Variable Selection in Archaeometry: the Statistical Analysis of Glass Compositional Data

C.M. Jackson
Dept. of Archaeology and Prehistory, University of Sheffield.
Northgate House, West Street, Sheffield S1 4ET, U.K.
E-mail: c.m.jackson@sheffield.ac.uk

M.J.Baxter
Dept. of Mathematics, Statistics and Operational Research, The Nottingham Trent University.
Clifton Campus, Nottingham NG11 8NS, U.K.
E-mail: mjb@maths.ntu.ac.uk

Introduction

In this note, we discuss some aspects of variable selection, in the multivariate, statistical analysis of glass compositional data in archaeometry. Modern analytical techniques, such as inductively coupled plasma spectroscopy (ICP), allow the concentrations of 20 or more elements to be determined, in a routine fashion (Heyworth, 1991; Jackson, 1992). Statistical methods, such as cluster or principal component analysis (PCA), can be used to investigate, whether or not there are distinct groups within the data. It is then the hope, that any grouping (or structure) can be interpreted in an archaeologically, informative fashion, relating, for example, to the provenance or technology of manufacture of the material studied. Similar procedures are widely used, in the study of archaeological ceramics and, less so, of metals (Baxter, 1994).

We have often observed that structure, revealed by a multivariate analysis can, in retrospect, be as readily displayed using far fewer variables, than those originally measured. There are purely statistical reasons, to be discussed in section 2, for preferring to work with a small number of variables. From the archaeological perspective, it is often easier to interpret and present results, if a relatively small number of variables is involved. Variable selection is, therefore, of importance.

Variable choice and selection has been recognised as an issue, ever since multivariate statistics began to be applied to archaeometric data, on a routine basis (Bieber *et al.*, 1976). However, it has not, to the best of our knowledge, been studied in a systematic fashion within archaeometry. In this note, we confine our discussion to the study of glass, since the considerations involved are specific to the materials studied and questions asked.

Statistical issues

The compositional data, generated by an analytical technique, can be collected in an n by p data table, or matrix, X, where n, is the number of specimens analysed, and p, is the number of elements determined. Many multivariate methods aim to display such data, in a lower number of dimensions (usually 2 or 3), in order that the structure or pattern in the data can be more readily seen. Often this objective can be achieved using $q << p$ variables, and the problem is how to select the q variables.

Krzanowski (1988) provides a succinct account of why, for large p, variable selection is of interest. Among these, with comments, are the following:

(a) If variables, that do not contribute to patterning in a data set, are used in an analysis they can actually obscure the pattern. It seems to be widely accepted that the more variables one can measure, the better (Harbottle, 1976; Pollard, 1986; Taylor and Robinson, 1996), but not all will necessarily be helpful in a statistical analysis.

(b) If the data are to be modelled statistically, and p is large, then very large n may be needed, for reliable estimates of parameters within the model. Even if n is large, computational difficulties can arise in modelling methodologies. While exploratory, as opposed to modelling, methodologies have predominated in archaeological studies (Baxter, 1994), there is increased interest in the latter (e.g. Buck *et al.*, 1996). Additionally, some standard techniques, used in provenance studies, centred on the use of Mahalanobis distance, are model-based, so variable selection and/or sample size is an issue.

(c) The fewer variables that are used, the easier it can be to interpret results. This will be the main focus of the rest of this paper.

Archaeological considerations and glass analysis

The analysis of most inorganic materials, in archaeology, is undertaken under two broad premises: to understand either the technology of the artefact, or, to decide upon its provenance. Early work on ceramics concentrated on provenance and was relatively successful in its application. This is because clays, and hence, ceramics, have the advantage of being linked to a particular geology.

Glass is more problematic, as distinguishing between technology and provenance is difficult chemically, because of the nature of the raw materials used, and the changes these undergo, during glass production. Ancient glasses are usually manufactured using a two component system – silica, the main glass former, and an alkali, in the form of a geological salt or plant ash, which acts as a flux to the silica. These two components can derive from a large number of organic or inorganic sources, each of which contributes many different elements to the overall glass chemistry. In neither case, can either component be linked to any specific source.

This makes glass chemistry particularly complex, and makes it imperative, in this case, that the choice of variables has to be very specifically linked, to the archaeological question posed.

