Three-dimensional Data Display Using Kernel Density Estimates.

Christian C. Beardah
Dept. of Mathematics, Statistics and Operational Research. The Nottingham Trent University
Clifton Campus, Nottingham NG11 8NS, U.K.
E-mail: c.beardah@maths.ntu.ac.uk

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

Introduction

Having looked at univariate and bivariate Kernel Density Estimates (KDEs) and their applications in previous CAA proceedings (Beardah and Baxter, 1996a and Beardah, 1998), it is natural to ask how three-dimensional KDEs may be utilised in archaeology. Data that are naturally three-dimensional do exist in archaeological applications; for example, we shall later look at the case of lead isotope ratio data. However, the utility of these methods is not restricted to such special cases. Often, higher dimensional data are analysed, by subjecting them to some dimension reduction technique, such as principal component analysis (PCA). KDE methods can then be applied to the first two or three components of the PCA scores, in an effort to identify structure.

For given trivariate data:

$$\underline{X}_1 = (x_1, y_1, z_1), \ldots, \underline{X}_n = (x_n, y_n, z_n)$$

a trivariate KDE is formed, by placing a four-dimensional "bump" at each data point. The value of the KDE, at any point $\underline{v} = (x, y, z)$ in space, is found by summing the "height" of bumps, that pass through the point $\underline{v}$. In the simplest terms, this is expressed mathematically as:

$$\hat{f}(x, y, z) = \frac{1}{nh_1h_2h_3} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_1}, \frac{y - y_i}{h_2}, \frac{z - z_i}{h_3} \right) \quad (1)$$

The shape of the bump is defined by the *kernel* function, denoted by $K(x, y, z)$. This is usually a trivariate probability density function (pdf) such as the trivariate normal pdf, given by

$$K(x, y, z) = (2\pi)^{-3/2} \exp(-\frac{1}{2}(x^2 + y^2 + z^2)).$$

Such functions have the property, that their volume is 1. The appearance of the KDE is not greatly influenced by the choice of kernel function (see Silverman, 1986, or Wand and Jones, 1995). We shall, therefore, use the trivariate normal, pdf, throughout this paper.

In equation (1), $h_1, h_2, h_3 > 0$ are called the *smoothing parameters* and control the amount of smoothing, in each of the three co-ordinate directions. The choice of these parameters can have a profound effect upon the appearance of the KDE, and hence, upon any conclusions drawn from an analysis, thereof. In the case of one-, and sometimes two-dimensional data, it is often possible to make a subjective choice of the smoothing parameters, at least as a basis for further refinement. However, as the dimensionality of the data increases, we, inevitably, become more reliant upon automatic, data-based choices. A simple method, of automatically selecting smoothing parameters in high dimensions (≥ 2), is to apply one of the several well-known univariate techniques (see Wand and Jones, 1995), to each of the variables in turn. Throughout this paper we have used this technique, individually applying the univariate, normal scale rule to each of the three variables. While, just as in the bivariate case, more general formulations than (1) exist, the problem of automatically selecting the smoothing parameters, in these (trivariate) cases, has not been fully addressed, and we do not consider them further, here. (See Wand and Jones, 1993 and Beardah, 1998 for details in the bivariate case.)

Examination of equation (1) immediately reveals a potential problem with trivariate KDEs. Namely, how do we display them? At each point, (x, y, z), in three-dimensional space, we have a corresponding estimated density value given by $\hat{f}(x, y, z)$, so we need *four* dimensions to display the KDE. One possible approach to this problem, is to use colour or shading to represent the density value, and to take three-dimensional slices through the four-dimensional density estimate. Figure 1(b) shows the results of such an approach, based upon a small dataset (n=8) for illustration.

An alternative, and in our view, better approach, is to form *percentage contour shells*, which for a given value, p, enclose the p% of the data, which is most dense. This is the three-dimensional implementation of a technique introduced by Bowman and Foster (1993) (see also Beardah and Baxter, 1996a, b and Baxter, et al., 1997, for some archaeological applications). In the case of two-dimensional data we have two-dimensional contour lines, which enclose the most dense p% of the data. For three-dimensional data, our contours are three-dimensional surfaces, which connect points with the same density. We refer to p, as the *level of inclusion* of the contour shell. For the small dataset, illustrated by the scatter plot of Figure 1(a) and Figure 1(c), a 75% contour shell is shown.

