Archeologia e Calcolatori
21, 2010, 249-260

FIT FOR PURPOSE? ARCHAEOLOGICAL DATA IN THE 21ST CENTURY

1. Where we are now

Archaeology has always been a prolific creator of data, and the rate of creation has increased significantly over the past 30 years or so. Despite the current economic downturn, growth is likely to continue in the long term, and whether it does or not, we still have a vast undigested backlog with which to deal. The nature of the data is changing too – we now have growing bodies of "born digital" data, alongside data that have been collected conventionally (which may have subsequently been digitised by its originators), and the outcomes of large projects to digitise blocks of "legacy" data. Some has been analysed, and some not, but there is growing concern that it should all be available for re-use.

Indeed, it is almost impossible to obtain funding these days unless the outcomes will be made available for re-use. Add to this the growing numbers of originators (especially commercial archaeological organisations), and becomes clear that we have a problem on our hands, at the very least one of data management, and probably several others too. I shall concentrate on just two of them here – data quality and data integration – other challenges will be left to another forum.

2. Data quality

We all know that all datasets contain errors to a greater or lesser extent – nothing is perfect, or at least nothing can be assumed to be perfect. But do we behave as if they do? Or do we blithely ignore the possibilities as we rush into analysis and interpretation? Although I have taught about the nature of error, and its possible effects, for many years, its impact was brought home to me some years ago, when a survey that I had designed was ruined by the fact that the data on which I had based the design (and which were only one year old) were completely unreliable and in some cases wildly untrue.

The Wiltshire County Council Collection Condition Survey was carried out in 2000 by a team from the Wiltshire Conservation Service, for whom I acted as design consultant; the results were analysed by Wessex Archaeology, using Access as a database and Excel to carry out the statistical calculations. The survey covered sixteen different types of collections distributed over fifteen museums; results were required for individual museums, for each of their collections, and for collection categories across the county as a whole. No museum had a collection of every category. The numbers of objects in

Conservation priority	1 = urgent			2 = high			3 = low			4 = little		
	% standard dev.			% standard dev.			% standard dev.			% standard dev.		
collection category		actual	target		actual	target		actual	target		actual	target
agriculture	1,1	*0,49*	0,35	9,0	*1,4*	0,95	53,5	*2,3*	1,66	36,2	*2,2*	1,60
archaeology	0,7	*0,50*	0,28	1,5	*0,68*	0,41	16,1	*2,1*	1,23	81,7	*2,2*	1,29
arms & armour	0	0	0,00	5,4	*0,91*	0,75	41,2	*1,9*	1,64	52,7	*1,9*	1,66
biology	0	0	0,00	1,4	*0,59*	0,39	1,8	*0,65*	0,44	96,5	*0,91*	0,61
ethnography	1,4	0	0,39	5,4	0	0,75	25,8	0	1,46	67,1	0	1,57
geology	0,40	0,17	0,21	2,0	0,31	0,47	5,1	*0,76*	0,73	92,3	0,85	0,89
maritime	0	0	0,00	0	0	0,00	42,9	0	1,65	57,1	0	1,65
medals	0	0	0,00	5,5	0,55	0,76	25,6	0,65	1,45	68,9	0,83	1,54
medical	0	0	0,00	1,2	0,28	0,36	23,4	1,2	1,41	75,4	1,2	1,44
music	0	0	0,00	0	0	0,00	30,0	0	1,53	70,0	0	1,53
numismatics	0	0	0,00	1,1	*0,43*	0,35	5,6	*0,91*	0,77	92,8	*1,0*	0,86
personalia	0	0	0,00	4,3	*2,6*	0,68	35,5	*5,2*	1,60	60,3	*5,4*	1,63
photography	0	0	0,00	0	0	0,00	60,5	0	1,63	39,5	0	1,63
science/industry	0	0	0,00	6,2	0	0,80	6,2	0	0,80	87,5	0	1,10
social history	0,50	0,23	0,24	3,2	0,54	0,59	30,4	*1,6*	1,53	65,9	*1,6*	1,58
transport	0	0	0,00	14,3	*2,4*	1,17	62,6	*2,4*	1,61	23,1	0	1,40
all	0,50	*0,14*	0,07	1,9	*0,19*	0,14	15,0	*0,54*	0,36	82,4	*0,57*	0,38

