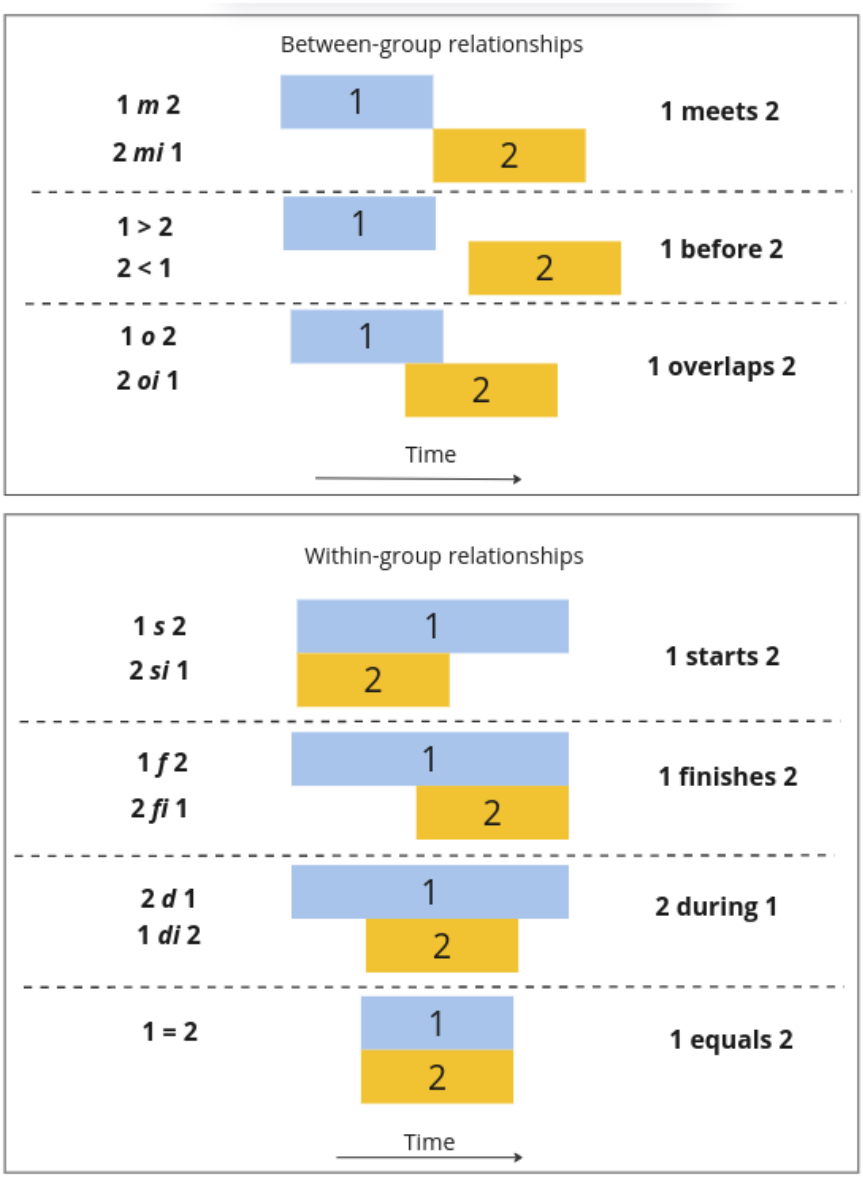


Figure 2.4: A diagram showing the 13 different Allen algebras (to the left) with a pictorial representation such that the rectangles represent blocks of time (middle) and the corresponding relationship between archaeological groups 1 and 2 that each Allen algebra represents (to the right) (Allen, 1983). Abbreviations used are defined in 2.1.

Chapter 3

Absolute dating evidence

Stratigraphic sequences, in addition to grouping, provide us with relative dating evidence for a given excavation. However, we can gather additional information from artefacts found during the excavation, commonly referred to as finds. Common types of finds include bones, pottery and coins. During single-context excavation, archaeologists excavate finds and artefacts and take samples as appropriate. Archaeologists record the context, and often the precise 3D location, in which the finds were discovered or where the sample was located. Certain types of finds and samples can be dated by absolute dating methods, e.g. organic materials such as bone and wood. Absolute dating gives a numerical estimate of the age of a sample or a find and, thus, an approximation of the calendar age of the context from which the find or sample originates, effectively anchoring the stratigraphic sequence to a point in time. However, this is only true if the find dates the context in which it was found. Note that this is not always the case, and how to address this problem is discussed in Chapter 6.

Scientific dates are often used to provide absolute dating evidence, but artefact experts such as coin or pottery experts can also provide absolute dates based on inspection of artefacts. Our research focuses on radiocarbon dating as a source of absolute dating evidence due to the interesting statistical problems it leads to. Radiocarbon dating provides an absolute date estimate for an organic sample and is a method of scientific dating carried out by specialist laboratories.

However, the radiocarbon determinations provided by radiocarbon dating laboratories are not directly equivalent to calendar year estimates. Therefore, conversion of such dates to the appropriate scale is required. Converting from the radiocarbon scale to the calibrated scale (an approximation of the calendar scale), which we refer to as calibration, requires a calibration curve that must be estimated. In this chapter, we provide motivations for such a curve and outline the history of constructing such calibration curves. Following these initial explanations, we outline the calibration of a single radiocarbon determination.

3.1 The birth of radiocarbon dating

Willard Libby first suggested the existence of radioactive carbon in organic material (Anderson and Libby, 1946) and used this to attempt to date organic samples (Libby et al., 1949), subsequently developing radiocarbon dating as a method for dating organic material (Libby, 1955). It is no exaggeration to call this discovery a revolution (Bayliss, 2009), which earned Libby the Nobel prize for Chemistry in 1960 (NobelPrize.org, 2022). His work on the radioactivity of organic materials and the development of radiocarbon dating theory had significant implications for archaeology. While stratigraphy allows us to construct a relative chronology, radiocarbon dating provides an anchor for that chronology, placing it in an approximate absolute

position on the calibrated scale.

3.1.1 Radioactive decay

To understand how radiocarbon dating works, we must introduce the concepts of isotopes and radioactive decay. Every chemical element consists of atoms that have a nucleus. The nucleus of an atom consists of positively charged protons and neutrons with no charge. Though the number of protons in the nucleus of a specific element is fixed and determines which element the atom comprises, the number of neutrons can vary, leading to isotopes of an element. An isotope is represented using the chemical symbol for its element, with a left-aligned superscript representing the combined number of protons and neutrons. For example, the element carbon has 6 protons in its nucleus and has three isotopes: ^{12}C, ^{13}C and ^{14}C, with six, seven and eight neutrons, respectively (see Bowman, 1990). Only a tiny part of a given sample of carbon will be ^{14}C, less than 10^{-12}% (Bronk Ramsey, 2008) However, the relative proportion of ^{14}C to ^{12}C in a particular sample can be estimated using methods discussed in the following section.

Some isotopes are unstable due to the excess energy stored in the nucleus. Such isotopes are termed radioactive and will decay over time into a more stable isotopic form; ^{14}C is an example of an unstable radioactive isotope (as such, it is frequently referred to as radiocarbon, hence the name radiocarbon dating). The law of radioactive decay, first proposed by (Rutherford and Soddy, 1903, Section 4) for all radioactive elements, allows us to model the decay of ^{14}C using the equation:

$$A = A_0 e^{-\lambda t}, \tag{3.1}$$

which gives us the number of ^{14}C atoms A in a sample at time t, as defined in

(Bowman, 1990, pg 11), such that A_0 is the number of ^{14}C atoms at time $t = 0$.

Further, we have the constant λ, which is unique for each isotope and corresponds to the rate at which a sample decays. Each radioactive isotope has a half-life, $T_{1/2}$, which is a measure of how long it takes for half the atoms in a sample of radioactive material to decay into a more stable atom. Section 3.1.4 discusses estimating the radiocarbon's half-life. The decay constant λ is proportional to the reciprocal of the half-life such that $\lambda = \ln(2)/T_{1/2}$ (Bowman, 1990, pg 11). Consequently, given Equation 3.1, provided we can obtain values for A, A_0 and the half-life $T_{1/2}$ (thus giving us λ), we can calculate the time elapsed since a sample of radioactive material began to decay. It was this theory that Willard Libby utilised when he proposed using radiocarbon dating to date organic matter.

3.1.2 Radiocarbon in living organisms

Until death every living organism (such as humans, plants and animals) interacts with the Earth's life-sustaining biosphere (see Thompson et al., 2022). Libby suggested that living organisms absorb the isotopes ^{12}C, ^{13}C and, most importantly, the radioactive isotope ^{14}C during this interaction (Libby, 1955). Libby further proposed that while alive, the absorption and decay rates of ^{14}C in an organism reach a state of equilibrium with their environment, but that upon death, only the decay process remains. Therefore, by calculating the extent of the decay of ^{14}C within an organic sample, using Equation 3.1 and letting the year of death of an organism correspond to $t = 0$, we can determine the time elapsed since death. Thus, radiocarbon dating was born.

3.1.3 Measuring the number of radiocarbon atoms

Determining the age of a sample requires the half-life of ^{14}C and being able to estimate the number of ^{14}C atoms both now and at death. The true number of ^{14}C atoms in a sample cannot be determined exactly, but experimental values can be derived to estimate it, as discussed in Bowman (1990) Chapter 3. Libby et al. (1949) utilised beta counting to detect ^{14}C in a sample. When ^{14}C decays into the stable isotope ^{14}N, it emits a beta particle which can be detected, allowing laboratories to estimate the number of ^{14}C atoms in a sample.

However, accelerator mass spectrometry (AMS) has since provided a far more precise method for estimating the number of ^{14}C atoms in a sample (Bronk Ramsey et al., 2004), uses the fact that different isotopes have different atomic weights. When applying a magnetic force to a carbon sample, lighter carbon isotopes are deflected more than heavier ones allowing detectors to count the number of atoms of each isotope.

Both AMS and beta counting have experimental errors. As Bowman (1990), Chapter 3 discusses, these errors are usually evaluated by repeating experiments a large number of times and measuring the variation in the measurements of ^{14}C atoms. However, due to the cost of repeating experiments and limited sample sizes, this is not feasible for radiocarbon dating. So instead, the laboratory errors are approximated. An in-depth discussion of how this is done is provided in Scott et al. (2007).

An important point to highlight is that, strictly speaking, the laboratory error on a radiocarbon determination is not symmetrical. However, for samples of most material obtained during excavation, this asymmetry is small. Therefore, it is standard practice for a laboratory to provide a radiocarbon determination as $x \pm \sigma$

such that x is the estimated radiocarbon age of the sample, and σ is the laboratory error assumed to be symmetrical about x and equivalent to 1 standard deviation. Further details of the methods used to estimate the number of ^{14}C atoms in a sample (and the errors these methods produce) are omitted here but can be found in (Bowman, 1990, Chapter 3). We outline the statistical relationship between a radiocarbon determination $x \pm \sigma$ and the true number of ^{14}C atoms in the sample in Section 3.2.3, but some additional theory is required first.

Whilst we cannot directly measure A_0, the number of ^{14}C atoms in the sample at death, Libby (1955) assumed that the concentration of ^{14}C in the atmosphere and biosphere of the earth had remained constant over time and space. Thus he proposed using modern samples to estimate the number of ^{14}C atoms in an organic sample at death (Libby, 1955, see Equation 5). Therefore, all that remains to utilise the law of radioactive decay for dating is to calculate the decay constant, which requires the half-life of ^{14}C.

3.1.4 Half-life of radiocarbon

To calculate the decay constant λ in Equation 3.1, Libby took the half-life of ^{14}C to be 5568 ± 30 years (Johnson et al., 1951), and this half-life (known as Libby's half-life) was used for over a decade. Researchers have since proposed more accurate estimates for the half-life of ^{14}C, and the currently accepted value is 5700 ± 30 (Kondev et al., 2021). However, using any half-life other than Libby's half-life produces inconsistent results compared to previously published radiocarbon ages. The research community agreed that they would proceed with the less accurate but, importantly, consistent half-life of 5568 ± 30 for continuity with previously dated samples (Godwin, 1962). While it may seem odd for the radiocarbon dating community to do this it is, in fact, of no consequence since we can correct the discrepancy in the half-life during the

step that converts radiocarbon determinations (and their errors) to the calibrated scale.

3.2 Calibration of radiocarbon determinations

The revolution of radiocarbon dating was not without its obstacles. Libby's theory relies on the concentration of ^{14}C in the atmosphere and biosphere remaining approximately constant over time so that we can estimate the number of ^{14}C atoms at $t = 0$ (death) using modern samples. Libby acknowledged early on that this assumption might not hold, and de Vries (1958) confirmed this by radiocarbon dating tree-ring samples of known calendar ages. Further studies from various laboratories provided further evidence (see Reimer, 2021, for details) of the variation of ^{14}C in the atmosphere and biosphere over time and space.

Due to the varying concentration of ^{14}C in the atmosphere and biosphere a sample's radiocarbon age, i.e. the time elapsed since the organism died (equivalent to t in equation 3.1), is not equivalent to the number of calendar years since death, which is its calendar age. Therefore, to be able to interpret radiocarbon determinations, a function is required to convert between the radiocarbon scale and the calendar scale. Let θ denote the true calendar age of a given sample (cal BP) and $r(\theta)$ denote its radiocarbon age (BP). A deterministic relationship exists between any given calendar age and its true underlying radiocarbon age. This is because any number of samples with the same calendar age will have experienced the same rate of ^{14}C decay and, therefore, will have the same radiocarbon age (Blackwell and Buck, 2008).

However, the factors that affect the concentration of ^{14}C in earth's biosphere are too complex to determine the exact function $r(\cdot)$, such that $r(\cdot)$ denotes a function which can calibrate any given value of θ to its true radiocarbon age $r(\theta)$ (see Bowman,

1990, Chapter 2). Instead, data and statistical curve-fitting methods are used to approximate $r(\theta)$. Estimates of the relationship between θ and $r(\theta)$ are known as radiocarbon calibration curves, and we denote such calibration curves as $\mu(\cdot)$, such that $\mu(\cdot)$ is an approximation of $r(\cdot)$.

3.2.1 Data used to construct calibration curves

To construct a calibration curve, dating methods that can provide reliable calendar age estimates for an organic sample are required. That same sample can also be radiocarbon dated, giving an associated radiocarbon determination. Hence, each sample can be represented by a data pair. Datasets of these data pairs can then be used as input for models to estimate calibration curves. Dendrochronology is a commonly used method of obtaining calendar age estimates for organic samples. This method compares the widths of tree rings resulting from annual growth patterns to a master chronology to provide a calendar age often accurate to one calendar year, for a sample of wood. For a comprehensive guide to dendrochronology, see Baillie (1982).

Initially radiocarbon-dendrochronology data pairs were used to produce tables of radiocarbon corrections (Stuiver and Suess, 1966). Following this, various local short-scale calibration curves were formed by radiocarbon dating wood samples that had been dendrochronologically dated and published as tables of corrections, e.g. Suess (1967, 1970) and Switsur (1973). While the early calibration curves were suitable for the Northern Hemisphere, further calibration curves were eventually provided for the Southern Hemisphere (Lerman et al., 1970) and marine samples (Stuiver et al., 1986) due to regional variations in atmospheric ^{14}C concentration.

Reimer (2021) outlines the evolution of all the calibration curves over the past

several decades. A critical development, however, was the first internationally agreed-on calibration curve (Stuiver and Pearson, 1986; Pearson and Stuiver, 1986), which led to the development of the IntCal working group (IWG). This group is dedicated to modelling internationally-agreed calibration curves for marine samples and Northern and Southern Hemispheres samples. These calibration curves are continually updated and published when improved data or curve-fitting methods become available, the most recent being the IntCal20 curve (Reimer et al., 2020), SHCal20 (Hogg et al., 2020), and MarineCal20 (Heaton et al., 2020b).

3.2.2 Key details of IntCal calibration curves

The data used by IntCal to produce curve outputs are often referred to as IntCal datasets. These datasets comprise pairs of calendar age estimates and radiocarbon determinations. Sources for obtaining such data are many and varied, so inferring an estimate of the curve from the resulting database is a major statistical undertaking in its own right, (Buck and Blackwell, 2004; Heaton et al., 2020a), but such methods are not discussed further here since we do not construct our own calibration curves in this thesis. Instead, we use the latest Northern Hemisphere calibration curve estimate, IntCal20 (Reimer et al., 2020).

It is important to note that for other sources of calendar age estimates, there are errors on both the calendar age estimates and the radiocarbon determinations which must be considered when estimating calibration curves. For initial estimates of the calibration curves, error on the calendar scale was not included in the modelling process. However, more recent curves have included errors from calendar age estimates within the IntCal datasets, and now this is standard practice (Buck and Blackwell, 2004; Reimer et al., 2004, 2013, 2020). Alternative methods for obtaining calendar date estimates are not discussed here since we focus on radiocarbon dating

but see Reimer (2021) for a discussion of other dating methods that are used to provide data from which calibration curves can be estimated, and Reimer et al. (2020) for a detailed discussion of the data used to construct the most recent Northern hemisphere calibration curve.

The calibration curves published by IWG are our best available estimates of $r(\theta)$. Strictly speaking, the Northern hemisphere, Southern hemisphere, and Marine calibration curve estimates correspond to approximations of distinct functions. We might denote these estimations $\mu_N(\theta), \mu_S\theta)$ and $\mu_M(\theta)$ respectively. However, to simplify notation, using $\mu(\theta)$ to represent a chosen estimate of a particular calibration curve is standard practice. A further note of importance is that, due to the uncertainty in the calendar age estimates in the IntCal data, the calibration curve output that corresponds to the calendar scale is, in fact, an approximation to the calendar scale which is why we refer to it as the calibrated scale. Since all approximations of true calendar ages for the remainder of this thesis will be obtained using a calibration curve, all future references to θ will refer to the true calibrated age of a sample.

All IntCal calibration curves are published as five columns of model output, with the first column consisting of selected calibrated radiocarbon ages on an irregular grid for which the IntCal team have provided curve output. For IntCal20, these are in the range of 55,000-0 cal BP and are yearly for the later part of the calibration curve but become more sparse for the earlier parts of the curve. Two of the five columns in the IntCal calibration curves correspond to curve output modelled in domains other than radiocarbon or calibrated age domains. See Heaton et al. (2020a), Section 3.2.1 for definitions of these domains and details of such conversions. However, since we work only in the radiocarbon age and calibrated age domains when calibrating

archaeological radiocarbon determinations, further discussion of domains other than radiocarbon and calibrated age is omitted.

For each calibrated age θ in the IntCal published curves, there are two columns which are in the radiocarbon age domain. The first of these represents a pointwise estimate of $r(\theta)$, $\mu(\theta)$. The second provides a pointwise error $\delta(\theta)$ for each $\mu(\theta)$. The pointwise error, assumed to be symmetric about $\mu(\theta)$, accounts for the uncertainty involved in estimating the calibration curves. This uncertainty is partly due to the uncertainty in the data used to estimate the curve - in particular, the laboratory error for radiocarbon dates, in addition to any errors on the calendar dates in the IntCal20 datasets. There is further uncertainty due to the curve fitting methods being statistical. Finally, the most recent curve estimate, IntCal20, provides variances that correspond to predictive intervals, as opposed to confidence intervals, to account for the uncertainty not able to be quantified in the laboratory errors (such as regional differences between samples used for data in the IntCal datasets) and any additional unseen variation in radiocarbon determinations that will be calibrated by the IntCal20 curve. The pointwise means and variances are provided for yearly intervals of θ for the early calibration curve and larger intervals for older parts of the curve.

Using linear interpolation on the published IntCal20 curve outputs, for sections where annual estimates are not provided, we obtained an interpolated output of the pointwise estimates $\mu(\theta)$ with error $\delta(\theta)$ for each integer calibrated age between 55,000-0 cal BP as seen in Figure 3.1. Generally, users will use linear or cubic interpolation on the IntCal curve outputs. We chose linear interpolation to avoid over-smoothing the curve. When plotting the interpolated curve, we see a non-monotonic relationship between the radiocarbon scale and the calibrated scale. This

relationship was, in fact, observed in early estimates of calibration curves, with Suess (1970) demonstrating what is referred to in the literature as 'wiggles' in the structure. His lack of justification for constructing such a curve resulted in a failure to convince early researchers. However, further dendrochronology dating confirmed that Suess' 'wiggle' structure was correct (Reimer, 2021).

Each true calibrated age θ has a deterministic relationship with its true radiocarbon age $\mu(\theta)$. However, due to the 'wiggles' in the calibration curve, a $\mu(\theta)$ value does not uniquely map to a value of θ on the calibrated scale. Figure 3.1 demonstrates a section of the interpolated IntCal20 calibration curve estimate (Reimer, 2021). Note that the inset in that figure highlights the lack of a one-to-one relationship between the radiocarbon scale and calibrated scale for the radiocarbon age 5300 BP. We will now discuss how to use the most recent calibration curve $\mu(\cdot)$ (IntCal20) to calibrate a single radiocarbon determination, taking into account the non-monotonic nature of the calibration curve and the uncertainty on the curve.

3.2.3 Calibration of a single radiocarbon determination

The one-to-many relationship between the radiocarbon scale and the calibrated scale complicates the calibration of radiocarbon determinations. However, it is relatively simple due to the assumptions about the nature and the various sources of uncertainty and by utilising likelihood functions. Given a sample with a true calibrated age of θ and true radiocarbon age of $r(\theta)$, a radiocarbon determination $x \pm \sigma$ is a realisation of a random variable X such that X is an approximation of $r(\theta)$. Given that our best approximation to $r(\theta)$ is the calibration curve $\mu(\theta)$, we must also take into account the error on the curve $\delta(\theta)$. Recall from Section 3.1.3, that the laboratory error σ on a radiocarbon determination x is assumed to be symmetrical. It is standard practice to assume that X is normally distributed, conditional on θ

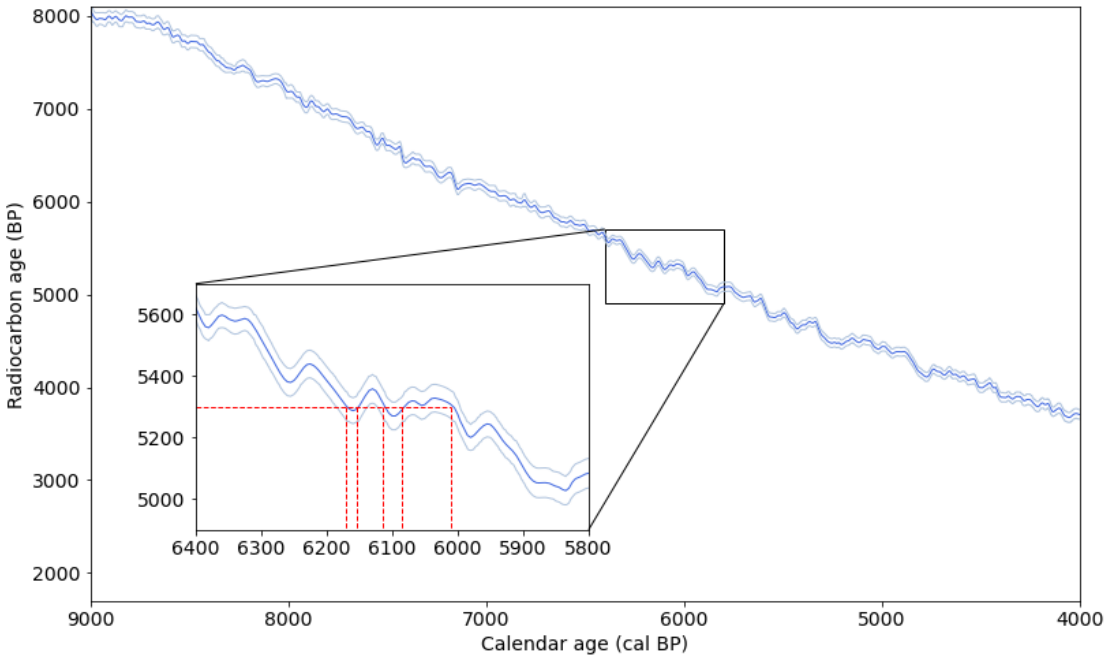


Figure 3.1: A section of the IntCal20 calibration curve $\mu(\theta)$ (navy) $\pm 2\delta(\theta)$ (light grey) given for calibrated ages θ between 9000-4000 cal BP. Inset zooms in on the calibration curve in the range 6400-5800 cal BP, showing a lack of unique mapping for the radiocarbon age 5300 BP to the calibrated scale

with mean $\mu(\theta)$ and variance $\sigma^2 + \delta(\theta)^2$ (Blaauw, 2010; Nicholls and Jones, 2001).

Given the assumptions mentioned above, we know that the probability distribution of data $x \pm \sigma$ given parameter θ is the probability density function of a normal distribution with mean $\mu(\theta)$ and variance $\sigma^2 + \delta(\theta)^2$. However, we are interested in the true value of θ. Therefore, we use the likelihood function $L(\theta; x)$, which gives us the likelihood of each possible value of θ being the true value, given fixed data. Both $L(\theta; x)$ and $p(x|\theta)$ are the same function up to proportionality. However, $L(\theta; x)$ is a function of θ instead of a function of x. Using the probability density function of

the normal distribution, we get

$$L(\theta; x) \propto \exp\left(-\frac{(x-\mu(\theta))^2}{2(\sigma^2+\delta(\theta)^2)}\right) \tag{3.2}$$

up to proportionality. Note that for simplicity of notation we avoid explicitly conditioning of σ, $\mu(\cdot)$ and $\delta(\cdot)$ in the likelihood notation $L(\theta; x)$ as in Christen (1994). However, in all notation from this point, the use of data x includes the associated laboratory error σ and a calibration curve, i.e. $\mu(\cdot)$ and $\delta(\cdot)$.

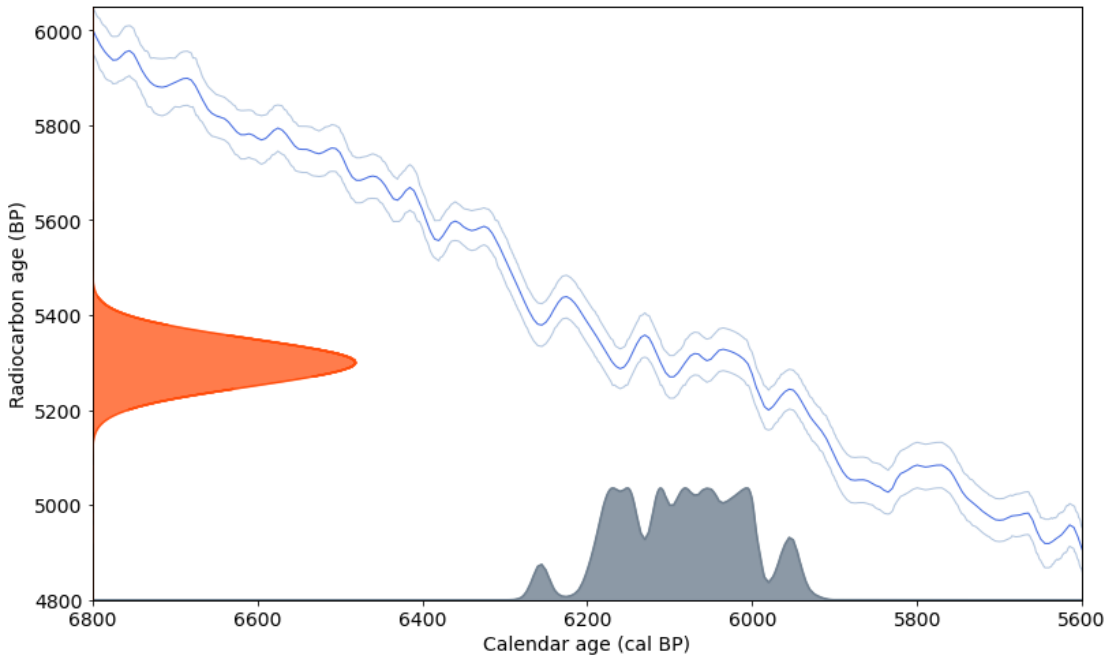


Figure 3.2: Plot of the IntCal20 curve, showing calibration of the (normally distributed) radiocarbon determination 5300±50 BP (orange), to a probability distribution on the calibrated scale (grey). Note, for both probability distributions, the curve's height represents probability density (with scale for probability not provided).

Summary statistics for calibrated radiocarbon determinations

By evaluating the likelihood function for all integer calibrated ages in the range 50,000-0 cal BP and normalising, we can obtain a discrete approximation to $p(\theta|x)$. However, calibrating the radiocarbon determination 5300 ± 50 BP, in Figure 3.2, results in a probability distribution on the calibrated scale that is now multi-modal. Therefore, summary statistics such as mean and variance are not particularly informative and should not be used.

Instead, we find the highest posterior density (HPD) interval R for the calibrated age θ. An HPD interval is the interval, or set of intervals, such that for all $\theta \in R$ and $\theta' \notin R$,

$$p(\theta|x) \geq p(\theta'|x),$$

meaning all calibrated ages within the interval(s) have a higher probability of being the true value of θ than any outside the interval, and the cumulative probability of all the values within the interval should sum to some limit set by the user. This limit is usually 0.95, corresponding to 95%, as it is with confidence intervals. Thus, calibrated ages are published after calibration as an HPD interval (or set of intervals) instead of an estimate and error.

The availability of calibration curves solved the problem of calibrating radiocarbon determinations to the calibrated scale. However, the HPD interval of the calibrated radiocarbon determination can be rather large. This is particularly true for older samples with larger errors or radiocarbon determinations calibrated by flatter parts of the calibration curve. Fortunately, archaeologists can use more than absolute dating evidence when constructing the chronology of a site. Chapter 5 introduces how both absolute and relative dating evidence can be utilised in chronology construction, which allows us to improve the precision of calibrated age estimates within

a chronology using the method of statistical inference known as Bayesian inference.

Stratigraphic sequences, along with scientific dates and, ideally, grouping, are the three components we required for the chronology construction carried out in our research, as discussed in this chapter and the previous one. Despite single-context excavations being standard in the UK and the widespread use of the Harris matrix for displaying stratigraphic sequences, these components are not always archived well, if at all, in digital archives. This lack of archived relative and absolute dating evidence presented a problem for our research since we required such evidence for testing and case studies. The following chapter explores the results of a data review searching for such dating evidence in digital repositories within the UK and the data that we proceeded to use for case studies.

Chapter 4

Obtaining archaeological dating evidence

Archaeological excavation often produces large volumes of physical finds/samples and data in analogue and digital formats. Increasingly, much of these data are recorded and managed digitally (so-called born-digital data) or are archived in a digital format when deposited in a repository. Relative and absolute dating evidence and the associated data are produced and augmented at various stages in the excavation and post-excavation process. Careful consideration is required upon archiving to ensure the data will be suitable for reuse.

For all disciplines, it is imperative that data are reusable since it allows for the reproducibility or reanalysis of results and data. However, the reusability of data is of the utmost importance in archaeology. By and large, archaeological excavation is destruction (Wheeler, 1956). Gathering data during an excavation predominantly destroys the source of that data permanently. Even if the data source is not entirely destroyed, such as samples used for scientific dating, any given sample is of finite size. Therefore, only a limited amount of scientific dating or other destructive analysis

can be carried out.

Many archaeologists in the research community urge that archived archaeological data must adhere to the FAIR principles (Wilkinson et al., 2016), with FAIR referring to data being Findable, Accessible, Interoperable and Reusable. Digital archives put significant effort into ensuring data can be found within their repositories, easily accessed from them, and that users can query metadata to ensure data are interoperable. Nicholson et al. (2023) Table 1 discusses in detail how to ensure data adhere to the FAIR principles. Nicholson et al. (2023) also discuss how the archaeological community might improve the reuse of archaeological data in general. In this chapter, we will examine how the process of ensuring digital archaeological data are findable, accessible and interoperable does not, in itself, automatically enable reusability. The resulting lack of reusable data in archaeology can prevent advancements in software and methodologies since well-understood existing data are needed for testing and case studies.

As discussed in Chapter 1, Section 1.2, the research undertaken for this thesis was funded via a Collaborative Doctoral Partnership (CDP) between Historic England and AHRC. One of the conditions of this funding was that we would carry out a systematic data review of the quality and utility of resources in digital heritage repositories in the UK. This data review complemented our research objectives well since a key goal of the project was to semi-automate the complete chronology construction process from data gathering on an archaeological excavation through to Bayesian chronology construction (see Figure 7.1 for an illustration of the process and Chapter 5 for theoretic details). We required case studies of previously excavated sites with well-archived relative dating evidence and, ideally, absolute dating evidence such as radiocarbon determinations to accomplish this. Since a wide-ranging and represent-

ative sample of case-study data sets was sought for testing the models and methods developed later in the research, it was decided that a systematic data review would be undertaken to search for the required data rather than an ad-hoc approach. This allowed us to provide a snapshot of the quality and quantity of the archaeological data available in the archaeological digital archives, as Historic England required.

This extensive data review formed an internal report for Historic England and was later published alongside additional case studies of digital archaeological data reuse in Moody et al. (2021). This chapter briefly examines recent literature on the reuse of digital data in archaeology, though a more detailed review is provided in Moody et al. (2021). In addition, we signpost the reader to examples of other research where the reuse of absolute and relative dating evidence has added value to an excavation. Following this, we describe the methodologies used in our data review and highlight key results from the report written for Historic England and the current author's contribution to Moody et al. (2021). Finally, we discuss how our research seeks to mitigate some of the issues of digital data management and reuse encountered during the data review.

4.1 Digital repositories in archaeology

Increased computing power has resulted in the ability to produce, manipulate and store large volumes of digital data during excavation and post-excavation research. All digital data would ideally be archived within private or public archives. However, some data can be kept by the excavating organisation and never archived. Moreover, the excavating organisation dictates when digital data are deposited, and they may deposit data in stages or wait until the end of a project, if at all. Long-term digital preservation is considered particularly important if excavation data is only recorded

digitally. Data such as GIS plan drawings, CAD section drawings, numerical data, spatial coordinates, or PDFs of archaeological reports are often recorded as digital data and thus kept in digital repositories. For digital archiving to be practical, any data should be secure, files should be in a format that is non-obsolescent, and it should be readily accessible in the future if needed.

The Archaeology Data Service (ADS) is a public digital repository for heritage data; they have continually worked to ensure digital archaeological data are archived to ensure the longevity and access of the files (Richards, 2008). Since its formation in 1996, the ADS has provided the leading public digital archive for archaeological data in Great Britain. They accept a wide range of archaeological data, including but not limited to databases, computer-generated images, photographs, audio files and reports in `PDF` format (The Archaeology Data Service, 2020).

In addition to the ADS online repository, the ADS launched the OASIS project in 2004. The OASIS (Online AccesS to the Index of archaeological investigationS) project is managed by the ADS in partnership with public heritage bodies in England, Scotland (and to lesser degrees Northern Ireland and Wales). It provides a central digital repository for archaeological grey (unpublished) literature. Additional repositories exist, such as DANS (Data Archiving and Networked Services) in Denmark (Hollander, 2017), tDAR (the Digital Archaeological Record) in the US (McManamon and Kintigh, 2017), and SND (Swedish National Data Service) in Sweden (Swedish National Data Service, 2023), see Nicholson et al. (2023) Table 2 for an extensive list of sources of digital archaeological data.

Despite making significant progress in preserving digital data, the ADS Director and colleagues acknowledge that more work is needed to ensure such data are reusable. Indeed, the ADS director, Richards (2017), describes how digital fieldwork archives

seldom have all the required data for further analysis or reuse. Richards (2017) also recognises that cost is a key factor in preventing the deposition of digital data with the ADS. However, he argues that depositing a comprehensive digital archive should be part of the standard workflow upon project completion. Since 2019, the SEADDA (Saving European Archaeology from the Digital Dark Age) project has been working on improving the archiving of digital archaeological data with partners, including staff from the ADS in over 26 countries (SEADDA, 2023).

The minimum data required to allow for reuse depends on the data type. In section 4.2.1, we outline the type of data required when chronological modelling is the goal. However, this is just one of an enormous range of post-excavation projects that might seek to reuse data. What is clear is that defined standards for the management and archiving of all types of digital archaeological data that can be produced during excavation are required to ensure the potential reuse of such data. These standards should consider all stages of the data cycle, as specified in Yakel et al. (2019), from collecting and recording the data during excavation to management and reuse during the post-excavation process.

4.1.1 Data management standards in archaeology

Additional research on digital data management in archaeology, such as Faniel et al. (2013) and Huggett (2018), agree that although standards have been established for storing and preserving digital data so that they are accessible in the future, reuse issues still need to be addressed. The lack of standardisation in how digital archaeological data are recorded and managed presents substantial problems. Archaeology in the UK is carried out in both the research and commercial sectors by at least ten major archaeological contracting organisations, a multitude of smaller operators, and many individual archaeological consultants and specialists. This results in widely

varying excavation, post-excavation, analysis and publication practices. Further, variation in data outputs is exacerbated by the many different types of archaeology encountered (e.g. the scale of excavation, the complexity of site stratigraphy, investigation methods, project type, and resources available). As a result, despite the ADS providing a central repository for depositing digital archaeological data, the data are not always in an optimal format for reuse, if deposited at all.

Throughout the process, from the start of the excavation to the publication of the final excavation report, there are multiple steps where data may be recorded, managed or stored in a way that facilitates or hinders reuse. Faniel et al. (2018) highlights a lack of formal archaeology database creation and management training. There is little incentive to improve this since training can be costly and time-consuming, with no immediate benefit to those working on an excavation. However, without sufficient training, it is difficult for those collecting and recording data to see how their decisions impact the usability and reusability of data.