The formation of such contour shells is fairly straightforward. After the KDE is formed, we evaluate the estimated density at each data point $\underline{X}_1,\ldots,\underline{X}_n$. Data points are then ranked in decreasing order of density. A density value, d_p say, that is exceeded by the estimated densities of $p\%$ of the data points, is found and used to define the contour shell, with level of inclusion p. Mathematically, this contour shell is given by points such that

$$\hat{f}(x,y,z) = d_p .$$

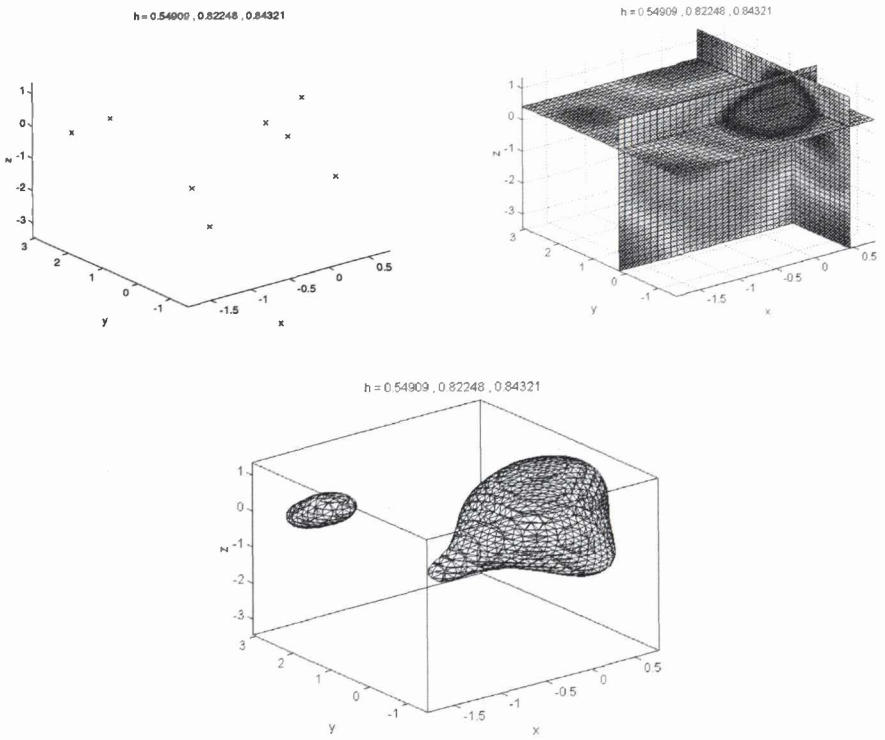

Figure 1. A small three-dimensional dataset (n=8) and visualisations of the corresponding trivariate KDE. Clockwise from the top-left: (a) a scatter plot of the data; (b) three-dimensional slices through the trivariate KDE, and (c) a 75% contour shell.

Motivational examples:

Example 1: Lead isotope ratio analysis.
Pollard and Heron (1996) discuss the use of lead isotope ratio analysis in archaeology. A sample from an ore-body, mined in antiquity, can be characterised by three lead isotope ratios. If n such samples are obtained, these can be used to estimate the lead isotope field for the ore-body, a three-dimensional construct, that delineates the isotopic compositional variation, within the ore-body. Isotopic compositions of artefacts can be compared with those, for sampled ore-bodies, to try and identify possible provenances (e.g. Sayre et al., 1992).

It is sometimes assumed that lead isotope fields have a trivariate normal distribution (Sayre et al., 1992), and it seems to be widely accepted that n = 20 is an acceptable value for statistical analysis (Pollard and Heron, 1996). Baxter and Gale (1998) and Baxter (1998) have cast serious doubt on the normality assumption, using data from ore-bodies, for which n > 20, and Westwood et al. (this volume) have shown that values of n, well in excess of 20, may be needed to detect quite clear non-normality. This last paper used univariate methods to demonstrate this; it is of some interest to investigate whether the direct use of 3-dimensional KDEs is helpful, in either detecting or displaying non-normality, for some of the data sets and sample sizes available.

