

ANALECTA
PRAEHISTORICA
LEIDENSIA

28

PUBLICATIONS OF THE INSTITUTE OF PREHISTORY
UNIVERSITY OF LEIDEN

INTERFACING THE PAST

COMPUTER APPLICATIONS AND QUANTITATIVE
METHODS IN ARCHAEOLOGY CAA95 VOL. I

EDITED BY
HANS KAMERMANS AND KELLY FENNEMA



UNIVERSITY OF LEIDEN 1996

contents

VOLUME I

Hans Kamermans Kelly Fennema	Preface
Data Management	
Jens Andresen Torsten Madsen	IDEA – the Integrated Database for Excavation Analysis 3
Peter Hinge	The Other Computer Interface 15
Thanasis Hadzilacos Polyxeni Myladié Stoumbou	Conceptual Data Modelling for Prehistoric Excavation Documentation 21
E. Agresti A. Maggiolo-Schettini R. Saccoccio M. Pierobon R. Pierobon-Benoit	Handling Excavation Maps in SYSAND 31
Alaine Lamprell Anthea Salisbury Alan Chalmers Simon Stoddart	An Integrated Information System for Archaeological Evidence 37
Jon Holmen Espen Uleberg	The National Documentation Project of Norway – the Archaeological sub-project 43
Irina Oberländer-Târnoveanu	Statistical view of the Archaeological Sites Database 47
Nigel D. Clubb Neil A.R. Lang	A Strategic Appraisal of Information Systems for Archaeology and Architecture in England – Past, Present and Future 51
Nigel D. Clubb Neil A.R. Lang	Learning from the achievements of Information Systems – the role of the Post-Implementation Review in medium to large scale systems 73
Neil Beagrie	Excavations and Archives: Alternative Aspects of Cultural Resource Management 81
Mark Bell Nicola King	The MARS Project – an interface with England's past 87

Archaeometry

- M.J. Baxter
H.E.M. Cool
M.P. Heyworth
Detecting Unusual Multivariate Data: An Archaeometric Example 95
- Jon Bradley
Mike Fletcher
Extraction and visualisation of information from ground penetrating radar surveys 103
- Gayle T. Allum
Robert G. Aykroyd
John G.B. Haigh
Restoration of magnetometry data using inverse-data methods 111
- W. Neubauer
P. Melichar
A. Eder-Hinterleitner
Collection, visualization and simulation of magnetic prospection data 121
- A. Eder-Hinterleitner
W. Neubauer
P. Melichar
Reconstruction of archaeological structures using magnetic prospection 131
- Phil Perkins
An image processing technique for the suppression of traces of modern agricultural activity in aerial photographs 139
- Statistics and Classification**
- Clive Orton
Markov models for museums 149
- Juan A. Barceló
Heuristic classification and fuzzy sets. New tools for archaeological typologies 155
- Kris Lockyear
Dmax based cluster analysis and the supply of coinage to Iron Age Dacia 165
- Christian C. Beardah
Mike J. Baxter
MATLAB Routines for Kernel Density Estimation and the Graphical Representation of Archaeological Data 179
- John W.M. Peterson
A computer model of Roman landscape in South Limburg 185
- Sabine Reinhold
Time versus Ritual – Typological Structures and Mortuary Practices in Late Bronze/Early Iron Age Cemeteries of North-East Caucasia ('Koban Culture') 195
- Leonardo García Sanjuán
Jesús Rodríguez López
Predicting the ritual? A suggested solution in archaeological forecasting through qualitative response models 203
- Johannes Müller
The use of correspondence analysis for different kinds of data categories: Domestic and ritual Globular Amphorae sites in Central Germany 217
- J. Steele
T.J. Sluckin
D.R. Denholm
C.S. Gamble
Simulating hunter-gatherer colonization of the Americas 223

- Paul M. Gibson An Archaeofaunal Ageing Comparative Study into the Performance of Human Analysis Versus Hybrid Neural Network Analysis 229
- Peter Durham Image Processing Strategies for Artefact Classification 235
Paul Lewis
Stephen J. Shennan
- Gijsbert R. Boekschoten A new tool for spatial analysis: "Rings & Sectors plus Density Analysis and Trace lines" 241
Dick Stapert
- Susan Holstrom Loving Estimating the age of stone artifacts using probabilities 251
- Oleg Missikoff Application of an object-oriented approach to the formalization of qualitative (and quantitative) data 263