To make glass, specific raw materials, that contain the elements which are the main glass formers, fluxes, or stabilisers are needed. For example, sand (or quartz) is an essential ingredient in glass manufacture, because it contains silica, which is a major glass former. In using sand, along with Si, other elements such as Fe, Al, Ti may be 'accidentally' introduced. In this sense, Si is deliberately introduced, via the sand, with other elements being included accidentally. Similarly, in glass manufacture, a source of alkali is needed. The use of different alkali-rich, raw materials will produce glasses, where the main flux is different – either Na or K (or both) - so that the presence of some elements may be deliberate or accidental, according to the choice of raw material. Whereas ingredients containing Si (or an element playing a similar role) and an alkali are essential, and intentionally added, other ingredients can be added intentionally, but are optional. These include elements that affect the appearance of the glass, such as its colour or opacity, and may include Co, Cu, Mn, Sb, Sn and Pb.

From the point of view of variable selection for statistical analysis, the consequence of this discussion is that different elements have different statuses. They perform different functions, may be intentional or accidental inclusions, and may be essential or optional components of the glass. A further complication is that, whether an element can be regarded as an intentional or accidental inclusion, will depend on the specific raw materials used in the glass making.

Case studies

Here, some glass data sets, that have been published and discussed in the literature (not necessarily using multivariate statistics), are studied briefly, to illustrate particular issues arising in relation to variable selection. For reasons of space, the presentation is in summary form; we hope to present a more detailed analysis elsewhere.

Expert selection

It is often the case that a choice of variables, to present the main features of a data set, is made without recourse to statistical methods, on the basis of expert knowledge, or because the main features are obvious. This may often be the case with highly coloured glasses.

Deep blue glass is coloured, using minerals containing the element cobalt. However, in addition to cobalt, many of these minerals contain impurities, eg., iron. Therefore, by producing a scatterplot of Fe vs. Co in the glasses, it may be possible to suggest the use of different sources of colorants, and hence, to define different groups of glasses.

This has been used to good effect by Henderson (1991, Figure 6), who used these variables to discriminate between a number of Iron Age, deep blue beads from Northern England, which had already been grouped, using typological criteria. He was able to show that three types of beads separated well, and was able to infer separate workshops, based upon the use of different cobalt sources.

In this case, a small number of easily selected variables, which could be correlated to intentional additives, could be predicted and used, without the need for more complex analysis.

Iterative analysis

It seems reasonable to suggest that the initial investigation of the data should always include those variables, which are essential to glass forming and which will define the basic glass type. However, it may not be necessary to include such variables in the later stages of analysis.

For example, in the analysis of French Medieval glass, given in a paper by Barrera and Velde (1989), in a sample of over 400 glasses, two groups are formed based on Na levels alone – glass high in Na and very low Na. The data, therefore, consists of two groups of glasses, formed using different alkalis.

If the initial groupings are separated and the data investigated further, for the low soda glass, three compositional groups can be defined, based upon K, Ca and Mg. These groupings can be linked to the original typological groups, identified by Barrera and Velde. As their initial aim was to investigate provenance of different types of glasses, based upon chemistry, this two-stage method was successful.

This suggests that in some cases, single-stage selection of variables may not be appropriate, and that some form of staged analysis, where different variables are used at different stages, may be appropriate.

Automatic variable selection vs. archaeological question

In the first two examples, clear groups in the data can be defined, using a small number of variables. In this next example, the same is true, but it is not obvious which, of several possible choices of variables, should be made.

Cox and Gillies (1986) analysed a number of blue Medieval glass samples, from the windows of York Minster and various archaeological excavations, and used cluster analysis, with 11 variables to establish that there were three main groups in the data. Baxter (1989) reanalysed the data, using principal component analysis, and the data have also been studied, more recently, using neural networks (Bell and Croson, 1998).

In fact, the structure in the data is so obvious, that it can easily be presented with a bivariate plot, using two variables; the problem is that this can be done for several choices of pairs of variables. For example, the choice of Na vs. K, Ca vs. K, or Si vs. Al shows three different groups, as does a plot of Cu vs. Fe (This is illustrated in Baxter and Henderson, submitted for publication, Figures 1 and 4).