Faniel et al. (2018) discusses the data management practices at two European excavations. They find that the aforementioned issues were present in their case studies, noting that a high staff turnover contributes to the lack of incentive to train staff thoroughly in database management. They also comment that humans do not "think like a database". As a result, it is difficult for them to understand how data-gathering practices can cause problems with data reuse. For example, humans can readily identify that the two identifiers "A203" and "a203" might relate to the same thing. However, a digital search algorithm may interpret these as two unique identifiers. Even if a human is interpreting the data, they may not be confident that the identifiers represent the same thing if they are unfamiliar with the original excavation. Typically, subtle issues like this accumulate and (taken together) result

in data unsuitable for reuse without significant effort.

Guidelines, such as Historic England's recording manual (Historic England, 2018), seek to ensure data collection practices are uniform within excavations conducted by their staff. However, those carrying out the excavations are responsible for deciding how they wish to record and manage their digital data. Moreover, even if data are collected and managed well on-site, management and manipulation of data during post-excavation analysis can also prevent reuse if, at the very least, a record of precisely how the data were used during the post-excavation process is not archived. Ideally, the raw digital data (as collected on-site) would always be archived within a digital repository, along with workflow summaries, models, algorithms, resulting outputs and notes for future users linking archived materials to any formal reports and/or grey literature.

Further, data stored in a proprietary format can prevent reuse, particularly if the software required to access the data must be purchased. The ADS open standards (The Archaeology Data Service, 2020) address this by allowing only specific file formats to be deposited. However, if data are kept with a private organisation, they may choose to store data in a proprietary format, even if they later make it publicly available. Additionally, if all the necessary data are not publicly available, this can also prevent reuse. These last issues particularly affect early-career researchers who may not have the necessary personal contracts or funds to locate or read the required data.

4.1.2 Value in reusing relative and absolute dating evidence

Although there are multiple scenarios in which it is useful to revisit archived digital archaeological data, in the remainder of this thesis, we focus on the value of reusing

chronological data, such as relative and absolute dating evidence. There are multiple reasons why such reuse might be of value. For example, increased computing power, or new statistical methodologies, can enable more complex analyses to be carried out, resulting in revised or new models. We might also wish to revisit data from an excavation if further scientific dating occurs, e.g. after new excavations have been undertaken or when the research focus changes in a follow-up stage of excavation. Any new dating augments the original data and might allow for a revised chronology to be provided, such as in Marciniak et al. (2015), where a new dating program changed the understanding of the Late Neolithic community of Çatalhöyük.

Furthermore, experts may disagree on the best chronological model for a given site. Therefore, it is essential to have data, associated models, and methods available in a trusted public digital repository to allow researchers to revise or revisit previous work, see Discamps et al. (2015); Higham and Heep (2019). Bayliss et al. (2014) remodelled the chronology of Buildings 1 and 5 from the north area of Çatalhöyük, initially explored by Cessford et al. (2005). The Cessford and Bayliss models are substantially different, resulting in quite different archaeological conclusions. Dye and Buck (2015) used this scenario to illustrate why it is beneficial to consider multiple chronological models for any given archaeological project (which is the focus of the research presented in Chapter 6 onwards). They proposed using mathematical graphs to represent and semi-automate the construction of such models. This work by Dye and Buck (2015), followed by the timely interest in the topic from Keith May, led to the research reported in this thesis being funded. The aim of this project is to provide prototype software to improve the management and modelling of relative and absolute dating evidence for use in Bayesian chronology construction.

We required data for testing our prototype software and for use in case studies

presented later in this thesis. An initial search by Keith May for data that might be useful for this PhD (during the project's design phase) using the front-end web page of ADS (https://archaeologydataservice.ac.uk/) proved slow, despite knowledge of archaeological sites likely to be suitable. We thus approached ADS, who generously spent time performing searches on their metadata to provide us with a subset of the files stored with the ADS that were most likely to contain the required data, which we systematically reviewed.

Figure 4.1: Diagram of the end-to-end process from collecting absolute and relative dating evidence to completing Bayesian chronology construction. The blue and yellow boxes represent the process of obtaining relative and absolute dating evidence, respectively. The red boxes outline the process of Bayesian inference.

4.2 Utility of relative and absolute dating evidence in ADS repositories

The primary objective for the data review described in this chapter was to seek all absolute and relative dating evidence within the ADS archive and OASIS on-

line repository that might be used to test prototype software being built as part of our research. Specifically, we sought stratigraphy (and grouping if available) and radiocarbon determination as examples of relative and absolute dating evidence, respectively. A secondary aim was to review the potential for reusing digital archaeological data stored within the ADS. It would not have been practical to review the potential for the reuse of all types of archaeological data within the ADS. Instead, the review provides a snapshot of the quality and quantity of data held by the ADS that might be used or reused in chronology construction (Moody et al., 2021).

4.2.1 Data sought

To be potentially useful for our research, the absolute and relative dating evidence needed (as a minimum) to consist of:

1. A table containing mutually consistent pairwise statements of the stratigraphic relationships between contexts (stratigraphic units) as they were observed in the field.

2. A table containing information about samples (or objects/finds) taken from specified contexts on the site, which might be suitable for scientific dating or that have already been dated.

3. A table containing groups and the contexts attributed to them. The relationships between any groups, declared using some form of specified descriptors such as Allen operators (Allen, 1983; May, 2020), was also desirable.

4.2.2 ADS files

To facilitate our exploration of the ADS Archive and OASIS repositories, Tim Evans and Jenny O'Brien of the ADS first provided us with 4,820 digital files in varying

formats deposited in the ADS archive. These files all had metadata containing at least one of the key-phrases (text strings or combination of text strings) in Table 4.1 suggesting that absolute or relative dating evidence was present within them.

Type of data sought	Key-phrase (combination of text strings)
Stratigraphy	*matr* OR *c14* OR *context* OR *phas*
Radiocarbon determinations	*radiocarbon* OR *c-14*
Phasing	*phasing* OR *phase*
Scientific dates	*dating* OR *date*

Table 4.1: A list of the key-phrases used to search the metadata of files deposited in the ADS archives to identify files that might contain the absolute and relative dating evidence needed for our research. Stemmed words were used to search for similar phrases with differing suffixes, such as phas being used to search for phase and phasing.

ADS chose the specific text strings used as those most likely to filter files that contained the following types of data: radiocarbon determinations, scientific dates, stratigraphy and grouping since the combination of radiocarbon determinations and either stratigraphy or grouping are the most common building blocks of a Bayesian chronology construction. As a contingency, in case there was a lack of radiocarbon determinations, we also requested files containing scientific dates.

Methodology

The methodology for searching the data in the archived files provided by ADS consisted of two stages. First, the metadata stored at ADS was queried for specific key-phrases as seen in Table 4.1, then the files identified were manually examined to see if they contained the kind of data we were hoping to find. A metadata search is only suitable if the files have appropriate metadata attached. Although ADS routinely add such metadata, this may not be true with all digital repositories. As such, anyone seeking to replicate our methodology should consider metadata quality

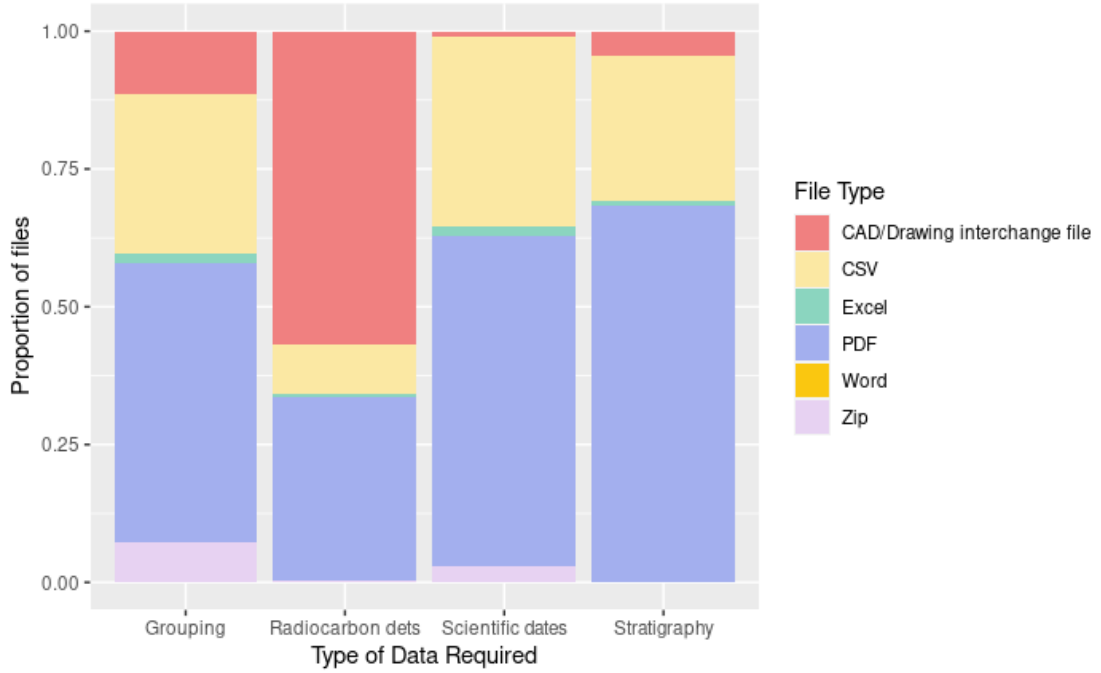


Figure 4.2: A stacked bar chart showing the proportions of file type relating to four categories (as indicated on the horizontal axis) held at ADS. The four categories of files were formed by searching ADS metadata for the key-phrases given in Table 4.1.

if it exists. Furthermore, checking each file by hand, as the author had to do, may not be plausible if there are many files or a strict time limit. If the volume of files is too large, see the methodology in Section 4.2.3.

Replication of results should also be considered. It is the subjective opinion of the person reading the files as to whether the content is reusable for their purpose. Therefore, what constitutes useful data should be clearly defined before examining files. For this research, it was defined that any stratigraphic information that was not given as a stratigraphic diagram or a table of contexts with above/below relationships would be deemed not useful. Even if files contained interpretations clearly derived from stratigraphic information, to work out the stratigraphic relationships correctly

from the interpretation would be extremely difficult and error-prone unless one were very familiar with the excavation. Similarly, files containing groups of contexts without reference to the archaeological contexts they contained were classed as not useful since a grouping alone (without contexts) could not provide suitable case studies for testing the software and models we were seeking to build.

In summary, the files deemed useful contained directly reusable relative and absolute dating evidence. That is not to say that other documents did not contain absolute and relative dating evidence. However, such evidence would have required additional work, such as contacting those who originally produced the data, to be able to utilise it. Thus, any results should only be interpreted as the volume of directly reusable relative and absolute dating evidence.

Results

The results of classifying the files can be seen in Figure 4.3 and Table 4.2. The current author manually classified the files provided to us by ADS in 3 ways: "useful data", "no useful data" and "not accessible". We chose to do this for two reasons. The first was that classifying the files allowed us to keep track of the files that might be useful for case studies. The second reason, and the reason we included the "not accessible" classification, was that we were seeking to quantify the volume of useful files and felt we should be forthright about which files were actually searched for data, in addition to which files contained useful data. Files that were not accessible required additional software to access them. Useful files were defined as those with directly reusable data, though additional work might be required first to convert them into an appropriate format for analysis. More detailed descriptions of useful data are available in Section 4.2.1.

Not accessed

The only files that were not accessible were those containing computer-aided drawings. These were provided as individual files and in zip files. Both the grouping and the radiocarbon determination categories had a large proportion of the files classified as not accessible as a result of the file formats being CAD and interchange files. However, examination of the file names and discussions with ADS suggested that the CAD files were plan and section drawings, which were unsuitable data for our case studies. Furthermore, despite section drawings showing individual sections of stratigraphy, it would not be feasible to obtain all section drawings and form a stratigraphic sequence from them without additional help from those who excavated the site. Therefore, as agreed with the ADS, they were not examined for relative or absolute dating evidence as part of this data review.

Useful data

Of the ADS files manually examined by the current author that we expected to contain scientific dating, grouping and stratigraphy, around half contained directly reusable data. Most files in the radiocarbon determination category that the author examined contained radiocarbon determinations with units. These were often given in the form of PDFs provided by the dating laboratories containing radiocarbon determinations along with other data (such as a laboratory reference number and notes from the laboratory), which is an example of good practice with regard to the data provided. However, since these documents were in PDF format, thus requiring additional work to convert the data into plain text format ready for reuse.

Around a quarter of the data archived and directly reusable were held in plain text files. For several reasons archiving data in plain text format, such as comma-separated (CSV) files, is good practice. The first is the longevity of this format.

Since it is so widely used, it is probable that software will be available to read these files for the foreseeable future. In addition, almost all software packages can read and produce text files, allowing for the interaction between multiple software packages. Finally, plain text files are easy to view and edit, even with moderate computer skills.

No useful data

Just under half of the files examined by the author, such that ADS's initial search on the files' metadata suggested they should have contained stratigraphy, did not. Further, half of those accessed, which we expected to contain grouping of contexts, did not. For these files, it was often the case that the information in the file would require additional interpretation to obtain the raw relative dating information, such as interpreting section drawings to obtain our own stratigraphic sequences. Furthermore, particularly for files where we expected to find stratigraphy, some files were empty plain text files which may have been a result of file corruption. Others were plain text files with table headings, suggesting that stratigraphic data would eventually be added, but this never occurred.

Files with metadata suggesting they contained scientific dating or radiocarbon determinations but which did not were often files that stated that scientific dating would be done in the future or that insufficient samples were found to use scientific dating methods. This was less common with files supposedly containing radiocarbon determinations, only 10% not containing radiocarbon determinations with units. Whereas 44% of files supposedly containing scientific dates did not.

Overall, the files provided by ADS did not contain relevant data for our research. Although some files contained grouping of contexts and stratigraphy, this relative dating evidence was often in the form of interpretive discussion derived from the

raw data obtained during an excavation or post-excavation research. However, the raw data itself for relative dating evidence was not provided, thus limiting reuse. Although sufficient radiocarbon determinations were available, they would not be useful for this research without the corresponding stratigraphy from the site.

	Number of files		
	Useful data	No useful data	Not accessed
Type of data			
Stratigraphy	1797	1472	144
Radiocarbon determinations	120	38	184
Grouping	216	346	129
Scientific dates	193	166	15

Table 4.2: Results of categorising files from ADS. Files classified as not accessed need propriety software to open, and useful data are all those that we found to be directly reusable for our research purposes.

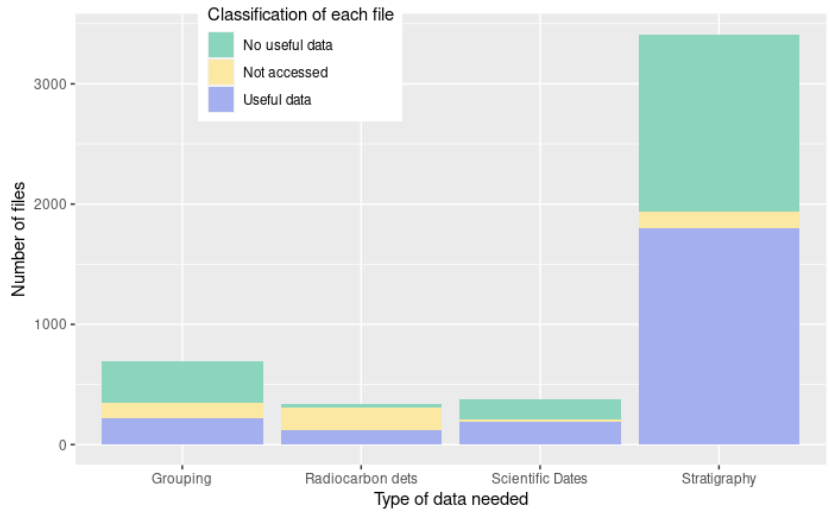


Figure 4.3: A stacked bar chart showing the number of files provided by ADS that have either directly reusable (useful) data, no directly reusable data or were not accessed. The data type that we expected to see in the files based on the metadata search carried out by ADS is seen on the horizontal axis.

4.2.3 OASIS reports

Despite a systematic search of the metadata by ADS, the files they provided from the ADS repository did not contain sufficient absolute and relative dating evidence for use in case studies. Discussions with ADS and Keith May suggested this might have been because those archiving the excavation records deemed sufficient information present in the text of the archaeological reports, and thus separately archiving digital data was not necessary. As such, we conducted further searches for data on the unpublished archaeological excavation reports held within OASIS.

Unpublished excavations reports, commonly referred to as grey literature, are common in archaeology due to the large number of excavations carried out. Much of the archaeology carried out in the UK is by private contractors as a result of government policy introduced in recent decades which dictates archaeological investigation should be carried out before developmental building work occurs (see Evans, 2015). Excavation reports may not contain data and results worth publishing but should, under this policy, be archived nonetheless. A report by Jones et al. (2003) stated that the reuse of such reports was hampered since users did not know where to find them. The purpose of OASIS is to provide a single online resource that enables access to the large volume of unpublished archaeological reports produced as a result of contract and research archaeology. The reports are made publicly available via the local Historic Environment Records for the relevant area or for download at `https://archaeologydataservice.ac.uk/archsearch/browser.jsf`.

The entire collection of unpublished (grey) literature stored in OASIS as of November 2018 was provided by ADS. As with the archived documents provided by ADS, the aim was to source absolute and relative dating evidence that might be suitable for the research in this thesis. However, due to the large volume of data (37,320 documents

as of November 2018) we would require a different methodology since it would not be feasible to check such a large volume of documents manually.

Methodology

Instead of searching the OASIS metadata for key-phrases as we did in the previous section, we searched the documents' text. We used a technique known as shell scripting within the `unix` computer operating system. Such scripting involves writing computer commands (in our case using the `BASH` shell (Free Software Foundation, Inc, 2019)) in text format that, when executed, can search plain text files for key-phrases (text strings or combination of strings) defined by the user and return a list of documents that contain them.

Careful consideration was needed when choosing the key-phrases to prevent spurious results. While searching the documents computationally for specific text strings removes human error, it disregards the setting where the words were used within the document. Consequently, documents may contain words used in a natural language setting with specific meanings in archaeology. For example, the text string *phase* is often used to describe a period in the timeline of excavation as well as a group of contexts. For our research, we were only interested in documents in which the term *phase* was used to refer to the grouping of archaeological contexts. Similarly, searching for *carbon* did not guarantee that the documents highlighted would contain radiocarbon determinations themselves as opposed to, for example, discussions of the interpretation of radiocarbon determinations or of carbon-rich deposits in the stratigraphy.

As with the key-phrases used to search the ADS metadata matches in Section 4.2.2, using combinations of text strings proved useful in preventing spurious matches,

especially for strings with multiple meanings. We first chose key-phrases closely matching those used to search the ADS metadata. These were *strat*, *phase*, *date* and *carbon*. However, since these text strings can have multiple meanings, we adopted an iterative approach to choosing our key-phrases. At each stage key-phrases were chosen, and a computational search of the documents was carried out. Then, samples of the documents filtered by the search (around 20 per key-phrase) were manually examined, and any false positives were identified, allowing the key-phrases to be further refined. For example, even if one searches for a combination of either *radiocarbon*, *carbon date* or *carbon dating*, one obtains 3300 documents. However, when a sample of the documents was examined, it was found that some discussed samples that were unsuitable for dating or the fact that dating would be done in the future, much as we had found when searching the ADS metadata. To reduce the occurrence of these false positives, we also searched documents for the character ±, since radiocarbon determinations are reported as an estimate plus/minus a laboratory standard error.

The full `Bash` scripts used to complete the computational search can be found at Moody (2019), and the entire process of choosing the key-phrases are outlined in Figure 4.4. The yellow boxes in Figure 4.4 and Table 4.3, alongside their key-phrase labels, are given the final key-phrases. We have allocated labels (presented in bold) to key-phrases to allow for their concise reference within this chapter.

Results

The results of the key-phrase search on the content of the OASIS files can be seen in the Venn diagram in Figure 4.5. Of the 37,320 OASIS documents, 102 contained all of the key-phrases given in Table 4.3, indicating they contained evidence of stratigraphy, a stratigraphic diagram, radiocarbon determinations and grouping of

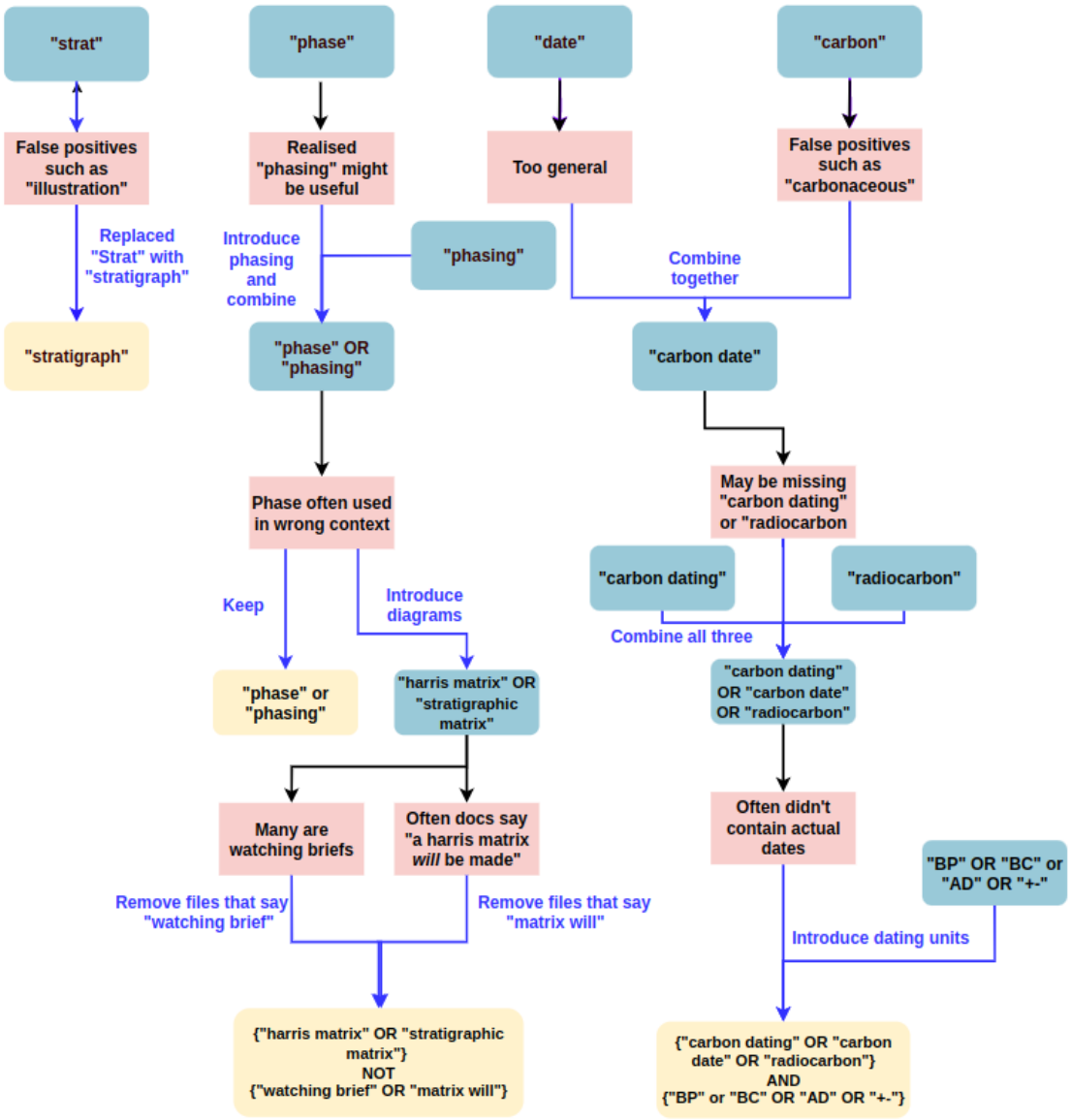


Figure 4.4: Diagram showing the process of choosing key-phrases for use in an automated key-phrase search on documents downloaded from the online repository OASIS. Teal boxes contain key-phrases used early in the process. Red boxes describe issues encountered, blue edges and their corresponding labels represent actions to resolve those issues, and yellow boxes contain the final key-phrases.

Key-phrase label	Text string/combination of text strings
Stratigraphy	*stratigraph*
Grouping	*phase OR phasing*
Stratigraphic diagrams	*harris matrix* OR *stratigraphic matrix* EXCLUDING *watching brief* OR *matrix will*
Radiocarbon determinations	*carbon date* OR *carbon dating* OR *radiocarbon* AND *Cal BP* OR *AD* OR ± OR *BC*

Table 4.3: Table defining key-phrases and the text string/combination of text strings they consist of. These key phrases were used to search the OASIS files as provided by ADS in November 2018. Each key-phrase has a label that is used to reference it in the main body of text.

contexts. These were examined, and most were indeed found to contain the kinds of data that could potentially be useful for early-stage testing of our prototype software but not complex enough for case studies. However, some were still false positives. Most of the false positives were due to the text string *phase* being used to denote the phase of an excavation rather than grouping contexts into archaeological phases as part of the post-excavation process.

Initially filtering by **Stratigraphy**, as defined in Table 4.3, gives a total of 13,114 documents, of which, 9,260 contained only **Stratigraphy**, but not **Stratigraphic Diagrams**, **Radiocarbon determinations** or **Grouping**. These results suggest that although many documents contained discussion of stratigraphy, most of these did not contain any formal diagrams to represent the stratigraphy. Nor do they contain formal chronology construction of the type we wished to replicate since this would usually require at least some radiocarbon determinations.

Most of the documents did, however, contain the data that the key-phrase search suggested they would. The documents that did not contain the data we required,

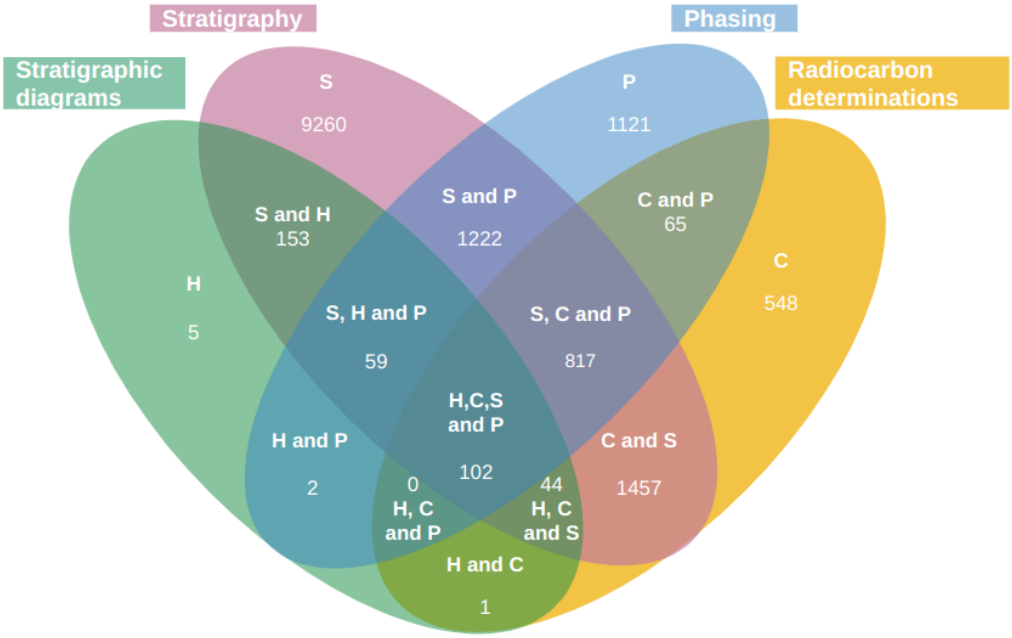


Figure 4.5: Venn diagram showing the total number of documents containing each combination of key-phrases, with each segment label corresponding to key-phrase labels as defined in table 4.3. Note that each total represents the number of documents containing that single key-phrase or combination of key-phrases only and none of the others.

despite containing all the key-phrases, often stated that the stratigraphy and information about the grouping of contexts were contained in the "stratigraphic archive" with no suggestion as to where we might find this archive.

4.3 Summary of data review

A detailed discussion of the data review results and specific recommendations to improve the reusability of digital chronological data are provided in Moody et al. (2021), Section 5. As such, we only include the key points in this section.

The data review highlighted that insufficient chronological data are archived from excavation projects to allow for reuse in further analysis. The incomplete or near-empty files discovered when searching the ADS files were surprising, given the cost of deposition. Further, we found a real lack of reusable stratigraphic evidence within the ADS archives, and the data we found in the OASIS documents did not contain stratigraphy complex enough for our requirements.

In addition, we had time generously given to us by the staff at the ADS to obtain data for us, which is currently not feasible for the broader research community. Even with the help of ADS, who could easily query the metadata of deposited files, these documents often did not contain directly reusable data. Huggett (2018) raised concerns that the FAIR principles fall short of providing plausible solutions for increasing the reuse of archaeological data, highlighting that making data findable, accessible and interoperable does not ensure the data are reusable. We found that this was indeed the case.

This PhD research aims to provide software that automates at least parts of the chronology construction process. The data review summarised in this chapter highlighted that we needed to ensure our software stored all the necessary data to reproduce analyses once such data were archived. Whilst automating the process of chronology construction was important, if a record of this process and all the data used and produced cannot be easily and quickly archived, we would only exacerbate

the problems highlighted in this chapter and Moody et al. (2021). Furthermore, it indicated early on in our research that alternative sources of complex stratigraphic data and associated grouping and radiocarbon determination for case studies were required.

4.4 Alternative sources of case study data

To obtain the data we required, we turned to publications of well-known sites that we knew had large stratigraphic sequences and for which extensive radiocarbon dating had been carried out. The first was the site of the Danebury Iron Age hillfort within the UK. Extensive excavations were carried out between 1969 and 1988 (Cunliffe and Poole, 1991). Numerous monographs were published presenting the results of the excavations. The Harris matrices for the stages of excavation were provided in these monographs, albeit they were printed on microfiche. However, the lack of other data sources meant the time required to convert the microfiche matrices to PDF and then convert them to mathematical graphs was deemed necessary. Furthermore, the volume of relative and absolute dating evidence was substantial, and all data required for chronological modelling were published, thus allowing us to test prototype software.

Further, we used data from excavations at Çatalhöyük (Marciniak et al., 2015) as case studies for very much the same reasons, i.e. a large volume of data was published with all the components required for archaeological dating. For smaller case studies, we used a stratigraphic sequence from St Veit-Kliglberg as given in Buck et al. (1996) Chapter 9. While having a more comprehensive range of datasets for case studies would have been useful, the aforementioned data sources, alongside hypothetical stratigraphic sequences, proved sufficient for testing that prototype software

for various stratigraphic sequences.

Up to this point, we have introduced the concept of chronology construction, the data used to carry it out, and the current issues with archiving and reusing such digital data. We will now detail the statistical theory behind chronology construction, demonstrating how the relative and absolute dating evidence discussed so far is utilised. Following this, we introduce the prototype software written for this research and how it seeks to address some of the issues motivated in this chapter, alongside the other purposes of the software, which are motivated in Chapter 6 and 7.

Chapter 5

Bayesian chronological modelling

Bayliss (2009) describes three radiocarbon revolutions. So far, we have introduced two of these: radiocarbon dating itself; and the introduction of calibration curves, which allow us to calibrate radiocarbon determinations to the calibrated scale. The third revolution she describes was the introduction of Bayesian inference in chronological modelling. In this section, we motivate why Bayesian inference is the obvious choice for constructing archaeological chronologies. Then, we provide a summary of the theory of Bayesian inference required before describing the history of Bayesian chronology construction. Finally, we showcase the use of Bayesian inference for two illustrative examples of chronological models.

In archaeology, we often obtain expert knowledge during an excavation from the relative dating evidence, which takes the form of the stratigraphic sequence and grouping of the contexts within the sequence. Such relative dating evidence is inherently subjective. The interpretation of a site and the resulting stratigraphic sequence and groupings may change depending on the expert, or indeed, an individual

expert's opinion may change over time following further excavation, analysis or scientific dating. Alternatively, an expert may consider multiple chronologies to be plausible but may be unable to pick a specific chronology to be the most likely. Nonetheless, this subjective knowledge, though dependent upon the expert, provides valuable information that informs our prior beliefs regarding the true chronology of a site.

As discussed in Chapter 2, Harris (1975) argues that the finds and artefacts that lead to absolute dates are only part of the information obtained when excavating. Thus, ignoring stratigraphic and grouping information would exclude valuable data from our inference. Since a stratigraphic sequence and grouping of contexts within that sequence provide a prior ordering of the calibrated ages in our chronology that we wish to estimate. Further, recall from Chapter 2 that the process of gathering archaeological data is non-repeatable, we may carry out further excavations at a site, but we can not re-excavate the same part of a site twice. It is essential that an appropriate statistical methodology is chosen to allow for the use of subjective and non-repeatable data in chronological models.

There are two main schools of thought in statistics: Bayesian and frequentist. Despite depending on much of the same underlying probability theory, the two methods of inference are opposing ways of thinking (Vallverdu, 2008). For example, suppose we have a model representing a population we wish to learn about; this model will have parameters. Frequentist statistics focuses on the probability of observing specific values of repeatable data sampled from the population, given fixed parameter(s) for the model. On the other hand, Bayesian inference treats parameters as random variables and, as such, allows us to make probabilistic statements about them (O'Hagan, 2008). Thus, in Bayesian statistics, we instead focus on the

probability of observing parameter values given fixed observed data and describe our (often subjective) prior beliefs about the true values of such parameters using probabilistic statements.

In chronology construction, our parameters of interest are the true calibrated dates of specific archaeological events, for which we often have a wealth of prior knowledge. As seen in Chapter 3, the likelihood $L(\theta; x)$ expresses (once normalised) the probability of any given calibrated age being the true calibrated age of a sample given our fixed data (a radiocarbon determination for the sample). Utilising Bayesian inference, we are able to formalise relative dating evidence as probabilistic statements, known as prior knowledge, about our parameters of interest (details of how to do so are provided in Section 5.3.2). This allows us to improve our estimate of the true calibrated age of a sample and use all available data obtained during an excavation.

Bayesian inference has, in fact, been adopted in many areas of archaeology, as discussed by Otárola-Castillo et al. (2023). Further, it has been widely accepted as the methodology of choice for archaeological chronology construction, Buck et al. (1996); Bronk Ramsey (2009); Bayliss and Bronk Ramsey (2004) motivates in detail the appropriateness and advantages of utilising Bayesian inference in chronology construction. Thus, further justification of our use of Bayesian chronology construction is omitted here and we refer the reader to the aforementioned publications for further detail. We now introduce the theoretical details of Bayesian inference for readers who are not familiar before discussing the development of Bayesian chronology construction.

5.1 Theory behind Bayesian inference

The simplest Bayesian chronological model would only consider the stratigraphic sequence as our prior knowledge and omit grouping. For such a model, our parameters are the set of N true calibrated ages we wish to estimate, which we denote as $\boldsymbol{\theta} = \{\theta_1, ..., \theta_i, ..., \theta_N\}$. Assuming we have radiocarbon determinations for every calibrated age we wish to estimate, these parameters will have associated data in the form of N radiocarbon determinations $\boldsymbol{x} = \{x_1, ..., x_1, ..., x_N\}$ with errors $\boldsymbol{\sigma} = \{\sigma_1, ..., \sigma_1, ..., \sigma_N\}$. Note that, just as all use of x in all previous notation implies conditioning on error σ, all use of $\boldsymbol{x}$ implies conditioning on $\boldsymbol{\sigma}$.

As seen in Chapter 3 Section 3.2.3, for a single radiocarbon determination, we have the likelihood function $L(\boldsymbol{\theta}; \boldsymbol{x})$ as a statistical function which defines the relationships between observed radiocarbon determinations $\boldsymbol{x}$ and unknown calibrated ages $\boldsymbol{\theta}$, as provided in Buck et al. (1992). Our prior beliefs about $\boldsymbol{\theta}$, derived from the stratigraphy and grouping, are expressed as the prior density function $p(\boldsymbol{\theta})$. By taking the product of the prior and the likelihood, we obtain up to proportionality the posterior as outlined in Figure 5.1. The posterior $p(\boldsymbol{\theta}|\boldsymbol{x})$ is a probability distribution that gives us our updated beliefs about the parameters we wish to estimate, having taken into account prior knowledge and data.

The exact relationship between the prior, likelihood and posterior derives from Bayes theorem (see e.g. Gelman et al., 2004, Chapter 1)

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)},$$

and can be expressed as

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{\theta})L(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x})}. \tag{5.1}$$

.

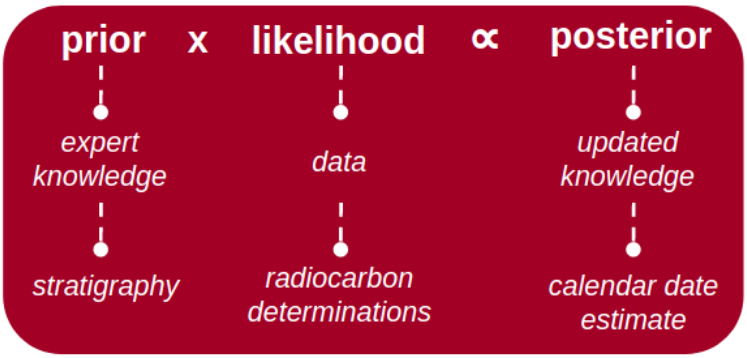


Figure 5.1: Illustration showing the relationship between prior, likelihood and posterior in Bayesian chronology construction

The denominator in Equation 5.1 is known as the normalising constant and is equal to:

$$p(\boldsymbol{x}) = \int p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{5.2}$$

However, given that it is a constant and the data are fixed, it is often omitted for ease of calculation. Therefore, we often consider our posterior up to proportionality as

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{x}), \tag{5.3}$$

and renormalise later if needed. For a more in-depth introduction to probability theory and Bayesian inference for chronology construction, we refer the reader to Buck et al. (1996). However, since the theory introduced in this section is well-established and routinely utilised, we omit further theoretical detail here.