VOLUME II

Geographic Information Systems I

- David Wheatley Between the lines: the role of GIS-based predictive modelling in the interpretation of extensive survey data 275
- Roger Martlew The contribution of GIS to the study of landscape evolution in the Yorkshire Dales, UK 293
- Vincent Gaffney Extending GIS Methods for Regional Archaeology: the Wroxeter Hinterland Project 297
Martijn van Leusen
- Trevor M. Harris Multi-dimensional GIS: exploratory approaches to spatial and temporal relationships within archaeological stratigraphy 307
Gary R. Lock
- Philip Verhagen The use of GIS as a tool for modelling ecological change and human occupation in the Middle Aguas Valley (S.E. Spain) 317
- Federica Massagrande The Romans in southwestern Spain: total conquest or partial assimilation? Can GIS answer? 325
- Shen Eric Lim Recent examples of geographical analysis of archaeological evidence from central Italy 331
Simon Stoddart
Andrew Harrison
Alan Chalmers
- Vincent Gaffney Satellite Imagery and GIS applications in Mediterranean Landscapes 337
Krištof Oštir
Tomaž Podobnikar
Zoran Staničič
- Yvette Bommeljé The long and winding road: land routes in Aetolia (Greece) since Byzantine times 343
Peter Doorn

- Javier Baena Preysler
Concepción Blasco Application of GIS to images and their processing: the Chiribiquete Mountains Project 353

Geographic Information Systems II: The York Applications

- Julian D. Richards From Site to Landscape: multi-level GIS applications in archaeology 361
- Harold Mytum Intrasite Patterning and the Temporal Dimension using GIS: the example of Kellington Churchyard 363
- A. Paul Miller Digging deep: GIS in the city 369
- Julian D. Richards Putting the site in its setting: GIS and the search for Anglo-Saxon settlements in Northumbria 379
- Jeffrey A. Chartrand Archaeological Resource Visibility and GIS: A case study in Yorkshire 389

Visualisation

- John Wilcock A description of the display software for Stafford Castle Visitor Centre, UK 405
- Christian Menard
Robert Sablatnig Pictorial, Three-dimensional Acquisition of Archaeological Finds as Basis for an Automatic Classification 419
- Katalin T. Biró Simple fun – Interactive computer demonstration program on the exhibition of the Szentgál-Tűzköveshegy prehistoric industrial area 433
- György Csáki
Ferenc Redő Documentation and modelling of a Roman imperial villa in Central Italy 437
- Maurizio Forte
Antonella Guidazzoli Archaeology, GIS and desktop virtual reality: the ARCTOS project 443
- Germà Wünsch
Elisabet Arasa
Marta Pérez Dissecting the palimpsest: an easy computer-graphic approach to the stratigraphic sequence of Túnel VII site (Tierra del Fuego, Argentina) 457
- David Gilman Romano
Osama Tolba Remote Sensing and GIS in the Study of Roman Centuriation in the Corinthia, Greece 461
- F.J. Baena
F. Quesada
M.C. Blasco An application of GIS intra-site analysis to Museum Display 469

Education and Publication

- Robin B. Boast
Sam J. Lucy Teaching with objects 479

- Martin Belcher
Alan Chalmers
Andrew Harrison
Simon Stoddart
Teaching the Visualisation of Landscapes – Approaches in Computer based learning for Archaeologists 487
- Anja C. Wolle
Stephen J. Shennan
A Tool for Multimedia Excavation Reports – a prototype 493
- G. Gyftodimos
D. Rigopoulos
M. Spiliopoulou
Exploring Archaeological Information through an Open Hypermedia System 501
- Martijn van Leusen
Sara Champion
Jonathan Lizee
Thomas Plunkett
Toward a European Archaeological Heritage Web 511
- Mike Heyworth
Seamus Ross
Julian Richards
Internet archaeology: an international electronic journal for archaeology 521
- Virgil Mihailescu-Bîrliba
Vasile Chirica
A Survey of the Development of Computer Applications in Romanian Archaeology 529
- Kris Lockyear
Computer-aided publication in practice 535

Detecting Unusual Multivariate Data: An Archaeometric Example

1 Introduction

There has been much recent interest in the statistical literature concerning the detection of outliers and other unusual cases in multivariate data. This has arisen, in part, because developments in computing power have made possible the application of methodology that is iterative and computer-intensive in nature.