The groups, defined by K and Na, suggest differences in alkalis, which may be related to provenance, those defined by Si and Al, by differences in sand sources, again, probably related to provenance, whilst those defined by Cu and Fe,

which can act as colorants, may be interpreted in terms of technology. Any choice is acceptable if one wishes simply to demonstrate there are three chemical groups in the data, but one or the other may be more appropriate, depending upon the archaeological question, that the analysis seeks to address. Automatic variable selection, in this instance, would miss this feature of the data.

Selection for interpretation

A related example, that introduces further considerations in variable selection, is based on data on Medieval glass compositions, collected by Heyworth (1991). One of the data sets is from Saxon Southampton (Hamwih) and consists of over 200 specimens, the majority of which are characterised, by Heyworth, as light green or light blue. Multivariate analyses, using 11 major and minor oxides, suggest two main (overlapping) concentrations in the data, corresponding to the two colours. The PCA plot that shows this is closely matched, by a plot of Fe vs. Mn, only, and the grouping can be clearly displayed, by a plot of the Fe/Mn ratio.

Since the grouping is colour related, this is not surprising, given the roles that Fe and Mn are known to play in glass coloration. Heyworth (1991) discusses this at length, noting that light-blue (the predominant colour) is the more difficult colour to achieve accidentally. He hypothesises that the predominance may reflect a deliberate attempt to achieve this colour, which may also be chronologically related.

Here a satisfying technological and archaeological explanation can be advanced, for the main patterning in the data, that only two variables are needed to reveal. For this, and other data sets, Fe is highly correlated with other elements, particularly Ti. If automatic variable selection, is used it would usually be the case that only a subset (possibly one) of such correlated variables would be needed, but there is no guarantee that the variables that aid archaeological interpretation will be chosen. In the example being discussed, Ti could equally well replace Fe in the plot, and would give the same pattern, but would not aid interpretation.

Hidden and unexpected archaeological information

Sometimes when variable selection may seem obvious, there are other factors which may be hidden in the data and whose interpretation is desirable. This can be illustrated, by reference to the comparison of blue-green and colourless Roman glasses, from Coppergate, York (Jackson, 1992).

Glasses are naturally coloured blue or green by iron impurities, derived from the sand. To produce a colourless glass, either antimony or manganese must be added. Initial variable selection is therefore easy, in this case, if differences in colour are to be investigated. Plotting Sb and Mn showed two distinct groups, based upon colour. However, subsequent examination of the data, by discriminant analysis, showed that these two groups could be defined by another suite of variables, not associated with the mineral additives, and which were not immediately obvious. These included Fe, Al, and P elements, correlated with the silica. This would indicate that for colourless glass, treatment of the raw materials, before glass production, or selection of specific sands, containing fewer impurities, had taken place.

In this example, an initial archaeologically driven choice of variables confirmed the existence of two chemically and visually distinct groups. Further multivariate analysis suggested other, interpretable differences, based on other variables, that were not immediately obvious. To fully understand the data, both sets of variables need to be taken into consideration.

Conclusions

In all the examples discussed, and many others that we have seen, a small number of the measured variables could be used to present and interpret the archaeologically important features of the data. This suggests that variable selection is important, if for no other reason than, it simplifies both the interpretation and communication of results.

In three of the examples (a, b, e), initial variable selection was based on expert opinion, rather than being guided by statistical analysis, while in three cases (c, d, e), multivariate statistical analyses either drew attention to patterns that could be more simply explained, or revealed features of interest that were not immediately apparent. In four cases (b, c, d, e), either different subsets of variables conveyed distinct archaeological information, or conveyed similar statistical information but differed in their archaeological importance.

The examples have been chosen to illustrate different aspects of the variable selection problem in a broad context. From a narrower perspective, our original interest was whether or not variable selection techniques, as developed in the statistical literature, might prove to be of use. We have not discussed the technicalities of these, but typically, a selection mechanism will result in the choice of a single subset of variables, intended to optimise some criterion of 'success' (e.g., group separation in discriminant analysis), or will result in a choice that is the outcome of some 'plausible' selection algorithm, having no obvious optimality properties. In general, different methods or algorithms result in different choices.