5.2 History of Bayesian chronology construction

Naylor and Smith (1988) first proposed using Bayesian inference to estimate the calibrated ages for samples of interest using radiocarbon determinations. They sought to compute the posterior of their parameters of interest exactly, requiring calculation of the integral defined in Equation 5.2, as well as any required summary descriptions. They did so using numerical methods (see Naylor and Smith, 1988, Section 3). However, as the number of parameters in the model increases, so does the complexity of the integral calculation and the numerical evaluation quickly becomes intractable.

Litton and Leese (1991) outlined the process of calibrating sequences of radiocarbon determinations when using Bayesian inference to incorporate prior information and sought to address some concerns the community had regarding the initial work by Naylor and Smith (1988). Further, they highlighted that the computational limitations that Naylor and Smith (1988) had encountered might be addressed by utilising advancements (at the time) in sampling methods used to approximate probability distributions such as Gibbs sampling (Geman and Geman, 1984), which had been shown to aid in the analysis of high-dimensional Bayesian models (see Gelfand and Smith, 1991). Such sampling methods only require us to sample from densities proportional to the posteriors distributions of interest and are part of a class of sampling methods called Markov Chain Monte Carlo (MCMC) methods (see Appendix A Section A.1 if unfamiliar with MCMC methods). These samples can then be used to obtain summary statistics, omitting the need to evaluate the high-dimensional integrals.

Buck et al. (1991) provided further illustrative examples of utilising Bayesian inference

in chronology construction before developing upon that work in Buck et al. (1992), which introduced for the first time the use of Gibbs sampling methods for approximating the posterior distributions of chronological models. In the subsequent years, further illustrative examples of sampling-based Bayesian inference for chronology construction were produced (see Buck et al., 1994; Christen, 1994; Zeidler et al., 1998, for some examples).

During this time, Buck et al. (1996) was published, which provided an in-depth introduction to Bayesian inference in archaeology, written for an archaeological audience. Buck et al. (1996) Chapter 9 focuses on using Bayesian inference in chronology construction with various case studies, such as the excavation at St. Veit-Klinglberg in Austria. Buck et al. (1996) demonstrated the processes of including the stratigraphic relationships observed on-site as prior knowledge when estimating the true calibrated ages of ten samples. They showed that there was a 35% reduction on average in the number of plausible calibrated ages included in the 95% HPD interval for the true calibrated age of ten samples when compared to simply calibrating the radiocarbon determinations for the samples alone. Software for Bayesian chronology construction followed these initial publications, for example OxCal (Bronk Ramsey, 1995) and BCal (Buck et al., 1999). We extensively review the available software in Chapter 7 but omit further discussion here.

Decades on, the chronological modelling community has widely adopted Bayesian inference. This is particularly true in the UK where as shown in Bayliss (2009), Figure 11, 73% of radiocarbon determinations obtained by Historic England between 1993 and 2009 were analysed within a Bayesian framework. This uptake of Bayesian chronological modelling has also been seen in the research community as a whole, as seen in Bayliss (2015), Figure 2, which demonstrates an upward trend of published

papers utilising Bayesian chronology construction between 1991 and 2014.

So far, we have provided a general introduction to Bayesian inference and outlined the history of Bayesian chronological modelling. The remainder of this chapter will outline precisely the form of the prior, likelihood and posterior for a model that only utilises stratigraphy as prior knowledge (without any grouping). We then demonstrate the value of Bayesian inference for a slightly more extensive case study. More complex chronological models incorporating grouping are discussed in Chapter 6.

5.3 A Bayesian model for chronology construction

To carry out Bayesian inference requires two components, the prior and the likelihood, to obtain a posterior up to proportionality and thus derive posterior estimates for our parameters of interest. Within this section we define the well-established form of these components as motivated in detail in Buck et al. (1996) Chapter 9.

5.3.1 Likelihood

The joint likelihood function $L(\boldsymbol{\theta}; \boldsymbol{x})$ is simply the product of the individual likelihood functions for each radiocarbon determination, as defined in Chapter 3 Section 3.2. Thus, in a stratigraphic sequence with N archaeological contexts, we have:

$$L(\boldsymbol{\theta}; \boldsymbol{x}) \propto \prod_{i=1}^{N} \exp\left(-\frac{(x_i - \mu(\theta_i))^2}{2(\sigma_i^2 + \delta(\theta_i)^2)}\right). \tag{5.4}$$

Note that this differs slightly from the likelihood provided in Buck et al. (1996) Chapter 9, with the inclusion of $\delta(\theta_i)$. This is because, at the time Buck et al. (1996) were writing, the errors on the calibration curve estimates were comparatively small

relative to the standard errors on the radiocarbon determinations for archaeological samples. However, this is no longer the case due to advancements such as AMS and also to improved estimates of the uncertainties on the calibration curve estimates, so both σ_i and $\delta(\theta_i)$ are now conventionally modelled.

5.3.2 Prior knowledge about true calibrated ages

The next component we require is the prior $p(\boldsymbol{\theta})$, which is the function that represents our prior knowledge about the parameters in $\boldsymbol{\theta}$, which for the type of models considered in this chapter, derives from the stratigraphic relationships. The stratigraphic relationships provide an a priori chronological ordering for contexts. As such, assuming a radiocarbon sample dates the context in which it was found, the stratigraphic sequence allows us to impose an ordering on the true calibrated ages of the samples. For example, suppose we have two contexts for which we have radiocarbon-dated samples and that the samples from context 1 and context 2 have true calibrated ages θ_1 and θ_2 cal BP, respectively. If the stratigraphic sequence tells us that context 1 is below context 2 and so is older, we can infer that $\theta_1 > \theta_2$ since older ages are larger on the cal BP scale (see Section A.2 for an implemented example of such a model).

To express the stratigraphic relationships as probabilistic statements, let C be the set of all possible values for $\boldsymbol{\theta}$ such that the ordering of the true calibrated ages satisfies the stratigraphic constraints recorded on site, then we define the indicator function

$$\mathbb{I}_C(\boldsymbol{\theta}) = \begin{cases} 1 & \boldsymbol{\theta} \in C \\ 0 & \text{otherwise,} \end{cases} \tag{5.5}$$

which sets to zero the probability of $\boldsymbol{\theta}$ equal to any set of calibrated ages that do

not satisfy the stratigraphic constraints. Therefore, we simply set our prior as this indicator function, giving $p(\boldsymbol{\theta}) = \mathbb{I}_C(\boldsymbol{\theta})$.

5.3.3 Posterior

To obtain our posterior, up to proportionality, we take the following product,

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto L(\boldsymbol{\theta}; \boldsymbol{x})p(\boldsymbol{\theta}) = \prod_{i=1}^{N} \exp\left(-\frac{(x_i - \mu(\theta_i))^2}{2(\sigma_i^2 + \delta(\theta_i)^2)}\right) \mathbb{I}_C(\boldsymbol{\theta}) \tag{5.6}$$

using the likelihood defined in Equation 5.4 and an indicator function as our prior. To calculate the joint posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ exactly would require us to calculate the normalising constant

$$p(\boldsymbol{x}) = \int \mathbb{I}_C(\boldsymbol{\theta}) \prod_{i=1}^{N} \exp\left(-\frac{(x_i - \mu(\theta_i))^2}{2(\sigma_i^2 + \delta(\theta_i)^2)}\right) \mathrm{d}\boldsymbol{\theta}. \tag{5.7}$$

To obtain a posterior estimate for a single calibrated age estimate is simple. Since it only requires integration in one dimension. However, as we begin to consider more parameters and, as such, begin to include prior knowledge, we must evaluate multidimensional integrals. As discussed, Naylor and Smith (1988) used numerical integration to accomplish this, but as Ghosh et al. (2006), Chapter 7 discusses, numerical integration becomes unstable as the dimensions of the integral get larger. As a result, statisticians now routinely use sampling methods to approximate probability distributions of interest.

We discuss the sampling methods used in this research, following the introduction of more complex chronological models in Chapter 6. However, for the illustrative example that follows, we utilised Gibbs sampling to obtain posterior results. Buck

et al. (1996) Chapter 9 describes the use of Gibbs sampling (see Geman and Geman, 1984) to approximate the posteriors of parameters of interest, providing a detailed algorithm for implementation. Since Gibbs sampling for chronological models is discussed in detail in the aforementioned research and is only used for the following example, we omit methodological details here. However, for readers unfamiliar with Gibbs sampling, pseudo-code and an example of the use of Gibbs sampling for a chronological model with two parameters are provided in Appendix A Section A.2.

5.4 Effect of plateaus in the calibration curve on posterior estimates

We now examine the results of estimating calibrated age estimates for seven hypothetical samples, which, without prior knowledge, have wide HPD intervals following calibration. Such results are obtained using Gibbs sampling, as discussed in the previous section.

5.4.1 Hallstatt Plateau

The width of the HPD intervals of calibrated radiocarbon determinations is influenced by the part of the calibration curve on which they lie, as well as the laboratory error on the radiocarbon date. In addition to the calibration curve being non-monotonic, calibration is also further complicated by the general trend of the curve in specific places on the calibrated scale. For example, in places on the curve where the general trend is very flat, radiocarbon determinations will often have very wide calibrated age ranges, even if they have a small laboratory error. One such example is the so-called Hallstatt plateau which is an area on the calibration curve between 2750

cal BP and 2350 cal BP, as seen in Figure 5.2

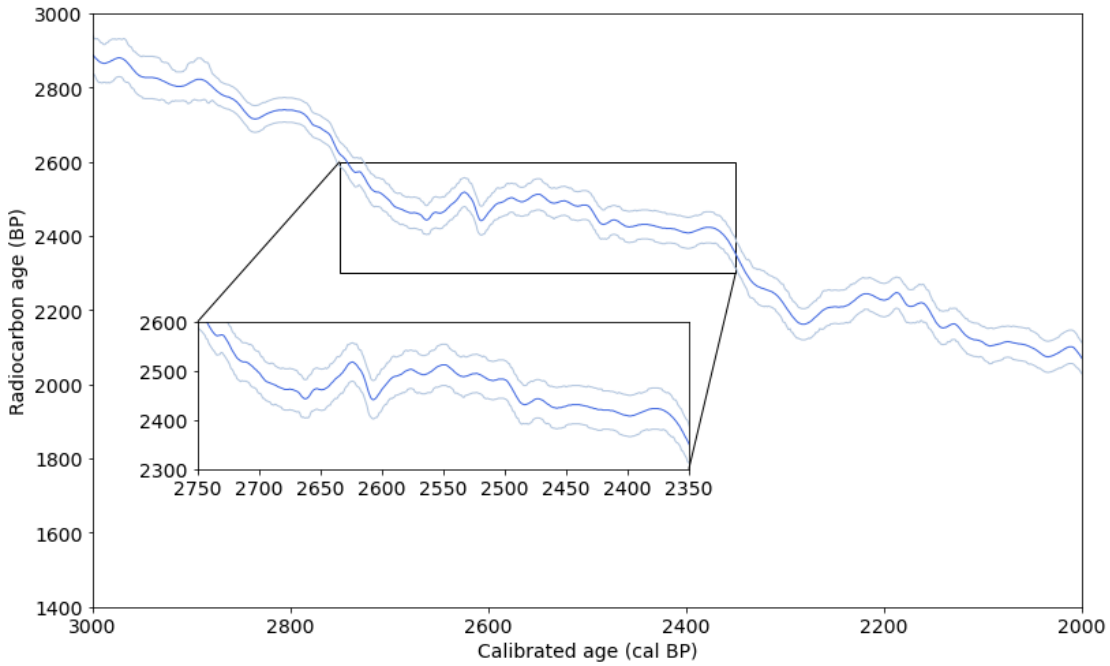


Figure 5.2: Section of the IntCal20 calibration curve (Reimer et al., 2020) from 3000 to 2000 cal BP. Inset highlights the Hallstatt plateau, which is a flat part of the calibration curve between 2750 and 2350 cal BP.

By including a priori relative dating evidence when calibrating determinations whose likelihoods lie on flatter parts of the calibration curve, we can improve the precision of the posterior calendar age estimates. For example, suppose we have seven contexts $h1$ to $h7$ and radiocarbon determinations for samples found in these contexts, as outlined in Table 5.1. Further, we have a stratigraphic sequence that provides us with the relationships for contexts $h1$ to $h7$ shown in the mathematical graph in Figure 5.3.

When calibrating the radiocarbon determinations for samples from contexts $h1$ to $h7$ without including prior knowledge, as seen in Table 5.1 column 3, the 95% HPD intervals for the posterior calibrated age estimate of each determination are identical,

Figure 5.3: Mathematical graph showing the stratigraphic sequence of a hypothetical site.

despite the radiocarbon estimates for the sample spanning 35 years. Furthermore, the HPD intervals span 350 years (cal BP) for each calibrated radiocarbon determination. The inclusion of prior knowledge in the form,

$$\theta_1 > \theta_2 > \theta_3 > \theta_4 > \theta_5 > \theta_6 > \theta_7,$$

as provided by the stratigraphic sequence seen in Figure 5.3, provides the HPD intervals shown in Table 5.1 column 4. As we can see, such prior knowledge reduces the span of all the HPD intervals of the posterior calibrated age estimates. In particular, the HPD intervals for the true calibrated age of contexts $h1$ and $h2$ are considerably shorter once prior knowledge is included. Presented in Figure 5.3 is a comparison between the marginal posterior distributions for true calibrated ages of contexts $h1$ to $h7$ when prior knowledge is and is not included, where a notable change to the marginal posterior is observed for all parameters when prior knowledge is included.

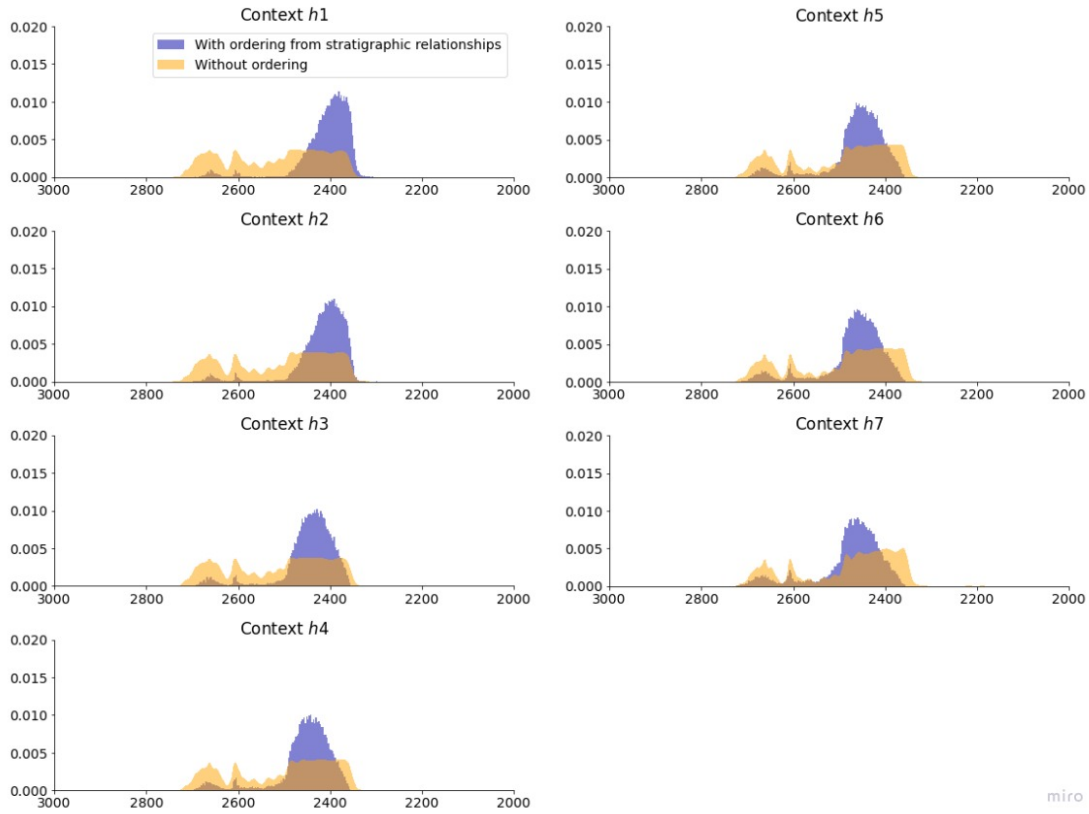


Figure 5.4: Comparison of the estimated marginal posterior densities for the calibrated ages (cal BP) of 7 samples when ordering is included as prior knowledge (blue histograms) and when it is not (yellow histograms).

Thus far, all the Bayesian chronological modelling theory and examples we have introduced have only considered stratigraphy as prior knowledge. However, grouping also provides informative prior knowledge. As such, the following chapter expands upon the theory introduced in this chapter and introduces the established theory used routinely in the chronological modelling community to include grouping in Bayesian chronological models.

Context	Radiocarbon determination	Calibrated age 95% HPD interval without ordering	Calibrated age 95% HPD interval with ordering
$h1$	2400 ± 50	2340-2540, 2561-2700 cal BP	2342-2485 cal BP
$h2$	2410 ± 50	2340-2540, 2561-2700 cal BP	2347-2487 cal BP
$h3$	2413 ± 50	2340-2540, 2561-2700 cal BP	2360-2511, 2647-2675 cal BP
$h4$	2420 ± 50	2340-2540, 2561-2700 cal BP	2361-2519, 2646-2682 cal BP
$h5$	2430 ± 50	2340-2540, 2561-2700 cal BP	2364-2540, 2600-613, 2645-2682 cal BP
$h6$	2425 ± 50	2340-2540, 2561-2700 cal BP	2365-2542, 2598-2614, 2645-2688 cal BP
$h7$	2435 ± 50	2340-2540, 2561-2700 cal BP	2364-2572, 2595-2617, 2644-2690 cal BP

Table 5.1: Table of 7 hypothetical radiocarbon determinations that are calibrated by Hallstatt plateau region of the IntCal20 calibration curve. HPD intervals are provided for calibrated ages after calibration, with the third column showing HPD intervals with no ordering on the calibrated ages and the fourth column showing posterior calibrated age estimates when the ordering and grouping seen in Figure 5.3 are included in the analysis.

Chapter 6

Hierarchical Bayesian chronological models

Thus far, we have focused on estimating the true calibrated age of archaeological organic samples obtained during excavation using Bayesian chronological models. Each archaeological sample is attributed to a context, as discussed in Chapter 2, Section 2.2. These contexts are often grouped, representing distinct periods of time in the chronology. In addition to estimating the age of specific samples, archaeologists are also interested in the true calibrated age of the archaeological events representing the start and end of these periods of time, which we refer to as group boundaries. This is despite not having absolute dating evidence for such boundaries.

This chapter introduces the structure of the Bayesian chronological model that is commonly used when grouping is considered and demonstrates how such models are implemented. Following this, we examine a variety of case studies showcasing a selection of chronological models that one might encounter when carrying out chronological modelling.

The very early work on Bayesian chronological modelling by Naylor and Smith (1988)

included grouping. They assumed that the true calibrated ages of unordered archaeological samples within a group were uniformly distributed, thus expressing the prior knowledge that the true calibrated ages are contained within a specific group, but we have no knowledge of where in that group they lie. This creates a hierarchical structure between the true calibrated ages of samples and their group boundaries, formally defined in Section 6.1. Such hierarchical models can also be used for groups with ordered calibrated ages of samples (such that the ordering derives from the stratigraphic sequence).

The following section formally defines the structure of the hierarchical Bayesian model used for chronology construction in this thesis which predominantly follows the model introduced in Nicholls and Jones (2001). We discuss how the model has changed from early definitions, such as Naylor and Smith (1988), and the sampling methods now routinely used for implementation.

6.1 Bayesian model for chronology construction

Consider an archaeological site such that the contexts (and, therefore, the true calibrated ages of the samples from those contexts) have been grouped into J groups; we define the group boundaries α_j and β_j for $j = 1, .., J$ as the calibrated age of the start and end of group j, respectively. For such a chronology, we are interested in estimating the sets of parameters $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_J\}$ and $\boldsymbol{\beta} = \{\beta_1, ..., \beta_J\}$, in addition to $\boldsymbol{\theta}$.

The introduction of grouping requires a slight change in the notation for the true calibrated ages for a set of samples $\boldsymbol{\theta}$. Suppose each group j contains N_j contexts, then we define $\theta_{i,j}$ to be the true calibrated age of the sample found in the ith context of the jth group, which has a radiocarbon determination of $x_{i,j} \pm \sigma_{i,j}$ for $i = 1, .., N_j$

and $j = 1, ..., J$.

6.1.1 Likelihood for hierarchical model

Following the approach taken in the previous chapter with redefined notation, given J groups with the jth group containing N_j archaeological contexts, the likelihood is given by:

$$L(\boldsymbol{\theta}; \boldsymbol{x}) \propto \prod_{j=1}^{J} \prod_{i=1}^{N_j} \exp\left(-\frac{(x_{i,j} - \mu(\theta_{i,j}))^2}{2(\sigma_{i,j}^2 + \delta(\theta_{i,j})^2)}\right). \tag{6.1}$$

Since we have no direct scientific dating evidence for group boundaries, all stratigraphic information regarding their relative or absolute dates is considered prior knowledge.

6.1.2 Prior for the hierarchical model

The prior $p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ can be broken down into its hierarchical components. Our prior knowledge of the group boundaries is independent, a priori, of the calibrated age of the samples within the model. However, $\boldsymbol{\theta}$ is conditional on the group boundaries $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. As such, we use conditional probability to express the prior as

$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\alpha}, \boldsymbol{\beta}), \tag{6.2}$$

where $p(\boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ is our prior knowledge about $\boldsymbol{\theta}$, given the groupings and $p(\boldsymbol{\alpha}, \boldsymbol{\beta})$ represents our prior knowledge of the group boundary ages only.

Prior knowledge about true calibrated ages of our samples

First, we define our prior on $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta})$. Our prior knowledge for $\boldsymbol{\theta}$ now derives from the stratigraphic relationships and the grouping of the contexts. In addition to

the prior knowledge on $\boldsymbol{\theta}$ based on the stratigraphic sequence, we now have additional prior knowledge about $\boldsymbol{\theta}$ conditional on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ since we know that any date $\theta_{i,j}$ is constrained by the ages of the boundaries of the group that it is in. Therefore, as first suggested by Naylor and Smith (1988) and defined in Buck et al. (1996) Chapter 9, we assume $\theta_{i,j}$ is uniformly distributed with respect to α_j and β_j since each true calibrated age within the group boundaries is equally likely to be the true calibrated age for all samples. Thus, we have

$$\theta_{i,j}|\alpha_j, \beta_j \sim U(\beta_j, \alpha_j) \text{ for } j \text{ in } 1, ..., N_j$$

As with the previous chapter, we use the indicator function $\mathbb{I}_C(\boldsymbol{\theta})$ to express stratigraphic relationships as probabilistic statements. In addition, another indicator function is required to ensure each $\theta_{i,j}$ is contained by its group boundaries within the model. Following notation in Buck et al. (1996), we thus define

$$\mathbb{I}_{(\beta_j,\alpha_j)}(\theta_{i,j}) = \begin{cases} 1 & \beta_j \leq \theta_{i,j} \leq \alpha_j \\ 0 & \text{otherwise,} \end{cases} \tag{6.3}$$

which sets to zero the probability of $\theta_{i,j}$ taking a value outside the calibrated ages of the group boundaries for group j. As such, our prior knowledge of $\boldsymbol{\theta}$ can be expressed as

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbb{I}_C(\boldsymbol{\theta}) \prod_{j=1}^{J} (\alpha_j - \beta_j)^{-N_j} \prod_{i=1}^{N_j} \mathbb{I}_{(\beta_j,\alpha_j)}(\theta_{i,j}). \tag{6.4}$$

Prior knowledge of the group boundaries for hierarchical model

As seen in Chapter 2, Section 2.2, archaeologists not only group contexts, they also provide ordering information for these groups using the between-group Allen

algebras. Much like the stratigraphic relationships for contexts, these between-group relationships impose ordering on the group boundaries, as illustrated in Table 6.1.

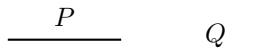
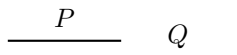
Allen algebra	**Group relationship**	**Diagram**	**Logical statement**
$P > Q$ $Q < P$	P before Q Q after P		$\alpha_P > \beta_P > \alpha_Q > \beta_Q$
$P\ m\ Q$ $Q\ mi\ P$	P meets Q Q is met by P		$\alpha_P > \beta_P = \alpha_Q > \beta_Q$
$P\ o\ Q$ $Q\ oi\ Q$	P overlaps Q Q overlapped by P		$\alpha_P > \alpha_Q > \beta_P > \beta_Q$

Table 6.1: Table shows 6 of 13 Allen algebras (Allen, 1983) in the first column, and how these translate to archaeological group relationships for groups P and Q (columns two and three). Column four shows the logical statements implied by the prior knowledge in column two, which is then used to construct $\mathbb{I}_G(\boldsymbol{\alpha}, \boldsymbol{\beta})$ (see equation 6.5).

Given such group relationships, this prior knowledge about group boundaries is expressed using the indicator function

$$\mathbb{I}_G(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \begin{cases} 1 & \boldsymbol{\alpha}, \boldsymbol{\beta} \in G \\ 0 & \text{otherwise,} \end{cases} \tag{6.5}$$

such that G is the set of values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that satisfy the logical constraints imposed on the group boundaries by the Allen algebra relationships.

Naylor and Smith (1988) set $p(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbb{I}_G(\boldsymbol{\alpha}, \boldsymbol{\beta})$ such that all group relationships were assumed to be abutting (equivalent to the 'meets' Allen algebra), which was intended to imply that the only information we have about the group boundary ages is the group orderings and abutting relationships between them. However, Nicholls and Jones (2001) later argued that while setting $p(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbb{I}_G(\boldsymbol{\alpha}, \boldsymbol{\beta})$ might

represent no prior knowledge about the individual group boundary ages other than their orderings, it inadvertently imposes a non-uniform prior on the span of the chronology for their case study on a model containing 6 groups and 7 unordered contexts. Nicholls and Jones (2001) define the span as $s(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \max(\boldsymbol{\alpha}) - \min(\boldsymbol{\beta})$. They demonstrated that the prior $p(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbb{I}_G(\boldsymbol{\alpha}, \boldsymbol{\beta})$ imposes a preference in the model for a span of length $2s$ by a factor of 30 when compared to a span of length s. This results from the marginal prior of $s(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is equal to $s(\boldsymbol{\alpha}, \boldsymbol{\beta})^{d-2}R - s(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Thus, there is a favouring of larger spans when we set $p(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbb{I}_G(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Therefore, they proposed an alternative prior,

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{s(\boldsymbol{\alpha}, \boldsymbol{\beta})^{2-d}}{R - s(\boldsymbol{\alpha}, \boldsymbol{\beta})}, \tag{6.6}$$

such that d is the number of group boundary parameters and $R = P - A$ is some span defined by the chronological modeller that is larger than the plausible span of the chronology. For example, a cautious modeller using IntCal20 (Reimer et al., 2020) might set $P = 55,000$ and $A = 0$, thus setting R to be $55,000$, representing the whole span of the calibration curve.

The changes to $p(\boldsymbol{\alpha}, \boldsymbol{\beta})$ provided in Nicholls and Jones (2001) were widely accepted by the chronological modelling community and implemented in BCal and OxCal (Millard, 2015). Thus in the remainder of this thesis, we also define our prior on the group boundaries to be

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{s(\boldsymbol{\alpha}, \boldsymbol{\beta})^{2-d}}{R - s(\boldsymbol{\alpha}, \boldsymbol{\beta})} \mathbb{I}_G(\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{6.7}$$

6.1.3 Posterior for the hierarchical model

To obtain our posterior $p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{x})$, up to proportionality, we take the following product,

$$L(\boldsymbol{\theta}; \boldsymbol{x})p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{s(\boldsymbol{\alpha}, \boldsymbol{\beta})^{2-d}}{R - s(\boldsymbol{\alpha}, \boldsymbol{\beta})} \mathbb{I}_G(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mathbb{I}_C(\boldsymbol{\theta}) \times \prod_{j=1}^{J} (\alpha_j - \beta_j)^{-N_j} \prod_{i=1}^{N_j} \mathbb{I}_{(\beta_j, \alpha_j)}(\theta_{i,j}) \exp\left(-\frac{(x_{i,j} - \mu(\theta_{i,j}))^2}{2(\sigma_{i,j}^2 + \delta(\theta_{i,j})^2)} \right). \quad (6.8)$$

The structure of this hierarchical model is provided in DAG format in Figure 6.1. Note that this is a graphical representation of the overall hierarchical model structure and not a graphical representation of prior knowledge as seen in this thesis thus far.

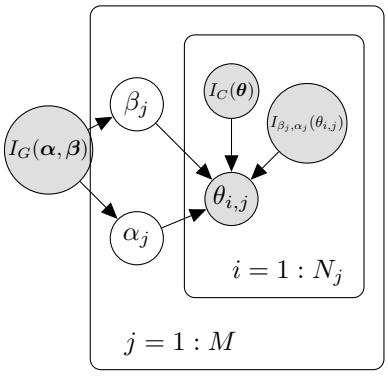


Figure 6.1: DAG representation of a Bayesian chronological model such that grouping, group relationships, and stratigraphic relationships are included as prior knowledge for N groups such that group j contains N_j contexts.

6.2 Sampling methods for hierarchical models

We utilise sampling methods to estimate the parameters' marginal posterior distributions in $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. However, Gibbs sampling requires us to express the conditional posterior distributions for each parameter within our model. Alternatively, another type of MCMC algorithm called Metropolis-Hastings only requires the use of the full joint density distribution for a given chronological model. Since we wanted to automate the process of chronology construction, it was preferable only to have to automate the calculations required to use one density distribution as opposed to the multiple conditional densities required for Gibbs sampling.

Furthermore, Nicholls and Jones (2001) provides pseudo-code for a Metropolis-Hastings algorithm that scales well for large models. Thus, we chose to implement their algorithm within this research. Since the algorithm closely follows the work of Nicholls and Jones (2001), we omit discussion from the main text. However, we include details of the algorithm in Appendix A Section A.3, where we outline subtle changes made to our algorithm to account for ordering samples within groups, which was not included in the original Nicholls and Jones (2001) model.

Further details of the computational process of Bayesian chronology construction are provided in Chapter 7. First, the remainder of this chapter demonstrates how DAGs can be used to represent chronological models and how a variety of plausible chronological models can arise from a single stratigraphic sequence.

6.3 Representing prior knowledge as mathematical graphs

Graph theory lends itself rather nicely to the purpose of representing stratigraphic data. As previously mentioned in Chapter 2, mathematical graphs comprise nodes and edges that connect some or all of the nodes. Although a relatively simple premise, the ability to impose specific conditions on mathematical graphs means we can represent complex data in a rigorous and logical format.

Recall also from Chapter 2 that, for this research, there are two conditions that the mathematical graphs we use must satisfy: that they are directed (because time flows in only one direction) and that they have no cycles (i.e. are acyclic). Consequently, all graphical representations of chronological models must be directed acyclic graphs (DAGs). Representing a stratigraphic sequence as a DAG is straightforward (see Chapter 2) but conveys only the prior knowledge contained in $\mathbb{I}_C(\boldsymbol{\theta})$. Dye and Buck (2015) noted the importance of also including information from $\mathbb{I}_G(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\mathbb{I}_{\alpha_j,\beta_j}(\theta_{i,j})$ to create a complete prior chronological DAG that might be suitable to aid in the semi-automation of Bayesian chronology construction.

To automate chronology construction, we have to formalise precisely when in the modelling process, different components of the prior knowledge are introduced and ensure archiving is facilitated at each step. To do this, we decided that first, all prior knowledge pertaining to physical relationships (i.e. the stratigraphic relationships) would be represented in a DAG (which we define as a stratigraphic DAG). Following this, the stratigraphic DAG would be converted into a chronological DAG, which would represent the temporal relationships implied by the stratigraphy and thus provide us with the prior for chronological modelling. From the chronological

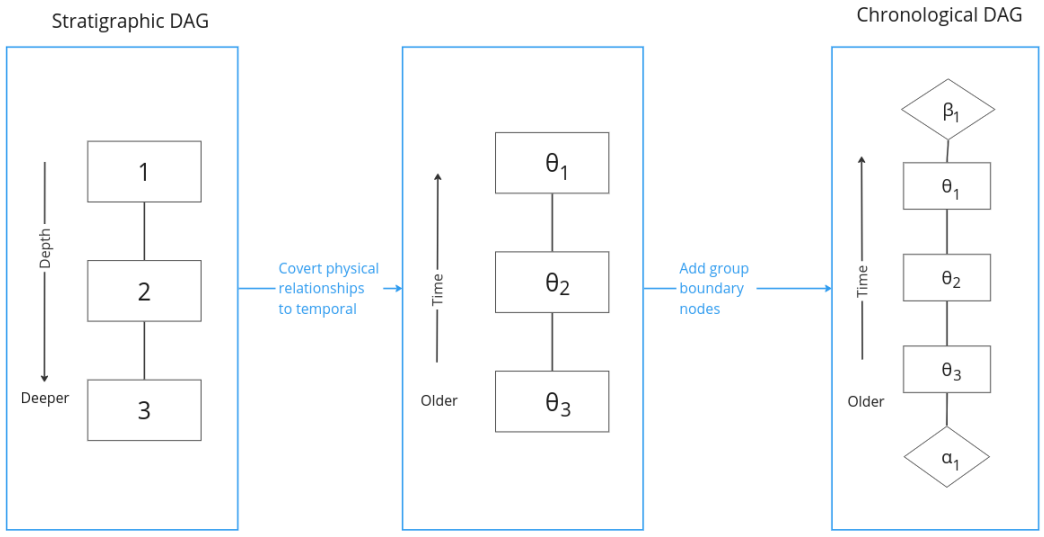


Figure 6.2: A diagram showing the conversion of a simple stratigraphic DAG with three contexts (far left) to chronological DAG (far right). The first step comprises converting the graph to temporal space rather than physical. In the second step, nodes are added to represent group boundaries (represented as diamond-shaped nodes to distinguish them from contexts), and edges are added to represent their relationship with the contexts in the original stratigraphic DAG.

DAG, we can then computationally extract the prior knowledge to pass to an MCMC algorithm and, thus, ultimately obtain posterior estimates for chronological parameters of interest.

As discussed in Chapter 2, physical stratigraphic relationships between contexts do not always directly correspond to temporal relationships between the true calibrated ages of samples found within those contexts. Therefore, to semi-automate the process of chronology construction, we had to clearly define when we are working in physical space or temporal space, to ensure the correct information is used to form a chronological model automatically and that the conversion from the stratigraphic record is clearly and consistently documented.

In its simplest form, this conversion comprises two steps, as demonstrated in Figure 6.2. In the first step, we convert from working with physical relationships to temporal relationships. Thus, the nodes no longer represent contexts that are physically above and below each other but the true calibrated ages of the samples found within these contexts. Furthermore, the edges no longer represent above-below relationships in physical space. In a chronological DAG, they now represent a before-after relationship in time. For example, in Figure 6.2, the stratigraphic DAG presents the information that context 1 is above context 2. Whereas the chronological DAG conveys that $\theta_1 < \theta_2$ (i.e. the sample in context 1 is younger than the one in context 2). Because of this change in what the edges represent, it is imperative that the samples date the contexts in which they were found. If this is not the case, then this must be represented in the chronological DAG. We demonstrate how to do this and discuss such examples in Section 6.4.1.

The second stage of conversion to the chronological DAG requires adding new nodes since a DAG that includes all the information provided by the aforementioned indicator functions requires nodes for all parameters in $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. In this thesis (as suggested by Dye and Buck, 2015), we represent group boundary parameters as diamond shapes nodes to distinguish them from nodes that represent calibrated ages of samples found in contexts (which are presented as rectangles and referred to henceforth as context nodes for brevity). An example of a chronological graph is provided on the right of Figure 6.2. Thus far, we have shown a simple example of obtaining a single chronological DAG from a stratigraphic DAG with grouping information. However, multiple chronological graphs can arise from any given stratigraphic DAG, depending on the prior information available.

6.4 Examples of hierarchical Bayesian chronological models

We will demonstrate how multiple chronological DAGs (which each represent a different plausible chronological model) can arise from a single small stratigraphic sequence. While archaeologists typically proceed with a single chronological model, these case studies demonstrate the impact on the results (and thus the archaeological interpretation) that can result from small changes to the chronological model. This is a form of sensitivity analysis, which is the practice of constructing alternative models and examining the impact on the posterior density estimates (see Bayliss and Marshall, 2022, Chapter 2). By doing this sensitivity analysis, we motivate the need to consider multiple plausible chronological models when there is uncertainty in how the archaeological record might be interpreted.

The stratigraphic and chronological DAGs shown in Figure 6.3 are taken from a stratigraphic sequence observed during excavations at St Veit-Klinglberg in Austria (Buck et al., 1994). This stratigraphic sequence will be used for the remainder of this chapter to showcase different hypothetical chronological DAGs that may arise from the same stratigraphic sequence. Larger sites (and thus larger stratigraphic sequences) are examined in Chapter 7. We chose this small stratigraphic sequence so that changes within the chronological DAG can be easily observed in diagrams.