Tab. 1 – The percentages of objects in each conservation priority in the Wiltshire Museums Survey, broken down by collection category. The actual and target standard deviations for each collection category are also shown.

collection category	museum code	*N*	$\hat{N}$
agriculture	L	4387	2926
	C	*100*	*31*
	S	*>700*	*100*
	W	*400*	*108*
archaeology	D	<150,000	44.100
	S	>100,000	38.250
	C	*>500*	*126*
arms & armour	W	557	45
	ME	129	62
	MA	*115*	*24*
	T	*350*	*15*
biology	D	40.000	11.850
numismatics	D	20.000	11.520
	S	>10,000	7776
	C	*300*	*30*
personalia	T	2000	155
	P	*200*	*26*
	W	*535*	*3*
transport	W	550	15
	D	*6000*	*28*
	MA	*183*	*6*
	P	*100*	*18*
	T	*150*	*4*

Tab. 2 – List of "problem" collections, for which the original population size greatly exceeds the survey population size. Collections shown in italics did not contribute significantly to a large standard deviation.

each category in each museum were obtained from a survey (the *Mapping Project*) carried by the South West Museums Council in 1999. Objects were to be placed into one of four categories, known as "priorities", according to their need of conservation treatment: "urgent", "high", "low" and "little" (Keene, Orton 1992).

The scope of the survey meant that results were required at various levels:

– Level 1: "global", i.e. the overall collection.
– Level 2: "museum", i.e. all collections in a single museum.
– Level 3: "collection", i.e. all collections of a single category.
– Level 4: a single collection category in a single museum.

The main aim was to estimate the proportion of objects in the collection that required conservation treatment (the "urgent" plus "high" priorities), which was thought likely to be of the order of 10%, at each of these four levels. The design was intended to provide the smallest overall sample that could achieve standard deviations of 1% on proportions of 10% at Level 3 and standard deviations of 2% on proportions of 10% at Level 4. It was expected to achieve standard deviations of about 0.3% at Level 1. The outcomes were disappointing: at Level 3, of the sixteen collection categories, two failed to meet their target standard deviation for the "urgent" priority and eleven of the others had proportions of zero in this priority; seven failed for the "high" priority and three of the others had zero proportions (see Table 1); all categories failed at Level 1. Failures are indicated in red (online version) or italic (printed version). Seven collections can be described as "problems" (see Table 2, which also shows the original population sizes N of each "problem" collection, as given by the *Mapping Project*, and the estimated population sizes $\hat{N}$ derived from the survey itself).

It can be seen that these collections are characterised by large discrepancies between these two sets of values, with N being much larger than $\hat{N}$. The overall effect of these changes in population numbers was to reduce the population from over 450,000 to about 160,000, and the sample from about 21,000 to about 10,000 objects. It is clear that the main, and probably the only, reason why the survey did not achieve its design criteria is that some population figures were inaccurate, often wildly so (see Orton 2003 for a fuller account).

It is not my purpose to criticise this survey, which is probably no worse than many others, but simply to point out the serious repercussions that errors in data can have, based on my own personal experience.

3. Data integration

The multiplicity of field and laboratory projects, carried out by a wide range of diverse organisations, means that any work of synthesis is likely to

engage with data from a range of sources. The days of collecting fresh data to answer every question are now over (if indeed they ever existed), and we are all encouraged to make the best use of existing datasets, adding to them only if necessary to remedy particular deficiencies. There are even organisations which exist to facilitate this process, for example the Archaeology Data Service (ADS) in the UK, based at the University of York. If, for example, one uses the Service's *ArchSearch* facility to discover sites of a certain type in a region, the information may come from a variety of sources: local (county) SMRs/HERs, English Heritage, Defence of Britain Project, etc. For this information to be useful to the researcher, the sources must, at least to some extent, "speak the same language", i.e. the same terms should have the same meanings wherever they are used. For some classes of data, this process is well advanced, e.g. the MIDAS Heritage UK Historic Environment Data Standard for site types (http://www.midas-heritage.info/), but for other classes, e.g. pottery wares (Tyers 1996) it is far less so. We shall see an example later of the effects that this can have.