With respect to glass analysis, automatic variable selection methods may prove to be unhelpful for various reasons.

(I) Some variables should always be included in an initial analysis (e.g., Na, Ca, K), because of their potential importance in determining the basic glass composition. Upon subsequent analysis, these may be omitted, if it is known that the glass of interest conforms to the same broad type.

(II) If a set of variables is highly correlated, then for statistical discrimination, only a subset of these may be needed. This may be satisfactory for descriptive purposes. For archaeological interpretation, however, one may need to take account of more variables, than a purely statistical approach would dictate (e.g., example (e)).

(III) Sometimes one or two variables can be selected to answer specific questions, but would not necessarily be those, selected on purely statistical grounds. Variable selection should not be based on automatic choice, but be case specific, and most important, it needs archaeological input.

(IV) Iterative analysis is important, if we are to gain a full understanding of the material under study, even if not all the variables are used in the final analysis. Such analysis may result in several different subsets being selected and interpreted. Automatic methods of variable selection do not usually allow for this.

We emphasise that we are not concluding that (multivariate) statistical analysis is unhelpful. Indeed, the opposite was the case in some of the case, studies discussed. Rather, the argument is that methods of variable selection, in which purely statistical considerations are the guiding force, are likely to prove unsatisfactory. Variable selection is important, but any statistical approach depends upon the material studied, and *more important*, needs to be firmly guided by archaeological understanding.

Stated thus, this perhaps seems obvious. However, in applying scientific, mathematical, and statistical methods to the analysis of archaeological materials, the obvious has, arguably, often been forgotten.

Bibliography

BARRERA, J. & VELDE B. (1989), "A study of French Medieval glass composition", *Archaeologia Medievale,* 19, pp. 81-130.

BAXTER, M.J. (1989), "Multivariate analysis of data on glass compositions: a methodological note", *Archaeometry,* 33, p. 29-41.

BAXTER, M.J. (1994), *Exploratory Multivariate Analysis in Archaeology*, University Press, Edinburgh.

BAXTER, M.J. & HENDERSON, J. (submitted for publication), "Variable selection for graphical plotting in archaeometry, with an emphasis on studies in ancient glass", *Archaeometry.*

BELL, S. & CROSON C. (1998), "Artificial neural networks as a tool for archaeological data analysis", *Archaeometry,* 40, pp. 139-151.

BIEBER, A.M, BROOKS D.W., HARBOTTLE G. & SAYRE E.V. (1976), "Application of multivariate techniques to analytical data on Aegean ceramics", *Archaeometry,* 18, pp. 59-74.

BUCK, C.E., CAVANAGH W.G. & LITTON C.D. (1996), *Bayesian Approach to Interpreting Archaeological Data*, Wiley, Chichester.

COX, G.A. & GILLIES K.J.S. (1986), "The X-ray fluorescence analysis of Medieval durable blue soda glass from York Minster", *Archaeometry,* 28, pp. 57-68.

HARBOTTLE, G. (1976), "Activation analysis in archaeology", in: NEWTON G.W.A. (ed.), *Radiochemistry* 3, Chemical Society, London , pp.33-72.

HENDERSON, J. (1991), "Industrial specialisation in late Iron Age Britain and Europe", *Archaeological Journal,* 148, pp. 104-148.

HEYWORTH, M.P. (1991), *An Archaeological and Compositional Study of Early Medieval Glass from North-West Europe*, unpublished PhD thesis, University of Bradford.

JACKSON, C.M. (1992), *A Compositional Analysis of Roman and Early Post-Roman Glass and Glassworking Waste from Selected British Sites*, unpublished PhD thesis, University of Bradford.

KRZANOWSKI, W.J. (1988), *Principles of Multivariate Analysis*, Clarendon Press, Oxford.

POLLARD, A.M. (1986), "Multivariate methods of data analysis", in: JONES R.E. (ed.), *Greek and Cypriot Pottery: A Review of Scientific Studies*, British School at Athens Fitch Laboratory Occasional, Paper 1, Athens , pp. 56-83.

TAYLOR, R.J & ROBINSON V.J. (1996), "Neutron activation analysis of Roman African Red Slip ware kilns", *Archaeometry,* 38, pp. 231-243.