

Reports from the Environmental Archaeology Unit, York 99/11, 6 pp.

Looking a gift horse in the mouth: the problems of using archived and published biometrical data.

by

Cluny Johnstone

Summary

A brief overview is presented of some of the problems associated with using archived and published biometrical data. Many of the problems were encountered whilst undertaking an undergraduate dissertation based on biometrical data for archaeological horses. Discussions with colleagues brought to light other problems relating to archived and published data. Locating sources, obtaining the data from them and actually making use of those data are considered here. There are problems associated with publications, archives, 'grey literature', microfiches, summary tables and unclear or absent methods statements. These problems excluded many potentially useful and interesting datasets from further analysis, lessening the value of undertaking a synthetic project of this type. In conclusion, some ways in which the problems may be obviated in the future are suggested.

KEYWORDS: PRE-PUBLICATION DRAFT; BIOMETRICAL DATA; ARCHIVES; PUBLISHED DATA; METHOD STATEMENTS

Author's address:

Palaeoecology Research Services
Environmental Archaeology Unit
Dept. of Biology
University of York
PO Box 373
Heslington
York YO10 5YW

Telephone: (01904) 434474
Fax: (01904) 433850

27th January 1999

Looking a gift horse in the mouth: the problems of using archived and published biometrical data.

Introduction

The old adage ‘not looking a gift horse in the mouth’, in case you discover it is an old nag, began to seem strangely appropriate whilst undertaking an undergraduate dissertation on horses from archaeological sites. The dissertation was based on gathering biometrical data from a wide variety of periods, and this paper highlights various problems encountered whilst trying to locate, obtain and use them. As will be seen, the problems encountered led to a number of potentially useful and interesting datasets being, in effect, ‘useless old nags’ for one or more reasons. The aim of this paper is not to point the finger of blame but to try and bring to the attention of those who publish or archive data the needs of others who may wish to use the data at a later date

Finding data

The first problem is that of locating data. The initial aim of the dissertation was to use all the available information, both published and archived, but it soon became apparent that obtaining data from archive sources would be an extremely time-consuming process. A large number of publication reports did not contain the raw data required for synthesis and analysis of measurements. In many cases these reports failed to state where the archives were kept, or if any actually existed, which for many older reports is perhaps unlikely.

Archived sources

The first point to be made here is that when archives are discussed, it is usually the paper or electronic records that are referred to, ignoring the fact that the primary archive is the bones themselves. Very rarely do reports state where the bone material (or any other finds) is to be kept, or even if the material is kept at all after a publication relating to it has been written. This information is crucial since researchers have to go back to the bones themselves when paper or electronic archives are either missing or unintelligible.

Having expended much time and effort locating an archive, what state is it in? Paper and/or electronic archives may exist for a given dataset but are they in a form from which information can be extracted? If abbreviations have been used during the recording process, have notes of what these mean been deposited along with the records? If the meaning of abbreviations has not been recorded then effectively the data have also been lost.

It should also be remembered that paper records may become illegible with time if good quality paper and permanent ink are not used. Electronic archives may also become unreadable through corruption of disks and also by the advance of technology. Data may already have been lost through the use of 5 ¼” disks or old mainframe systems for which the technology to retrieve the data has become unavailable.

Published sources

For many published sites, the only route to obtaining data from archives involved approaching the publishers to obtain authors' addresses - potentially an extremely lengthy process. Even this long-winded way of obtaining data is not available for the area of semi-published 'grey literature'. In effect, this means that the data are lost to most people wishing to them at a later date.

Where data are published, they are is very often presented in the form of summary tables containing the minimum and maximum values, number of cases, mean and standard deviation. This is particularly frustrating from the point of view of a future researcher, it's like 'putting a carrot in front of a donkey': the table indicates that numerous measurements were taken but there is no way of getting to the raw data from these tables.

Publications that contain raw data confine them either to an appendix, which is the best solution, or put them on to microfiche. Although useful up to a point, microfiche is not really an acceptable solution as many people don't have access to a microfiche reader and even less have access to a machine for enlarging and printing microfiche to paper (mainly restricted to those with access to university libraries). Another problem with microfiche is that with repeated use the fiches become very scratched making them very difficult to read (a small scratch on a fiche becomes a gigantic blur when blown up 20-25 times to a readable size). All this presupposes that the microfiche you require is still in the pocket at the back of the book! It seems that the problems with microfiche are now being taken on board since, after the initial enthusiastic rush of publications containing

microfiche, the numbers appears to have trailed off in recent years.

The other point to make about publications is that, if we are being brutally honest, only fellow specialists read specialist reports in detail, so why not include data of use to future research in a user-friendly format. The limitations of allotted publication space are well known. However, as far as I am aware, appendices are not usually as restricted as the main text if it is specified one is needed at the beginning of a project. However, a word of caution on this subject; using an incredibly tiny, and hence unreadable, font size is not really an acceptable solution to space problems either.