Beyond this, there are technical issues, such as file formats and software preferences, which are in principle solvable but in practice can cause serious problems. A classic example is that of the archive of the Newham Museum Archaeological Service, which was deposited with the ADS when the Service closed in 1998, but parts of which proved to be unreadable because of either (a) obsolete graphic file formats or (b) coding systems for which the key had been lost (http://ads.ahds.ac.uk/newsletter/issue6.html).

4. What we teach about data quality and integration

In my experience, archaeology undergraduates are taught about the different types of data (nominal, ordinal, interval and ratio) and the implications for analysis of the different types. They are also taught about the different types of error that can arise: random error, systematic error (bias) and gross errors (outliers). It is to be hoped that such topics are taught universally. A less common topic, perhaps, is that of misclassification – the use of the "wrong" label for a particular archaeological object – and coding errors, where an object is correctly classified but the wrong code is entered when the data are recorded. Altogether more subtle, and outside our scope here, are errors in the models which we use to analyse our data – for example, are our data really "Normal", and does it matter if they are not?

On the topic of integration, students should learn about the different sources of data to which they can have recourse, and a visiting lecture from a representative of the ADS or a similar organisation should feature on every syllabus. This would lead naturally to the question of the standardisation of terminology: it seems to be necessary in order to facilitate communication and

the integration of data, but does it also fossilise a topic? How do we achieve change in a standardised system?

5. Case studies

Here I present three case studies, one each on different aspects:

– quantitative (ratio) data – measuring objects;
– qualitative (nominal) data – classifying pottery;
– integration of data from different sources within a region.

5.1 *Measuring objects*

As a class exercise, intended to illustrate aspects of variability and error, a class of ten students was asked to measure the lengths and widths of a sample of 35 flint axes from the Humbla Collection in the UCL Institute of Archaeology Collections. Students were deliberately not given any specialist equipment for this task, nor were they given formal definitions of "length" and "width" of such objects, since part of the exercise was to draw out differences in interpretation and method.

The outcomes are shown in Table 3; clear gross errors are shown in red (online version) or bold (printed version), and probable errors in blue (online version) or italic (printed version). Two errors fall into the former category: student F has over-estimated the length of axe no. 4 by about 30 mm, and student A has transposed the length and width of axe no. 26. There are several smaller errors, all except one of which (student J, axe no. 20) are negative. This suggests a source of bias, which is particularly clear in the case of the widths measured by student J, which are almost consistently low. The class attributed this bias to a tendency to measure from the end of the ruler instead of the end of the scale on the ruler (a difference of about 5 mm). Another problem that arose was that student F failed to complete the task in the time available, creating a problem of "missing data".

Thus, of the 700 measurements taken, three (0,4%) are clearly gross errors, 25 (3,6%) are probable (though less pronounced) gross errors, and seven (1%) are missing. The rest express a level of variability, measured in terms of the range, usually of between 4 and 10% of the mean value. Ranges of above 10% of the mean value seem to indicate the possibility of outliers. The remaining variation comprises small "random" errors. There are no strong correlations between range or percentage range and size, whether measured by length or width, which suggests that shape plays an important part in the scale of errors. It is obviously not possible to extrapolate directly from these figures, but on the other hand there is no reason to suppose that they are out of the ordinary.