Note that in the chronological DAG in Figure 6.3, and indeed all chronological DAGs for the remainder of this thesis, nodes representing $\boldsymbol{\theta}$ parameters are now represented by their context labels. It is important to note that in a chronological DAG any given rectangular node with label c still represents the parameter θ_c as it did in Figure 6.2. The use of context labels only is to aid in the visualisation of larger chronological

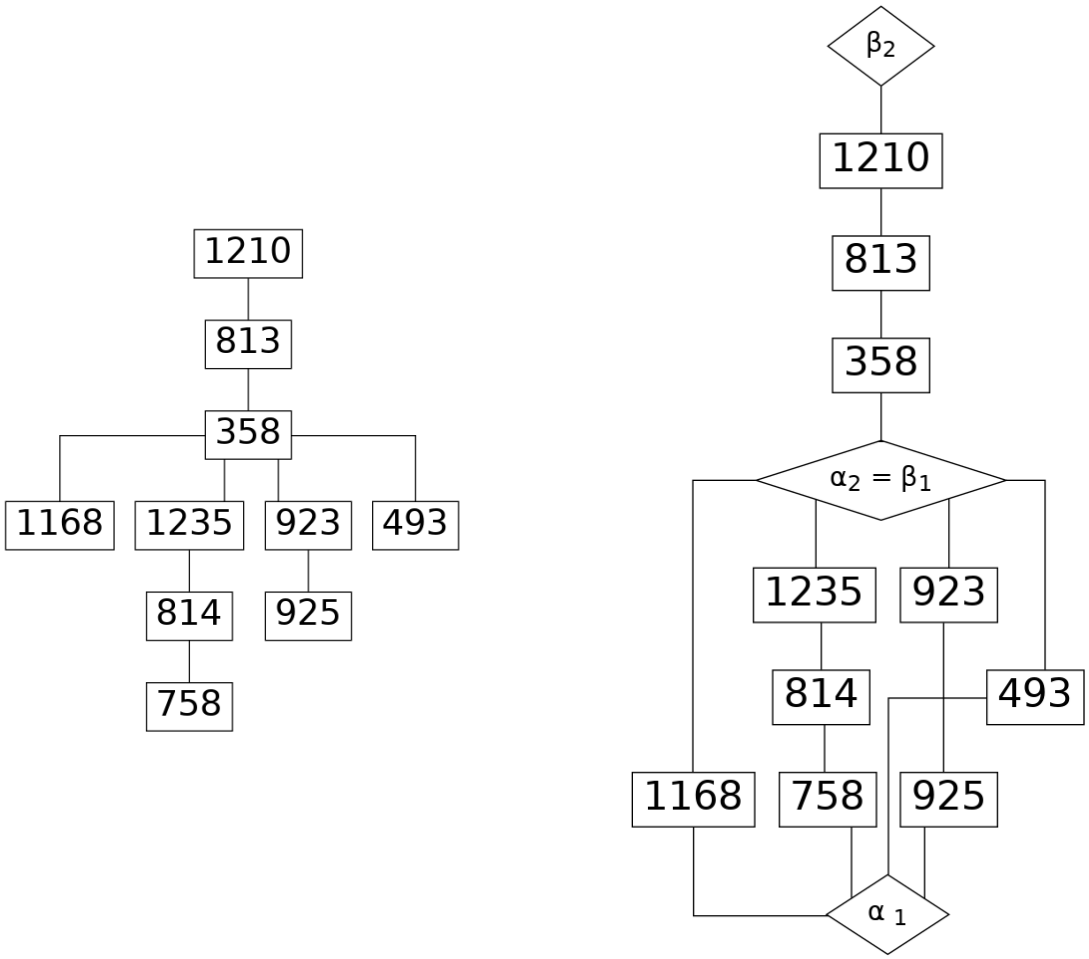


Figure 6.3: *left*: stratigraphic DAG representing the physical stratigraphic relationships in a sequence from excavations at St Veit-Klinglberg (see Buck et al., 1996, Chapter 9); *right*: chronological DAG representing the chronological sequence implied by the stratigraphy as well as grouping information for the contexts. Note diamond-shaped nodes represent group boundary parameters, and rectangular nodes represent calibrated ages of samples found in contexts. For any given context c in a chronological DAG, the node represents the parameter θ_c.

models (as we will see in Chapter 7). Chronological DAGs can be distinguished from stratigraphic DAGs within this thesis by the inclusion of diamond shape group boundary nodes in chronological DAGs.

6.4.1 Case study one: residual and intrusive samples

In Figure 6.2, we demonstrated the conversion of physical relationships in a stratigraphic DAG to temporal relationships in a chronological DAG, such that each physical above-below relationship directly corresponded to a temporal before-after relationship. A key assumption of using the stratigraphic sequence to infer the ordering of the true calibrated ages of samples is that the sample dates the contexts in which it was found. However, this is not always true. There are two reasons why a sample might not date the context in which it was found; either it is older than the context or younger. Older samples that contaminate younger contexts are referred to by archaeologists as residual samples, and younger samples that contaminate older contexts are called intrusive samples. Though residual and intrusive samples cannot be used directly to estimate the calibrated age of the contexts in which they were found, they still provide useful dating evidence.

For example, a radiocarbon determination for an organic sample that is deemed residual provides the oldest possible calibrated age of that context. Indeed, the true age of the context must be younger than the sample found within the context. The age of an intrusive sample provides what is referred to as terminus ante quem (TAQ), an age which is the youngest possible calibrated age of a context. Similarly, residual samples provide us with a terminus post quem (TPQ), which gives the oldest possible calibrated age for the context in which it was found, these in turn lead to changes in the edges in the chronological DAG. Incorrectly assuming a sample within a context is not residual/intrusive can lead to constraints on other parameters within

the chronological model that are not supported by archaeological evidence and thus lead to erroneous posterior density estimates.

Archaeologists take great care to avoid carrying out scientific dating on residual and intrusive samples. For some samples, it is clear when excavating that the sample might be residual or intrusive. For example, the 'old wood' problem (Schiffer, 1986) is a problem encountered by archaeologists when trying to radiocarbon date charcoal samples that originate from the burning of old wood. Radiocarbon dating such a sample will result in an estimate for the calibrated age of the tree. If this happened long before it was used for fuel, then the charcoal sample cannot reliably date the context in which it was found and is, therefore, residual. However, it is not always the case the samples are identified as potentially residual or intrusive on-site. Instead, this may be suspected after scientific dating is carried out (as discussed in Dye and Buck, 2015, Section 5).

Intrusive sample in St Veit-Klinglberg stratigraphic sequence

Buck et al. (1994) postulated that two contexts in the St Veit-Klinglberg sequence contained intrusive samples, contexts 1168 and 493. In this example, we examine the impact on our posterior estimate of the true calibrated age of context 1168 if it is indeed intrusive. While the stratigraphic DAG remains the same (since context 1168 is stratigraphically below context 358), in the chronological DAG, we cannot be sure the age of the sample found in context 1168 is older than context 358. Therefore, θ_{1158}, the true calibrated age of the sample found in context 1158, is only constrained in the Bayesian model by α_1, which is older than it and cannot be constrained by parameters below it according to the stratigraphic sequence. This can be seen in Figure 6.4 by the removal of all edges connecting 1158 (θ_{1158}) to any parameters other than α_1.

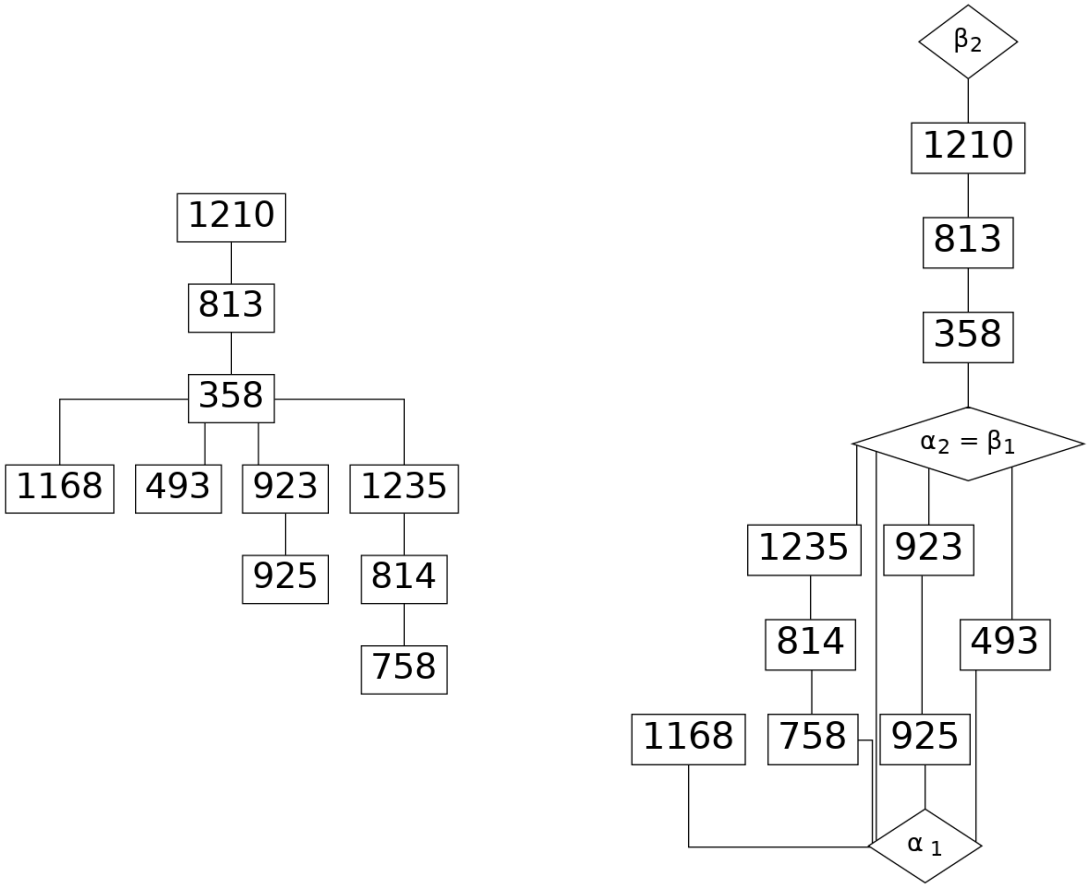


Figure 6.4: *left*: stratigraphic DAG representing the physical stratigraphic relationships in a sequence from excavations at St Veit-Klinglberg (Buck et al., 1994); *right* chronological DAG representing the chronological sequence implied by the stratigraphy as well as grouping information for the contexts when the sample found in context 1168 is considered intrusive. Note that diamond-shaped nodes represent group boundary parameters, and rectangular nodes represent calibrated ages of samples found in contexts. For any given context c in a chronological DAG, the node represents the parameter θ_c.

We obtain two marginal posterior estimates for θ_{1168}, using a model when θ_{1168} is considered residual and one where it is not, using the Metropolis-Hastings algorithm outlined in Appendix A Section A.3. The posterior probability densities of these estimates are provided in Figure 6.5. As shown by plots in that figure, the posterior estimate of θ_{1168} tends towards the younger calibrated ages supported by the likelihood when it is treated as intrusive as opposed to when it is not. Further, the marginal posterior estimate for θ_{1168} is notably more precise when it is not considered intrusive. The change observed in the results in Figure 6.5 exemplifies the impact residual/intrusive samples can have on the results of the Bayesian chronological modelling. Thus, if there is uncertainty as to whether a sample is intrusive/residual or not, we argue that both models should be produced and documented.

6.4.2 Case study two: Different type of group relationships

In addition to residual and intrusive samples, different chronological DAGs (representing different chronological models) may arise from the same stratigraphic DAG, depending on the Allen algebra relationships between the groups in the model. Recall from Table 6.1 that we consider three types of group relationships: abutting (meets), gap and overlapping. The contexts in the St Veit-Klinglberg stratigraphic sequence were separated into groups 1 and 2 such that group 1 is older than group 2. Buck et al. (1994) defined the relationship between the two groups as abutting. However, as an illustrative example, we now investigate the impact on the posterior results if alternative group relationships are considered. The chronological DAGs for the three possible Allen algebra group relationships between groups 1 and 2 are provided in Figure 6.6.

To examine the impact of different group relationships on our results, demonstrating that our choice of model is important, we consider the group boundary parameter for

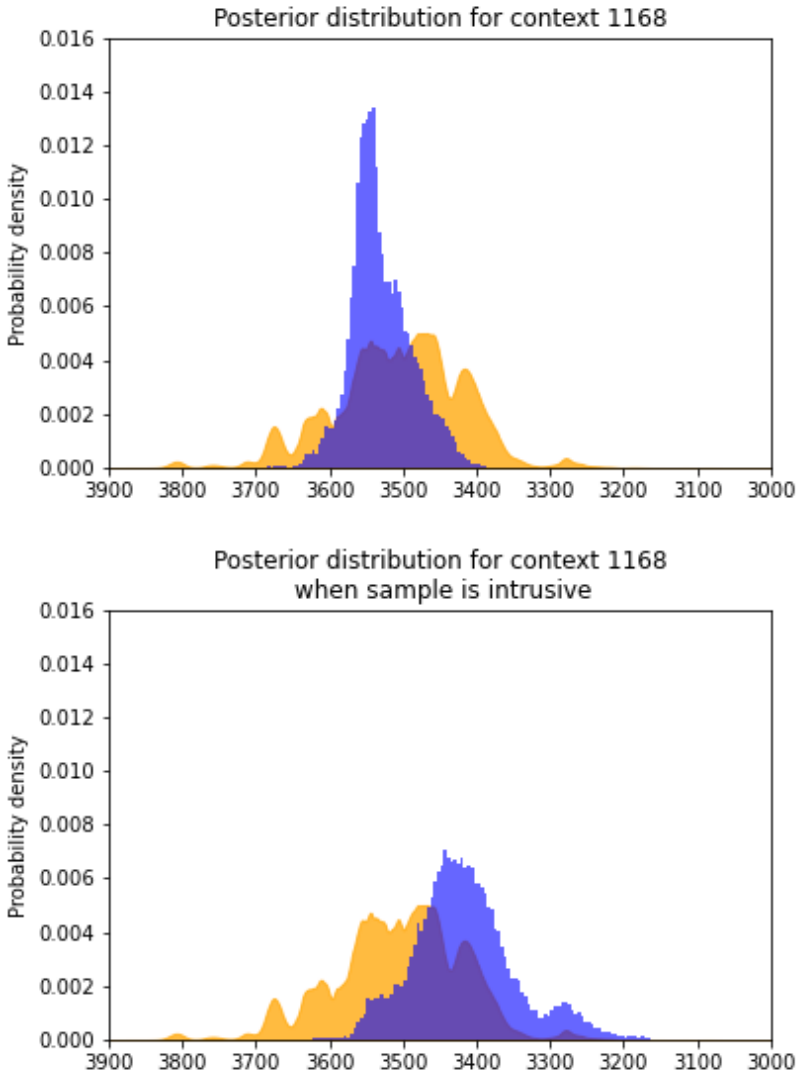


Figure 6.5: Estimated marginal posterior density plots (blue histograms) for the true calibrated age of a sample found in context 1168 at St Veit-Klinglberg with radiocarbon determination 3435 ± 60 BP. *(top)*: if the sample is not considered intrusive, and *(bottom)* if the sample is considered intrusive. A marginal posterior density when no prior knowledge is included (yellow plots) for the true calibrated age of the sample is provided in both plots for reference.

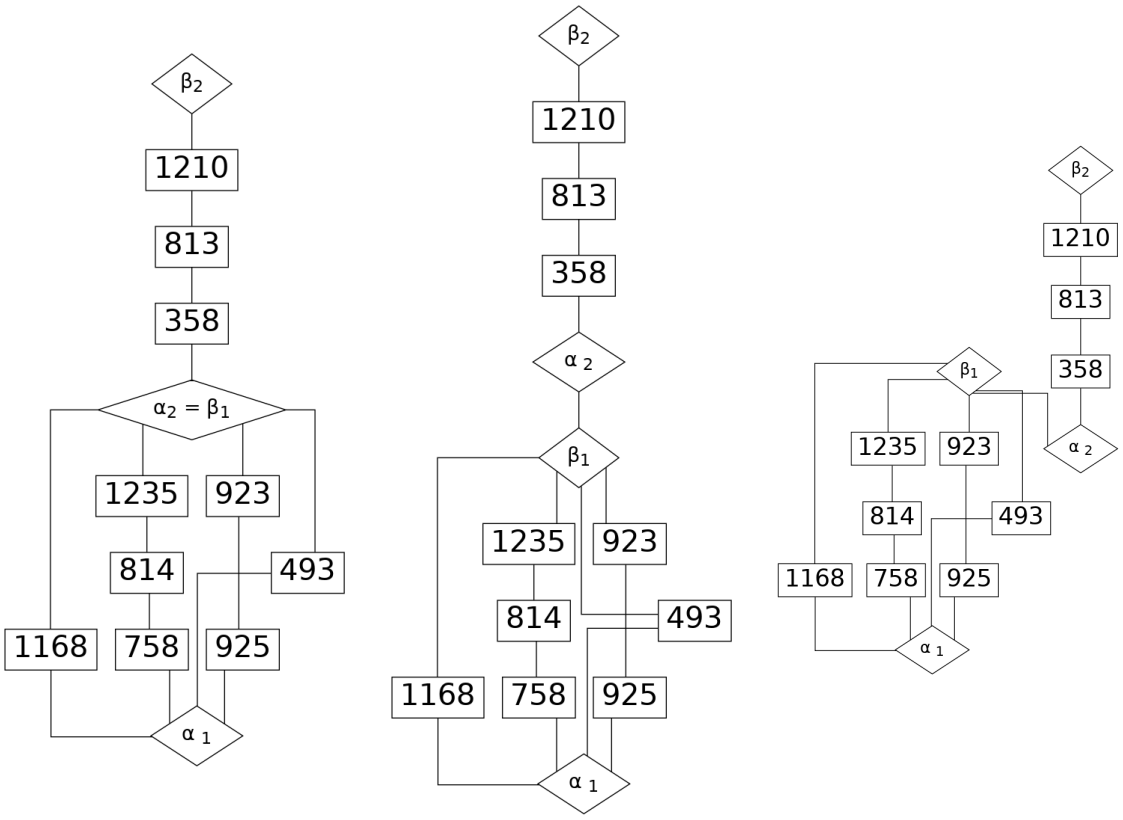


Figure 6.6: Three chronological DAGs showing the possible relationships between groups 1 and 2 for St Veit-Klingberg chronological model presented in Buck et al. (1994) Chapter 9. Relationships shown are when groups 1 and 2 are *left:* abutting, *middle:* gap, *right:* overlap. Note that diamond-shaped nodes represent group boundary parameters, and rectangular nodes represent calibrated ages of samples found in contexts. For any given context node c in a chronological DAG, the node represents the parameter θ_c.

the start of group 2 α_2. Plots of the posterior marginal densities for this parameter for the three different models are provided in Figure 6.7. As can be seen in the plots, when group 2 is considered to overlap with group 1, this leads to older ages on the calibrated scale being deemed plausible for α_2. Just as the previous case study showed how intrusive samples affect posterior results for a chronological model, different group relationships can change the results of the marginal posterior estimates of specific parameters on the calibrated scale by order of 100 years, a considerable change to the results, especially if there is no (or weak) archaeological evidence to support one type of group relationship over another.

Though the indicator functions in the prior allow us to include useful relative dating evidence in our Bayesian chronological models, they impose strict orderings on the parameters in the model, with no accounting for uncertainty. If such uncertainty exists, producing and documenting all chronological models that an archaeologist considers plausible for a given site provides a record of that uncertainty.

6.5 Conclusion

In this chapter, we have introduced all of the remaining established theory required for the interdisciplinary research discussed in this thesis. So far, we have introduced the concepts of stratigraphy and grouping and how they provide relative dating evidence. Further, we have introduced radiocarbon dating and how to take a radiocarbon determination and convert it to the calibrated scale. Following this, we introduced research (carried out by the author) into the quality and utility of relative dating evidence within digital repositories at ADS. Next, we discussed using Bayesian inference in chronology construction and the various components required to implement it: a calibration curve, expert knowledge and radiocarbon determ-

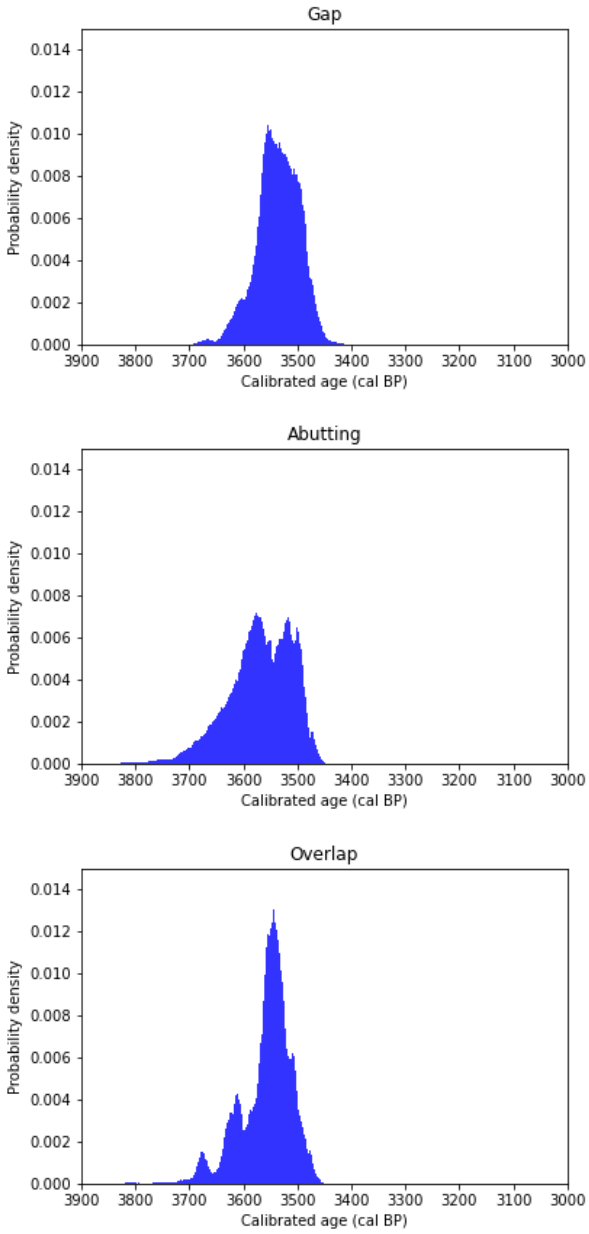


Figure 6.7: Plots showing estimated posterior density plots for the marginal distribution of α_1 for three different chronological models when the two groups in the model are considered to *top graph:* have a gap between them, *middle graph:* be abutting, *bottom graph:* have an overlap

inations. Finally, the concept of using mathematical graphs to represent relative dating evidence has been introduced. We have shown how multiple chronological models can arise from a single small stratigraphic sequence. For the remainder of this thesis, we use all the previous theory to demonstrate how we can automate the process of chronology construction from site records to final chronological interpretation, allowing for the production and analysis of multiple plausible chronological models.

Chapter 7

PolyChron

Following the results shown in the case studies in Chapter 6, we argue that when carrying out chronological modelling, we should document any uncertainty pertaining to which chronological models are most appropriate and formally archive any plausible models along with the results of any analysis of these models. However, when the research in this thesis commenced, no software existed to allow for the semi-automated construction of multiple Bayesian chronological models for a given site. Therefore, in this chapter, we extend upon the theory in Chapter 6, Section 6.3 and demonstrate that we have successfully implemented the formalisation of the process of handling relative dating evidence using DAGs in prototype software.

When semi-automating the process of building chronological models, a fundamental choice was whether to produce new purpose-built software or produce code to connect existing (but disconnected) software for different parts of the chronological modelling process. We chose the former writing software in `Python` (Van Rossum and Drake, 2009), including writing the underlying `Python` code that runs the MCMC

algorithm needed to obtain our calibrated age estimates for the parameters within our chronological model, as seen in Chapter 6 Section 6.2. Recall that we wrote the code for that algorithm, utilising pseudo-code from Nicholls and Jones (2001), despite other off-the-shelf software being available to carry out Bayesian analysis for chronological models.

We do not wish to reinvent the wheel (or wheels in this case), and we are aware of various types of software for chronological modelling and managing relative and absolute dating evidence. However, as we demonstrate in this chapter, our research is akin to proving that we can use all the wheels together to build a vehicle for end-to-end chronological modelling. Metaphors aside, as demonstrated in Figure 7.1, there are many stages in the process of chronology construction. We sought to demonstrate that it is possible to manage relative and absolute dating evidence (from revising relative dating evidence post-excavation to obtaining posterior calibrated estimates) in a singular piece of software or by connecting existing software. As it is increasingly seen as best practice in research (Open Research Data Task Force, 2019), it was important that any new or existing software was open-source. Further, we wanted the software to be free as we felt this would allow archaeologists to utilise it within the strict budget constraints that excavations and post-excavation research are often subject to.

In this chapter, we review the existing software for building Bayesian chronological models alongside software for managing stratigraphic data and motivate why new purpose-built software that manages the whole process of chronological modelling was required. From this, we introduce our prototype software, `PolyChron`, which provides a graphical user interface to our MCMC algorithm and thus allows users to manage stratigraphic data and automatically render the results of Bayesian chrono-

logical modelling with no further data input required by the user. Finally, we discuss the benefits and limitations of `PolyChron` and how we hope to develop it.

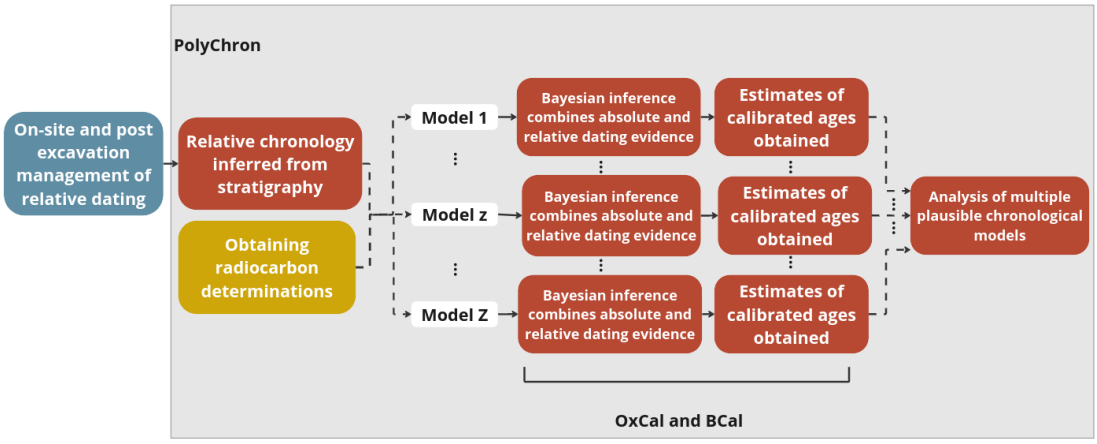


Figure 7.1: Overview of the process of building multiple plausible Bayesian chronological models. The dotted arrows highlight the part of the process that `PolyChron` automates as demonstrated in Section 7.3. All parts of the process that are carried out in `PolyChron` lie in the grey box.

7.1 A history of chronological modelling software

Before the introduction of chronology construction in radiocarbon dating in the late 1990s, there were two types of software for chronology construction. The first was for calibrating single radiocarbon determinations using the calibration curves introduced in Chapter 3, and the second was for managing and visualising relative dating evidence.

Following the development of radiocarbon calibration curves, software was developed in the 1980s to calibrate radiocarbon determinations. The first widely used piece of software was CALIB (Stuiver and Reimer, 1993), which allowed users to calibrate radiocarbon determinations, produce probability distribution plots and carry out

statistical tests for contemporaneity of radiocarbon determinations (using the methods of Ward and Wilson, 1978). Following the introduction of Bayesian chronological modelling as discussed in Chapter 5, the calibration of radiocarbon determinations was subsequently incorporated into Bayesian chronology construction software such as `OxCal` (Bronk Ramsey, 2023) and `BCal` (Buck et al., 1999).

Software for managing relative dating evidence

Following the introduction of the Harris matrix (Harris, 1975) in the early 90s, the community developed various programs to display stratigraphic data as a Harris matrix; see Herzog (2006) for a detailed summary. However, as Herzog (2006) highlighted, most of these are no longer maintained. Currently, the most commonly referenced software for managing stratigraphic data are:

- `ArchEd` which is available at `https://arched.software.informer.com/`,
- `jnet`, which was previously known as `gnet` (Ryan, 2001),
- `Harris Matrix Composer` (Traxler and Neubauer, 2008),
- and `Stratify` (Herzog, 2006).

Additionally, as discussed in Chapter 4, commercial and academic units may also use their in-house software or general software, such as Microsoft Excel, to produce diagrams of stratigraphic sequences stored as PDFs. The Museum of London Archaeology (MOLA) use the `Bonn Archaeological Software Package (BASP)` (Herzog, 1993), which allows the user to construct large stratigraphic sequences. However, this software is not open source and, thus, would not have provided the flexibility required for this project.

`ArchEd` and `jnet` allow users to construct a Harris matrix and manipulate the

diagram before exporting the resulting image. However, it is unclear if `jnet` is still maintained since previously published links to the software are now broken. `ArchEd` is still being maintained, but its functionality is limited to producing Harris Matrix diagrams. Further, the user manually enters stratigraphic relationships individually, which we felt would have proved too time-consuming when visualising large stratigraphic sequences.

`Stratify` has additional functionality, allowing users to define phases, check for cycles in their stratigraphic relationships, and produce stratigraphic diagrams. A particular strength of `Stratify` is that the user can incorporate additional site data, such as site photographs and drawings. However, despite having additional functionality, `Stratify` does not have point-and-click functionality to allow the user to produce multiple stratigraphic sequences quickly and easily, upon which the later work in this thesis relies.

The `Harris Matrix Composer` provides similar functionality to `Stratify` and is scalable for larger sites. However, the free version of this software is limited to 50 contexts. As such, this was not conducive to our aim of free, open-source software since we sought to be able to automate the chronology construction process for stratigraphic sequences of all sizes. Therefore, it was important that any software used be free and open-source for all parts of the chronology construction process, whether we used one singular new piece of software or a combination of existing software with new code to connect them.

In addition to the previously mentioned software, there exists hm (Dye, 2019), which is purpose build command line software that allows the user to display stratigraphic and chronological DAGs (as seen in Chapter 6 Section 6.3). This software scales very well for large stratigraphic sequences, results of which are demonstrated in Dye and

Buck (2015). However, we wished for the software we chose to be suitable for the whole chronological community rather than just those comfortable with command-line programming. Thus, we decided software with a graphical user interface would be more appropriate.

Keith May is currently leading a team that is working on software `PHASER` that allows users to manage a wealth of archaeological dating evidence using a graphical user interface (May et al., 2023). This software was not available at the start of this PhD project, but the authors do seek to investigate the potential and benefits of interfacing `PHASER` with `PolyChron` in the future.

All of the programs previously mentioned provided certain functionality that we required and provided inspiration for our work. However, none allowed us to produce stratigraphic and chronological DAGs visualisations while providing point-and-click functionality to adjust such mathematical graphs, thus producing varying plausible chronological models. Therefore, we decided early in the project that we would produce a graphical user interface (GUI) that allows the user to import relative and absolute dating evidence, with the goal of displaying and manipulating stratigraphic and chronological DAGs. In addition, due to the lack of reusable data found in the ADS archives (as discussed in Chapter 4), we felt that building a purpose-built GUI would allow us to write code that would support users in the archiving on their work once complete.

Review of Bayesian archaeological chronology construction software

Purpose-built software for chronology construction is not limited to archaeological chronology construction. Packages such as BChron (Haslett and Parnell, 2008) and BACON (Blaauw and Christen, 2011) also allow for the calibration of radiocarbon

determinations and the statistical modelling of the true calibrated ages of samples. However, these packages are used for paleo-environmental modelling and primarily focus on modelling the relationship between the calibrated age of samples and their depth in a vertical sediment profile (Blaauw and Christen, 2011). In archaeology, we obtain radiocarbon determinations when datable organic samples are found in archaeological deposits that are removed in the reverse order of their deposition. Thus, although the depth of contexts is recorded during excavation, only the relative depth of archaeological deposits with respect to other deposits is used in the modelling process, rather than the absolute depth value for any given sample. Thus, in what follows, we only discuss software and packages used for Bayesian archaeological chronology construction.

`OxCal` is the most commonly used software for Bayesian archaeological chronology construction and is the software of choice by Historic England (Bayliss and Marshall, 2022). As acknowledged in the user manual (Bronk Ramsey, 2023), `OxCal` requires the user to have a reasonable understanding of the implications of their modelling decisions. They state in their manual that there is the potential for accidentally introducing bias if the user is not experienced in statistical modelling. `BCal` (Buck et al., 1999), on the other hand, has reduced functionality when compared to OxCal, following a more linear approach as it guides the user through the GUI to build their model. As a result, there is less opportunity for the user to produce statistically implausible models.

`OxCal` and `BCal` enable the user to calibrate a single radiocarbon determination and produce the corresponding probability density plots on the calibrated scale, effectively replacing the functionality of CALIB. In addition, they allow the user to build Bayesian chronological models. Other software such as DateLab (Jones and

Nicholls, 2002) and, more recently ChronoModel (Lanos and Philippe, 2018) have also been developed for Bayesian chronological modelling. Though, as discussed in Chapter 5, the Bayesian model introduced in Jones and Nicholls (2002) is now implemented in `BCal` and OxCal. Further, DateLab itself does not appear to be maintained any longer. On the other hand, ChronoModel is maintained and available for download on some variants of all major desktop operating systems. However, they assume a uniform prior on the calibrated ages conditional on the ages of the group boundaries (see Lanos and Philippe, 2018, Equation 5). As previously discussed, this assumption produces an artificially wide time span between the group boundaries. Thus, only `OxCal` and `BCal` are discussed in what follows.

`OxCal` formalises the construction of chronological models using its own chronological query language (CQL). This formal language allows users to define what type of absolute dating evidence they use and the relationships between the true calibrated dates of the samples used to obtain such evidence. See Bronk Ramsey (1997) Table 1 for an outline of the main functions one can use when writing in CQL. `OxCal` implements a GUI so that users do not have to be able to program directly in CQL. However, users still require sufficient statistical knowledge to understand the implications of the model they are building for the results they obtain.

Levy et al. (2020) utilise mathematical graphs in their `CHRONOLOG` software, which provides a GUI that can construct CQL scripts for interpretation by OxCal. Much like our research, this `CHRONOLOG` uses mathematical graphs to manage chronological information, allowing the user to visualise their relative and absolute dating evidence and run algorithms to check for logical inconsistencies in the information provided.

A useful feature for chronology construction `CHRONOLOG` provides is their *tightening function*, which allows the user to put in all prior information regarding the start,

end and duration of periods of archaeological activity in calendar years. According to prior information, the software calculates more accurate start and end calendar ages for such periods. However, the tightening function relies on written historical records that can provide a duration for archaeological events or start and end ages for such events. Much of British pre-history does not have such records. As such, we sought to provide a GUI focusing on managing relative and absolute dating evidence of the type seen in Chapter 2 and 3, where evidence in the form of written historical records are seldom used.

`BCal` also guides users through the chronological modelling process via a GUI within a web browser, providing questions specific to their project, which prompt the user to input their relative and absolute dating evidence (Buck et al., 1999). Once this has been done, `BCal` takes that information and automatically converts it to a format that can be interpreted by the Bayesian radiocarbon calibration software `mexcal` (written in C++). `BCal` does not require the user to directly construct the input to `mexcal`, and the linear method of inputting data makes it challenging to build implausible statistical models and thus may be more suitable for users with limited programming or statistics experience. However, it is limited compared to `OxCal` with respect to the types of absolute and relative dating evidence that can be incorporated into the model. Furthermore, users can only build a single chronological model at a time in both `OxCal` and `BCal`.

Despite both `OxCal` and `BCal` being well-established and utilised by the archaeological community, we chose to write our algorithm (utilising the algorithm provided in Nicholls and Jones (2001)) to build Bayesian chronological models automatically. Of all the issues previously discussed, our primary research motivation for doing this was to be able to build multiple chronological models automatically from user input.

To achieve this in OxCal or BCal would have required converting user input into CQL or mexcal, respectively. Whilst this would have been possible as Levy et al. (2020) have demonstrated, it would have limited all future development of PolyChron to use the functionality available in OxCal or BCal. Furthermore, neither OxCal nor BCal is open-source. Therefore, our end-to-end implementation of semi-automated chronological construction would not have been fully open source. Further discussion of the functionality of BCal and OxCal is reserved for Section 7.6, where we compare the functionality of OxCal, BCal and PolyChron.

7.2 Motivation for building PolyChron

PolyChron is prototype software required to be installed locally on a user's machine, written entirely by the author, allowing users to produce multiple chronological models. In addition, it allows the user to obtain graph theoretic representations of their stratigraphic sequences and prior knowledge within a given hierarchical Bayesian chronological model and save the raw digital data (as collected on-site) along with graph theoretic representations of their models, resulting outputs and supplementary notes produced during such modelling on their machine, thus facilitating future archiving of a complete site archive. In the following sections, we will explore the specific motivations for building the software the way we did and introduce the functionality of PolyChron. We also outline the software's layout, demonstrating its modular structure and why such structure is essential for the function and longevity of PolyChron. This section specifically discusses the goals we sought to achieve when producing PolyChron and how they correspond to this project's research objectives.

Improved archiving of relative and absolute dating evidence

A key goal when building `PolyChron` was to produce software that allowed for better archiving of all aforementioned data and information produced during the process of chronological modelling. It was particularly important that the software would enable us to address at least some of the issues with archiving the data (and supplementary information about the data) used and produced when building chronological models. As motivated in Chapter 4, issues such as incomplete site archives or lack of contextual information as to how or why dating evidence may have been augmented from its original form during the modelling process prevent the reuse of such dating evidence.