Figure 2 shows 30%, 50%, 70% and 90% contour shells based upon the trivariate KDE, formed when n=59 ore samples, from the Lavrion field (Stos-Gale, et. al., 1996, Table 2), are considered. It is clear from Figure 2, that when using percentage contour shells as an exploratory technique, care needs to be taken to examine contours, based upon a variety of different levels of inclusion; otherwise, evidence supporting certain types of data structure may be missed. In this case, p values in the approximate range, 35 to 60, result in the appearance of two clear groups. Small p values (<35) reveal only the more prominent of the two groups. On the other hand, large p values (>60) yield contour shells, which change little in appearance as p is increased, and support the overall impression of highly non-normal structure.

A possible disadvantage of the contour shell approach, to the display of trivariate KDEs, is that without advanced graphics facilities, it is impossible to effectively display contour shells, corresponding to several p values on the same axes. Scott (1992) gives some examples of what can be done with

advanced graphics, in particular with transparent and/or partially "peeled" contour shells. However, such displays can sometimes be difficult to interpret, and we feel that the use of subplots, as in Figure 2, provides a simple and easily interpretable alternative. Another widely available tool, that can be used for investigating the appearance of contour shells, at various levels of inclusion, is animation. By stringing together a sequence of images, of contour shells at varying levels of inclusion we can obtain a smooth animation which can give a good impression of the structure, or lack thereof, exhibited by the data. In addition, "fly-by" or rotational animations can be used to view three-dimensional contour shells, from a variety of viewpoints.

An important question, in the context of this particular example, is whether the visual evidence seen in Figure 2 supports the hypothesis, that these data are from a normal population. We suggest that, to the contrary, the contour

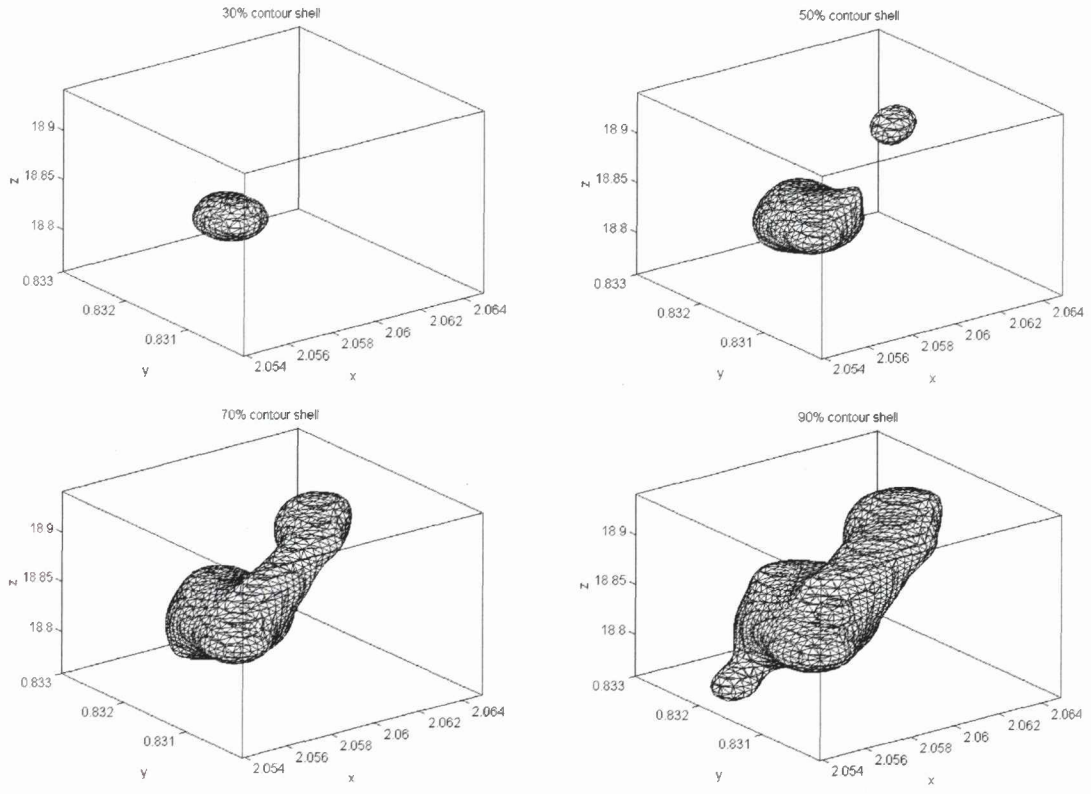

Figure 2. 30%, 50%, 70% and 90% contour shells based upon n=59 ore samples from the Lavrion field

shells indicate the presence of two separate clusterings, and certainly suggest non-normality.