lengths (mm)												raw data			corrected data		
	student																
axe number	A	B	C	D	E	F	G	H	I	J		average	range	% range	average	range	% range
1	119	120	120	119	122	122	120	120	120	122		120,4	3	2,5	120,4	3	2,5
2	125	129	128	127	130	131	129	126	128	133		128,6	8	6,2	128,6	8	6,2
3	128	128	132	131	130	132	131	130	131	134		130,7	6	4,6	130,7	6	4,6
4	107	106	108	107	111	140	107	108	109	110		111,3	33	29,6	108,1	5	4,5
5	121	120	122	122	123	124	122	121	123	125		122,3	5	4,1	122,3	5	4,1
6	127	128	128	126	129	130	127	128	121	125		126,9	9	7,1	126,9	9	7,1
7	119	120	122	122	123	122	121	122	121	124		121,6	5	4,1	121,6	5	4,1
8	84	83	85	85	86	86	83	83	85	80		84,0	6	7,1	84,0	6	7,1
9	75	74	76	76	76	77	77	76	75	76		75,8	3	4,0	75,8	3	4,0
10	122	120	130	126	128	130	128	128	125	131		126,8	9	7,1	126,8	9	7,1
11	110	113	*105*	111	116	112	114	114	112	115		112,2	11	9,8	113,0	6	5,3
12	111	110	112	116	113	*103*	110	112	110	113		111,0	13	11,7	111,9	6	5,4
13	88	80	82	82	83		82	82	81	84		82,7	8	9,7	82,7	8	9,7
14	92	90	94	95	94	96	95	94	92	95		93,7	6	6,4	93,7	6	6,4
15	93	92	95	95	96	96	95	95	94	94		94,5	4	4,2	94,5	4	4,2
16	94	94	95	92	96		97	95	97	99		95,4	5	5,2	95,4	5	5,2
17	90	90	92	91	92	94	92	91	92	95		91,9	5	5,4	91,9	5	5,4
18	119	120	120	120	123	124	120	120	120	127		121,3	7	5,8	121,3	7	5,8
19	123	127	124	125	126	128	130	125	124	133		126,5	10	7,9	126,5	10	7,9
20	116	120	119	123	120		120	118	118	*128*		120,2	12	10,0	119,3	7	5,8
21	127	132	129	126	132	131	130	129	128	134		129,8	8	6,2	129,8	8	6,2
22	126	126	128	126	129	130	126	127	127	131		127,6	5	3,9	127,6	5	3,9
23	99	100	106	104	103	107	104	102	103	106		103,4	8	7,7	103,4	8	7,7
24	101	107	104	105	103		104	103	102	104		103,7	6	5,8	103,7	6	5,8
25	113	117	112	115	117	118	*105*	115	117	118		114,7	13	11,3	115,8	6	5,2
26	43	95	98	100	101	99	98	98	99	102		93,3	59	63,2	98,9	7	7,1
27	88	91	91	92	93	94	87	91	90	95		91,2	7	7,7	91,2	7	7,7
28	128	127	130	129	130	132	132	130	130	135		130,3	8	6,1	130,3	8	6,1
29	119	120	121	117	123	121	122	120	121	122		120,6	6	5,0	120,6	6	5,0
30	121	124	123	119	123		122	120	123	124		122,1	5	4,1	122,1	5	4,1
31	94	100	105	110	95	100	99	101	100	104		100,8	11	10,9	100,8	11	10,9
32	134	136	142	137	141	141	138	140	138	144		139,1	10	7,2	139,1	10	7,2
33	126	130	132	134	129	130	126	130	127	130		129,4	8	6,2	129,4	8	6,2
34	139	136	143	140	142	140	142	142	140	147		141,1	11	7,8	141,1	11	7,8
35	156	160	163	156	157			159	158	155		158,0	8	5,1	158,0	8	5,1
average	110,8	113,3	114,7	114,3	115,3	116,9	112,8	114,1	113,7	117,0		114,3	9,7	8,9	114,5	6,8	6,0