Methods of chemical analysis, such as inductively coupled plasma spectroscopy, increasingly generate large multivariate data sets, of artefact compositions for example, that are subjected to 'standard' methods of statistical analysis such as cluster analysis, principal component analysis (PCA) or discriminant analysis (Baxter 1994).

The performance of these analyses can be affected by unusual cases, such as outliers, in the data. It is good practice to screen the data in advance of applying such methods in order to identify cases that may affect their performance. How cases that are unusual are treated will depend on the context of a study, but it is often sensible to remove unusual cases from subsequent analyses, in order to study the structure in the bulk of the remaining data.

The present paper has arisen as part of a wider programme of study looking at approaches to the statistical analysis of large archaeometric data sets. Here we look selectively at a number of approaches for identifying unusual cases in such data sets, with a view to raising questions about aspects of some of the methodologies that are available.

The data used are the chemical compositions of 250 specimens of glass found in a single post-Roman context in excavations at Winchester. Most of the glass is window glass, mainly light blue-green in colour but including other more distinctive pieces representative of other colours in the assemblage of several thousand specimens. Some samples that were possibly vessel glass were also selected for comparative purposes. For the purpose of this paper the major and minor oxides only, based on the elements Al, Ca, Fe, K, Mg, Mn, Na, P and Ti, will be used. It was postulated, in advance of chemical analysis, that most of the glass would be reasonably homogeneous with respect to such oxides. We have also looked statistically at analyses

based on nine trace elements, and on the full set of eighteen variables, and will refer in passing to some of the results of these analyses.

The glass had not previously been analysed statistically and, as part of an experiment, was studied typologically independently of the statistical analysis. The aim was to see whether unusual cases detected statistically were also typologically unusual, without the interpretation of the statistical analysis being affected by a knowledge of the typology.

In the next section a brief review of some approaches to detecting unusual multivariate data is given. We have not attempted a comprehensive review, and refer the reader to the original publications and references given there for technical details. The application of some of the methodologies to the data set described above, and its relationship to the typological analysis, is then discussed.

The concluding discussion, rather than attempting to reach definitive conclusions about the structure of this particular data set, is more concerned to raise issues about how unusual data should be identified and handled. We wonder to what extent fairly 'simple' approaches will often suffice for practical purposes. This is not a question that can be answered without more practical experience of the methodologies discussed here, and others that 'compete' with them.

2 Detecting unusual data

Approaches to detecting unusual multivariate data include the following.

1. Univariate and bivariate data inspection.
2. Inspection of the first and last few principal components from a principal component analysis (PCA) (e.g. Hawkins/Fatti 1984).
3. Influence analyses to identify those cases that have the greatest effect on some specific technique such as PCA (e.g. Brooks 1994).
4. Cluster analysis.
5. The use of Mahalanobis distance, d_j , for the j 'th case, where the square of d_j is given by

$$(x_j - \bar{x})^T S^{-1} (x_j - \bar{x})$$

and \bar{x} and S are the multivariate mean and covariance matrix of the $n \times p$ data matrix X of which \mathbf{x}_j is the j 'th row. Large values of d_j are intended to identify points remote from the bulk of the data.

Many other statistics have been proposed for detecting unusual data, but Mahalanobis distance, or variants of it, has received the most attention in practice. Its major disadvantage is that \bar{x} and S are themselves affected by unusual observations so that d_j is affected by the cases it is designed to detect, and may fail to do so. Principal component analyses, which are based on an eigen-analysis of S , possibly after standardising the data, suffers from a similar problem.