Figure 7.2 outlines a hierarchy of the importance we ascribe to digital archaeological data specific to chronology construction in archaeology. We classify the context label (or identifier) as the primary data since chronological modelling cannot be carried out without a set of contexts. Each context then has supplementary data, which together constitutes the complete information pertinent to chronological modelling, such as radiocarbon determinations, contexts, grouping, and group relationships. Within `PolyChron`, we seek to ensure that all of the data types outlined in Figure 7.2 are automatically saved to the user's machines as and when they are produced/input by the user. Further, it is important that it is clear what supplementary data corresponds to a given context so that following subsequent archiving of site data, there is the potential for the reuse of such data for further chronological modelling or other purposes.

Further, we wanted changes from the physical stratigraphic relationships in the stratigraphic sequence to the temporal relationships that comprise part of the chronological model to be a distinct and recorded step in the chronological modelling process

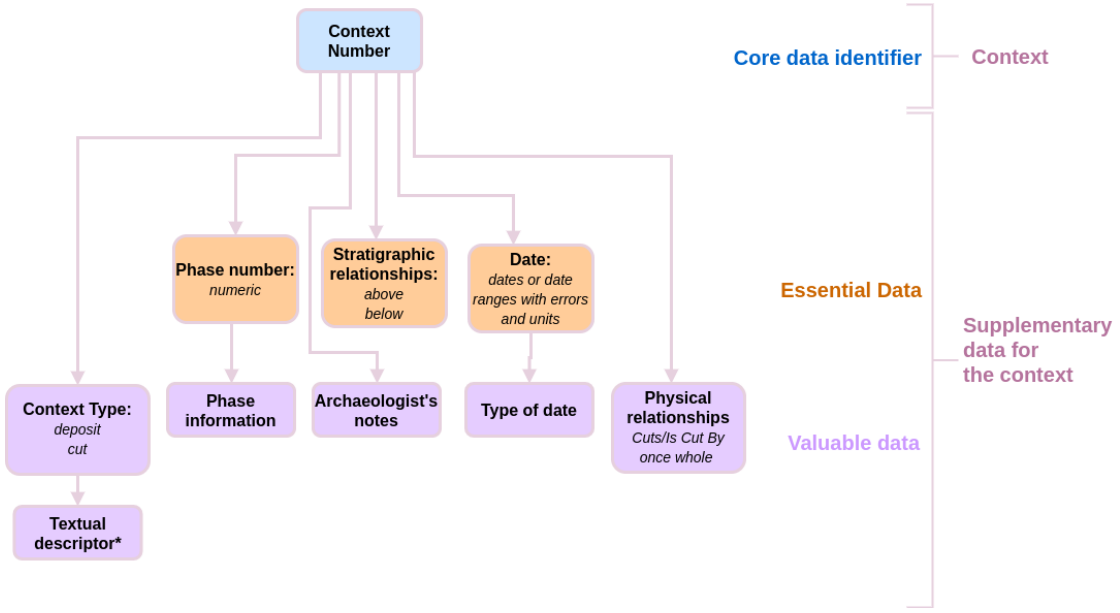


Figure 7.2: Diagram showing the types of data and supplementary data we seek in order to carry out chronological modelling. The hierarchical structure indicates the relative importance of each data type for the reusability of the information below. *Textual descriptor should be a brief description of the context using controlled vocabularies such as those defined by Historic England (2021).

as outlined in Chapter 6 Section 6.3. As such, a given site's stratigraphic DAG and chronological DAG needed to be archived as a record of specific steps in the chronological modelling process.

Automating the initial construction of chronological models

We have referenced the automation of chronology construction (precisely the steps highlighted in Figure 7.1) throughout this thesis as a specific goal of the PhD. More precisely, however, we wanted to be able to read into the software any given record of a stratigraphic sequence (provided as a table of pairs of contexts and their above-below relationships), grouping data and radiocarbon determinations and construct the chronological model (and obtain posterior estimates for all parameters within the model) automatically, regardless of the complexity or size of the stratigraphic sequence. This, in turn, leads to a large combination of possible stratigraphic sequences, numbers of groups and combinations of group relationships. Moreover, this is before one considers additional factors that can change the chronological model, such as residual and intrusive samples.

However, though the stratigraphic and group relationships expressed by the prior knowledge can vary widely, the hierarchical structure of the Bayesian model stays the same. Refer back to Chapter 6 Figure 6.1 for a graphical representation of the hierarchical Bayesian model defined in Chapter 6 Section 6.1.3.

In any given model, the true calibrated ages of contexts are constrained by their respective group boundaries and the ordering imposed by the stratigraphic sequence. Therefore, to be able to automate the construction of Bayesian chronological models given the fixed structure of the hierarchical model, the potential changes we need to consider are the number of parameters and the information that defines the indicator

functions $\mathbb{I}_{\alpha_j,\beta_j}(\theta_{i,j})$, $\mathbb{I}_G(\boldsymbol{\alpha},\boldsymbol{\beta})$, and $\mathbb{I}_C(\boldsymbol{\theta})$.

Since the indicator functions comprise collections of logical statements between the parameters in the model, we can extract that information from DAG representations of the stratigraphic sequence by automating the conversion of the relationships between the nodes into 'less than', 'greater than', or 'equals' relationships, as demonstrated in Chapter 6, Figure 6.2.

More specifically, to automate the construction of Bayesian chronological models, the goal was to find a way to consistently and reliably convert user input to a stratigraphic DAG and then subsequently convert the information represented by the stratigraphic DAG, in conjunction with further user input about residual and intrusive samples and group relationships, to logical statements for use in a purpose-built MCMC algorithm for Bayesian chronology construction.

Rapid implementation of multiple plausible models for a given site

A further goal was to allow for the rapid construction of multiple plausible Bayesian chronological models with little to no extra input or work required from the user. In particular, we wanted to ensure that this automation would work for large and complex stratigraphic sequences for which, with existing software, considering more than one plausible chronological model would not be feasible. As with automatically constructing a single model, we also wanted to ensure that all data, model output and supplementary notes for each plausible chronological model are saved in an appropriate format and that it is clear which models correspond to the same project.

Ability to develop and improve the software in future

It was important for `PolyChron` to have the potential for development in the future, such as adding other types of statistical models or additional data management functionality. Within the time constraints of the PhD, we were aware that we could not implement all possible Bayesian chronological models within the software that users may require, nor all desirable components of the GUI. However, we wanted to ensure that we built the software using a modular structure and that each component provided output in a standardised format that could be used as input to other components. As a result, adding new functionality to the software would be relatively simple by introducing new modular components and removing or replacing less functional or redundant components.

7.3 `PolyChron` structure

This section outlines the modular structure of `PolyChron`. We separate this structure into four components labelled: project loading, prior elicitation, MCMC algorithm, and post-MCMC analysis, which each correspond to a subsection within this section. Figure 7.3 outlines the functionality of each component and how they interact with one another, and a link to a video demonstration is provided in Appendix B Section B.1. In this section, we highlight the key functionality of each component, the structure of the information between them, and define terminology specific to `PolyChron`, which is used throughout the supplementary video demonstration.

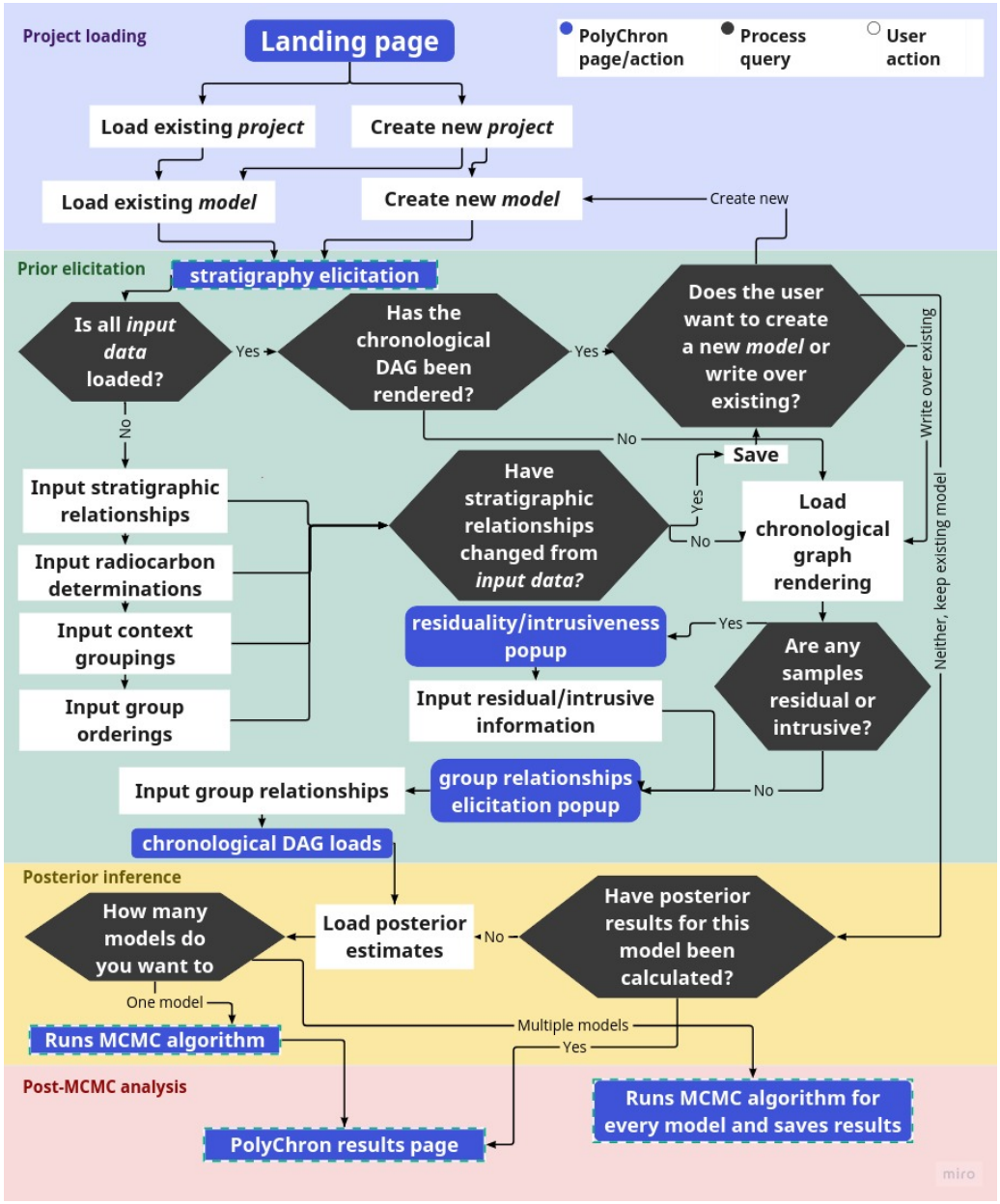


Figure 7.3: Diagram showing a simplified version of `PolyChron`'s structure and the user's process to obtain results for a Bayesian chronological model. Boxes with green and white borders are stages where data are automatically saved locally to the user's machine by `PolyChron` without the user needing to save project progress.

7.3.1 Project loading

The project loading module is the smallest of the four, and this part of the GUI is presented to the user upon loading `PolyChron`. Here, they set up their initial data structures. The highest level data structure that a user can make is a *project*, which is a directory that holds all files and directories containing data or information specific to a given archaeological site. Within each *project*, the user can set up *models*, where here, a *model* refers to the data structure that holds all the data and information for a single chronological model. Users can produce as many *models* as they require and all *model* directories are stored within the *project* directory. Note that all use of terminology such as *project* and *model*, which refer to specific data structures or functionality in `PolyChron`, will be presented in italics for the remainder of this thesis.

The project loading component allows the user to load new or existing *projects*, and within each *project*, they can load new or existing *models*. At present, the GUI only supports the display of one *model* at a time. Once the user has chosen the *project* and *model* to work on within the GUI, they can load the prior elicitation component of `PolyChron`. Upon doing this, for new `projects`, the software will create a *project* directory within the directory containing `PolyChron`. Within the *project* directory will be a directory for any existing *models*. Each *model* directory contains four folders that store data and information from various parts of the modelling process, as presented in Figure 7.3. We now introduce the other components of `PolyChron` before returning to discuss the file structure and data-saving process in more detail.

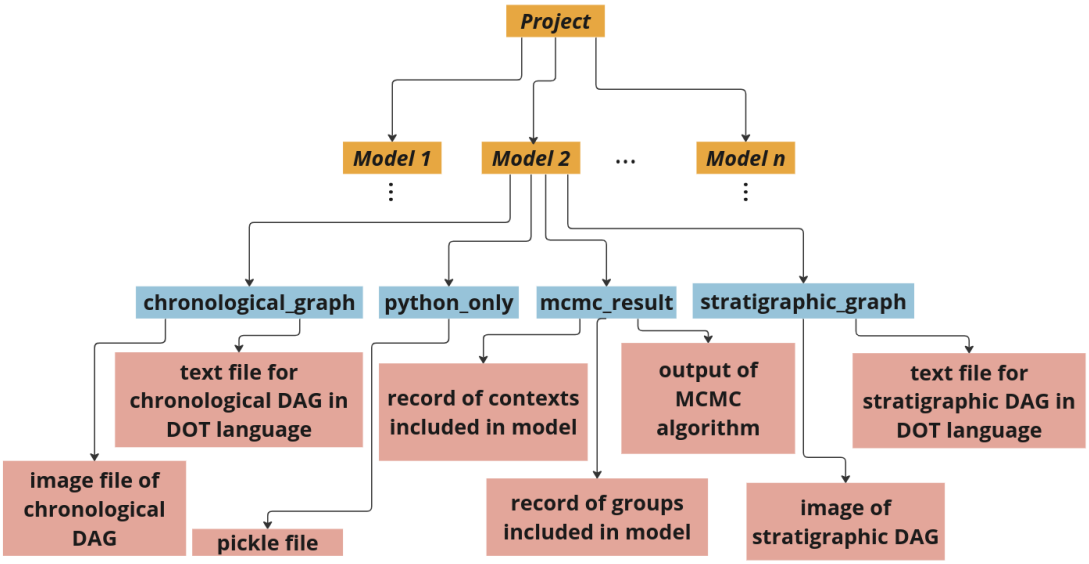


Figure 7.4: Directory structure that `PolyChron` builds when a user forms a new *project*. Directories are given in orange, folders in blue and individual files of various types in red. A pickle file is a file used to store Python objects in a single file that can be loaded back into Python at some future point.

7.3.2 Prior elicitation component

The prior elicitation component of `PolyChron` is part of the GUI, where the user reads in plain text files containing all the absolute and relative dating evidence they wish to use for their Bayesian chronological modelling. All functionality for this component is provided on a single page in the GUI. To build a Bayesian chronological model in `PolyChron`, users must load the following four types of data:

1. stratigraphic relationships,
2. context grouping,
3. group orderings,
4. and radiocarbon determinations.

We define these four files together as *input data* that provide the minimum information that `PolyChron` requires to allow users to construct a Bayesian chronological model. Users can also optionally load a file that outlines any contexts that have been deemed equal or contemporary. The required data format is provided in Appendix B Section B.2 and shown in the supplementary video demonstration.

Users must load the stratigraphic relationships file first, which `PolyChron` then converts to a stratigraphic DAG which is then displayed in the GUI. All data management in the prior elicitation component is done via mathematical graphs, specifically DAGs, using the `NetworkX` package (Hagberg et al., 2008). Once the file defining the stratigraphic relationships is loaded, `PolyChron` converts that data to a stratigraphic DAG. For DAGs constructed using the `NetworkX` package, each node (context) has an associated dictionary. A dictionary in `Python` is a data structure that can store any data type within Python. Data in the `Python` dictionary are attributed to labels called keys. In our case, the keys are the context identifiers. Any associated data for each context that have been input by the user, such as radiocarbon determinations and stratigraphic relationships, are stored in the dictionary for each key (context label).

Context grouping and radiocarbon determinations can be loaded in any relative order, but only after the file that provides the stratigraphic relationships has been read into `PolyChron`. Users can only include one level of grouping at present, though it is not uncommon for subgrouping to be available in the stratigraphic record and there is the possibility of adding this to `PolyChron` in the future.

Visualisation and manipulation of stratigraphic DAGs

Once the stratigraphic relationship file has been read into `PolyChron`, a stratigraphic DAG is automatically displayed on the GUI, users can then manipulate the stratigraphic DAG using point-and-click functionality on the `PolyChron` GUI. They can visualise and adjust the stratigraphic DAG as required by zooming in on the DAG and moving it around, which allows for easy visualisation of large stratigraphic sequences.

At present, the layout of the stratigraphic DAG (and indeed all DAGs displayed in `PolyChron`) is determined by a layout algorithm in the `NetworkX` package which seeks to minimise the crossing of edges. However, suppose the stratigraphic sequence that has been read into `PolyChron` is, in fact, made up of multiple disjoint stratigraphic sequences (and thus disjoint stratigraphic DAGs). In that case, this can lead to a given context in one stratigraphic sequence being placed higher than contexts in a disjoint stratigraphic sequence such that archaeological evidence suggests those contexts come from a younger group. The user can force `PolyChron` to display the stratigraphic DAG so that any given context must be displayed below all contexts in a group younger than that context. However, forcing the formatting in this way can lead to cluttered placement of edges, so we avoid using that functionality for all examples displayed within this thesis and seek to improve this functionality in the future.

Users can also query or make changes to the stratigraphic DAG by right-clicking on a specific context and choosing from the following functionalities in the drop-down menu:

- adding new contexts or deleting existing contexts in the stratigraphic DAG,

- adding or removing stratigraphic relationships by adding or removing edges in the stratigraphic DAG,
- setting contexts to be equal to each other (if they are once-whole or contemporary),
- and querying or changing supplementary data (as defined in Figure 7.2).

The user must save any changes to the stratigraphic DAG, which can either be saved as a new *model* or by overwriting the existing *model* by using the file menu on the GUI toolbar. When the user makes a change to the stratigraphic DAG, `PolyChron` loads a text box in which the user is required to input the reason for making the changes before they can progress. This data is automatically saved to the user's machine when they save their *model*, thus ensuring any changes to the data originally read into the software are accounted for. An example of this process is provided in the supplementary video that is provided in Appendix B Section B.1.

Converting to the chronological graph

In chapter 6 Figure 6.2, we outlined the process of converting a stratigraphic DAG to a chronological DAG, such that all physical relationships are converted to temporal relationships, any residual/intrusive samples are accounted for, and nodes for group boundary parameters are then added. `PolyChron` follows the same process, requiring additional data input from the user, which we now outline.

When the user chooses to load the chronological DAG, they are prompted with a message querying if any of the samples from the contexts in the stratigraphic DAG are residual or intrusive. If they click *yes*, they are taken to a new page where they can input which contexts are residual and intrusive and whether to include them in the modelling using point-and-click functionality on the stratigraphic DAG to allow for easy and efficient input of information. Once the users have highlighted

all appropriate contexts as residual/intrusive, the GUI requires the user to classify residual or intrusive contexts as either terminus post quem (TPQ) or terminus ante quem (TAQ), respectively or exclude them from the modelling entirely. If they choose TPQ or TAQ, the appropriate edges are removed from the chronological DAG to account for the sample potentially being younger or older than the context for intrusive and residual samples, respectively.

The user is then taken to the part of the GUI where they can input any group relationships; they are also taken straight to this screen if they click *no* at the message prompt regarding residual/intrusive samples. Whereas the group relationships that the user loads at the start are strictly before and after in time, the group relationships GUI allows the user to expand on this by declaring whether the before-after relationship between adjacent groups is abutting, has a gap, or overlaps. To do this, the user is shown a canvas on the GUI with rectangles representing all the groups within the model. The user can then place the rectangles in different positions on the horizontal axis to show any combination of gap, overlap or abutting relationships between the groups within the model (selecting them, of course, based on archaeological evidence). The group relationships are then stored automatically by `PolyChron`, ready to pass to the posterior inference module, and the chronological DAG is loaded on the post-MCMC analysis screen within the GUI.

Changing the model

Just as the user can save changes to the stratigraphic DAG in the form of different *models*, any changes to the chronological DAG prompts the user to save such changes as a new plausible *model*, rather than overwriting the existing one. Changes to the chronological DAG occur because additional prior knowledge is elicited when rendering the chronological graph (in the form of residual/intrusive samples and/or

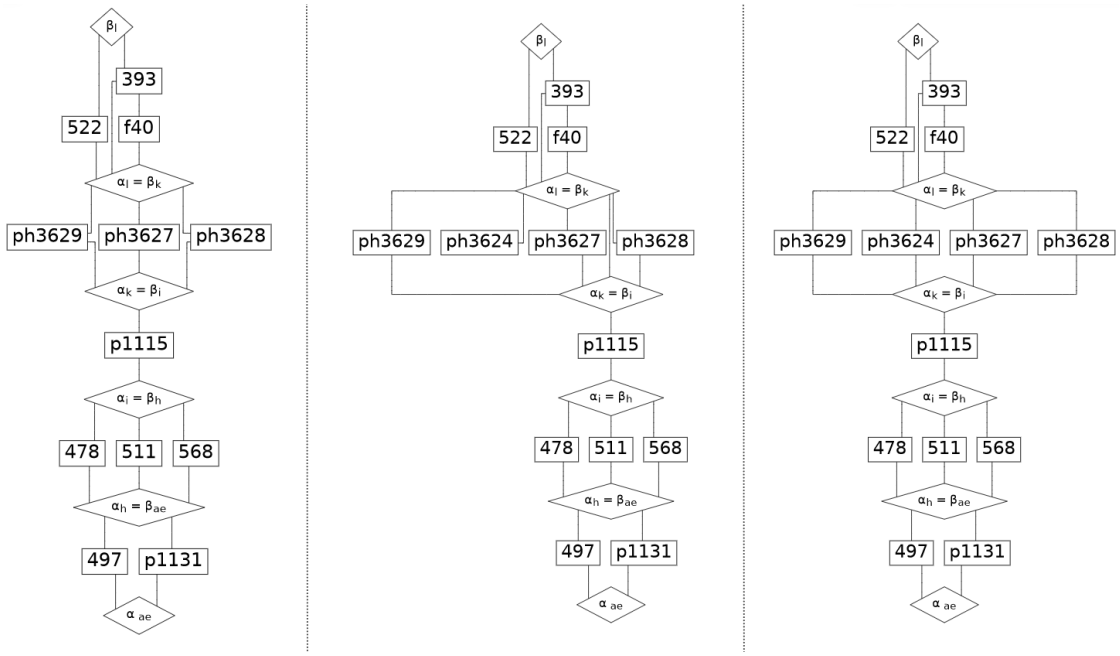


Figure 7.5: Three different chronological graphs from the same stratigraphic DAG, using data from a small stratigraphic sequence at Danebury (Cunliffe and Poole, 1991). *to the left* Context is ph3624 residual and removed, *middle* Context ph3624 is residual but kept in the model and treated at TPQ, (*to the right* No contexts are residual.

changes in the group relationship types); one stratigraphic DAG may lead to multiple chronological models (that are represented by distinct chronological DAGs) that the user may deem plausible as demonstrated in the previous section. Any changes to the chronological DAG, even for the same stratigraphic DAG, should be saved as a new *model* within `PolyChron` since the final chronological model differs. There are two ways the user can do this. The first is to rerender the chronological DAG, which forces the user to either overwrite the existing chronological DAG in the current *model* directory, resulting in only one *model* and the loss of all the data attributed to the original model. The preferred method is to save the changes in a new model directory, thus recording both models that the user considered plausible.

Users are able to produce multiple plausible chronological models rapidly, using the same initial *input files*. For example, we were able to produce the chronological DAGs in Figure 7.5 for a single stratigraphic sequence from the same *input files*. To do this, we simply rerendered the chronological DAG three times, inputting that: context ph3624 is residual and should be treated as TPQ, that ph3624 is residual and should be removed from the modelling, and that ph3624 is not residual. Each time we saved the changes as a new *model* for which `PolyChron` creates a new directory. Producing these three distinct chronological DAGs took around five minutes in `PolyChron`. As they are loaded, `PolyChron` saves all the required data ready to pass to the MCMC algorithm (see Table 7.1 for details), requiring no additional effort from the user to obtain posterior results.

7.3.3 Posterior inference module

Once the user has loaded the chronological DAG for a given *model* and is sure it represents the chronological model they intended to build, they simply click the *Render MCMC* option and `PolyChron` will compute joint posterior estimates for all parameters in their model using our purpose built `Python` module which automates the construction of the input required for the MCMC algorithm (which we introduced in Chapter 6 and provide details of in Appendix A Section A.3). The input to the MCMC algorithm is automatically constructed from the data that has been stored during the process of rendering the stratigraphic and chronological DAGs in the GUI. The input, specifically the ten `Python` variables required for the MCMC algorithm, is defined in Table 7.1. These variables are then input to a single function within the posterior inference module. Thus, for any additional Bayesian chronological models to be implemented in `PolyChron` in future, the code for producing such a model would simply need to take in some collection of these ten variables to be connected

to the GUI.

Users are able to obtain posterior results for a single *model* or all *models* in a project for which a chronological DAG was previously rendered. If they choose the option to render multiple models, `PolyChron` automatically checks which model directories have a chronological DAG rendered (and all the required *input data* for the MCMC algorithm already exists) and asks the user which of these models they wish to run. The software then sequentially obtains results for all models the user selects.

At present, running multiple models is done sequentially. Thus, the time taken to run a model scales linearly with the number of models. There is the possibility of exploring other options for improving the efficiency of this process. However, the time taken to obtain posterior estimates is computational time and not the additional time required from the user and so speeding this process has not been a priority during this PhD project.

The automated construction of the input to the MCMC algorithm is made possible by storing all information for a given archaeological site in a dictionary which we refer to as the *site dictionary*. This dictionary stores all data required for sampling within the MCMC algorithm in a standardised format as outlined in Figure 7.6. Additional prior information can be easily added to the *site dictionary* if required for future development of `PolyChron`. For example, Heaton et al. (2020a) discusses how covariance estimates between calendar years are available as model output for IntCal20. Though this does not affect the calibration of single radiocarbon determinations, such model output may potentially be used when calibrating multiple determinations together as a Bayesian chronological model. The IntCal team has planned work to explore how best to include this information in calibration software, and this information could easily be added to the site dictionary if required.

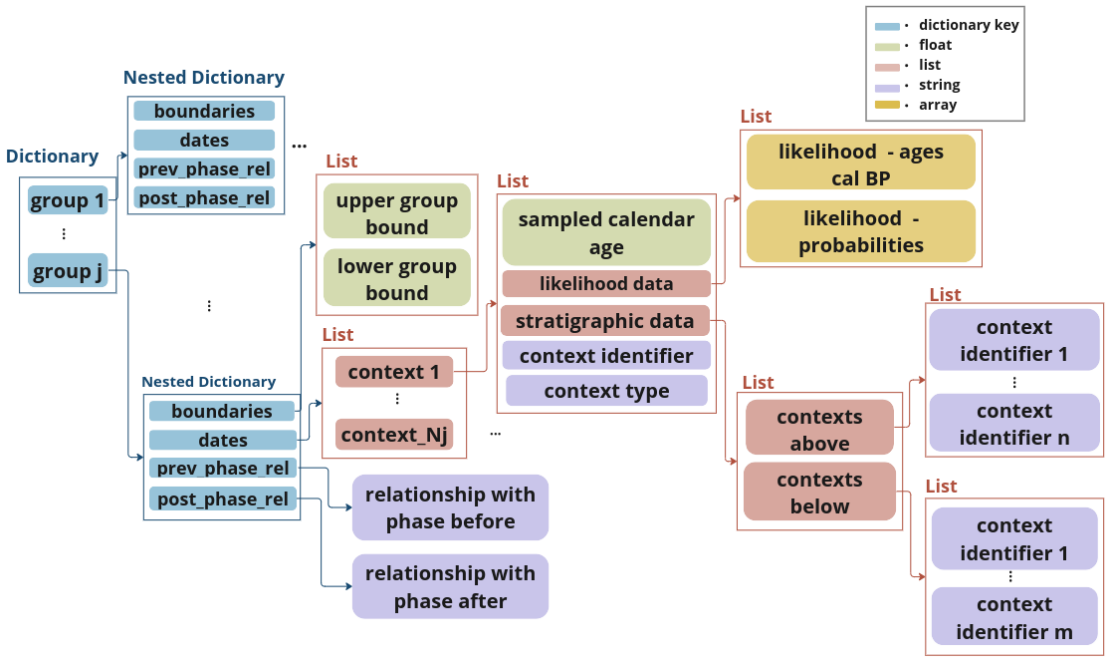


Figure 7.6: Structure of a site dictionary within `PolyChron`. The colours of the boxes correspond to the `Python` data types in which the data are stored. Arrows indicate nested data. For example, the dictionary on the left-hand side has j keys (representing group labels), which correspond to j nested dictionaries containing further data about contexts within each group and that group's relationships with its adjacent groups.

7.3.4 Viewing dating results

Once the posterior inference module has run the MCMC algorithm and reached the minimum number of samples (set to 50,000 by default) and had checked that the sampler has reached convergence using the checks outlined in Appendix A Section A.5, `PolyChron` moves to the next component in the software, post-MCMC analysis. This loads a new window in `PolyChron`, which allows users to use point-and-click functionality to choose which contexts they wish to view posterior results for. This allows for efficient viewing of marginal HPD intervals and posterior probability density plots for any parameters (or collection of parameters) within the chrono-

logical model.

Users can also view HPD intervals and posterior probability density plots for the time elapsed between any two parameters in a chronological model. As previously discussed in Chapter 2 Section 2.2, archaeologists are often more interested in the time elapsed between two archaeological events than they are in the dates of the events themselves. For example, they might want to know the time elapsed between α_j and β_j to estimate the archaeological activity duration for group j.

7.4 `PolyChron`: Case studies

Thus far, we have motivated `PolyChron` and discussed functionality included in the software which allows for the semi-automated construction of chronological models and require very little manual input from the user once the initial *input data* has been loaded into the software. This section provides two case studies of `PolyChron` being used in practice, demonstrating how we have utilised `PolyChron` to further our research.

7.4.1 Rendering larger models in `PolyChron`

A key focus when building `PolyChron` was scalability. We sought to ensure that stratigraphic sequences of all sizes and complexity could be rendered in the software and posterior results produced for their corresponding Bayesian chronological models. Within the software, there were two features that we anticipated might cause issues for scalability. The first was the point-and-click functionality for the stratigraphic DAG, and the second was the conversion of the chronological DAG into the input format for the MCMC algorithm.

Due to the lack of complete data archives available for larger stratigraphic sequences than those utilised thus far, we opted to test the scalability of `PolyChron` using a stratigraphic sequence from Danebury (Cunliffe and Poole, 1991). The sequence we used was the stratigraphic sequence for the 1984-1985 southeast quadrant at Danebury. This stratigraphic sequence was published as an image on microfiche, and so considerable work was required to convert the relative dating evidence within this diagram to the required format of the stratigraphic relationships input file for `PolyChron`, as outlined in Appendix B Section B.2. It consists of 142 contexts which have been grouped into nine groups. Absolute dating evidence was unavailable for most of the contexts in the stratigraphic sequence. Instead, we used simulated radiocarbon determinations (see Appendix B Section B.3 for details) to test how reliably `PolyChron` can covert the data stored in a stratigraphic DAG into a chronological model and subsequent posterior density estimates for the parameters within the model.

Using a larger stratigraphic sequence as *input data* to `PolyChron` than those demonstrated previously did not impact the functionality or speed of the project loading component of the software. Within the prior elicitation component of the software, we observed that the images of the stratigraphic and chronological DAGs render slightly slower (taking around a second rather than a perceived instant result). However, the stratigraphic DAG's zooming, panning and point-and-click functionality were not impacted. See Figure 7.7 for the stratigraphic DAG for the Danebury stratigraphic sequence.

Further, the conversion of the data represented in the chronological DAG into variables required for the MCMC algorithm (see Table 7.1) showed no difference in speed or functionality. Manual inspection of the variables produced for input

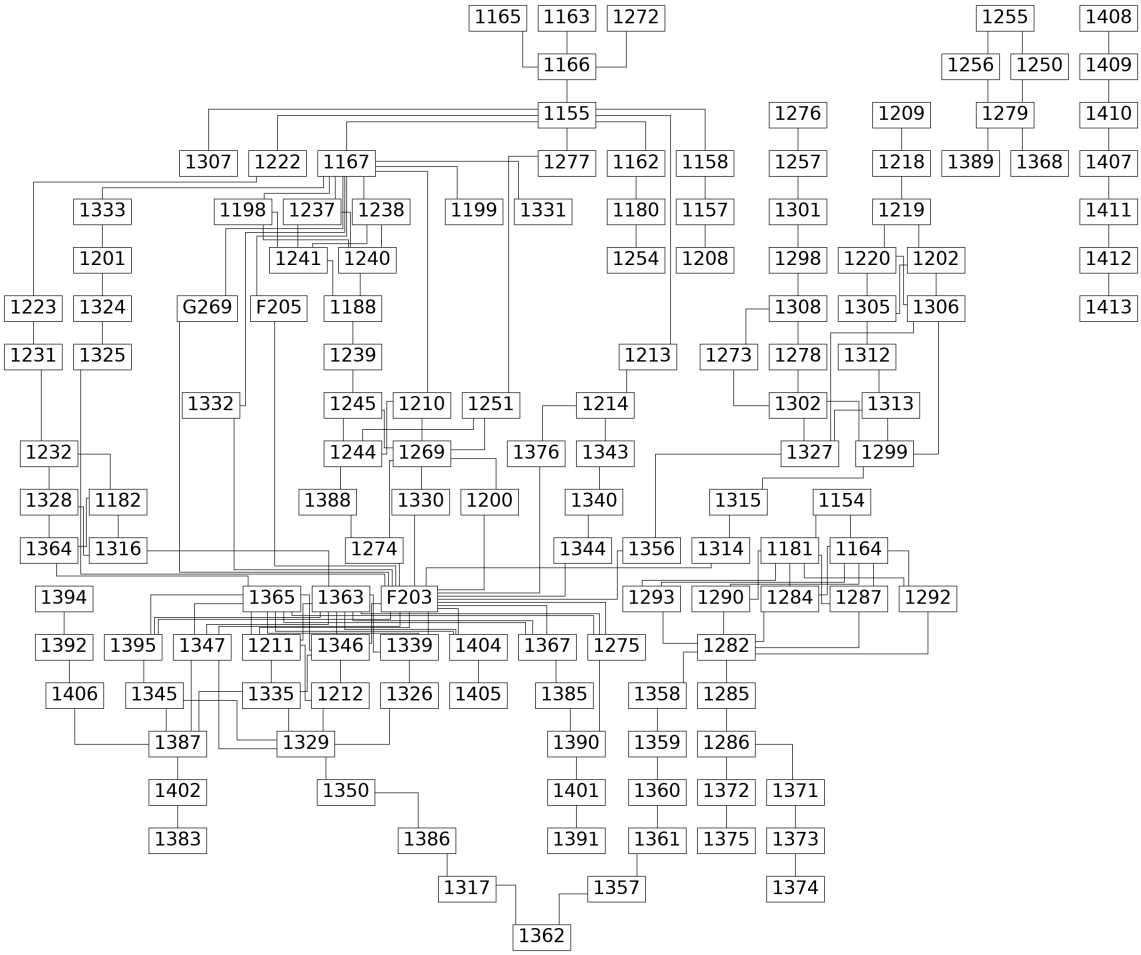


Figure 7.7: Stratigraphic DAG rendered in `PolyChron` for the 1984-1985 South-East quadrant stratigraphic sequence (Cunliffe and Poole, 1991).

into the MCMC algorithm confirmed that all variables had been correctly converted from the chronological DAG. We obtained posterior density estimates for all 152 parameters (10 group boundaries and 142 simulated radiocarbon determinations), which corresponds to obtaining a minimum of 20,000 samples for each parameter for two runs of the MCMC algorithms before running the appropriate convergence checks (see Appendix A Section A.5). Running the MCMC algorithm for this model took around 17 minutes on the author's laptop. Given the size of the strati-

graphic sequence, we felt this time was acceptable for proof-of-concept software and that it demonstrated `PolyChron`'s ability to handle larger stratigraphic sequences. In addition, it confirmed the posterior inference module's ability to reliably and automatically covert chronological DAGs to model output, even for complex models with many parameters.

7.4.2 Multiple plausible chronological diagrams

Initial development work on `PolyChron` confirmed that DAG could be used to semi-automate chronology construction. Following this, we sought to investigate if existing algorithms for analysing mathematical graphs could provide us with information about the importance of a given context (and its associated sample) within the chronological model.

Among the wealth of algorithms that exist for analysing mathematical graphs, a subset of algorithms highlight 'nodes of influence' in a mathematical graph i.e. nodes that have the highest influence on the other nodes in the graph. A crude measure of this might be the degree of a node, which is a measure of many nodes any given node is connected to via edges (see Zhan et al., 2017). This would correspond to the number of parameters in a chronological model that a given parameter is directly connected to in the chronological DAG for that model. However, more complex measures also exist, such as eigenvector centrality (see Zaki and Meira, 2014, Chapter 4) and Katz centrality measures (Katz, 1953). These measures take into account not only the number of nodes a given node is connected to but also how influential the nodes are that a given node is connected to. Katz centrality in particular is recommended for directed graphs (see Zhan et al., 2017).

As part of the research within this PhD, we sought to investigate if such algorithms

could indicate which contexts have the most influence within a chronological DAG for a given model and the subsequent impact on the posterior results of that Bayesian chronological model. We wanted to investigate if this could highlight the most appropriate organic samples to be sent for radiocarbon dating if there are more samples than the budget allows dating for. Existing research into sample selection for choosing a subset of optimal samples, not based on the quality of the samples, (Christen, 1994, see) utilises computationally intensive simulation methods. Thus, we wanted to explore if graph theory algorithms could achieve this without requiring simulation methods.