widths (mm)												raw data			corrected data		
	student																
axe number	A	B	C	D	E	F	G	H	I	J		average	range	% range	average	range	% range
1	47	47	48	48	49	50	47	48	49	*35*		46,8	15	32,1	48,1	3	6,2
2	55	56	57	56	60	57	58	57	57	55		56,8	5	8,8	56,8	5	8,8
3	46	45	42	45	46	47	46	47	46	41		45,1	5	11,1	45,1	5	11,1
4	37	35	35	36	36	37	34,5	36	36	34		35,7	3	8,4	35,7	3	8,4
5	46	45	45	47	47	49	48	46	48	*36*		45,7	12	26,3	46,8	4	8,6
6	49	48	48	49	50	50	48	48	48	46		48,4	4	8,3	48,4	4	8,3
7	47	48	48	48	49	48	48	48	48	47		47,9	2	4,2	47,9	2	4,2
8	44	45	46	46	47	47	45	46	47	*38*		45,1	9	20,0	45,9	3	6,5
9	35	34,5	34	35	36	35	34	35	35	33		34,7	3	8,7	34,7	3	8,7
10	66	65,5	68	66	69	70	67	66	68	*59*		66,5	11	16,6	67,3	4	5,9
11	58	58	58	59	59	61	59	59	58	*49*		57,8	12	20,8	58,8	3	5,1
12	55	56	56	56	58	59	55	56	57	*51*		55,9	8	14,3	56,4	4	7,1
13	38	38	38	38	39		35	38	38	*31*		37,0	8	21,6	37,8	4	10,6
14	50	47	49	51	50	59	49	51	50	*42*		49,8	17	34,1	50,7	4	7,9
15	45	44	44	45	45	46	44,5	45	45	*32*		43,6	13	29,9	44,8	2	4,5
16	39	39	41	41	41		39	40	38	37		39,4	4	10,1	39,4	4	10,1
17	43	43	43	44	45	45	44	45	45	*37*		43,4	8	18,4	44,1	2	4,5
18	52	49,5	49	50	51	50	50	50	50	*44*		49,6	18	36,3	50,2	3	6,0
19	47	47	47	45	49	49	48	47	47	49		47,5	4	8,4	47,5	4	8,4
20	50	50	51	52	52		50	51	52	49		50,8	2	3,9	50,8	2	3,9
21	51	49	49	50	49	51	49	50	50	51		49,9	2	4,0	49,9	2	4,0
22	47	48	48	47	49	50	48	48	49	48		48,2	3	6,2	48,2	4	8,3
23	52	51,5	53	52	53	56	52	54	53	*42*		51,9	4	7,7	52,9	4	7,6
24	46	46	46	46	48		47	47	46	43		46,1	5	10,8	46,1	5	10,8
25	40	40	40	42	42	41	40	40	40	39		40,4	3	7,4	40,4	3	7,4
26	88	42	43	43	43	44	42	42	43	*29*		45,9	59	128,5	42,8	2	4,7
27	43	43	43	43	44	45	43	44	42	*37*		42,7	8	18,7	43,3	2	4,6
28	47	47	46	46	48	48	48	47	48	43		46,8	5	10,7	46,8	5	10,7
29	50	50	50	56	52	54	50	50	51	*39*		50,2	17	33,9	51,4	6	11,7
30	55	55	55	55	56		55	55	55	*43*		53,8	13	24,2	55,1	1	1,8
31	53	53	53	54	56	56	53	53	*49*	55		53,5	7	13,1	53,5	3	5,6
32	52	50,5	51	51	53	*58*	51	52	50	50		51,9	8	15,4	51,9	3	5,8
33	48	48	49	47	48	48	47	49	49	*38*		47,1	11	23,4	48,1	2	4,2
34	53	52	54	54	55	55	53,5	54	52	*43*		52,6	3	5,7	53,6	3	5,6
35	53	52	52	52	54			53	52	*45*		51,6	9	17,4	52,6	2	3,8
average	49,3	47,6	48,0	48,4	49,4	50,5	47,9	48,5	48,3	42,6		48,1	9,1	19,1	48,4	3	6,9

Tab. 3 – The lengths (a) and widths (b) of 35 flint axes from the Humbla Collection in the UCL Institute of Archaeology Collections, as measured by a class of ten students.