Example 2: Glass composition data.

As a second example, data on the chemical composition of specimens, of early Medieval glass, from excavations at Southampton, extracted from Heyworth (1991), are used. A PCA was undertaken, using standardised values of eleven major/minor oxides, of the composition for (a) the whole dataset (n=271) and (b) those specimens, characterised by Heyworth as light blue or light green (n=227). In previous analyses of the dataset (a) (Beardah and Baxter, 1996b) bivariate component plots showed two main clusters in the data, that largely corresponded to the two colours. This grouping could also be seen in one dimension, using the ratio of iron to manganese in the glass (the effect of chemistry on colour is discussed in Heyworth, 1991). Iron, and the variables, that are highly correlated with it, dominated the first PC; manganese does not feature strongly on the first two PCs but dominates the third. It is thus of interest to see how strongly the pattern in the data was revealed, by looking at the first three, as opposed to two, PCs.

Contour shells, based upon the first three PCs of the whole dataset, showed, just as when the first or the first two PCs were used, the presence of two clear groupings. Of course, not all values of p revealed such structure, and, as mentioned previously, care needs to be taken when examining a succession of contour shells, at different levels of inclusion, possibly by making use of animation. In this case, p values in the range 43 to 63 resulted in contour shells, that split into two separate sub-shells, with no overlap. Small p values (<43) revealed only the more prominent of the two groups, largely associated with light blue glass. On the other hand, larger values of p (>63) yielded contour shells, that, while suggestive of two groupings, were no longer split into two separate sub-shells. Figure 3 shows an example of this behaviour. Here, p=65 was used. As p is increased further, the "bridge" between the two groups widens, however p values, as high as 75, are generally supportive of the existence of two groupings, within these data.

As stated above, Heyworth subjectively assigned a colour to each glass specimen. Figure 4 shows separate, 50% contour shells, based upon the first three PCs, of the n=62 glass fragments, identified as light green, and the n=165

fragments, identified as light blue. Clearly, Figure 4 supports the earlier observations, regarding these data. Complete separation between the two contour shells occurs for p values, as large as 65%. Furthermore, although some overlap occurs for higher levels of inclusion, the groups remain largely distinct, even for p values of about 80.

Trivariate simulations: sample size issues

The examples in the previous section illustrate that trivariate KDEs can provide a useful tool for exploratory data analysis and visualisation. However, KDEs do not simply display the data as given, but involve estimation of the structure of the population, from which the data are sampled - the better to see non-obvious patterns in the data. Like all statistical estimates, KDEs are subject to uncertainty, in the form of variance, and they are also subject to bias. In general, as sample size increases, the properties of the estimate improve. The question arises as to how large a sample is needed, for three-dimensional KDEs to be useful. In this section, we present a discussion of some experiments undertaken to

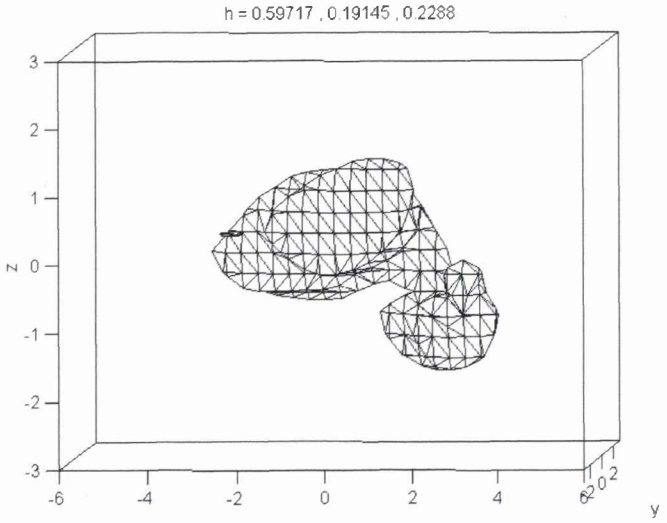

Figure 3. A 65% contour shell, based upon consideration of the whole glass assemblage (n=271).

investigate the practicality of the methodology, for realistic sample sizes and data structures.