For this case study, we again used the Danebury stratigraphic sequence (presented in Figure 7.7) to maximise the data points we had, to allow us to investigate if there was a relationship between node importance and change to posterior results. To test this theory, a node's influence on the original stratigraphic DAG was calculated using Katz centrality (see Appendix B Section B.4 for details). Following the initial work to produce the *input data* as discussed in Section 7.4.1, the author wrote additional code which, for each node parameter in the chronological graph automatically rendered a new chronological DAG with that specific parameter node removed but with all other temporal relationships between parameters preserved and saved this to a new *project* directory on the user's machine. The results of the chronological model were then automatically passed to the posterior inference module in `PolyChron` for each new chronological DAG, which for the Danebury site produced 142 plausible chronological models corresponding to 142 contexts that could be removed.

Once the posterior results had been obtained for each model, we then measured the similarity between marginal posteriors of each parameter in a given model when compared to the original chronological model with no parameters removed and then

found the average of this measure for the whole chronological model. Thus giving a measure of how different, on average, the posterior results were when each parameter was removed from the model.

Measuring the similarity between probability distributions is not trivial, and there are a wealth of metrics that seek to quantify such similarity. To ensure our results were not influenced by the specific metrics we used, we considered four potential metrics: Kullback–Leibler divergence (see Joyce, 2011), Bhattacharyya distance (Bhattacharyya, 1943), Hellinger distance (Enclyclopedia of Mathematics, 2020) and an overlapping metric proposed by Pastore and Calcagnì (2019). These metrics are formally defined in Appendix B B.5 and are some variation of a calculation summing a measurement of the difference between two probability distributions for each plausible value for a given parameter.

There was no evidence of a relationship between the influence of a node within the stratigraphic DAG and a change in posterior results of the model when the context corresponding to that node is no longer included in the model. However, the key point is that 142 plausible chronological models were able to be rendered by `PolyChron`, allowing us to test the aforementioned theory pertaining to node importance without any additional time being required from the user once the initial chronological model had been rendered. These 142 plausible chronological models were rendered over the course of three days while the author continued with other tasks.

Despite obtaining negative results, very little time had to be invested by the author to test this theory. Therefore, we were able to investigate other avenues of research for analysing multiple plausible chronological graphs as discussed in Chapter 8. The automatic rendering of posterior results for Bayesian chronological models from the

chronological DAG exemplified that future research questions can be explored simply by the user producing the appropriate stratigraphic and chronological DAGs within `PolyChron`, or by adding additional functionality that automates the production of a large number of plausible chronological DAGs, as we did to test the aforementioned hypothesis.

7.5 Limitations of `PolyChron`

Though `PolyChron` is now a fully functioning prototype, several areas for improvement should be addressed in future work before the wider community utilises it. Currently, the main limitation is that the GUI is written in Python, specifically in `Tkinter`, which means the user is required to download `Python` all appropriate modules and only then, `PolyChron` can be used on the user's machine. However, due to the modular nature of the software, it is possible to convert the GUI into a different format, say HTML and keep the posterior inference module written in Python. A benefit of this is that it would allow users to load `PolyChron` in their web browser without needing to download `Python` and all the required packages to their machine, which could prove difficult for users with little to no programming experience.

Further, the current method of running the MCMC algorithm sequentially for all the models the user wishes to evaluate could be improved. Other options could be explored, such as parallel programming. For the author's use, parallel programming would have been a suitable solution since their device had multiple cores. However, since this is prototype software and may yet migrate to web-based software as opposed to desktop software, we chose not to explore this route at present. Since if `PolyChron` did remain as desktop software, potential users may not have devices with multiple cores. Thus, a more elegant solution may be required.

At present, PolyChron has limited functionality compared to other software such as OxCal, BCal or Stratify with regards to the type of absolute and relative dating evidence that can be managed within the software. However, the goal when building PolyChron was not to replicate all functionality for chronological data management software and Bayesian chronological modelling software. Instead, we wanted to show that prototype software can be written to manage the whole process. However, additional functionality will certainly benefit users of PolyChron, and we now move to discuss what functionality is missing when compared to existing software and which functionality could be included in future.

7.6 Future development of PolyChron

The planned development of the functionality of PolyChron is broadly split into two categories: managing relative dating evidence and the Bayesian modelling capability of the software. At present, additional site data, such as plans and section drawings, photographs etc. (which can be included in software such as Stratify), cannot be included in PolyChron. Consequently, data archived from PolyChron at present can not provide a complete site archive. Therefore, future development should ensure that all data pertaining to a given stratigraphic sequence can be elegantly managed and subsequently archived from PolyChron.

The development of the Bayesian modelling capability of the software will seek to incorporate much of the functionality seen in OxCal and BCal. At present, the only absolute dating evidence that can be included is radiocarbon determinations. Future development would seek to allow for other absolute dating, such as luminescence dates and dendrochronology dates and dating evidence derived from cultural finds experts (e.g ceramicists). Though we must note that further work to formalise

the prior elicitation of dating evidence from finds experts is required before such information can be included in automated chronology construction.

In addition, we would like to add functionality to the GUI to allow users to include subgrouping in their model if desired. At present, this is possible to do in `BCal` using their 'floating parameters' option or in `OcxCal` by defining a 'boundary()' event in a CQL script. `OxCal` also allows for a variety of potential probability distributions for modelling a priori knowledge about the deposition of datable material within a given group, conditional on the group boundary parameters. Within `PolyChron`, we only consider the probability function in Chapter 6 Equation 6.7 to represent such prior knowledge. Other potential distributions could be added to the posterior inference module in the future, such as those outlined in Table Table 1 Bronk Ramsey (2008). Were we to do so, to ensure adequate archiving, we would require the user to provide the rationale for such modelling choices, which would then be saved within the *model* directory as supplementary information.

Further development of `PolyChron` would not seek to exactly replicate either `BCal` or `OxCal`. At present, `PolyChron` provides more flexibility for the user to visualise and alter their model than `OxCal` or `BCal` due to the point-and-click functionality and use of DAGs to present information, but it is currently missing some of the modelling options that `OxCal` and `BCal` provide. Any of the extra functionality added to `PolyChron` in the future must be done in such a way as to avoid users producing statically implausible models or allowing users to make modelling choices without understanding the implications of the overall model. For example, adding too many subgroups that are only useful for visualisation purposes (for which calibrated date estimates are not required) can lead to overparametisation of the model and result in inefficient computation of posterior distributions.

7.7 Summary of Chapter

Within this chapter, we have motivated a need for and described the functionality of our prototype software, `PolyChron`. This software demonstrates that directed acyclic graphs can be used in practice to visualise and manage relative dating evidence and to structure data resulting from such evidence in a format suitable for input to Bayesian chronological modelling algorithms. As demonstrated in our case studies and in the supplementary video, `PolyChron` can be used to semi-automate the construction of a variety of Bayesian models, including those with complex stratigraphic sequences or those with residual/intrusive samples. Further, it can reliably produce multiple plausible Bayesian chronological models, allowing users to document any models they deem plausible rather than proceeding with a single chronological model simply due to time or software constraints.

Variable	Label	Description
calibration curve estimate	calibration	data frame containing data from the IntCal20 calibration curve. Contains three columns corresponding to: calendar age, radiocarbon age and radiocarbon error.
topological sort	topo_sort	list of contexts in the chronological graph order by a topological sort [a] on the chronological graph.
stratigraphic relationships	strat_vec	A list of tuple. Each tuple contains the contexts above and below each context such that the contexts are in the same order as in topo_sort
radiocarbon determination estimate	rcd_est	list of radiocarbon determinations (measurement only no error), the order corresponds to contexts in topo_sort
radiocarbon determination error	rcd_err	list of radiocarbon errors corresponding to measurements in rcd_est, order corresponds to contexts in topo_sort
context groups	key_ref	list of group allocation ordered for each context in topo_sort
context type	context_type	list of context type (normal, residual or intrusive)
groups	phi_ref	list of all the groups in the model from oldest to youngest
previous group relationships	prev_phase	list of group relationship types between a group and the the group directly older than it, for each group in phi_ref, in the same order.
subsequent group relationships	post_phase	list of group relationship types between a group and the group directly older than it, for each group in phi_ref in the same order.

Table 7.1: Input `Python` variables required to be passed from the prior elicitation component to the posterior inference module within `PolyChron`

[a] A topological sort as defined by `NetworkX` (Hagberg et al., 2008) is "a nonunique permutation of the nodes of a directed graph such that an edge from node 1 to node 2 implies that node 1 appears before node 2 in the topological sort order".

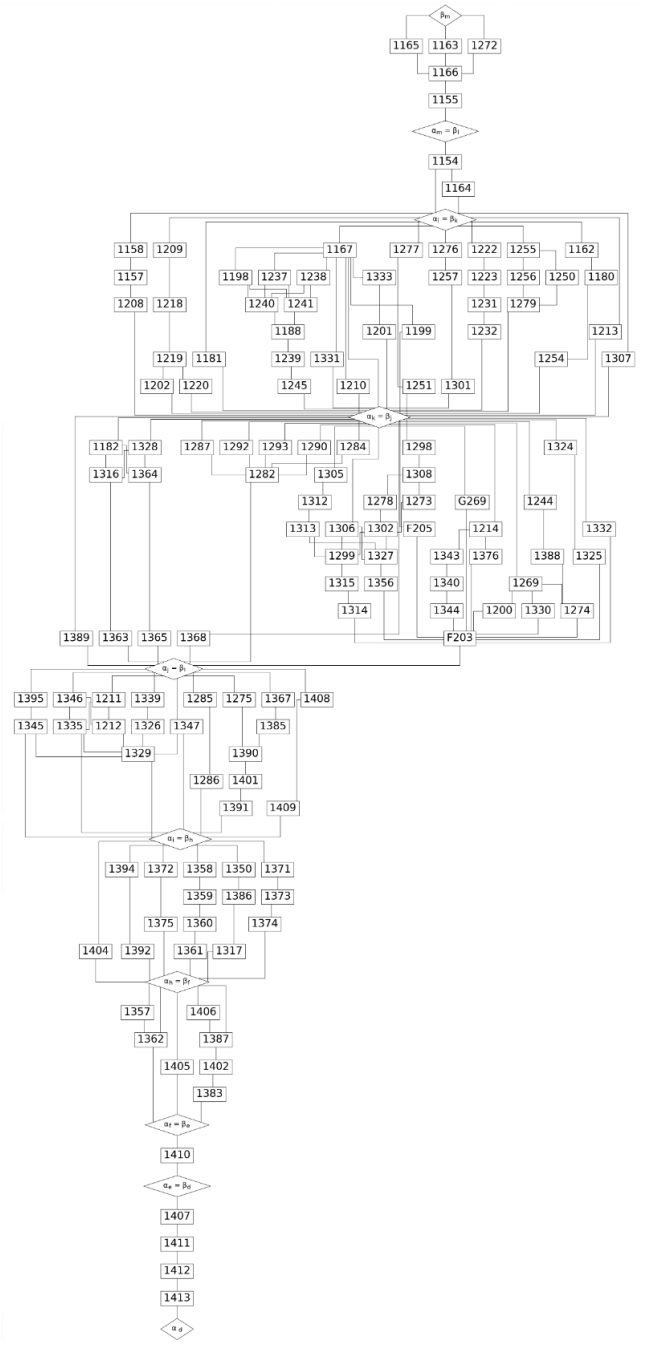


Figure 7.8: Chronological DAG rendered in `PolyChron` from the stratigraphic DAG in Figure 7.8 for the 1984-1985 South-East quadrant stratigraphic sequence (Cunliffe and Poole, 1991), such that all group relationships are assumed to be abutting.

Chapter 8

Additional analysis of chronological models

In the previous chapter and the supplementary video demonstration provided with the thesis, we demonstrated PolyChron's ability to rapidly produce multiple plausible chronological models, even for large models. The discussion now turns to obtaining posterior estimates for parameters of interest when considering multiple plausible chronological models, in such a way that the results make statistical and archaeological sense.

At present, unlike the established theory underlying the process of building Bayesian chronological models discussed in Chapter 5, there is not a single standard methodology for analysing multiple chronological models. Nonetheless, others have considered it. For example, it is discussed at length in Buck and Millard (2004), with Chapter 5 (Sahu, 2004) in particular focusing on model choice (albeit for a type of archaeological chronology construction not examined in this thesis). Broadly speaking, previous analysis of multiple models in archaeological chronology construction fit into three categories: model selection/comparison, model averaging and sensitivity

analysis.

Model selection/comparison methods use varying statistical methodologies to provide metrics to allow users to compare competing statistical models. Model selection is used in both frequentist and Bayesian statistics. However, the metrics used differ. As described by Wasserman (2000), model selection methods in Bayesian inference seek to use the data to select a single model from a set of plausible models. A commonly used metric for model selection is the Bayes factor (Kass and Raftery, 1995), which is the ratio of the posterior to prior odds ratios (see Bernardo and Smith, 1994, Defintion 6.1). However, for improper priors (i.e. priors that are statistical functions but not probability distribution functions and thus do not integrate to 1), Bayes factors are not well defined. Further, for complex Bayesian models, Bayes factors are often hard to compute since they require integrating over high dimensions (though increasingly MCMC methods are used to approximate Bayes factors, just as they are used to approximate posterior distributions). Both the aforementioned situations that complicate the calculation of Bayes factors arise commonly for Bayesian chronological models. OxCal provides a model comparison metric in the form of their agreement indices, which they refer to as a type of pseudo-Bayes factor, though Karlsberg (2006) argues they are a ratio of posterior predictive densities [1] (see Karlsberg, 2006, page 57 for a formal proof of this).

However, a more pressing issue is that Bayes factors, the OxCal agreement indices, and indeed other model selection metrics we might have considered within this research all assume that a true model exists for any given problem and that the true model is contained within a set of specified models. As occasionally discussed in the literature, however, this assumption does not commonly hold for sets of chrono-

[1] A posterior predictive density is a distribution of future observations, dependent upon the data we have already observed (Gelman et al., 2004, Chapter 2)

logical models. For example, Bayliss and Bronk Ramsey (2004) states that there is no 'right' model for any given chronological site. The current author is also not willing to assume that a given set of Bayesian chronological models contains a single true model for several reasons. First, one can never fully excavate a site. Therefore a chronological model can only ever represent a partial representation of the true stratigraphic sequence of a site. Further, given the wealth of factors that can affect the reliability of both the prior and the likelihood, such as residuality, intrusiveness or contaminated samples, even if one were to excavate a site fully, we deem it highly unlikely that one would obtain the true model.

The assumption that the true model is within a given set of models is also made for model averaging methods. Given this, no further discussion of model averaging is provided here, but we refer the reader to Hoeting et al. (1999) for further details. Sensitivity analysis is not a method of model comparison but is instead used to assess how sensitive a given posterior inference might be to changes in the prior or likelihood (see Bayliss and Marshall, 2022, Chapter 2 for a discussion with respect to chronology construction). As outlined in Chapter 6, Sections 6.4.1 and 6.4.2, this is an important step in the Bayesian modelling process and can be carried out in `PolyChron`. However, sensitivity analysis only informs us about how robust a model's posterior results are to changes in the parameters or structure of our prior or likelihood and does not provide information about the reliability of the posterior results themselves.

In this chapter, we explore a rather different approach to analysing multiple plausible chronological models and present the results of a simple application before discussing more complex applications of the methodology, which we seek to explore in the future.

8.1 Quantifying uncertainty in archaeological dating evidence

Thus far, we have expressed our subjective prior chronological knowledge as probabilistic statements, primarily in the form of indicator functions. For example, the indicator functions $\mathbb{I}_{\mathcal{G}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, $\mathbb{I}_{C}(\boldsymbol{\theta})$ and $\mathbb{I}_{\alpha_j,\beta_j}(\theta_{i,j})$ defined in Chapter 6 Section 6.1.3 ensure that our posterior results satisfy the ordering imposed by stratigraphic and grouping information from the excavation.

Proceeding with a single model based on one such set of indicator functions omits the uncertainty pertaining to the prior knowledge derived from the relative dating evidence. As discussed in Bayliss and Bronk Ramsey (2004) Section 2.3.3, relative chronological relationships represented by the indicator functions are definitive. For example, a context either has a stratigraphic relationship with another or does not. However, as previously discussed, such definitive statements rarely hold up to much scrutiny.

In most other applications of Bayesian inference, we account for uncertainty in prior knowledge using a carefully chosen probability distribution. However, as Bayliss and Bronk Ramsey (2004) section 2.3.4 remarks, it is not clear that an archaeologist (even those who are highly numerate) could provide adequate prior knowledge to express uncertainty about stratigraphic and group relationships as a probability distribution. Indeed, this problem was encountered in the early research on Bayesian chronological modelling, where a seemingly suitable choice of a uniform distribution to represent a uniform prior knowledge about group boundary parameters led to unintentionally informative prior knowledge being implied with regard to the time elapsed between group boundaries (see Chapter 6 Section 6.1.2).

Williamson and Goldstein (2015) argue that in situations like ours, where we have several plausible models with quite different structures, Bayes linear methods (Goldstein and Wooff, 2007) can be used to quantify our posterior uncertainty more coherently that the methods discussed in the previous section. Bayes linear methods use summary statistics such as expectation, variance and covariance to quantify prior beliefs about quantities for which defining a full probability distribution (or probabilistic functions) is not feasible or is impractical. Discussion and motivation of the theory behind Bayes linear methods are beyond the scope of this thesis, but we refer the reader to Goldstein and Wooff (2007) for a comprehensive account. Here we only discuss the theory directly relevant to our own application.

We must first introduce the concept of parameters of interest. Within a Bayesian model, we may have multiple parameters that are used to obtain the most informative posterior estimates possible. Often, however, it is only a subset of these parameters that we are, in fact, interested in. For example, within Bayesian chronological modelling, $\boldsymbol{\theta}$ parameters are incredibly important for obtaining the most informative posterior estimates of the group boundary parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, since it is the $\boldsymbol{\theta}$ parameters that have associated absolute dating evidence. However, in practice, it is not uncommon for the posterior estimates of the group boundaries to be the parameters that the archaeologists are ultimately interested in.

An additional concept we must motivate is that of judgements, as defined in Williamson and Goldstein (2015). Suppose we have a set of M plausible chronological models, for which we have carried out a full Bayesian analysis (as detailed in Chapter 6) and obtained marginal posterior estimates for each of the parameters in our M models, then V_m is the set of judgements for model m that encompasses all the prior knowledge, modelling decisions and computational methods that result in model

m. Typically, in chronological modelling, we proceed with a single model and thus only consider one set of judgements, V_0, to be correct. However, Williamson and Goldstein (2015) argues that for complex models (such as those studies considered in this thesis), it is unlikely that those carrying out the analysis truly believe that the V_0 is the only set of judgements that can adequately represent our scientific insights and beliefs about a given problem. Given this, they propose a method that they call posterior belief assessment (PBA).

8.1.1 Overview of posterior belief assessment

Posterior belief assessment, proposed by (Williamson and Goldstein, 2015), involves expressing prior beliefs in two stages. First, we make probabilistic statements that comprise the set of judgements, V_m, for model m, which correspond (in chronological modelling) to the prior beliefs discussed in Chapters 5 and 6 that are used to form the prior $p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{x})$. These prior beliefs are expressed for all model parameters and used to carry out the Bayesian inference seen in Chapter 6. The second step in PBA is to represent prior knowledge about the parameter(s) of interest; these are the parameters for which we are most interested in posterior estimates for. The prior knowledge in this step should represent our true beliefs with respect to our parameter(s) of interest and do not require the specification of full probability distributions, only expectations and variances.

When carrying out Bayesian inference, our judgements must be sufficiently representative of the situation one is trying to model. However, it is not unusual for certain modelling decisions to be made, not because the model best represents our prior beliefs but because it results in computationally efficient posteriors for which to calculate estimates. Williamson and Goldstein (2015) argues that Bayes linear methods, which focus only on summary statistics such as (expectation and variance) are

more suited to quantifying our prior beliefs about parameters within more complex models and do not require such simplifications of our prior beliefs.

We now define PBA for a single arbitrary parameter of interest, a group boundary parameter α_j, for which we have M Bayesian chronological models we consider plausible. Given α_j, let $\mathcal{G}$ be the vector of estimated marginal posterior expectations of α_j for each of the M models we consider plausible,

$$\mathcal{G} = (\mathrm{E}\left[\alpha_j|\mathbf{x}; V_1\right], \mathrm{E}\left[\alpha_j|\mathbf{x}; V_1\right], ..., \mathrm{E}\left[\alpha_j|\mathbf{x}; V_M\right]), \tag{8.1}$$

then Williamson and Goldstein (2015) defines the posterior belief assessment for α as:

$$\mathrm{E}_{\mathcal{G}}\left[\alpha_j\right] = \mathrm{E}\left[\alpha_j\right] + \mathrm{Cov}\left[\alpha_j, \mathcal{G}\right] \mathrm{Var}\left[\mathcal{G}\right]^{-1} (\mathcal{G} - \mathrm{E}\left[\mathcal{G}\right]). \tag{8.2}$$

Here, the quantity $\mathrm{E}_{\mathcal{G}}\left[\alpha_j\right]$ (referred to forthright as our adjusted beliefs) is at least as close to α_j (according to our prior beliefs and the Bayesian inference carried out to obtain the set $\mathcal{G}$) as the posterior expectation $\mathrm{E}\left[\alpha_j|\mathbf{x}; V_i\right]$ for i in $1, ..., m$. For proof of this, please see Williamson and Goldstein (2015) Theorem 1. An important distinction here is that posterior belief assessment provides a framework to allow statisticians to assess a given expert's uncertainty in the Bayesian inference they have carried out, and the above result strictly holds conditional upon the expert's beliefs.

PBA for Bayesian chronological models

An important assumption of PBA is that it does not require that the set of M Bayesian chronological models that we consider contains the true model. Instead, it only requires that we believe each model m is plausible for the situation we are trying to represent. Furthermore, within the posterior belief assessment framework,

it is possible to consider an infinite number of plausible models. While we do not consider large numbers of plausible models within the initial investigation presented within this chapter, this could readily occur in archaeology, and so should be the subject of future work.

To carry out posterior belief analysis on a given parameter or set of parameters within a Bayesian chronological model, we focus on their posterior expectations and variances of posterior estimates, as defined in Equation 8.2. For parameters with multimodal distributions such as those relating to the likelihood component of the posterior distribution for θ parameters in Chapters 5 and 6, expectations are not informative summary statistics with respect to the true value of the parameter, as discussed in Chapter 3 Section 3.2.3. Group boundary parameters, by contrast, typically have unimodal likelihoods, so (despite their hierarchical relationships with the $\boldsymbol{\theta}$ parameters) their posterior distributions are commonly unimodal. Thus, we chose to proceed with investigating the use of PBA for Bayesian chronological models but restricted such research to carrying out PBA on group boundary parameters. Indeed, much of the published literature on Bayesian chronology construction includes a focus on the posterior estimates of the calibrated ages of group boundary parameters (see Bayliss et al., 2015; Krus et al., 2015, for examples).

8.2 Case study - Posterior belief assessment in archaeology

To investigate the use of PBA on parameters within a chronological model, we utilised part of the relative and absolute dating evidence obtained during excavations at Çatalhöyük, in particular, the East Mound (Bayliss et al., 2015). Since 1961,

various periods of excavations have been carried out at Çatalhöyük. As discussed in Bayliss et al. (2015), excavations at the East Mound focused on estimating the start date of the occupation of Çatalhöyük, to examine the temporal relationship between the settlement at Çatalhöyük and the occupation of a nearby settlement Boncuklu Höyük. Further, Bayliss et al. (2015) wanted to investigate the early use of pottery within the Middle East for which a start date for the settlement at Çatalhöyük is informative. Given the focus on a specific parameter within a larger chronological model, namely the oldest phase boundary within the model, we felt this was a suitable case study to examine the use of PBA for a single parameter.

East Mound

The DAG in Figure 8.1 is derived from the stratigraphic sequence provided in Bayliss et al. (2015) Figure 3. Within the stratigraphic DAG, there are 31 contexts divided into 7 groups. Bayliss et al. (2015) discuss in detail which dates correspond to samples that do not necessarily date the contexts in which they were found. For demonstrative purposes (since the number of plausible models increases the more of such contexts we consider) we focused on two of these contexts:

- 5328, which contained a sample that was considered likely to be a statistical outlier due to a laboratory error,
- and 4826, which contained a sample that was considered to be residual potentially.

Within this case study, we consider multiple plausible chronological models that arise from the fact that the samples from contexts 5328 and 4826 do not necessarily date the contexts in which they were found. For the potentially residual sample found in context 4826, there are three courses of action: assume it dates the context, assume

it is residual and remove it from the chronological models, or assume it is residual and treat it as terminus post quem (TPQ). For each of those three courses of action, there are two options for the sample from context 5328: assume the sample is not a statistical outlier or assume that it is and remove it from the modelling (though for such outliers, we recognise that removing the sample from the modelling process is a rather drastic course of action and not ideal). This resulted in six plausible chronological models that are illustrated in Figure 8.3 and outlined in Table 8.1. The very oldest phase boundary parameter in all six models is α_{X_11D}, which is the parameter we focus on for the PBA in the following section. For neatness of notation, we refer to α_{X_11D} simply as α_X, for the remainder of this thesis.

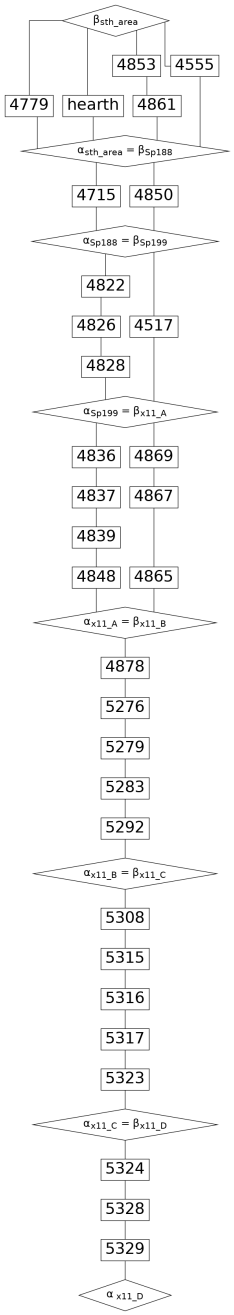


Figure 8.1: Chronological DAG constructed using part of the relative and absolute dating evidence obtained during excavations at the East Mound, Çatalhöyük (Bayliss et al., 2015). Note that the node at the bottom of the DAG is the parameter for which we are carrying out PBA on.

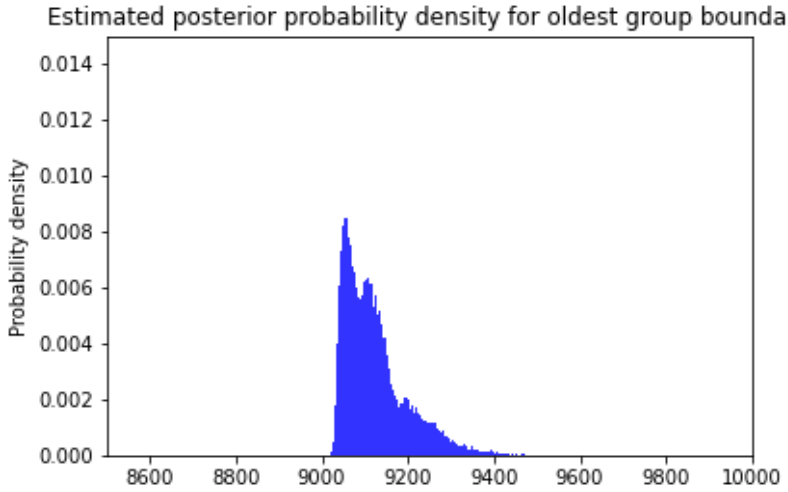


Figure 8.2: Estimated posterior density of α_X under the model illustrated in Figure 8.1. Note that though the distribution is not strictly unimodal, it does not exhibit multiple distinct modes as is observed for marginal posterior estimates of parameters with likelihoods derived directly from radiocarbon determination. As such, we deemed the marginal posterior expectation to be an adequate numerical summary to allow us to proceed with PBA of α_X.

8.2.1 Prior belief specification for oldest boundary parameter

In order to carry out PBA for α_X and thus obtain our posterior belief assessment $\mathrm{E}_{\mathcal{G}}[\alpha_X]$, as defined in Equation 8.2, we must evaluate the following quantities: $\mathrm{E}[\mathcal{G}]$, $\mathrm{Var}[\mathcal{G}]$ and $\mathrm{Cov}[y, \mathcal{G}]$. Further, we must obtain the set of posterior expectations

$$\mathcal{G} = (\mathrm{E}\left[\alpha_X|\mathbf{x}; V_1\right], \mathrm{E}\left[\alpha_X|\mathbf{x}; V_2\right], ..., \mathrm{E}\left[\alpha_X|\mathbf{x}; V_6\right]),$$

that result from carrying out Bayesian inference for our six plausible models.

In what follows, we present a hypothetical elicitation of $\mathrm{E}[\mathcal{G}]$, $\mathrm{Var}[\mathcal{G}]$ and $\mathrm{Cov}[y, \mathcal{G}]$, as proof of concepts, but in future seek to elicit them from experts.

Model	Sample from context 4826	Sample from context 5328
1	excluded	excluded
2	excluded	included
3	TPQ	excluded
4	TPQ	included
5	included	excluded
6	included	included

Table 8.1: Summary of the combinations of how prior knowledge pertaining to contexts 4826 and 5328 are included in the chronological model.

Elicitation of key prior quantities

To elicit $\mathrm{E}[\mathcal{G}]$, $\mathrm{Var}[\mathcal{G}]$ and $\mathrm{Cov}[\alpha_X, \mathcal{G}]$ requires two further quantities $\mathrm{E}[\alpha_X]$ and $\mathrm{Var}[\alpha_X]$. and we obtain these values using a hypothetical process of elicitation which is referred to as prior belief specification. Thus, to carry out PBA for α_X requires elicitation of the following key quantities:

- $\mathrm{E}[\alpha_X]$: the prior expected value of α_X according to the experts beliefs,
- $\mathrm{Var}[\alpha_X]$: the measure of spread about $\mathrm{E}[\alpha_X]$ according to the experts beliefs,
- $\mathrm{E}[\mathcal{G}]$: the prior expected value of the posterior expectations of α_X according to the expert's beliefs, prior to Bayesian inference being carried out,
- $\mathrm{Var}[\mathcal{G}]$: the prior measure of spread about $\mathrm{E}[\mathcal{G}]$ according to the expert's beliefs,
- $\mathrm{Cov}[\alpha_X, \mathcal{G}]$: the prior measure of the relationship between the expert's prior knowledge of α_X and their prior knowledge pertaining to the set of posterior expectations $\mathcal{G}$ which results from the Bayesian inference the expert has carried out.

Elicitation of $\mathrm{E}[\alpha_X]$ and $\mathrm{Var}[\alpha_X]$ from experts would require formal protocols to be

established and training provided so that knowledge could be reliably elicited. This was not feasible in the timeframe of this PhD project and so for proof of concept, we chose a value for $\mathrm{E}[\alpha_X]$ that was equal to the estimate of α_X in the literature prior to the additional dating being carried out by Bayliss and Bronk Ramsey (2004), which was around 9250 calibrated years as given in Cessford et al. (2005). Following this, we set the value of $\mathrm{Var}[\alpha_X]$ to be equal to the square of the maximum error of the radiocarbon determinations used within the analysis which gave $\mathrm{Var}[\alpha_X] = 9604$ calibrated years. We acknowledge that this seems a somewhat large and arbitrary value for the variance. However, we do seek to elicit such values within future research and examine the use of PBA for differing values of $\mathrm{Var}[\alpha_X]$ in Section 8.3.

Since we have already defined $\mathrm{E}[\alpha_X]$, and we do not believe that the M plausible models within set $\mathcal{G}$ are biased, i.e. we expect the posterior expectations within G to on average, accurately predict α_X, we set $\mathrm{E}[\mathcal{G}] = \mathrm{E}[\alpha_X] = 9250$ cal BP. When determining a suitable value of $\mathrm{Var}[\mathcal{G}]$ for testing purposes, we set $\mathrm{Var}[\mathcal{G}]$ to be greater than $\mathrm{Var}[\alpha_X]$, since we felt there was additional uncertainty due to the data (our set of radiocarbon determinations) that we needed to account for. Such uncertainty can arise for a number of reasons, such as contamination of samples used for radiocarbon dating or uncertainty in our prior knowledge, as previously discussed. Thus we set $\mathrm{Var}[\mathcal{G}] = 10000$ cal BP. We would like to remind the reader that when determining the values of these key quantities, we were simply seeking to obtain feasible values for such quantities that would allow us to test the use of PBA on archaeological chronological models. However, future research would need to ensure that these values are formally elicited from experts.

Finally, we needed to determine a value for $\mathrm{Cov}[\alpha_X, \mathcal{G}]$. It is difficult for anyone, even those who are highly numerate, to reliably quantify covariance between two

variables. Thus, we instead (upon the advice of Prof Daniel Williamson) chose to exploit the following relationship:

$$\text{Var}[\alpha_X - \mathcal{G}] = \text{Var}[\alpha_X] + \text{Var}[\mathcal{G}] - 2\text{Cov}[\alpha_X, \mathcal{G}], \tag{8.3}$$

since it is considerably easier to quantify the variance of the difference between two variables than to quantify the covariance between two variables. Furthermore, we have already obtained values for $\text{Var}[\alpha_X]$ and $\text{Var}[\mathcal{G}]$.

However, we must ensure that the value we (hypothetically) elicit for $\text{Cov}[\alpha_X - \mathcal{G}]$ does not result in a value of $\text{Cov}[\alpha_X, \mathcal{G}]$ that does not reliably represent our prior beliefs. By the definition of covariance,

$$\text{Cov}[\alpha_X, \mathcal{G}] = \text{Cor}(\alpha_X, \mathcal{G})\sqrt{\text{Var}[\alpha_X]\text{Var}[\mathcal{G}]}.$$

We expect $\text{Cor}(\alpha_X, \mathcal{G})$, the correlation between our beliefs regarding the true value of α_X and the output of the Bayesian inference provided in $\mathcal{G}$, to be positive, and so we obtain the following inequality:

$$0 \leq \text{Cov}[\alpha_X, \mathcal{G}] \leq \sqrt{\text{Var}[\alpha_X]\text{Var}[\mathcal{G}]}. \tag{8.4}$$

Substituting the information held in Equation 8.4 into Equation 8.3 gives,

$$\text{Var}[\alpha_X] + \text{Var}[\mathcal{G}] - 2\sqrt{\text{Var}[\alpha_X]\text{Var}[\mathcal{G}]} \leq \text{Var}[\alpha_X - \mathcal{G}] \leq \text{Var}[\alpha_X] + \text{Var}[\mathcal{G}] \tag{8.5}$$

In practice, we expect the interval of plausible values for $\text{Var}[\alpha_X - \mathcal{G}]$ to be even smaller than that defined above. For example, a correlation of either 0 or 1 between α_X and $\mathcal{G}$ is highly unlikely. Indeed, a correlation of 0 would result in our adjusted

expectation $\mathrm{E}_G[\alpha_X]$ being equal to $\mathrm{E}[\alpha_X]$, from the definition of PBA given in Equation 8.2 which can be seen in the practice in Figure 8.5.

For this case study, using the inequality defined in Equation 8.5 we determined that

$$4 \leq \mathrm{Var}[\alpha_X - \mathcal{G}] \leq 19,604,$$

as calculated using Equation 8.5. Note that these are simply limits and we need to obtain an expectation of $\mathrm{Var}[\alpha_X - \mathcal{G}]$ to calculate $\mathrm{Cov}[\alpha_X, \mathcal{G}]$ Thus, we decided that for proof of concept, we would take the variance of the difference between α_X and $\mathcal{G}$ to be 50^2 which is approximately the square of the average error of the radiocarbon determinations within the model and lies between the two limits of the equality given in Equation 8.5. Thus, this results in us setting $\mathrm{Cov}[\alpha_X, \mathcal{G}] = 8852$.

The set of posterior expectations $\mathcal{G}$ was obtained by building the six plausible chronological models detailed in Table 8.1 within PolyChron, then calculating the posterior expectation of α_X from the marginal posterior distribution of α_X for each model.

8.3 Results of PBA for a single parameter

Now that we have obtained values for the key quantities of interest, we can carry out the posterior belief assessment for the parameter α_X. We utilise existing code, written in `R` by Astfalck (2022), which calculates adjusted expectations and variances for our prior belief specification which has then been updated by $\mathcal{G}$. This code only requires specification of the key quantities outlined in Section 8.2.1 to allow the user to obtain adjusted beliefs for key quantities of interest. Within this section, we present the results of carrying out PBA using the quantities outlined in the section above. Following that, we carry out a form of sensitivity analysis, exploring how the

results of PBA change as we change the elicited values of key quantities.

In Figure 8.4 we illustrate the adjusted beliefs that result from the posterior belief assessment observing a notable reduction in variance about the expectation of α_X when compared to the posterior expectation for the same parameter from each of our plausible models, despite the fact our variance in the prior belief specification was rather large in comparison to the posterior variances of each of the models.

Though the initial result of carrying out PBA for a single parameter was promising, since our key values of interest were not formally elicited from an expert, and were rather arbitrarily chosen, we sought to explore how sensitive the results of PBA were to changes in the prior belief specification.

Poor prior specification of $\mathrm{E}[\alpha_X]$

First, we consider the impact on the adjusted beliefs when the elicited value of $\mathrm{E}[\alpha_X]$ is in poor agreement with the data. As mentioned in the previous section, when the elicited value of $\mathrm{Cov}[\alpha_X, \mathcal{G}]$ tends towards 0, $\mathrm{E}_{\mathcal{G}}[\alpha_X]$ becomes $\mathrm{E}[\alpha_X]$ as seen in the bottom plot of Figure 8.5. Consequently, for poor estimates of $\mathrm{E}[\alpha_X]$, a covariance equal to 0 results in a value of $\mathrm{E}_{\mathcal{G}}[\alpha_X] = \mathrm{E}[\alpha_X]$ which is consequently also in poor agreement with the data.

