RADIOCARBON, Vol 53, Nr 4, 2011, p 559–562 © 2011 by the Arizona Board of Regents on behalf of the University of Arizona

MODELS, DATA, STATISTICS, AND OUTLIERS—A STATISTICAL REVOLUTION IN ARCHAEOLOGY AND ^{14}C DATING

E Marian Scott

School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QW, United Kingdom.
Email: marian.scott@glasgow.ac.uk.

ABSTRACT. Increasingly, the uses of data are becoming more and more sophisticated as the archaeological and chronological questions being asked become more complex. Statistical models and tools for inference are a routine part of an archaeological investigation encouraged through the availability of software, and with each release of that software, additional functionality is being added. This comes with enormous benefit but also at a cost—the dreaded *black box*. Therefore, this article, as the first in a series of short articles, will attempt to cover some of the things one needs to know to make the most of the power of the statistical revolution, while avoiding the pitfalls.

INTRODUCTION

Increasingly, the modern world of archaeological science is making more and more use of models, and the basis of our understanding of past cultures and past environments is becoming more and more quantitative. Whenever we measure features or attributes of an artifact or study a set of artifacts, or if we are interested in creating a chronology, then we create numerical information. If we are in the fortunate position of having a number of such artifacts (e.g. pots from a single site or several sites) or a series of dates, then we need to have tools to summarize the information that we have collected and which then can be generalized to the wider population (the process of statistical inference using models). In this introductory paper to the series, we discuss some of the commonly encountered quantitative issues in dating and subsequent modeling.

WHAT IS A STATISTICAL MODEL?

"*All models are wrong, some are more useful than others*" G E P Box

A model is a conceptual description of the processes that generate the data we observe. It is then formulated mathematically in a series of equations, and which requires us to estimate the parameters that appear in these equations using the data. A statistical model will also include a random component. In recent years, within archaeology there has been a major adoption of Bayesian models, characterized by the formal inclusion of the concept of "prior" knowledge, or beliefs (Bayliss 2009). A Bayesian model has 3 components; formally, they are described as the **prior**, the **likelihood**, and the **posterior**. Bayes' theorem is used to define the posterior in terms of the likelihood (involving the data) and the prior (expressed in terms of the model parameters).

But not all models need to be Bayesian. Modeling is required partly because it is becoming increasingly uncommon that single, one-off samples are being dated. Users (^{14}C) want high precision (often greater than is achievable in the laboratory), so different dating strategies are being developed and the power of models is being used to help deliver the desired precision. Additionally, users often have qualitative or semi-quantitative information about relative relationships between samples or the events they relate to and expectations of the "ages," and it is this extra information that makes Bayesian modeling so persuasive, offering many benefits.

WHAT ABOUT THE DATA?

All data are uncertain; some are more uncertain than other. The data (e.g. our ^{14}C dates) form part of our evidence base. They are characterized by variability, or in other words, all measurement is sub-

https://doi.org/10.1017/S0033822200039035 Published online by Cambridge University Press

ject to uncertainty, which can be considered to have several components, one resulting from the analytical measurement procedure and the other arising from the sample selection process. The uncertainty of a measurement describes the spread of values that might have been observed were the measurement repeated under *identical* conditions. Outliers or "odd" values are troublesome, but they often are an entirely natural manifestation of the variability in our measurements—how we recognize and deal with them, however, can present a challenge.

First, it is important to consider the underpinning nature of measurement and what the commonly quoted ^{14}C error represents. Analytical uncertainty reflects that every time a ^{14}C measurement is made (under identical conditions), the result is different (Scott et al. 2007). A ^{14}C quoted error fundamentally represents the analytical uncertainty; it describes the spread of values that might be observed were the measurement repeated. However, many laboratories also introduce additional sources of uncertainty in the final quoted error, and many also provide experimental verification of the quoted uncertainty based on repeated measurements (replication) of a standard or reference material. We can expect then that in our final conclusions, our interpretation must reflect this uncertainty, and that there will be a language of uncertainty expressed in terms of confidence or probability.

TWO EXAMPLES

The 2 examples below give a flavor of the nature and complexity of the archaeological questions and the data being used to answer them.

1. The Scythians (Zaitseva et al. 2004)

An expedition from the Central Asiatic Archaeology Department (K V Chugunov) of the Hermitage Museum in St Petersburg and the Eurasian Department of German Archaeological Institution (Prof G Parzinger and A Nagler) discovered the unique royal burial mound Arzhan-2, grave 5 in 2001. The grave had not been robbed; there were 2 buried people dressed in richly decorated clothes with accompanying golden artifacts in typical Scythian animal style.

The objectives of the ensuing investigation were

- to date this new monument;
- to integrate the new monument into the chronology that existed for Eurasian Scythian monuments.