This has led to a variety of proposals for robust analyses in which estimates of \bar{x} and S are determined, usually iteratively, that are unaffected by outliers. There are two broad ways in which this can be done, either by defining weighting schemes that downweight extreme cases, or by identifying a subset of data uncontaminated by extremes and basing calculations on this. This last idea forms the basis of the paper by Atkinson and Mulira (1993), whose approach is used here. The Atkinson/Mulira approach is similar in spirit to other approaches that have been proposed while being simpler and more practical. It is aimed at the detection of multivariate outliers, rather than directly at the robust estimation of S . Other approaches, not discussed here, aim to estimate S robustly, and then use the estimate in a principal component analysis to detect outliers, for example. The basic idea is as follows.

1. Select $p + 1$ observations and calculate \bar{x} and S using these.
2. Calculate d_j , increment the sample size by some small integer k , and select a new sample to consist of those cases with the smallest values of d_j .
3. Use the new sample to recalculate \bar{x} and S and then repeat stage 2, selecting $(p + 1 + sk)$ cases at stage s , until the data set is exhausted.
4. Identify and display results for a suitable choice of s .

Reference can be made to Atkinson and Mulira (1993) for a discussion of methods of displaying the results. Here we shall use an index plot of d_j , where calculations for S are based on about 80% of the data. The results we give are not sensitive to variation about this value of 80%, and were virtually the same for several different choices of initial sample, including one specifically designed to include the most obvious outliers. Atkinson and Mulira (1993) suggest statistical guidelines for identifying outliers that assume that the sample from which S is calculated is multivariate normal. As shall be seen this turns out to be far from the case with our data, and we have interpreted the plots subjectively.

3 Analysis

Using the 250×9 data set described in the introduction, box-and-whisker plots and dot-plots were used to identify 'obvious' univariate outliers. Let H denote the inter-quartile range; then such outliers were defined to be points more than $3H$ from the upper or lower quartiles and visibly separate from the rest of the data. This last criterion was imposed to avoid identifying as outliers points in the tail of a long-tailed distribution. According to this definition the following points clearly stand out.

1. (87, 127, 179, 232) with high Ca and Ti;
2. (18, 31, 93, 118) with high Fe;
3. 98 with high Mn and Ti;
4. (141, 234) with low Ca.
5. Additionally it was noted that 20 cases, of which (225, 242) were most prominent, lay in the tails of Mg, K and P, which all had long tailed distributions. These were not designated as outliers but form a distinct cluster of points. The question of whether or not the data should have been transformed prior to analysis is raised by this observation, and we return to it in the final section.

Figure 1 shows two index plots of d_j , the upper plot using S calculated from all the data, and the lower plot using the (hopefully) uncontaminated 80% of the data determined by the Atkinson/Mulira procedure. The lower plot quite clearly suggests 9 cases as unusual and these correspond to the first 9 cases identified by univariate inspection above. The upper plot, which is affected by the unusual data is less clear cut and less easy to interpret.

Note that values of d_j for the unusual cases are much more extreme in the lower diagram.

A plot of the 4'th and 5'th principal components of standardised data also clearly identified the same 9 cases as unusual. An average-link cluster analysis of standardised data with Euclidean distance as the measure of dissimilarity, shown in figure 2, if 'cut' at about 6, separates out the 11 points noted in 1 to 4 above from the rest. If cut at 5 the majority of the cases noted in 5 above are also separated out.

The 9 'obvious' unusual points were removed from the data set and analyses repeated. Cases 141 and 234 then stood out, particularly on the second principal component, and were also removed.

Figure 3 shows the principal component analysis plot for the first two components, undertaken after this removal, and suggests some clustering of the data, with 242 as possibly unusual. The cases in the cluster to the lower right of the plot all belong to the group noted in 5 above.

In summary the Atkinson/Mulira approach, tailored to the identification of multivariate outliers, identifies much the same points as a simple univariate analysis or a cluster

