Beyond the histogram - improved approaches to simple data display in archaeology using kernel density estimates

M.J. Baxter and C.C. Beardah†

Department of Mathematics, Statistics and Operational Research
The Nottingham Trent University, Nottingham NG11 8NS, United Kingdom
† email: mat3beardcc@uk.ac.ntu

1 Introduction

Histograms are among the most widely used methods of data presentation in archaeology. They are a particular example of a *density estimate* and their appearance depends on both the choice of origin of the histogram and the width of the intervals used. The origin is the lower boundary of the first interval in the histogram and it is assumed in what follows that the interval widths, also known as bin-widths, are equal. Whallon (1987) has expressed concern about this dependence, particularly since the choice of origin and bin-width may affect the archaeological conclusions drawn from a histogram.

As Orton (1988) noted, in commenting on Whallon's article, alternatives to the histogram that avoid some of their problems exist but have not been exploited by archaeologists. This remains the case. In another context Cleveland (1993) has suggested that for comparative purposes - a common archaeological application - histograms are usually inefficient, and better approaches exist.

In the present paper an effective alternative to the histogram, kernel density estimation, is discussed and illustrated. Silverman's (1986) book helped popularise the ideas involved and an up-to-date account is available in Wand and Jones (1995). Apart from our own work (Baxter and Beardah, 1995; Beardah and Baxter, 1995) we are not aware of applications of these ideas to archaeological data presentation.

Given the pervasiveness of the histogram in archaeological data presentation archaeologists ought to be interested in these ideas. That this does not seem to be so is, we suspect, a consequence of a lack of accessible software to implement the methodology. Additionally, there are technical problems, concerning the choice of bin-width, that have only been resolved relatively recently. One of us (CCB) has developed a comprehensive set of routines within the MATLAB package that can carry out kernel density estimation using the methodologies described in Wand and Jones (1995). The present paper describes and illustrates some of these ideas.

Our hope is that archaeologists will be convinced about the value of kernel density estimation for archaeological data presentation. While our software can be made freely available to anyone who is interested we recognise that not everyone will have access to our resources (hardware and software). It is likely, however, that the methodology that we describe will become increasingly available, and we hope that the present paper will encourage archaeologists to experiment with it when it does so.

In the next section we illustrate the ideas of kernel density estimation by example. Technical aspects are discussed informally in Section 3 and more mathematically in an appendix, where we concentrate on the potentially critical choice of bin- or window-width. Some further illustrative examples are given in Section 4. For simplicity of illustration the paper concentrates on univariate kernel density estimation, which is likely to be of most immediate interest to archaeologists. Extensions, including adaptive estimates, bounded estimates and two dimensional estimates are noted in the final section.

2 Initial example

To illustrate both the problems with histograms and their resolution using kernel density estimation Figure 1 may be consulted. This is inspired by the example of Whallon

(1987) and was produced using the S-plus package using a routine given in Venables and Ripley (1994, pp. 134-5). The histograms shown are based on the radii of 81 Danish Neolithic pots (Madsen, 1988, p. 18). In each case a bin-width of 1.5 is used; the origin of the first bin is 2.9 for the first histogram and increases by increments of 0.2 thereafter.

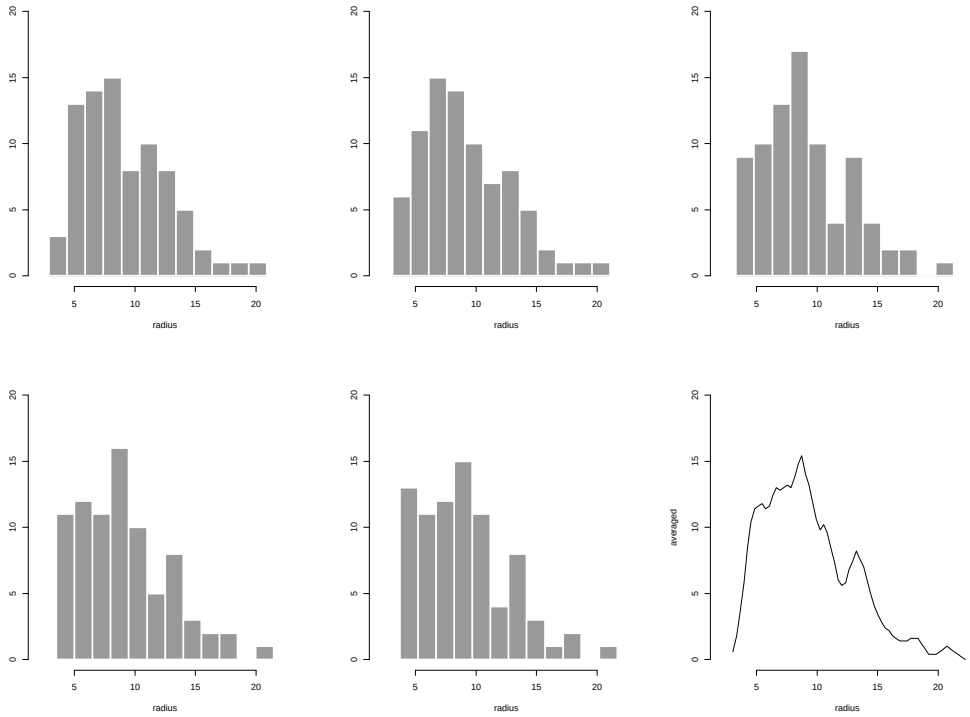

Figure 1: Five 'shifted' histograms and their average for the radii of 81 Danish Neolithic pots.

