GRAPHICAL APPROACHES TO MULTIVARIATE DATA ANALYSIS USING ARCHAEOLOGICAL DATA

KATHERINE J. BIBBY

A thesis submitted in partial fulfilment of the requirements of The Nottingham Trent University for the degree of Master of Philosophy

FOR REFERENCE ONLY

Department of Mathematics, Statistics and Operational Research

March 1997

ACKNOWLEDGEMENTS

For Francis, Mum, family and friends, for all your love and support over the last 25 years. Thank you for always being there for me, giving me guidance and encouragement.

To Barrie, for all the happy times spent teaching at Nottingham Trent. Thank you for all those words of wisdom.

Many thanks to Mike Baxter and Neville Davies for their invaluable assistance and numerous other colleagues for all their help and advice. Thanks also to Caroline Jackson, Hilary Cool and Mike Heyworth for allowing me to use their data.

DECLARATION

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or institute of learning.

ABSTRACT

This thesis investigates the application of some recently developed statistical methodology to problems arising in the analysis of multivariate archaeological data. Specifically, data sets on the chemical composition of glass fragments found in archaeological contexts are used.

The statistical methods often used to investigate such data (for the existence of groups, etc.) are sensitive to the presence of outliers. One focus of the thesis is a comparison of the performance of different methods of outlier detection with such data, including recently developed methodology.

Removal of outliers makes it easier to detect patterns in the data. Standard methods such as principal components analysis and cluster analysis are used for this purpose. Results are displayed using kernel density estimates (KDE's) in a variety of ways. Although KDE's are now an established statistical technique, their application to archaeological problems is comparatively novel.

For the data sets used here the newer outlier detection methods usually differed little from the methodologies they are supposed to improve on, in terms of outliers detected. They are also not ideally suited to data sets having the kind of structure exhibited by those used. The KDE's proved valuable in displaying structure in the data, and it was often possible to provide a substantive explanation for the structure in terms of glass chemistry and colour. It was also observed that the statistical outliers detected could, in retrospect, be recognised to be archaeologically or physically unusual with respect to colour.

The substantive analyses raise a number of interesting questions. For example, observed structure may be related to quite subtle colour difference (reflected in the chemistry), but colour is often not recorded. Even where it is, sample sizes smaller than those used here may not allow the detection of structure. Finally, one analysis revealed structure that will require further archaeological investigation.

Contents

List of Figures

List of Tables

1. Introduction

1.1 Aims

Outlier detection is common and important in archaeometry, both because of the way in which outliers can affect the procedures used to process data statistically, and because such specimens may be of interest in their own right. Throughout the course of the thesis an outlier in a set of data is defined to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data, (Barnett and Lewis, 1994). In this thesis we aim to evaluate the performance of different methods of multivariate outlier detection, in the particular context of the analysis of archaeometric data sets. We look at the development and exploitation of new methods of data display for visualising groups in archaeometric data. A substantive investigation is made of the role played by colour in the formation of compositionally distinct groups, revealed by multivariate data analysis after outlier removal. Finally, an assessment is made of the extent to which patterns in the data can be displayed using simpler graphs based on a small subset of the variables which are possibly suggested by multivariate data analysis.

We conclude that data visualisation is important with large data sets where conventional plots may be difficult to interpret, and that in turn, the use of kernel density estimates make it clear that for some of the data we examine, although not all, the colour of the glass plays an important role in the grouping. Although, using smaller data sets, or where colour is not recorded, KDE's are not as useful. In some of the analyses, the structure of the data can often be seen using just a subset of those oxides that particularly influence colour and so it is of interest to see if this is the case across all the data. For one of the data sets, where there is colour separation, the plots based on iron and manganese do very well, but the same also appears to be true for another data set, where there is no colour separation. After the elimination of outliers and carrying out multivariate data analysis it is possible to see the main structure in the data using somewhat simpler plots than those obtained from principal components analysis. Such plots also have much simpler interpretation.

1.2 Background

Initially, a computer based learning (CBL) module, funded by the Teaching and Learning Technology Programme (TLTP) was developed for use in both the teaching and learning of methods of exploratory data analysis, namely the histogram. The CBL histogram module was successfully completed and a paper entitled 'STEPS for Learning Statistics', published in 'Teaching Statistics'. Work began on a CBL kernel density estimate module, a version of which was demonstrated at the 1995 Computer Applications in Archaeology (CAA) Conference in Leiden. This was centred on a set of MATLAB routines written by Dr C. Beardah of the Dept. of Maths, Stats and OR at The Nottingham Trent University. At the time of writing these routines it was not anticipated how rapidly development would go as a direct result of Dr Beardah's collaboration with other experts on aspects of data presentation in archaeology, and the development of the CBL module was abandoned. Dr Beardah's routines, outlined and discussed in Beardah and Baxter (1996) and Baxter, Beardah and Wright (1995), now form the basis of Chapter 3 in this thesis where we describe the use of kernel density estimation (KDE) for detecting features in both univariate and multivariate data. In Chapters 5 and 6 the routines are then used to identify features in multivariate archaeological data. A paper written by Beardah and Baxter (1996), 'The archaeological application of kernel density estimates', has appeared in the first Internet Archaeology journal in September 1996 and all routines and help facilities are readily available.

Early on in the course of the work, several multivariate data sets relating to the chemical composition of archaeological glass became available, some of which had not previously been subjected to detailed statistical analysis. These data sets had originally been analysed by the archaeologists involved mainly for their chemical content using inductively coupled plasma spectroscopy (ICPS). These, along with others, were originally to be used to investigate various aspects of the graphical analysis of multivariate data, such as outlier detection. However as a result of my analyses and observations, it became apparent that there were matters of substantive interest that could be addressed. Thus the emphasis of the thesis has shifted, in part, from purely methodological comparisons to an investigation of these substantive issues. The substantive issues include the following. After outlier removal and using KDE's, there is clear grouping in the largest data set that is associated

with glass colour. This might be expected when analysing glass data but colour is often not recorded, so patterns which are detected in glass data, might be attributable to this but are not obvious. Many data sets collected are much smaller than the one mentioned above, thus multivariate analyses might not detect the differences. Therefore one focus of this thesis is the extent to which multivariate analysis seems to separate the glass for what appear to be reasons of colour. Given this separation it became clear that it was also possible to summarise the main patterns in the data using just two of the variables, one of which was highly correlated with several others. We examine if this is a consistent pattern across all data sets. As mentioned above, the thesis addresses both methodological and substantive issues. On the methodological front. PCA and KDE are standard techniques but the latter has been little used in archaeology. The outlier detection methods used in the thesis are relatively new and are applied to data where there are potential problems. Therefore we try to assess the practical similarities and differences between the results arising from different methods of outlier detection. why such differences occur and which methods are preferred.

1.3 Structure

In Chapter 2 the manufacturing process of ancient glass is outlined and we describe the different data sets used. In Chapter 3 we discuss principal components and cluster analysis and, in particular, their uses in the detection of outliers. We describe kernel density estimation and its uses for detecting features in both univariate and multivariate data. We also describe how kernel density estimation can be used for identifying multimodality, or groupings in multivariate data. In Chapter 4 we present a thorough investigation into the varying outlier detection methods that are available, presently an important area of research, and compare their functionality. This chapter includes a section, 4.8, where a particular problem arising in the detection of outliers is described. The problem takes the form of the differing results observed from a particular outlier detection method if distinct groupings are present in the data. An assemblage of glass excavated from York Minster is used to illustrate this. In Chapter 5 we turn our attention to a particular data set, an assemblage excavated at Southampton. Each technique and method described in Chapters 3 and 4 is applied to this real data, giving us insight into outlier detection methods by

considering their effect on a data set. In Chapter 6, four ancient glass data sets are analysed in depth using the differing methods, described in Chapters 3 - 4. The substantive issues raised after initial analyses of the five data sets are also discussed in Chapters 5 and 6. Chapter 7 has two main strands, discussion of the different methods for detecting outliers, and discussion and conclusions of the substantive results, showing to what extent the methodology used is useful for exploring glass compositional data sets.

2. Glass chemistry and the data sets used

2.1 Raw materials of glass

Ancient glass is mainly composed of *silica*, and the main source of silica for glass manufacture is sand, from both unconsolidated deposits and sandstones. The physical and chemical nature of a sand deposit, whether consolidated or not, is influenced by the type of rock from which it is derived. For example sediments resulting from the weathering of crystalline rocks tend to have a higher feldspar and heavier mineral content. Feldspars frequently contain some form of alkali - aluminium, sodium and potassium - and these alkalis will be incorporated into the final mix. For colour control the most important undesirable impurity is iron, in the form of ferric oxide, Fe_2O_3. In the manufacture of colourless glass Fe_2O_3 is most undesirable and sands suitable for making colourless glass should have a low Fe_2O_3 level so the glass will not have a characteristic green tinge.

Alkalis are added to the glass to reduce the melting temperature. From the type of alkali alone, ancient glasses can be divided into two very broad compositional groupings based on the concentrations of K, Na and Ca. Glass made with plant alkalis, also known as plant ash glass, is high in K and Ca and low in Na. This type of glass is less durable, prone to weathering, becoming opaque and often disintegrating. Glass made with salt water plant alkalis is high in Ca and Na, and often Mg.

Most ancient glasses which do not contain any deliberately added colorant, show a pronounced 'natural' green colour or tinge. This is because they contain appreciable levels of *Iron*, usually 0.3 - 1.5 % as oxide, which is present as a contaminant of the raw materials. Sand often contains variable amounts of Fe, which is still present after refining. Some plant ashes also contain Fe (up to 0.4 %). The colour Fe imparts ranges from bluish aqua → green → yellow green → olive → brown, which is related to the proportions of ferrous and ferric Fe present respectively. This in turn also depends on the atmospheres which prevail when the glasses are melted and upon the presence of various redox species present within the melt. When Fe levels are increased in the glass so are the Al levels - this indicates either the use of a purer sand source or a strict refining process.

Phosphorus has an influence on colour and the presence of large amounts of P_2O_5 can decolorize the yellow ferric ion into a colourless complex ferric phosphate. However in most Roman glass there is too little P_2O_5 in the solution to decolorize the glass.

Manganese can act as both a colorant and decolorizer. It can oxidise Fe and, by its own colour, compensate for the green shade which iron produces in the glass. At low levels, Mn is better known as a decolorant in Roman glasses, where it oxidises the blue/green ferrous ions in the glass.

Antimony, Sb, is a stronger oxidising agent and thus a more efficient decolorizer than Mn, producing a more brilliant glass. Its use for decolorising glass is based on raising the internal oxygen pressure of the melt to the extent that FeO is oxidised to Fe_2O_3 and gaseous oxygen is liberated. (Henderson, 1985).

2.2 Decolorizers

Addition of Mn and Sb falls into two groups; high concentrations around 1% and low concentrations around 0.1%. The low concentration can be attributed to the result of the normal inclusion of impurities, and the higher concentration in colourless glasses, the result of deliberate addition. The introduction of these oxides to glass would primarily be to remove discoloration due to the Fe through oxidation and chemical complexing. Introduction of Sb was correlated with the production of low Mn and K glasses - Sb later gradually being replaced by Mn. Glasses excavated from the late 2nd - 3rd centuries have high levels of Sb and glasses from the late 3rd - 4th centuries have high Mn levels thus indicating the shift from the use of Sb to Mn, (Heyworth, 1991). The addition of 0.1% Sb will have a much greater effect on the colour of glass than 0.1% Mn as it is a much stronger decolorizer. Thus the shift from Sb to Mn is unknown, although it was thought to relate to the chemicals that were readily and more easily available during the 2nd - 4th centuries, (Heyworth, 1991). Amounts of Mn and Sb, if added deliberately to decolorize glass, will to some extent depend on the amount of Fe in the initial glass mix - higher amounts of Fe require increasing amounts of decolorizer to successfully produce colourless glass.

2.3 Colour

The production of colour in glass not only depends upon inclusion of a specific metal oxide and the way the colorant is mixed into the batch, but it also depends upon the presence of other oxides in the batch, the furnace temperature, length of firing, type of fuel and the state of oxidation or reduction in the furnace (gaseous atmosphere). It must be noted that the vivid green colour of many glass fragments may also be linked to the presence of both Cu and Fe, and a blue coloration can also be due to a combination of Cu, Fe and Co. The minimum amounts of Cu and Co required to produce a blue colour are 0.6 and 0.02% respectively. The majority of the glass found in Britain in the Roman or immediate post-Roman period is not strongly coloured but light green or light blue. It is presumed to not have colorants or decolorants deliberately added, therefore the colour is dependent primarily upon the concentration of Fe present in the raw materials. Blue/green glass is the most simple and inexpensive to produce requiring less skilled labour, less control of the furnace conditions to influence specific colours and presumably no addition of colorants/decolorants. The main mean intra-site compositional differences between the groups of light blue and light green glass (although slight) appear to be between levels of Fe, Mn, SbO and Cu.

2.4 Relationship between Fe and Mn

In glass containing both Fe and Mn, Mn acts as an oxidising agent changing the FeO to Fe_2O_3, itself being reduced to MnO. In many cases there appears to be a correlation between the colour and the concentrations of the Fe and Mn in the glass. However it was first noted by Geilmann *et al* (1955), who analysed 39 Medieval glasses containing a range of Fe and Mn concentrations, that sometimes there appeared to be no correlation between the colour and the concentrations of the oxides. This work was extended by Sellner *et al* (1979), and it was concluded that the individual colours were produced by particular oxidation states of the colouring oxides which are not revealed by chemical analyses. Colour can be affected in three ways - by the variations in composition, the time spent in the molten condition and by the atmosphere in the furnace. A major contributor is the redox equilibrium between Fe and Mn, the ratio being found in glass at around 0.4%. To successfully decolorize the glass, the Mn would need to be added in quantities

approximately equal in weight to the Fe already present. The mean ratio of Fe:Mn seen in the glasses varies between 1.1:1 and approximately 2.1:1, indicating a possible role as a decolorizer, since an excess of Fe:Mn produces a light blue colour, whilst an equal or excess of Mn:Fe produces a light green colour, (Jackson, 1992),.

2.5 Different furnaces - oxidizing and reducing

The correlation between the glass and the ceramic colour may also be a function of differences in the redox conditions within the furnace. The ceramic colour of the melting pots, used to melt the glass, may themselves have contained colourants which during the melting process slowly 'leaked; into the glass mixture. Glass melted in an oxidising atmosphere and cooled slowly after subsequent melts would tend towards a greener colour. Glass formed under reducing conditions would tend to be a blue/green in colour and would stay this colour unless induced to change by the addition of an oxidising agent. When Fe is the only abundant transition element, increased melting times and temperature lead to a light blue colour, whilst a yellow tint depends upon short melting times.

2.6 General technical points on glass/glass-making procedure

Looking at the composition of glass, in particular those elements which appear to colour the glass, Fe and Mn influence the colour of glass batches. Fe is usually found as a minor oxide which enters the glass batch as an impurity, and the colour of iron-containing glass is strongly influenced by the furnace atmosphere during melting. A blue colour is produced in strongly reducing conditions, a blue/green or green colour in less strongly reducing conditions and a yellow/brown colour in oxidising conditions. Mn is again usually found as a minor oxide entering as an impurity. This oxide can deliberately be added to glass to act as a decolorizer, where the glass contains high levels of Mn (>0.5%), this is the result of deliberate addition, (Heyworth, 1991). When looking at lightly tinted glass there is a further complication, this being the ratio between Fe and Mn. Since both Fe and Mn would be present as unrecognised impurities of components in a glass batch, the resulting ratio of Fe:Mn would have been an important factor in the final tint of the glass.

In conclusion it may be assumed that the oxides Fe and Mn played an important role in the Medieval glass industry. Mn was indeed added as a decolorant to light green glass batches but in turn the actual furnace conditions also played a large part in the glass-making

process. The following analyses take into account the above details of the Medieval glass-making industry.

2.7 Description of the five glass assemblages

In Chapters 5 and 6 five glass assemblages excavated from various sites across the UK will be analysed in depth. The description of the data used is as follows.

Southampton glass - this assemblage consists of 271 specimens which date back to the early Medieval period, 9/10th century AD (Heyworth, 1991). The content levels of 11 of the major/minor oxide components of glass will be analysed, namely Al_2O_3, Fe_2O_3, MgO, CaO, Na_2O, K_2O, TiO_2, P_2O_5, MnO, PbO and SbO. The majority of the pieces are coloured light blue or light green and, after in depth analyses, the data separate into two main groupings based on colour.

Winchester vessel glass - this assemblage consists of 102 specimens which date from the late Roman period, 4th century AD (Heyworth, 1991). The colour of this assemblage is predominantly light green, but also includes some blue pieces. As with the Southampton glass, the same 11 major/minor oxides are analysed. Again the data can be separated on the basis of colour, those specimens coloured green/light green and those coloured blue.

Winchester window glass - this assemblage consists of 44 specimens which date from the 7th century AD, up to and including the 11th century AD (Heyworth, 1991). The data fall into four typological groups (Heyworth, 1992) : durable glass of 'early' type, 7-9 centuries AD; durable glass of 'late' type, 9-11 centuries AD; durable blue glass, later than the 10th century AD; non-durable glass, later than the 10th century AD. As with the Southampton glass, the same 11 major/minor oxides are analysed. After analysis the data do fall into four distinct groupings, which may also separate on the basis of colour, with three groups consisting of light blue pieces and one group consisting of blue pieces.

Winchester cullet glass - this assemblage consists of 250 specimens of glass which date from the 7th century AD. The 250 pieces were selected visually to be representative of the pieces found in a pit thought to be an ancient cullet (or waste glass) bank. Analyses are

undertaken using 9 major/minor oxides Al_2O_3, Fe_2O_3, MgO, CaO, Na_2O, K_2O, TiO_2, P_2O_5, MnO and the trace elements Ba, Co, Cr, Cu, Li, Ni, Sr, V, Zn. A majority of the pieces are coloured blue-green and much of the analysis concentrates on this blue-green glass only. As with the Winchester window glass, specimens of the same colour separate into distinct groupings.

Coppergate glass - this assemblage consists of 233 specimens which date from the early Roman period, 1st - 4th centuries AD (Jackson, 1992). The fragments are mainly blue-green in colour but the batch does include light blue, light green, and colourless fragments. The Coppergate assemblage also includes a collection of crucible waste glass which dates to a different period. This is removed from any further analyses and we concentrate primarily on the blue/green glass. The content levels of 11 of the major/minor oxide components of glass will be analysed, namely Al_2O_3, Fe_2O_3, MgO, CaO, Na_2O, K_2O, TiO_2, P_2O_5, MnO, PbO and SbO.

York Minster window glass - analysed in Chapter 5. This assemblage consists of 27 specimens of window glass mainly from York Minster but also from excavations, and dates from the Medieval period. Analyses are undertaken using the 11 major/minor oxides Al_2O_3, Fe_2O_3, MgO, CaO, Na_2O, K_2O, TiO_2, P_2O_5, MnO, PbO and SbO. After analysis the data separates into three distinct groupings. It is used in Chapter 5 to illustrate one of the problems concerned with outlier detection methods.

3. Data exploration and display

3.1 Principal component analysis

In archaeological practice for artefacts made of ceramic, glass, metal, etc., the chemical composition is often informative about the source of the artefact or the technology used in its production. Chemical analysis leads to data collection which in turn leads to statistical analysis. It is difficult to directly see structure, (e.g. groups), in the data so methods of multivariate data analysis are used to obtain a clearer picture. Two of the most common are principal component analysis (PCA) and cluster analysis and firstly we discuss PCA.

Measurements are typically available on the concentration of p oxides/elements in each sample of n objects/observations. If univariate or bivariate exploratory data analysis identifies clear outliers in the data, then it would be sensible to omit such observations from a subsequent PCA. PCA is a methodology for exploratory data analysis of multi-dimensional data and it involves the construction of p new uncorrelated variables or components that are linear combinations of the original variables. Often, a PCA is used to obtain a 2 or 3 dimensional picture from p-dimensional data, where p>3 and the data cannot easily be visualised directly. It is hoped that a plot based on the most important 2 or 3 components will reveal important features, e.g. the presence of outlying observations and the presence (or absence) of chemically distinct groups in the data. In practice, if the first 2 or 3 components account for a 'good' percentage, e.g. 60%, of the variation in the data, then the component plot will be reasonably informative, (Jackson, 1991). The raw data consist of measurements, usually in %'s, of the major/minor oxide composition of an object and/or trace element compositions measured in parts per million (ppm). The number of oxides/elements, p, can range from 3 to 30 (Baxter, 1993; 1994).

3.2 Notation and Theory

The p variables measured will be denoted by $X_1, X_2, \ldots, X_p$. The measurement for the j'th variable on the i'th object is x_{ij}. Usually the x_{ij} are transformed and/or standardised in some way before a PCA. The variables that result from such a transformation/standardisation will be denoted by $Z_1, Z_2, \ldots, Z_p$. The $n \times p$ matrix of the raw data is X, and of the modified data is Z. In the standard approach to PCA, new variables of the form

$$Y_i = a_{i1}Z_1 + a_{i2}Z_2 + \ldots + a_{ip}Z_p$$

(3.2.1)

are defined with the property that they are uncorrelated and Y_1 has the highest variance, Y_2 the second highest variance, and so on. The a_{ij} can be obtained from a singular value decomposition of Z, or from an eigen-analysis of the covariance or correlation matrix of Z. In the latter case the a_{ij} are just the coefficients of the i'th eigenvector. and the variance of Y_i, σ^2_i, is the i'th eigenvalue. The eigen analysis approach is briefly described below .

3.3 Eigen Analysis of the covariance (correlation) matrix

We define

$$z_{ij} = \frac{\left(x_{ij} - \overline{x_j}\right)}{s_j^{\alpha}}$$

(3.3.1)

Where s_j is the estimated standard deviation of X_j and define

$$S = \frac{Z'Z}{(n-1)}$$

(3.3.2)

Now if $\alpha = 0$ then S is the estimated covariance matrix of the data and if $\alpha = 1$ then S is the estimated correlation matrix. Since this is standard methodology see Jolliffe (1986) and Jackson (1991) for further reference.

3.4 Standardisation

The importance of a variable in a PCA is related to the variance of that variable, s_j^2. With widely differing variances, a PCA will be determined by a subset of the variables with the large variances. For this reason it is common to standardise the variables to have a zero mean and unit variance, by dividing by s_j, the estimated standard deviation of X_j. The standardised value of an observation, z_{ij}, is as shown in (3.3.1) where $\alpha = 1$. Since standardisation gives the variables equal variance, therefore in turn equal weight, each variable may then potentially contribute to the PCA. If the data are measured in different units then standardisation is essential.

In some circumstances, standardisation of the variables before analysis may improve chances of a simple interpretation. All measurements of the data used in this thesis are made on the same units and therefore it could be argued that the covariance matrix might be more appropriate, but the correlation matrix is used because all the variables are then treated on an equal footing. Logarithmically transforming data can also aid the interpretation process and in Chapter 6 log-transformed data have been used due to the structure of the data being clearer using this rather than standardised data.

3.5 Usage of PCA in this thesis

An objective of this thesis is to investigate techniques for detecting multivariate outliers. According to Jolliffe (1986, 174) one major problem in detecting such outliers is that an observation may not be extreme in any of the original variables, but it can still be an outlier because it does not conform to the correlation structure of the remainder of the data. It is impossible to detect such outliers by looking at the original variables individually (Jolliffe 1986, 174). The first few, or the last few principal components can be used in order to detect outliers. These principal components will detect different types of outlier and, in general, the last few are more likely to provide additional information which is not available in plots of the original variables. As discussed in Gnanadesikan and Kettenring (1972), the outliers which are detectable from a plot of the first few principal components are those which inflate variances and covariances. If an outlier is the cause of a large increase in one or more of the variances of the original variables, then it will be extreme on those variables and thus detectable simply by looking at plots of the original variables. Alternative tests for the detection of outliers are available and these are discussed further in Chapter 4. Principal components analysis is additionally used in Chapters 5 and 6 to investigate compositional structure in the data and to investigate if archaeological types cluster together on a component plot based on the chemical data.

3.6 Cluster analysis

Cluster analysis can also be used for detecting outliers and for finding groups in data. Distinct groups of objects are known as *clusters*, and the aim of cluster analysis is to discover them. Since there is no unique definition of a cluster. in fact there are several kinds, a diversity of algorithms are available for performing cluster analysis and displaying the results graphically. In archaeology, according to Baxter (1994, 140), "the most common use of cluster analysis is to classify a set of 'individuals' (e.g. artefacts, assemblages, graves. etc.) into subgroups such that individuals within a group are similar to each other in some sense and different from individuals in other groups.". One approach of cluster analysis discussed in this thesis is to compute distances between objects belonging to different groups in the data, in order to quantify their degree of dissimilarity (Kaufman and Rousseeuw, 1990). It is necessary to compute a distance for each pair of

objects i and j. The measure of distance most commonly used is the *Euclidean* distance, (Everitt, 1993), defined as follows :

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \qquad \textbf{(3.6.1)}$$

Where x_{ik} and x_{ij} are the variable values for individuals i and j and p is the number of variables. In a cluster analysis the variables are usually standardised, i.e. Z_{ij} with $\alpha = 1$, see (3.3.1). Other approaches are available, for further discussion see Everitt (1993).

Four procedures which are all examples of hierarchical agglomerative clustering techniques are outlined as follows :

(a) *Single linkage* (or the nearest-neighbour method). The distance between the groups is defined as the distance between the closest pair of individuals, where only pairs consisting of one individual from each group are considered. Single linkage tends to identify outliers within a data set rather than clusters.

(b) *Complete linkage* (or furthest-neighbour). This is the opposite of single linkage in that the distance between groups is defined as the most distant pair of individuals, one from each group. It tends to produce small compact clusters.

(c) *Average linkage*. Here the distance between two clusters is defined as the average of the distances between all pairs of individuals that are made up of one individual from each group.

(d) *Ward's method.* This is similar in origin to average linkage and amalgamates clusters on the basis of similarity between groups rather than just between a pair of individuals.

Output for all the above-mentioned techniques is in the form of a dendrogram, which shows graphically how objects link up and at what level of similarity.

Single linkage is used infrequently, compared to average linkage or Ward's method, in that totally separate clusters can be amalgamated because of a close similarity between just two individuals, one from each cluster. Complete linkage, on the other hand, may identify too many clusters since it tends to divide the data up into many small clusters. For the purpose

of this thesis, cluster analysis is used primarily for outlier detection, i.e. detecting a cluster consisting of a single value or just 2 or 3 values, but is also used to define separate clusters in the data in Chapters 5 and 6. Rather than relying on a single cluster analysis method it is sensible to examine competing methods. If these produce similar results this will verify the reality of the outliers. Single, complete and average linkage are further discussed with illustrative examples in Chapter 5.

3.7 Kernel density estimation

The histogram is the oldest and most widely used density estimator. A density estimator is a technique whereby, rather than assuming a distributional form, such as the normal, for the data, the most appropriate density is empirically estimated from the sample values. For the presentation and exploration of data, histograms are an extremely useful class of density estimates, particularly in the univariate case. However, even in one dimension, the choice of origin can have quite an effect, where 'origin' is defined as the point at which the histogram is started. The appearance also depends on the width of the intervals. Although the histogram is a useful tool for data presentation, many other alternative density estimates are available. Kernel density estimates (KDEs) for univariate data can be thought of as smoothed histograms that are not dependent on a choice of origin. Silverman (1986) discusses kernel density estimation in detail. Like the histogram, the kernel estimator can be used to investigate univariate data, but unlike the histogram, many of the important applications of density estimation are to bivariate data. The bivariate histogram can be used but interpretation of which is extremely difficult. Density estimation is the construction of an estimate of the density function from observed data and can, for example, give an indication of skewness and multimodality in the data. In some cases obvious conclusions will be reached, while in others they will point the way to further analysis and/or data collection. Although the histogram remains a valuable tool for univariate data examination and presentation, problems may arise when histograms are constructed using the same data and the same bin widths but different origins.

3.8 The univariate kernel density estimator

For univariate data the kernel density estimator is simply a sum of 'bumps' placed at the observations, X_1, X_2, ... X_n. The shape of the bump is defined by a mathematical function, the kernel *K(x)*, and the width of the bump is determined by a window-width or smoothing parameter, *h*, which is analogous to the interval width of the histogram. The amount of detail, spurious or otherwise, in the density estimation is determined by the smoothing parameter, *h*. The shape of the resulting KDE does not depend on the choice of origin and the choice of the kernel function is usually unimportant compared to the choice of *h*. The most common kernel is the normal probability density function. General formulation for a density estimate is as follows.

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h})$$

(3.8.1)

Where $\hat{f}(x)$ is an estimate of the density the data is assumed to follow. *K* is the kernel function, *h* is the window-width and n is the size of the sample. As previously mentioned, the amount of detail in the density estimation is determined by *h*, where large values of *h* can over-smooth the data, while small values can under-smooth. The value of *h* can be selected using a variety of differing approaches. Many objective or data-driven choices for *h* are described in detail in Wand and Jones (1995), but for the purpose of this thesis just two of the methods are outlined here, and later used in Chapters 5 and 6.

If the data is thought as a sample of n, taken from an underlying and unknown true density, *f(x)*, then it is possible to define a measure of closeness between the KDE and the true density, where the chosen estimate of *h* 'maximises' this closeness. If the true density of a sample is normal, then the choice for *h* is

$$h = 1.06\hat{\sigma}n^{-\frac{1}{5}}$$

(3.8.2)

Silverman (1986), where $\hat{\sigma}$ is an estimate of σ, the standard deviation of the normal distribution. This is the normal scale rule and will typically over-smooth the data if the underlying density is not normal.

The closeness of a KDE to the true density can be defined in terms of the asymptotic mean integrated square error, AMISE, and the value of h, which minimises this has the form

$$h = \left[\alpha(K)\beta(f'')n\right]^{-\frac{1}{5}} \tag{3.8.3}$$

Where $\alpha(K)$ is a function of the known kernel and

$$\beta(f'') = \int_{x\in\Re} f''^2 dx \tag{3.8.4}$$

is a function of the unknown true density that can be interpreted as its roughness.

Another approach known as the 'solve the equation' (STE) method, is further discussed by Beardah and Baxter (1995) and Wand and Jones (1995). In this case, an equation that relates h to a function of the unknown density is defined. An initial estimate of h leads to an estimate of the density, that in turn leads to a new value for h and a new density estimate. The process continues until the estimate of h converges. Initially the starting point is, in effect, the formula for h (AMISE), seen in (3.8.3).

$$h = [\alpha(K)\beta(f'')n]^{-\frac{1}{5}} \tag{3.8.5}$$

which is a product of two terms, the first being a function of the kernel and the second depending upon the roughness of the second derivative of f. An initial estimate of h is used to form a KDE, $\hat{f}_0$, from which a new value of h_1 can be calculated via

$$h_1 = [\alpha(K)\beta(\hat{f}_0)n]^{-\frac{1}{5}} \tag{3.8.6}$$

(3.8.6) can then be extended to an iteration where

$$h_{i+1} = [\alpha(K)\beta(\hat{f}_i)n]^{-\frac{1}{5}} \tag{3.8.7}$$

and for i = 0, 1, ... each KDE, $\hat{f}_i$, is calculated using window-width h_i.

The STE approach is a general one that can be implemented in several ways. The routines of Beardah, used in this thesis, use a method described by Sheather and Jones in Wand and

Jones (1995). Where to find descriptions of these routines is given in the Appendix. Since the STE approach does not depend on the unknown underlying density being normal, the data tend not to be over-smoothed.

3.9 The adaptive kernel estimator

The basic idea of an adaptive kernel estimate is to construct a kernel estimate consisting of 'bumps' or kernels placed at the observed points, but to allow the window widths of the kernels to vary from one point to another. This is the most effective way to deal with long-tailed densities - to use a broader kernel in regions of low density. The adaptive kernel estimate, $\hat{f}(x)$, is defined by

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} h^{-d}\lambda_i^{-d} K\left\{\frac{(x - X_i)}{h\lambda_i}\right\}, \qquad \textbf{(3.9.1)}$$

where λ_i is a local bandwidth factor and d is the number of dimensions.

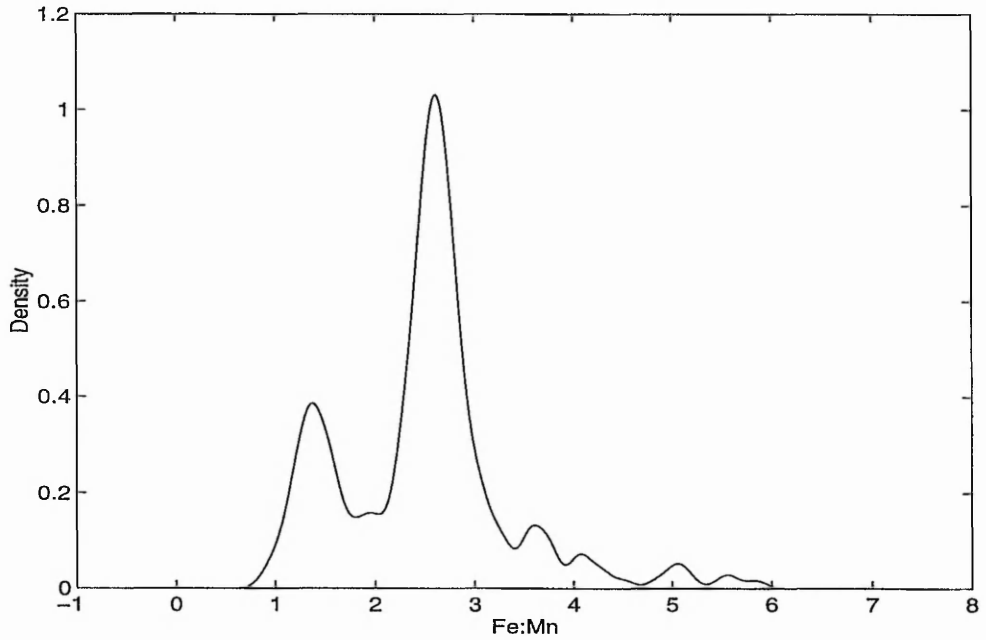
$$\lambda_i = \left\{\frac{\tilde{f}(X_i)}{g}\right\}^{-\alpha} \qquad \textbf{(3.9.2)}$$

g is the geometric mean of a pilot estimate $\tilde{f}(X_i)$ and α is a sensitivity parameter, a number satisfying $0 \leq \alpha \leq 1$. In general, the value of h is determined from the original pilot estimate, $\tilde{f}(X_i)$, and depends on the method used to first obtain this. A natural pilot estimate would be a fixed kernel estimate with bandwidth chosen by reference to a standard distribution. The local bandwidth factors then depend on a power of the pilot density, the larger the power, α, the more sensitive the method will be to variations in the pilot density, and the more difference there will be between bandwidths used in different parts of the sample. The value of $\alpha = 0$ will reduce the method back to a fixed width kernel approach. The adaptive kernel method and its usage is discussed further in Silverman (1986, p. 100-110). In the approach we use in this thesis, h is the value from the pilot estimate and $\alpha = 1/d$, where d = dimensionality.

3.10 Example of univariate kernel density estimation

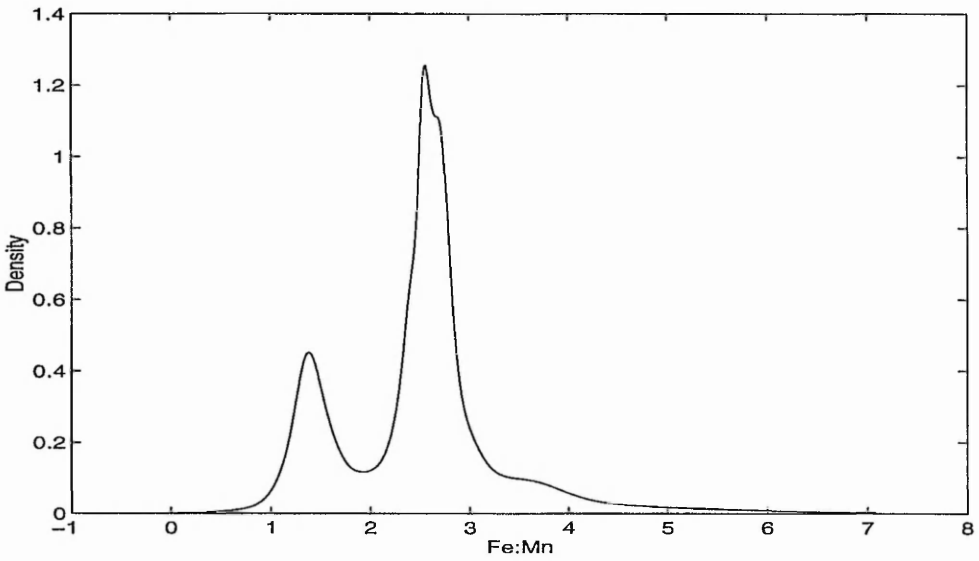
In order to illustrate univariate KDE we present the Fe:Mn ratio of the Southampton data, after the removal of outliers. We use the Fe:Mn ratio since it is used in Chapter 5 to illustrate how KDEs can be useful for identifying modes in the data that have an archaeological interpretation. An initial KDE indicated some observations to be extreme outliers, having very high Fe:Mn ratios, thus obscuring features in the data. These observations were removed and the data re-analysed using the STE method in order to smooth. The STE method is used since the underlying density is not normal and the normal method may over-smooth the data.

Figure 3.10.1 Univariate KDE for the Fe:Mn ratio of the Southampton glass data, using the STE method for the selection of h - after the removal of those observations with high Fe:Mn ratios.



The STE method used in Figure 3.10.1 does well at picking up isolated observations in the tail of the distribution and the larger class widths will occur in the less dense parts of the distribution, for example the anti-mode at 2 is in a less dense area which could have a larger h). The univariate KDE plot of Figure 3.10.2 uses the adaptive STE method to calculate the varying values of h. This adaptive estimate identifies two modes, thus emphasising 2 main groupings in the data.

Figure 3.10.2 Univariate KDE for the Fe:Mn ratio of the Southampton glass data, using the adaptive STE method for the selection of *h* - after the removal of those observations with high Fe:Mn ratios.



3.11 The bivariate kernel density estimator

For bivariate data the bivariate KDE is used. In this case the n points in a plane are defined by the co-ordinates $\underline{X}_i = (X_i, Y_i)$, for i = 1, 2, ..., n. Locating a 'bump' at each point corresponds to centering a three-dimensional bump at each point, and then, at each point in the plane, summing the height of the bumps. Looking at the bivariate normal distribution, if a single smoothing parameter, *h*, is used, then the version of the kernel placed on each data point will be scaled equally in all directions. Formulation for the bivariate KDE is given by

$$\hat{f}(\underline{x}) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left\{\frac{\underline{x} - \underline{X}_i}{h}\right\} \qquad \textbf{(3.11.1)}$$

defined for 2-dimensional $\underline{x}$. The choice of window-widths for the bivariate case is not as well developed as for the univariate case. The formulation used for smoothing in both the x and the y direction is as follows

$$\hat{f}(x, y) = \frac{1}{nh_1h_2}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_1}, \frac{y - Y_i}{h_2}\right) \qquad \textbf{(3.11.2)}$$

where h_1 and h_2 are the window-widths in the x and y directions. Where to find a description of the methods used for calculating h_1 and h_2 is included in the Appendix.

3.12 Example of bivariate kernel density estimation

The methods mentioned above are best illustrated using a range of examples. The Southampton glass data, Chapter 5, are used for illustration.

Figure 3.12.1 A KDE estimate, using the normal scale rule for the selection of h_1 and h_2, for the Southampton data - observations coloured light blue and light green only. Where h = 0.726, 0.2982 refers to h_1 = 0.726, the amount of smoothing in the x-direction and h_2 = 0.2982, the amount of smoothing in the y-direction.

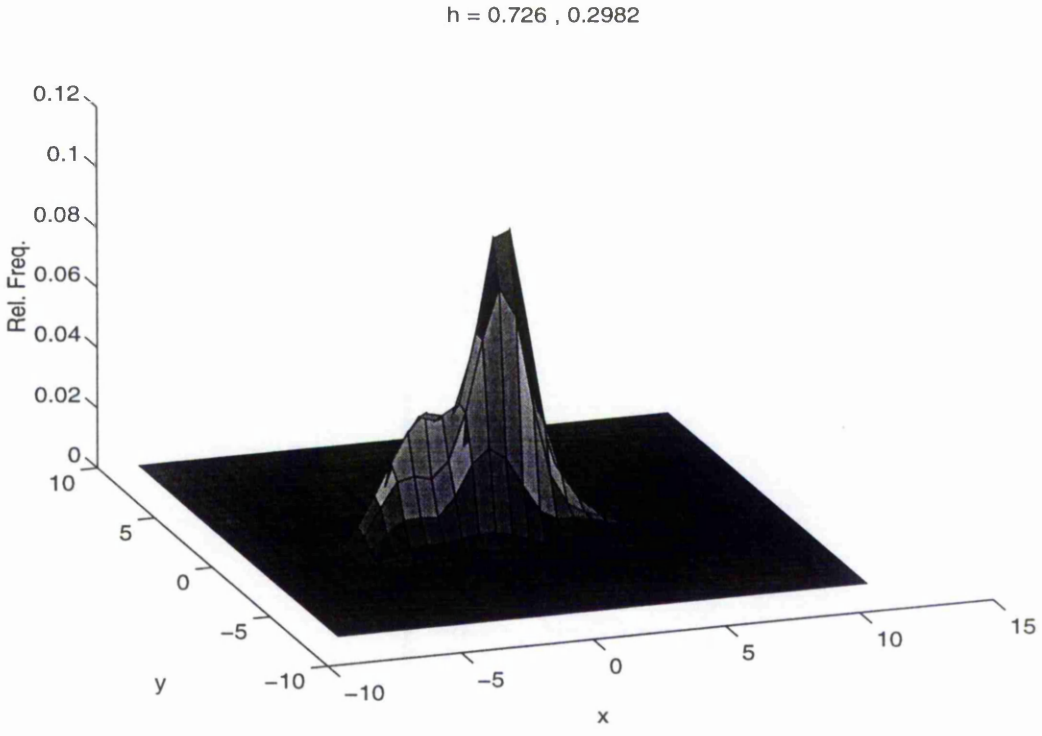


Figure 3.12.2 A KDE estimate, based on the STE rule for the selection of h_1 and h_2, for the Southampton data - observations coloured light blue and light green only. Where $h_1 = 0.3522$ and $h_2 = 0.2616$.