As usual, in this type of experiment, we make use of normal mixture densities (NMDs). These are, simply, mixtures of several normal densities. The basic idea is to draw samples of varying size from NMDs, chosen to exhibit interesting structure, in this case, multi-modality. We then compare the structure of KDEs, based upon such samples, with the known structure of the population, from which the sample was drawn. For the purposes of this paper, the comparison is done on a purely subjective basis, by visually inspecting percentage contour shells, associated with KDEs of samples.

A trivariate NMD (denoted NMD_1), that simulates the perceived structure found in the Lavrion field data (see motivational example 1 above), has been created. NMD_1 is a bi-modal mixture of two trivariate normal densities. Figure 5 illustrates that KDEs, based upon samples of size n=60 from this NMD, exhibit similar structure to KDEs, based upon the data itself. We are interested in the following question:

- *In practice*, how big a sample size do we need, to reproduce the true qualitative (bi-modal) structure of NMD_1?

Or, looking at the problem from a slightly different perspective:

- *In practice*, how small does a sample from the standard trivariate normal density have to be, to fail to reproduce this normality (in the sense that the KDE is visually misleading and suggests multi-modality)?

The "curse of dimensionality" means that bigger samples are needed in three dimensions, than in the case of one- or two-dimensional data. For example, Silverman (1986) states that the sample sizes needed to achieve, in some sense, an acceptable error, when approximating a normal density by a KDE, are n_1=4, n_2=19, n_3=67, in one, two and three dimensions, respectively. Other authors, using different measures of "acceptable error", report findings which are similar in spirit, to those of Silverman (for example, see Scott, 1992). That is, as the dimensionality of the data increases, the sample size required to adequately reproduce the true structure of a population, also increases, and at a much faster rate.

Summary of simulation results

For sample sizes of n=20, 30, 45, and 60, respectively, samples were drawn repeatedly from NMD_1. Each sample was analysed by inspection of 20%, 40%, 60%, and 80% contours, and a subjective judgement was made, as to whether the sample exhibited uni-modal, bi-modal, or some other (usually multi-modal) structure. The results are summarised in Table 1. The nature of the investigation militates against a large number of repetitions. In spite of this, for each sample size shown in Table 1, 100 repetitions were made.

Sample size, n	Uni-modal structure	Bi-modal structure	Other structure
20	40	53	7
30	30	64	6
45	25	70	5
60	15	75	10

Table 1. Results obtained by repeated sampling from NMD_1. For each sample size, 100 samples were taken. The number of samples, exhibiting uni-modal, bi-modal, or "other" structures were recorded.

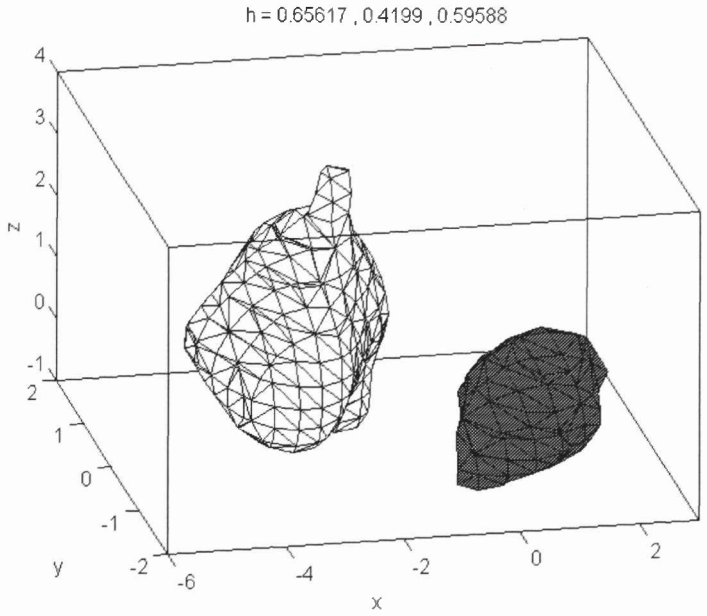

Figure 4. 50% contour shells, based upon separate consideration of the n=62 glass fragments identified as light green (light contour shell), and the n=165, identified as light blue (dark contour shell).