5.2 *Classifying pottery*

There seems to be relatively little information about the accuracy of the nominal data recorded in archaeology, although it has been studied in other areas. A lesson can be learnt from the work done by Robinson as long ago as 1976 (Robinson 1979). Supported by the Medieval Pottery Research Group, she looked for reliable and simple visual ways of characterising ceramics, in the context of the descriptive paradigm set by Peacock (1977) and implemented at, for example, the Museum of London (Orton 1979).

She tested three methods of description:

1. Freestyle written description.
2. Questionnaire-type description (tick boxes).
3. As 2, supported by mounted examples and/or photographs.

She visited 23 organisations in Britain with six groups, each of five sherds and each from a different part of the country, and asked four or five people at each organisation (divided into experienced and inexperienced observers) to describe them by each method. Some minor changes were made to method 3 about halfway through the experiment. The questionnaires were based on seven attributes: colour, hardness, feel, fracture, inclusion, surface treatment and manufacture (method 1) or glaze (methods 2 and 3). A further difference between methods 2 and 3 was the use of a simple 12-colour chart in method 3, in contrast to the full Munsell colour system in method 2. Results for each sample were archived (Robinson 1978) and summary tables were published (Robinson 1979, figs. 11, 12).

She analysed the outcomes by counting as a success any combination of sample sherd and attribute for which 70% or more of the observers agreed on the descriptive category. Different attributes and methods were compared by calculating the percentage of sample sherds which achieved success for any particular combination of attribute and method. She concluded that the results of method 1 were too diverse to be useful, but that the results of methods 2 and 3 were "reasonably good", with method 3 usually (but not always) performing better than method 2. The use of photographs as standards seems to have been as successful as mounted samples.

Her conclusions seem to be rather optimistic, given her data. While 70% agreement is sufficient to demonstrate statistically that the results could not have come about by pure chance (i.e. by guesswork on the part of the observer), no one would suppose that that might have been a valid explanation. The proportions achieving even this, to my mind quite modest, level of agreement are low. Table 4 shows the proportions of the samples achieving this level in method 3 for the "best" of the attributes, for experienced observers and for all observers.

This and similar experiences have led to the creation of high-quality visual guides (e.g. Tomber, Dore 1998) and to suggestions that these could

attribute	percentage success experienced	all
colour: core	73	49
: int	61	55
: ext	64	61
fracture: samples	59	27
: photos	40	30
: both	46	15
inclusions: sorting	64	55
manufacture	55	36
: samples and photos	80	60
glaze: presence/absence	97	97
: colour ext	85	87
: colour int	74	58

Tab. 4 – The percentages of the samples in Robinson's experiment which achieved at least 70% agreement in method 3 for the "best" of the attributes, for experienced observers and for all observers.

be made available online (e.g. Lange 2004). Sherd-by-sherd detailed verbal description, which was probably brought about in the 1970s by a lack of confidence in newly-appointed ceramic specialists, combined with the blandishments of over-confident computer data analysts, has given way to the idea of standards to which examples can be matched, hopefully with more consistency than in this experiment. However, the basic underlying conclusions remain: archaeologists do not seem to be very good at agreeing on basic visual descriptions.

There are two possible reactions to this conclusion: either (a) an appreciable proportion of descriptions are simply wrong, or (b) attribute states are in the eye of the beholder, and «there is no such thing as a right or wrong answer» (Robinson 1979, 28). In either case, inter-site (or strictly, inter-observer) comparisons are likely to be unreliable.

5.3 *Integrating regional data*

As we have seen above, Roman pottery in Britain can frequently be found under a range of "aliases", i.e. different names given to the same ware by different researchers or organisations (Tyers 1996, 85). What might the practical implications be? Some years ago, I was asked to referee a paper about excavations carried out in advance of the building of the Croydon Tramlink, a new tram route in South London. The work had been carried out by Oxford Archaeology (OA), one of the largest and most respected archaeological companies in the UK. My particular interest was in the ceramic reports, of which there were two: Roman and post-Roman. The latter had been writ-

ten by a freelance specialist, who had used the fabric codes employed by the (then) Museum of London Archaeology Service (MoLAS). The former had been written by OA's in-house Roman ceramic specialist, who naturally had used OA's own ceramic codes, which were not the same as MoLAS's codes.