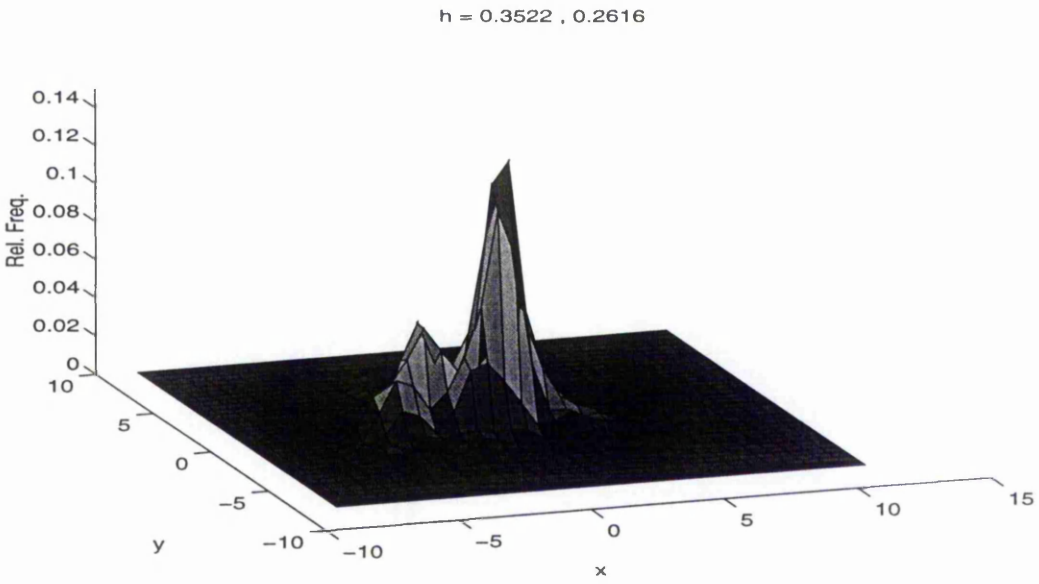


Figure 3.12.1 and Figure 3.12.2 are both KDE's based on the first two principal components of the Southampton glass data for those observations coloured light blue/light green only. Since KDE's can be applied to composite variables, such as those derived in a principal components analysis, this allows us to look at the data in 2 dimensions. This form of data presentation is used in Chapters 5 and 6 when analysing archaeological data, alongside univariate methods, principal components analysis and cluster analysis. Figure 3.12.1 is a KDE estimate based on the normal scale rule for the selection of h_1 and h_2. The normal scale estimate oversmooths the data, (where $h_1 = 0.7260$ and $h_2 = 0.2982$), in the x-direction, and misses the smaller mode, to the left of the plot, suggested by the STE approach of Figure 3.12.2, (where $h_1 = 0.3522$ and $h_2 = 0.2616$). The modes are associated with the different colours, the smaller mode to the left of the plot indicates those observations coloured light green and the larger mode to the right indicates light blue specimens, i.e. the modes represent a real phenomenon and are not an artefact of the methodology.

KDE's can also be used as a basis for producing contour plots of the data. Baxter, Beardah and Wright (1995) discuss the use of contouring with archaeological data. After a bivariate

KDE has been obtained each two dimensional data point is associated with a density height that can be ranked from largest to smallest. For example, the first 50% of the ranked observations may be used to define contours that enclose the densest 50% of the data. The level of contouring can be varied to contain any specified proportion of the data. The following plots show how a particular contour level, in this case 50%, may be selected and drawn to reveal groupings, if any, in the data.

Figure 3.12.3 A KDE of the all the Southampton glass data, excluding outliers, using the normal scale rule. The contour is for the 50% inclusion level

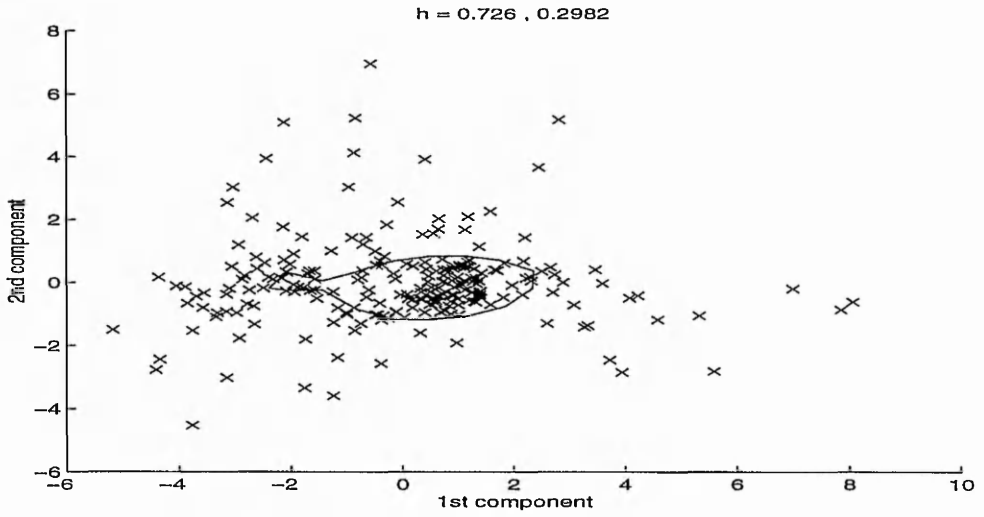
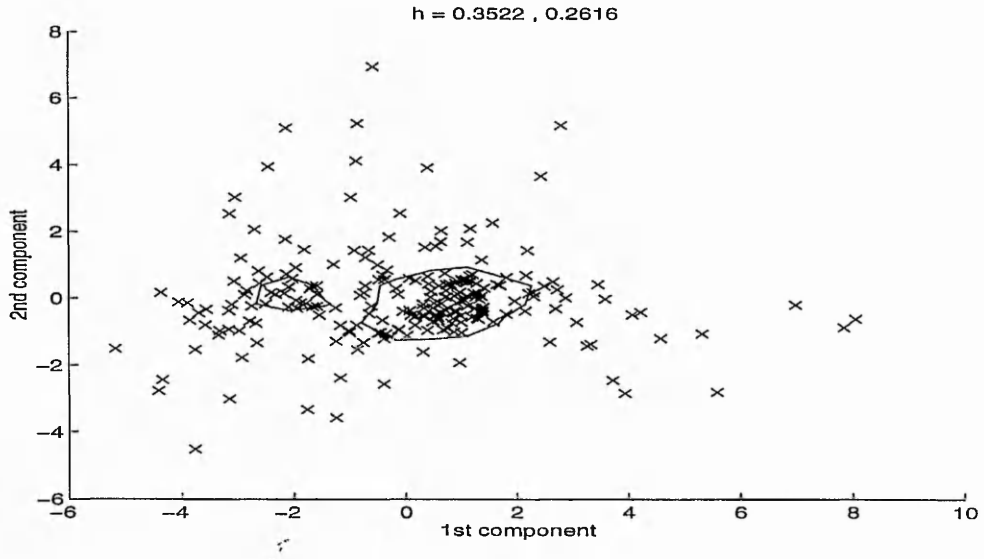


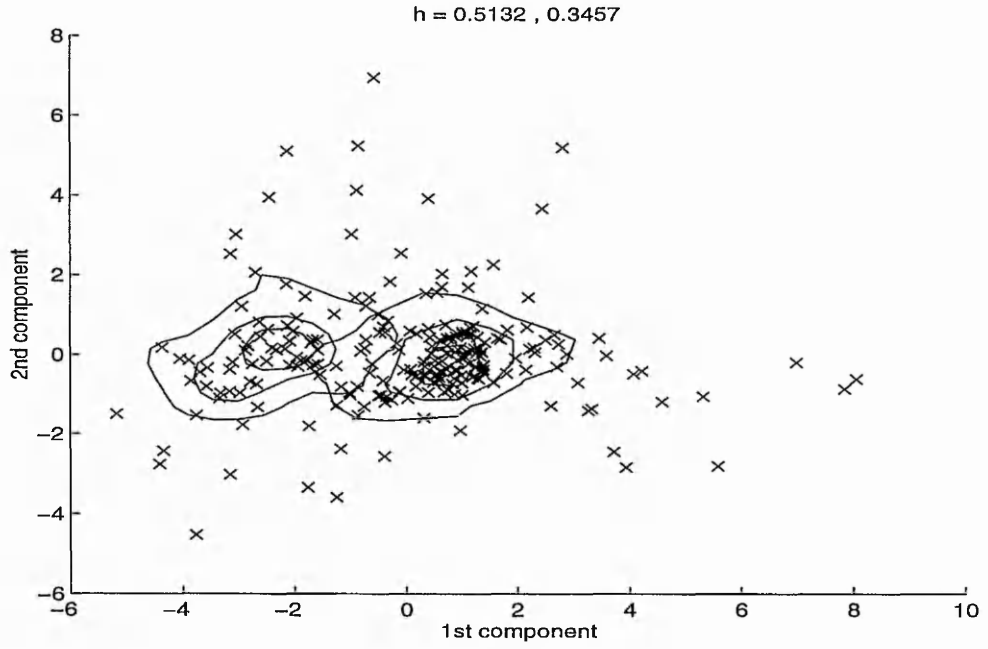
Figure 3.12.4 A KDE of the all Southampton glass data, excluding outliers, using the STE rule. The contour is for the 50% inclusion level.



The two main groupings identified in Figure 3.12.2 have also been identified by the contour plot of Figure 3.12.4, where selection of h_1 and h_2 is based on the STE method. As

previously seen, in Figure 3.12.1, the normal scale estimate contour plot of Figure 3.12.3 appears to oversmooth the data in the x-direction as it does not suggest the two groupings as readily. Contouring can be used in many forms and one such usage is separate contouring for groups present in the data, Bowman and Foster (1993). When constructing a separate contour plot the data belonging to each group are separated into individual data sets. Each group is then contoured separately and plotted on the same contour plot. Figure 3.12.5 shows two groups via separate contouring corresponding to those specimens coloured light blue and those coloured light green, (Southampton data). The selection of h_1 and h_2 is based on the STE method. The separate contour plot defines the contours which enclose the densest n% of the data. In Figure 3.12.5 contours correspond to the 25, 50 and 75% respectively for each group. Separate contouring is further used in Chapters 5 and 6 as a visual technique for displaying groups in the data.

Figure 3.12.5 Separate contour plot using the STE method for selection of h_1 and h_2. Observations coloured light blue are encapsulated in the contour to the right of the plot and observations coloured light green in the contour to the left of the plot



Bivariate kernel density estimation, in this thesis, is used primarily to enhance the interpretation of principal component analyses. Since it can prove very useful for showing

concentrations of observations or modes in the data, this approach to data presentation and interpretation is used effectively in Chapters 5 and 6.

4. Multivariate outlier detection

One main purpose of initial data analysis is to screen the data for possible unusual values. If such values are present they should be investigated before any detailed analysis of the data is undertaken. If anomalous values are genuine, i.e. not recording errors, then thought must be given as to whether to retain these values for further analysis. One definition of an outlier is an 'extreme' observation that lies 'far away' from the rest of the data values, i.e. a value which is atypical of those in the rest of the dataset. With multivariate data, atypicality can arise in a number of different ways, and different aspects of atypicality will in general require different techniques for their detection.

Classical outlier detection methods, using the Mahalanobis distance (d_j^2), defined as a measure of distance that takes into account correlation in the data, are powerful when the data contain only one outlier.

$$d_j^{\,2} = (\mathbf{x}_j - \bar{\mathbf{x}})' S^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}) \quad (j = 1, \ldots, n) \tag{4.1}$$

Note that S is the covariance matrix (α = 0 in 3.3.1). However, the usefulness of these methods decreases drastically if more than one outlying observation is present in the data, this is usually due to what are known as masking and swamping problems. For example, a large value of MD may indicate that the corresponding observation is an outlier, but two problems arise in practice. Firstly, outliers do not necessarily have large values for MD. For example a small cluster of outliers will attract the mean estimate. $\bar{\mathbf{x}}$ (4.1.1), and will inflate the covariance matrix standard deviation estimate, S (4.1.2). in its direction, yielding small values for MD. This problem is known as the masking problem because the presence of one outlier masks the appearance of another outlier. Secondly, not all observations with large MD values are outliers. For example, a small cluster of outliers will attract $\bar{\mathbf{x}}$ and inflate S in its direction and away from some other observations which belong to the pattern suggested by the majority of observations, thus yielding large MD values for these observations. This problem is known as the swamping problem. (Krzanowski and Marriott, 1994)

4.1 Notation and Theory

Let the n × p data matrix be X, with typical element x_{ij} and of which $\mathbf{x}_i$ is the i'th row. Barnett (1976) discusses the ordering of multivariate data and identifies four different types: (a) *Marginal ordering* (M-ordering). This is basically an inspection of the marginal samples using dotplots, box and whisker plots, etc. This type of sub-ordering may serve as an introduction to some further sub-ordering process.

(b) *Partial ordering* (P-ordering). This type of sub-ordering lies in examining the numbers of sample points which lie in different regions of the sample space after it has been partitioned in some manner.

(c) *Conditional ordering* (C-ordering). This sub-ordering principale for multivariate data is one in which ordering is conducted on one of the marginal sets of observations *conditional* on selection. The process is repeated sequentially through all the marginal sets of observations to produce statistically equivalent blocks.

(d) *Reduced ordering* (R-ordering). In this case each multivariate observation is reduced to a single value by means of some combination of the component sample values. These single values can then be ordered univariately.

Of these all these ordering procedures R-ordering is the one most suitable for definition of 'extremes',

Siotani (1959) bases his definition of extremeness of a multivariate observation x, on its 'distance value'. This distance value is the 'Mahalanobis distance'.

If $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is a random sample of multivariate observations from a population with unknown location and dispersion parameters μ and Σ, then suitable quantities to use as estimates of them would be

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \qquad (4.1.1)$$

and S, (3.3.2).

4.2 Basic statistics for outlier detection

Krzanowski and Marriott (1994) list the following statistics taken from Gnanadesikan and Kettenring (1972), where $\bar{x}$ and S are defined as in (4.1.1) and (3.3.2).

$$q^2_j = (\mathbf{x}_j - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}}) \quad (j = 1, \ldots, n) \tag{4.2.1}$$

$$t^2_j = (\mathbf{x}_j - \bar{\mathbf{x}})' S (\mathbf{x}_j - \bar{\mathbf{x}}) \quad (j = 1, \ldots, n) \tag{4.2.2}$$

$$u^2_j = \frac{(\mathbf{x}_j - \bar{\mathbf{x}})' S (\mathbf{x}_j - \bar{\mathbf{x}})}{(\mathbf{x}_j - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}})} \quad (j = 1, \ldots, n) \tag{4.2.3}$$

$$v^2_j = \frac{(\mathbf{x}_j - \bar{\mathbf{x}})' S^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})}{(\mathbf{x}_j - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}})} \quad (j = 1, \ldots, n) \tag{4.2.4}$$

$$d^2_j = (\mathbf{x}_j - \bar{\mathbf{x}})' S^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}) \quad (j = 1, \ldots, n) \tag{4.2.5}$$

$$d^2_{jk} = (\mathbf{x}_j - \mathbf{x}_k)' S^{-1} (\mathbf{x}_j - \mathbf{x}_k) \quad (j < k = 1, \ldots, n) \tag{4.2.6}$$

Each of these statistics identifies the contribution of the individual observations to specific effects as follows, Krżanowski and Marriott (1994, 51)

- q^2_j isolates observations which excessively inflate the overall scale
- t^2_j determines which observations have the greatest influence on the orientation and scale of the first few principal components of S
- u^2_j is similar to t^2_j but puts more emphasis on orientation and less on scale
- v^2_j measures the relative contributions of the observations on the orientations of the last few principal components
- d^2_j uncovers those observations which lie far away from the general scatter of points
- d^2_{jk} has the same objectives as d^2_j but provides far more detail of inter-object separation.

The statistic most generally used is d^2_j, which is Mahalanobis distance. Large values of d_j are intended to identify points remote from the bulk of the data. Mahalanobis distance is used in Wilk's (1963) test statistic, which also formed the basis of a graphical method described by Bacon-Shone and Fung (1987). A more recent method is described by Caroni

and Prescott (1992), this being the sequential application of Wilk's multivariate outlier test statistic. The procedure of Caroni and Prescott is outlined below.

4.3 Wilk's multivariate outlier test statistic

As above, we assume that $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ is a multivariate sample of size n, where the mean μ and covariance matrix Σ, are unknown. So the standard estimates of (4.1.1) and (3.3.2) are assumed.

Wilk's test statistic is given by $W_j = \dfrac{|A^{(j)}|}{|A|}$ where A is as defined as in (3.3.2) and $A^{(j)}$ is the corresponding matrix with $\mathbf{x}_j$ eliminated from the sample.

The potential outlier is that point, with index l, whose removal leads to the greatest reduction in $|A|$, i.e. the point for which this ratio is minimised. Wilk's statistic is then defined as

$$D_1 = \min_j(W_j) = \frac{|A^{(l)}|}{|A|} \qquad \textbf{(4.3.1)}$$

D_1 can also be written as

$$D_1 = 1 - \frac{n}{n-1}(\mathbf{x}_l - \bar{\mathbf{x}})' A^{-1}(\mathbf{x}_l - \bar{\mathbf{x}}) \qquad \textbf{(4.3.2)}$$

for ease of computation (Caroni and Prescott, 1992).

Once the most extreme observation x_l, has been identified, it is removed from the analysis and Wilk's procedure is applied to the reduced sample of n - 1 multivariate observations. D_2 may be defined as

$$D_2 = 1 - \frac{n-1}{n-2}(\mathbf{x}_m - \bar{\mathbf{x}}^{(l)})'(A^{(l)})^{-1}(\mathbf{x}_m - \bar{\mathbf{x}}^{(l)}) \qquad \textbf{(4.3.3)}$$

Where m is the index of the second most extreme outlier and $\bar{\mathbf{x}}^{(l)}$ is the vector of the sample means with $\mathbf{x}_{[l]}$ eliminated. This procedure is then repeated to identify a series of potential outliers x_l, x_m, ..., corresponding to a series of Wilk's statistics $D_1, D_2, \ldots$. In order to decide which observations are actual outliers we compare the $D_1, D_2, \ldots D_k$ against appropriate critical values $\lambda_1, \lambda_2, \ldots, \lambda_k$ in turn. The number of outliers declared by the

sequential procedure is the lowest value r for which $D_r > \lambda_r$ is true. To find the critical values is just simply a case of looking at tables given by Wilks (1963, 425).

4.4 Rousseeuw and Van Zomeren's algorithm for outlier detection

Rousseeuw and van Zomeren (1990) propose to avoid the masking effect by computing distances based on very robust estimates of location and covariance which themselves are based on the minimum volume ellipsoid covering half the data. Rousseeuw and van Zomeren (1990) use the minimum volume ellipsoid estimator (MVE). Robust distances are computed for each data point x_i, using the following

$$RD_i = \sqrt{(\mathbf{x}_i - T(X))' C(X)^{-1} (\mathbf{x}_i - T(X))} \tag{4.4.1}$$

where T(X) and C(X) are robust location and scale estimators. T(X) is the centre of the minimum volume ellipsoid, C(X) is determined by the same ellipsoid and w_j is the weight function.

$$T(X) = \frac{\sum_{i=1}^{n} w_i \mathbf{x}_i}{\sum_{i=1}^{n} w_i} \tag{4.4.2}$$

$$C(X) = \frac{\sum_{i=1}^{n} (\mathbf{x}_i - T(X))(\mathbf{x}_i - T(X))'}{\sum_{i=1}^{n} w_i - 1} \tag{4.4.3}$$

Having calculated the robust distances we now need to identify those observations thought to be outlying. Here a cut-off point is used which is the maximum expected value from a sample of n chi-squared random variables on p degrees of freedom, approximated by :

$$E(\max \chi_p^2) = \chi_p^2 \{(\frac{n - 0.5}{n})\} \tag{4.4.4}$$

If $RD_i > E(\max \chi^2_p)$ then these observations can be thought of as possible outliers.

4.5 Hadi's algorithm for outlier detection

Hadi (1992; 1994) suggests a method for identifying outliers in multivariate samples which is less susceptible to masking and swamping problems. This is to use robust estimators of the location and dispersion parameters ($\bar{x}$ and S) which are unaffected by outliers. Hadi (1992; 1994) proposes a procedure for the detection of multiple outliers in multivariate data.

Let X be an n × p data matrix representing a random sample of size n from a p-dimensional population. Initially, the n observations are rearranged in ascending order according to the chosen robust distance

$$D_i(C_M, S_M) = \sqrt{\{(\mathbf{x}_i - C_M)' S_M^{-1} (\mathbf{x}_i - C_M)\}} \tag{4.5.1}$$

where C_M is a vector containing the co-ordinate medians and S_M is defined to be :

$$S_M = \frac{\sum_{i=1}^{n} (\mathbf{x}_i - C_M)(\mathbf{x}_i - C_M)'}{n - 1} \tag{4.5.2}$$

A weight function is then defined as

$$v_i = \begin{cases} 1 & \text{if } i \le \text{integer part of } (n+p+1)/2 \\ 0 & \text{otherwise} \end{cases}$$

and finally

$$D_i(C_R, S_R) = \sqrt{\{(\mathbf{x}_i - C_R)' S_R^{-1} (\mathbf{x}_i - C_R)\}} \tag{4.5.3}$$

is calculated where C_R and S_R are robust location and covariance matrix estimators, defined by:

$$C_R = \frac{\sum_{i-1}^{n} v_i \mathbf{x}_i}{\sum_{i=1}^{n} v_i} \tag{4.5.4}$$

$$S_R = \frac{\sum_{i=1}^{n} v_i (\mathbf{x}_i - C_R)(\mathbf{x}_i - C_R)'}{\sum_{i=1}^{n} v_i - 1}$$

(4.5.5)

Note that (4.5.3) to (4.5.5) are the same as (4.4.1) to (4.4.3) except for the choice of the weight function, hence Hadi's algorithm and Rouseeuw and Van Zomeren's are very similar in approach.

Now the observations are rearranged in ascending order according to $D_i(C_R,S_R)$. The rearranged observations are then divided into two subsets : one subset containing the first p+1 observations and the other containing the last n-p-1 observations. The robust distances $D_i(C_b,S_b)$ are then calculated as in (4.5.1), by setting $C_M=C_b$ and $S_M=S_b$, where C_b and S_b are the mean and covariance matrix of the basic subset. The observations are rearranged in ascending order according to $D_i(C_b,S_b)$. Let r be the number of observations in the current basic subset. The observations are then divided into two subsets - a basic subset containing the first (r+1) observations and another subset containing the remaining (n-r-1) observations.

An explanation of the procedure is as follows. Initially the n observations are ordered using $D_i(C_M,S_M)$ which is based on robust estimators of location and dispersion. Now outliers are more likely to appear in the second subset containing (n-r-1) observations and the initial basic subset, containing (r+1) observations, is highly unlikely to contain outliers. The procedure is repeated as defined above and stops when a stopping criterion is met, most commonly when the basic subset contains h observations, where h = (n+p+1)/2. Once the stopping criterion has been met the location and covariance matrix estimator based on the observations included in the final subset are used to compute the robust distances :

$$D_i(C_b, S_b) = \sqrt{\{(\mathbf{x}_i - C_b)'(c_b S_b)^{-1}(\mathbf{x}_i - C_b)\}}$$

(4.5.6)

where c_b is a correction factor (Hadi 1994), which is used to obtain consistency when the data come from a multivariate normal distribution.

$$c_b = \left(1 + \frac{2}{n-1-3p} + \frac{p+1}{n-p}\right)^2$$

(4.5.7)

The observations with large values of $D_i(C_b,S_b)$ above are then declared as outliers, i.e. if $D_i(C_b,S_b) > \chi^2_{p,\alpha/n}$. Again the cut-off point is based on the chi-squared distribution.

4.6 The Atkinson and Mulira forward algorithm

Atkinson and Mulira (1993) propose a method which is aimed at the detection of multivariate outliers using the Mahalanobis distance. rather than directly at the robust estimation of S. The Atkinson and Mulira method uses the standard estimates given by (4.1.1) and (3.3.1), but from a subset of m observations chosen to be unlikely to contain outliers.

- First randomly define a starting position, z, by selecting m < n observations, (normally m = p − 1, the smallest number from which the distances can be calculated), and calculate $\bar{x}$ and the covariance matrix S. It is worth noting that m = p + 1 doesn't always work, due to singularity problems, so it may be necessary to find a set of (p + 1) observations that does work as an initial starting position or use more than (p + 1) observations.
- Using these estimates, n Mahalanobis distances can be calculated. If the m observations are outlier free, any outliers will give rise to large Mahalanobis distances.
- These values of MD are sorted in ascending order. the sample size is incremented by some small integer *s* and the m + *s* observations with the smallest distances are used to calculate new estimates of the mean and covariance matrix.
- The above 2 steps are repeated until m = n.
- Outliers will only be included as m approaches n. when no non-outlying observations remain to be introduced into the fit.
- The result of this analysis is an (n - p) × n matrix of Mahalanobis distances. Note that the final result may depend on the initial starting position. so it is sensible to start from several random starting positions in order to check that the same results are obtained.

There are various methods which can be used for displaying the results.

- *The stalactite plot.* This plot shows how the pattern of suspected outliers changes with m. Those observations with large Mahalanobis distances can be classed as outliers. The cut-off point used to define an outlier is the maximum expected value from a sample of n chi-squared random variables on p degrees of freedom, approximated as in (4.4.4).
- *The index plot.* This is an index plot of the Mahalanobis distances. typically when m is 80% or 90% of the sample size n. As in the point above, the same cut-off point can be used to define an outlier, Atkinson and Mulira (1993, 29).
- *The probability plot.* As with the index plot, the probability plot is useful when wanting to look at 80% or 90% of n. This plot is also helpful in interpreting the magnitudes of the distances for suspected outliers.

4.7 Discussion

One major disadvantage of calculating Mahalanobis distances using an entire sample of data is that $\bar{x}$ and S are themselves adversely affected by outliers. This is because $\bar{x}$ and S are calculated using all the observations, i.e. outliers are included in the initial calculations. This therefore leads to the breakdown of the approach for detecting outliers using Mahalanobis distance since d_j is affected by the cases it is designed to detect and hence may fail to detect them. Also it is worth noting that the statistics which act on the first few principal components, namely d^2_j, t^2_j and u^2_j, tend to detect those outliers which inflate variances, covariances or correlations in the data. Therefore it may be concluded that the outliers identified by Mahalanobis distance could be detected by simple uni and bi-variate analyses. For example, by inspection of dotplots, box and whisker plots and plots of the first two principal components.

Wilk's statistic is based on the change of the determinant of the sample scatter matrix after some observations are eliminated from the sample. But as with d^2_j, the initial calculations of $\bar{x}$ and S are based on all the observations in the dataset (including possible outliers). Therefore, in the same way, Wilk's statistic can be affected by the cases it is designed to detect. One major defect of Wilk's statistic is that since the number of outliers present is uncertain, it does not necessarily lead to an outlier-free dataset.

The method of Rousseeuw and van Zomeren uses the minimum volume ellipsoid to provide robust estimates of the mean and covariance matrix of the data. In this approach, $\bar{x}$ (T(X), (4.4.2)) is approximated by the centre of the minimum volume ellipsoid (MVE) covering half of the observations and S (C(X), (4.4.3)) is estimated from this same ellipsoid. Rousseeuw and van Zomeren give a warning against using the method unless there are at least five observations per dimension, in order to avoid 'the curse of dimensionality' which may lead to an unrepresentative MVE. The comments of Cook and Hawkins (1990) heavily criticise the use of the MVE approach saying that this method demands excessive computation and it may also produce misleading answers leading to an excess of outliers.

The Hadi (1992) and Atkinson and Mulira (1993) algorithms are based on similar ideas to the one proposed by Rousseeuw and van Zomeren (1990), but they improve on this algorithm in many ways. Hadi (1992) states that although the MVE approach has a breakdown point of 50%, which means that T(X) will remain bounded and the eigenvalues of C(X) will stay away from zero and infinity when less than half the data are replaced by arbitrary values, it is computationally expensive, dependent on resampling, and it may not even be computationally feasible to find the MVE. Another problem arises when the volume of the ellipsoid is 0 (where the rank of the subsample p+1 is less than p) and the distances RD_i in (4.4.1) cannot be computed. Rousseeuw and van Zomeren avoid this problem by simply omitting any subsample with a nearly singular covariance matrix which is, as Hadi (1992) points out, "A method which searches for a MVE and ignores ellipsoids with zero volumes seems to defeat its own purpose.". To overcome these problems, Hadi suggests using the robust distance $D_i(C_b,S_b)$. He states this is effective in identifying multivariate outliers, and in dealing with masking and swamping problems, because it is based on robust estimators of location and the covariance matrix. Hadi points out that an advantage $D_i(C_b,S_b)$ has over RD_i is that T(X) and C(X), the mean vector, (4.4.2), and covariance matrix, (4.4.3), are based on only p+1 observations whereas C_b and S_b are based on h=(n+p+1)/2 observations. Therefore they are more accurate estimators than T(X) and C(X). Hadi uses the same forward algorithm as used by Atkinson and Mulira (1993), but starting from robust estimates of the means and covariances for calculation of the initial Mahalanobis distances. Hadi's forward search terminates when m is the median of the

number of observations (h=(n+p+1)/2) when allowance is made for the effect of fitting. In the procedure proposed by Atkinson and Mulira, the method continues until m=n. Therefore we would expect this to be more reliable than the Hadi method since all the observations are being taken into account. Atkinson and Mulira essentially use the same algorithm as the one described by Hadi but the complex 'start-up' procedures of Hadi are not needed.

In theory, the initial sub-sample of the Atkinson and Mulira algorithm needs to be clear of outliers, so that unbiased estimates are obtained of means and covariances for calculating the distances. Since the initial sub-sample is selected at random, one or more outliers could well be included, perhaps leading to a subsequent failure to identify some of the outliers. But examples contradicting this statement can be seen in Atkinson and Mulira (1993) where, using the Hawkins *et al* (1984) data, n=75 and p=3, in one search where the initial sub-sample contained the 14 outliers in 75 observations they showed that the method correctly identified the outliers even from a starting position consisting of just the 14 outliers. Thus the starting point for the forward algorithm may not be crucial. This is further illustrated in Chapter 5 in an analysis of the Southampton glass data, but, as discussed in the next sub-section, we show that the starting position can indeed be crucial depending on the type of data used. This forward calculation of Mahalanobis distances by the resampling method of Atkinson and Mulira (1993) provides an alternative to the methods of Hadi (1992; 1994) and Rousseeuw and van Zomeren (1990), although as already mentioned above these three methods are all based on similar ideas and algorithms but as Hadi improves on the procedure described by Rousseeuw and van Zomeren in various ways mentioned, so Atkinson and Mulira improve on Hadi's algorithm.

Atkinson (1994) investigates the above-mentioned robust methods for detecting multiple outliers using the robust estimators based on the minimum ellipsoid. He concludes that the forward algorithm rapidly leads to the detection of multiple outliers, and exact calculation of robust parameter estimates does not seem to be necessary for outlier detection. One problem with the forward algorithm is the presence of grouping. The Atkinson and Mulira method assumes that a majority of the data form a reference group against which unusual data may be judged. If distinct groups are present in the data, then there is not a natural reference group to which unusual data can be related. In this case the outcome of the

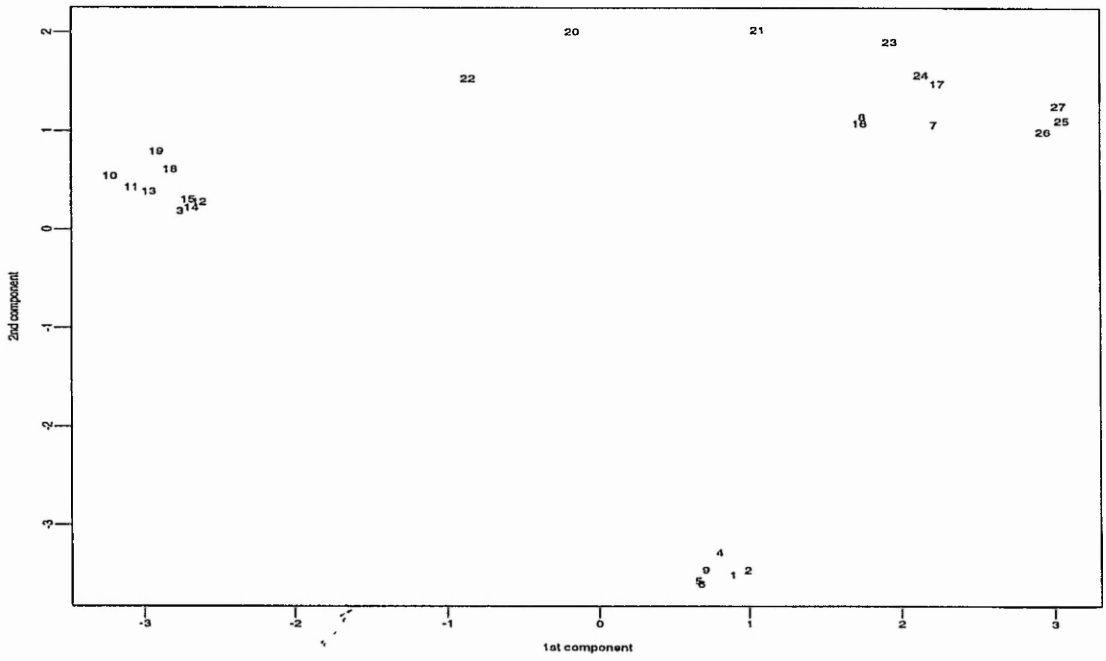
forward algorithm is dependent on the initial choice of cases from which the Mahalanobis distances are calculated. This is also a potential problem with the Rousseeuw and van Zomeren and Hadi algorithms and it is illustrated by example in the next sub-section.

4.8 Discussion of the Atkinson and Mulira method of outlier detection using window glass from York Minster

The only method we will demonstrate in this section is that of Atkinson and Mulira (1993). This is because, although all the above mentioned methods are based on similar ideas and algorithms, the Atkinson and Mulira method is more sophisticated and any problem which this method is unable to deal with would not be overcome by the other methods.

The 27 specimens of window glass are taken from an assemblage of blue glass from windows in York Minster and subsequent excavations. Having looked at the original data using univariate plots for each of the 11 oxides, Al_2O_3, Fe_2O_3, MgO, CaO, Na_2O, K_2O, TiO_2, P_2O_5, MnO, PbO, SbO, it would appear that PbO will have little or no effect on the analysis as there is no variation in its value. Therefore PbO is removed from any further analyses. An initial principal component analysis of standardised data produces the following plot.

Figure 4.8.1 Plot of the first two principal components using standardised data based on the correlation matrix



Looking at Figure 4.8.1 the data appear to fall into three distinct groupings consisting of the following observations, see Table 4.8.1, although observations 20 and 22 appear to be a little more 'detached' from a clear group.

Table 4.8.1 Table showing observations belonging to the three distinct groupings

Group	Observations
1	1 2 4 5 6 9
2	3 10 11 12 13 14 15 18 19
3	7 8 16 17 20 21 22 23 24 25 26 27

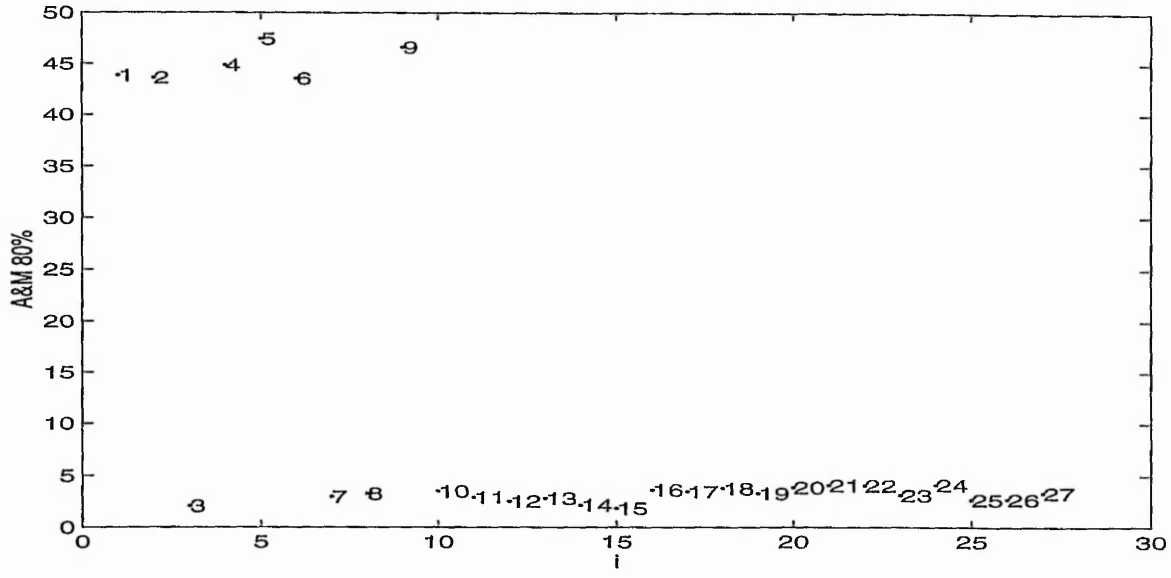
If the Atkinson and Mulira method is used on this data problems may occur due to the grouping in the data. The outcome of the forward algorithm is dependent on the initial starting position, z, from which the Mahalanobis distances are calculated. Atkinson and Mulira suggest using random starts and recommend that several be 'tried out' before making any decisions concerning outliers. To illustrate potential problems with this method however, we will first use non-random starts.

Taking an initial starting position, z, (m=p+1, where p=10 in this case), those observations found in one of the groupings indicated by Figure 4.8.1. If z contains the observations from group 3, (7, 8, 16, 17, 20, 21, 22, 23, 24, 25, 26), then when 80% of the data have been included in the calculation of the Mahalanobis distances, as suggested by Atkinson and Mulira (1993, 29), the index plot of Figure 4.8.2 is produced. When using the Atkinson and Mulira method, a cut-off point can be used in order to help define the outliers, based on χ^2, see (4.4.4).

Table 4.8.2 Table indicating χ^2 values for n = 27 and p = 10

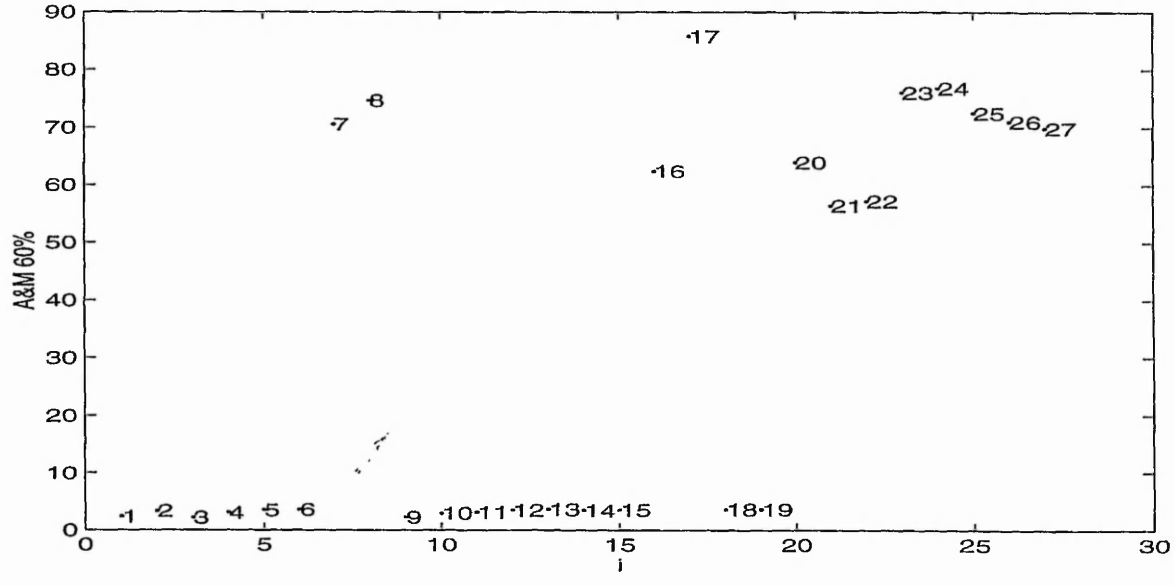
	χ^2
n=100% (27), p=10, 5%	17.97
n=100% (27), p=10, 1%	22.78

Figure 4.8.2 Index plot with 80% of the York Minster data



The index plot of the Mahalanobis distances of Figure 4.8.2 indicates that the observations of group 1 : 1, 2, 4, 5, 6 and 9 are outliers, since all observations have d^2 values well in excess of the critical values 17.97 and 22.78, for the 5% and 1% significance levels respectively. Next the following observations are taken as the starting position, (3, 10, 11, 12, 13, 14, 15, 18, 19, 1, 2). The first nine observations listed are observations which make up group 2 and the last two are found in group 1. When 60% of the data have been included, the following plot is obtained. We use 60% rather than 80% for illustrative reasons.

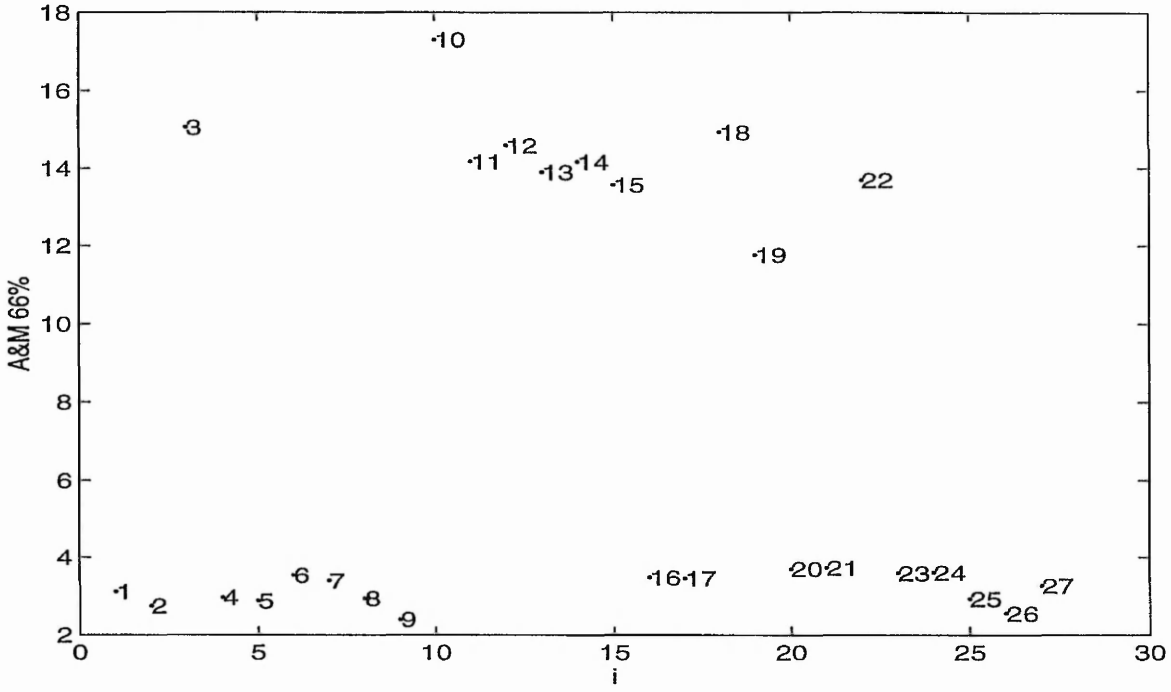
Figure 4.8.3 Index plot with 60% of the York Minster data



This indicates that the observations of group 3 are outliers, since all observations have d^2 values well in excess of the critical values 17.97 and 22.78, for the 5% and 1% significance levels respectively.