The data available to the study were

- ^{14}C measurements (on artifacts and some of the logs that were used in the burial chamber construction);
- stylistic information on the type of burial, animal designs on the artifacts.

2. The ARCANE Project (2011)

One of the objectives of this large pan-European project was

- to define a chronology that links the important regional sites in the Near East.

The data available included

- ^{14}C measurements from artifacts from excavations at different sites, often with associated stratigraphy, measured in different laboratories, but also some new measurements;
- historical records, king lists from stone tablets, etc.

https://doi.org/10.1017/S0033822200039035 Published online by Cambridge University Press

Both examples illustrate well the nature of the objectives, specifically describing temporal relationships and constructing chronologies. They also reflect the varied nature of both the qualitative and quantitative data available.

WHAT CAN GO WRONG IN GENERAL?

There are many things that can go wrong or disappoint in multidisciplinary scientific endeavors, such as creating a chronology for a site or a region. Starting from the early stages, the first potential difficulty is that the ^{14}C measurements and modeling results do not match up or agree with the expectations. We can consider why this might happen.

Problem 1: "Poor" samples, not associated with the events of interest (uncertainty over archaeology or stratigraphy...);

Problem 2: Expectations are not well founded (*poor* archaeological model used to formulate the statistical model);

Problem 3: Measurements are insufficiently accurate or precise (*poor* measurements). We are not able to answer the archaeological question (e.g. to answer the questions requires us to pin down an event to within 25 yr and this may not be possible), or ages are inconsistent with archaeological expectations.

HOW CAN WE DEAL WITH THESE DIFFICULTIES?

All 3 problems are intrinsically linked in different ways. There may be difficulties with a measurement once the sample reaches the laboratory, but laboratories are concerned about the quality of their measurements and will have procedures in place to ensure that good practices are followed (Scott et al. 2010). On the other hand, there may be perfectly sound measurements that are inconsistent with the archaeological expectations, and in the same way archaeologists will try to ensure that the sample quality and context is sound and as tightly controlled as possible.

Moving on, we may encounter situations where the model results are unexpected, difficult to reconcile with our perceived wisdom, or simply do not deliver the required precision. It might be possible to try another model, but one must remember, all models are wrong, just some are better/more useful than others. It is always worthwhile to consider the evidence or scientific basis for the model. Simple is good, but sometimes we need more complex structures. A sensitivity analysis to explore how one's results are affected by changes in the model or in the data can often be very informative. However, the key point is that one should keep testing and challenging one's understanding.

CONCLUSIONS

Our view of the past can only be seen hazily; there is uncertainty and to gain precision and a clearer view, we need to introduce other information. Bayesian chronology construction provides a framework to incorporate "extra" information describing the relationships in a series of results with potential precision gains for the user. However, there are challenges in how to describe the nature of the relationship between the samples/events and its representation in a mathematical form. As users/consumers, we need to understand the implications of the choices we make, and have an understanding even at a basic level of the sophisticated computer codes that exist and what they do and of the nature of the data.

https://doi.org/10.1017/S0033822200039035 Published online by Cambridge University Press

In the forthcoming series of short articles, we will cover the sorts of problems that commonly crop up and their statistical nature, including variability and distributions; outliers and unusual observations; summary statistics including means, medians, standard deviations, weighted means, and robustness; errors and uncertainties; inferential tools, regression, and calibration; and finally, Bayesian methods.

ACKNOWLEDGMENTS

This paper is based on a plenary lecture given at the Radiocarbon and Archaeology conference in Cyprus, April 2011. I would like to thank the organizers for the opportunity to deliver that lecture.

REFERENCES

ARCANE. 2011. ARCANE Project: Synchronizing Cultures and Civilizations of the Ancient Near East and the Eastern Mediterranean in the Third Millennium BC. URL: http://www.arcane.uni-tuebingen.de/welcome.html. Accessed October 2011.

Bayliss A. 2009. Rolling out revolution: using radiocarbon dating in archaeology. *Radiocarbon* 51(1):123–47.

Scott EM, Cook GT, Naysmith P. 2007. Error and uncertainty in radiocarbon measurements. *Radiocarbon* 49(2):427–40.

Scott EM, Cook GT, Naysmith P. 2010. The Fifth International Radiocarbon Intercomparison (VIRI): an assessment of laboratory performance in Stage 3. *Radiocarbon* 52(3):859–65.

Zaitseva GI, Chugunov KV, Dergachev VA, Nagler A, Parzinger G, Scott EM, Sementsov AA, Vasiliev S, van Geel B, van der Plicht J, Lebedeva LM. 2004. Chronological studies of the Arzhan-2 Scythian monument in Tuva (Russia). *Radiocarbon* 46(1): 277–84.

https://doi.org/10.1017/S0033822200039035 Published online by Cambridge University Press