The appearance of the histogram is clearly sensitive to the choice of origin. For example, the second histogram appears to be skew and (almost) uni-modal, whereas there are relatively pronounced secondary modes in the final three histograms. In the final histogram the distribution is relatively uniform over the approximate range 4-11, contrasting with the clear peak in the third histogram for example.

The final diagram in the figure is the average of the five histograms, and is an example of an average shifted histograms (ASH) (Scott, 1992). In contrast to the histograms its appearance does not depend on a particular choice of origin. It does depend on on the choice of bin-width and, while smoother than the histograms is not as smooth as we would like.

The ASH can be regarded as an approximation to a kernel density estimate (KDE) and provides one way of calculating a KDE. At their simplest KDEs can be thought of as providing smoothed versions of histograms that are dependent on bin-width but not on choice of origin. The smoothness is an advantage from a presentational viewpoint and makes it easier to compare several histograms. The main problem in practice is to obtain a sufficiently smooth representation of the data while also retaining its main features. In Figure 1, for example, there appears to be a secondary mode at about 13 but this might disappear if a larger bin-width were used. The choice of bin-width is

clearly critical. In our previous work (e.g. Beardah and Baxter, 1995) we have chosen the bin-width (or window-width as it is called) subjectively, taking as a starting point an estimate that tends to over-smooth the data and systematically reducing this value thereafter. More recently we have incorporated a range of methods for automatic choice of window-width into the MATLAB routines that have been developed by the second author. The underlying ideas are described and illustrated in the next two sections.

3 Univariate kernel density estimates

It is possible to think of KDEs in different ways. In the previous section the KDE was presented as the average of a set of shifted histograms. An alternative viewpoint is as follows. Given n points $X_1, X_2, \ldots, X_n$ situated on a line a KDE can be obtained by placing a 'bump' at each point and then summing the height of each bump at each point on the X-axis. The shape of the bump is defined by a mathematical function, the kernel $K(x)$, that integrates to 1. The spread of the bump is determined by a window- or band-width, h, that is analogous to the bin-width of a histogram. The kernel is usually a symmetric probability density function. The KDE is usually insensitive to the choice of kernel so that in what follows the normal density function is assumed.

Mathematically, this gives the KDE as

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

where $\hat{f}(x)$ is an estimate of the density underlying the data.

Compared to the histogram the shape of $\hat{f}(x)$ does not depend upon the choice of origin, but is affected by the bandwidth h. Large values of h over-smooth, while small values under-smooth the data. In the examples in Baxter and Beardah (1995) the value of h was varied and the final estimate chosen subjectively. Following Wand and Jones (1995) our MATLAB routines have since been modified to allow a more 'objective' choice of h.

The idea underlying the options available is as follows. We want a kernel density estimate which is in some way 'optimal'. In other words we aim to choose a value of h which makes our KDE as 'close' as possible to the true underlying density, $f(x)$ (which is unknown). The measure of 'closeness' is the asymptotic mean integrated square error (AMISE) which can be shown to have the form

$$AMISE(\hat{f}) = \frac{1}{nh} A + \frac{1}{4} h^4 B. \qquad (1)$$

Fuller details are given in the appendix. The terms A and B in equation (1) are dependent on the known kernel, while B is also dependent on the integral of the squared second derivative of the unknown $f(x)$ that we shall denote by $R(f'')$. This last term can be thought of as measuring the 'roughness' of the underlying density.

Note that the two terms within equation (1) have opposite effects as h is varied. For small h the first term (connected with the variance of $\hat{f}$) is large while the second term (the bias) is small, and vice versa for large h. This illustrates the importance of h and also suggests that there is an optimal value of h which minimises the AMISE. This minimising value is easily shown to be

$$h_{AMISE} = \left[\frac{A}{nB}\right]^{1/5}. \qquad (2)$$

This expression, which through B depends upon the second derivative of the unknown density f, is the starting point for many methods for automatic selection of h.

If it is assumed that the true density f is normal, equation (2) can be used to get the so-called *normal scale* estimate of h_{AMISE} that we will denote by $\hat{h}_{NS}$ (see the appendix). This depends on the standard deviation of the underlying density, which must be estimated. The normal scale estimate provides a quick and simply calculated value of h which works reasonably well if the data are close to normal in structure. However, for non-normal data (exhibiting modes for example) $\hat{h}_{NS}$ will tend to be too large and hence over-smooth.

To avoid this problem one approach is to estimate the roughness of the true density, $R(f'')$, on which B depends and use this in estimating h via (2). This approach gives rise to a family of *direct plug-in* (DPI) estimates that are used in the examples of the next section. Details are given in the appendix.