Alternatively, high values of $\mathrm{Cov}[\alpha_X, \mathcal{G}]$ result in adjusted beliefs that are closer to the posterior results observed in $\mathcal{G}$. However, the variance of the adjusted beliefs is unrealistically small as observed in the top plot of Figure 8.5. When setting the covariance to a small value, but not the limit of $\sqrt{\mathrm{Var}[\alpha_X]\mathrm{Var}[\mathcal{G}]}$ we obtain adjusted beliefs that are in agreement with the posterior expectations in $\mathcal{G}$, as seen in the middle plot of Figure 8.5. However, the variance of the adjusted beliefs is considerably smaller than we would expect for prior beliefs and model outputs that

disagree with each other so much, which we suspect to be due to the contrary prior information that we believe α_j is highly correlated with G, as implied by a value of $\text{Cov}(\alpha_j, G)$ close to $\sqrt{\text{Var}[\alpha_X]\text{Var}[\mathcal{G}]}$, despite our values of $\text{E}[\alpha_j]$ and $\text{E}[G]$ being so far apart.

In summary, PBA is sensitive to any prior belief specification that results in a value of $\text{E}[\alpha_X]$ that is in poor agreement with the data, regardless of the choice of covariance. Thus, it should only be used on Bayesian chronological models when tangible archaeological evidence can be used to inform elicitation of $\text{E}(\alpha_X)$, such as dating evidence from previous excavations.

Prior specification of variance

PBA performed better when $\text{E}(\alpha_X)$ was in good agreement with the data. We have seen in this section thus far that 'extreme' values of $\text{Cov}[\alpha_X, \mathcal{G}]$ affect the results of PBA. Now we turn our attention to the impact on the results of PBA when $\text{E}(\alpha_X)$ is in good agreement with the data and when the value of $\text{Var}[\alpha_X]$ increases or decreases, representing more or less certainty from the expert in their specified value of $\text{E}[\alpha_X]$.

As illustrated in Figure 8.6, smaller elicited values of variance, such as $\text{Var}[\alpha_X] = 2604$ and $\text{Var}[\mathcal{G}] = 3000$ have a high influence on the variance of the adjusted beliefs, suggesting that small values of $\text{Var}[\alpha_X]$ should be avoided unless they are supported by strong archaeological evidence. Whereas larger variances such as $\text{Var}[\alpha_X] = 29604$ and $\text{Var}[\mathcal{G}] = 30000$ did not lead to exceedingly large variances for adjusted beliefs when compared to the initial values $\text{Var}[\alpha_X] = 9604$ and $\text{Var}[\mathcal{G}] = 10000$, as demonstrated in the top plot of Figure 8.6.

8.4 Discussion and future work

Within this chapter, we have presented an initial exploration into the use of posterior belief assessment for parameters in a chronological model. We believe that PBA might be a potential methodology for allowing the consideration of multiple plausible chronological models, thus allowing further quantification of uncertainty in prior knowledge that is not well handled presently. However, considerable additional research would be required before it could be used in practice.

A limitation of PBA, specific to chronology construction, is that this method can allow us to carry out an analysis which utilises multiple plausible chronological models, but only when an informative prior belief structure can be elicited. While experts working at sites that have had previous dating carried out may have a wealth of prior knowledge suitable for this, more work is needed to explore other sources of prior knowledge and whether such knowledge can be reliably elicited.

8.4.1 Future work

Future work exploring the use of PBA for Bayesian chronological models will focus on three objectives. The first will be to investigate the use of PBA for multiple parameters. In the case study demonstrated in this chapter, the research focus was a specific parameter within the chronological model. However, archaeologists are often interested in the duration (or span) of a specific period of archaeological activity. In such situations, we would need to examine the use of PBA on at least pairs of parameters and their differences.

The second avenue of future research would focus on considering a larger volume of plausible models. There are two scenarios to consider. The first would be a large

number of plausible models, for which we believe, a priori, that all have similar expectations for the parameters of interest. The second would be a large number of plausible models for which the expert believes certain subsets of models lead to different estimations for the parameters of interest. For example, uncertainty about the Allen algebra relationships between groups may result in multiple plausible models that we believe, a priori, lead to differing posterior expectations for group boundary parameters.

The PBA methodology Williamson and Goldstein (2015) utilises the concept of coexchangeable classes. Given we have K coexchangable classes, the k^{th} coexchangable class is a set of H_k posterior expectations,

$$C_k = (\mathrm{E}\left[\alpha|\mathbf{x}; V_{k,1}\right], \mathrm{E}\left[\alpha|\mathbf{x}; V_{k,2}\right], ..., \mathrm{E}\left[\alpha|\mathbf{x}; V_{k,H_k}\right]),$$

such that we believe that any given posterior expectation within the class can be expressed as

$$\mathrm{E}\left[\alpha|\mathbf{x}; V_{k,h}\right] = M(C_k) + R_h(C_k),$$

such that $M(C_k)$ is the common posterior expectation of the parameters of interest for all models within that class k and $R_h(C_k)$ is a residual term attributed to the h^{th} expectation within class k, which accounts for the difference between the common posterior expectation of the class, $M(C_k)$ and the actual posterior expectation of the parameters of interest for model h (Astfalck et al., 2021). We are interested in investigating the use of coexchangable classes when carrying out PBA since it would allow us to group large numbers of plausible models into classes that we believe, a priori will provide the same adjusted beliefs of the parameter (or parameters) of interest.

The final objective, given research from the two aforementioned objectives proves successful, would be to establish protocols and training for the elicitation of prior beliefs about the parameters for which we are carrying out PBA. A particular focus would be on incorporating prior knowledge that may previously have been excluded since it was not possible to elicit a formal probability distribution from such expert knowledge, but eliciting expectations and variances from such knowledge seems plausible. An example of such prior knowledge might be calendar ages for cultural finds such as coins and pottery, for which approximate calendar ages can be inferred from inscriptions and cultural style respectively. We wish to explore whether formal elicitation processes can be determined to elicit such prior knowledge.

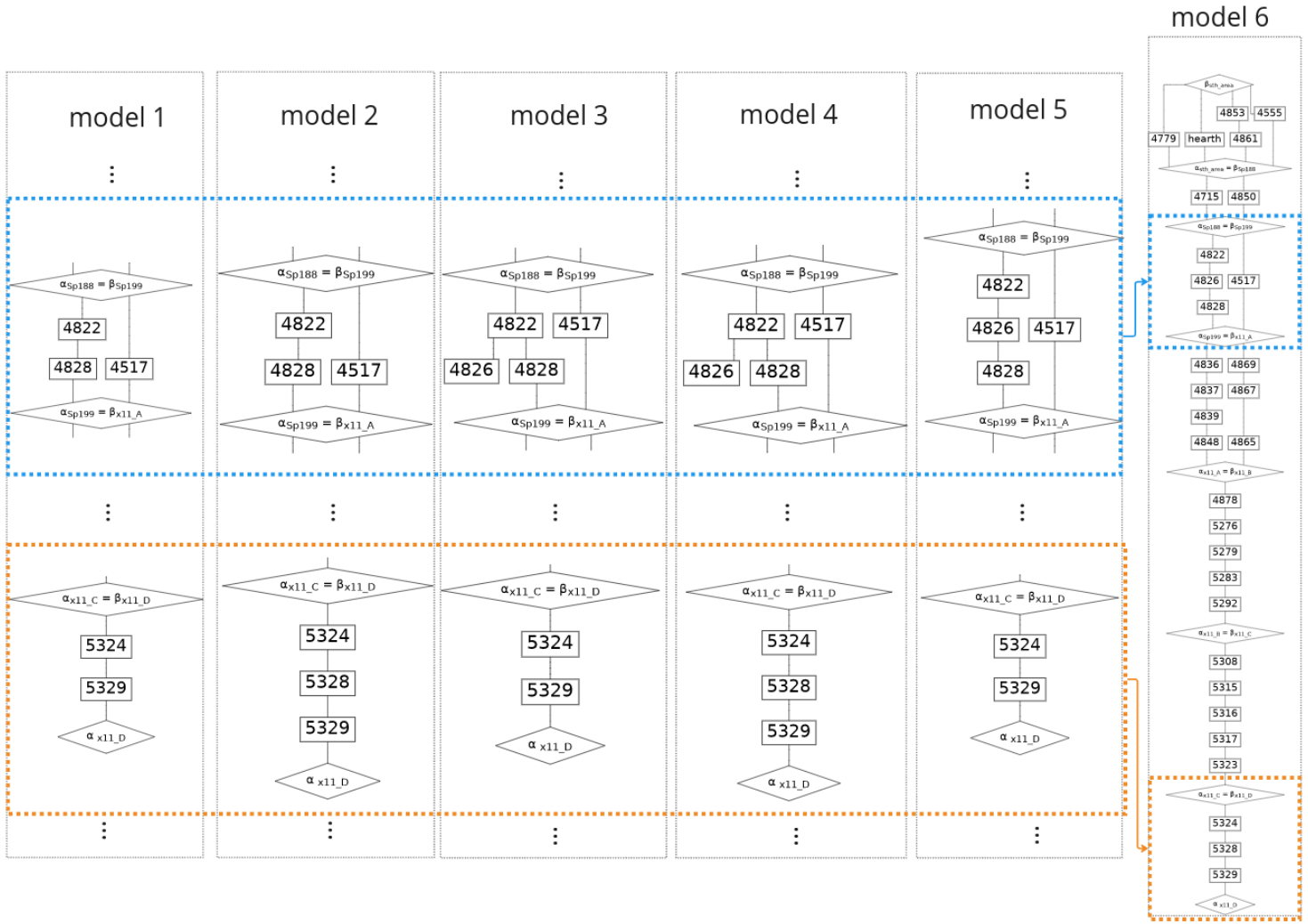


Figure 8.3: Chronological DAGs corresponding to the six plausible models outlined in Table 8.1. For visualisation purposes, only the parts of the chronological DAGs that change for each model are provided for models 1 to 5. For a larger scale render of model 6, please refer back to Figure 8.1.

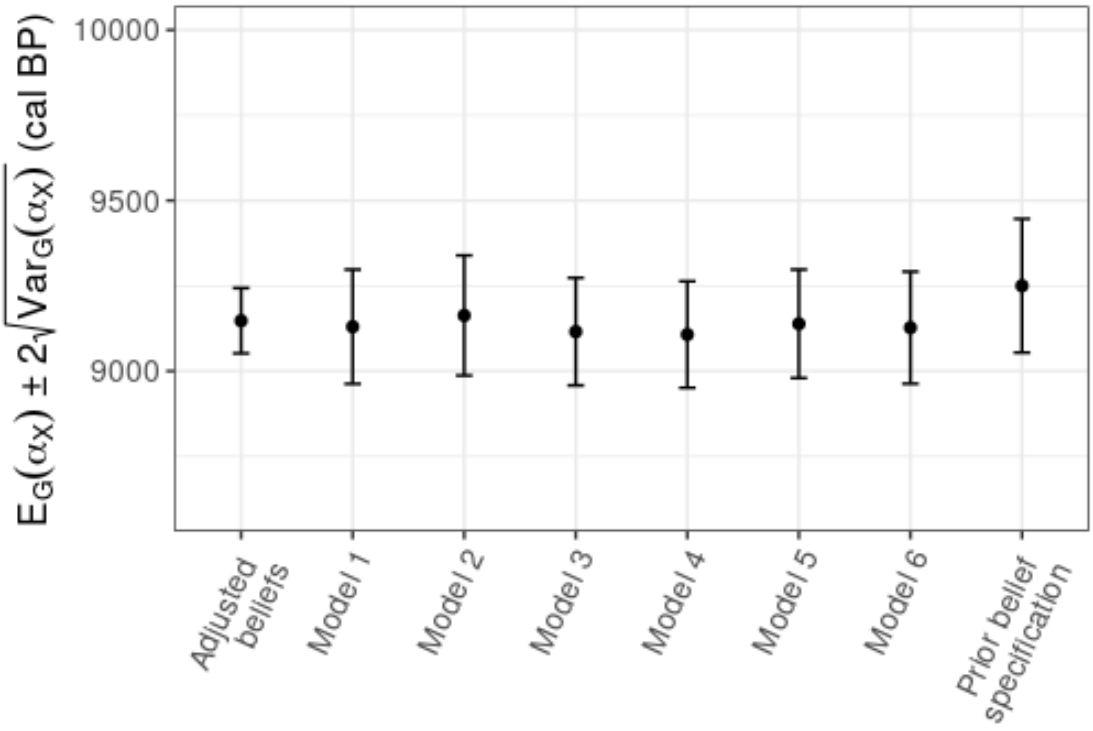


Figure 8.4: Results of carrying out PBA for α_X, as seen in Figure 8.1, such that the prior belief specification gives $\mathrm{E}[\alpha_X] = 9250$ and $\mathrm{Var}[\alpha_X] = 9604$. G is the set of six posterior expectations for α_X corresponding to the six models seen in Figure 8.3. Further, $\mathrm{E}[\mathcal{G}] = 9250$, $\mathrm{Var}[\mathcal{G}] = 10000$ and $\mathrm{Cov}[\alpha_X, \mathcal{G}] = 8852$. A 76% reduction is seen in the variance of the prior beliefs for the adjusted beliefs

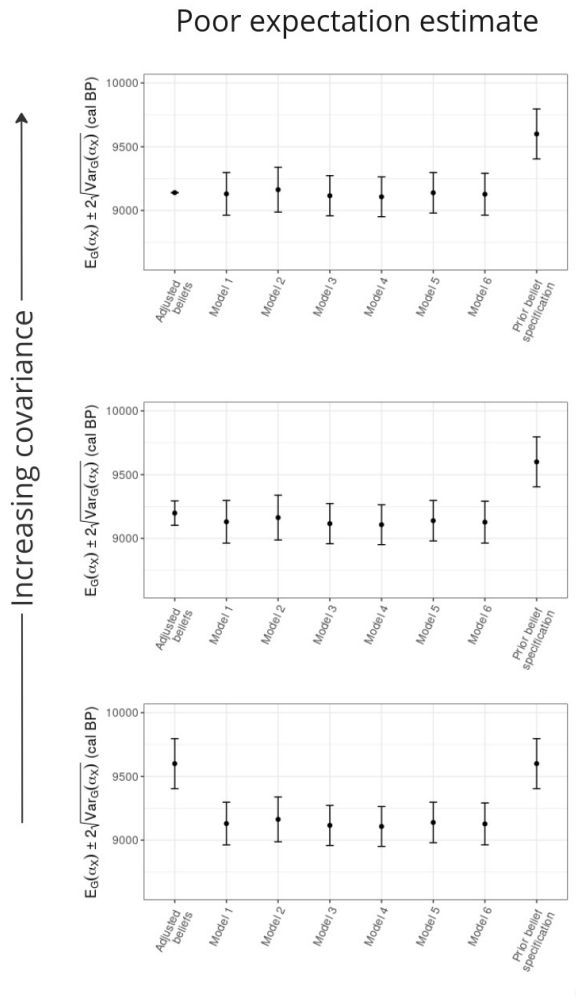


Figure 8.5: Results of carrying out PBA for α_X, as seen in Figure 8.1, such that the prior belief specification gives $\mathrm{E}[\alpha_X] = 9600$ and $\mathrm{Var}[\alpha_X] = 9604$. G is the set of six posterior expectations for α_X corresponding to the six models seen in Figure 8.3. Further, $\mathrm{E}[\mathcal{G}] = 9600$, and $\mathrm{Var}[\mathcal{G}] = 10000$ and we vary $\mathrm{Cov}[\alpha_X, \mathcal{G}]$. *top:* A unrealistically negligible variance for the adjusted beliefs is observed for $\mathrm{Cov}[\alpha_X, \mathcal{G}] = \sqrt{\mathrm{Var}[\alpha_X]\mathrm{Var}[\mathcal{G}]}$, *middle:* A $\mathrm{Cov}[\alpha_X, \mathcal{G}] = 8852$ suggests that α_X and G are highly correlated which is in disagreement with their expectations, resulting in an unrealistically small variance of the adjusted beliefs. *bottom:* $\mathrm{Cov}[\alpha_X, \mathcal{G}] = 0$ which results in the adjusted beliefs being set equal to the prior belief specification.

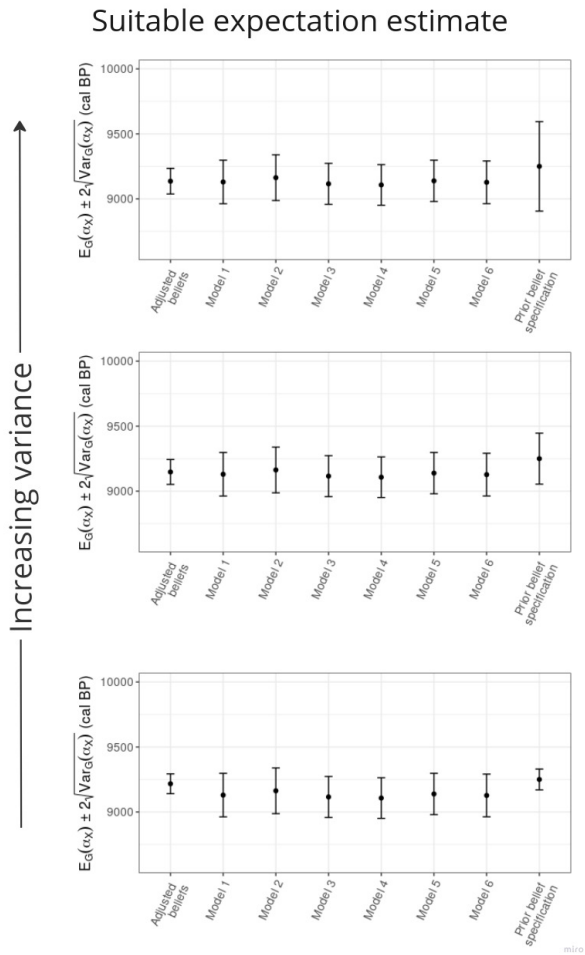


Figure 8.6: Results of carrying out PBA for α_X, as seen in Figure 8.1, such that the prior belief specification gives $\mathrm{E}[\alpha_X] = 9250$ and $\mathrm{Cov}[\alpha_X, \mathcal{G}] = 8852$. G is the set of six posterior expectations for α_X corresponding to the six models seen in Figure 8.3. We vary $\mathrm{Var}[\alpha_X]$ and $\mathrm{Var}[\mathcal{G}]$. *top:* Results of setting $\mathrm{Var}[\alpha_X] = 29604$ and $\mathrm{Var}[\mathcal{G}] = 30000$ with a notable reduction in the variance of our adjusted beliefs still observed, *middle:* results of setting $\mathrm{Var}[\alpha_X] = 9960$ and $\mathrm{Var}[\mathcal{G}] = 10000$ as seen in Figure 8.4 and provided here for comparison, *bottom:* Results of setting $\mathrm{Var}[\alpha_X] = 2604$ and $\mathrm{Var}[\mathcal{G}] = 3000$ which shows a very small variance for our adjusted beliefs.

Chapter 9

Conclusions and future work

Within this thesis, we have demonstrated research that built upon initial work by Dye and Buck (2015), who first suggested the use of graph theory to semi-automate the management and modelling of absolute and relative dating evidence. We have demonstrated that utilising directed acyclic graphs allows for the efficient transfer of digital information pertaining to absolute and relative dating evidence throughout the chronology construction process. In particular, we have demonstrated how our novel software, `PolyChron` can reliably manage a large volume of complex stratigraphic relationships and grouping, in addition to radiocarbon determinations. Thus, enabling the semi-automation of chronology construction.

In this chapter, we provide conclusions about the novel research carried out for this thesis, before outlining future work that will improve upon the aforementioned research.

9.1 Conclusions for novel research

In Chapter 4, we presented the author's research into the quality and utility of digital archaeological dating evidence within UK digital repositories (Moody et al., 2021). We focused on digital archaeological data archived at ADS, since this is the archive predominately used by both academic and commercial archaeologists in the UK. What we observed as a result of the data review was insufficient archiving of the digital dating evidence used and produced during the chronological modelling process, thus hindering future reuse of such data.

In particular, we found a lack of data pertaining to large and complex stratigraphic sequences, which we postulated (following discussion with the user community) was due to such data being held in paper format due to a lack of appropriate software. This research has been widely disseminated within the archaeological research community, including at international meetings organised by the SEADDA (Saving European Archaeology from the Digital Dark Age) project.

In Chapter 7, we reviewed existing software for both managing relative dating evidence and for Bayesian chronology modelling, demonstrating that in order for the novel software we used/wrote to be open-source and free to use, it required us to write new purposed-built software for the automation of chronology construction. We then highlighted key features of our novel software `PolyChron`, which are expanded upon in the supplementary video. We demonstrated the different stages of building a Bayesian chronological model within `PolyChron` and discussed the management of data at each stage.

We demonstrated the (largely) automated saving of data and supplementary information pertaining to Bayesian chronological modelling to the user's machine, which we

hope would improve archiving of such data in digital repositories when `PolyChron` is introduced to the user community. Further, we showcased the use of directed acyclic graphs for the rapid building of Bayesian chronological models and provide input to appropriate implementation algorithms with very little user input and effort.

Further, we introduced the modular format of `PolyChron`, highlighting how this format will allow us to continually improve and change the software as required by any advances in the statistical or computational theory utilised.

This chapter also provided case studies showing how we used `PolyChron` to semi-automate the construction of multiple plausible chronological models for both small and large stratigraphic sequences. These case studies demonstrated that DAGs can be used to manage dating evidence within PolyChron and reliably provide input to the posterior inference module, thus obtaining posterior estimates for models with varying features and prior knowledge, such as:

- removed contexts,
- contexts with residual/intrusive samples,
- and a variety of group relationships.

In particular, our case study demonstrating the use of PolyChron to build multiple Bayesian chronological models for relative dating evidence from Cunliffe and Poole (1991) exemplified the potential for the user community to use `PolyChron` to manage large stratigraphic sequences and all other associated dating evidence.

Chapter 8 built upon this research and showcases a novel use of an existing statistical methodology called posterior belief assessment, which allowed us to incorporate multiple plausible chronological models into our analysis and thus obtain posterior estimates that better represent our prior knowledge and uncertainty for a hypothet-

ical parameter of interest within the chronological model, taken from an existing chronological model from excavations at the East Mound of Çatalhöyük.

We found that provided sufficient prior knowledge was available to allow us to obtain informative prior expectations and variances for a subset of parameters of interest within a chronological model, PBA provided a suitable methodological framework for quantifying uncertainty in our Bayesian analysis when we consider multiple chronological models plausible, for a single parameter of interest. Further work is required, however, to investigate the use of PBA on multiple parameters of interest.

9.2 Future work

Future planned work in this thesis has been discussed in detail in Chapter 7 Section 7.6 and Chapter 8 Section 8.4.1. Here, we highlight the immediate focus of our future research.

The primary avenue of future research is the conversion of PolyChron to a format that could be considered a minimal viable product. The focus with this would be on error handling, in particular, on providing the user with informative error messages when problems occur. Further, the graphical user interface for PolyChron will likely require conversion to an online interface to improve usability.

In addition, we would like to improve the capacity for PolyChron to handle additional data and information obtained during excavation and post-excavation research. This includes, but is not limited to:

- supplementary data for radiocarbon determinations (at a minimum providing the data that satisfies conventions 4-6 for reporting radiocarbon determinations provided by Millard (2014)),

- sub-grouping of contexts, as discussed in Chapter 2 Section 2.2,
- additional sources of scientific dating,
- and dates deriving from cultural finds (e.g. pottery and coins).

We seek to ensure that, in addition to allowing users to rapidly semi-automate the construction of multiple Bayesian chronological models, all required data and supplementary information used and produced during this process is automatically saved to the user's machine. As a result, this would provide all data required for a complete site archive for a given archaeological site, thus improving the potential for the reuse of such data.

Bibliography

Allen, J. F.

1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843. xiv, xxiii, xxiv, 15, 17, 44, 81

Anderson, E. C. and W. F. Libby

1946. Atmospheric helium three and radiocarbon from cosmic radiation. *Phys. Rev.*, 69:671–672. 20

Astfalck, L.

2022. *R package for Bayes linear analysis*. 152

Astfalck, L., D. Williamson, N. Gandy, L. Gregoire, and R. Ivanovic

2021. Coexchangeable process modelling for uncertainty quantification in joint climate reconstruction. *Journal of the American Statistical Association*. 156

Baillie, M.

1982. *Tree-ring Dating and Archaeology*. London: Croon Helm. 26

Barker, P.

1993. *Techniques of archaeological excavation*, Batsford studies in archaeology, 3rd ed., fully revised. edition. Batsford. 1

Bayliss, A.

2009. Rolling out revolution: Using radiocarbon dating in archaeology. *Radiocarbon*, 51:123–47. 14, 20, 61, 67

Bayliss, A.

2015. Quality in Bayesian chronological models in archaeology. *World Archaeology*, 47(4):677–700. 67

Bayliss, A., F. Brock, S. Farid, I. Hodder, J. Southon, and R. E. Taylor

2015. Getting to the bottom of it all: A Bayesian approach to dating the start of Çatalhöyük getting to the bottom of it all: A Bayesian approach to dating the start of Çatalhöyük. *Source: Journal of World Prehistory*, 28:1–26. xix, 144, 145, 147

Bayliss, A. and C. Bronk Ramsey

2004. *Pragmatic Bayesians: a decade of integrating radiocarbon dates into chronological models*, chapter 2, Pp. 25–41. London: Springer-Verlag. 63, 139, 140, 150

Bayliss, A., S. Farid, and T. Higham

2014. Time will tell: practising Bayesian chronological modelling on the east mound. In *Çatalhöyök excavations: the 2000–2008 seasons*, I. Hodder, ed., Pp. 53–90. Los Angeles: (CA): Cotsen Institute of Archaeology. 42

Bayliss, A. and P. Marshall

2022. *Radiocarbon Dating and Chronological Modelling: Guidelines and Best Practice*. Historic England. 88, 105, 139

Bernardo, J. M. and A. F. M. Smith

1994. *Bayesian theory*. Wiley. 138

Bhattacharyya, A.

1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109. 130

Blaauw, M.

2010. Methods and code for 'classical' age-modelling of radiocarbon sequences. *Quaternary Geochronology*, 5:512–518. 31

Blaauw, M. and J. A. Christen

2011. Flexible paleoclimate age-depth models using an autoregressive gamma process. *Bayesian Analysis*, 6:457–474. 104, 105

Blackwell, P. and C. E. Buck

2008. Estimating radiocarbon calibration curves. *Bayesian Analysis*, 3. 25

Bowman, S.

1990. *Radiocarbon Dating*, Blum, Fred H. University of California Press. 21, 22, 23, 24, 25

Bronk Ramsey, C.

1995. Radiocarbon calibration and analysis of stratigraphy: The oxcal program. *Radiocarbon*, 37(2):425–430. 67

Bronk Ramsey, C.

1997. Probability and dating. *Radiocarbon*, 40(1):461–474. 106

Bronk Ramsey, C.

2008. Radiocarbon dating: Revolutions in understanding. *Archaeometry*, 50:249–275. 21, 133

Bronk Ramsey, C.

2009. Dealing with outliers and offsets in radiocarbon dating. *Radiocarbon*, 51(3):1023–1045. 63

Bronk Ramsey, C.

2023. *OxCal 4.4 Manual*. 102, 105

Bronk Ramsey, C., T. Higham, and P. Leach

2004. Towards high-precision ams: Progress and limitations. *Radiocarbon*, 46(1):17–24. 23

Buck, C., C. Litton, and A. Smith

1992. Calibration of radiocarbon results pertaining to related archaeological events. *Journal of Archaeological Science*, 19(5):497–512. 64, 67

Buck, C. E. and P. G. Blackwell

2004. Formal statistical models for estimating radiocarbon calibration curves. *Radiocarbon*, 46:1093–1102. 27

Buck, C. E., W. G. Cavanagh, and C. D. Litton

1996. *Bayesian Approach to Interpreting Archaeological Data*, chapter 9, Pp. 201–252. Chichester, England: Wiley. xvi, 58, 63, 65, 67, 68, 70, 80, 89

Buck, C. E., A. Christen, and G. N. James

1999. Bcal: an on-line Bayesian radiocarbon calibration tool. *Internet Archaeology*. 67, 102, 105, 107

Buck, C. E., J. A. Christen, J. Kenworthy, and C. Litton

1994. Estimating the duration of archaeological activity using 14C determinations. *Oxford Journal of Archaeology*, 13:229–240. xvii, 67, 88, 91, 92, 93, 95

Buck, C. E., C. D. Litton, J. B. Kenworthy, and A. F. Smith

1991. Combining archaeological and radiocarbon information: A Bayesian approach to calibration. *Antiquity*, 65:808–821. 66

Buck, C. E. and A. R. Millard

2004. *Tools for constructing chronologies. Crossing disciplinary boundaries*. 2, 137

Bédard, M.

2008. Optimal acceptance rates for Metropolis algorithms: Moving beyond 0.234. *Stochastic Processes and their Applications*, 118(12):2198–2222. 198

Carver, M.

2014. *Excavation Methods in Archaeology*. New York, NY: Springer New York. 1

Cessford, C., M. Blumbach, H. G. Akoğlu, T. Higham, P. I. Kuniholm, S. W. Manning, M. W. Newton, M. Ozbakan, and A. M. Ozer

2005. Absolute dating at Çatalhöyök. In *Inhabiting Çatalhöyük: Reports from the 1995–1999 Seasons*, volume 4, Pp. 65–99. Cambridge and London: McDonald Institute for Archaeological Research and British Institute of Archaeology at Ankara. 42, 150

Christen, J. A.

1994. Summarizing a set of radiocarbon determinations: A robust approach. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43:489–503. 32, 67, 129

Cunliffe, B. and C. Poole

1991. *The Excavations 1979-1988: The Site*, volume 4. Council for British Archaeology. xix, 58, 121, 126, 127, 136, 165

de Vries, H.

1958. *Variation in concentration of radiocarbon with time and location on Earth.* Akademie Van Wet. 25

Discamps, E., B. Gravina, and N. Teyssandier

2015. In the eye of the beholder: contextual issues for Bayesian modelling at the Middle-to-Upper Palaeolithic transition. *World Archaeology*, 47(4):601–621. 42

Dye, T. S.

2019. *Write Graphviz dot files for archaeological sequence diagrams and Bayesian chronological models.* 103

Dye, T. S. and C. E. Buck

2015. Archaeological sequence diagrams and Bayesian chronological models. *Journal of Archaeological Science*, 63:84–93. i, 12, 42, 85, 87, 91, 103, 163

Enclyclopedia of Mathematics

2020. Hellinger distance. http://encyclopediaofmath.org/index.php? 130

Evans, T.

2015. A reassessment of archaeological grey literature: semantics and paradoxes. *Internet Archaeology.* 51

Faniel, I., A. Austin, E. Kansa, S. Witcher Kansa, P. France, J. Jacobs, R. Boytner, and E. Yakel

2018. Beyond the archive: Bridging data creation and reuse in archaeology. *Advances in Archaeological Practice*, 6(2):105–116. 40

Faniel, I., E. Kansa, S. Whitcher-Kansa, J. Barrera-Gomez, and E. Yakel

2013. The challenges of digging data: A study of context in archaeological

data reuse. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, Pp. 295–304, New York, NY, USA. Association for Computing Machinery. 39

Free Software Foundation, Inc

2019. *Bash manual page*. Ubuntu. 52

Gelfand, A. E. and A. F. M. Smith

1991. Gibbs sampling for marginal posterior expectations. *Communications in Statistics-theory and Methods*, 20:1747–1766. 66

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin

2004. *Bayesian Data Analysis*, 3rd ed. edition. Chapman and Hall. 64, 138, 191, 192

Gelman, A. and D. B. Rubin

1992. Inference from iterative simulation using multiple sequences. *Science*, 7:457–472. 198

Geman, S. and D. Geman

1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741. 66, 71

Ghosh, J. K., M. Delampady, and T. Samanta

2006. *An Introduction to Bayesian Analysis Theory and Methods / by Jayanta K. Ghosh, Mohan Delampady, Tapas Samanta.*, Springer Texts in Statistics, 1st ed. 2006. edition. New York, NY: Springer New York : Imprint: Springer. 70, 189

Godwin, H.

1962. Half-life of Radiocarbon. *Nature*, 195(4845):984. 24

Goldstein, M. and D. Wooff

2007. *Bayes linear statistics: Theory and methods*. John Wiley & Sons. 141

Hagberg, A., P. Swart, and D. S Chult

2008. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States). 117, 135

Harris, E., M. Brown, and G. Brown

1993. *Practices of archaeological stratigraphy*. London: Academic Press. 9

Harris, E. C.

1975. The stratigraphic sequence: A question of time. *Archaeology*, 7:109–121. 62, 102

Harris, E. C.

1989. *Principles of Archaeological Stratigraphy*, 2nd edition. London: Academic Press. 6, 8, 10

Haslett, J. and A. Parnell

2008. A simple monotone process with application to radiocarbon-dated depth chronologies. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 57:399–418. 104

Heaton, T. J., M. Blaauw, P. G. Blackwell, C. Bronk Ramsey, P. J. Reimer, and E. M. Scott

2020a. The IntCal20 approach to radiocarbon calibration curve construction: A new methodology using Bayesian splines and errors-in-variables. *Radiocarbon*, 62:821–863. 27, 28, 123

Heaton, T. J., P. Köhler, M. Butzin, E. Bard, R. W. Reimer, W. E. N. Austin, C. Bronk Ramsey, P. M. Grootes, K. A. Hughen, B. Kromer, and et al.

2020b. Marine20—the marine radiocarbon age calibration curve (0–55,000 cal BP). *Radiocarbon*, 62(4):779–820. 27

Herzog, I.

1993. *Computer-aided Harris Matrix generation*, Pp. 201–217. Academic Press. 12, 102

Herzog, I.

2006. No news from stratigraphic computing? 102

Herzog, I. and I. Scollar

1990. A new graph theoretic oriented program for Harris Matrix analysis. In *Computer applications and quantitative methods in archaeology 1990*, Pp. 53–59. Southhampton, UK: Computer Applications and Quantitative Methods in Archaeology. 12

Higham, T. F. G. and G. S. Heep

2019. Reply to: 'In the eye of the beholder: contextual issues for Bayesian modelling at the Middle-to-Upper Palaeolithic transition', by Discamps, Gravina and Teyssandier (2015). *World Archaeology*, 51(1):126–133. 42

Historic England

2018. *Historic England recording manual*. Historic England, Fort Cumberland, UK. 41

Historic England

2021. Fish vocabularies. xviii, 110

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky

1999. Bayesian model averaging: A tutorial. *Statistical Science*, 14:382–417. 139

Hogg, A. G., T. J. Heaton, Q. Hua, J. G. Palmer, C. S. Turney, J. Southon, A. Bayliss, P. G. Blackwell, G. Boswijk, C. Bronk Ramsey, and et al.

2020. Shcal20 southern hemisphere calibration, 0–55,000 Years cal BP. *Radiocarbon*, 62(4):759–778. 27

Hollander, H.

2017. Saving treasures of the world heritage of mankind at the digital archive DANS. *Internet Archaeology*. 38

Huggett, J.

2018. Reuse remix recycle. *Advances in Archaeological Practice*, 6(2):93–104. 39, 57

Jensen, P.

2017. Where are we? reviewing the integration of complex spatial data in current field archaeology. *Internet Archaeology*. 6

Johnson, F., F. H. H. Roberts, R. F. Heizer, J. B. Griffin, W. S. Webb, W. A. Ritchie, H. de Terra, J. Bird, H. L. Movius, R. J. Braidwood, T. Jacobsen, R. A. Parker, S. Weinberg, R. F. Flint, E. S. Deevey, F. Rainey, and D. Collier

1951. Radiocarbon Dating: A Report on the Program to Aid in the Development of the Method of Dating. *Memoirs of the Society for American Archaeology*, 8(8):1–65. 24

Jones, M. and G. Nicholls

2002. New radiocarbon calibration software. *Radiocarbon*, 44:663–674. 105, 106

Jones, S., A. MacSween, S. Jeffrey, R. Morris, and M. Heyworth

2003. From the ground up: The publication of archaeological projects: a user needs survey. a summary. *Internet Archaeology*. 51

Joyce, J. M.

2011. Kullback-leibler divergence. In *International encyclopedia of statistical science*, Pp. 720–722. Springer. 130

Karlsberg, A. J.

2006. *Flexible Bayesian Methods for Archaeological Dating*. PhD thesis, University of Sheffield. 138

Kass, R. E. and A. E. Raftery

1995. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795. 138

Katz, L.

1953. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43. 128

Kondev, F. G., M. Wang, W. J. Huang, S. Naimi, and G. Audi

2021. The NUBASE2020 evaluation of nuclear physics properties. *Chinese Physics C*, 45(3):030001. 24

Krus, A. M., R. Cook, and D. Hamilton

2015. Bayesian chronological modeling of sunwatch, a fort ancient village in dayton, ohio. *Radiocarbon*, 57:965–977. 144

Lanos, P. and A. Philippe

2018. Event date model: A robust Bayesian tool for chronology building. *Communications for Statistical Applications and Methods*, 25:131–157. 106

Lerman, J., W. Mook, J. Vogel, and I. Olsson

1970. C14 in tree rings from different localities. In *Radiocarbon Variations and Absolute Chronology. Proceedings, XII Nobel Symposium. New York: Wiley. p*, Pp. 275–301. 26

Levy, E., G. Geeraerts, F. Pluqet, E. Piasetzky, and A. Fantalkin

2020. Chronological networks in archaeology: A formalised scheme. *Journal of Archaeological Science*, Pp. 105–225. 106, 108

Libby, W.