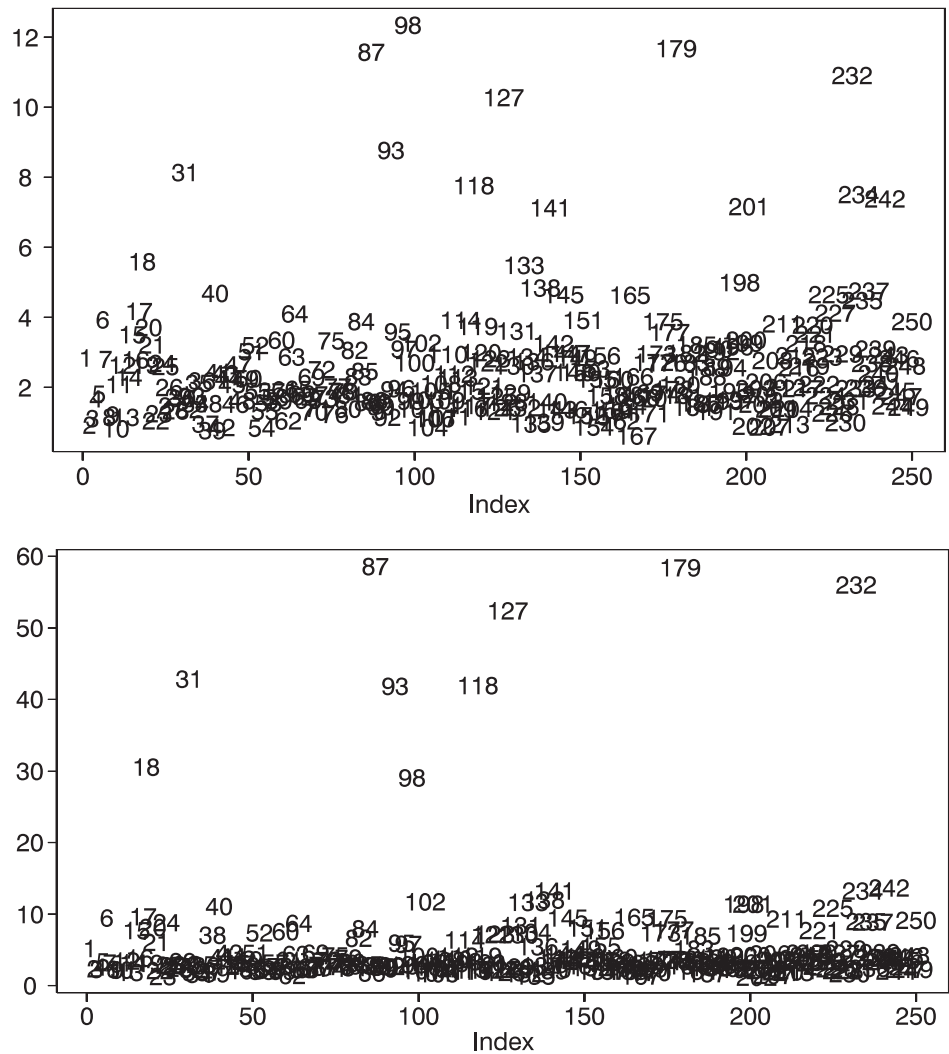


Figure 1. Index plots of Mahalanobis distance for the oxide data with calculations based on all the data (upper figure) and 80% of the data identified by the Atkinson/Mulira procedure. Points are labelled by their index for convenience of identifications.

analysis for these data. Additionally, the apparent clustering in figure 3 suggests that there is not a single homogeneous group against which other specimens can be judged to be ‘unusual’. This raises a number of issues about the utility of the multivariate methodology to which we will return after briefly noting the relationship of the statistical analysis to the typological analysis.

4 Typological analysis

The statistical and typological analyses were initially carried out independently of each other. This is not recommended as a general practice but, in the present case,

it was of interest to see whether or not the two approaches produced compatible results.

Three of the first group noted previously, (87, 127, 179), are Roman vessel or bottle glass while 232 is heat-affected so a clear identification is not possible. That is three, if not all four, of these cases are typological outliers, compared to the bulk of light blue-green post-Roman window glass.

The same is true of 98, which is an unusual specimen of vessel glass, possibly of Mediterranean origin. The other group of four, (18, 31, 93, 118), is also highly distinctive in terms of colouring (emerald) and stands out from all other specimens.