Now we use a starting position of (1, 2, 4, 5, 6, 9, 7, 8, 16, 17, 20). The first six observations listed make up group 1 and the latter five are found in group 3. When 66% of the data have been included, the following is produced.

Figure 4.8.4 Index plot with 66% of the York Minster data



The index plot of Figure 4.8.4 identifies the observations of group 2 as different from the rest but only observation 10 has a d^2 value in excess of the critical value 17.97 for the 5% significance level.

4.9 Discussion

Essentially there are three groups in the data and by judicious choice of starting position we are able to highlight the observations in each group as outliers. The values in Table 4.9.1 verify this point and the corresponding value for m, i.e. the percentage of data needed to be

included in the calculation of the Mahalanobis distances, is selected in order to best view the data and the outlying values.

Table 4.9.1 Table indicating initial z, with corresponding outliers

Initial starting position, z	Outliers detected
7, 8, 16, 17, 20, 21, 22, 23, 24, 25, 26	m = 80%, observations 1, 2, 4, 5, 6, 9
3, 10, 11, 12, 13, 14, 15, 18, 19, 1, 2	m = 60%, observations 7, 8, 16, 17, 20, 21, 22, 23, 24, 25, 26, 27
1, 2, 4, 5, 6, 9, 7, 8, 16, 17, 20	m = 66%, observations 3, 10, 11, 12, 13, 14, 15, 18, 19, 22

It is also of interest to carry out runs from random starts to see whether or not consistent results are obtained.

Table 4.9.2 Table to show outliers detected after 10 random starts

Random starting position, z	**Outliers detected**
1, 3, 7, 8, 9, 12, 22, 23, 24, 26, 27	No extreme values detected
1, 2, 3, 4, 5, 10, 13, 14, 18, 23, 27	6, 7, 8, 16, 17, 20, 21, 22, 24 **(predominantly group 3)**
4, 7, 8, 12, 15, 16, 17, 19, 22, 25, 26	6, 18, 20, 21, 23, 24, 27 **(predominantly group 3)**
3, 6, 7, 9, 10, 11, 14, 15, 20, 21, 23	13, 16, 17, 18, 19, 22, 24, 25, 26 **(group 2 and group 3)**
4, 5, 13, 16, 17, 18, 19, 22, 24, 25, 26	6, 7, 8, 20, 21, 23, 27 **(group 3)**
2, 3, 4, 10, 11, 12, 14, 15, 18, 19, 21	7, 8, 16, 17, 20, 21, 22, 23, 24, 25, 26, 27 **(group 3)**
1, 2, 3, 6, 9, 14, 15, 18, 22, 24, 25	4, 5, 8, 10, 17, 19, 20, 21, 23, 26, 27 **(predominantly group 3)**
4, 8, 15, 16, 17, 19, 21, 22, 23, 25, 26	1, 3, 5, 6, 10, 11, 12, 13, 18, 20, 24 **(predominantly group 1 and 2)**
6, 9, 11, 12, 15, 17, 18, 19, 20, 21, 22	3, 7, 10, 13, 23, 24, 25, 26, 27 **(group 2 and 3)**
7, 8, 10, 12, 20, 22, 23, 24, 25, 26, 27	1, 2, 4, 5, 6, 9, 16, 17, 18, 19 **(predominantly group 1)**

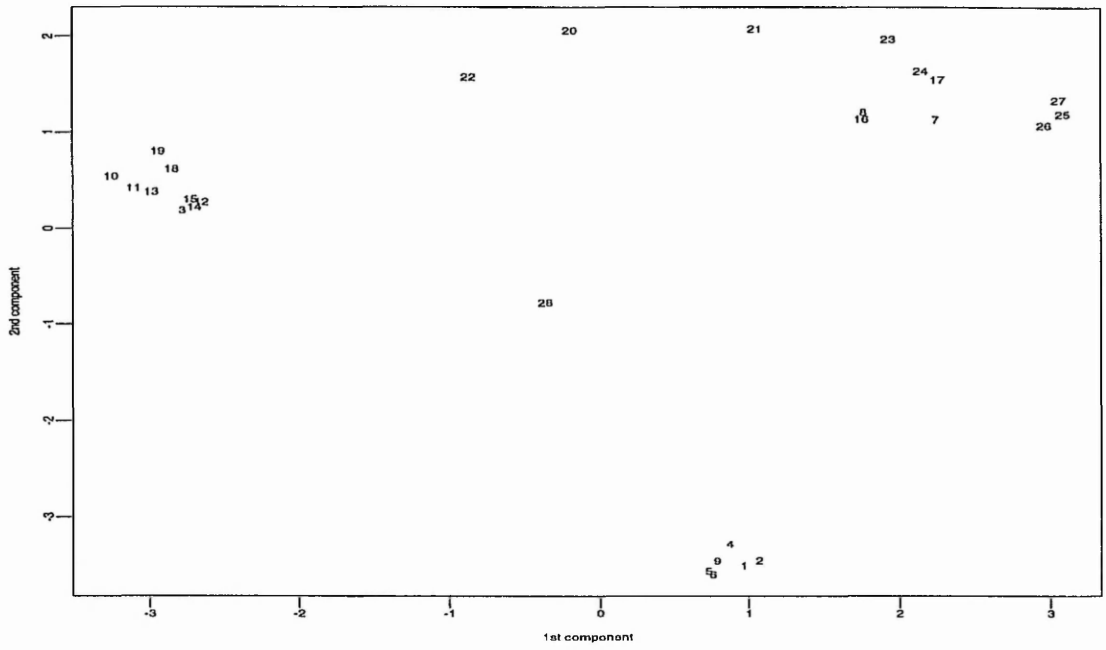
The outliers detected after 10 random starts, outlined in Table 4.9.2, show that very different results are recorded according to the different starting positions, so that there is no consistent pattern.

4.10 Introduction of a simulated outlier

Next a simulated observation is added to the above dataset and we are able to see how the Atkinson and Mulira method deals with an internal outlier (i.e. one which lies in the centre of the three groups). It is hoped that this observation will be identified as an unusual value,

but the presence of grouping may have an effect on the analysis, thus illustrating the problem of the forward algorithm. A principal component analysis, carried out on standardised data, produces the following plot.

Figure 4.10.1 Plot of the first two principal components using standardised data based on the correlation matrix, after the inclusion of a simulated internal outlier



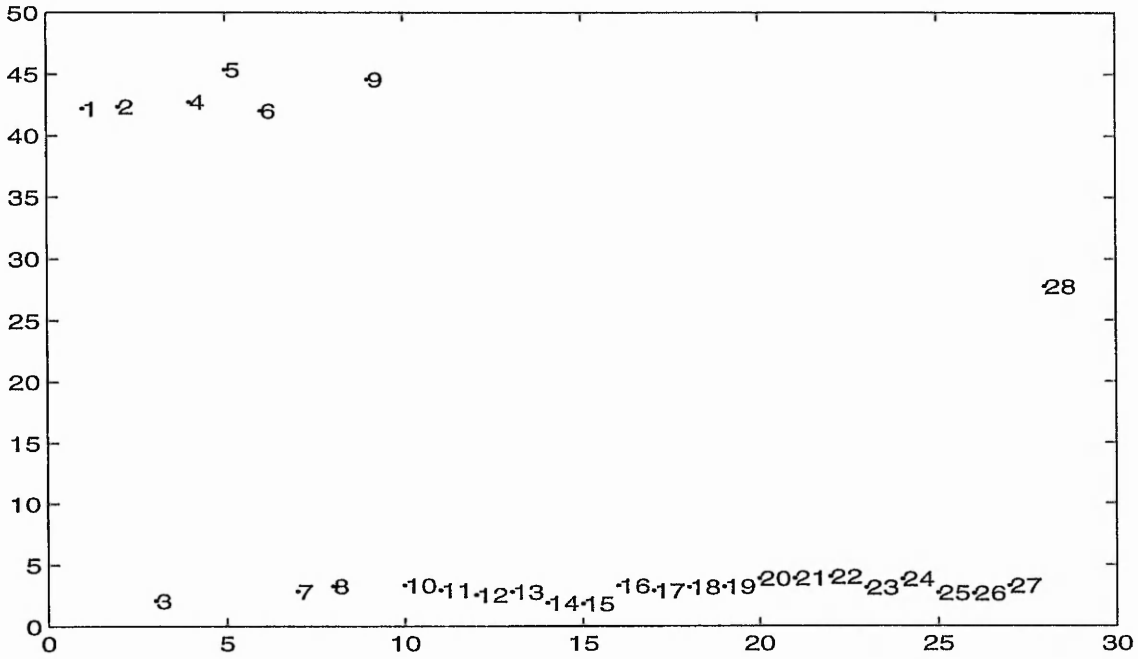
Looking at Figure 4.10.1, it can be seen that observation 28 does indeed appear to be an internal outlier since it lies away from the three distinct groupings. We now perform the Atkinson and Mulira outlier detection method on this new data. The following table shows the 1% and 5% cut-off points against which d^2 is measured.

Table 4.10.1 Table indicating χ^2 values for the following n and p values

	χ^2
n=100% (28), p=10, 5%	17.98
n=100% (28), p=10, 1%	22.79

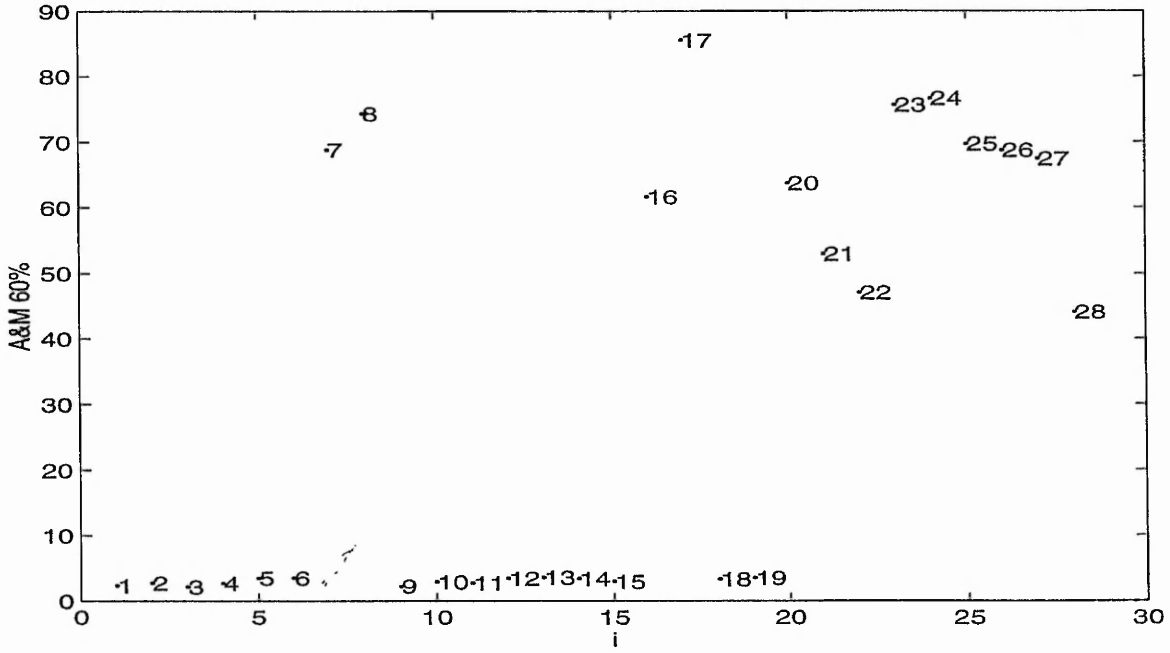
Taking as the initial starting position (7, 8, 16, 17, 20, 21, 22, 23, 24, 25, 26), these observations are found in group 3. When 80% of the data have been included in the analysis, the following index plot is produced.

Figure 4.10.2 Index plot with 80% of the York Minster data, after the inclusion of the internal outlier



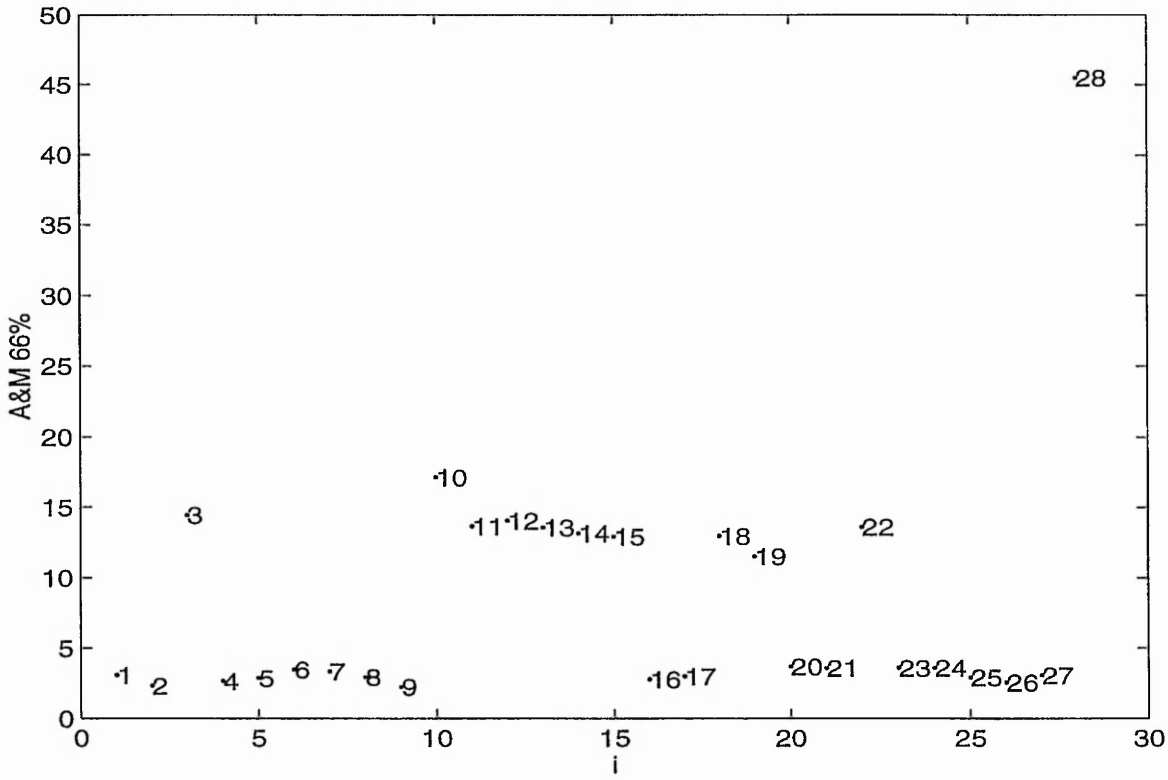
The index plot of Figure 3.10.2 identifies observation 28 as an outlier and it also identifies the observations of group 1 to be outlying since each observation has a d^2 value which exceeds the 5% and the 1% the cut-off points, 17.98 and 22.79 respectively. This plot corresponds to that of Figure 4.8.2. Next taking (3, 10, 11, 12, 13, 14, 15, 18, 19, 1, 2), the plot of Figure 4.10.3 is obtained, where 60% of the data has been included in the analysis.

Figure 4.10.3 Index plot with 60% of the York Minster data, after the inclusion of the internal outlier



As shown in Figure 4.8.3, where the same starting position was used, Figure 4.10.3 indicates that observation 28 is an outlier and the observations belonging to group 3 are also identified, since each observation has a d^2 value which exceeds the 5% and the 1% the cut-off points, 17.98 and 22.79 respectively. Now we use the final starting position of (1, 2, 4, 5, 6, 9, 7, 8, 16, 17, 10). When 66% of the data have been included in the calculation of the Mahalanobis distances, Figure 4.10.4 is obtained. The index plot identifies the observations of group 2 to be outlying as well as identifying observation 28 as an extreme outlier. Although the observations of group 2 are identified as being different from the rest, they do not lie beyond the 5% cut-off point. Observation 28 in this case can be classed as an extreme outlier since it has a d^2 value which exceeds both the 5% and the 1% points.

Figure 4.10.4 Index plot with 66% of the York Minster data, after the inclusion of the internal outlier



As seen previously, by subjective choice of starting position, we are able to highlight each group as an outlier alongside the 'real' outlier, observation 28.

Table 4.10.2 Table indicating initial z, with corresponding outliers

Initial starting position, z	Outliers detected
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	m = 95%, observation **28**
7, 8, 16, 17, 20, 21, 22, 23, 24, 25, 26	m = 80%, observations 1, 2, 4, 5, 6, 9, **28**
3, 10, 11, 12, 13, 14, 15, 18, 19, 1, 2	m = 60%, observations 7, 8, 16, 17, 20, 21, 22, 23, 24, 25, 26, 27, **28**
1, 2, 4, 5, 6, 9, 7, 8, 16, 17, 20	m = 66%, observations 3, 10, 11, 12, 13, 14, 15, 18, 19, 22, **28**

Again it will be of interest to carry out runs from random starts to see whether or not the results obtained are consistent.

Table 4.10.3 Table to show outliers detected after 10 random starts, after the inclusion of an internal outlier

Random starting position, z	Outliers detected
4, 7, 8, 9, 10, 14, 17 18, 21, 22, 28	6, 13, 16, 20, 24, 24, 26 (predominantly group 3)
1, 2, 4, 6, 17, 18, 19, 20, 22, 27, 28	3, 7, 8, 10, 11, 12, 13, 14, 15, 21, 25, 26 (predominantly group 2 and 3)
7, 8, 12, 13, 15, 18, 21, 22, 24, 25, 26	1, 2, 4, 5, 6, 9, **28** (group 1)
2, 9, 10, 14, 15, 16, 19, 20, 21, 23, 27	**28**
6, 7, 8, 9, 17, 18, 22, 23, 24, 25, 27	**28**
1, 5, 7, 8, 11, 12, 16, 20, 22, 27, 28	18, 24, 25, 26
5, 10, 11, 14, 18, 19, 20, 24, 25, 27	6, 22, **28**
3, 5, 8, 11, 13, 14, 18, 21, 22, 23, 26	6, 24, **28**
6, 9, 10, 13, 17, 19, 22, 23, 24, 26, 27	**28**
1, 2, 4, 5, 9, 16, 20, 21, 24, 27, 28	3, 7, 8, 10, 11, 12, 13, 14, 15, 18, 19, 22 (predominantly group 2 and 3)

4.11 Discussion

It should be noted that detection is dependent on whether or not observation 28 is in the random starting sample. This is illustrated further in Table 4.10.3 where different results are obtained, depending on the initial value of z. We are able to conclude that methods of outlier detection are useful when dealing with 'ordinary' data as they appear to detect outliers correctly, but they have extreme difficulty when they are faced with groupings, clusters and/or unusual points in the data.

5. Illustration - analysis of the Southampton glass

We now turn our attention to a particular data set, the assemblage excavated at Southampton. Each technique and method described in Chapters 3 and 4 is applied to this real data, giving us insight into outlier detection methods by considering their effect on a data set.

5.1 Univariate methods for outlier detection

First we analyse the Southampton glass data using a variety of univariate methods, dotplots and box and whisker plots, for each oxide. These univariate methods are ideal for detecting those observations which lie far away from the rest of the data. The outliers listed in Table 5.1.1 have been interpreted as such and go some way to identifying possible outliers in the data for each oxide.

Table 5.1.1 List of outliers detected using univariate methods

Oxide	Outliers (in order of severity)
Al_2O_3	122
Fe_2O_3	205, 122
MgO	133, 195, 154
CaO	122, 195
Na_2O	
K_2O	154
TiO_2	122, 258
P_2O_5	133
MnO	258
PbO	133,5
SbO	108,42

Observations 122 and 133 feature more prominently as outliers, having high (Al_2O_3, Fe_2O_3, CaO, TiO_2) and (MgO, P_2O_5, PbO) content levels respectively.

5.2 Multivariate methods for outlier detection

Having looked at a variety of multivariate outlier detection methods available we are now able to actually use some the statistics outlined, namely q^2_j, t^2_j, u^2_j, v^2_j, d^2_j, Wilk's, Hadi's and the Atkinson and Mulira method. The statistics have been obtained using the software package MATLAB. Of all the methods described, all but the Rousseeuw and van Zomeren technique have been used. Some of the statistics, namely d^2_j, Wilk's and the Atkinson and Mulira method, where m=n, produce the same plots, the only difference being their

orientation and/or their scaling. Outlier detection in the following plots is visual since the extreme outliers tend to lie far away from the rest of the observations but the resulting plots of Atkinson and Mulira use a cut-off point obtained via a chi-square critical value. The first set of outlier detection statistics to be used consists of q^2_j, t^2_j, u^2_j, v^2_j, and d^2_j, (Krzanowski and Marriott, 1994).

Figure 5.2.5 Index plot of q^2_j for Southampton glass

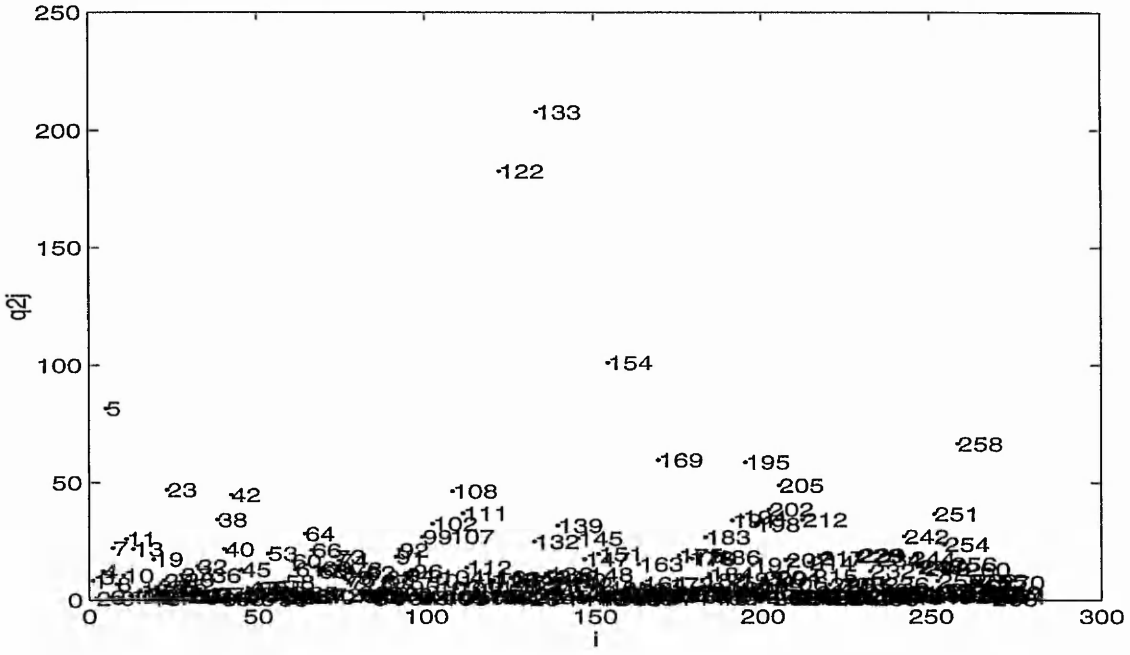


Figure 5.2.6 Index plot of t^2_i for Southampton glass

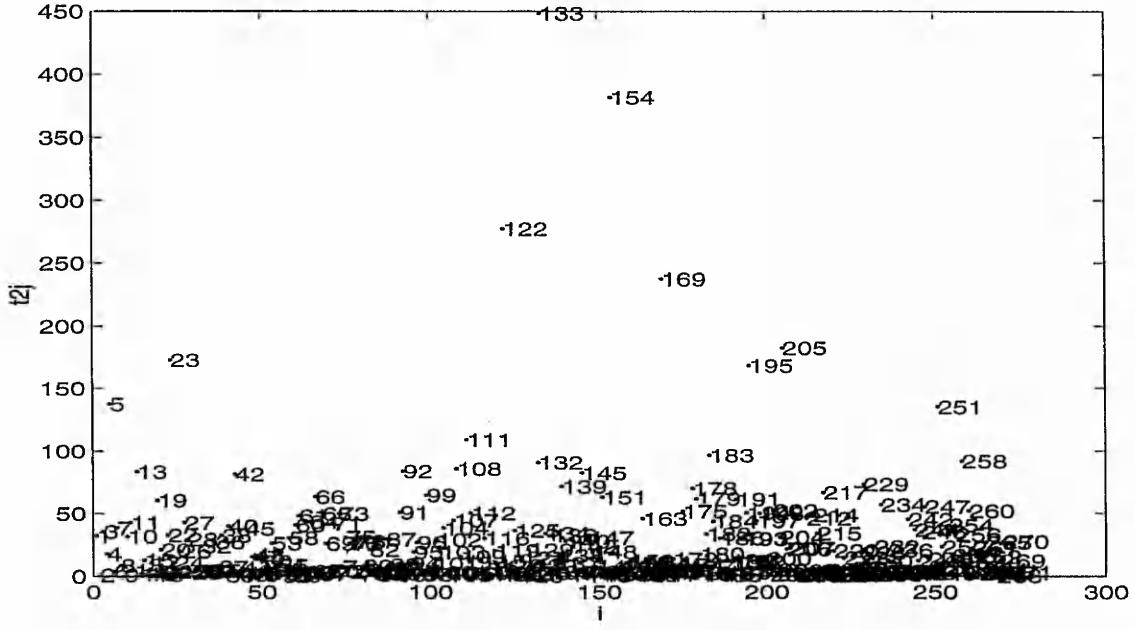


Figure 5.2.7 Index plot of u^2_i for Southampton glass

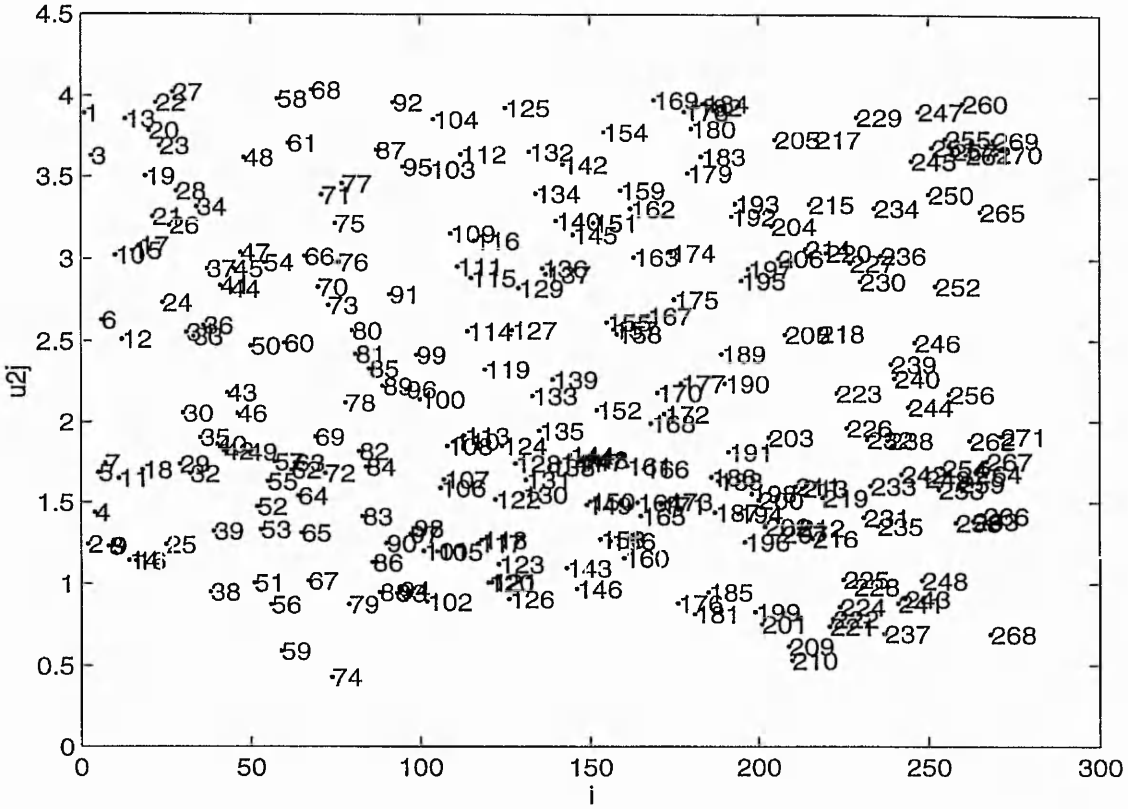


Figure 5.2.8 Index plot of v^2_i for Southampton glass

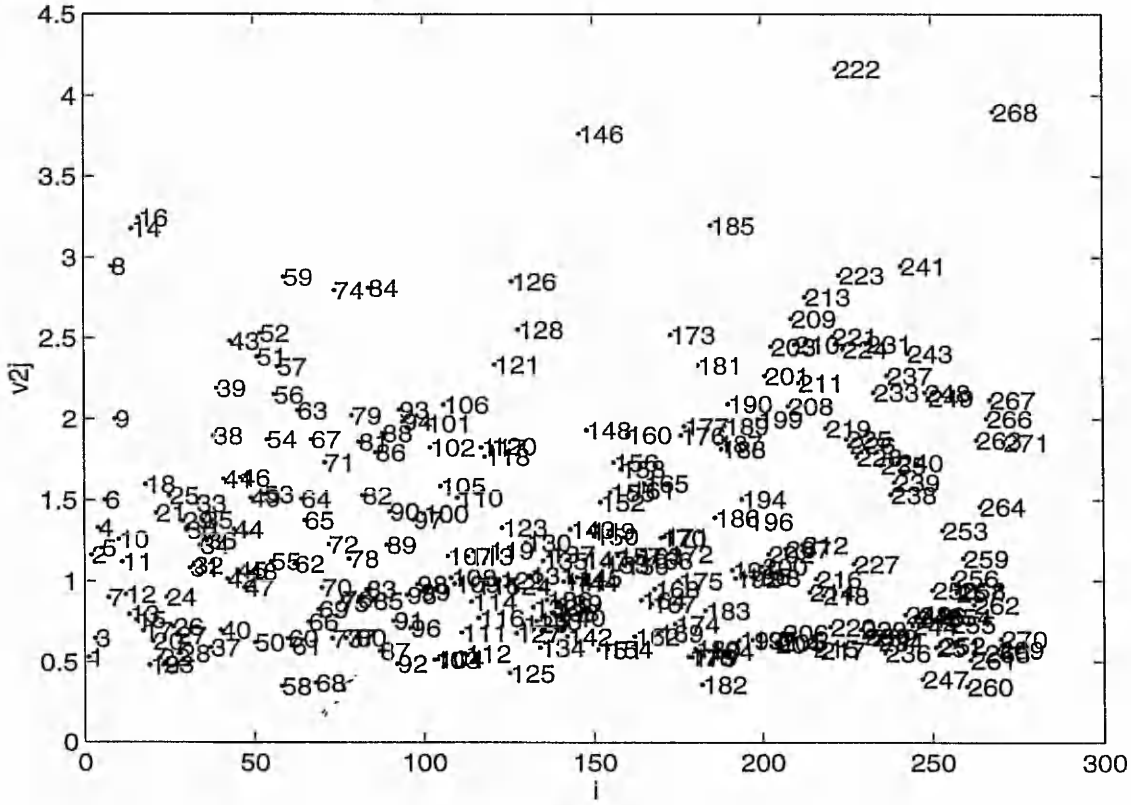
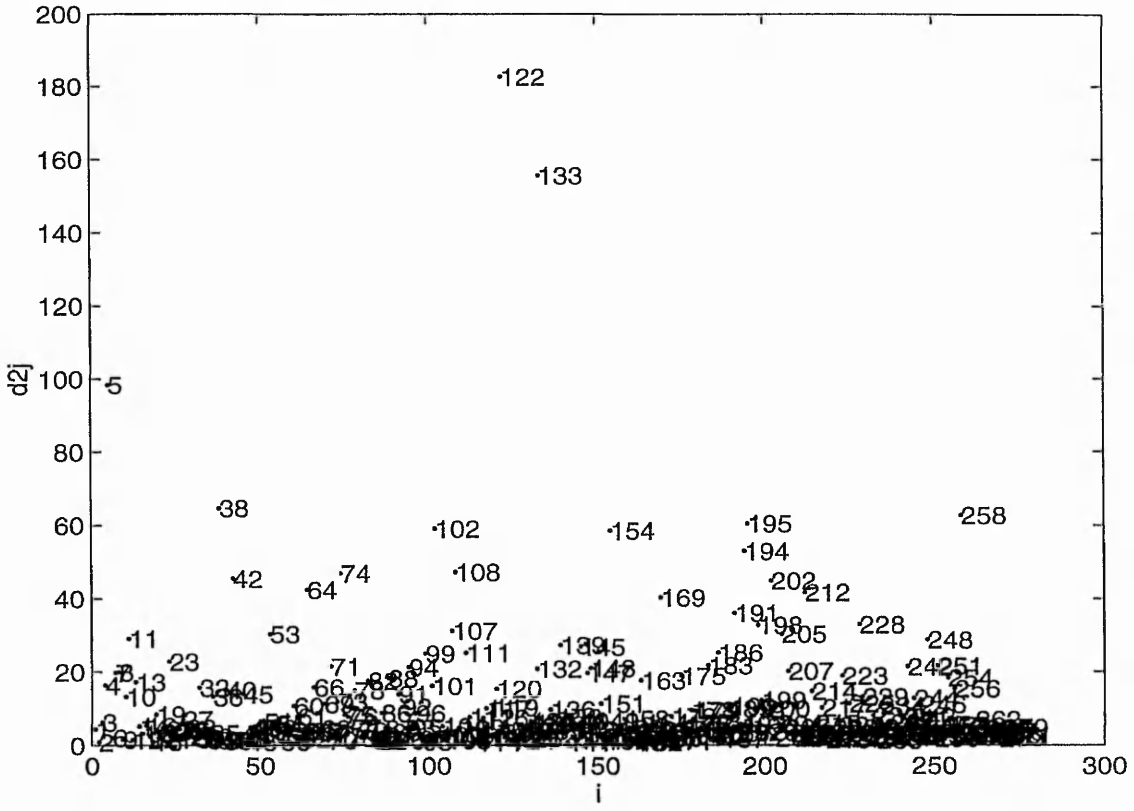


Figure 5.2.9 Index plot of d^2_i for Southampton glass



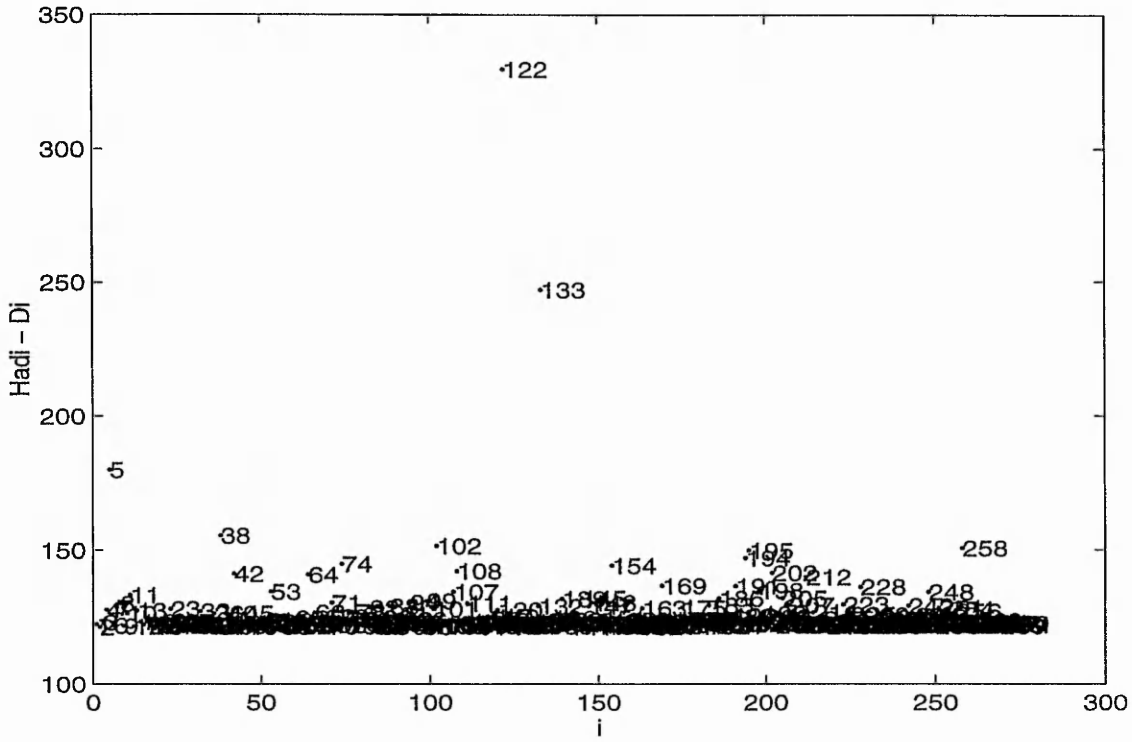
The index plot of q^2_j, Figure 5.2.5, identifies that observations 133 followed by 122 excessively inflate the overall scale of the plot. The statistics t^2_j, u^2_j and d^2_j measure outliers according to the first few principal components and will identify those values which inflate variances, covariances or correlations as outliers in the data. The index plot of t^2_j, Figure 5.2.6, shows observations 133, 154, 122 and 169 are plainly visible outlying values with 133 being the most extreme. The index plot of u^2_j,

Figure 5.2.7, does not give any indication of outlying values. This is due to the fact that u^2_j puts more emphasis on the orientation of the first few components. According to the index plot of d^2_j, Figure 5.2.9, observations 122, 133 and 5 have high values of d^2_j and can therefore be classed as outliers. The remaining statistic, v^2_j, measures outliers according to the last few principal components and the plot of Figure 5.2.8 does not identify any extreme outlying observations.

Wilk's statistic for detecting outliers produces the same plot as that of d^2_j, seen in Figure 5.2.9, apart from orientation and scaling.

Figure 5.2.10 shows those outliers detected using the Hadi method. Observation 122 followed by 133 has the largest value of $D_i(C_b, S_b)$, (4.5.1).

Figure 5.2.10 Index plot of the distances derived from Hadi's algorithm for Southampton glass



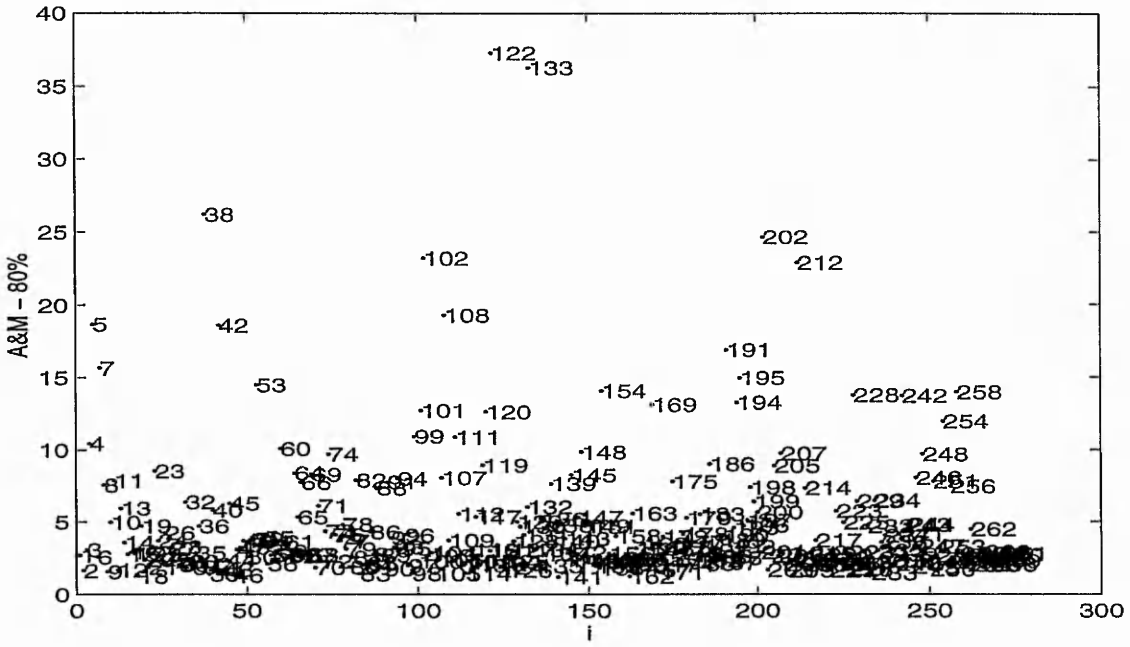
The index plot of Figure 5.2.11 shows the Mahalanobis distances when m=80% of n, that is 217, where 217 observations have been used to calculate the mean and covariance matrix. Here a cut-off point can be used to define the outliers, see (4.4.4).

Table 5.2.1 Indicating χ^2 values for the following n and p values

	χ^2
n=271, p=11, 5%	18.27
n=271, p=11, 1%	23.17

At the 1% significance level observations 122 and 133 both have d^2 values well in excess of this critical value, therefore identifying them as outliers. Two other observations also have d^2 values slightly greater than the critical value, namely 38 and 202.

Figure 5.2.11 Index plot where 80% of the Southampton glass data have been included in the calculation of the Mahalanobis distances



Now we look at when all 271 observations have been used to calculate the mean and covariance matrix, i.e. 100% of the data have been used to calculate the Mahalanobis distances. At the 1% significance level observation 122 has a d^2 value in excess of this critical value and at the 5% level both 122 and 133 are anomalous observations. The 100% Atkinson and Mulira plot is identical to the plot of d^2_j observed in Figure 5.2.9.