Using data

After overcoming all the obstacles above and having located raw measurement data, there are still a number of points to consider before using them. The most crucial question i.e. how were the measurements taken, should be included in the methods statement or main text in a clear enough manner to enable other workers to replicate the measurements. If von den Driesch's (1976) guidelines were used this should be specified, preferably at the top of the table of measurements as well as in the methods section of text. When non standard measurements are used, illustrations and a full description should be provided.

Problems that can occur when inadequate descriptions are given have been illustrated recently when using shaft diameter measurements on tibiae. Data from two separate reports were used and both stated that SD was the shaft diameter without specifying how the measurement was taken. Von den Driesch indicates this is measured in the medio-lateral plane of the bone, but other

workers use the anterior-posterior plane (which is narrower in most mammals). Combining the two datasets with different ways of measuring the shaft diameter produced false results showing very convincing groups of animals of approximately the same height, but one set with gracile limbs and the other with more robust limbs (Figure 1.).

Trusting other people

This next question is guaranteed to be controversial but needs to be made none the less - Can we trust the measurements taken by other workers? The implication of this are not quite as drastic as it may at first seem, and can be clarified into a number of issues. Firstly, were the measurements being taken with acceptable accuracy? In other words, were callipers with at least 0.1 mm graduations used. Secondly, is the person interpreting specific measurement protocols in exactly the same way you are? Thirdly has a clear enough description of non-standard measurements been given to enable another worker to follow the procedure exactly?

Whilst an undergraduate, an experiment was conducted during a laboratory session to show inter- and intra-observer error. A class of approximately 35 students were each asked to measure a single bone five times, showing that a fairly limited degree of intra-observer error was possible. However, when all 35 averages were compared, a much greater degree of inter-observer error was noted between the measurements. This has many implications for future data users and should be borne in mind when taking measurements.

For the dissertation mentioned earlier, a large number of calculations of withers heights were

made, and the accuracy of the measurements became a crucial issue. It was interesting to note that a difference of 2 mm in the greatest length of a horse metacarpal could make as much as 25 mm difference in the estimated withers height - which could mean the difference between an individual being interpreted as a horse or a pony.

Suggestions for the future

A number of small adjustments in working practice, and in particular the way reports are structured, could make an enormous difference to workers trying to use data at a later date, and this could be achieved without expending much extra time. We should all be using data from other sites for comparative purposes, and it would be so much easier and less time consuming to do this if the data were included in the reports in the first place. The amount of time taken to construct a table of measurements, and thus the cost of doing so, is less in the long run than having to track down the data you require from an author or institution.

If the data are not to be included in the report, at least stating the institution where the work took place and/or authors' addresses, would give data-gatherers a starting point. In fact it should be standard practice always to include an address to which enquiries about a report can be sent, even for 'glossy' publications and books. Zooarchaeological archives should ideally be located with both the main archaeological archive and the author/s of the report so that enquiries addressed to either place can be answered satisfactorily.

The problem of how to publish raw data is most easily solved by the use of tables in an appendix. Microfiche is not accessible to so many people that it is not really an

acceptable solution. If a summary table has to be included, it should be in addition to the raw data.

Including an adequate description of how the measurements have been taken is absolutely vital and cannot be expressed too strongly. Figure 1. illustrates the kind of false results, and hence interpretation, that could result from a misunderstanding of what measurement has been taken. Including this kind of information, as well as in the method statement, with the appendix of raw data reduces the chances of misinterpretation for those who merely photocopy the part of the report they require.

Two additional data storage methods, CD ROM and the Internet, have become more widely available and cost effective in the last few years, and offer previously unavailable data storage and dissemination pathways. CD writers have now come within the price range of most small organisations and even individuals, so the ability to store and disseminate information has been increased. However, it should be borne in mind that these may become unavailable at some point just as 5 1/4" disks have.

Internet access has also become much more widespread and the introduction of Internet journals offers a new means of allowing others to access your data. The space available for publishing data on the web is not restricted in the same way that paper journals are, so it is feasible to publish entire datasets along with the article analysing them.

At some point in the future a central national repository for archaeological data on the Internet would be a very useful tool. The Archaeology Data Service (based in the Department of Archaeology, University of York) provides just such a service. However, it is not compulsory to deposit archives with

them so the range of data available at present is limited (and it has not been running for long). This service is offered as a possible way forward to allow wider dissemination of information and access to that information. The centralised deposition of information is unlikely to happen in the near future, if ever! It would be useful if those people who write specifications for archaeological interventions were to include a statement about central archiving as a means of accomplishing this. A starting point would be for all state-funded projects to be centrally archived rather than dispersed around the country, and perhaps then others would follow suit.

Conclusions

The aim of this paper was to provoke thought amongst fellow workers just how accessible our data are to each other. I hope these thoughts may lead to some consideration of how in future practices can be changed so as to ensure that it is easier to access biometrical data in the future.

References

Driesch, A. Von den. 1976. A guide to the measurement of animal bone from archaeological sites. *Peabody Museum Bulletin* 1. Harvard University, Cambridge, Mass.

Figure 1. Possible false results produced by unclear or absent method statements.