Sample size, n	Uni-modal structure	Bi-modal structure	Other structure
20	62	32	6
30	64	30	6
45	72	14	12
60	88	12	0

Table 2. Results obtained, by repeated sampling from the standard trivariate normal density. For each sample size, 100 samples were taken.

It can be seen from Table 1, that even samples of size n=45 and n=60 share the true bi-modal structure of the population, only 70-75% of the time. Also, as may be expected, as the sample size is reduced, samples become less successful in reproducing the true structure of the population. Indeed, samples of size n=20, truly reflect the structure of the population in only approximately 50% of cases.

For sample sizes of n=20, 30, 45, and 60, respectively, samples were drawn repeatedly from the standard trivariate normal density. The results are summarised in Table 2. Again, 100 repetitions were made for each sample size, shown in Table 2.

In this case, the true structure of the population is uni-modal. It can be seen from Table 2 that samples of all sizes share the true structure of the population, most of the time. Again, as the sample size is reduced, samples become less successful in reproducing the true structure of the population. It is noticeable, that the samples are generally more successful at reproducing the true structure of the population, in this case. However, samples of size n=20, still prove inadequate, for truly reflecting the structure of the population, in about 40% of cases. Tables 1 and 2, both suggest, that sample sizes, somewhat in excess of 60, are needed to have a high level of confidence, that the KDEs reflect the true structure in the data.

Summary and conclusions

In the case of univariate data, we have argued elsewhere (Beardah and Baxter, 1996a, b and Baxter, et. al., 1997), that KDEs provide a useful alternative to the histogram. For bivariate data, the case for using KDEs is even stronger, as the method has distinct presentational advantages over both,

two-dimensional histograms and the scatter plot. The statistical theory, that underpins the technique, can be used to provide guidance for appropriate amounts of smoothing, and powerful contouring methods follow easily, from the definition of the KDE, as a mathematical function.

In the case of three-dimensional data, many of the aforementioned advantages still apply. The only drawbacks are the increased computing power, required to cope with both, the computation and the graphical presentation of the output, and the well-known difficulties, associated with the presentation of three- and four-dimensional functions, on a two-dimensional screen. Bearing these difficulties in mind, we have found contouring to be the most successful method of visualising trivariate KDEs. We have illustrated that such methods can be effectively used, in exploratory data analysis. However, for each KDE, it is wise to look at contours with several levels of inclusion. In this respect, animations, showing how the percentage contour shells vary from low to high levels of inclusion, are a very useful tool.