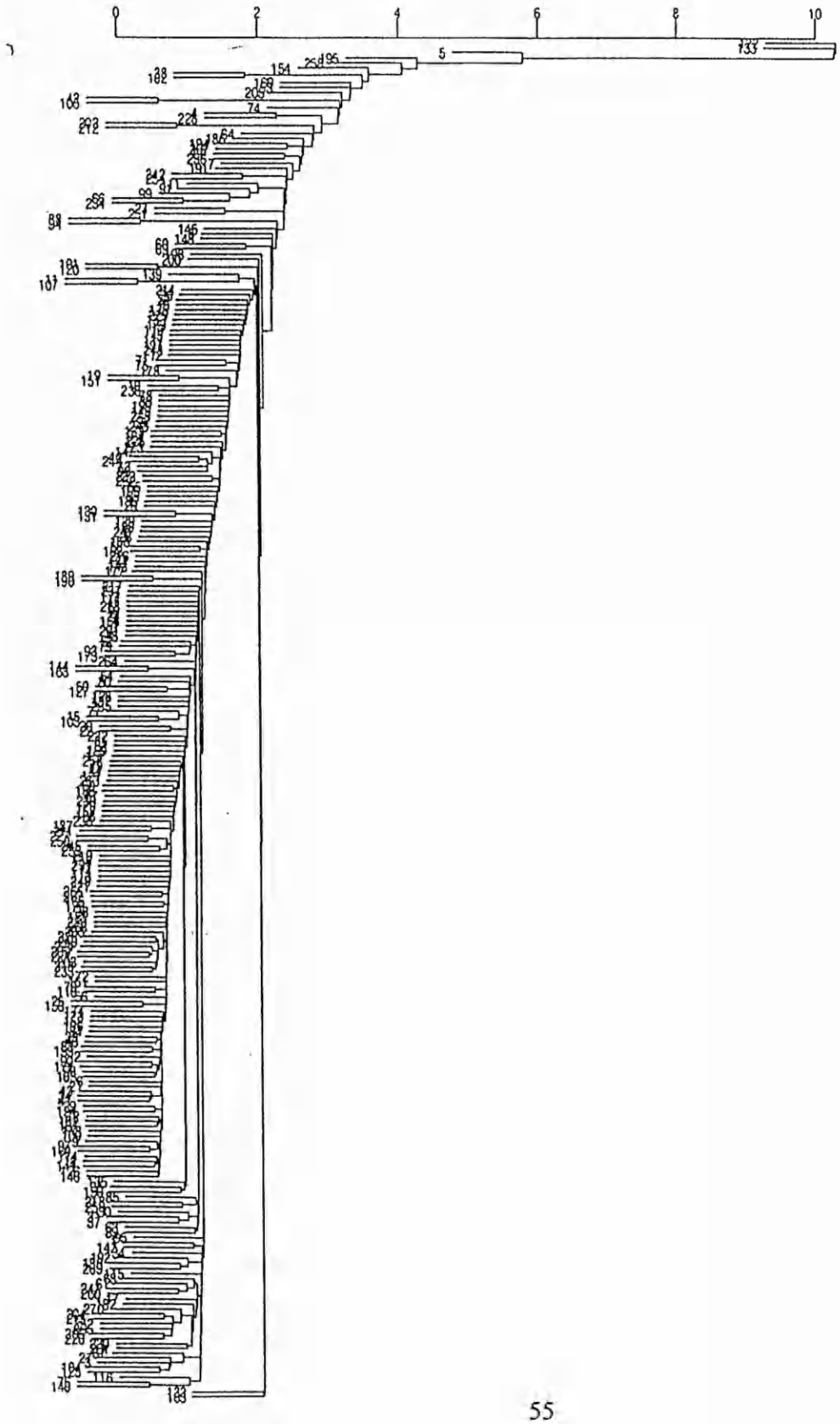
Figure 5.2.11 is based on a starting position of m = 12 (p + 1), where the initial subset contains the following 12 observations: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. The method is repeated with various starting positions: [5, 122, 133, 258, 38, 42, 102, 108, 202, 212, 191, 195] and [1, 2, 3, 4, 5, 122, 133, 154, 169, 195, 258, 12]. The anomalous observations seen originally are included in subsequent analyses as the starting position and the same results are recorded. This is unlike the results observed in Chapter 4, where the results obtained are very much dependent on the starting position. In this case the Southampton data does not have distinct groupings which are totally separate from each other, as seen in the York Minster data, therefore the Atkinson and Mulira method does not have difficulty in identifying those observations which are outlying.

5.3 Cluster analysis as an outlier detection method

As well as using the outlier detection methods discussed in this chapter, cluster analysis may be used as a method which, as well as detecting clusters or groupings in the data, will also identify possible outliers. Average linkage, according to work done and conclusions made in the 1960's, is the most useful of all the hierarchical techniques (Everitt, 1993). It must be noted that cluster analysis in this thesis is used only as a 'back-up' to the other univariate and multivariate techniques used throughout the course of the thesis, and therefore the results obtained are for reference purposes only. Figure 5.3.1 identifies observation 133 as the most extreme outlier followed by observation 122. Observations 5, 258, 195 and 154 are also identified as separate clusters, i.e. outliers. Baxter (1994, p. 182) states that single linkage is of limited practical use when working with archaeological data and looking for possible groupings, because of chaining. However it is potentially useful if one is interested in outlier detection. The dendrogram of Figure 5.3.2 does identify those observations which have previously been detected, namely observations 122, 133, 5, 195, 258 and 154. Complete linkage is not used as it is not useful for outlier detection, see Chapter 3.

Figure 5.3.1 Average link cluster analysis of the Southampton glass

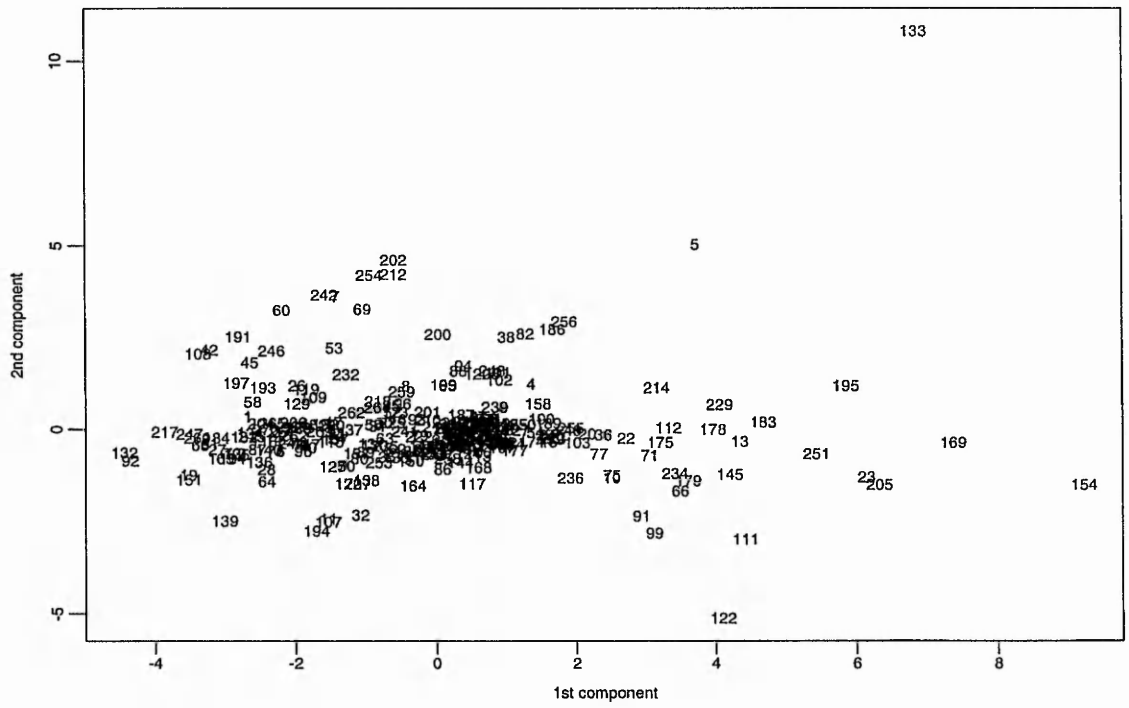
Figure 5.3.2 Single link cluster analysis of the Southampton glass



5.4 Principal components analysis as an outlier detection method

Another method which can prove useful in detecting outliers is the principal component plot. Figure 5.4.1 is a plot of the first two principal components using standardised data. Observation 133 appears to lie furthest away from the rest of the data followed by 5, 154 and 122. Observations 169 and 195, which were identified as possible outliers by the index plot of t^2_j and the average link dendrogram respectively, are also possible outliers on this PCA plot.

Figure 5.4.1 Plot of the first two principal components using the correlation matrix based on standardised data labelled according to observation number



It is of some interest to look at the component scores of the initial principal components analysis, since the higher order components may indicate observations which are clearly outliers, but are not evident on the PCA plot.

Table 5.4.1 Table listing outliers detected by each component

Outliers	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
5					×	×					
38						×					
64							×				
108					×						
122		×	×	×					×		×
133		×			×				×		
145								×			
154	×										
169	×										
195								×			
223										×	
258			×								

Table 5.4.1 lists those observations which appear to be outlying on univariate plots of the higher order components. Observations 122 and 133 are detected in plots of PC's 2 - 5, 9 and 11. Additionally, observations 154 and 169 are actually detected in a plot of the first principal component, thus indicating they are outlying values. Observations 195 and 258 are also detected in the higher order component plots, namely PC8 and PC3 respectively. Additional observations are also indicated on some of the higher order plots but are not substantiated by the multivariate outlier detection methods outlined above. This could be due to the fact that the methods are not perfect and will not necessarily identify all outlying values.

The statistics q^2_j, t^2_j, u^2_j, v^2_j, d^2_j and the Wilk's, Hadi and the Atkinson and Mulira methods used in Figure 5.2.5 - Figure 5.2.11 have been used on the original, i.e. unstandardised, data. In the following table '×' indicates an outlier detected by the statistic stated.

Table 5.4.2 Table showing outliers suggested by the various methods

Obs	Univariate methods	q^2_j	t^2_j	d^2_j	Hadi	A&M 80%	Ave	Sin	PCA
5	×						×	×	×
122	×	×	×	×	×	×	×	×	×
133	×	×	×	×	×	×	×	×	×
154	×		×						×
169			×						×
195	×						×		×
258	×						×		
42	×								
108	×								
205	×								

Note: A&M - Atkinson and Mulira method

Ave. - Average link cluster analysis

Sin. - Single link cluster analysis

Com. - Complete link cluster analysis

5.5 Discussion

The statistics d^2_j and t^2_j and the method of Atkinson and Mulira appear to be the most useful and straightforward of all the outlier detection methods outlined above. This is because they are simple to use and produce easily interpretable results, more so than Hadi's method which has a complicated starting procedure to find robust estimators of location and covariance matrix. As already stated, the Rousseeuw and van Zomeren method has been greatly improved upon by Hadi and then subsequently by Atkinson and Mulira so it is fruitless to return to this method of outlier detection. Wilk's statistic produces identical results to both d^2_j and Atkinson and Mulira, where m = n, and cluster analysis is useful in verifying outliers that have already been identified using the 'true' outlier detection methods, although in the case of the Southampton glass these analyses discover additional outliers. PCA again verifies all those outliers that have already been detected by each of the above methods of outlier detection and the higher order component plots also detect some observations not necessarily clear outliers on the plot of the first two components. Univariate methods are ideal for detecting those observations which lie far away from the rest of the data and these identify the same outlying observations.

Typological analyses were initially carried out on the data by Heyworth (1991). Descriptions of the outliers are given in Table 5.5.1

Table 5.5.1 Colour/chemical descriptions of the outliers

Observation	Description
5	Red fragments, high contents of lead of nearly 5%
122	Brown/Yellow, high contents of iron
133	Red fragments, high lead contents of nearly 5%
154	Light blue
169	Dark opaque fragment, high iron contents of nearly 5%
195	Green fragment
258	Green fragment

The majority of the glass specimens excavated at Southampton are lightly tinted blue or green. Six of the seven outliers are strongly coloured and it is possible that they date to a

different period from the rest of the assemblage (Heyworth 1991). We would therefore expect these observations to be detected as unusual when compared to the rest of the data. Table 5.5.2 lists those oxides which are found to be of a high content in the outliers detected.

Table 5.5.2 Table listing the high content levels of the outliers

Outlier	Element
5	PbO
122	Al_2O_3, Fe_2O_3, CaO, TiO_2
133	MgO, P_2O_5, PbO
154	MgO, K_2O
169	Fe_2O_3
195	MgO, CaO
258	TiO_2, MnO

It is the high content levels of the oxides which cause the corresponding observations to be strongly coloured and thus outlying in the various analyses. Observations 122 and 133 are very obvious outliers on all the plots used, including univariate dotplots and box and whisker plots. This would suggest these are univariate outliers since they are detectable by looking at plots of the original variables. A univariate outlier is one which causes a large increase in one or more of the variances of the original variables, thus it will be extreme on those variables. The PCA and cluster analysis suggest observation 5 is an outlier, but this is not detected by any of the other multivariate methods, although it does stand out in many of the plots. One reason for this is that it is detectable by the higher components, namely PC5 and PC6. This suggests that observation 5 is a multivariate outlier. Observations 195 and 258 are both strongly coloured green and, as seen with observation 5, are both detected by cluster analysis and by the higher order component plots, PC3 and PC8 respectively, again suggesting multivariate outliers. Observations 154 and 169 are both detected on the first principal component.

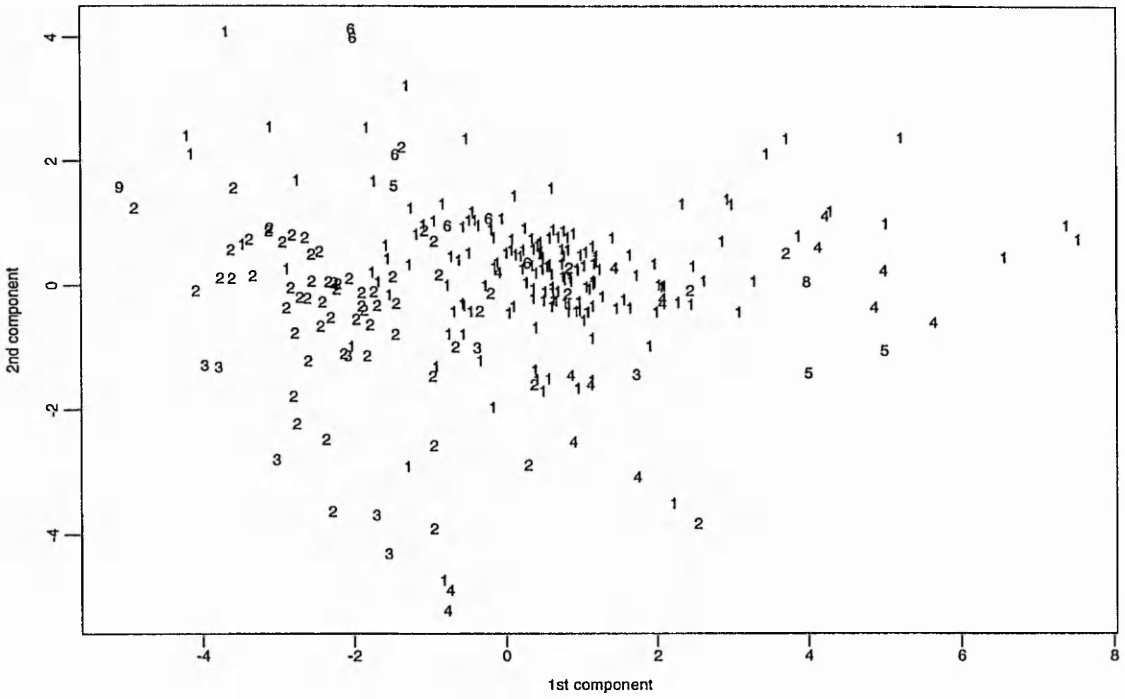
As observed, principal components analysis can be useful when used as a method for outlier detection. Further usage of PCA will be explored in the analyses of the Southampton glass data in the next sub-section.

5.6 Substantive analysis of the Southampton glass

The content levels of 11 of the major/minor oxide components of the Southampton glass assemblage, namely Al_2O_3, Fe_2O_3, MgO, CaO, Na_2O, K_2O, TiO_2, P_2O_5, MnO, PbO, SbO have been measured. From now on the oxides will be referred using the chemical symbol of the corresponding element. Univariate analyses of these data identify some outlying specimens, refer to Table 5.1.1. Observations 122 and 133 feature more prominently as extreme outliers, having high (Al, Fe, Mg, Ca, Ti) and (Mg, P, Pb) content levels respectively, but observations 5, 154, 195 and 258 are also identified as outlying. Table 5.4.2 lists those outliers identified using the various multivariate outlier detection methods, including principal components analysis and cluster analysis, outlined in Chapter 4. Observations 122 and 133 appear to feature as outliers in the analyses undertaken. It may also be of some interest to look at the component scores of the principal components analysis, since the higher order components may indicate observations which are clearly outliers, but are not evident on the PCA plot, see Table 5.4.1.

Having focused on methods of detecting outliers we have found the following observations to be physically distinct: 5, 122, 133, 154, 169, 195 and 258. These observations are very strongly coloured, (5-red, 122-green/yellow, 133-red, 169 - strongly coloured opaque red, 195-green and 258-green), unlike the majority of the fragments which are lightly tinted light blue/light green, see Table 5.5.1. Removing these from the analysis leaves 264 specimens in all. Now if we look at a plot of the first two principal components after the removal of the seven outliers and label according to colour, Figure 5.6.2 indicates a separation of the data on the basis of colour, (1-light blue, 2-light green, 3-blue, 4-green, 5-red, 6-green/yellow, 7-brown/yellow, 8-polychrome, 9-dark opaque), showing that a majority of the fragments are light blue in colour, followed by a large proportion which are tinted light green. There is also a lot of dispersion and overlap with a majority of the 4's, 5's, 8's and 9's round the periphery of the plot. We now present a substantive analysis of the glass after the removal of the outliers.

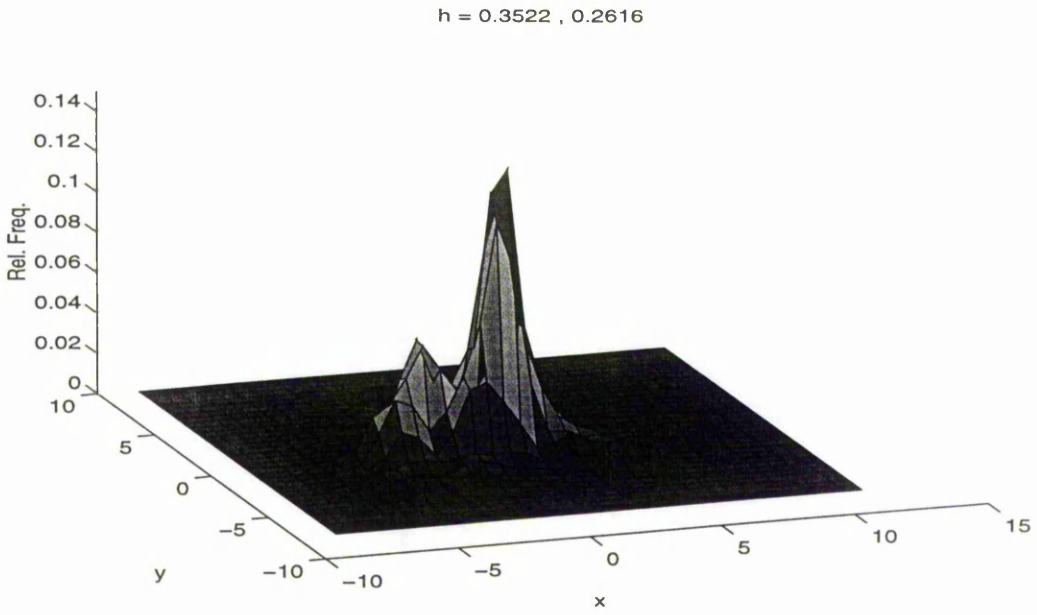
Figure 5.6.2 Plot of the first two principal components labelled according to colour - after the removal of the seven outliers



The first component accounts for 42% of the variation in the data. Several elements have component coefficients of around 0.4, namely Fe, Al, Mg and K. Table 5.6.4 shows the correlations between the components and the elements and shows these same elements to have an r value in excess of 0.8, showing high positive inter-correlation. Table 5.6.4 also shows that the second principal component correlates with P, Pb and Sb and the third principal component with Mn. Together, the first two components account for 58% of variation in the data and the first four components are needed to 'explain' 80% of the variation.

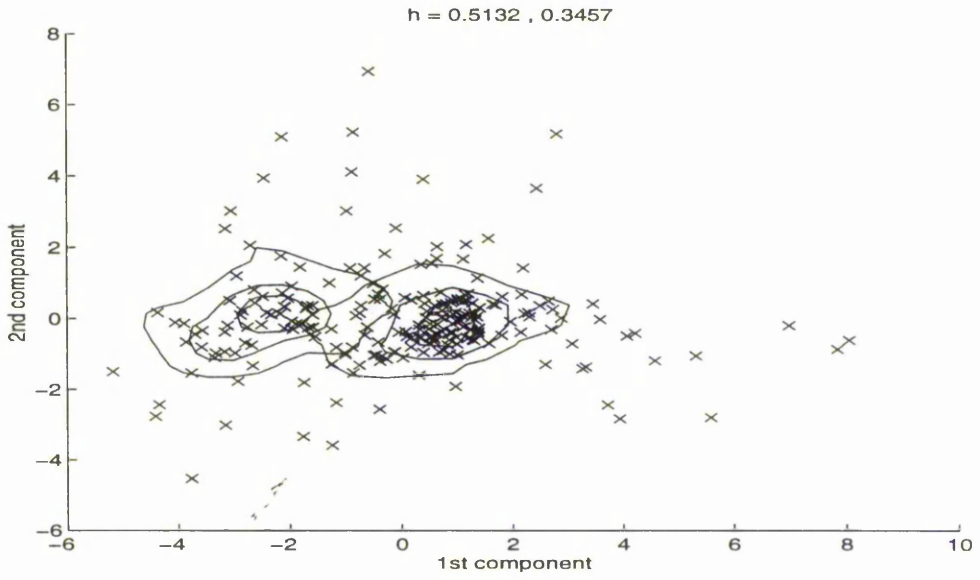
Now only those observations coloured light blue (1) and light green (2) have been plotted in order to make it easier to see the separation. The KDE plot, using the STE method with $h_1 = 0.3522$ and $h_2 = 0.2616$, of Figure 5.6.3 gives an alternative view of the data, showing the clear bimodality.

Figure 5.6.3 Kernel density estimate plot of the first two principal components using observations coloured light blue and light green only



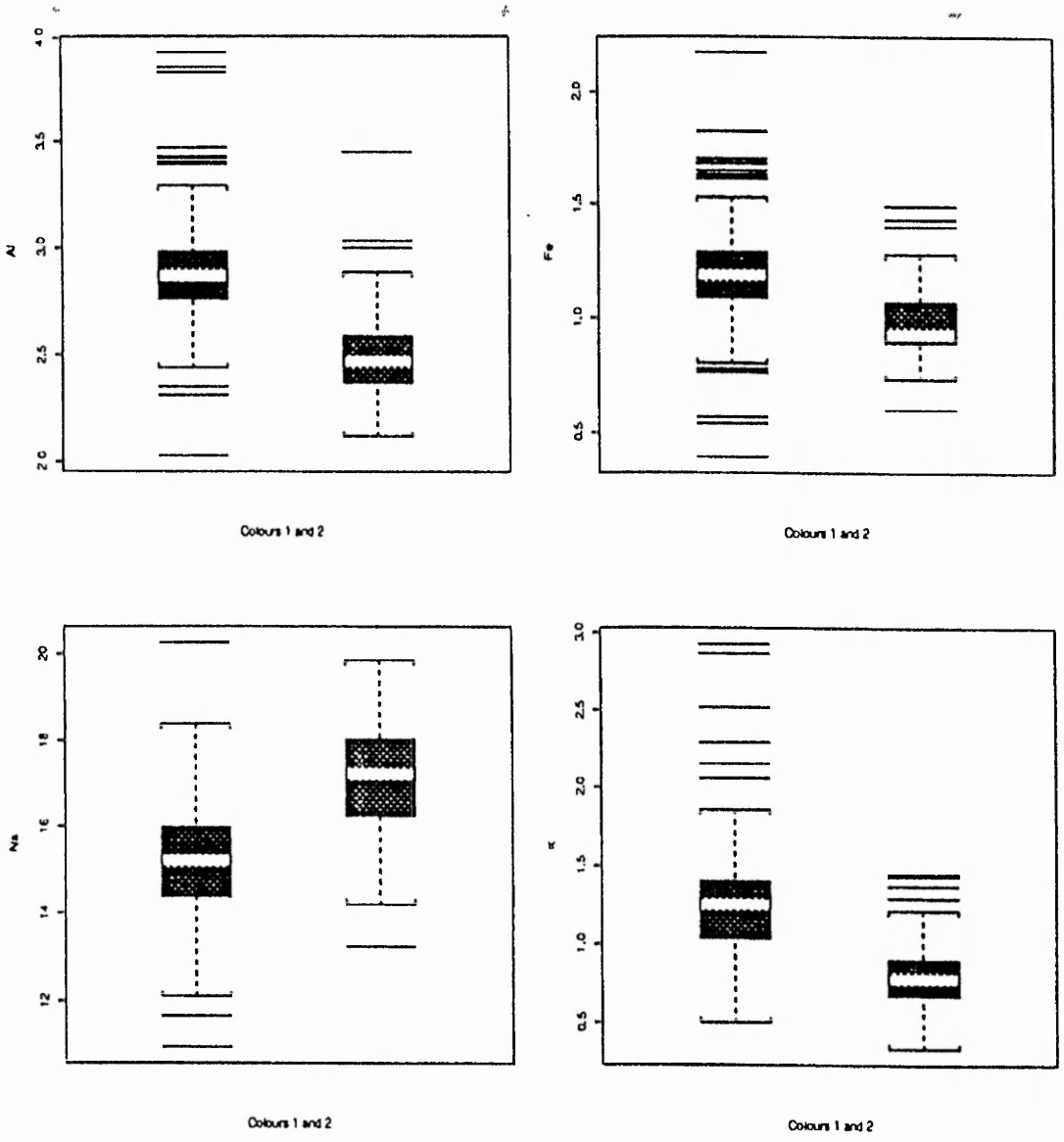
The separate contour plot of Figure 5.6.4, using those observations coloured light blue and light green only, shows the two main groupings quite clearly. Here the data have been treated as two separate groups and each group has been contoured separately.

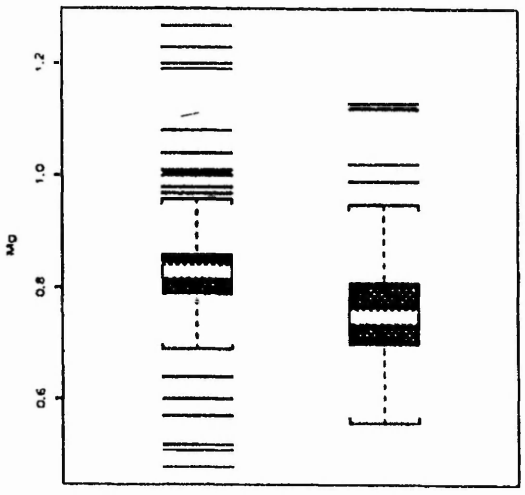
Figure 5.6.4 Separate contour plot for those observations coloured light blue(1) and light green(2). Observations coloured light blue are encapsulated in the contour to the right of the plot and observations coloured light green in the contour to the left of the plot. Contours at 25,50 and 75%.

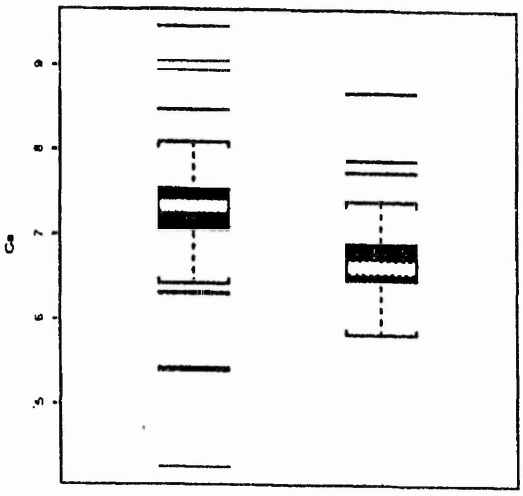


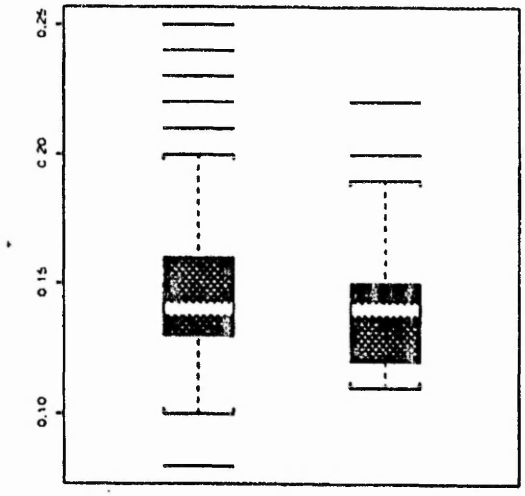
The STE method for the selection of h_1 and h_2 has been used for the contour plot of Figure 5.6.4, where $h_1 = 0.5132$ and $h_2 = 0.3457$. Contours correspond to the 25, 50 and 75% levels for each group.

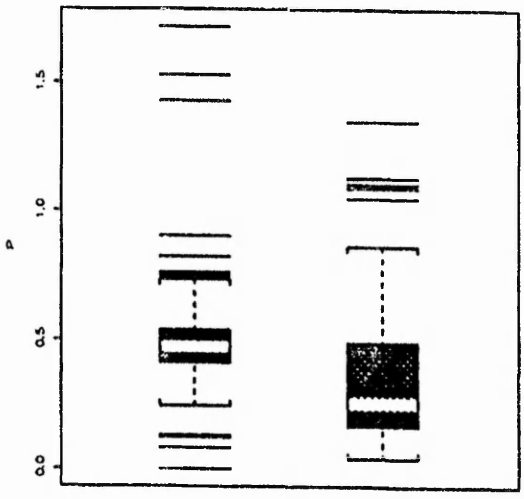
Figure 5.6.5 Boxplots showing the chemical composition of the observations coloured light blue(1) and light green(2) only

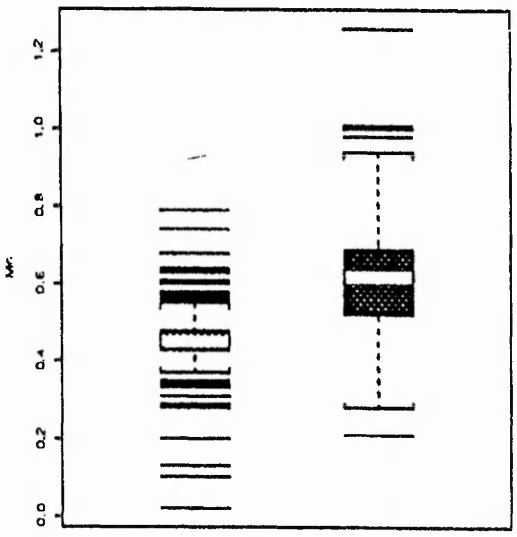


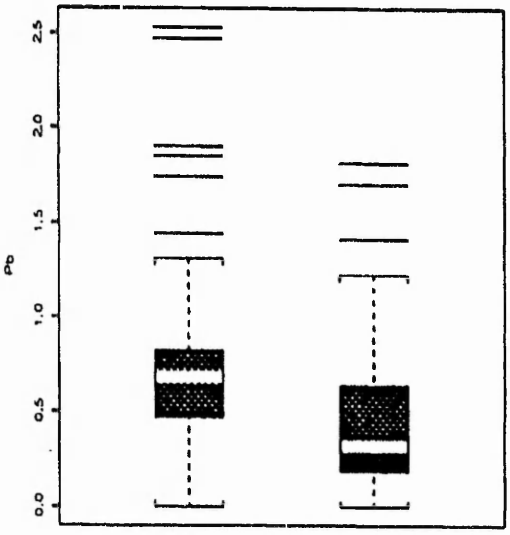

Mg
0.6
0.8
1.0
1.2
Colours 1 and 2

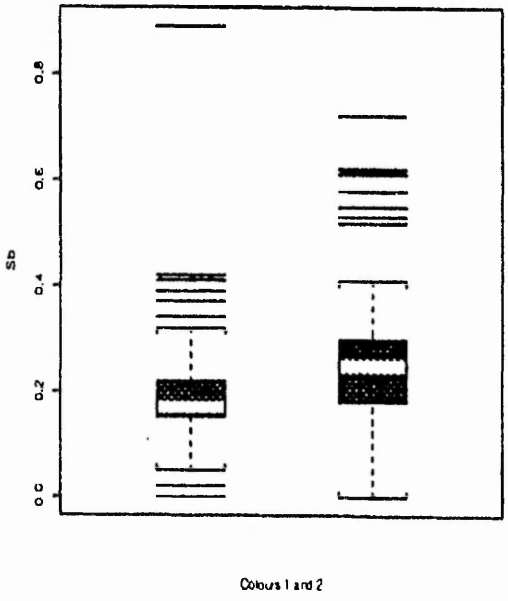

Ca
5
6
7
8
9
Colours 1 and 2


C 20
0.10
0.15
0.20
0.25
Colours 1 and 2


P
0.0
0.5
1.0
1.5
Colours 1 and 2


Mn
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Colours 1 and 2


Pb
0.0
0.5
1.0
1.5
2.0
2.5
Colours 1 and 2


Sb
0.0
0.2
0.4
0.6
0.8
Colours 1 and 2

The boxplots of Figure 5.6.5 show the chemical compositions of the two colour groups. It is evident that the two main concentrations are colour related but these colour groups are also compositionally different. The light blue glass has a relatively high content of K and Ca but low levels of Na, thus indicating the glass is possibly made with plant alkalis. It also has high levels of Fe, Al and P. This is to be expected because they all enter the process as a complex via the silica, so the higher the amount of Al the higher the amounts of Fe and P and vice versa. The level of Mn (approx. 0.45%) can be attributed to the result of the normal inclusion of impurities since higher amounts would be needed to effectively decolorize the glass. The level of Fe (approx. 1.3%) shows that it is present as a contaminant of the raw materials. The light blue colour appears to be a result of the redox conditions in the furnace, since glass formed under reducing conditions tends to a blue colour. The light green glass has lower levels of Al, Fe and P, but it has high levels of Na and Ca suggesting a possible alkali source of saltwater plants. Mn appears to be acting as a redox element which was added to the melt during the glass making process since the higher level of Mn (approx. 0.7%) suggests deliberate addition. It is also possible that the light green glass was formed under oxidising conditions and cooled slowly after subsequent melts. In both the light blue and light green glass the level of Sb is low, suggesting it was not used as a decolorizer in the assemblage found at Southampton.

The results of discriminant analysis with cross validation verify that the two groupings associated with colour are present, since 89% of observations are correctly classified into these two colour groups.

The correlations of the elements, seen in Table 5.6.3, are for all the data excluding the seven outliers. It can be seen that Fe is highly positively correlated with Al, (and K and Ti to a lesser extent), and Fe and Mn are uncorrelated. Since the data are not homogeneous it is impossible to state why these elements may or may not be closely related, but suggestions have been made that Fe and Al may have entered the batch together and that Fe and Mn did not enter together, suggesting that Fe entered as a contaminant of the raw materials but Mn was added separately as a redox element (Jackson, 1992).

Table 5.6.3 Correlations of all the data - excluding the seven outliers

	Al	Fe	Mg	Ca	Na	K	Ti	P	Mn	Pb
Fe	**0.73**									
Mg	0.59	0.71								
Ca	0.54	0.56	0.71							
Na	-0.50	-0.43	-0.30	-0.43						
K	0.67	0.67	0.70	0.61	-0.57					
Ti	0.61	0.71	0.59	0.25	-0.20	0.50				
P	0.15	0.37	0.24	0.07	-0.04	0.11	0.04			
Mn	-0.45	**-0.02**	0.15	-0.08	0.45	-0.20	0.26	-0.06		
Pb	0.17	0.41	0.36	0.29	-0.12	0.16	0.07	0.39	-0.03	
Sb	-0.25	0.01	-0.12	-0.11	0.33	-0.24	-0.29	0.30	0.06	0.19

The correlations of the elements and the principal components may give more insight into which, if any, element dominates the analysis. Table 5.6.4 shows that the first principal component is dominated by Fe, although Al, K and Mg also have coefficients in excess of 0.8, and Mn dominates the third component, suggesting the distinction of the two main concentrations, (light blue and light green colours), could be based wholly upon the differing levels of these two important elements, see, Figure 5.6.6 which shows a plot of the first and the third components.

Table 5.6.4 Correlations of the elements and the first three principal components

	pc1	pc2	pc3
Al	-0.85	-0.21	0.13
Fe	-0.89	0.26	-0.05
Mg	-0.83	0.24	-0.24
Ca	-0.74	0.02	0.04
Na	0.61	0.44	-0.35
K	-0.85	-0.16	-0.00
Ti	-0.68	0.02	-0.56
P	-0.26	0.63	0.38
Mn	0.18	0.44	-0.81
Pb	-0.37	0.57	0.30
Sb	0.24	0.68	0.31

Figure 5.6.6 Plot of the 1st and the 3rd principal components, after the removal of the seven outliers

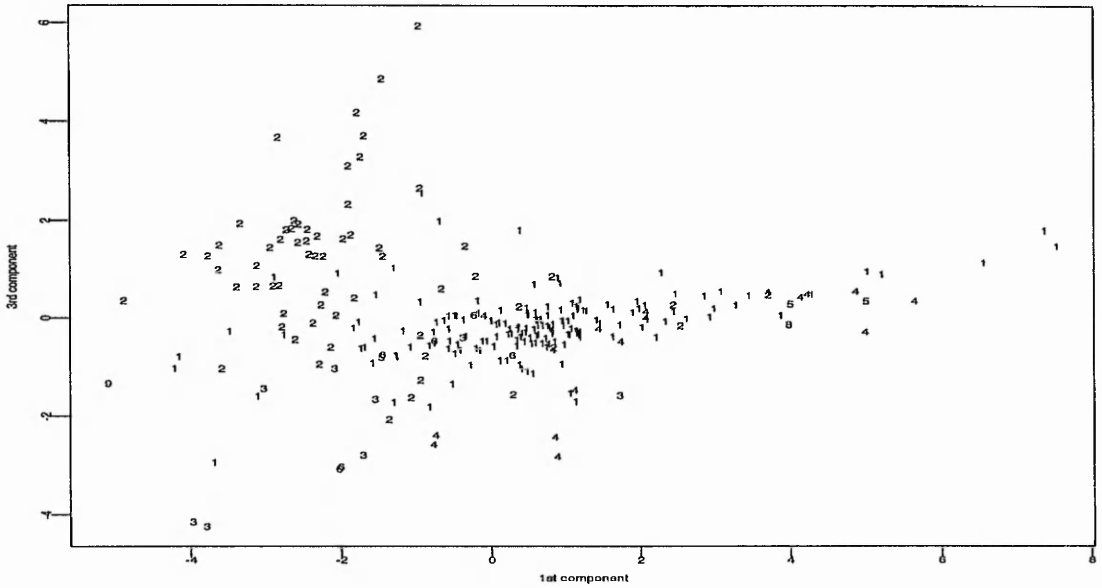
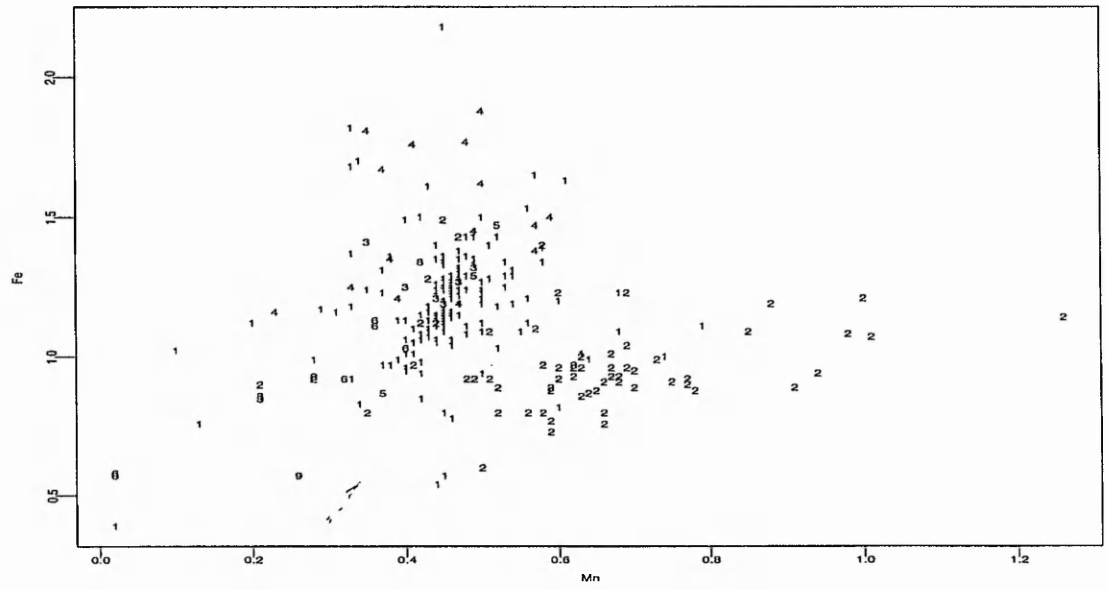


Figure 5.6.7 is a plot of Fe against Mn, using the original data after the removal of the seven outliers. As already suggested the distinction of the two main colour concentrations is based on the Fe and Mn content of the glass. The plot indicates that just by looking at a plot of Fe vs Mn, we are able to see the data separating into two main concentrations based on colour.

Figure 5.6.7 Plot of Fe against Mn, after the removal of seven outliers



Typological analyses were initially carried out on the data by Heyworth (1991) and the following descriptions of the outliers are given in Table 5.5.1. Most of the fragments excavated are lightly tinted. Six out of the seven outliers are found to be strongly coloured with high contents of Pb, Fe and Al. It is these high contents which cause the corresponding observations to be outlying in the various analyses and they also indicate that these fragments possibly do not date back to the early Medieval period. Overall we are able to conclude that the two main concentrations are colour related, and although these colour related groups do appear to be compositionally distinct, there is some overlap.

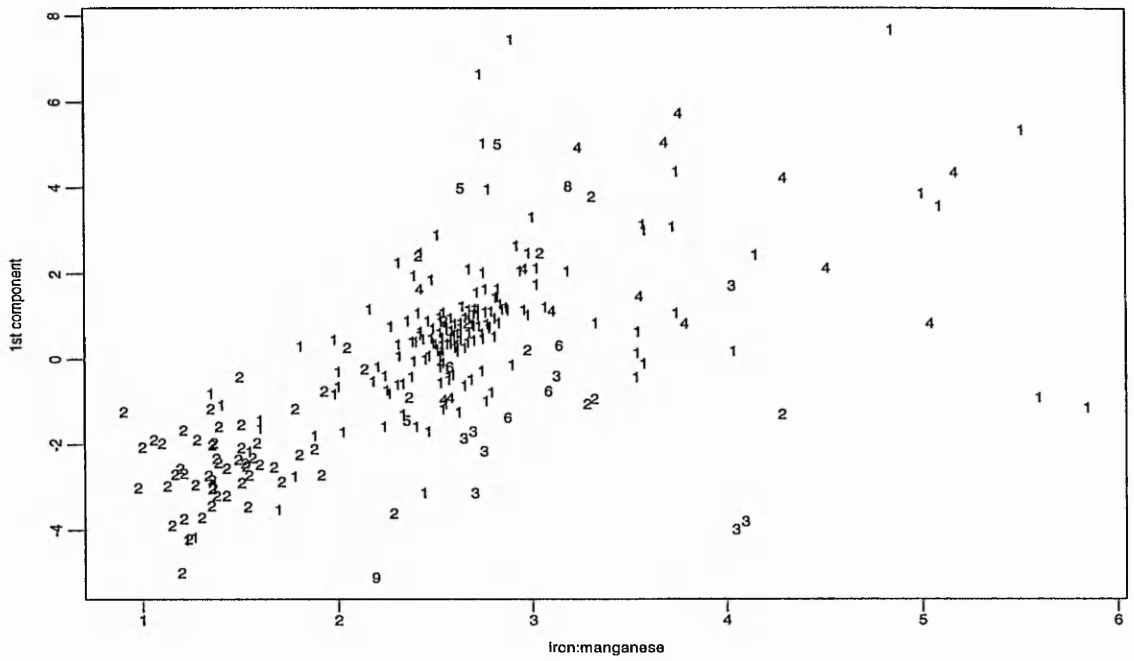
Since the two main colour concentrations appear to be based on the Fe and Mn content of the glass it is appropriate to also look at a plots of the Fe:Mn ratio. The plot of the Fe:Mn ratio against the first principal component of the data, (after the removal of the seven outliers), labelled by colour and shown in Figure 5.6.8, again suggests the two main concentrations as identified in Figure 5.6.2 - Figure 5.6.4. Additional observations, namely 11, 107, 139, 194, have been removed from the following analyses since they all have very high Fe:Mn ratios which obscure the plot.