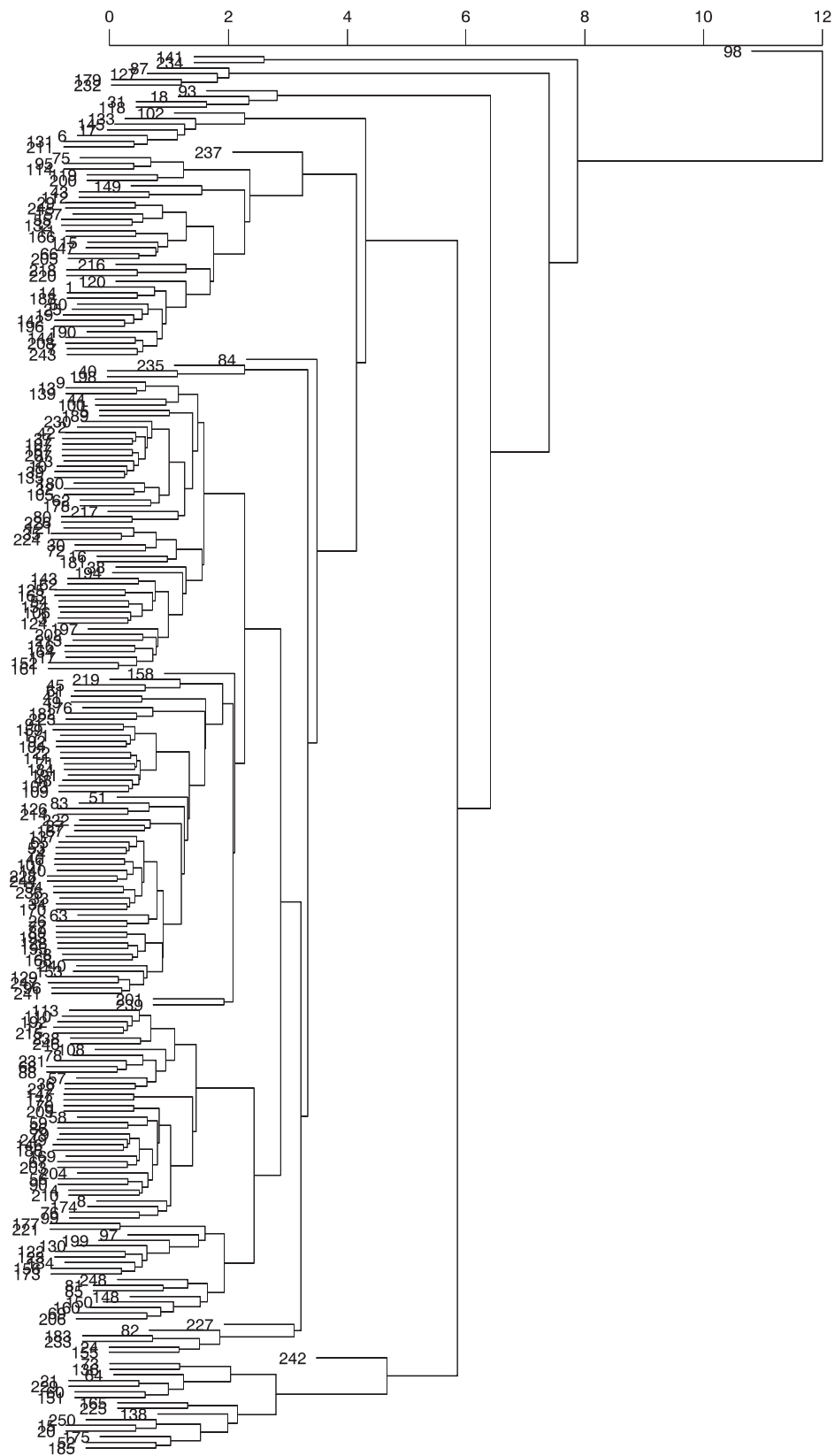


Figure 2. Average link cluster analysis of oxide data.