1955. *Radiocarbon Dating*. Chicago, USA: University of Chicago Press. 20, 22, 24

Libby, W. F., E. C. Anderson, and J. R. Arnold

1949. Age determination by radiocarbon content: World-wide assay of natural radiocarbon. *New Series*, 109:227–228. 20, 23

Litton, C. and M. Leese

1991. Some statistical problems arising in radiocarbon calibration. In *Computer Applications and Quantitative Methods in Archaeology 1990*, S. Rahtz and K. Lockyear, eds., BAR International Series 565, Pp. 100–109, Oxford. Tempus Reparatum. 66

Lucas, G.

2015. Archaeology and contemporaneity. *Archaeological Dialogues*, 22:1–15. 15

Marciniak, A., M. Barański, A. Bayliss, L. Czerniak, T. Goslar, J. Southon, and R. Taylor

2015. Fragmenting times: Interpreting a Bayesian chronology for the Late Neolithic occupation of Çatalhöyök East, Turkey. *Antiquity*, 89:154–176. 42, 58

May, K.

2020. The Matrix: Connecting time and space in archaeological stratigraphic records and archives. *Internet Archaeology*, 55. 14, 15, 44

May, K., J. Taylor, and C. Binding

2023. Stratigraphic analysis and the matrix: connecting and re-using digital records and archives of archaeological investigations. *Internet Archaeology*. 104

McManamon, F. and K. Kintigh

2017. tdar: A cultural heritage archive for twenty-first-century public outreach, research, and resource management. *Advances in Archaeological Practice*, 5. 38

Millard, A.

2015. *An introduction to model building in OxCal*, 1.0 edition. 82

Millard, A. R.

2014. Conventions for reporting radiocarbon determinations. *Radiocarbon*, 56(2):555–559. 166

Moody, B. C.

2019. *Code for the Quality and Utility of Resources in Digital Heritage Repositories*. GitHub. 53

Moody, B. C., T. Dye, K. May, H. Wright, and C. E. Buck

2021. Digital chronological data reuse in archaeology: Three case studies with varying purposes and perspectives. *Journal of Archaeological Science: Reports*, 40:103188. i, 3, 37, 44, 57, 58, 164

Naylor, J. C. and A. F. M. Smith

1988. An archaeological inference problem. *Journal of the American Statistical Association*, 83:588–595. 66, 70, 77, 78, 80, 81

Nicholls, G. and M. Jones

2001. Radiocarbon dating with temporal order constraints. *Journal of the Royal Statistical Society Series C*, 50:503–521. 31, 78, 81, 82, 84, 100, 107, 193, 195

Nicholson, C., S. Kansa, N. Gupta, and R. Fernandez

2023. Will it ever be fair? making archaeological data findable, accessible, interoperable, and reusable. *Advances in Archaeological Practice*, 11:63–75. 36, 38

NobelPrize.org

2022. The Nobel Prize in Chemistry 1960. https://www.nobelprize.org/prizes/chemistry/1960/summary/. 20

O'Hagan, A.

2008. *Handbook of Probability: Theory and Applications*, Pp. 85–100. Thousand Oaks, California: SAGE Publications, Inc. 62

Open Research Data Task Force

2019. Realising the potential final report of the open research data task force executive summary. Technical report, UK Government. 100

Otárola-Castillo, E., M. G. Torquato, and C. E. Buck

2023. The Bayesian inferential paradigm in archaeology. In *Handbook of Archaeological Sciences*, Pp. 1193–1209. John Wiley & Sons, Ltd. 63

Pastore, M. and A. Calcagnì

2019. Measuring distribution similarities between samples: A distribution-free overlapping index. *Frontiers in Psychology*, 10. 130, 205

Pearson, G. W. and M. Stuiver

1986. High-precision calibration of the radiocarbon time scale, 500–2500 BC. *Radiocarbon*, 28(2B):839–862. 27

Reimer, P., W. Austin, E. Bard, A. Bayliss, P. Blackwell, C. Bronk Ramsey, M. Butzin, H. Cheng, R. Lawrence Edwards, M. Friedrich, P. Grootes, T. Guilderson, I. Hajdas, T. Heaton, A. Hogg, K. Hughen, B. Kromer, S. Manning, R. Muscheler, J. Palmer, C. Pearson, J. van der Plicht, R. Reimer, D. Richards, E. Scott, J. Southon, C. Turney, L. Wacker, florian adolphi, U. Büntgen, M. Capano, S. Fahrni, alexandra Fogtmann-Schulz, R. Friedrich, F. Miyake, jesper olsen, F. Reinig, M. Sakamoto, A. sookdeo, and S. Talamo

2020. The IntCal20 Northern Hemisphere radiocarbon age calibration curve (0-55 cal kBP). *Radiocarbon*, 62(4):725. xv, 27, 28, 72, 82

Reimer, P. J.

2021. Evolution of radiocarbon calibration. *Radiocarbon*, P. 1–17. 25, 26, 28, 30

Reimer, P. J., M. G. L. Baillie, E. Bard, A. Bayliss, J. W. Beck, C. J. H. Bertrand, P. G. Blackwell, C. E. Buck, G. S. Burr, K. B. Cutler, P. E. Damon, R. L. Edwards, R. G. Fairbanks, M. Friedrich, T. P. Guilderson, A. G. Hogg, K. A. Hughen, B. Kromer, G. McCormac, S. Manning, C. Bronk Ramsey, R. W. Reimer, S. Remmele, J. R. Southon, M. Stuiver, S. Talamo, F. W. Taylor, J. van der Plicht, and C. E. Weyhenmeyer

2004. IntCal04 terrestrial radiocarbon age calibration, 0-26 cal kyr BP. *Radiocarbon*, 46(3):1029–1058. 27

Reimer, P. J., E. Bard, A. Bayliss, J. W. Beck, P. G. Blackwell, C. Bronk Ramsey, C. E. Buck, H. Cheng, R. L. Edwards, M. Friedrich, P. M. Grootes, T. P. Guilderson, H. Haflidason, I. Hajdas, C. Hatté, T. J. Heaton, D. L. Hoffmann, A. G. Hogg, K. A. Hughen, K. F. Kaiser, B. Kromer, S. W. Manning, M. Niu, R. W. Reimer, D. A. Richards, E. M. Scott, J. R. Southon, R. A. Staff, C. S. M.

Turney, and J. van der Plicht

2013. IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon*, 55:1869–1887. 27

Richards, J. D.

2008. Managing digital preservation and access: The Archaeology Data Service. In *Managing Archaeological Resources*, F. P. McManamon, A. Stout, and J. A. Barnes, eds., One World Archaeology, Pp. 173–94. Left Coast Press. 38

Richards, J. D.

2017. Twenty years preserving data: A view from the United Kingdom. *Advances in Archaeological Practice*, 5(3):227–237. 38, 39

Robert, C. P. and G. Casella

1999. *Monte Carlo statistical methods*, Springer texts in statistics, 2nd ed. edition. New York: Springer. 189, 193

Roskams, S.

2002. Excavation. *Journal of Anthropological Research*, 58(2):270–272. 5, 9

Rutherford, E. and F. Soddy

1903. Lx. radioactive change. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 5:576–591. 21

Ryan, N. S.

2001. Jnet, a successor to gnet. In *Archaologie und Computer, Workshop 6*. 102

Sahu, S. K.

2004. *Applications of Formal Model Choice to Archaeological Chronology Building*, Pp. 111–127. London: Springer London. 137

Schiffer, M. B.

1986. Radiocarbon dating and the "old wood" problem: The case of the hohokam chronology. *Journal of Archaeological Science*, 13(1):13–30. 91

Scott, E. M., G. T. Cook, and P. Naysmith

2007. Error and uncertainty in radiocarbon measurements. *Radiocarbon*, 49:427–440. 23

SEADDA

February 2023. Homepage. 39

Spade, D. A.

2020. Chapter 1 - Markov chain Monte Carlo methods: Theory and practice. In *Principles and Methods for Data Science*, A. S. Srinivasa Rao and C. Rao, eds., volume 43 of *Handbook of Statistics*, Pp. 1–66. Elsevier. 196

Spence, C.

1990. *Archaeological Site Manual*. London. 9

Stuiver, M. and G. W. Pearson

1986. High-precision calibration of the radiocarbon time scale, ad 1950–500 bc. *Radiocarbon*, 28(2B):805–838. 27

Stuiver, M., G. W. Pearson, and T. Braziunas

1986. Radiocarbon age calibration of marine samples back to 9000 Cal Yr BP. *Radiocarbon*, 28(2B):980–1021. 26

Stuiver, M. and P. J. Reimer

1993. Extended 14C data base and revised CALIB 3.0 14C age calibration program. *Radiocarbon*, 35:215–230. 101

Stuiver, M. and H. E. Suess

1966. On the relationship between radiocarbon dates and true sample ages. *Radiocarbon*, 8:534–540. 26

Suess, H. E.

1967. Bristlecone pine calibration of the radiocarbon time scale from 4100 to 1500 BC. *Radioactive Dating and Methods of Low-Level Counting. Vienna, International Atomic Energy Agency*, Pp. 143–151. 26

Suess, H. E.

1970. Bristlecone pine calibration of the radiocarbon time-scale 5200 BC to the present. *Radiocarbon variations and absolute chronology, Proceedings of XIIth Nobel Symposium, 1970*, Pp. 303–308. 26, 30

Swedish National Data Service

February 2023. Homepage. 38

Switsur, V.

1973. The radiocarbon calendar recalibrated. *Antiquity*, 47(186):131. 26

The Archaeology Data Service

July 2020. Guidelines for depositors. 38, 41

The SciPy community

2023. *Integration - Manual page*. 191

Thompson, M. B., J. N. Thompson, and D. M. Gates

2022. biosphere Encyclopedia Britannica. 22

Traxler, C. and W. Neubauer

2008. The Harris Matrix Composer - A New Tool to Manage Archaeological

Stratigraphy. In *Digital Heritage, Proceedings of the 14th International Conference on Virtual Systems and Multimedia*, Pp. 13–20. VSMM Commitee. 102

Vallverdu, J.

2008. The false dilemma: Bayesian vs. frequentist. *Electronic Journal for Philosophy*. 62

Van Rossum, G. and F. L. Drake

2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace. 99

Ward, G. and S. Wilson

1978. Procedures for comparing and combining radiocarbon age determinations: A critique. *Archaeometry*, 20:19–31. 102

Wasserman, L.

2000. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44:922107. 138

Wheeler, M.

1956. *Archaeology from the Earth*, Pelican books A356. Penguin Books. 35

Wilkinson, M., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. O. Bonino da Silva Santos, P. Bourne, J. Bouwman, A. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. Evelo, R. Finkers, and B. Mons

2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3. 36

Williamson, D. and M. Goldstein

2015. Posterior belief assessment: Extracting meaningful subjective judgements

from Bayesian analyses with complex statistical models. *Bayesian Analysis*, 10. 140, 141, 142, 143, 156

Yakel, E., I. M. Faniel, and Z. J. Maiorana

2019. Virtuous and vicious circles in the data life-cycle. *Information Research*, 24. 39

Zaki, M. J. and W. Meira

2014. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. USA: Cambridge University Press. 128

Zeidler, J. A., C. E. Buck, and C. D. Litton

1998. Integration of archaeological phase information and radiocarbon results from the Jama River Valley, Ecuador: A Bayesian approach. *Antiquity*, 9:160–179. 67

Zhan, J., S. Gurung, and S. P. K. Parsa

2017. Identification of top-k nodes in large networks using katz centrality. *Journal of Big Data*, 4. 128

Appendix A

A.1 Sampling methods for approximating posterior distributions

Monte Carlo methods are sampling methods that allow the approximation of integrals by sampling from an appropriate function, which for large samples will approximate an integral of interest (Robert and Casella, 1999, Chapter 3). In Bayesian inference, this requires being able to sample from a function proportional to the posterior. However, as Ghosh et al. (2006) Chapter 7, Page 215 highlights often in Bayesian inference, we cannot explicitly state and sample from the exact form of the posterior. As a result, it is common practice to use a specific class of Monte Carlo methods called Monte Carlo Markov Chain (MCMC) methods instead. For detailed theoretical background on Markov chains, for the underlying theory of MCMC methods, see Robert and Casella (1999) Chapter 4 and Chapters 6-7, respectively.

One such method of MCMC simulation is Gibbs sampling. We demonstrate the

use numerical integration and Gibbs sampling to approximate two calibrated ages of interest below for users unfamiliar with Gibbs sampling.

A.2 Bayesian chronology construction with two parameters

Suppose we have a sample from each of context 1 and context 2, and these samples have true calibrated ages θ_1 and θ_2. Further, the samples from contexts 1 and 2 have radiocarbon determinations $x_1 \pm \sigma_1 = 1100 \pm 50$ and $x_2 \pm \sigma = 1000 \pm 50$, respectively. Prior information tells us that θ_1 is below θ_2 in the stratigraphic sequence and, therefore, $\theta_1 > \theta_2$ on the calibrated scale.

Therefore, our prior is

$$\mathbb{I}_C(\theta_1, \theta_2) = \begin{cases} 1 & \theta_1, \theta_2 \in C \\ 0 & \text{otherwise} \end{cases}. \tag{A.1}$$

such that $C = \{\theta_1, \theta_2 | \theta_1 > \theta_2\}$ and our posterior becomes

$$p(\theta_1, \theta_2 | x_1, x_2) \propto \mathbb{I}_C(\theta_1, \theta_2) \exp\left(-\frac{(x_1 - \mu(\theta_1))^2}{2(\sigma_1^2 + \delta(\theta_1)^2)}\right) \exp\left(-\frac{(x_2 - \mu(\theta_2))^2}{2(\sigma_2^2 + \delta(\theta_2)^2)}\right) \tag{A.2}$$

with normalising constant

$$\int\int \mathbb{I}_C(\theta_1, \theta_2) \exp\left(-\frac{(x_1 - \mu(\theta_1))^2}{2(\sigma_1^2 + \delta(\theta_1)^2)}\right) \exp\left(-\frac{(x_2 - \mu(\theta_2))^2}{2(\sigma_2^2 + \delta(\theta_2)^2)}\right) \mathrm{d}\theta_1 \mathrm{d}\theta_2, \tag{A.3}$$

which follows directly from Equation 5.6.

A.2.1 Numerical integration

Calculating the posterior exactly for the two-sample example requires calculating the integral outlined in Equation A.3. This can be done using numerical integration, which evaluates the integral for a given function over a set number of points to approximate the integral. The more points used, the more accurate the approximation (Gelman et al., 2004, Chapter 10). Using the `integrate` package within the `SciPy` library in `Python` (The SciPy community, 2023), we calculated our normalising constant. Following this, obtaining the probability distribution for a single parameter requires finding the posterior marginal distribution of that parameter. We calculate this by integrating the joint posterior distribution over all other parameters within the distribution. For θ_1 and θ_2 the posterior marginal distributions are equal to

$$p(\theta_1|x_1) = \int p(\theta_1, \theta_2|x_1, x_2)\mathrm{d}\theta_2 \tag{A.4}$$

$$p(\theta_2|x_2) = \int p(\theta_1, \theta_2|x_1, x_2)\mathrm{d}\theta_1 \tag{A.5}$$

respectively. Plots of exact posterior marginal distributions $p(\theta_1|x_1)$ and $p(\theta_2|x_2)$ can be seen in Figure A.1. However, as previously discussed, numerical integration becomes unreliable for larger dimensions. As a result, Gibbs sampling or other MCMC sampling methods are used to obtain posterior marginal distributions.

A.2.2 Sampling algorithm for a model with two parameters

Gibbs sampling comprises a sampling scheme in which we sample from the condition distributions of each parameter. A conditional distribution gives the probability distribution for a parameter, conditional on observed values of all other parameters

and data within the model. For θ_1 and θ_2, these are

$$p(\theta_1|\theta_2, x_1) = \mathbb{I}_{C_1}(\theta_1) \exp\left(-\frac{(x_1 - \mu(\theta_1))^2}{2(\sigma_1^2 + \delta(\theta_1)^2)}\right) \tag{A.6}$$

$$p(\theta_2|\theta_1, x_2) = \mathbb{I}_{C_2}(\theta_2) \exp\left(-\frac{(x_2 - \mu(\theta_2))^2}{2(\sigma_2^2 + \delta(\theta_2)^2)}\right) \tag{A.7}$$

respectively, such that $C_1 = \{\theta_1|\theta_2 < \theta_1 < 50,000\}$ and $C_2 = \{\theta_2|0 < \theta_2 < \theta_1\}$. By iteratively sampling from the conditional distributions of the parameters for a large number of iterations, we obtain an approximation to the marginal posterior distributions for our parameters of interest.

To carry out Gibbs sampling, we need initial values for our parameters. For our current example this requires initial values $\theta_1^{(0)}$ and $\theta_2^{(0)}$ that satisfy the condition $\theta_1 > \theta_2$. These are usually sampled from the respective likelihood of the parameters subject to ordering, which ensures the starting values for each parameter are plausible given our data. To remove the influence of the initial values on our final samples, we discard the first b samples, called a burn-in period (see Gelman et al., 2004, Chapter 11).

Algorithm 1 Gibbs sampler for estimating two true calibrated ages

1: **sample** $\theta_1^{(0)}$ and $\theta_2^{(0)}$
2: **select** sample size N
3: **for** k in 1 to N **do**
4: **sample** θ_1^k from $p(\theta_1|\theta_2^{k-1}, x_1)$
5: **sample** θ_2^k from $p(\theta_2|\theta_1^k, x_2)$
6: **end for**
7: **select** burn in period b
8: **return** last $b + 1$ to N samples

We obtained approximations to the marginal posteriors for θ_1 and θ_2 using the Gibbs sampler described in Algorithm 1. The results of which can be seen in

Figure A.1, demonstrating that the Gibbs sampler (which is easily extended to large numbers of parameters unlike numerical integration) very closely estimates the marginal posteriors of θ_1 and θ_2 (see yellow histograms in Figure A.1) in comparison to the results from numerical integration (see blue curves in Figure A.1).

A.3 Metropolis Hastings algorithm

Metropolis-Hastings algorithms are also a type of MCMC simulation, and are a form of rejection sampling which allow us to form a sequence of random samples which, for each parameter, approximates the marginal posterior. We add to the sequence for each parameter by sampling values for that parameter from a proposal distribution (an appropriately chosen probability distribution) and then accepting or rejecting the sampled value based on a pre-defined acceptance probability (see Robert and Casella, 1999, Chapter 6 Section 6.2) This process is repeated, creating a sequence of accepted random samples for each parameter. For a large enough number of samples, given appropriate starting values for each parameter, this will provide posterior estimates for our parameters. Metropolis-Hastings algorithms have many variations, and we now outline the sampling process for the specific algorithm we use in this research which closely follows the pseudocode outlined in Nicholls and Jones (2001).

The Metropolis-Hastings algorithm we use, as detailed by Nicholls and Jones (2001), consists of four update steps, with distinct proposal distributions q_1 and q_2 for steps 1 and 2 and steps 3 and 4 consist of translating and scaling the previous samples on the calibrated scale, respectively. Steps 3 and 4 improve the acceptance rate of the algorithm overall. Let N_p denote the minimum number of samples accepted for a parameter p in the combined set of all parameters $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and t denote

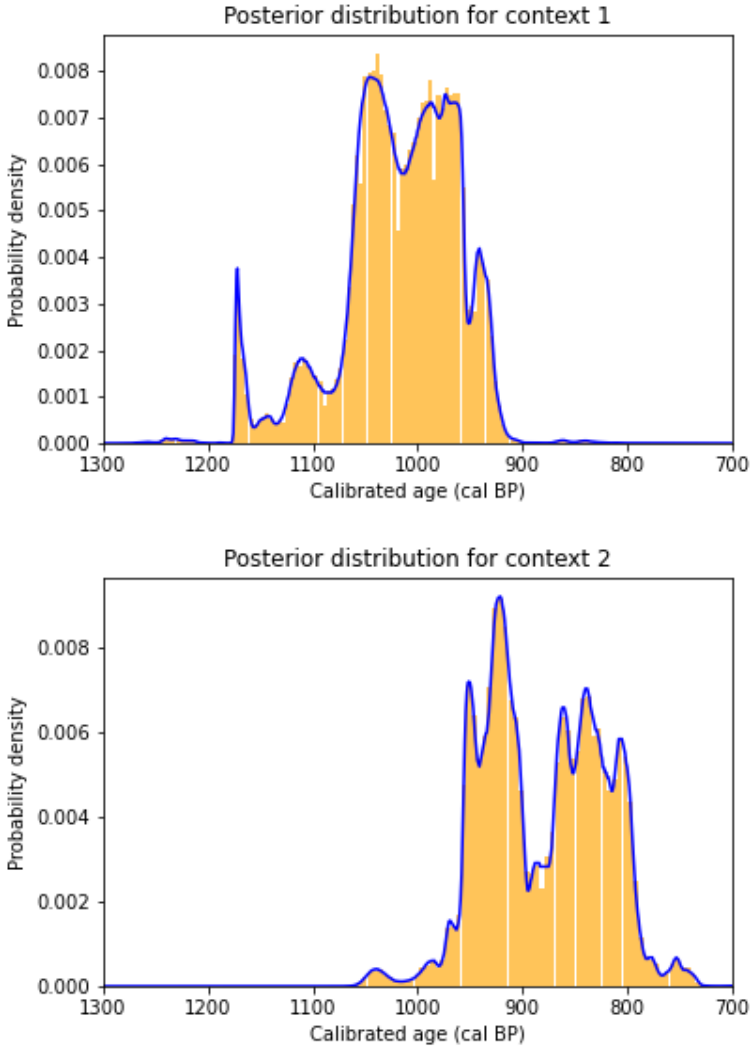


Figure A.1: Marginal posterior probability density plots for the true calibrated ages of samples found in contexts 1 and 2, with radiocarbon determinations of 1100 ± 50 and 1000 ± 50, respectively. The blue curves show exact calculations of the marginal posterior distributions using numerical integration, and the yellow histograms show the same marginal posterior densities that were obtained using Gibbs sampling.

the number of samples generated in total. In addition, let $l_{\theta_{i,j}}$ and $u_{\theta_{i,j}}$ be the sets of θ in $\boldsymbol{\theta}$ immediately younger and older than $\theta_{i,j}$, respectively, according to $\mathbb{I}_C(\boldsymbol{\theta})$. Similarly, let $l_{\alpha_{i,j}}$ and u_{α_j} represent the set of parameters within the model

that are directly younger and older than α_j respectively, according to $\mathbb{I}_G(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\mathbb{I}_{(\beta_j, \alpha_j)}(\theta_{i,j})$.

Following this notation, the complete algorithm for running the initial steps before the Metropolis-Hastings algorithm is provided in Algorithm 2. The four sampling steps for the algorithm itself (following closely the algorithm for models with only grouping and no stratigraphy introduced by Nicholls and Jones (2001)) are provided in algorithms 2-6, with acceptance probabilities a_v and proposal distributions q_1 and q_2 are defined.

A.4 Pseudo-code for Metropolis-Hastings algorithm

Algorithm 2 Initial steps for Metropolis-Hastings algorithm

$A = \max(\boldsymbol{x}) + 10 \cdot \max(\boldsymbol{\sigma})$
$P = \min(\boldsymbol{x}) - 10 \cdot \max(\boldsymbol{\sigma})$
$R = P - A$
for $1 < j < J$, $1 < i < N_j$ **do**
  sample $\theta_{i,j}^{(0)}$ from $L(\theta_{i,j}; x_{i,j})\mathbb{I}_C(\boldsymbol{\theta})$
end for
for $j = 1$ **do**
  sample $\alpha_j^{(0)}$ uniformly at random in $[P,\ \max(\theta_{i,j}^{(0)})]$
  sample $\beta_j^{(0)}$ uniformly at random in $[\min(\theta_{i,j}^{(0)}),\ \max(\theta_{i,j+1}^{(0)}, A)]$
end for
for $j = J$ **do**
  sample $\alpha_j^{(0)}$ uniformly at random in $[\min(\theta_{i,j-1}^{(0)}, P),\ \max(\theta_{i,j}^{(0)})]$
  sample $\beta_j^{(0)}$ uniformly at random in $[\min(\theta_{i,j}^{(0)}), A]$
end for
for $2 < j < J - 1$ **do**
  sample $\alpha_j^{(0)}$ uniformly at random in $[\min(\theta_{i,j-1}^{(0)}, P),\ \max(\theta_{i,j}^{(0)})]$
  sample $\beta_j^{(0)}$ uniformly at random in $[\min(\theta_{i,j}^{(0)}),\ \max(\theta_{i,j+1}^{(0)}, A)]$
end for

Algorithm 3 Sampling step 1

select $\theta_{i,j}$ uniformly at random in $\boldsymbol{\theta}$
q_1: **sample** $\theta^*_{i,j}$ uniformly at random in $\left[\max\left(l^{(t-1)}_{\theta_{i,j}}, \beta^{(t-1)}_j\right),\ \min\left(u^{(t-1)}_{\theta_{i,j}}, \alpha^{(t-1)}_j\right)\right]$
$\boldsymbol{\theta}^* = \left\{\theta^{(t-1)}_{1,1}, ...\theta^*_{i,j}, ...\theta^{(t-1)}_{J,N_j}\right\}$
sample $h \sim U(0,1)$
$a_1 = \min\left\{1, \dfrac{p(\boldsymbol{\theta}^*, \boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}|\boldsymbol{x})}{p(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}|\boldsymbol{x})}\right\}$
if $a_1 > h$ **then**
    $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$
    $N_s = N_s + 1$
else
    $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$
end if
$\boldsymbol{\alpha}^{(t)} = \boldsymbol{\alpha}^{(t-1)}$
$\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$
$t = t + 1$

A.5 Convergence checks

When using MCMC algorithms such as Metropolis Hastings, it is important to check the output for convergence. This can be done manually by plotting sample values for a given parameter on the vertical axis and iteration number t on the horizontal axis. If the sample for a parameter has converged to its marginal posterior distribution, we expect to see a plot showing random movement around a central value (Spade, 2020). An example of a trace plot that shows convergence is provided in Figure A.2.

Given that trace plots have to be manually inspected, we also utilised other convergence checks that could be easily automated within our MCMC algorithm. The first convergence check we use when running the algorithm is, for every $10,000$ samples (i.e. every $10,000^{th}$ value of t), we calculate the acceptance rate N_p/t for all parameters to ensure that for all parameters, the acceptance rate is greater than 0.1. This

Algorithm 4 Sampling step 2

select α_j or β_j uniformly at random in $\boldsymbol{\Phi} = \boldsymbol{\alpha} \cup \boldsymbol{\beta}$
if α_j selected **then**
    q_2: **Sample** α_j^* uniformly at random in $\left[\max\left(l_{\alpha_j}^{(t-1)}\right),\ \min\left(u_{\alpha_j}^{(t-1)}\right)\right]$
    $\boldsymbol{\alpha}^* = \left\{\alpha_1^{(t-1)}, ...\alpha_j^*, ...\alpha_J^{(t-1)}\right\}$
    Sample $h \sim U(0,1)$
    $a_2 = \min\left\{1, \dfrac{p(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^{(t-1)}|\boldsymbol{x})}{p(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}|\boldsymbol{x})}\right\}$
    if $a_2 > h$ **then**
        $\boldsymbol{\alpha}^{(t)} = \boldsymbol{\alpha}^*$
        $N_s = N_s + 1$
    else
        $\boldsymbol{\alpha}^{(t)} = \boldsymbol{\alpha}^{(t-1)}$
    end if
end if
if β_j selected **then**
    q_2: **sample** β_j^* uniformly at random in $\left[\max\left(l_{\beta_j}^{(t-1)}\right),\ \min\left(u_{\beta_j}^{(t-1)}\right)\right]$
    $\boldsymbol{\beta}^* = \left\{\beta_1^{(t-1)}, ...\beta_j^*, ...\beta_J^{(t-1)}\right\}$
    sample $h \sim U(0,1)$ $a_2 = \min\left\{1, \dfrac{p(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\beta}^*|\boldsymbol{x})}{p(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}|\boldsymbol{x})}\right\}$
    if $a_1 > h$ **then**
        $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^*$
        $N_s = N_s + 1$
    else
        $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$
    end if
end if
$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$
$t = t + 1$

Algorithm 5 Sampling step 3

sample s uniformly at random in $[0,\ \max(\boldsymbol{\sigma})]$
$\boldsymbol{\alpha}^* = \{\alpha + s \text{ for all } \alpha \in \boldsymbol{\alpha}^{(t-1)}\}$
$\boldsymbol{\beta}^* = \{\beta + s \text{ for all } \beta \in \boldsymbol{\beta}^{(t-1)}\}$
$\boldsymbol{\theta}^* = \{\theta + s \text{ for all } \theta \in \boldsymbol{\theta}^{(t-1)}\}$
sample $h \sim U(0, 1)$
$a_3 = \min\left\{1, \dfrac{p(\boldsymbol{\theta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^* | \boldsymbol{x})}{p(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\beta}^{(t-1)} | \boldsymbol{x})}\right\}$
if $a_3 > h$ **then**
    $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$
    $\boldsymbol{\alpha}^{(t)} = \boldsymbol{\alpha}^*$
    $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^*$
    $N_s = N_s + 1$
else
    $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$
    $\boldsymbol{\alpha}^{(t)} = \boldsymbol{\alpha}^{(t-1)}$
    $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$
end if
$t = t + 1$

threshold is slightly lower than the recommended threshold of 0.234 (Bédard, 2008). However, for models with many parameters, the acceptance rate lowers, and manual testing (checking trace plots) showed an acceptance rate of 0.1 to be sufficient for models of this size. If the acceptance rate is lower than 0.1, the code halts the MCMC algorithm and restarts the algorithm with new initial values for the parameters. Again, testing showed this to improve the acceptance rate effectively.

Once $50,000$ accepted samples have been obtained for all parameters, we re-run the MCMC algorithm again for the same model and use the two sequences of random samples obtained for each parameter to calculate the Gelman-Rubin convergence diagnostic, accepting a value lower than 1.1 as an indication of convergence which is the threshold defined in Gelman and Rubin (1992). (see Gelman and Rubin, 1992, Section 2.2 specifically for details of how to calculate the diagnostic value).

Algorithm 6 Sampling step 4

sample ρ uniformly at random in $[2/3, 3/2]$
$\boldsymbol{\alpha}^* = \{\rho\alpha - (\rho-1)av(\boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\theta}^{(t-1)}) \text{ for all } \alpha \in \boldsymbol{\alpha}^{(t-1)}\}$
$\boldsymbol{\beta}^* = \{\rho\beta - (\rho-1)av(\boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\theta}^{(t-1)}) \text{ for all } \beta \in \boldsymbol{\beta}^{(t-1)}\}$
$\boldsymbol{\theta}^* = \{\rho\theta - (\rho-1)av(\boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\theta}^{(t-1)}) \text{ for all } \theta \in \boldsymbol{\theta}^{(t-1)}\}$
sample $h \sim U(0,1)$
$a_4 = \min\left\{1, \frac{L(\boldsymbol{\theta}^*; \boldsymbol{x})}{L(\boldsymbol{\theta}^{(t-1)}; \boldsymbol{x})} \times \frac{R - (\max(\boldsymbol{\alpha}^*) - \min(\boldsymbol{\beta}^*))}{R - \rho(\max(\boldsymbol{\alpha}^{(t-1)}) - \min(\boldsymbol{\beta}^{(t-1)})} \times \frac{1}{\rho}\right\}$
if $a_4 > h$ **then**
    $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$
    $\boldsymbol{\alpha}^{(t)} = \boldsymbol{\alpha}^*$
    $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^*$
    $N_s = N_s + 1$
else
    $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$
    $\boldsymbol{\alpha}^{(t)} = \boldsymbol{\alpha}^{(t-1)}$
    $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$
    $N_s = N_s + 1$
end if
$t = t + 1$

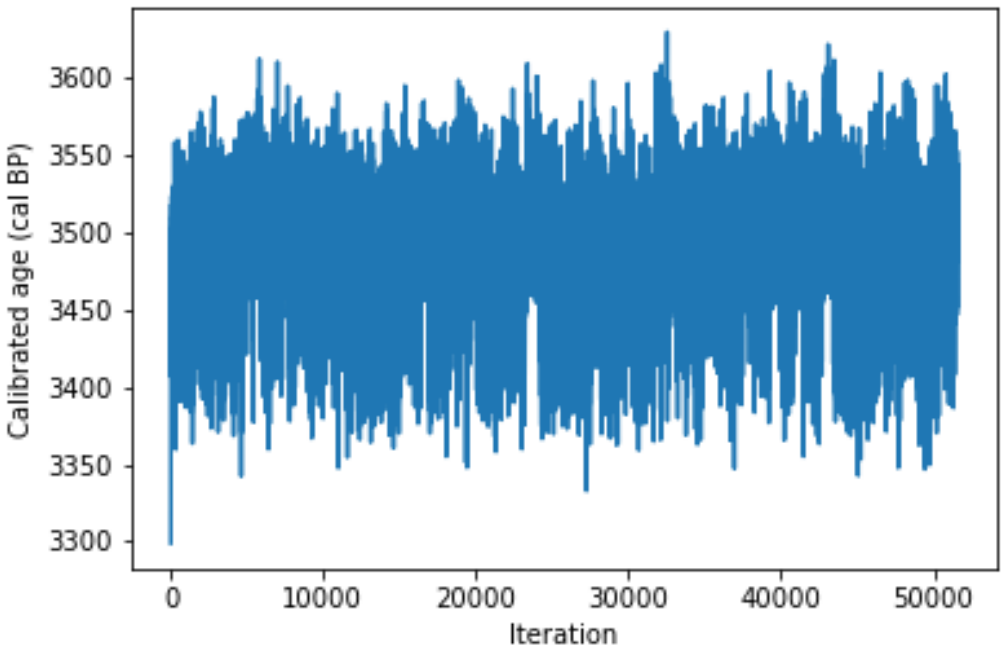


Figure A.2: An example of a trace plot for the sample of an arbitrary parameter obtained using a Metropolis-Hastings algorithm, which shows convergence.

Appendix B

B.1 Link to PolyChron video

The following link provides a demonstration of the use of `PolyChron` version 1.0, to aid in the examination of this thesis.

https://youtu.be/2IrcGBix6I8

B.2 Required format of input files

`PolyChron` requires four files as input as a minimum to allow users to construct a chronological model. We outline the file structure for each file below:

above	below
814	758
1235	814
358	1235
813	358
1210	813
358	493
923	925
358	923
358	1168
758	
925	
493	
1168	

Table B.1: Example of stratigraphic relationships input file. Contexts in the left column are stratigraphically above the contexts in right.

context	date	error
758	3275	75
814	3270	80
1235	3400	75
493	3190	75
925	3420	65
923	3435	60
1168	3160	70
358	3340	85
813	3270	75
1210	3200	70

Table B.2: Example of radiocarbon determination input file. Column 1 contains context labels, column 2 contains radiocarbon determination estimates and column 3 contains the corresponding laboratory error for each estimate.

context	group
758	1
814	1
1235	1
493	1
925	1
923	1
1168	1
358	2
813	2
1210	2

Table B.3: Example of context grouping input file. Column 1 contains context labels and column 2 contains each contexts corresponding group.

above	below
2	1

Table B.4: Example of group ordering input file. Column 1 contains groups that are older than the corresponding group in context 2.

B.3 Simulating radiocarbon determinations

To simulate radiocarbon determinations for n contexts within a chronological model, we utilised the following algorithm:

Algorithm 7 Initial steps for Metropolis-Hastings algorithm

sample n calibrated ages uniformly at random between A and P
sort the n calibrated ages into descending order
sort the n contexts using a topological sorting algorithm

for each calibrated age in ascending order **do**
    Back-calibrate each calibrated age to the radiocarbon scale to obtain a simulated radiocarbon estimate
    Add an error of 50 BP
end for

B.4 Katz centrality and variants of this measure

Katz centrality is a measure used in graph theory to quantify the importance of the nodes within a mathematical graph. Centrality measures are determined by both the number of nodes a given node is connected to and the importance of those nodes.

The Katz centrality measure can be expressed using the following formula:

$$C(i) = \alpha \sum (A(i,j)C(j)) + \beta$$

such that:

- $C(i)$ is the Katz centrality score for node i,
- $A(i,j)$ is the element in the adjacency matrix of the graph. It takes the value of 1 if there is an edge between nodes i and j and 0 otherwise,
- $C(j)$ is the Katz centrality score for node j, which is a neighbour of node i.
- α is called the attenuation factor and it reduces the weight of the centrality the further away we are in the graph from the node of interest
- β is a constant term which ensures that nodes in a directed graph with no incoming edges are not given a score of 0.

This formula is iteratively applied, with all nodes being given an initial centrality score of 1 and will eventually converge to the centrality score for each node in the network. The centrality scores reflect the relative importance of nodes, with higher scores indicating higher importance within the graph.

B.5 Metrics for measuring overlap of two probability distributions

Below are four measures for measuring the difference between two distributions p and q for parameter θ.

Kullback–Leibler divergence:

$$\int p(\theta) \log \left(\frac{p(\theta)}{q(\theta)} \right) d\theta$$

Bhattacharyya distance:

$$\int \sqrt{(p(\theta)q(\theta)} d\theta$$

Hellinger distance:

$$\frac{1}{\sqrt{2}} \sqrt{\int \left(\sqrt{p(\theta)} - \sqrt{q(\theta)} \right)^2 d\theta}$$

Overlapping metric proposed by Pastore and Calcagnì (2019):

$$\int \min(p(\theta), q(\theta)) d\theta$$