The outcome was that the Roman ceramics from the Tramlink sites could not be compared to those from other local sites, which had been excavated by MoLAS. There was, so to speak, an island of OA codes in a sea of MoLAS codes. My recommendation was that the Roman ceramics should be catalogued according to the MoLAS codes, and if that was not practical (and I appreciated there would be a cost), then a translation table between the two sets of codes should be provided. This was in no way critical of either system, it just highlighted the need to make the data useful in their regional context.

6. Discussion

I started to write this paper in the expectation that it would be mainly about errors in counting and measuring, with questions of classification and integration included for the sake of completeness. However, both counting errors (the Wiltshire survey) and measurement errors (the flint axes) seem to be pointing towards problems of classification and definition as important components. Some of the discrepancies in the counts are so severe that they cannot reasonably be ascribed to simple counting errors. For example, some of the noted discrepancies for the "arms and armour" category are over 90% of the original figure. It seems far more likely that the difference in number is due to a difference in definition; in other words, "what is an object?" Is a composite object (such as a suit of armour), one object or several? The important thing is that there is an agreed definition. The other source of large errors appears to be guesswork; for example, a figure of 100,000 looks suspiciously round, and immediately suggests that no-one has actually counted the objects.

The measurements, too, seem to point towards issues of definition as a contributory factor to the creation of errors. Giving students latitude in deciding what they meant by length and width has produced a range of subtly different interpretations, many of which are probably implicit. By contrast, osteologists have devised quite sophisticated devices and protocols for measuring lengths of long bones and key dimensions on skulls (Klein, Cruz-Uribe 1984, 22; Von den Driesch 1976).

7. Where do we go from here?

The thrust of this paper is that archaeologists need to be more aware of the potential for errors in their data, and of the problems that they may cause. This would lead to a greater concern for (a) preventing errors, (b) detecting

errors once they have occurred, and (c) living with errors that have escaped all our filtering processes.

7.1 *Preventing errors*

Action needs to be taken at the levels of (a) within organisations, and (b) between organisations. Within organisations, care needs to be taken in matching individuals to tasks, and in ensuring proper training and motivation. The stigma of data entry appearing to be a low-grade task must be avoided; for example, treating it as an activity for site staff when rain prevents excavation is to invite trouble, as I discovered working in Novgorod. Checks can be built into software, for example by providing lists of acceptable terms for nominal data and credibility limits for numerical data (creating error messages for other terms and for data outside the limits). Such checks often form part of data-entry software, but the extent to which they are used in archaeology is not known.

Between organisations, as we have seen above, the issue is one of ensuring consistency of terminology and approach. The provision of standards is well advanced in some areas, but there seems to be little to ensure that they are adhered to. There remains the problem of data that were collected before standards became available. The balance between ensuring standard terminology and allowing the discipline to develop is still a difficult one.

7.2 *Detecting errors*

Once errors have become embedded in a dataset, they are unlikely to be noticed until the data are analysed. Detecting errors in a table of numbers is extremely difficult, even to the trained eye, unless they are typographically obvious (e.g. a missing decimal point). Sometimes, a check sum will help (e.g. do the percentages in a composition sum to 100?). It may therefore be more productive to use graphical means of inspecting data, such as bar charts, histograms and scatter diagrams. It is possible for a data point to show up as an outlier on a scatter diagram even though its value on each axis could be considered typical. Multivariate outliers (i.e. with more than two dimensions) are more difficult to detect, and can escape even a battery of two-way scatter diagrams. However, tests are available if needed (e.g. Rousseeuw, Van Zomeren 1990). A guiding principle is that "if it looks wrong, it probably is wrong".