Table 5.6.5 Colour/chemical descriptions of the additional four outliers with high Fe:Mn ratios

Outlier	Description
11	Brown/yellow -
107	Brown/yellow - thought to be from the same vessel as obs. 11
139	Light blue
194	Light blue

It is interesting to note that the outliers listed in Table 5.6.5 are chemical outliers which are not detected by any other methods.

Figure 5.6.8 Plot of the Fe:Mn ratio against the 1st principal component for all data labelled according to colour - excluding outliers



The univariate KDE plot already seen in Figure 3.11.2, using the adaptive STE method for the selection of h, uses the Fe:Mn ratio of all the data excluding the original seven outliers, and the additional four observations with high Fe:Mn ratios. This plot shows clear bimodality, the peak to the left corresponding to those observations coloured light green and that to the right corresponding to those coloured light blue. The separation of these two colour groups can be seen more clearly in the KDE plot of Figure 5.6.9. This is an overlay plot of the KDE for the Fe:Mn ratio of those observations coloured light blue, (solid line), and a second KDE for the ratio of those coloured light green, (dashed line).

Figure 5.6.9 Two KDE's (superimposed) using the Fe:Mn ratio for light green (dashed line, h = 0.1296) and light blue (solid line, h = 0.9152) - excluding original seven outliers and those observations with large Fe:Mn ratios. Using the adaptive STE method for selection of h

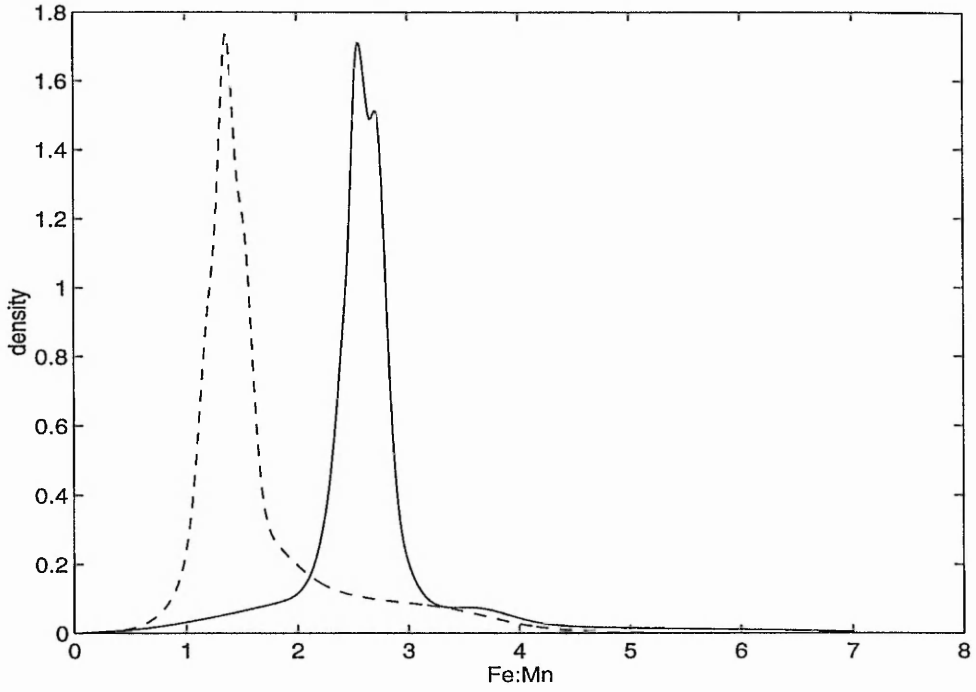


Figure 5.6.8 indicates that the Fe:Mn ratio is related to the first principal component which in turn can be interpreted as showing two main concentrations that are related to colour. Figure 5.6.9 highlights that there is grouping based on the Fe:Mn ratio.

In the case of the Southampton glass the data appear to fall into two groups which are strongly associated with colour. The glasses tinted light green contain, on average, less Fe and more Mn than the light blue glasses. Chemical analyses show that the Fe:Mn ratio for the two colour groups is of the order 2.1:1 for light blue glass and 1.1:1 for light green tinted glass (Heyworth, 1991). Overall, the above analyses do indicate that the lower the Fe:Mn ratio, the greater proportion of Mn to Fe, resulting in light green glasses and the higher the ratio, the greater proportion of Fe to Mn, therefore resulting in lightly tinted blue glass.

After carrying out multivariate analysis, it would seem that the main patterns in the data can be captured with far fewer variables than first thought. For example, a simple plot of Fe against Mn reveals the two concentrations clearly, see Figure 5.6.7. Also the Fe:Mn ratio reveals this bimodality very clearly. This will be discussed further in Chapter 7.

6. Results - application to glass data sets

Four ancient glass data sets, see Chapter 2 for further details, are now analysed in depth using the differing methods described in Chapters 3 and 4. The substantive issues raised after initial analyses of these are also discussed.

6.1 Winchester Vessel glass

The 102 specimens of Winchester vessel glass were initially analysed using univariate plots (box-and-whisker and dotplots). Antimony (Sb) has little or no effect on the analysis as there is little variation, and it is removed from further analyses.

Table 6.1.1 List of outliers suggested using outlier detection methods

Obs	Univariate methods	q^2_j	t^2_j	d^2_j	Hadi	A&M 80%	Ave	Sin	PCA
2	×	×	×	×	×	×	×	×	×
17	×	×	×	×	×	×	×	×	×
30	×	×		×	×	×	×	×	×
35	×	×	×	×	×	×	×	×	×
48	×	×	×	×	×	×	×	×	×
Additional									
8	×								
10	×								
15	×								
32	×								
58	×		×	×					
77	×								

Table 6.1.1 lists those observations suggested to be outliers by the various univariate techniques and multivariate outlier detection methods. Observations 2, 17, 30, 35 and 48 feature as the most recurring outliers by all methods. Univariate analyses pick up additional univariate outliers, but these are not suggested by the multivariate methods, apart from observation 58.

Figure 6.1.1 shows a plot of the first two principal components using standardised data of the remaining 10 variables.

Figure 6.1.1 Plot of the first two principal components using standardised data

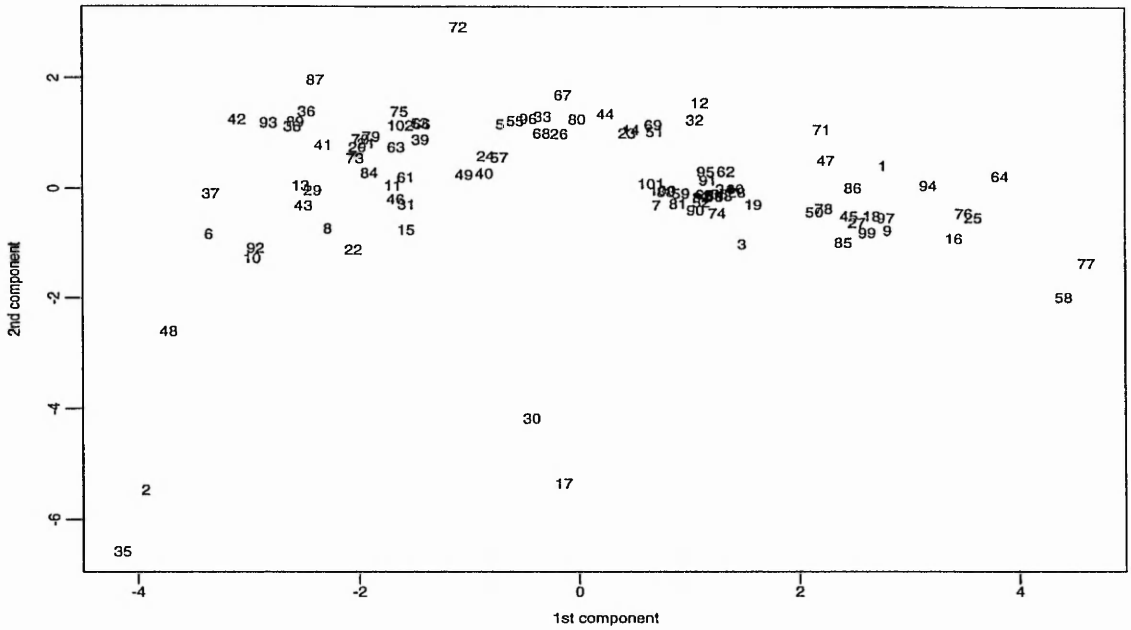


Figure 6.1.1 identifies two extreme outliers observations 2 and 35. Observation 48 is a possible outlier as it lies away from the general scatter of points. Observations 17 and 30 form a small outlying group. Table 6.1.2 lists those observations evidently outlying on univariate plots of the higher order components. The 2nd principal component identifies those observations which are outlying on the principal component plot of Figure 6.1.1.

Table 6.1.2 Table listing outliers detected by the higher order components

Outliers	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
2		×	×					×		
17		×		×			×			
30		×		×						
35		×								
48			×							

Table 6.1.3 indicates those elements which are found to be of high content in the corresponding observations.

Table 6.1.3 Table listing the high content levels of the outliers

Outlier	Element
2	P, Pb
17	Mg, K
30	Mg, K
35	K, P, Pb
48	Pb

Observations 2 and 35 (the most extreme outliers of Figure 6.1.1) both have high contents of P and Pb. Observation 48 also has a high content of Pb and observations 17 and 30, which form a small outlying group in Figure 6.1.1, both have high contents of Mg and K. Looking at Figure 6.1.1 there appears to be a separation of the data into 2 or more clusters. Observations 2, 17, 30, 35 and 48 are now removed as these are the most prominent and recurring outliers in all the analyses undertaken, (Table 6.1.1).

After the removal of the five outliers, the first component accounts for 53% of the variation in the data. Together the first two components account for 69% of the variation and the first three components are needed to 'explain' 83% of the variation.

Table 6.1.4 Correlations of all the data, after the removal of the five outliers

	Al	Fe	Mg	Ca	Na	K	Ti	P	Mn
Fe	0.33								
Mg	0.26	0.88							
Ca	0.26	-0.59	-0.56						
Na	-0.12	0.51	0.56	-0.60					
K	0.08	-0.45	-0.42	0.46	-0.28				
Ti	0.39	0.96	0.80	-0.56	0.46	-0.50			
P	-0.17	-0.37	-0.46	0.26	-0.39	0.47	-0.41		
Mn	0.27	0.91	0.81	-0.54	0.52	-0.53	0.86	-0.52	
Pb	-0.27	-0.24	-0.33	0.04	-0.25	0.25	-0.28	0.89	-0.37

Table 6.1.4 shows the correlations of the elements after the removal of the five outliers. It can be seen that Fe, Mg, Ti and Mn are all highly correlated, suggesting these elements may have entered the batch together via the raw materials. P is also highly correlated with Pb. As with the Southampton data, the Winchester vessel data are not homogenous therefore it is impossible to state why these elements are closely related. By looking at a principal components analysis using standardised data after the removal of the outliers, we are able to see if the clusters are separating on the basis of colour.

Figure 6.1.2 Plot of the first two principal components, after the removal of the above-mentioned outliers, labelled according to colour

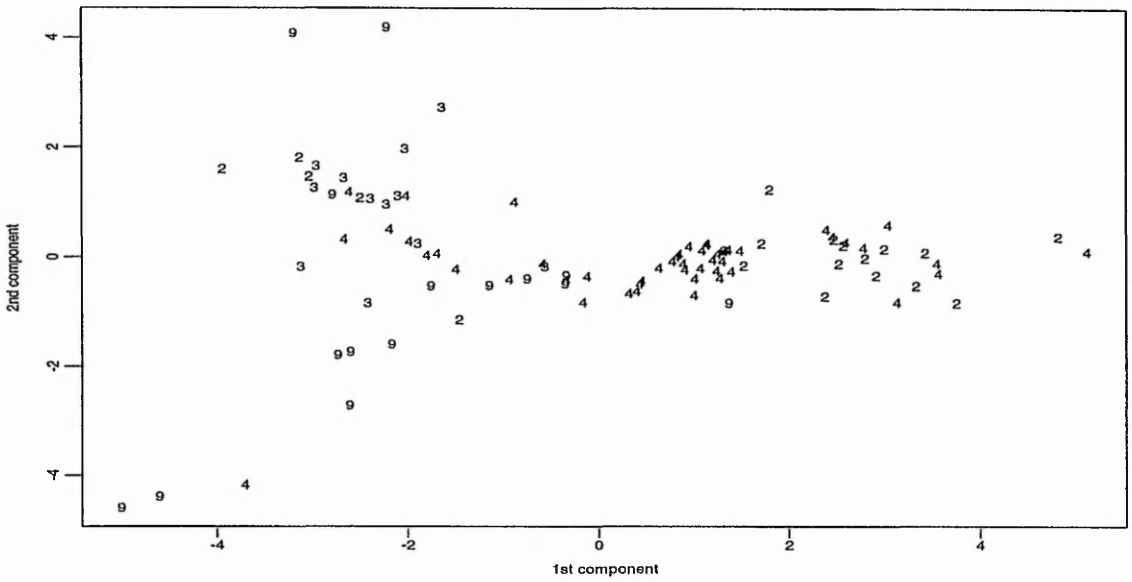
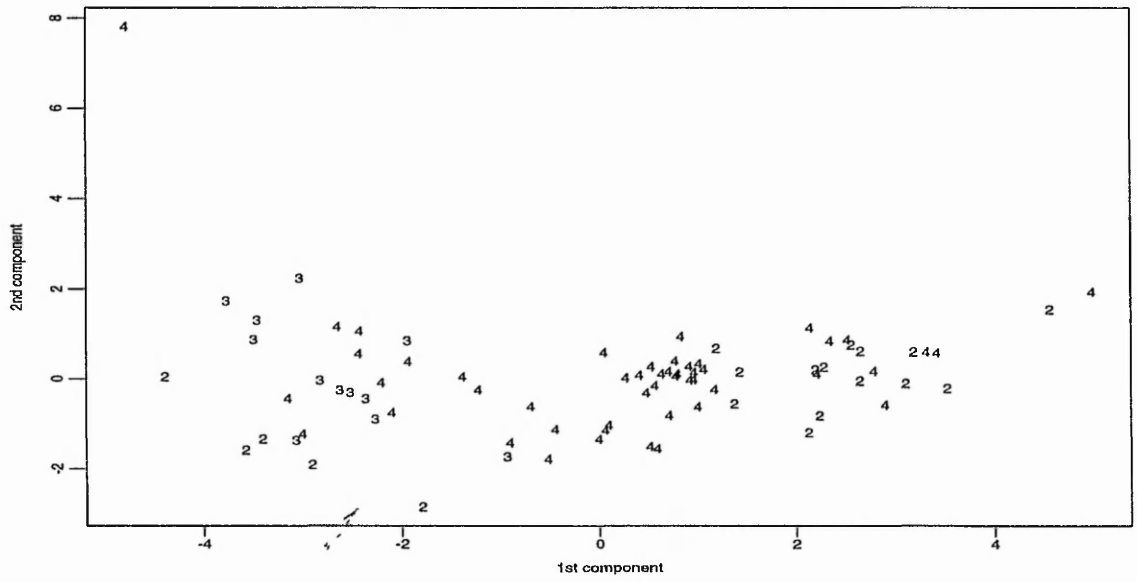


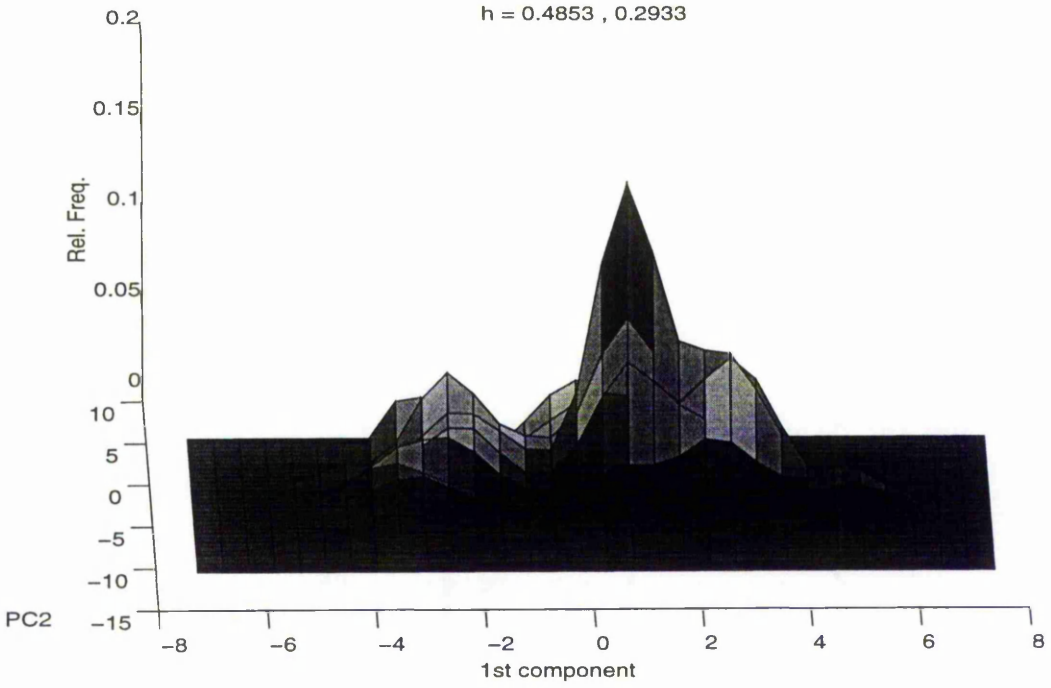
Figure 6.1.2 indicates that the majority of the fragments are light green/green in colour, followed by a small group to the left of the plot that are tinted blue. Carrying out a principal components analysis using standardised data based on the colours light green (2), blue (3) and green (4) only (after the removal of the outliers), the following plot, Figure 6.1.3, is produced.

Figure 6.1.3 Plot of the first two principal components using standardised data - for those specimens coloured light green (2), green (4) and blue (3) only



There does appear to be a separation of the data into three main concentrations related to colour. An outlier, relative to the remaining observations, has now appeared to the top left of the plot. This observation will not be removed from further analyses since we are now looking at colour separation. The blue fragments lie to the left of the plot and those tinted green/light green lie predominantly to the middle and right of the plot. The KDE plot of Figure 6.1.4, based on the specimens coloured light green (2), green (4) and blue (3) only, shows this separation quite clearly.

Figure 6.1.4 Kernel density estimate plot using the specimens coloured light green (2), green (4) and blue (3) only (after removal of outliers)



The separate contour plot, (25, 50, 75%), of Figure 6.1.5, where the light green and green specimens have been analysed together, separately from the blue specimens does indicate that the data form two distinct groups.

Figure 6.1.5 Plot of the separate contours (25, 50, 75%) based on those specimens coloured light green and green together and those coloured blue

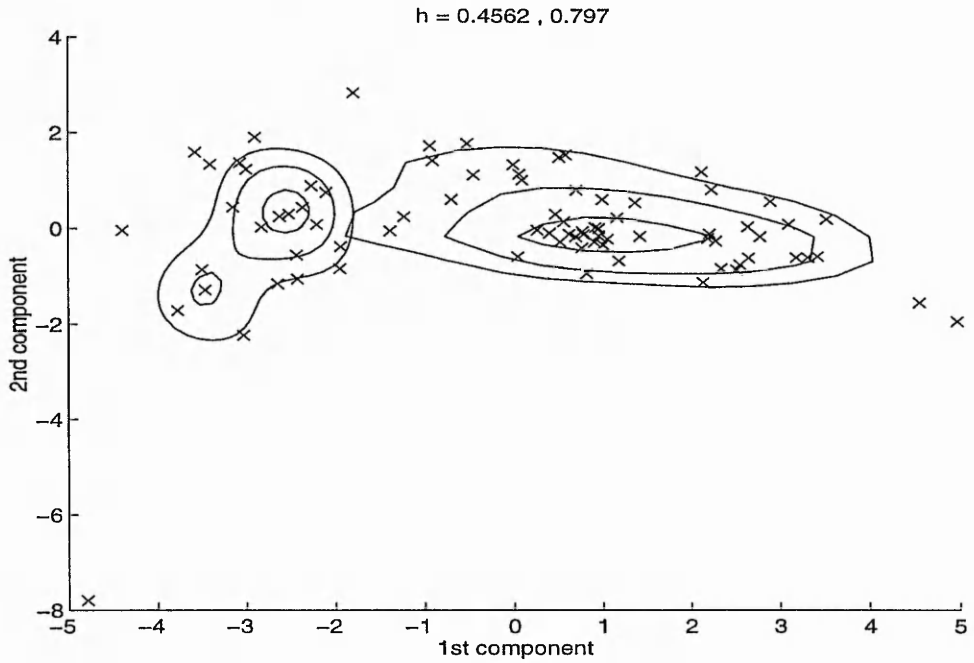
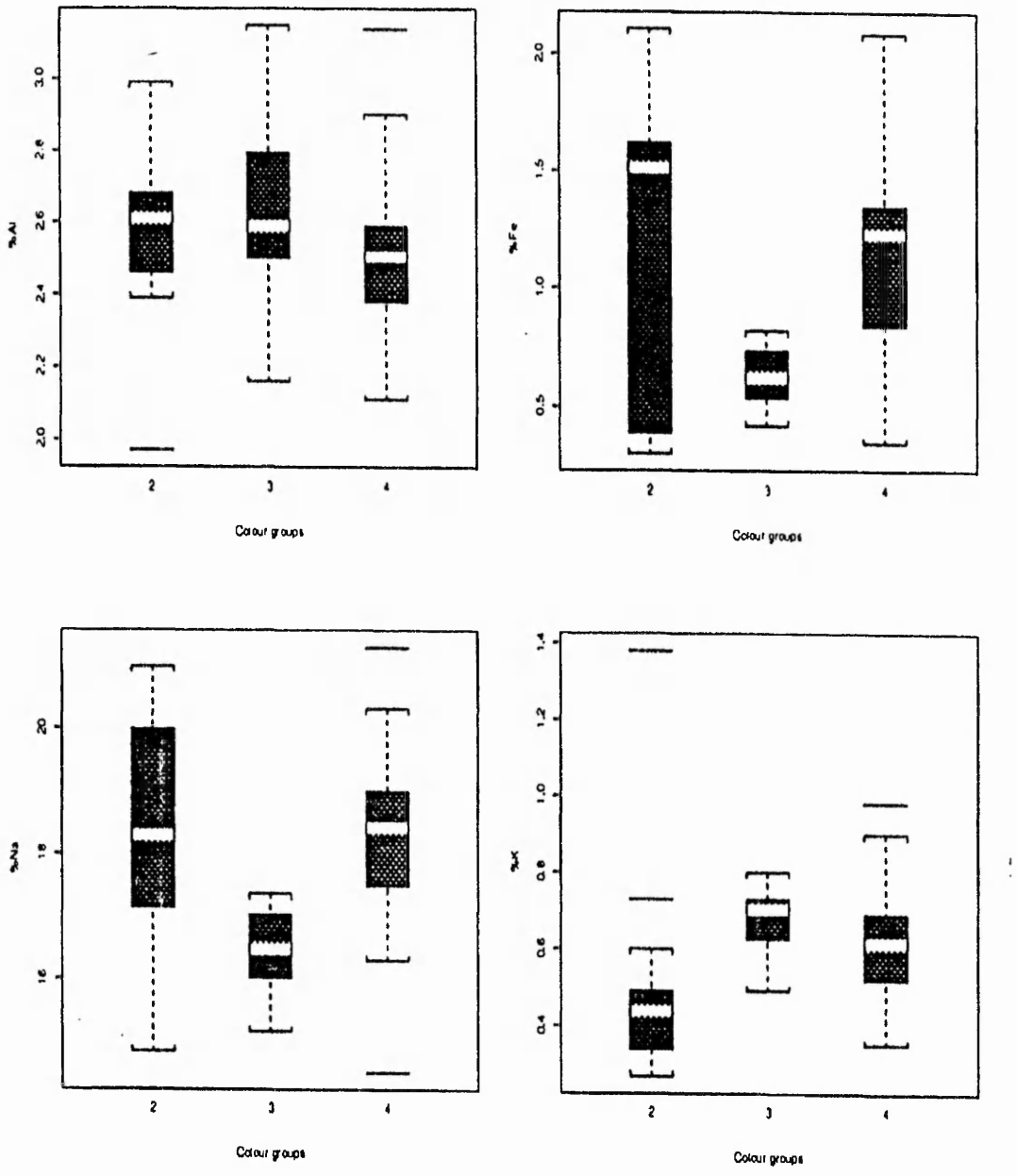
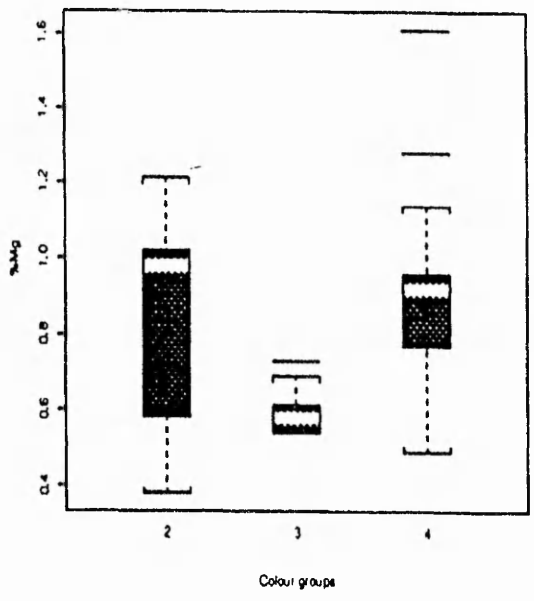


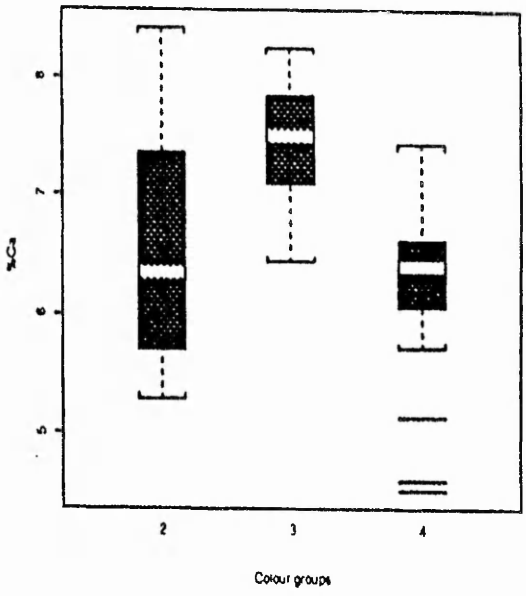
Figure 6.1.4 and Figure 6.1.5 each use the STE method for the selection of h_1 and h_2. The corresponding values of h_1 and h_2 can be seen on each plot.

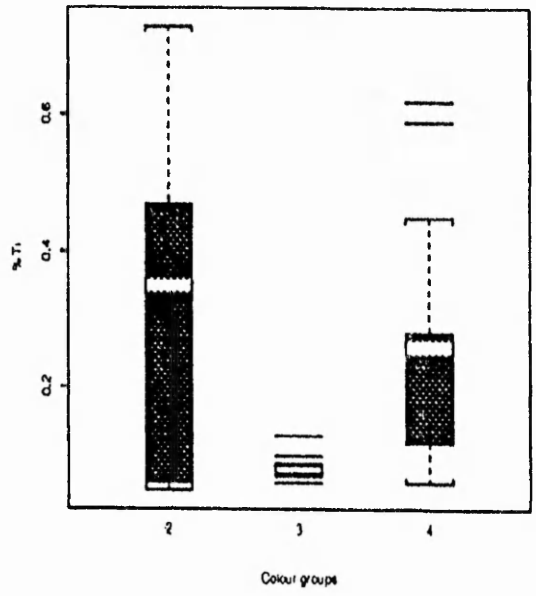
Also observed by the boxplots of Figure 6.1.6 the sub-division of the data is due to the colour of the sherds and not the different sites within Winchester (Heyworth, 1992). The data appear to form two separate groups, the first consisting of the light green and green coloured specimens and the second consisting of the blue specimens. Discrimination between the two groups is based mainly on the Fe content, with Mg, Ti and Mn also at different levels. Note, these are the correlated elements noted earlier in Table 6.1.4. Both colour groups 2 and 4 have a fairly high level of Fe, Mg and Ti and corresponding high Mn level. (Note : the mean ratio of Fe:Mn with respect to colour groups 2 and 4 is approx. 1:1.1, which is consistent with recent published work (Heyworth, 1992), as an equal or excess of Mn produces a light green colour). In colour group 3, although having a lower Fe, Mg and Ti content and a corresponding lower Mn content, the Fe:Mn ratio is actually 2:1, which is what we would expect to observe in blue glass, as an excess of Fe:Mn produces a blue colour in glass.

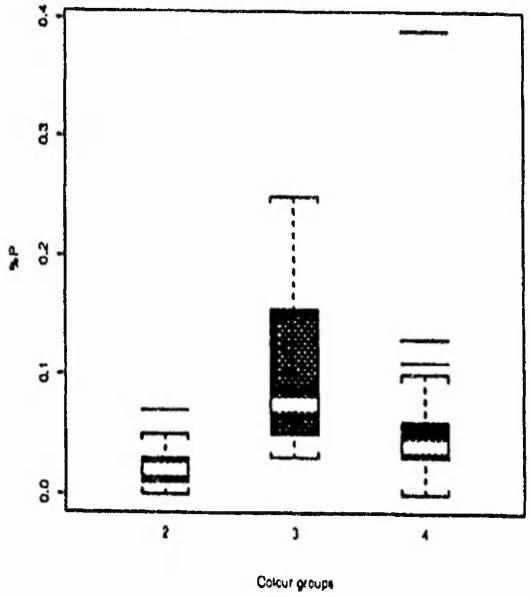
Figure 6.1.6 Boxplots showing the chemical composition of specimens coloured light green (2), blue (3) and green (4)




%Mg
0.4
0.6
0.8
1.0
1.2
1.4
1.6
2
3
4
Colour groups


%Ca
5
6
7
8
2
3
4
Colour groups


%Ti
0.2
0.4
0.6
2
3
4
Colour groups


%P
0.0
0.1
0.2
0.3
0.4
2
3
4
Colour groups

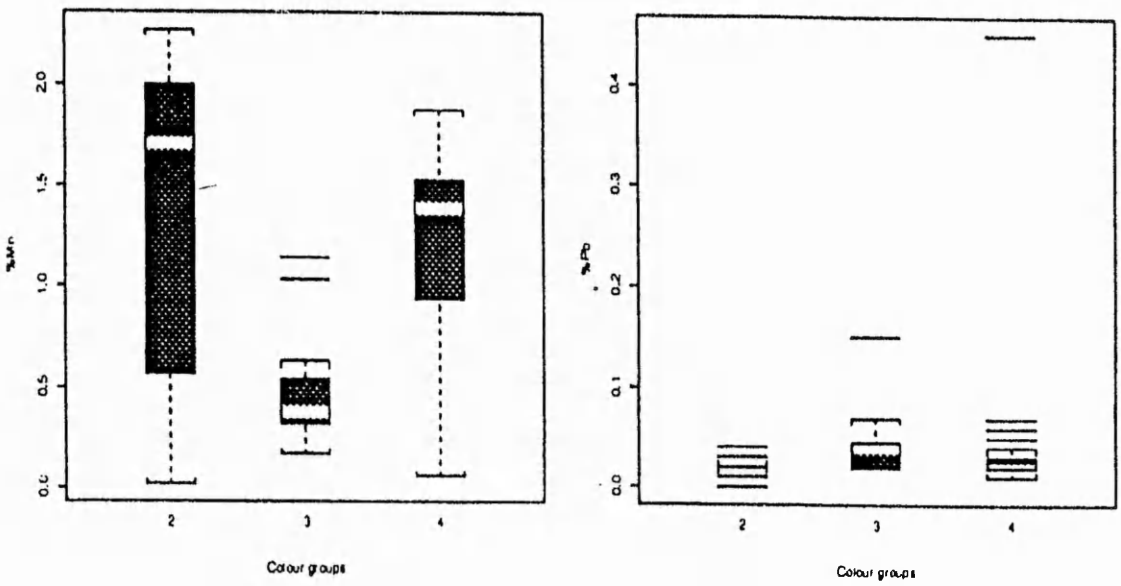

%Mn
0.0
0.5
1.0
1.5
2.0
2
3
4
Colour groups
%Pb
0.0
0.1
0.2
0.3
0.4
2
3
4
Colour groups

Table 6.1.5 shows the correlations between the components and the elements.

Table 6.1.5 Correlations of the elements and the principal components

	pc1	pc2	pc3	pc4	pc5	pc6	pc7	pc8	pc9	pc10
Al	-0.11	0.50	0.55	-0.03	0.39	0.43	0.26	-0.17	-0.05	0.02
Fe	**-0.40**	-0.09	0.28	0.01	-0.14	-0.07	-0.24	0.17	-0.07	**0.80**
Mg	**-0.39**	-0.05	0.16	-0.13	-0.19	-0.36	**0.70**	0.34	0.09	-0.18
Ca	0.29	0.44	0.18	0.05	0.28	-0.72	-0.21	0.14	0.13	0.04
Na	-0.28	-0.23	-0.24	-0.60	0.66	-0.08	-0.09	-0.00	-0.03	0.00
K	0.27	0.09	0.27	-0.77	-0.46	0.05	-0.14	-0.08	0.13	-0.04
Ti	**-0.39**	-0.03	0.27	0.11	-0.03	0.14	**-0.54**	0.36	0.25	**-0.51**
P	0.28	-0.41	0.44	0.01	0.10	-0.08	-0.03	0.19	**-0.69**	-0.16
Mn	**-0.40**	-0.01	0.13	0.05	-0.15	-0.36	-0.12	-0.76	-0.21	-0.20
Pb	0.21	-0.56	0.38	0.13	0.19	-0.05	0.11	-0.25	**0.60**	0.08

As previously discussed Fe, Mg, Ti and Mn are all highly correlated and P is also highly correlated with Pb. These same elements appear to dominate the correlations between the elements and the principal components, the coefficients are shown in bold in Table 6.1.5. The coefficients of Fe, Mg, Ti and Mn are all in excess of 0.35 on the first component. Also these same elements, and P and Pb, dominate the higher order components. Therefore, as the boxplots suggest discrimination between the two colour groups is based mainly on the Fe content, with Mg, Ti and Mn also at different levels, so the correlations show the extent to which these same elements dominate different analyses.

We can conclude that there are two main concentrations in the data and they are strongly associated with colour. These colour groups are also compositionally distinct with respect to a few variables, namely Fe, Mg, Ti and Mn.

6.2 Winchester Window glass

Forty four specimens of window glass excavated from a site at Winchester were analysed. Having looked at univariate plots of the elements and analysed the data using the various multivariate methods, Table 6.2.1 lists those outliers suggested by the different methods.

Table 6.2.1 List of outliers suggested by the various methods

Obs	Univariate methods	q^2_j	t^2_j	d^2_j	Hadi	A&M 80%	Ave	Sin	PCA
5	×		×				×	×	×
10	×	×	×				×		×
11	×	×		×	×	×	×		
24	×						×	×	
37	×	×	×	×	×	×	×	×	×
41	×	×	×	×	×	×	×	×	
42	×		×				×	×	×
43	×	×	×	×	×	×	×	×	
Additional									
4	×								
18	×								
23	×								
32	×								
33	×								
44	×	×	×				×		

The univariate methods pick out additional outliers to those suggested by the multivariate methods, although these are not substantiated by the other methods outlined, with the exception of observation 44. Table 6.2.2 lists those elements which are of a high content in the corresponding outliers, suggesting why some of the observations are outlying especially on the univariate plots.

Table 6.2.2 Table listing the high content levels of the outliers

Outlier	Element
5	Mg, Ca
10	Mg, Ca, K
11	Pb
24	Mn
37	Al, Mg, Ca, Na, K, P
41	Ti, Mn
42	P
43	Al, Mn

The principal components analysis using standardised data produces the following plot.

Figure 6.2.1 Plot of the first two principal components using standardised data

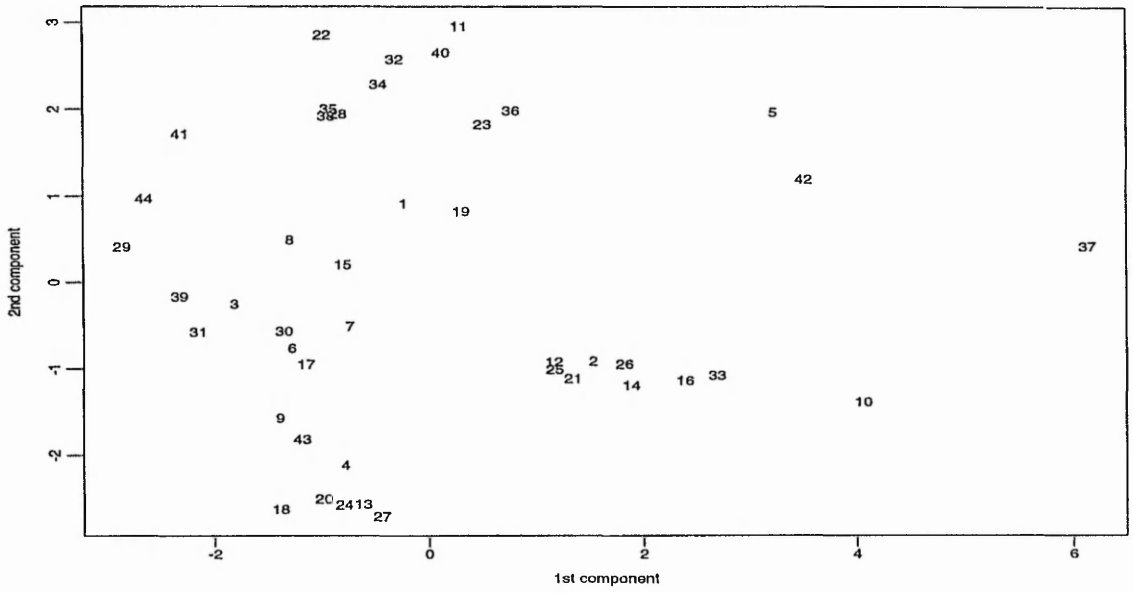


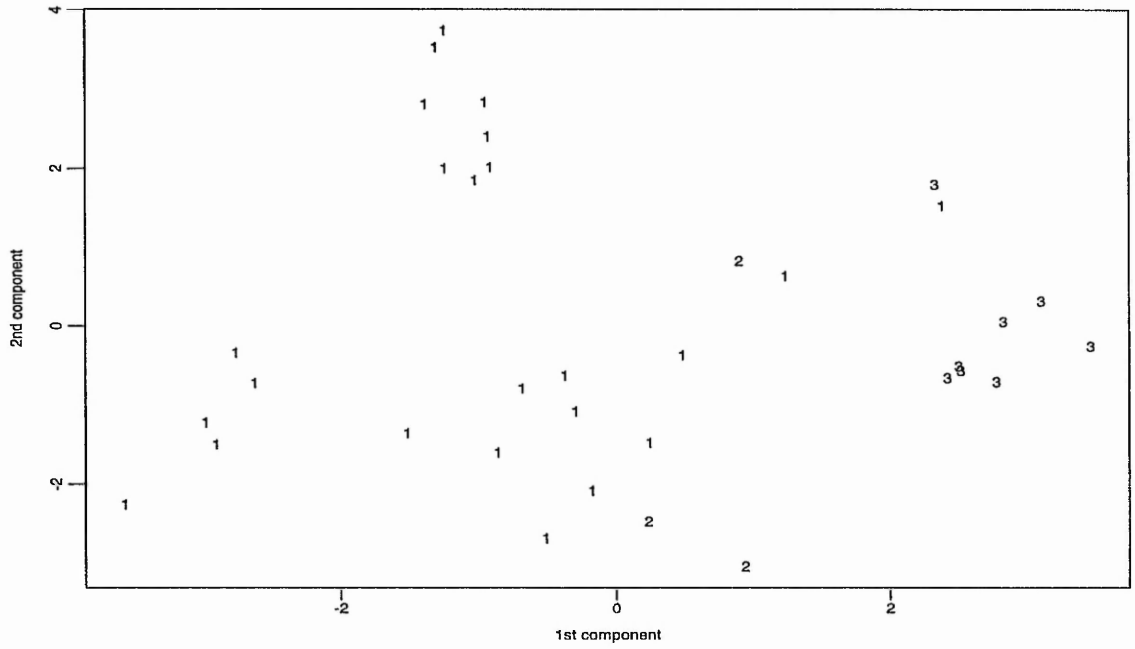
Figure 6.2.1 identifies four extreme outliers, observations 37, 5, 42 and 10, (in order of severity). Table 6.2.3 lists those observations which appear to be outlying on univariate plots of the higher order components. The higher order components appear to identify those outliers which are evident on the index plots of the various outlier detection methods, but are not actually evident on the PCA plot of Figure 6.2.1, namely observations 11, 24, 41 and 43, therefore suggesting they are multivariate outliers

Table 6.2.3 Table listing outliers detected by the higher order components

Outliers	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
5						×				×	
10											
11					×						
24					×						
37	×										×
41			×	×							
42											
43			×		×						

Observations 5, 10, 11, 24, 37, 41, 42, 43 are removed from further analyses. Figure 6.2.2 is a plot of the first two principal components after the removal of the above outliers labelled according to colour, where 1-light blue, 2-light green and 3-blue.

Figure 6.2.2 Plot of the first two principal components, after removal of outliers, labelled according to colour



Now, the first component accounts for just 35% of the variation in the data. The first two components account for 65% and the first three are needed to 'explain' 80% of the variation. This is not perfect, but it is high enough for the component plot to be reasonably informative about structure in the data. A majority of the specimens are light blue in colour, and form three small, distinct clusters. A small grouping of darker blue specimens is also apparent to the right of the PCA plot. Since only three of the remaining specimens, after the removal of the outliers, are coloured light green, the data are re-analysed using only those specimens coloured light blue and blue. The PCA plot of Figure 6.2.3 is based on the light blue and blue specimens only.