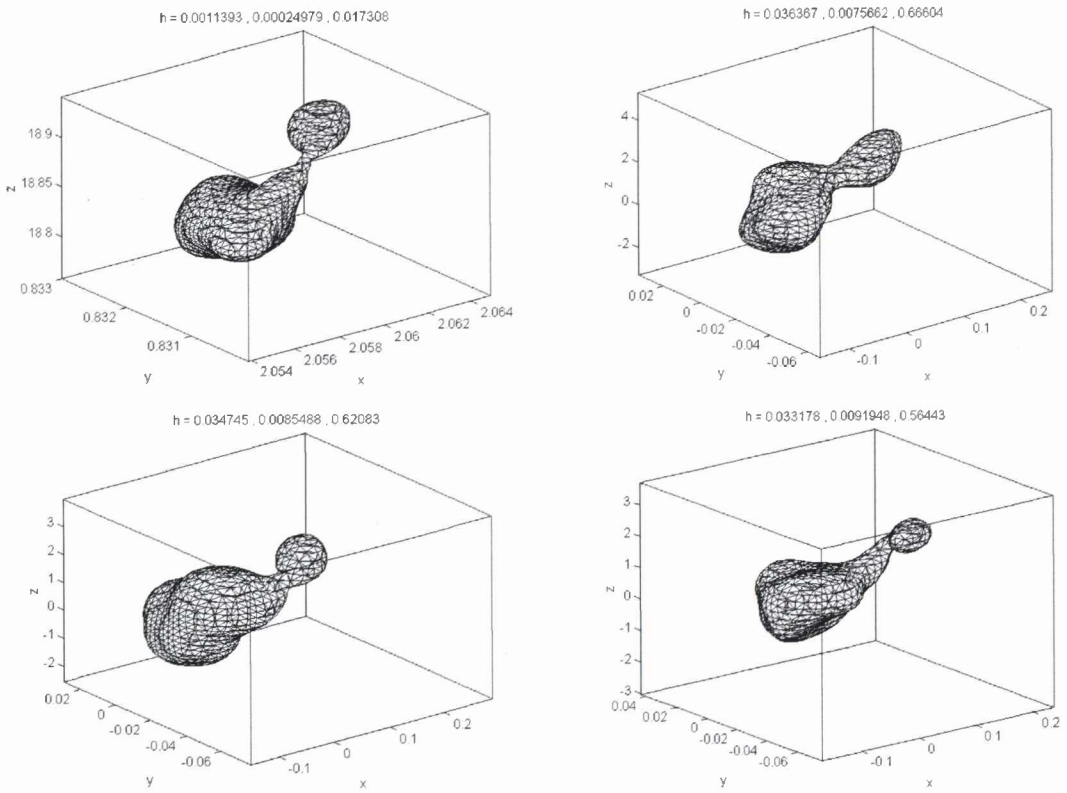

Figure 5. 60% contours, based upon the Lavrion field data (top left), and three samples of size n=60, from NMD_1.

The methods discussed here are not only applicable to naturally three-dimensional data, such as that found in lead isotope analysis. Three-dimensional KDEs are a useful tool, for the analysis of datasets, with dimension greater than three, provided that some type of dimension reduction technique (e.g., PCA) is used.

In the particular case of lead isotope ratio analysis, we have illustrated how KDEs, used as a purely exploratory technique, may provide visual evidence, that casts doubt on the assumption of normality, and can display the form of non-normality, if this is established by more formal approaches. Further, we have provided additional evidence, supporting the inadequacy of samples of size n=20. If the population from which the sample is drawn is non-normal (in this case bi-modal, in a form that can be modelled as a mixture of normal distributions, with some overlap between the components), then sample sizes, in the range 40 to 60, or larger, may be necessary, in order to reflect this structure.

Software

The techniques discussed in this paper have been implemented in the MATLAB package, by the first named author, and are freely available (email: christian.beardah@ntu.ac.uk). The routines include the facility to import and analyse the user's own data. All the illustrations were generated using this software.

Bibliography

BAXTER, M.J. (1998), "On the Multivariate Normality of Data Arising from Lead isotope fields", *Journal of Archaeological Science*, to appear.

BAXTER M.J. & GALE N.H. (1998), "Testing Multivariate Normality, with Applications to Lead Isotope Data Analysis in Archaeology", *Journal of Applied Statistics*, to appear.

BAXTER, M.J., BEARDAH C.C. & WRIGHT R.V.S. (1997), "Some Archaeological Applications of

Kernel Density Estimates", *Journal of Archaeological Science,* 24, pp. 347-354.

BEARDAH, C.C. (1998), "Uses of Multivariate Kernel Density Estimates in Archaeology", *Computer Applications and Quantitative Methods in Archaeology 1997*, Birmingham, to appear.

BEARDAH, C.C. & BAXTER M.J. (1996[a]), "MATLAB Routines for Kernel Density Estimation and the Graphical Presentation of Archaeological Data", *Analecta Prehistorica Leidensia 28, Interfacing the Past, Computer Applications and Quantitative Methods in Archaeology 1995*, Edited by H. Kammermans and K. Fennema, Leiden.

BEARDAH, C.C. & BAXTER M.J. (1996b), "The Archaeological use of Kernel Density Estimates", *Internet Archaeology*, 1, (http: //intarch.ac.uk/journal/issue1/beardah_index.ht.ml).

BOWMAN, A. & FOSTER P. (1993), "Density Based Exploration of Bivariate Data", *Statistics and Computing* 3, pp. 171-7.

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

POLLARD A.M. & HERON C. (1996), *Archaeological Chemistry*, Royal Society of Chemistry, Cambridge.

SAYRE E.V., YENER K..A., JOEL E.C. & BARNES I.L. (1992), "Statistical Evaluation of the Presently Accumulated Lead Isotope Data from Anatolia and Surrounding Regions", *Archaeometry,* 34, pp. 73-105.