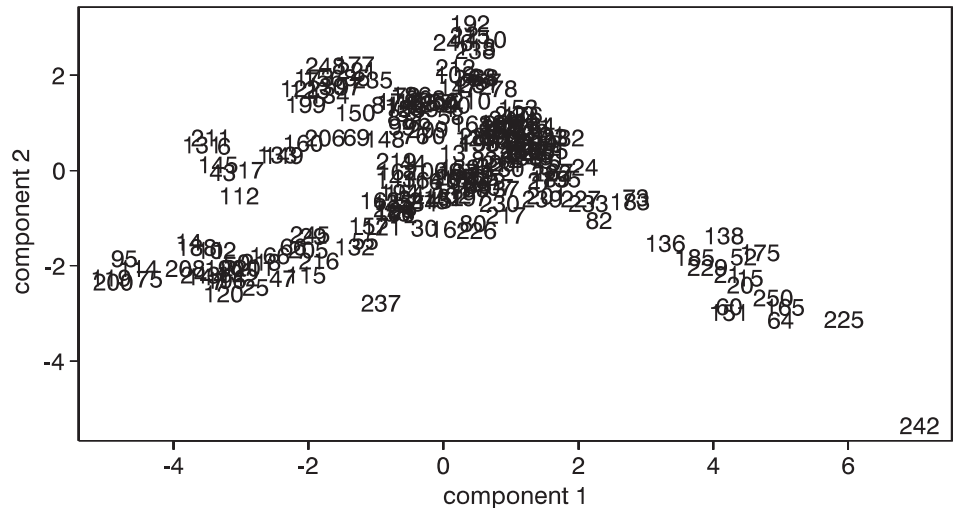


Figure 3. Principal component plot of the first two components using standardised data, after removal of the 11 clearest outliers identified in the text.

Cases (141, 234) were not originally singled out as typologically distinct. On re-inspection they are clearly window glass but the colouring, pale lime-green, is quite unusual.

The 20 or so specimens with high Mg, K and P, which were noted in the univariate and PCA analyses were also not identified as distinctive on a first examination. On re-inspection it appeared that they were mostly more ‘bubbly’ than other specimens, and this may reflect the interaction between composition and aspects of furnace technology connected with the speed at which the glass was heated.

The typological analysis also identified other cases that were either atypical (because they were Roman and/or non-window glass), or which formed small, distinctively coloured, sub-sets of the sample. Though not identified by an analysis of the major/minor oxides the majority of such cases were identified by a similar analysis of the trace elements.

5 Discussion

The work reported here is part of a broader programme that is examining approaches to the analysis of large archaeometric data sets. Only one data set has been discussed and any conclusions drawn from this can only be suggestive. What follows draws on this, and work as yet unreported. It is intended to suggest areas of study which would benefit from further research, and general guidelines that may help in the analysis of large and complex data sets.

1. The statistical analyses reported, and those conducted on the trace elements, have done rather well in identifying typologically unusual cases. It is also the case that most of the unusual data was identified using the simple univariate approaches, or ‘standard’ approaches such as PCA.

Identification of unusual multivariate data is a technically challenging problem. The simpler techniques identify the really obvious univariate and bivariate outliers, and it may be sensible to omit these from the data before attempting to identify genuine multivariate outliers. In the present example the simpler methods are all that is needed.

2. Cluster analysis can be bad at identifying clusters in archaeometric data, in the sense that the results are often method-dependent. It may, however, be quite good at suggesting outliers. It uses information on all pairs of distances between cases, rather than the distance of a case from the centroid of a sample of data so perhaps this is not surprising.

In the version of this paper presented at the CAA95 conference it was suggested that cluster analysis, though not as ‘exciting’ as the development of new methodology, could be more widely used for the detection of multivariate outliers. Some discussants noted that, in their experience, it was widely used for this purpose but rarely reported. Commonly, cluster analysis is used to detect clusters, and any outliers detected in the process are noted *en passant*. If cluster analysis is indeed used directly, and often, for multivariate outlier detection its

wider reporting would be welcome. We have seen references to the use of single-link, complete-link and average-link cluster analysis as suitable for outlier detection, but know of no systematic study comparing their merits. In particular a comparison with some of the newer methodology that is being proposed would be of interest.

3. The Atkinson/Mulira approach is easy to apply but, like other approaches based on Mahalanobis distance, assumes that a majority of the data form a coherent reference group against which the 'unusual' nature of other data may be judged. Ideally this reference group will have a multivariate normal distribution.

Identifying the reference group, which may then be used as a basis for robust analyses, is an aim in some of the literature. A theoretical ideal in some cases is that methodologies should be able to deal with up to 50% (almost) of cases that are outlying or unusual with respect to the reference group. In practice, outside of the context of simulated data, the idea of 50% of data being 'unusual' does not seem very realistic, and 20% or so may be a more reasonable limit. Leese and Main (1994), suggest a similar limit in their paper on the detection of outliers using probability plotting methods. They deal with the problem of detecting outliers relative to a known reference group, and such known grouping is not assumed in this paper.