7.3 *Living with errors*

Despite all our precautions, we must expect that errors will from time to time get through to our final datasets and be used for analysis. How can we minimise their effects on our outcomes and interpretation? We need analytical

techniques that are not unduly affected by "rogue" values; the jargon word for such techniques is "robust". Non-parametric techniques are often more robust than their parametric counterparts, as they do not depend on models (e.g. a Normal distribution) which may not be appropriate for a particular dataset (see, for example, Siegel, Castellan 1988). Further possibilities are offered by the bootstrap and jack-knife techniques (Baxter 2003, 148-154).

8. Conclusions

If we are to make progress with the application of quantitative methods in archaeology in the 21st century, we need to pay more attention to the quality of our data, and to ensuring the compatibility of data from different sources. Archaeologists need to be made more aware of the issues involved, and of the practical steps that they can take to minimise their effect. Otherwise we run the risk of building on insecure foundations.

Clive Orton
Institute of Archaeology
University College London

REFERENCES

Baxter M. 2003, *Statistics in Archaeology*, London, Hodder Arnold.

Keene S., Orton C. 1992, *Measuring the condition of museum collections*, in G. Lock, J. Moffett (eds.), *Computer Applications and Quantitative Methods in Archaeology 1991*, BAR International Series 577, Oxford, Tempus Reparatum, 163-166.

Klein R.G., Cruz-Uribe K. 1984, *The Analysis of Animal Bones from Archaeological Sites*, Chicago, Chicago University Press.

Lange A.G. 2004 (ed.), *Reference Collections Foundations for Future Archaeology,* Amersfoort, ROB.

Orton C. 1979, *Dealing with the pottery from a 600-acre urban site*, in M. Millett (ed.), *Pottery and the Archaeologist*, Institute of Archaeology Occasional Publication, 4, London, 61-72.

Orton C. 2003, *Keeping everybody happy: Museum surveys with multiple objectives*, in M. Doerr, A. Sarris (eds.), *The Digital Heritage of Archaeology CAA2002. Computer Applications and Quantitative Methods in Archaeology (Heraklion, Crete, 2002)*, Athens, Archive of Monuments and Publications, Hellenic Ministry of Culture, 391-397.

Peacock D.P.S. 1977, *Ceramics in Roman and Medieval archaeology*, in D.P.S. Peacock (ed.), *Pottery and Early Commerce*, London, Academic Press, 21-33.

Robinson A.M. 1978, *Medieval Pottery Fabrics. A Detailed Comparison of Methods of Description*, MA Dissertation, Bradford, University of Bradford.

Robinson A.M. 1979, *Three approaches to the problem of pottery descriptions*, «Medieval Ceramics», 3, 3-36.

Rousseeuw P.J., Van Zomeren B.C. 1990, *Unmasking multivariate outliers and leverage points*, «Journal of the American Statistical Association», 85, 411, 633-651.

Siegel S., Castellan J. 1988, *Nonparametric statistics for the behavioral sciences*, New York (2nd ed.), McGraw-Hill.

Tomber R.S., Dore J. 1998, *The National Roman Fabric Reference Collection: A Handbook*, London, Museum of London Archaeology Service.

Tyers P.A. 1996, *Pottery in Roman Britain*, London, B.T. Batsford Ltd.

Von den Driesch A. 1976, *A guide to the measurement of animal bones from archaeological sites*, «Bulletin of the Peabody Museum of Archaeology and Ethnology», 1, 1-136.

ABSTRACT

Archaeology continues to generate large amounts of data, in a growing range of formats and media. Old datasets have been or are being digitised, and there is increasing emphasis on the re-use of old datasets, and on preparing new datasets with re-use in mind. That being so, surprisingly little attention has been paid to the prevention and detection of errors in archaeological data, and in acquiring or developing robust methods of analysis. The sorts of errors that can be encountered in different types of data are approached and discussed through a series of case studies, dealing with counting errors, measurement errors, and classificatory errors. They are linked to another obstacle to the re-use of data: the lack of standardised terminology between different originators. Strategies for mitigating these problems (which cannot be totally overcome) are discussed.