Figure 6.2.3 Plot of the first two principal components, after removal of outliers and those specimens coloured light green, labelled according to colour

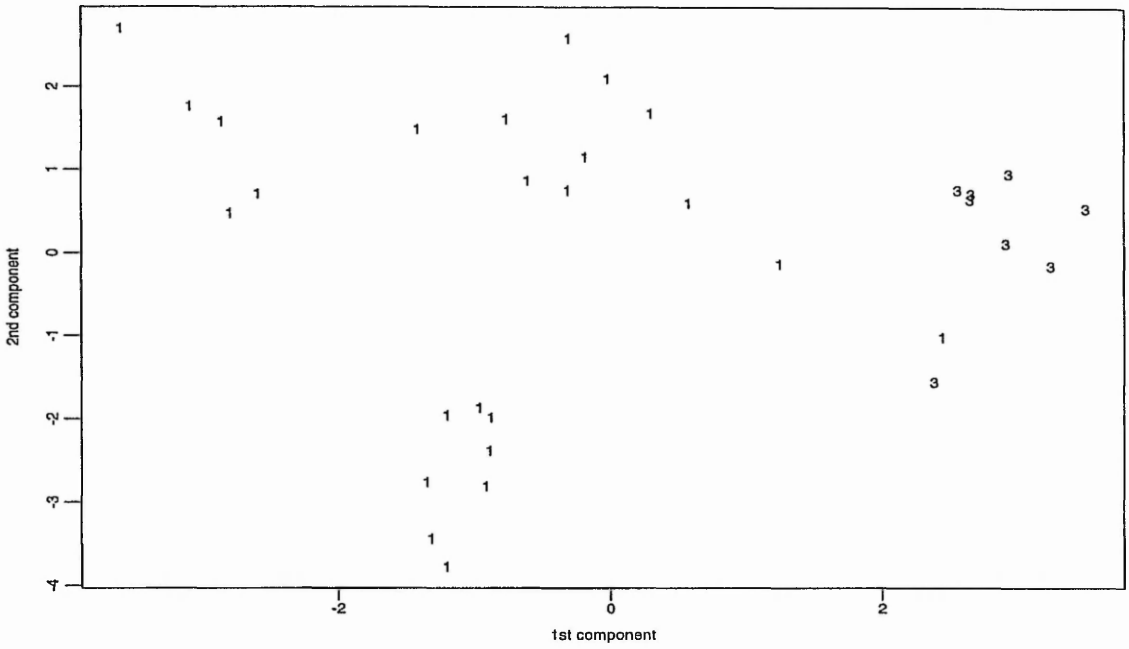


Table 6.2.4 lists the correlations of all the elements, excluding outliers and those specimens coloured light green. It can be seen that Fe is highly positively correlated with Pb, Sb and Ti. Also Mg is highly correlated with K and P.

Table 6.2.4 Correlations of all the data, excluding outliers and those specimens coloured light green

	Al	Fe	Mg	Ca	Na	K	Ti	P	Mn	Pb
Fe	-0.12									
Mg	-0.14	0.10								
Ca	0.37	-0.36	0.48							
Na	-0.45	0.23	-0.46	-0.71						
K	-0.08	-0.11	0.87	0.50	-0.48					
Ti	0.07	0.65	0.27	-0.34	0.09	0.06				
P	-0.13	0.50	0.74	0.20	-0.31	0.51	0.40			
Mn	-0.52	0.38	0.40	0.14	0.04	0.22	0.17	0.51		
Pb	0.06	0.76	-0.02	-0.29	0.10	-0.24	0.55	0.54	0.31	
Sb	-0.47	0.69	-0.07	-0.64	0.50	-0.24	0.30	0.36	0.31	0.62

To define the four groups more clearly, an average link cluster analysis was run on the remaining standardised data, shown in Figure 6.2.6, labelled according to colour. A four cluster breakdown was selected.

Figure 6.2.4 Average link cluster using only those specimens coloured light blue and blue, labelled according to colour

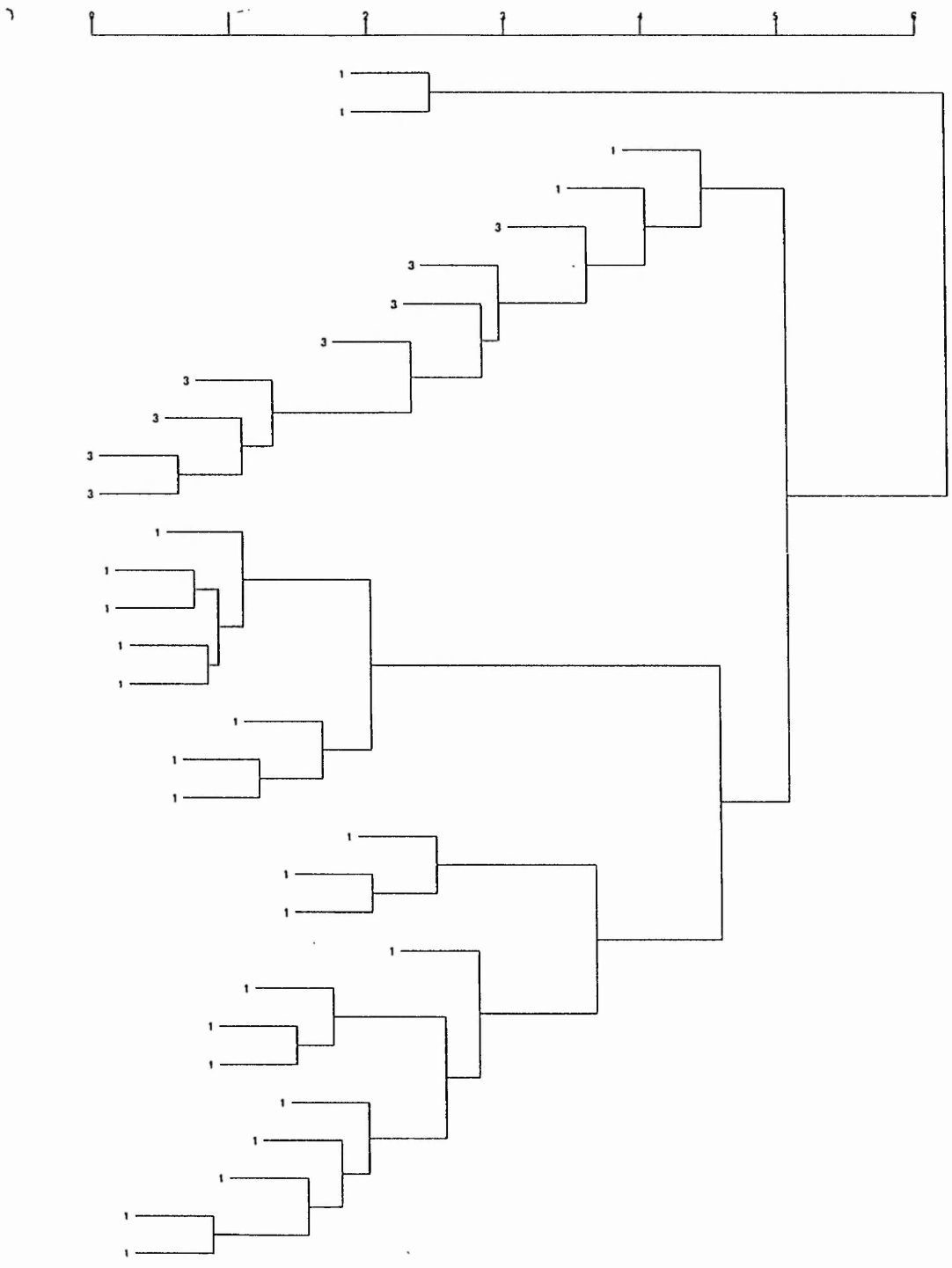


Figure 6.2.5 is a plot of the first two principal components labelled according to the groups defined by cluster analysis. The four groups, first observed in Figure 6.2.3, appear to be compositionally distinct, although two specimens of group 'c' seem to be associated with three 'b' specimens in Figure 6.2.5

Figure 6.2.5 Plot of the first two principal components, after removal of outliers and light green coloured specimens, labelled according to groups a - d

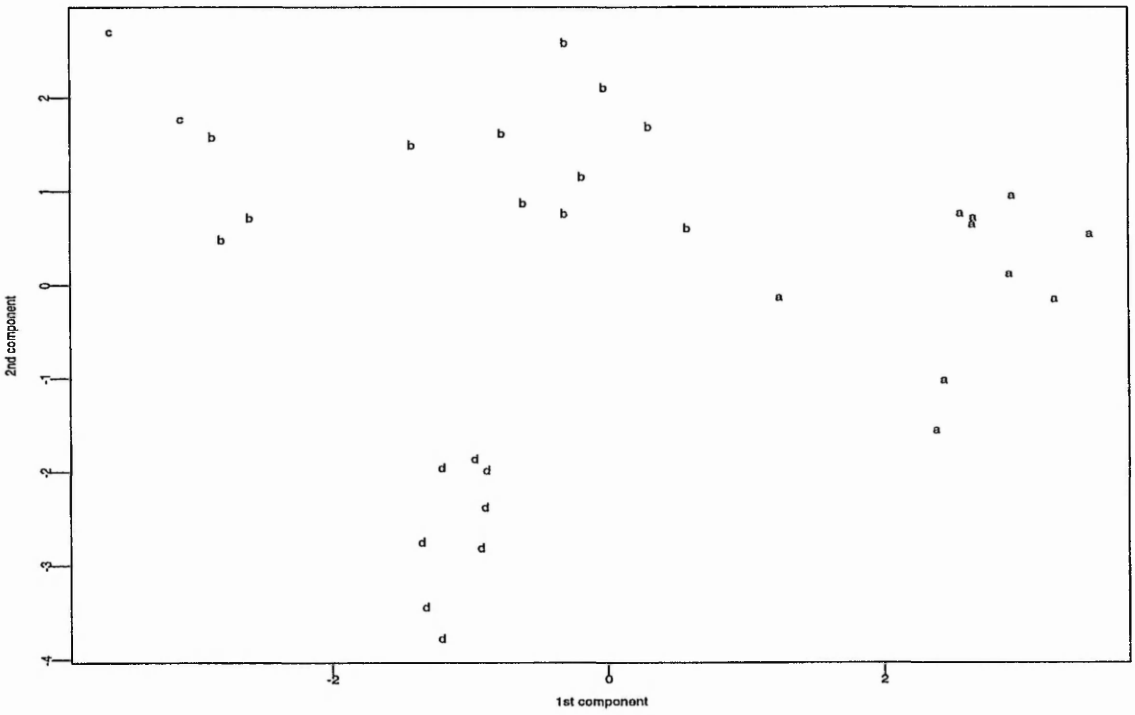
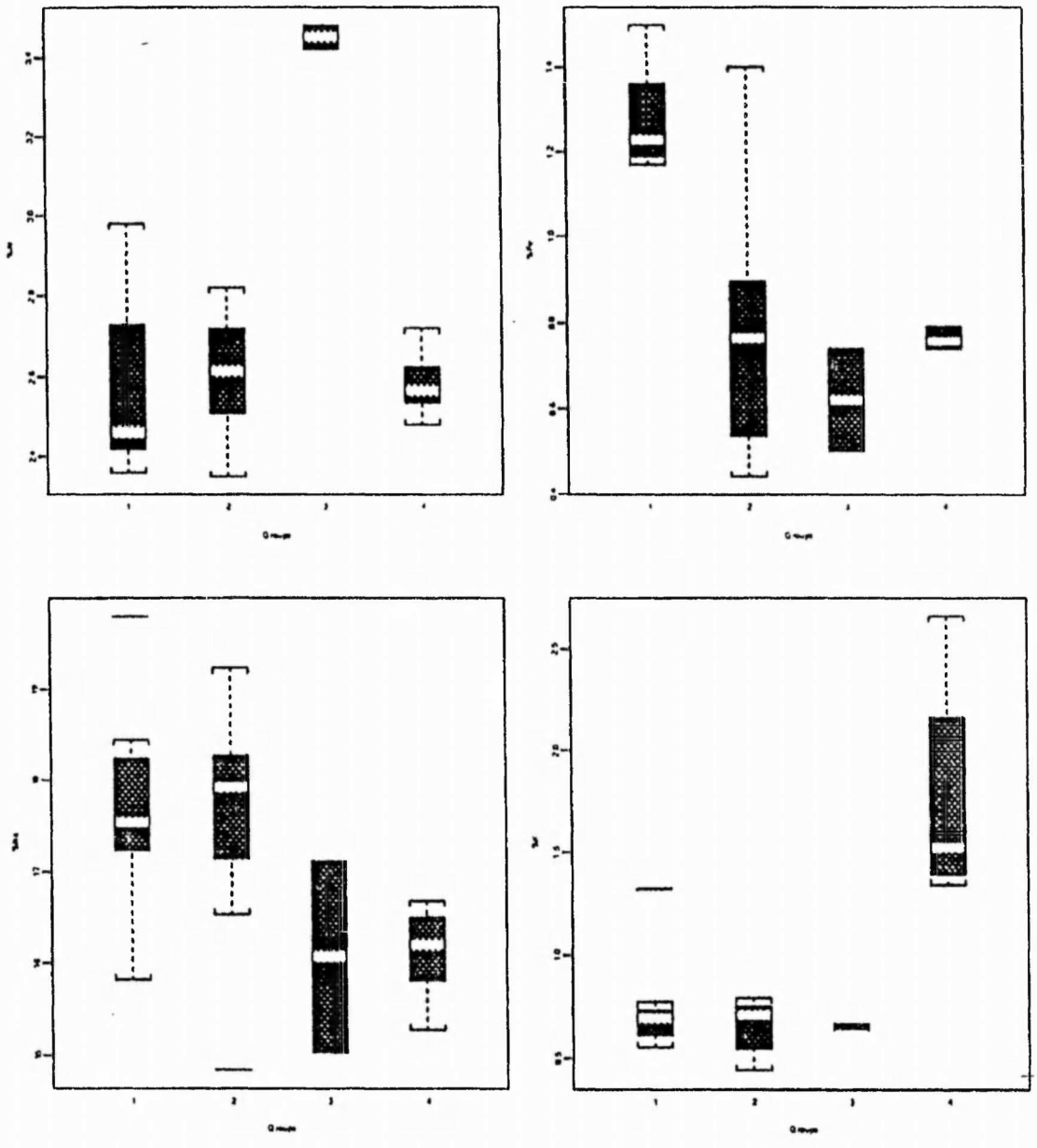
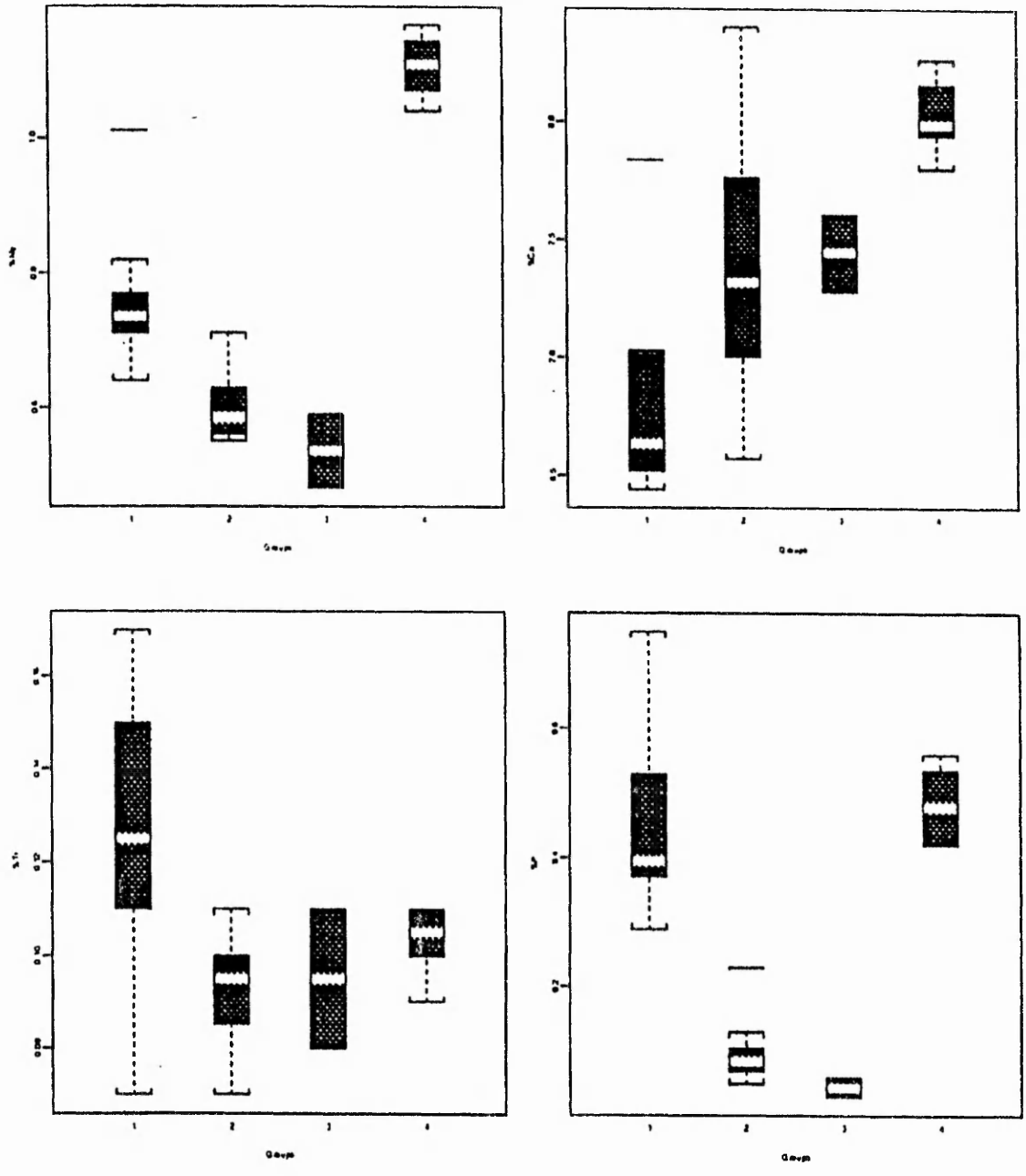
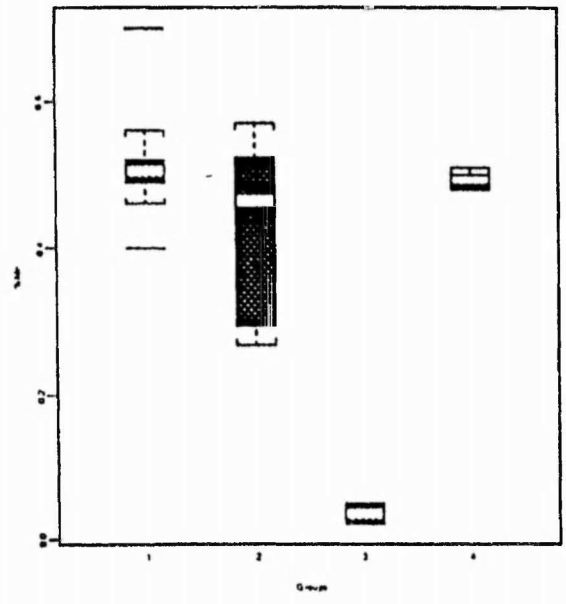


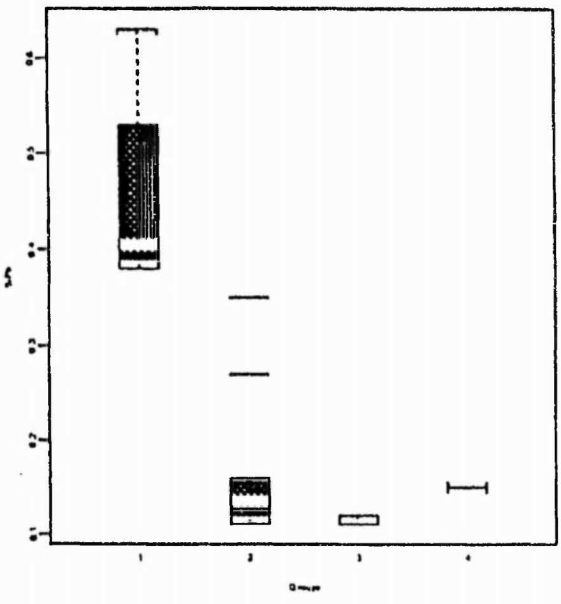
Figure 6.2.6 shows the boxplots for each element and groups a - d. Each group can be discriminated from the rest, with respect to the levels of a few variables. Group a (coloured blue) - high Fe, Ti, P, Pb, Sb; group b (coloured light blue) - high Ca, Na, low K, P; group c (coloured light blue) - high Al, low Mg, K, P, Mn, Pb, Sb; group d (coloured light blue) - high Mg, Ca, K, low Pb. The high level of Sb found in specimens belonging to group a may suggest the use of Sb as a decolorizer. This in turn could also suggest that the glass dates back to an earlier period when Sb was used for decolorising purposes. Group a, as with the other groups, has a Fe:Mn ratio of approx. 2.1:1, but the levels of Fe and Mn are much higher in this case, thus accounting for the darker blue colour.

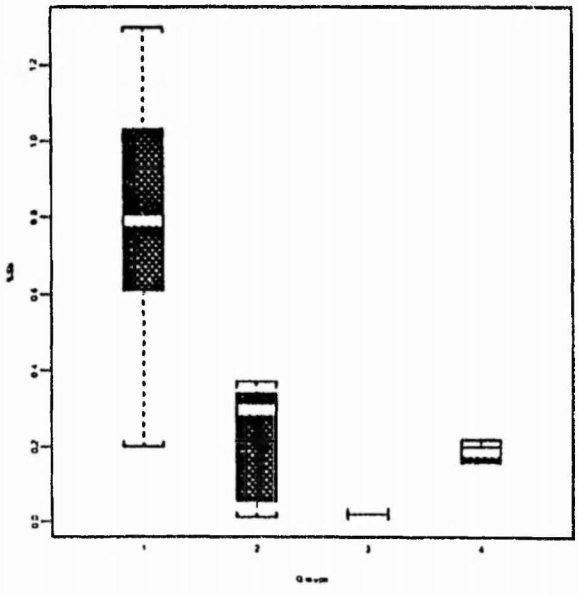
Figure 6.2.6 Boxplots of the chemical composition of the groups a to d, where a = 1, b=2, c=3, d=4











The correlations of the elements and the first three principal components also show the extent to which the same elements dominate different analyses.

Table 6.2.5 Correlations of the elements with the first three principal components

	pc1	pc2	pc3
Al	0.21	-0.01	**0.68**
Fe	**-0.44**	-0.09	0.21
Mg	-0.01	**-0.51**	-0.13
Ca	0.32	-0.34	0.07
Na	-0.25	0.35	-0.28
K	0.11	**-0.45**	-0.19
Ti	-0.32	-0.13	0.32
P	-0.25	**-0.44**	0.05
Mn	-0.24	-0.26	-0.34
Pb	**-0.40**	-0.07	0.35
Sb	**-0.45**	0.07	-0.12

Table 6.2.5 shows that the first principal component correlates with Fe, Pb and Sb, the second with Mg, K and P and the third component with Al, since they all have coefficients in excess of 0.4 .

The fragments appear to fall into four typological groups, as discussed by Heyworth, 1991.

Typological groups	Date
Durable blue glass	> 10th century AD
Non durable glass	> 10th century AD
Durable glass of early type	7 - 9 centuries AD
Durable glass of late type	9 - 11 centuries AD

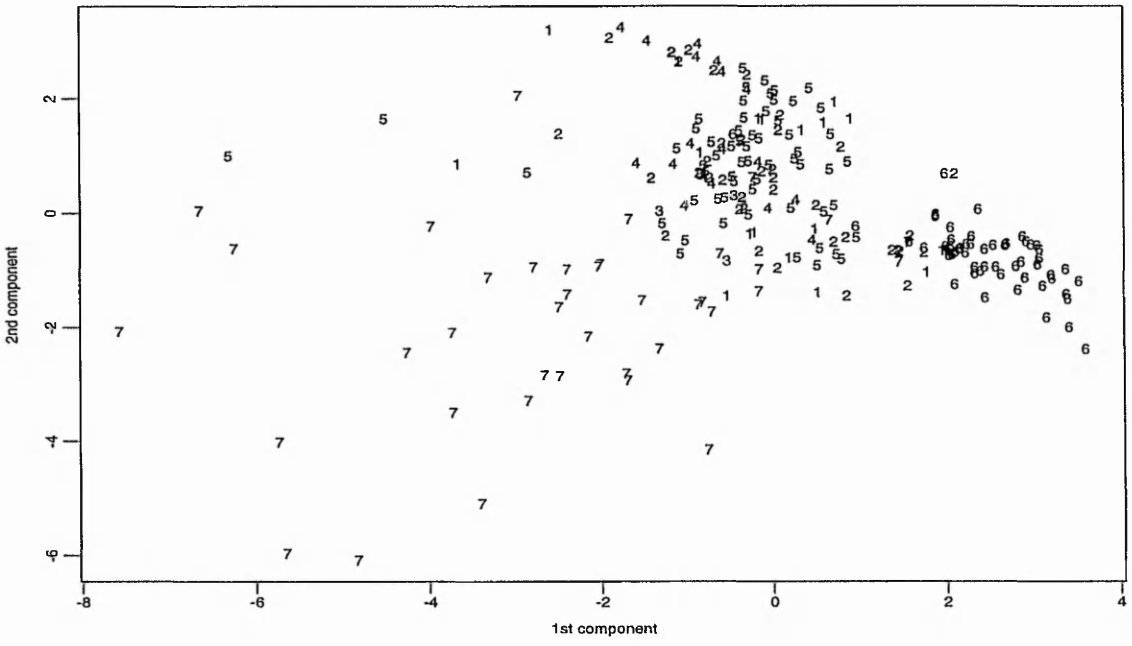
These typological groups can be related to the four compositionally distinct groups identified in Figure 6.2.3 to Figure 6.2.5. The durable blue glass relates to group a. This group is darker than the other three groups and it has a higher content of Fe and Ti suggesting its darker blue colour. The non durable glass relates to group d. This is suggested because this is possibly plant ash glass, see Chapter 2. It is high in Mg and Ca and lower in Na, therefore made with plant alkalis. This type of glass is less durable and prone to weathering, (Jackson, 1992). The remaining two groups b and c relate to the remaining two typological groups, durable glass of 'early' type and durable glass of 'late' type. Group b appears originate from salt water plant alkalis as it is high in Ca and Na and lower in K. This also has a higher level of Mn than group c, suggesting group b could be

of the 'late' type indicating the shift from the use of Sb to Mn. It must be made clear that the above are only suggestions.

6.3 Coppergate glass

Initially 233 specimens were analysed. The following plot of Figure 6.3.1 is of the first two principal components labelled according to colour. 1 - yellow/green, 2 - colourless with light green tinge, 3 - blue-green, 4 - blue-green with tendency towards green, 5 - blue-green with tendency towards blue, 6 - colourless, 7 - crucible waste glass from glass melting pots, coloured mainly green/light green. Looking at Figure 6.3.1 the crucible waste glass is outlying from the rest of the data. A further two groups also appear on the plot, made up of those observations labelled 1 - 5 (mainly 5's), the blue-green glass, and those labelled 6, the colourless glass, indicating separation of data on the basis of colour.

Figure 6.3.1 Plot of the 1st two principal components labelled according to colour, where 1 - yellow/green, 2 - colourless with light green tinge, 3 - blue-green, 4 - blue-green with tendency towards green, 5 - blue-green with tendency towards blue, 6 - colourless, 7 - crucible waste glass from glass melting pots



The 1st principal component accounts for just 37% of the variation in the data and a single variable does not appear to dominate the analysis. Together the first two components account for 60% of the data. The first five components are needed to 'explain' 85% of the variation in the data.

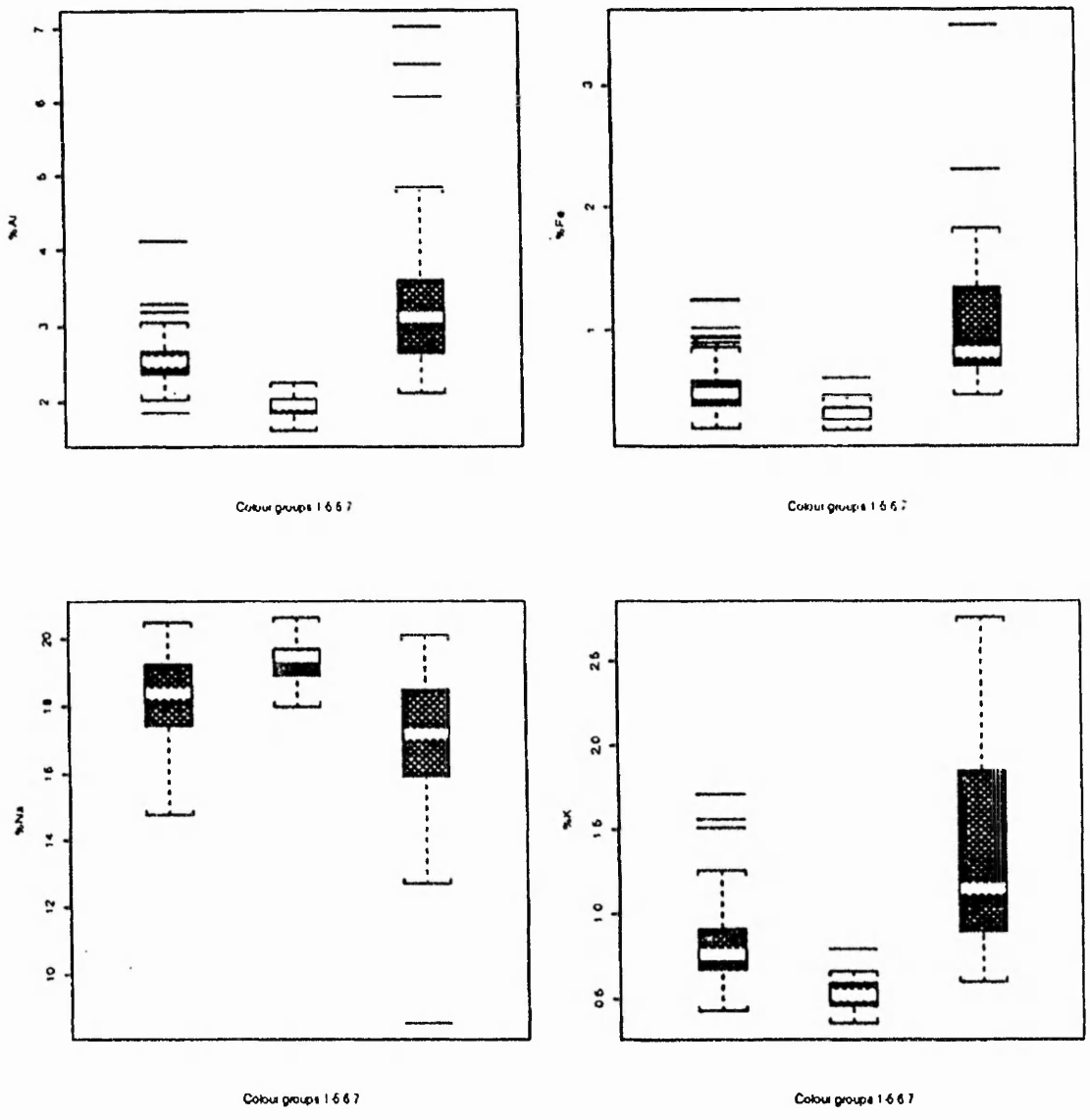
Table 6.3.1 Correlations of the elements for all the data

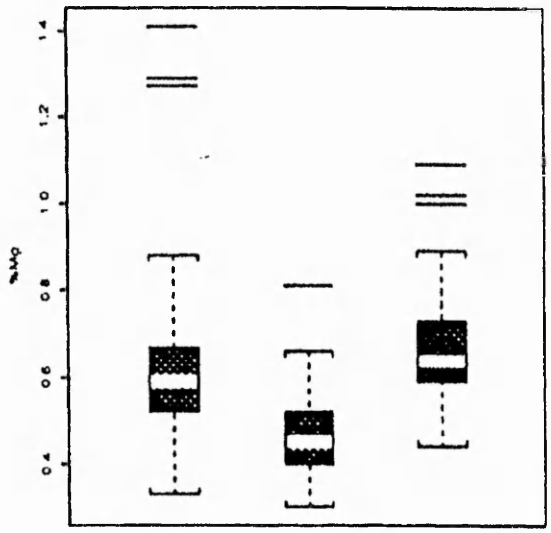
	Al	Fe	Mg	Ca	Na	K	Ti	P	Mn	P
Fe	**0.63**									
Mg	0.25	0.45								
Ca	-0.03	-0.19	0.28							
Na	-0.38	-0.35	-0.25	-0.23						
K	0.44	**0.53**	**0.57**	0.04	-0.51					
Ti	**0.76**	**0.71**	**0.59**	-0.23	-0.27	**0.56**				
P	0.35	0.31	**0.57**	0.46	-0.51	**0.56**	0.32			
Mn	0.16	0.04	0.40	0.49	-0.15	0.04	0.13	0.29		
Pb	0.16	0.21	0.01	-0.19	-0.08	0.15	0.14	-0.01	-0.05	
Sb	-0.25	0.02	-0.22	-0.58	0.35	-0.07	-0.04	-0.48	-0.59	0.25

The correlations of the elements, shown in Table 6.3.1, are for all the data. It can be seen that Al and Fe are both highly positively correlated with Ti. Also Al and Fe have a strong correlation. This is as we would expect since Fe and Al enter the batch together via the silica and it is known that Ti can enter with the Fe. As with the Winchester data sets, Mg, K and P are also all highly correlated. Due to this high inter-correlation among elements, a single element does not dominate the analyses.

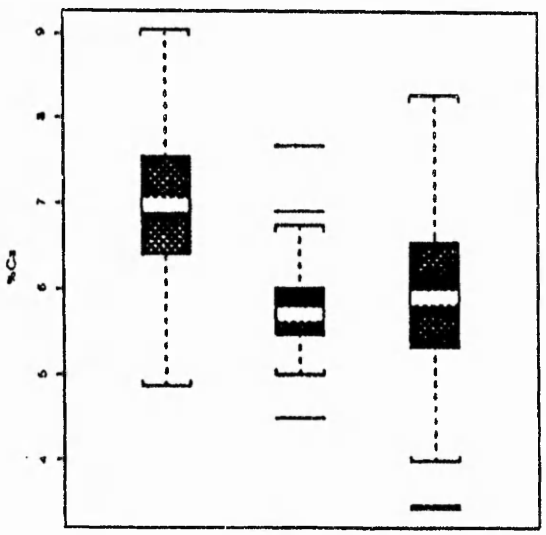
The boxplots of Figure 6.3.2 show the chemical compositions of the three colour groups that appear to be separating in Figure 6.3.1. The first box of each plot shows the chemical composition of all those observations coloured 1 - 5, mainly blue-green glass, the second box shows the chemical composition of all the colourless glass (6) and the third box the chemical composition of the crucible waste glass (7), mainly green/light green in colour. These boxplots establish that the colour groups are compositionally distinct, thus verifying what Figure 6.3.1 shows, that the data appear to be separating on the basis of colour. The glass coloured 1 - 5 has a fairly high content of Mg, Ca and Na and lower content of K suggesting a possible alkali source of salt water plants. This colour group also has a very high content of Mn, and much lower content of Sb thus indicating that Mn was added to the melt to act as a decolorizing agent. The colourless glass has a very high content of Na and fairly high Ca content and low K content therefore we can assume that, as with the blue-green glass of the first colour group, the alkali source could possibly be salt water plants. This colour group also has a very low content of Al, Fe, P and Mn but a very high content of Sb, suggesting that Sb was used to decolorize the colourless glass found at Coppergate. The crucible waste glass is very different from the rest of the glass in the assemblage due to the fact that this group is thought to date to a different period (Jackson 1992). This group has a very high content of Al, Fe, K and Ti and a lower content of Na and Mn. In order to obtain a clearer view of the data it is necessary to remove this 'waste' group and perform analyses on the remaining glass.

Figure 6.3.2 Boxplots showing chemical composition : first boxplot colours 1-5 together, second boxplot colour 6 only and the third boxplot colour 7 only. Where 1 - yellow/green, 2 - colourless with light green tinge, 3 - blue-green, 4 - blue-green with tendency towards green, 5 - blue-green with tendency towards blue, 6 - colourless, 7 - crucible waste glass from glass melting pots

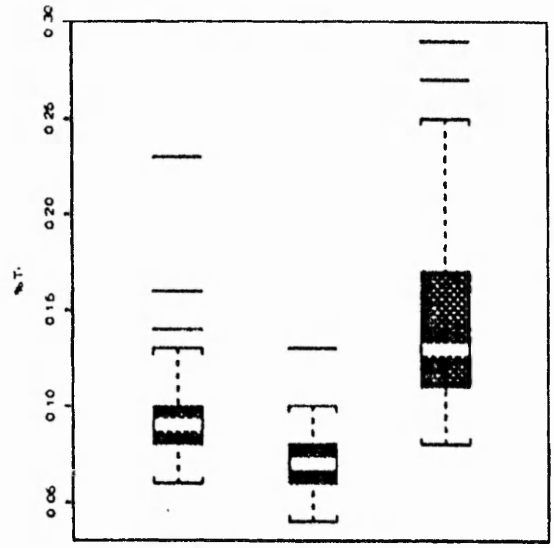




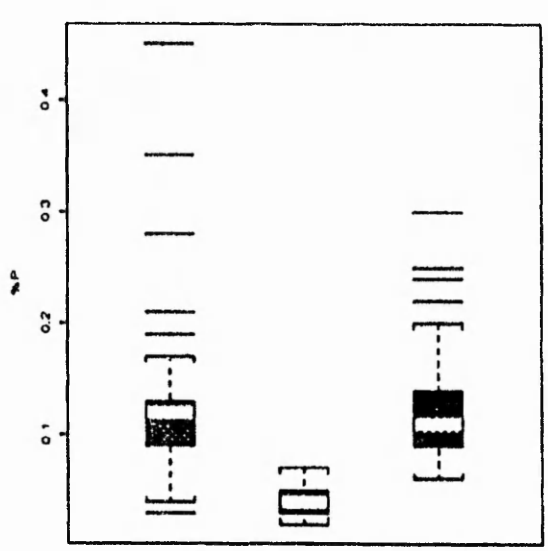
Colour groups 1 6 6 7



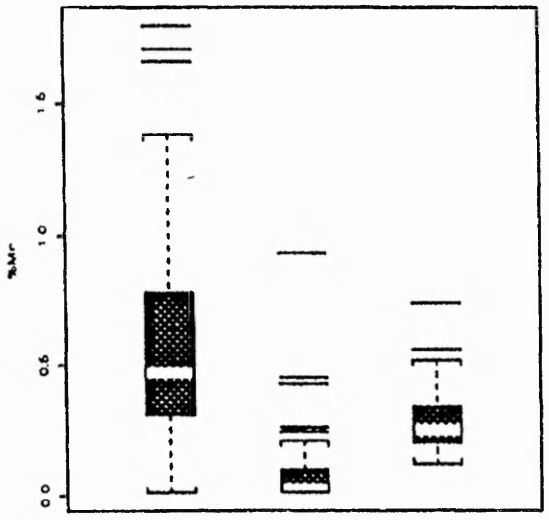
Colour groups 1 6 6 7

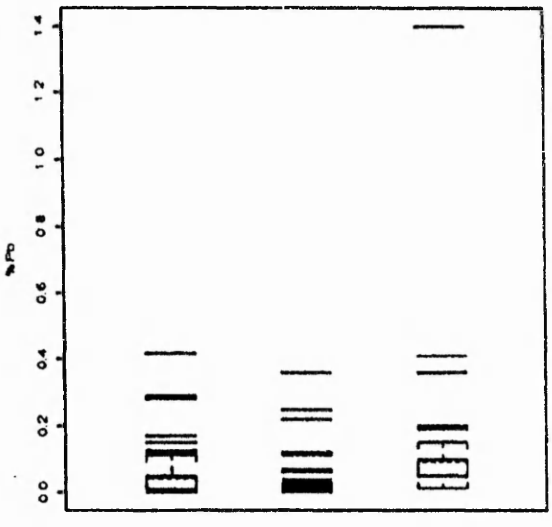


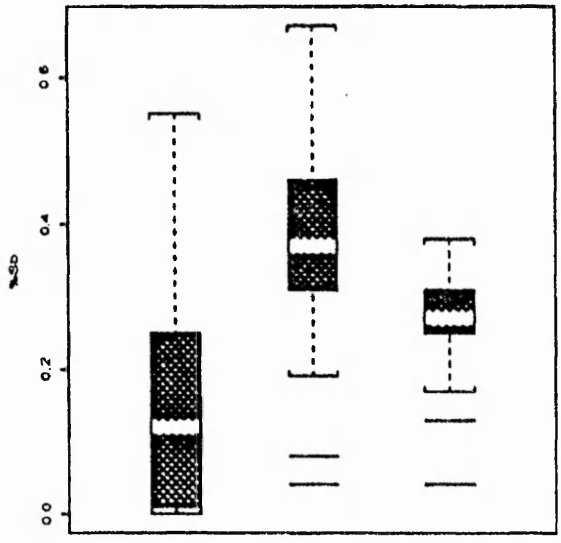
Colour groups 1 6 6 7



Colour groups 1 6 6 7


%Mn
0.0
0.5
1.0
1.5
Colour groups 1 6 6 7


%Pb
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Colour groups 1 6 6 7


%Sb
0.0
0.2
0.4
0.6
Colour groups 1 6 6 7

The correlations of the elements in the colourless glass can be seen in Table 6.3.2. Ti is highly correlated with Mg, Fe and Ca. This is what we would expect since Ti can enter the batch with Fe and Mg and usually Al as a complex via the silica. Mg is relatively highly correlated with Fe and Ca. One explanation is that Mg and Ca can enter the batch together via the alkali source, in this case possibly a salt water alkali. Sb is not significantly correlated with any other element thus suggesting that Sb was added to the colourless glass batch separately as a relatively pure decolorizer (Jackson 1992).

Table 6.3.2 Correlations of the colourless glass (6)

	Al	Fe	Mg	Ca	Na	K	Ti	P	Mn	P
Fe	0.54									
Mg	0.55	**0.67**								
Ca	0.49	0.58	**0.69**							
Na	0.19	0.25	0.26	0.36						
K	0.47	0.26	0.09	0.33	0.18					
Ti	0.50	**0.66**	**0.85**	**0.66**	0.19	0.10				
P	0.42	0.59	0.30	0.37	0.09	0.68	0.41			
Mn	0.45	0.56	0.42	0.47	0.10	0.45	0.52	0.61		
Pb	0.18	-0.34	-0.27	-0.05	0.02	0.19	-0.30	-0.23	-0.11	
Sb	-0.29	-0.53	-0.32	-0.25	0.08	0.10	-0.32	-0.28	-0.39	0.47

The correlations in

Table 6.3.3 are of the elements of the remaining glass coloured 1 - 5. Ti is highly positively correlated with Mg and Fe, suggesting these elements entered as a complex. No other elements appear to have strong correlation.