The presence of grouping, unknown in advance of analysis, is a problem in the application of the Atkinson/Mulira and similar approaches, as suggested by figure 3. Real data are frequently clustered, rather than the bulk forming a coherent group. As a hypothetical example, if we had three equal sized, equally dispersed, and equidistant groups of data in multivariate space there is not a natural reference group against which unusual data may be judged. The outcome of the application of the Atkinson/Mulira method in this case is dependent on the initial choice of cases from which Mahalanobis distance is calculated. One possible way round this difficulty might be to identify any distinct groups in the data, in the first instance, and then identify those cases which are outlying relative to all the groups so identified. This is not a trivial task and also raises the question of sample sizes.

4. The issue of sample size has not been mentioned so far, but is a non-trivial one. Even in ideal circumstances (i.e. a single multivariate-normally distributed set of data) the ratio n/p should be in the region of 3-5 or more

for techniques of the kind discussed here to 'work'. (Recommendations vary according to context.)

Analytical techniques now available will often produce data sets with $p > 20$, and obtaining samples with large n may be costly. If the samples that are obtained have a clustered structure, and so need to be broken down into smaller sub-samples in order to apply methods for multivariate outlier detection, this will exacerbate the problem.

5. The important issues of variable selection and data transformation have been ignored for the purposes of this paper. We have, in fact, looked at the analysis of different variable subsets and logarithmic transformation of the data and found that they give rise to different results in terms of the unusual cases identified. In general none of these results are 'wrong'; it is simply that, depending on the data treatment, different 'unusual' cases are being identified. For example, analysis of the trace elements identifies small and highly coloured groups rich in Cu and Ni, that are not distinctive with respect to the major and minor oxides. Whether such specimens are to be regarded as unusual will depend on the objectives of the research; those cases just noted are unusual in terms of their appearance but not in terms of their major/minor oxide composition.

We offer one further thought here. Data transformation, to normality, is often advocated as desirable without discussion. For some of our analyses this would down-weight the visual impact of specimens in the tail of a distribution, since if the transformation is successful then the specimens will lie in the tail of a normal distribution, and not be worthy of note. However, such specimens may be of distinct archaeological interest (e.g. the 'bubbly' group noted earlier) but may be less evident in analyses where the data have been transformed. The 'bubbly' group is much more evident on a PCA of the untransformed data than on one where the data are log-transformed, for example, but this is another story.

Acknowledgements

The British Academy is thanked for funding the analyses through the Fund for Applied Archaeological Science; Caroline Jackson is thanked for the sample preparation; Dr J.N. Walsh at Royal Holloway and Bedford New College, University of London, is thanked for providing the analytical data; and Katherine Bibby is thanked for undertaking some preliminary data analyses.

references

- Atkinson, A.C.
H.-M. Mulira 1993 The stalactite plot for the detection of multivariate outliers, *Statistics and Computing* 3,
27-35.
- Baxter, M.J. 1994 *Exploratory multivariate analysis in archaeology*. Edinburgh: Edinburgh University Press.
- Brooks, S.P. 1994 Diagnostics for principal components: influence functions as diagnostic tools, *The
Statistician* 43, 483-494.
- Hawkins, D.M.
L.P. Fatti 1984 Exploring multivariate data using the minor principal components, *The Statistician* 33,
325-338.
- Leese, M.N.
P.L. Main 1994 The efficient computation of unbiased Mahalanobis distances and their interpretation in
archaeometry, *Archaeometry* 36, 307-316.

M.J. Baxter
Dept. of Mathematics
Statistics and OR
Nottingham Trent University
Clifton Campus
Nottingham NG11 8NS
United Kingdom
e-mail: mat3baxtemj@nottingham-trent.ac.uk

H.E.M. Cool
Archaeological consultant
16 Lady Bay Road
West Bridgford
Nottingham NG2 5BJ
United Kingdom

M.P. Heyworth
Council for British Archaeology
Bowes Morrell House
111 Walmgate
York YO1 2UA
United Kingdom
e-mail: m.heyworth@dial.pipex.com