Table 6.3.3 Correlations of the glass coloured 1 - 5, where 1 - yellow/green, 2 - colourless with light green tinge, 3 - blue-green, 4 - blue-green with tendency towards green, 5 - blue-green with tendency towards blue

	Al	Fe	Mg	Ca	Na	K	Ti	P	Mn	P
Fe	-0.16									
Mg	-0.09	0.57								
Ca	0.55	-0.23	0.04							
Na	-0.35	0.16	0.18	-0.52						
K	0.07	0.33	0.41	-0.11	0.22					
Ti	-0.10	**0.62**	**0.79**	-0.25	0.34	0.41				
P	0.16	0.18	0.33	0.28	-0.32	0.41	0.07			
Mn	0.23	0.03	0.32	0.22	-0.14	-0.10	0.25	-0.06		
Pb	-0.15	0.17	0.03	-0.26	0.22	0.21	0.16	-0.09	-0.08	
Sb	-0.33	0.39	0.11	-0.43	0.47	0.19	0.25	-0.21	-0.38	0.49

Figure 6.3.3 is a plot of the first two principal components of the remaining glass, coloured 1 - 5, after the removal of the waste and the colourless glass. In this case the first principal component accounts for just 30% of the variation in the data, the first two account for 52% and the first five are needed to 'explain' 80% of the variation. Also the principal components do not appear to be dominated by an individual element. It is worth noting

that some outlying observations can be seen to the bottom right of the plot, namely 20, 169 and 173.

Figure 6.3.3 Plot of the 1st two principal components using glass coloured 1-5, labelled according to specimen number

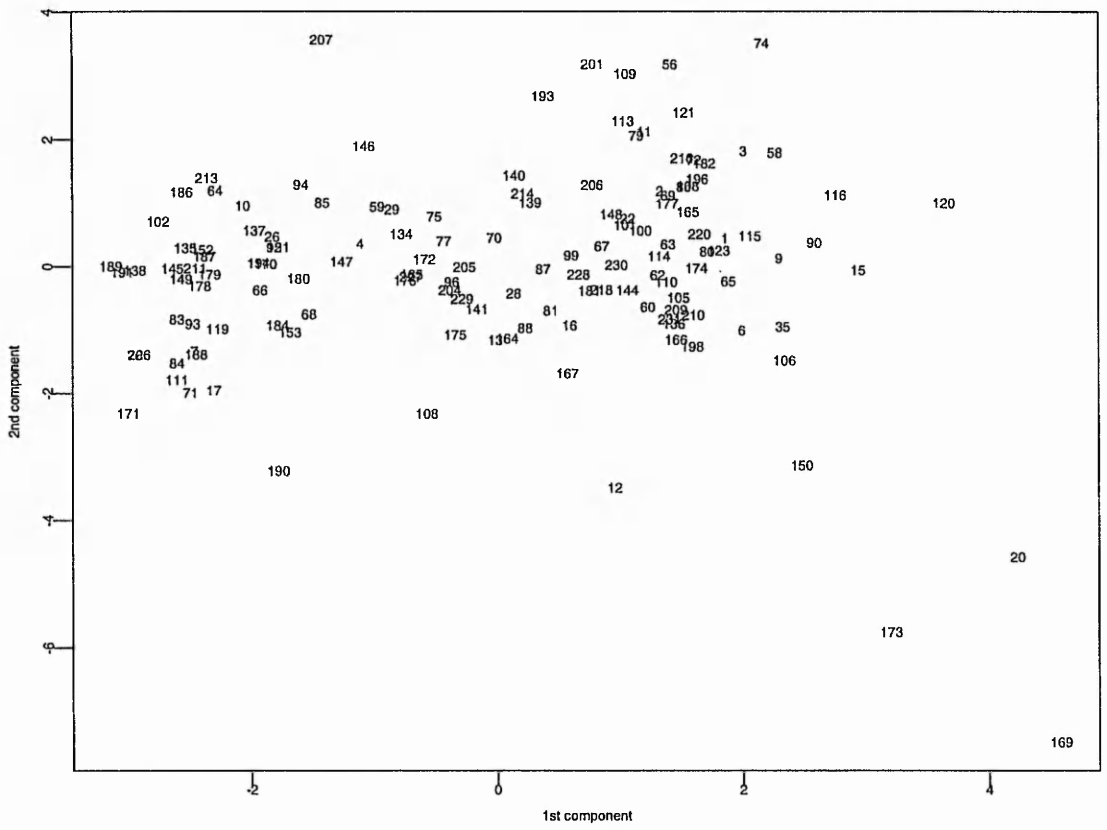


Table 6.3.4 Table showing outliers suggested by the various univariate and multivariate methods

Obs	Univariate methods	q^2_j	t^2_j	d^2_j	Hadi	A&M 80%	Ave	Sin	PCA
20	×	×	×	×	×	×	×	×	×
120	×	×	×	×	×	×	×		×
153	×	×		×	×	×	×	×	
169	×	×	×	×	×	×	×	×	×
173	×	×	×	×	×	×	×	×	×
Additional									
12									×
68	×							×	
108	×						×		
150	×						×		×

Table 6.3.4 lists those outliers identified using both univariate techniques and multivariate outlier detection methods, including the above principal components analysis. Observations 20, 120, 153, 169 and 173 have been identified by all the methods undertaken and are now removed from the analysis. Additional observations are also suggested by univariate methods but are not substantiated by the further analyses. A principal components analysis, with the removal of outliers, produces the following plot.

Figure 6.3.4 Plot of the 1st two principal components using glass coloured 1-5, labelled according to colour, where 1 - yellow/green, 2 - colourless with light green tinge, 3 - blue-green, 4 - blue-green with tendency towards green, 5 - blue-green with tendency towards blue, after the removal of observations 20, 120, 153, 169 and 173

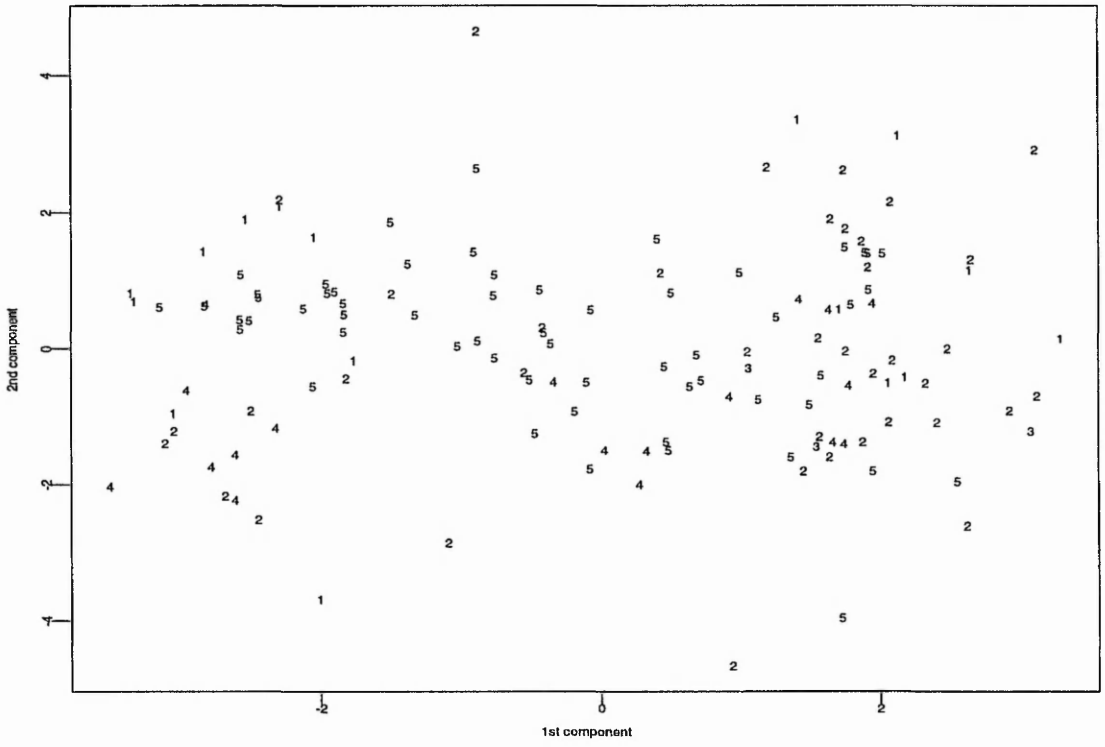


Figure 6.3.4 indicates that the data is separating into three possible concentrations, with obvious overlap, the majority of those observations which are colourless with a light green tinge (2) lie to the right of the plot and those coloured blue-green with tendency towards blue (5) lie to the middle and left of the plot.

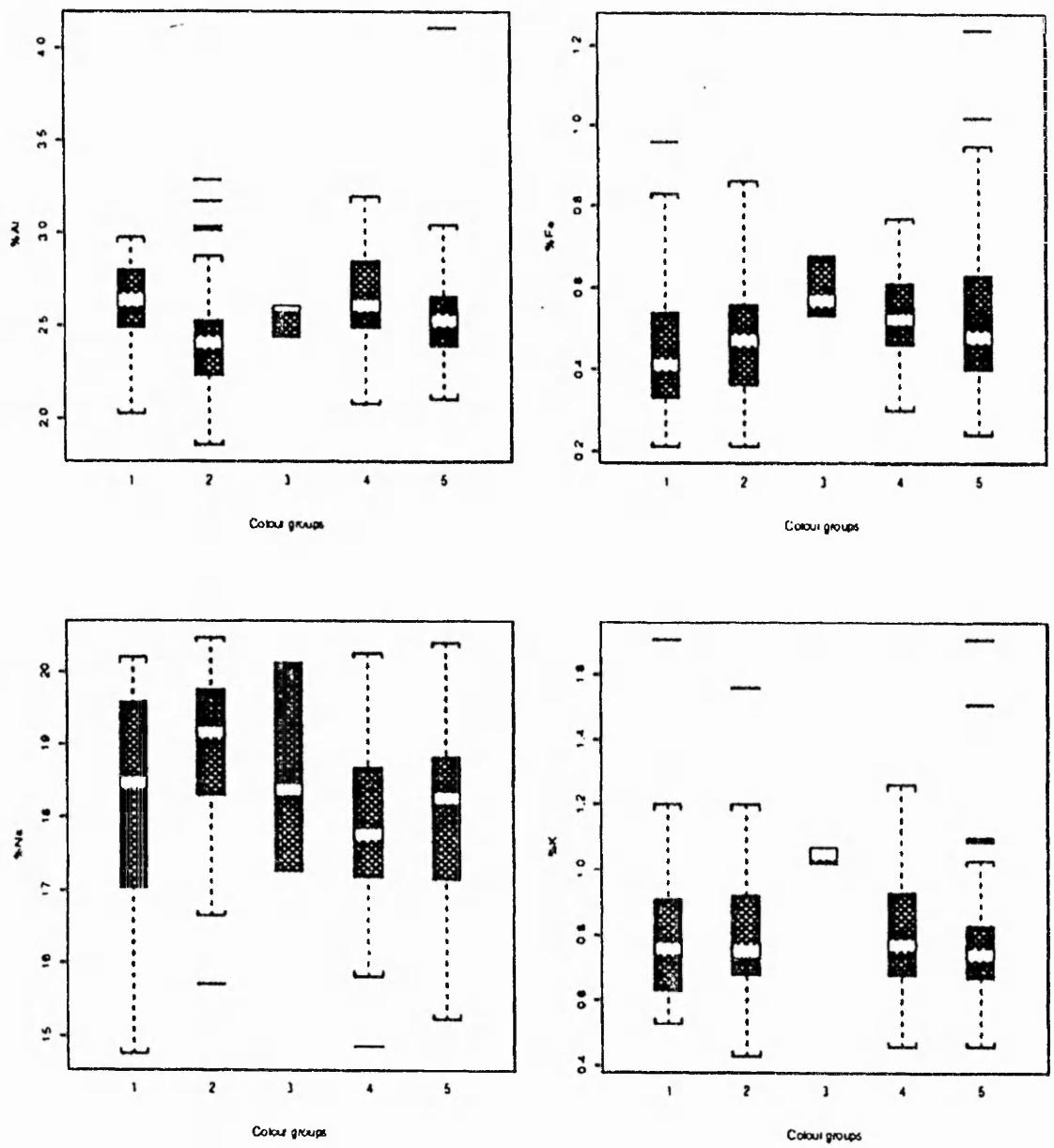
Table 6.3.5 Correlations of the elements and the first three principal components

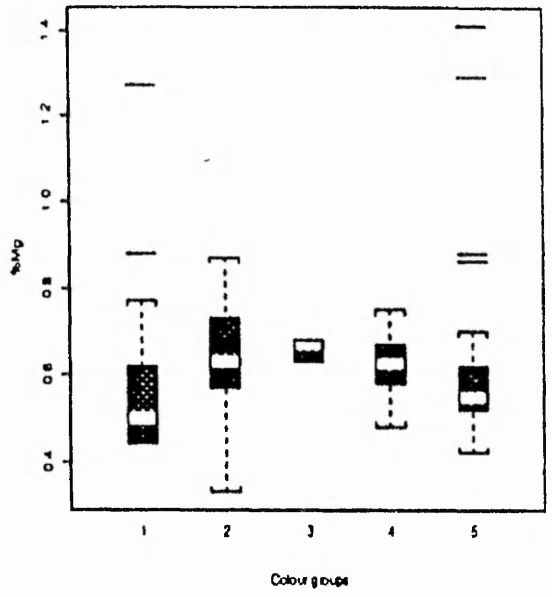
	pc1	pc2	pc3
Al	-0.11	-0.44	-0.40
Fe	0.27	**0.71**	-0.29
Mg	**0.90**	0.27	-0.17
Ca	**0.86**	-0.34	0.10
Na	-0.74	0.46	-0.12
K	**0.79**	0.28	0.09
Ti	0.27	0.68	-0.56
P	**0.95**	-0.08	0.18
Mn	-0.00	-0.18	**-0.83**
Pb	-0.08	0.55	0.22
Sb	-0.17	**0.78**	0.36

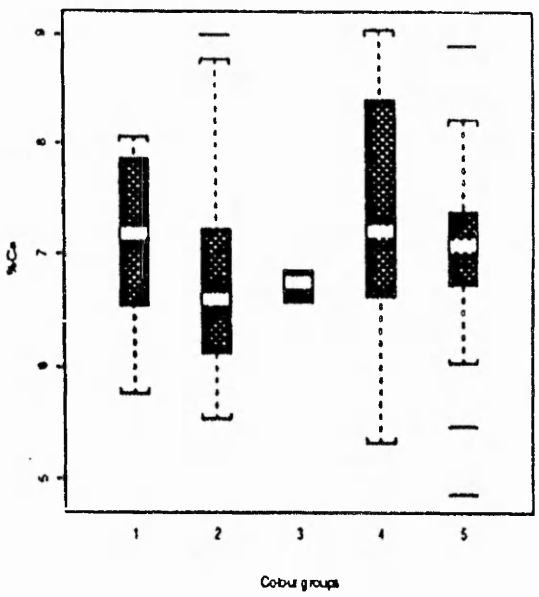
Table 6.3.5 shows that the first principal component correlates with P, Mg, Ca and K, the second with Fe and Sb and the third component with Mn, since they all have coefficients in excess of 0.7.

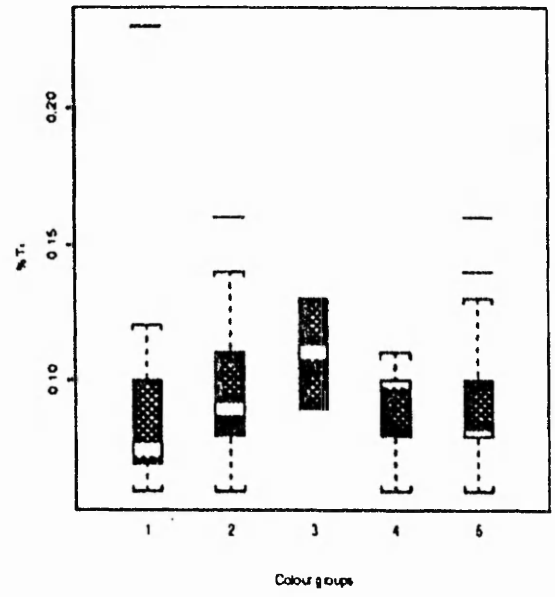
Figure 6.3.5 shows boxplots for each of the five colours found in the first colour group, observed in Figure 6.3.2. The colourless glass with light green tinge (2) has a lower Fe but higher Mn content, thus indicating the deliberate addition of Mn as a decolorizer. The blue-green with tendency towards green glass (4) has a higher Fe content than the light green glass, and a relatively high content of Mn, again indicating it's use as a decolorizer. When looking at the blue-green with tendency towards blue glass (5) and the blue-green glass (3), these both have a fairly high Fe content and a lower Mn content than the light green and tendency towards green glass which is what we would expect to find in a Roman blue-green glass.

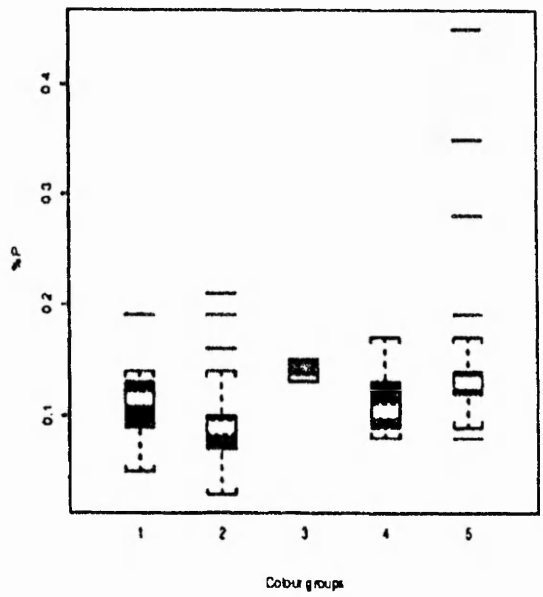
Figure 6.3.5 Boxplots showing the chemical composition of the observations coloured 1-5, where 1 - yellow/green, 2 - colourless with light green tinge, 3 - blue-green, 4 - blue-green with tendency towards green, 5 - blue-green with tendency towards blue

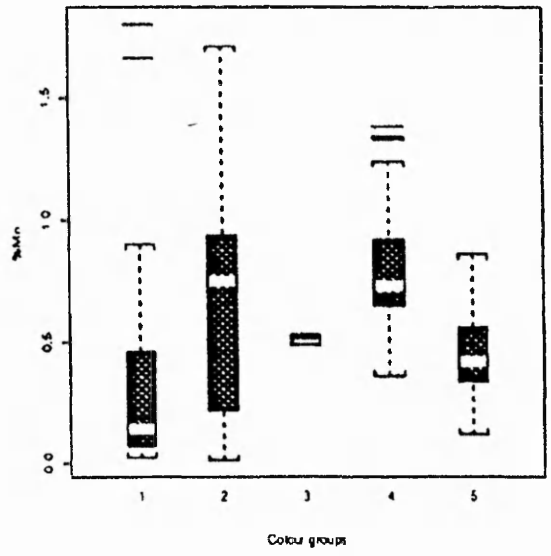


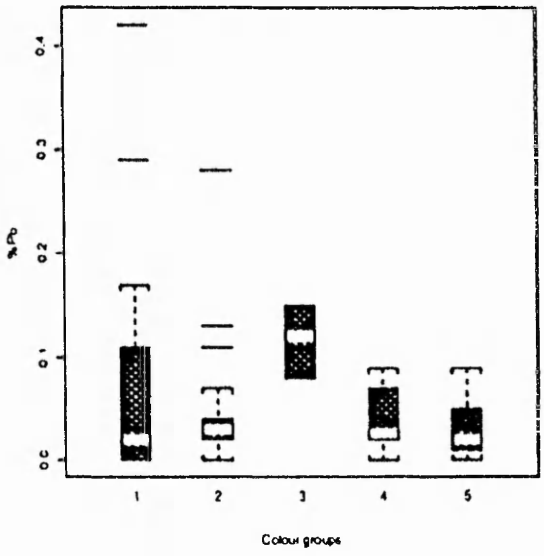

%Mg
0.4
0.6
0.8
1.0
1.2
1.4
1
2
3
4
5
Colour groups

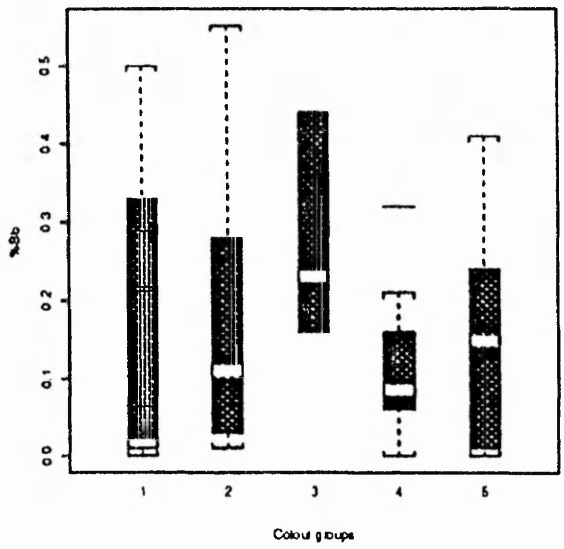

%Ca
5
6
7
8
9
1
2
3
4
5
Colour groups


%Ti
0.10
0.15
0.20
1
2
3
4
5
Colour groups


%P
0.1
0.2
0.3
0.4
1
2
3
4
5
Colour groups







To conclude, the Coppergate data appears to separate into two groups which are associated with colour.

6.4 Winchester cullet glass

Two hundred and fifty specimens of window glass found in a 'cullet bank' were taken from a site in Winchester and analysed. These 250 pieces were selected visually to be representative of the pieces found in the ancient cullet bank. Specimens of unusual colour were deliberately over-sampled. Analyses were undertaken using the major/minor oxides Al_2O_3, Fe_2O_3, MgO, CaO, Na_2O, K_2O, TiO_2, P_2O_5, MnO and the trace elements Ba, Co, Cr, Cu, Li, Ni, Sr, V, Zn.

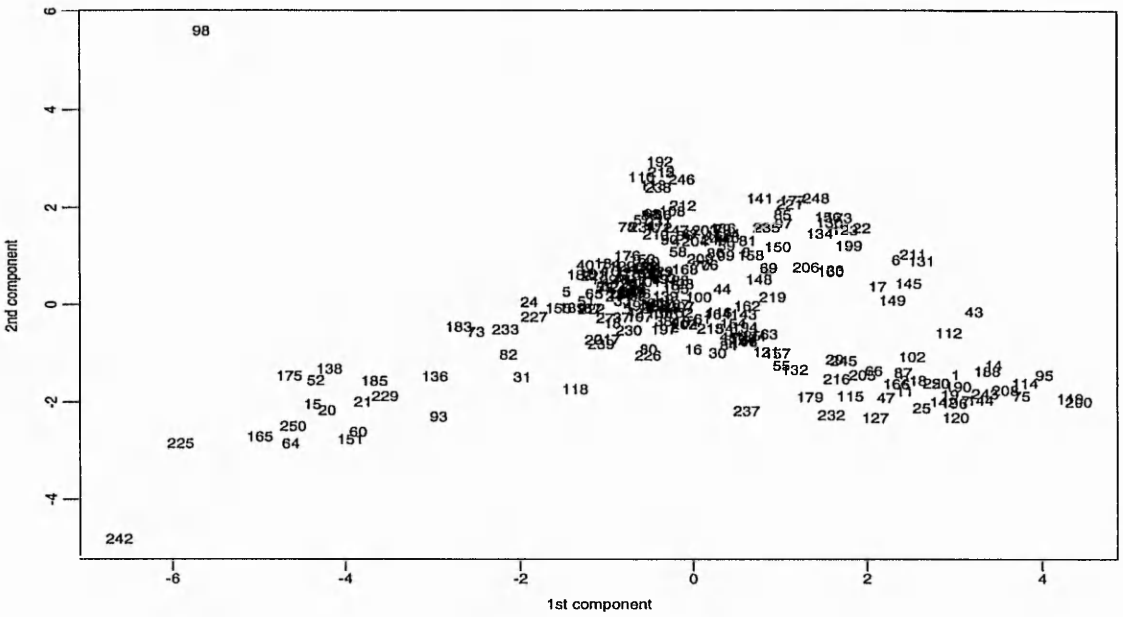
The following table lists the observations detected as outliers using various univariate and multivariate outlier detection methods on both the major and minor oxides.

Table 6.4.1 Table listing outliers suggested by various univariate and multivariate methods

Obs	Univariate methods	q^2_j	t^2_j	d^2_j	Hadi	A&M 80%	Ave	Sin	PCA
98	×	×	×	×	×	×	×	×	×
242	×	×	×				×	×	×
Additional									
84	×								
87	×								
179	×								
225	×								×
234	×					×			

Figure 6.4.1 shows a plot of the first two principal components using standardised data based on the correlation matrix. Observations 98 and 242 are seen to be the most outlying observations.

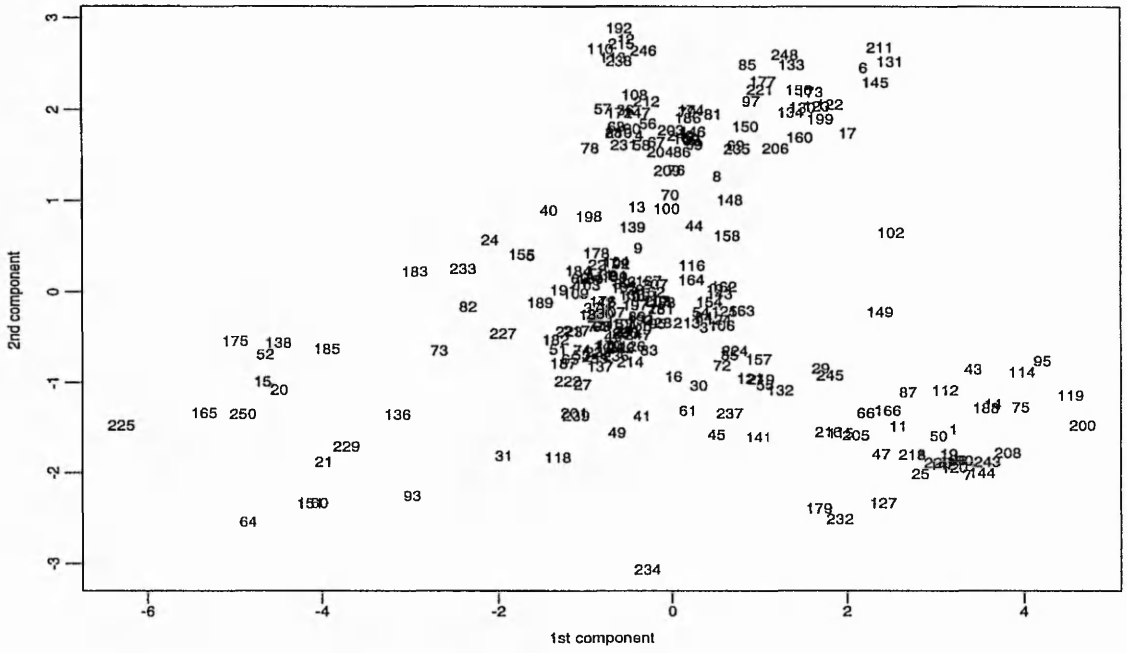
Figure 6.4.1 Plot of the first two principal components using standardised major/minor oxides based on the correlation matrix



As seen in Figure 6.4.1, observation 98 is the most extreme outlier. This observation, on re-inspection was found to be typologically distinct, being vessel glass of possibly near-Eastern origin. Observation 242, the next most extreme outlier, has unusually high values of K and P. Observations 98 and 242 are removed from the analyses as these are the most prominent and recurring outliers in all the above analyses undertaken, (see Table 6.4.1). A principal components analysis using standardised data after the removal of the two outliers produces the following plot, Figure 6.4.2.

The first component accounts for just 39% of the variation in the data and does not appear to be dominated by a single variable. The first and second components account for 60% of the variation in the data. The first four components are needed to 'explain' 86% of the variation in the data.

Figure 6.4.2 Plot of the first two principal components using standardised major/minor oxides based on the correlation matrix after the removal of observations 98 and 242



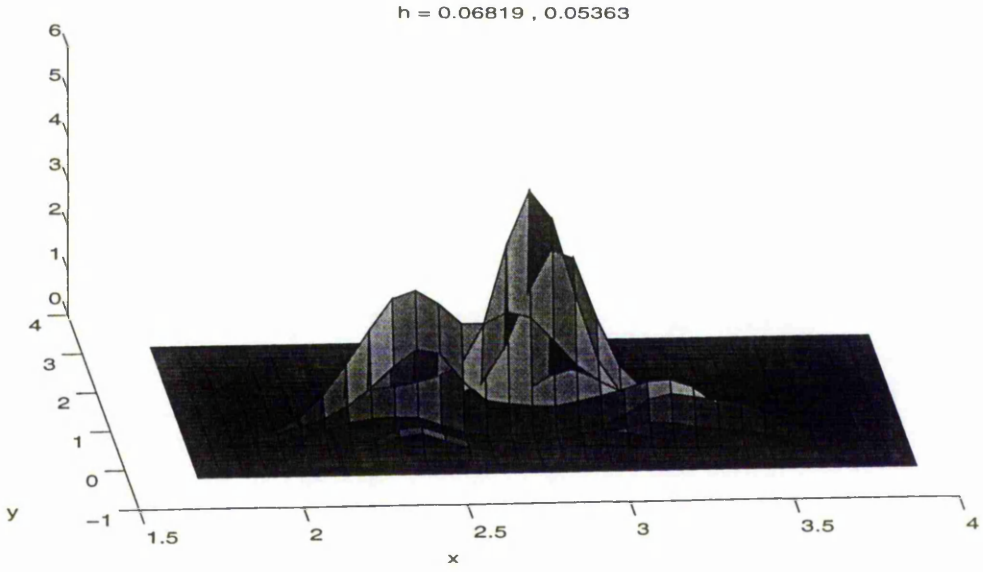
Correlations of the elements, listed in Table 6.4.2, are for all the data excluding the two outliers. It can be seen that Fe, Mg, P and K are all highly correlated. This relates to analyses of the Winchester window glass, section 6.2, as these same elements were also highly correlated.

Table 6.4.2 Correlations of the elements, after the removal of the two outliers

	Al	Fe	Mg	Ca	Na	K	Ti	P
Fe	-0.05							
Mg	-0.10	0.51						
Ca	-0.10	-0.18	-0.08					
Na	-0.27	0.06	0.27	-0.50				
K	-0.01	0.49	0.85	-0.07	0.10			
Ti	0.24	0.42	0.40	-0.20	0.27	0.10		
P	-0.17	0.57	0.79	0.13	0.04	0.86	0.16	
Mn	-0.69	0.11	0.46	-0.04	0.50	0.23	-0.05	0.35

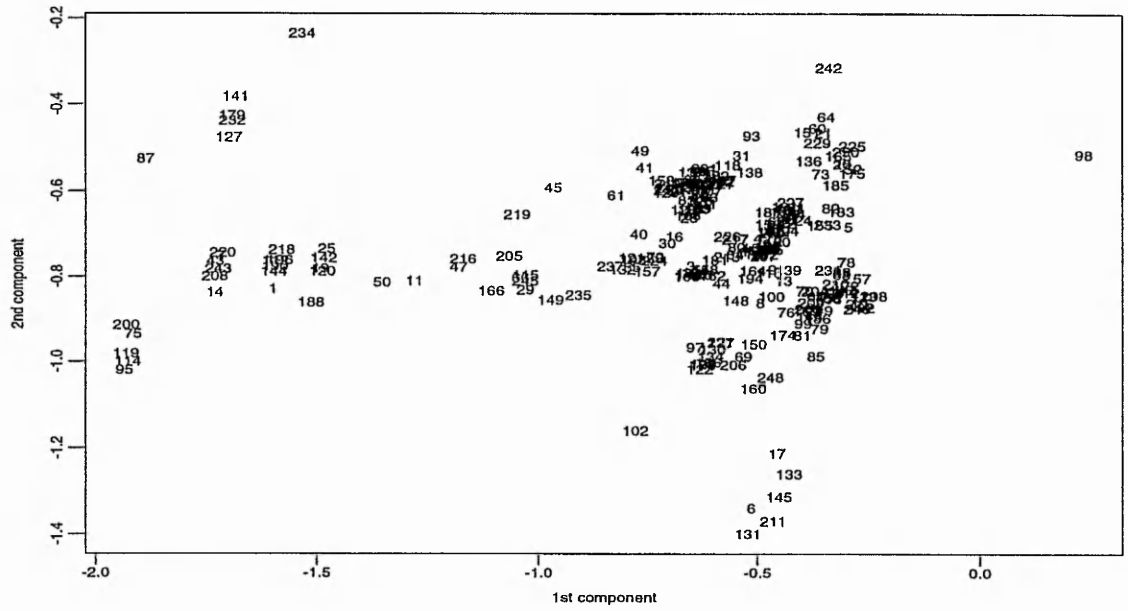
Figure 6.4.2 suggests the data is separating into 3 possible clusters, a closer inspection using a KDE identifies 3 groups. Figure 6.4.3 uses the STE method for the selection of h_1 and h_2, where $h_1 = 0.06819$ and $h_2 = 0.05363$.

Figure 6.4.3 Kernel density estimate plot of the first two principal components for all the data, after the removal of observations 98 and 242



This cullet glass has not previously been studied in great detail, unlike the other data sets, which have been the focus of some PhD theses. We have therefore decided to pursue the analyses further and differently than for the other sets. From the point of view of substantive interpretation, the log analysis of the data seemed potentially interesting. We use a log transformation since, in contrast to other analyses, the structure of the data is clearer using this type of data. A principal components analysis using logarithmically transformed, but unstandardised, data suggests several different groupings, with many of the specimens falling into 3 or 4 close but distinct clusters, see Figure 6.4.4.

Figure 6.4.4 Plot of the first two principal components using transformed major/minor oxides based on the covariance matrix



The separation into several distinct clusters can be seen more clearly than as observed in the 'standard' analysis of the data, (using standardised data). Using this transformed data, the first component accounts for 77% of the variation in the data and is entirely dominated by Mn with a coefficient of 0.97. The first two components account for 90% of the variation.

Table 6.4.3 Correlations of the elements and the first three principal components, using log-transformed data

	pc1	pc2	pc3
Al	-0.07	-0.12	-0.06
Fe	0.11	-0.58	-0.02
Mg	0.09	-0.21	-0.11
Ca	-0.01	0.09	-0.16
Na	0.04	-0.01	0.09
K	0.15	-0.33	-0.67
Ti	0.04	-0.63	0.57
P	0.13	-0.24	-0.39
Mn	0.97	0.19	0.14

As previously seen in Table 6.4.2, now using log transformed data, the same elements - Fe, Mg, P and K are all highly positively correlated.

Figure 6.4.5 shows a plot of the first two principal components, after the removal of observations 98 and 242, using logarithmically transformed data. As seen in Figure 6.4.4, many of the specimens fall into several distinct clusters or groupings. It is interesting to see that the plot of Fe against Mn, using log-transformed data is very similar to the PCA plot due to the fact that Fe and Mn are highly correlated, this is shown in Figure 6.4.6. This plot also shows the data separating into three or four main, and several minor, groups. As with the Southampton glass data, the main patterns in the data can be captured with far fewer variables.

Figure 6.4.5 Plot of the first two principal components using transformed major/minor oxides based on the covariance matrix after removal of observations 98 and 242

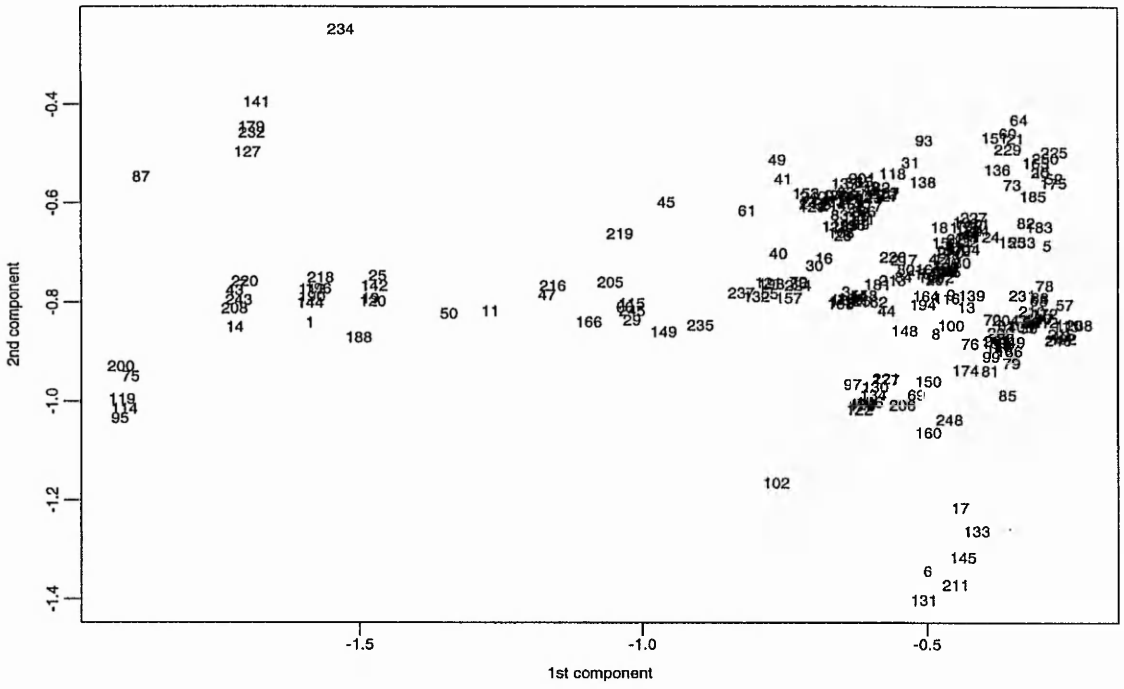
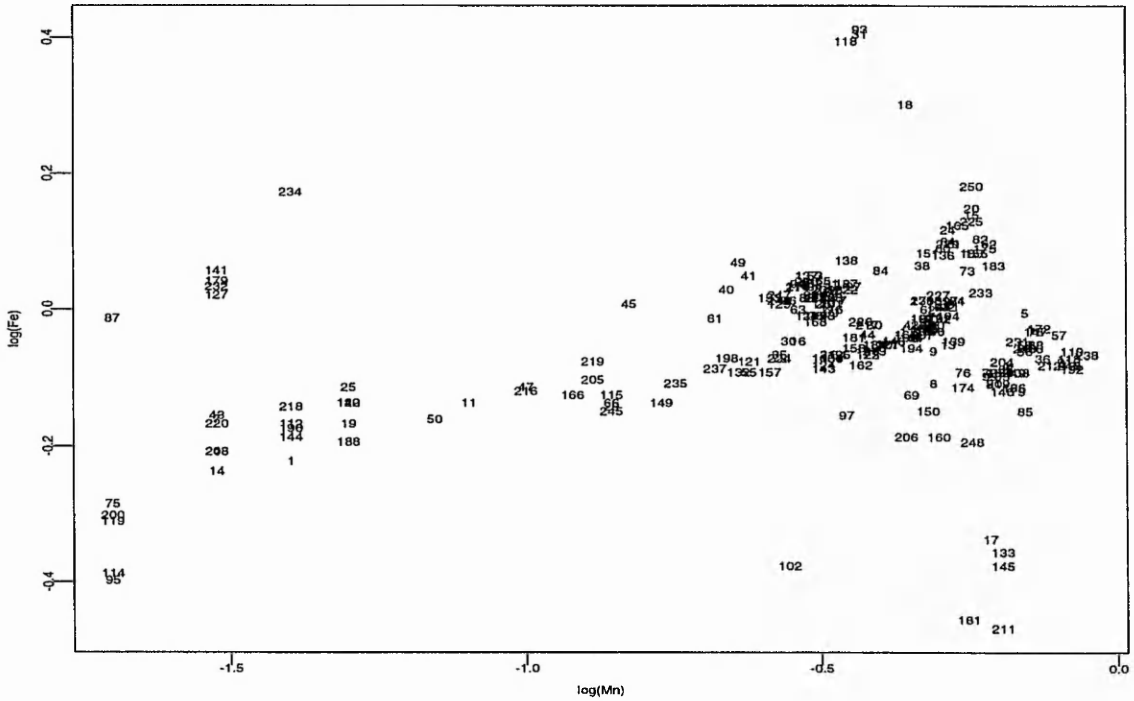


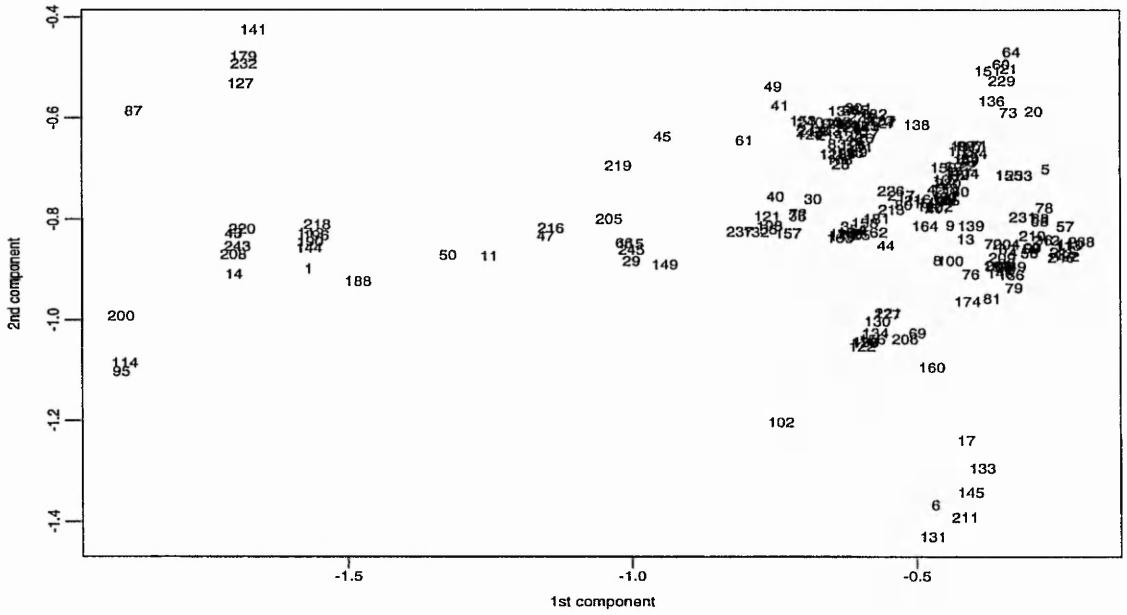
Figure 6.4.6 Plot of log Mn vs log Fe, after the removal of observations 98 and 242



Many of the smaller, distinct clusters or outlying observations seen on the periphery of the plots are strongly coloured pieces - emerald, turquoise and peacock blue. A majority of the pieces are coloured blue-green and the following analyses concentrate on this blue-green glass only, after the removal of observations 98 and 242. There are 208 specimens in all, some of them have red streaks in them but no evidence has emerged to suggest that they form a compositionally distinct group.

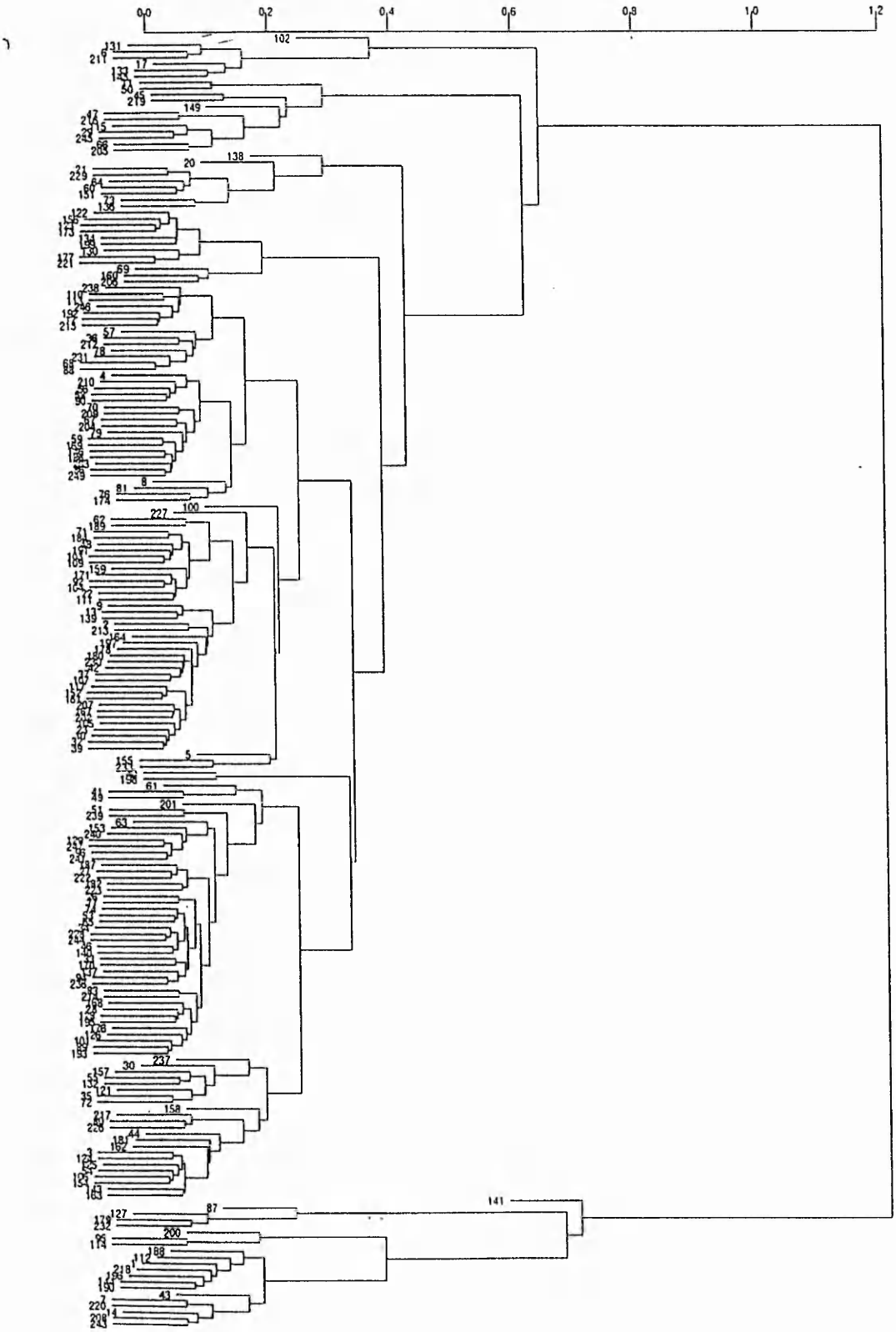
For the blue-green glass only, Figure 6.4.7 is a plot of the first two principal components using logarithmically transformed unstandardised data. Again the first two components 'explain' 91% of the variation in the data, with the first component being entirely dominated by Mn with a coefficient of 0.97.

Figure 6.4.7 Plot of the first two principal components using transformed major/minor oxides based on the covariance matrix - blue/green glass only



To define the groupings more clearly, an average link cluster analysis was run on the unstandardised log-transformed data, Figure 6.4.8.

Figure 6.4.8 Average link cluster analysis showing 23 cluster breakdown



A 23 cluster breakdown was selected simply to separate out the 3 or 4 main clusters evident in the upper right of the plot in Figure 6.4.7. The 3 or 4 main concentrations have 23 or more members, five of the smaller clusters have 6 or more members, with a further eight singleton clusters, three pairs and three triplets. These 23 'outliers' will be lumped together into a single 'miscellaneous' category. The four largest groups account for 141 (68%) of the blue-green specimens. Figure 6.4.9 is a plot of the first two principal components, labelled according to the groups defined by cluster analysis. Groups a, b, c, d, e, f, g, h, and k relate to groupings and group x relates to the 23 'outliers' lumped together into the miscellaneous category

Figure 6.4.9 Plot of the first two principal components using transformed major/minor oxides based on the covariance matrix - blue/green glass only, labelled according to group a-x

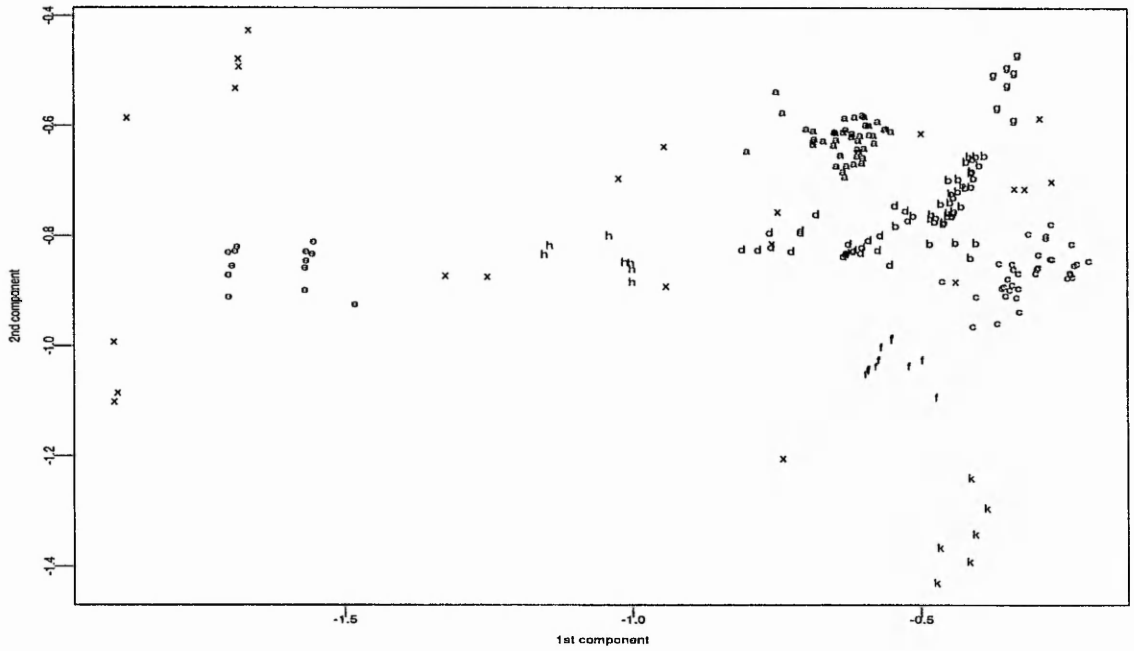


Figure 6.4.9 also suggests that groups c, d and possibly e could be sub-divided. The following graph, Figure 6.4.10, shows a plot of the first two principal components of the trace elements using log-transformed data. Again this differs from the treatment of the other data sets where the trace elements were not used in the analysis. It is hoped that an analysis of the trace elements will be consistent with that of the major/minor oxides. As seen in earlier analyses, observations 98 and 242 have been removed. The observations of the trace elements have been labelled according to the groupings a - h, k and x given in

Figure 6.4.9. The clusters seen in Figure 6.4.10 are distinct, although not as distinct as in the plot based on the major/minor oxides, and the same observations appear to fall into the same groupings, with some obvious overlap. Therefore we are able to conclude the trace elements analysis is consistent with the analysis of the major and minor oxides.

Figure 6.4.10 Plot of the first two principal components using transformed trace elements based on the covariance matrix - blue/green glass only, labelled according to group a - x

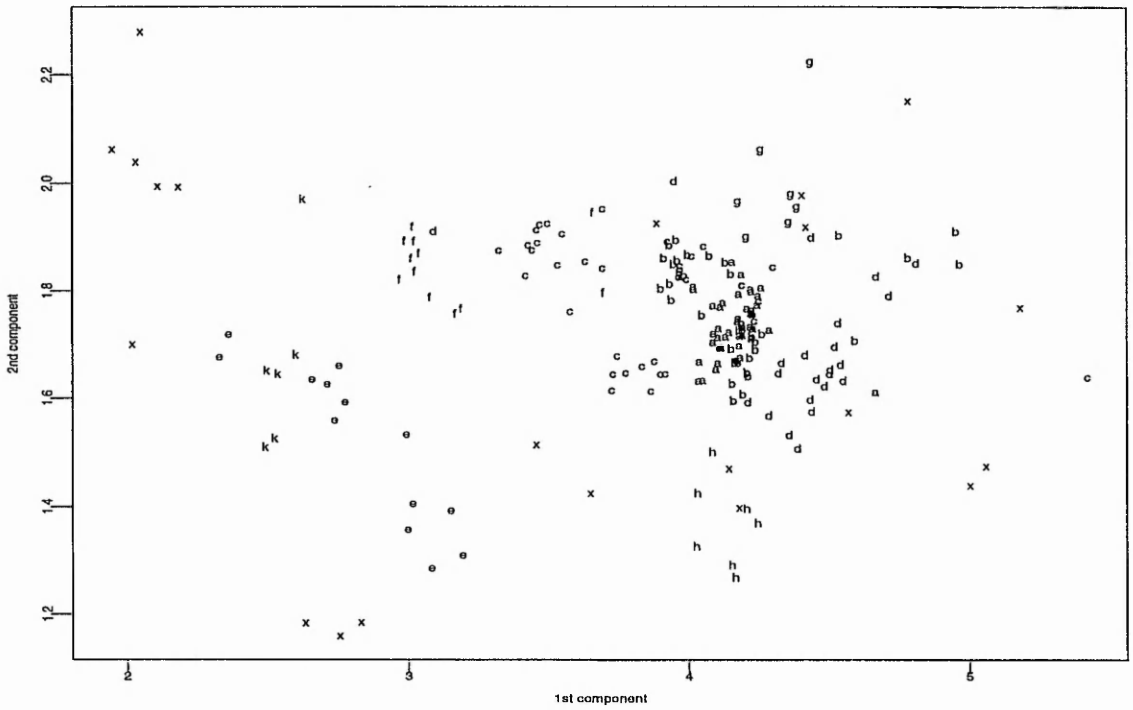
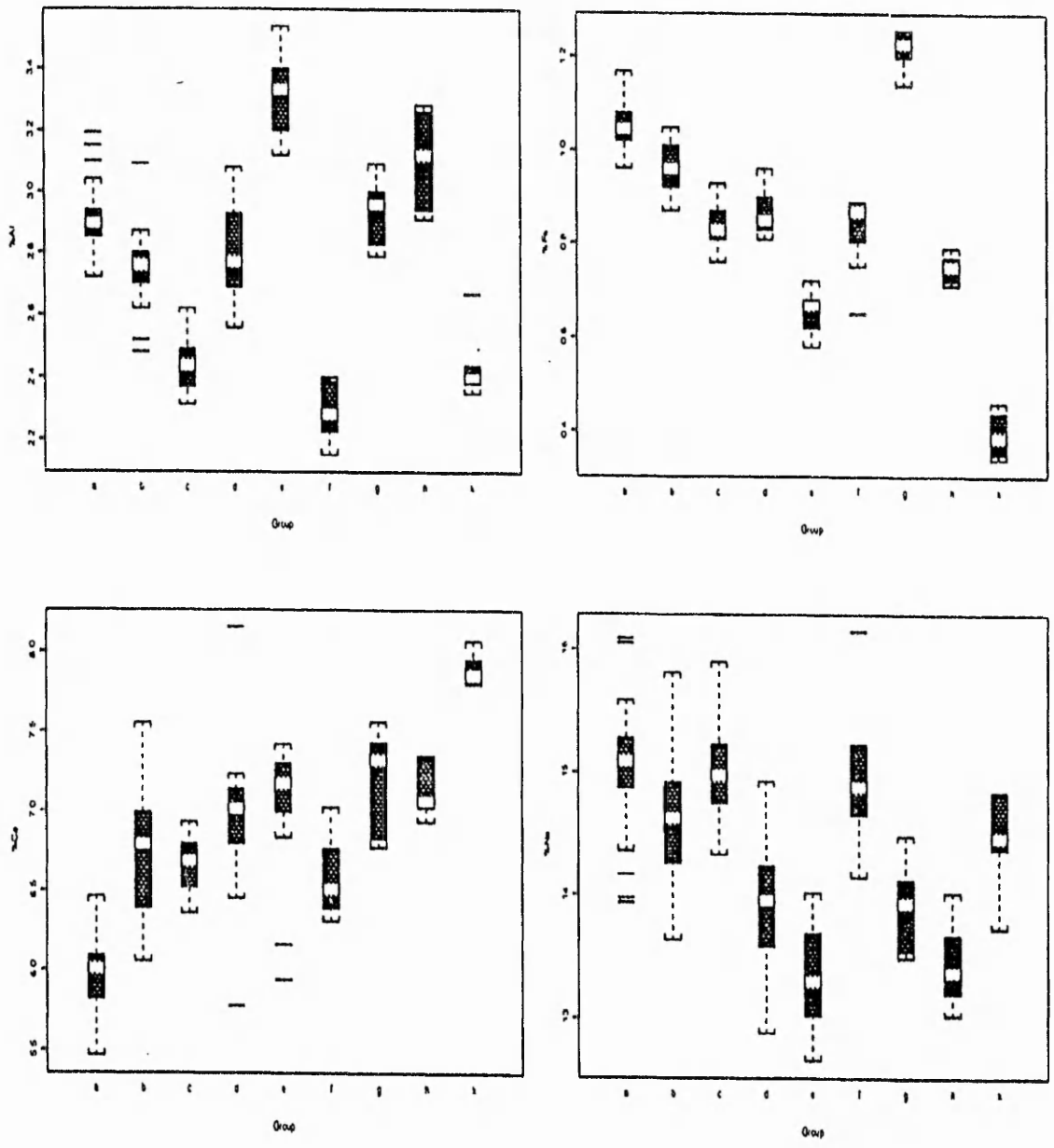
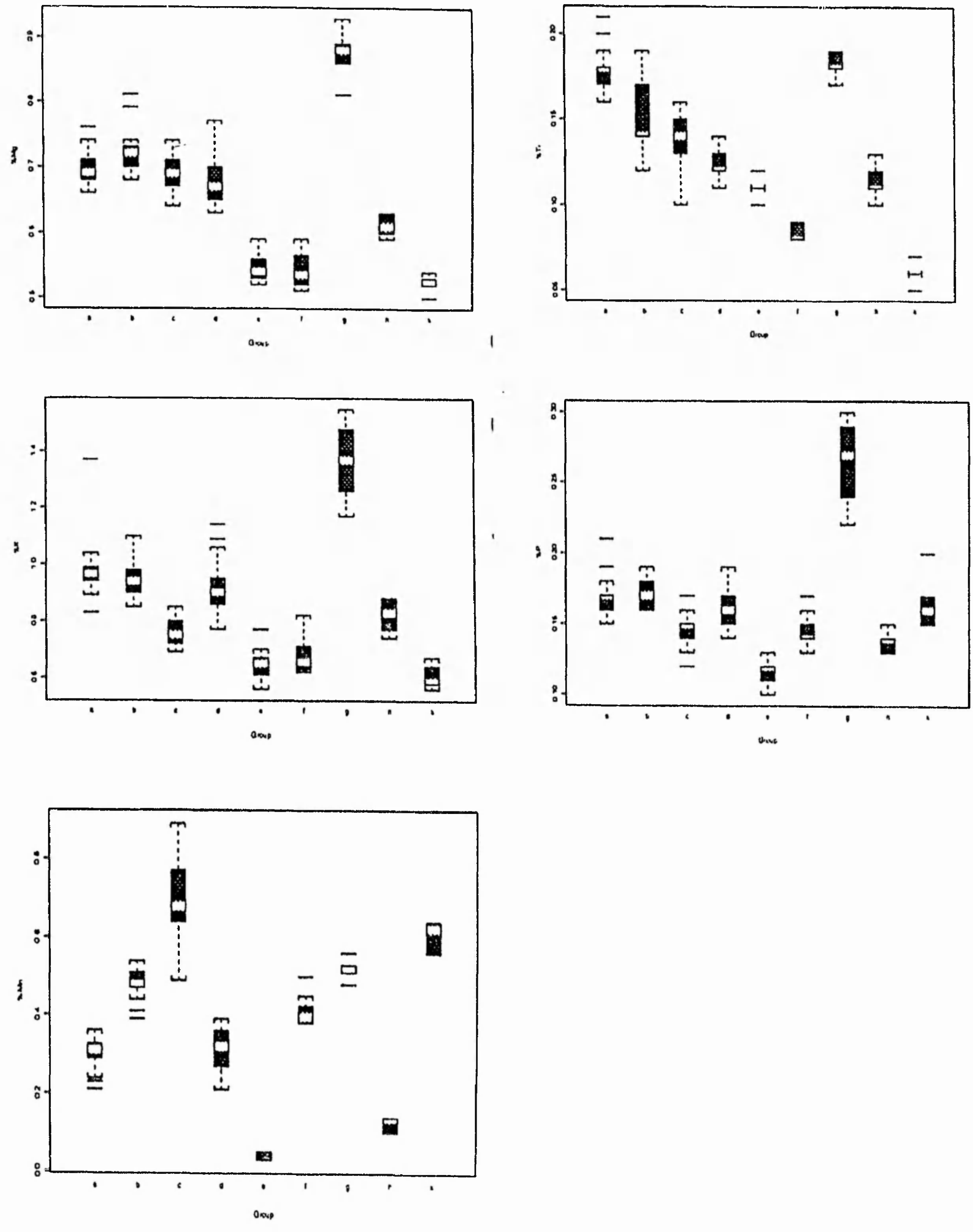


Figure 6.4.11 shows the boxplots for each major and minor oxide and groups a - k. Firstly, of the three largest groups a, b and c, the level of Mn almost completely separates them. Group a has high Fe and low Mn, group b has lower Fe and higher Mn and group c has a low level of Fe but high Mn. Group d is distinct from groups b and c with respect to the level of Mn; and from group a with respect to Fe and Ti. The smaller groups, e - k, can be discriminated from the rest, completely or nearly so, with respect to the levels of a few variables (e.g. Group e - low Mn, P, high Al; group f - low Al, Mg and Ti; group h - low Mn; group g - high Fe, Mg, K, P; group k - low Fe, Mg, Ti, high Ca, Mn).

Figure 6.4.11 Boxplots of the chemical composition of the groups a - x of the major/minor oxides - blue-green glass only





In order to analyse the data further the observations belonging to the larger groups (a - d) have been analysed separately. Figure 6.4.12 is a plot of the first two principal components labelled according to group a - d. As seen in the earlier analyses, these four groups separate out very distinctly with possible further sub-division of groups d, c and even b.

Figure 6.4.12 Plot of the first two principal components using transformed major/minor oxides based on the covariance matrix - blue/green glass only, labelled according to group a - d

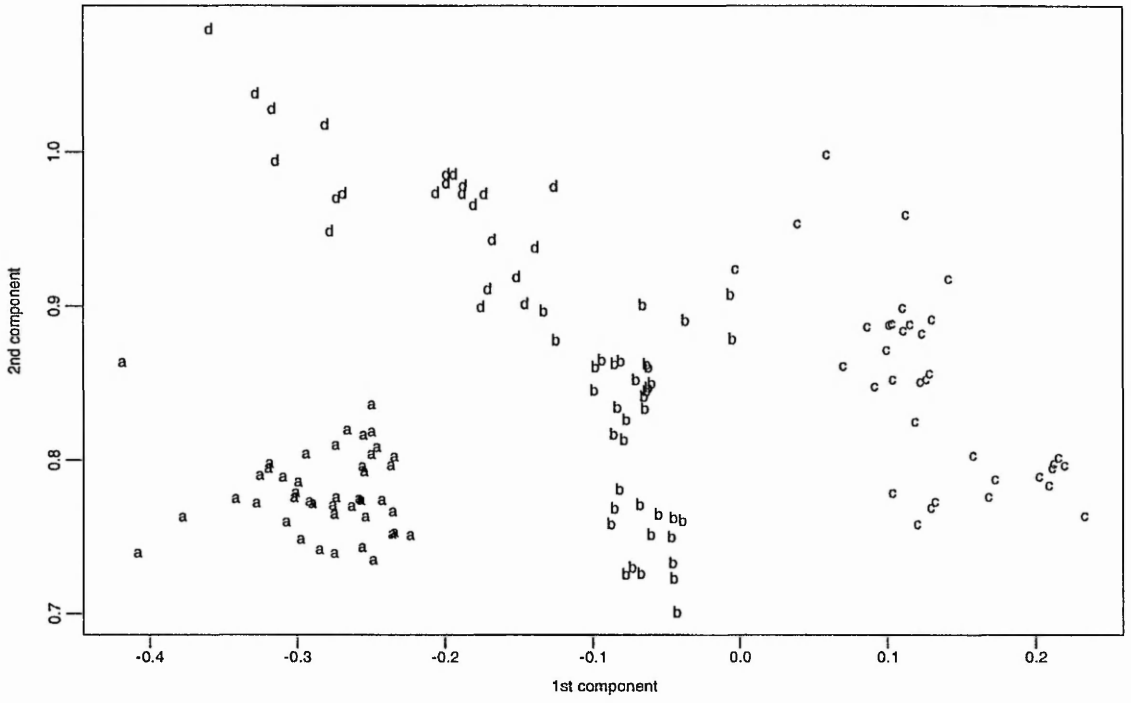
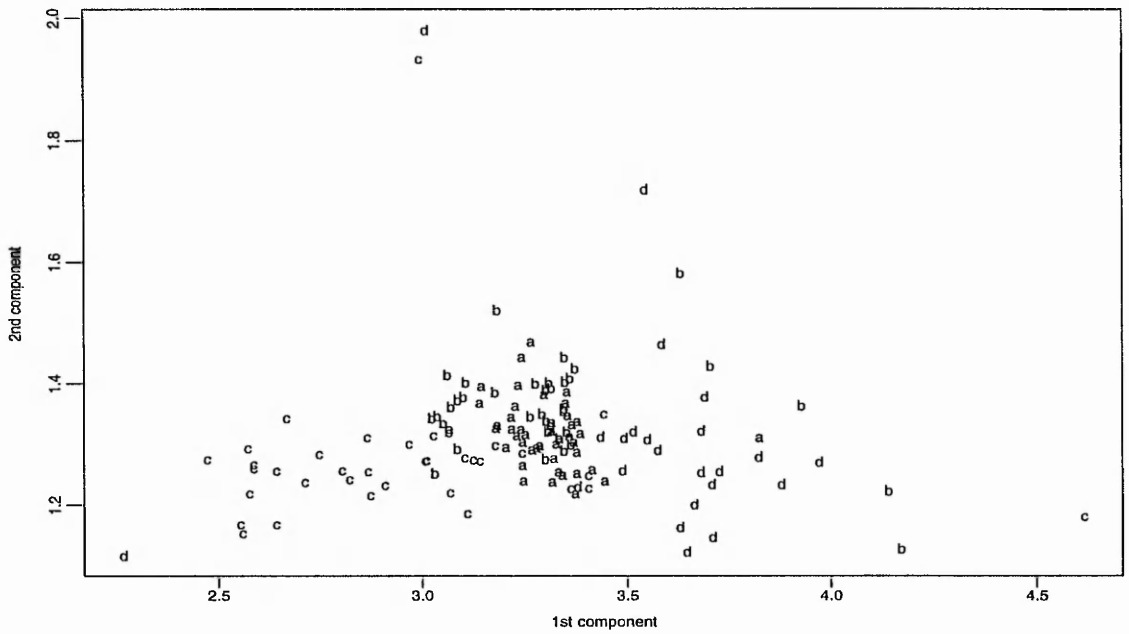


Figure 6.4.13 is a plot of groups a - d of the trace elements. The groups are not as distinct here as when using the major and minor oxides and also the trace elements may suggest some possible outliers not evident from the major/minor oxide analysis.

Figure 6.4.13 Plot of the first two principal components using transformed trace elements based on the covariance matrix - blue/green glass only, labelled according to group a - d



Due to the distinct separation of the 3 or 4 groups, which are also compositionally distinct, identified in the boxplots of Figure 6.4.11, it is possible that the different groups correspond to glass from different panes, therefore a relatively small number of windows.

The assemblage excavated at Winchester is slightly different, in respect to colour separation, than the other data sets that have been analysed. The glass does appear to separate on the basis of colour but since most of the specimens are blue-green we have concentrated on this to a further extent. As the majority of the specimens were blue-green glass the groups may correspond to different panes of glass or glass made in the same batch, whereas the other assemblages, found at Southampton, Coppergate and the Winchester vessel glass, appear to separate on the basis of colour. Another important feature of this data set, as with the Southampton glass, is that Fe is an important oxide in determining patterns in the data.

This cullet glass analysis has suggested some interesting groupings, the detailed interpretation of which is beyond the scope of this thesis, but will be pursued with the archaeologists involved.

7. Results and Conclusions

7.1 Introduction

Having analysed the five different glass data sets, results and conclusions are outlined below. Throughout the course of the thesis two issues have been raised, methodological and substantive. Firstly we discuss the methodological issues and take a look at the similarities and differences between the different outlier detection methods used. In the substantive discussion we observe how Fe (and the suite of associated oxides) appears to play a major role in the separation of the various data sets into groupings.

7.2 Methodological discussion and conclusions

The work presented here represents an overview of the theory and methods available for use with multivariate data. We have summarised different methods that can be used to graphically analyse a set of data and introduced different outlier detection techniques. The univariate methods examined suggest far more outliers for each data set than the multivariate methods. The actual nature of interpretation for the univariate analyses is slightly different from that of the multivariate analyses. Having looked at box and whisker plots, those observations detected as outlying by these plots (i.e. lying > 3.5 box widths away from the rest of the data) have been classed as outliers. In order to help verify this, the dotplots have also been analysed to see if any other observations are brought to light. Within this thesis we are obviously constrained as to how much graphical output can be included and so only tables illustrating our findings have been included. A subset of those outliers suggested by the univariate techniques are also picked up by the average linkage cluster analysis method and in turn the single linkage method tends to identify a subset of, or equivalent, outliers to those identified by average linkage.

Of the multivariate methods, d^2_j, Hadi and Atkinson and Mulira, all appear to produce identical results for each of the five data sets analysed. This is surprising in that d^2_j is the Mahalanobis distance and the other methods are, in theory, meant to improve on this.

Various approaches to multivariate data analysis have been discussed. In particular the advantages and disadvantages of outlier detection have been described and these methods are thought to have certain computational advantages over other analyses which aim to detect outlying observations. One disadvantage is outlined in Chapter 5 when detecting outliers amidst dense groupings in the data. It may be concluded that the methods considered here have great promise. Even at the early stages they perform better than existing approaches, such as univariate analyses, but great care must be taken when working with 'unusual' data, i.e. data which lies away from normality, since conclusions drawn from the analyses and possible outliers detected may be unjustified.

Another focus is that, after initial MVA, it is possible to summarise the main patterns in the data using just two of the variables. We examine this to see if it is consistent across the other data sets. As a result we are able to conclude that for the Southampton and Winchester Cullet glass data sets plots of Fe vs Mn will reveal the main structures in the data. due to colour separation for Southampton and chemical composition for Winchester Cullet.

We have also summarised kernel density estimation and its uses for graphically analysing both univariate and multivariate data, especially after outlier removal. This approach is relatively innovative in archaeology and the advantages are obvious for use with archaeological data. The methodology appears to be most useful for large data sets, where the conventional 2-D scatterplot may not reveal important features of the data. KDEs lend themselves naturally to contouring (Baxter, Beardah and Wright, 1995), in that it is possible to divide data into sub-groups and plot selected contours for each sub-group separately in order to examine their similarities and differences.

7.3 Comparisons of the glass analysed

The assemblage excavated at Southampton dates to the early Medieval period, 9th/10th century AD. After various analyses, the data appear to separate on the basis of colour. This assemblage consists of two colour groups, light blue and light green, which also

appear to be compositionally distinct. The levels of Fe and Mn and their ratio, Fe:Mn, are consistent with what we would expect to see in ancient glasses, 2.1:1 for light blue glass and 1.1:1 for light green glass, although it is thought that the light blue glass composition was initially light green in colour and gradually altered from light green to light blue as the furnace atmosphere became more reducing (Heyworth, 1992). The light blue glasses may also originate from a plant alkali source since this colour group has a high K and Ca content and lower Na content. The light green glasses have a higher Na content, indicating possible saltwater alkalis.

The Winchester vessel glass is thought to date from the late Roman period, 4th century AD, and the colour of this assemblage is predominantly light green. The data form two groups and discrimination between these two groups is based mainly on the Fe content. The light green green glass has a high Mn content with the Fe:Mn ratio at approx. 1.1:1 and, as seen in the light green Southampton glass, a high Na content indicating possible saltwater alkalis. The blue vessel glass found at Winchester has a Fe:Mn ratio of approx. 2.1:1 which is also seen in the light blue Southampton glass.

In conclusion, both the Southampton and the Winchester vessel glass form two separate colour groups consisting of light blue/blue and light green/green glass. In both cases it is thought that Mn has been added deliberately to act as a decolorizer.

The Winchester window and Winchester cullet glass can be viewed slightly differently because, although they separate into groups which appear to be colour-related, specimens of the same colour also separate into distinct groupings. In the case of the Winchester window glass those specimens coloured a darker blue form a separate group to those coloured light blue, but the light blue specimens then separate into additional clusters. The dark blue glass has a Fe:Mn ratio of approx. 2.1:1, with corresponding higher contents of Fe and Mn than those found in the light blue glass (although this glass also has a Fe:Mn ratio of approx. 2.1:1). The Winchester cullet glass is mainly blue/green in colour and additional analyses on just those specimens coloured blue/green have been performed. The data separate into four main clusters, where the level of Mn discriminates them. Looking at the Fe:Mn ratio, for each group this is approx. 2.1:1, but as with the Winchester window glass, one group has corresponding higher levels of Fe and Mn, thus indicating a darker

blue colour. It has been suggested that the groups seen in the Winchester cullet assemblage correspond to different panes of glass and with this in mind it is possible that the Winchester window glass also separates according to different panes of glass. The Coppergate glass also differs slightly as it contains a large amount of 'waste' glass which is possibly different in date to the rest of the glass found in this assemblage. It also contains a number of colourless specimens which have a high Sb content. The Sb appears to not be significantly correlated with any other oxides, suggesting it was added separately as a relatively pure decolorizer, (i.e. not premixed with the sand). Those specimens which do exhibit high levels of manganese (>1 %), tend to be either light green or light green-colourless glasses. On removing the colourless and waste glass, two groupings appear that are colour-related. The light green/green colour group is found to have a higher content of Mn and lower Fe than the second blue-green/green colour group, which has a higher Fe and lower Mn content level.

It must also be taken into account that the above assemblages have been colour coded by different archaeologists, Heyworth - Southampton, Winchester vessel and Winchester window, Cool - Winchester cullet and Jackson - Coppergate, so although they are rather subjective, it is also hoped they are consistent.

Table 7.3.4 lists a summary of each data set analysed. As discussed above, three of the five data sets analysed do relate to colour separation, namely Southampton, Winchester Vessel and Coppergate. These tend to separate according to the same colours light green/green and light blue/blue. Again these conclusions are to be expected for archaeological data due to the nature of the glass, the chemical composition of the glass and the glass-making process. The two data sets which do separate into groups which are not colour-related, Winchester Window and Winchester Cullet, do tend to separate for the same reasons as the above data sets - the chemical composition, namely the content of Fe, and the associated suite of oxides (Ti and Mg), and the Mn content.

Table 7.3.4 Summary of groupings identified in the archaeological glass data sets analysed

Glass	Groupings	Colour Related	Colours	Principal Oxides
Southampton	Y - 2	Y	Light blue, Light green	Fe
Winchester Vessel	Y - 2	Y	Light Green & green, blue	Fe
Winchester Window	Y - 4	N		Fe
Coppergate *	Y - 2	Y	Light green, blue/green & blue	
Winchester Cullet	Y - 3/4 main groups and some smaller groupings	N		Mn

* after removal of crucible waste and colourless glass

7.4 Substantive issues and conclusions

Throughout the analyses, Fe appears to play a large part in the separation of the different data into colour-related groupings. Table 7.4.5 lists the correlations of the remaining 10 oxides with Fe (rounded to 1dp).

Table 7.4.5 Correlations of the remaining oxides with Fe (rounded to 1 dp)

Oxide	Southampton	Winchester Vessel	Winchester Window	Coppergate (after removal of crucible waste and colourless glass)	Winchester Cullet
Al	0.7				
Mg	0.7	0.9		0.6	0.6
Ca	0.6	-0.6			
Na		0.5			
K	0.7				0.5
Ti	0.7	1.0	0.7	0.6	0.6
P			0.5		0.6
Mn		0.9			
Pb					*
Sb		*			*

* = Not applicable Blank = |r| = < 0.45

According to Table 7.4.5, for all the data sets, Fe and Ti are generally highly correlated, followed by Fe and Mg. As previously mentioned these three oxides are thought to enter the glass mixture together via the silica. It must be noted that, although Fe and Ti are highly correlated, there is no other consistency apart from this noted.

Table 7.4.6 Correlations of each oxide with the 1st principal component

Oxide	Southampton	Winchester Vessel	Winchester Window	Coppergate	Winchester Cullet
Al	0.8				
Fe	0.9	0.4	0.4		
Mg	0.8	0.4		0.9	
Ca	0.7		-0.3	0.9	
Na	-0.6			-0.7	
K	0.8			0.8	
Ti	0.7	0.4	0.3		
P				0.95	
Mn		0.4			0.97
Pb			0.4		*
Sb		*	0.4		*
	>0.5	>0.3	>0.3	>0.7	>0.2

*Not Applicable

Table 7.4.6 indicates that Fe is among the oxides most strongly related to the first principal component for three of the data sets, namely Southampton, Winchester Vessel and Winchester Window glass. The other oxides which are highly correlated with the first principal component (with some exceptions) tend to be those highly correlated with Fe, see Table 7.4.5. The Winchester Cullet glass is an exception, bearing in mind this was also analysed differently using logged data, in that Mn dominates. It also must be noted here that Fe and Ti actually dominate the second principal component, see Table 7.4.7. The final row in each table, Table 7.4.6 - Table 7.4.8, relates to the coefficient above which each oxide listed has a significant bearing on the analysis. This is subjectively chosen for each data set rather than having the same conditions for each.

Table 7.4.7 Correlations of each oxide with the 2nd principal component

Oxide	Southampton	Winchester Vessel	Winchester Window	Coppergate	Winchester Cullet
Al		0.5			
Fe				0.7	0.6
Mg			0.5		
Ca		0.4			
Na	0.4				
K			0.4		
Ti				0.7	0.6
P	0.6	-0.4	0.4		
Mn	0.4				
Pb	0.6	-0.6			
Sb	0.7	*		0.8	
	>0.4	>0.4	>0.3	>0.6	>0.4

*Not Applicable

Table 7.4.8 Correlations of each oxide with the 3rd principal component

Oxide	Southampton	Winchester Vessel	Winchester Window	Coppergate	Winchester Cullet
Al		0.6	0.7		
Fe					
Mg					
Ca					
Na	0.4				
K					-0.7
Ti	-0.6				0.6
P	0.4	0.4			-0.4
Mn	-0.8			-0.8	
Pb			0.4		
Sb		0.4			
	>0.4	>0.4	>0.4	>0.6	>0.3

For the Winchester Cullet and the Southampton glass, where Mn dominates the first and third principal component respectively, the principal component plots (Figure 6.4.5 and Figure 5.6.2) are very similar to the Fe vs Mn plots (Figure 6.4.6 and Figure 5.6.7). These both show structure, which is related to colour for the Southampton glass but not for the Winchester Cullet glass. The Winchester Cullet glass is to be further examined (by archaeologists) in light of the groupings obtained from the analyses here.

For the Winchester Vessel glass, Mn and Fe are both highly correlated with each other and with the first principal component. For the Winchester Window glass, Mn does not feature in any of the principal components.

There is no simple pattern, but the regular appearance of Fe (and associated oxides) is worthy of note. Going back to the chemical composition of archaeological data, the content of Fe and the ratio of Fe:Mn, from the graphical analyses of Chapters 5 and 6 we are able to make a connection between the differing colour of glass and this ratio. Those glasses which are a lighter blue in colour appear to have a higher Fe:Mn ratio than those glasses which are light green. This in turn can be related, possibly, to the sand source and the alkalis, two of the raw materials used in the glass making process. In many archaeological texts this colour separation is not discussed and it may be concluded that the various graphical approaches used in this thesis reveal features, in archaeological glass data, which previously went undetected. One reason that colour is not often taken into consideration when analysing archaeological data is that it is not often recorded. So although patterns which are detected in glass data might be attributable to colour separation we may never be certain. This does not however offer an explanation for the separation of the Winchester Cullet and Window glass, which do separate according to this Fe/Fe:Mn content, but have no relation to the colour.

In conclusion the Southampton, Winchester Vessel and Coppergate assemblages separate according to the same colours light green/green and light blue/blue. Again these conclusions are to be expected for archaeological data due to the nature of the glass and the chemical content of Fe, Mn and Fe:Mn. On the other hand the Winchester Window and Winchester Cullet although not related to colour, do tend to separate for the same reasons as the above data sets - the chemical composition, namely the content of Fe, and the associated suite of oxides (Ti and Mg), and the Mn content.

7.5 Future Work

The Winchester cullet glass analysis suggested some interesting groupings, the detailed interpretation of which is beyond the scope of this thesis, but will be pursued with the archaeologists involved.

During the course of this research it has become clear that colour plays an important role when analysing ancient glass. Colour can be affected in three ways - by the variations in composition, the time spent in the molten condition and by the atmosphere in the furnace. It is already thought that a major contributor is the content of Fe and Mn and also the redox equilibrium between Fe and Mn. Any further work on evidence to suggest this may ascertain if the ancient glass-makers where chemists who actually experimented with differing levels of oxides or if the colouring in glass can simply be related to the sand source and the alkalis.

Appendix

All MATLAB routines for kernel density estimation used throughout this thesis can be downloaded from the Internet. These are found in 'Internet Archaeology' in an article by Beardah and Baxter (1996).

The location is as follows :
http://intarch.ac.uk/journal/issue1/beardah/kdeia6#xtocid28219

Any interested parties can obtain the software, and any updates, by contacting Dr. Christian C. Beardah.
e-mail : ccb@maths.ntu.ac.uk

7.6 References

Atkinson, A.C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, **Vol. 89**, No. 428.

Atkinson, A.C. and Mulira, H.M. (1993). The stalactite plot for the detection of multivariate outliers. *Statistics and Computing*, **3**, 27-35.

Bacon-Shone, J. and Fung, W.K. (1987). A new graphical method for detecting single and multiple outliers in univariate and multivariate data. *Applied Statistics*, **36**, 2, 153-162.

Barnett, V. (1976). The ordering of multivariate data (with discussion). *Journal of the Royal Statistical Society, Series A,* **139**, 318-355.

Barnett, V. and Lewis, T. (1994). *Outliers in statistical data. 3rd edition.* John Wiley and Sons Ltd. N.Y.

Baxter, M.J. (1993). Principal component analysis in archaeometry. *Research Report - The Nottingham Trent University*, No. 1/93.

Baxter, M.J. (1994). *Exploratory multivariate analysis in archaeology*. Edinburgh University Press, Ltd.

Baxter, M.J., Beardah, C.C. and Wright, R.V.S. (1995). Some archaeological examples of kernel density estimates. *Research Report - The Nottingham Trent University*, No. 12/95.

Beardah, C.C. and Baxter, M.J. (1996). MATLAB routines for kernel density estimation and the graphical representation of archaeological data. *Publications of the institute of prehistory. University of Leiden*, **28**.

Beardah, C.C. and Baxter, M.J. (1996). The archaeological use of kernel density estimates'. *Internet Archaeology*, **Vol. 1, 5.1**. (http://intarch.ac.uk/journal/issue1/beardah/kdeia6#xtocid28219)

Bibby, K and Davies, N. (1995). STEPS for Learning Statistics. *Teaching Statistics*, **17**, 107-110.

Bowman, A. and Foster, P. (1993). Density based exploration of bivariate data. *Statistics and Computing*, **3**, 171-177.

Caroni, C. and Prescott, P. (1992). Sequential application of Wilk's multivariate outlier test. *Applied Statistics,* **41**, No. 2, 355-364.

Cook, R.D. and Hawkins, D.M. (1990). Comment on Rousseeuw and van Zomeren (1990). *Journal of the American Statistical Association* **85**, 640-644.

Everitt, B.S. (1993). *Cluster Analysis*, 3rd Edition. John Wiley & Sons, Inc. N.Y.

Geilmann *et al.* (1955). Beitraege zur Kenntnis alter Glaeser III: Die chemische Zusammensetzung einiger alter Glaeser, insbesondere deutscher Glaeser des 10. Bis 18. Jahrhunderts. *Glastech.* Ber. **28**, 146-156.

Gnanadesikan, R. and Kettenring, J.R. (1972). Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics*, **28**, 81-124.

Hadi, A.S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series B*, **54, No. 3**, 761-771.

Hadi, A.S. (1994). A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society, Series B*, **56, No. 2**, 393-396.

Hawkins, D.M., Bradu, D. and Kass, G.V. (1984). Location of several outliers in multiple-regression data using elemental sets. *Technometrics*, **26**, 3, 197-208.

Henderson, J. (1985). The raw materials of early glass production. *Oxford Journal of Archaeology*, **4(3)**.

Heyworth, M.P. (1991). An archaeological and compositional study of early medieval glass from north-west Europe. *PhD Thesis, University of Bradford.*

Jackson, C. (1992). A compositional analysis of roman and early post-roman glass and glassworking waste from selected British sites. *PhD Thesis, University of Bradford.*

Jackson, J.E. (1991). *A user's guide to principal components.* John Wiley & Sons, Inc. N.Y.

Jolliffe, I.T. (1986). *Principal Components Analysis.* Springer-Verlag, New York.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding groups in data.* John Wiley & Sons, Inc. N.Y.

Krzanowski, W.J. and Marriott, F.H.C. (1994). *Multivariate analysis. Part 1 Distributions, ordination and inference.* Edward Arnold, London.

Rousseeuw, P.J. and van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, **Vol. 85**, No. 411.

Sellner *et al.* (1979). An investigation of the relation between composition, colour and furnace atmosphere in early glass (forest glass) by atomic absorption spectroscopy and electron spin resonance (ESR). *Reports on Glass Technology* **52(12)**, 59-89.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis.* London : Chapman and Hall.

Siotani, M. (1959). The extreme value of the generalised distances of the individual points in the multivariate normal sample. *Annals of the Institute of Statistical Mathematics, Tokyo*, **10**, 183-208.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. London : Chapman and Hall.

Wilks, S.S. (1963). Multivariate statistical outliers. *Sankhy$\bar{a}$*. 25, 407-426.