The roughness, $R(f'')$, can be written in terms of the fourth derivative of f. An estimate of this depends on the sixth derivative, however; while an estimate of the sixth derivative depends on the eighth derivative and so on. These derivatives are unknown, but functions of them can be estimated by using the normal scale rule, for example.

If we choose to estimate (a function of) the sixth derivative and work backwards from there we get the so-called 1-stage direct plug-in (DPI-1) band-width estimator. If we begin by estimating a function of the eighth derivative and working back from there a 2-stage DPI estimate is obtained, and so on.

Repeatedly taking simulated samples from a known density can show the effect of varying the number of stages in the DPI method. Such studies show that while the average value of h so obtained becomes closer to the true optimal h_{MISE} value as the number of stages is increased, the variability h also increases. Again we have a trade off between bias and variance. Wand and Jones (1995) suggest that the 2-stage method provides the best compromise.

In the MATLAB routines written by the second author DPI-1, DPI-2 and DPI-3 estimates are available as well as the normal scale rule estimator. Related to the DPI estimates and also available is the *solve the equation* (STE) method. The essence of this approach is that an initial estimate of roughness is made, allowing an initial estimate of h. The associated estimate of f allows the roughness, and hence h, to be re-estimated. The process is repeated until h converges. Other approaches, based on the idea of cross-validation, are available and noted in the appendix.

4 Further examples

As a first illustration of the ideas of the previous section consider Figure 2. This is based on the rim diameters of 60 Bronze Age Italian cups. The data is given by Baxter (1994, pp. 233-4) and is a subset of material originally published by Lukesh and Howe (1978).

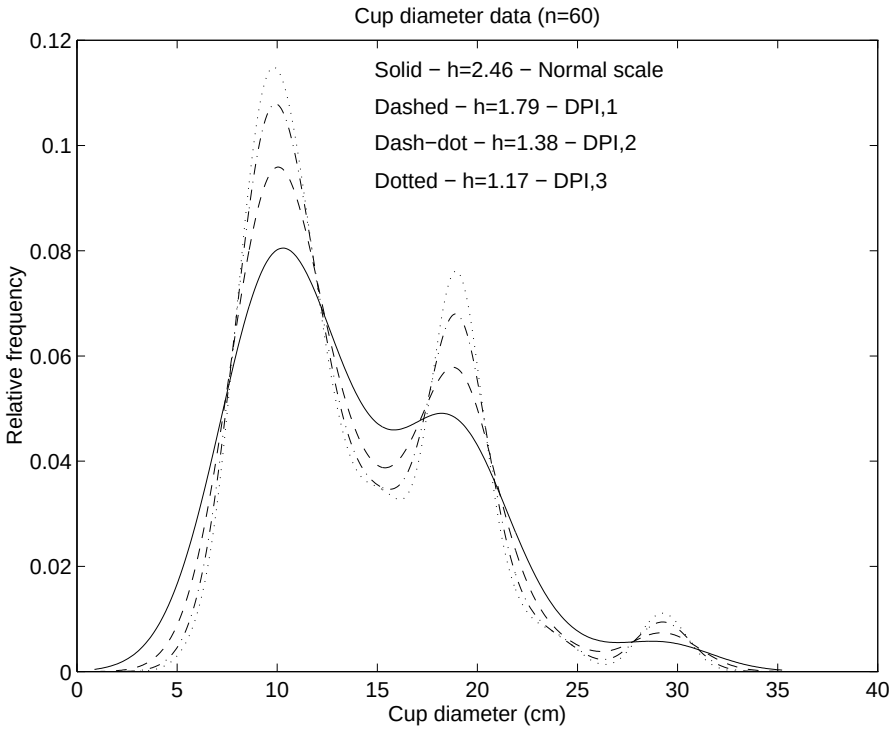

Figure 2: Four different h selection strategies generating KDEs based upon data representing the diameters of 60 Bronze Age cups from Italy.

Four estimates are shown with their associated values of h; these are the normal scale method and DPI-1, 2 and 3. These last three methods show that the data are clearly tri-modal, with the final mode a rather small one. As we move from DPI-1 to DPI-3 the tendency is for h to decrease and for the modes and troughs to be accentuated. The normal scale rule smooths the data too much so that the clear division between the first two modes is obscured and the third mode is missed.

To illustrate a type of use where KDEs may be particularly valuable consider Figures 3 and 4. Figure 3 shows histograms of the percentage content of calcium oxide in specimens of medieval French glass from four sites. The data are a subset of that given by Barrera and Velde (1989) and their site numbering is used. While it is of substantive interest to look at the distribution of calcium oxide, since this can reflect different glass-making traditions, the main point about the present example is that it typifies many such archaeological uses where histograms are used for comparative purposes.

To facilitate comparison scales are the same in each histogram. Absolute numbers have been plotted, though relative frequencies might equally well have been used. Comparisons reading down the page and within a column are straightforward; for example, it is clear that the calcium concentration of specimens from site 9 are typically greater than for site 2. Effecting comparisons reading across the page, or diagonally is much less straightforward, though it can be done, and difficulties of comparison would be magnified as the number of histograms increases (this is a small example compared with some in the literature).

A better, and potentially space-saving, way of effecting comparisons is shown in

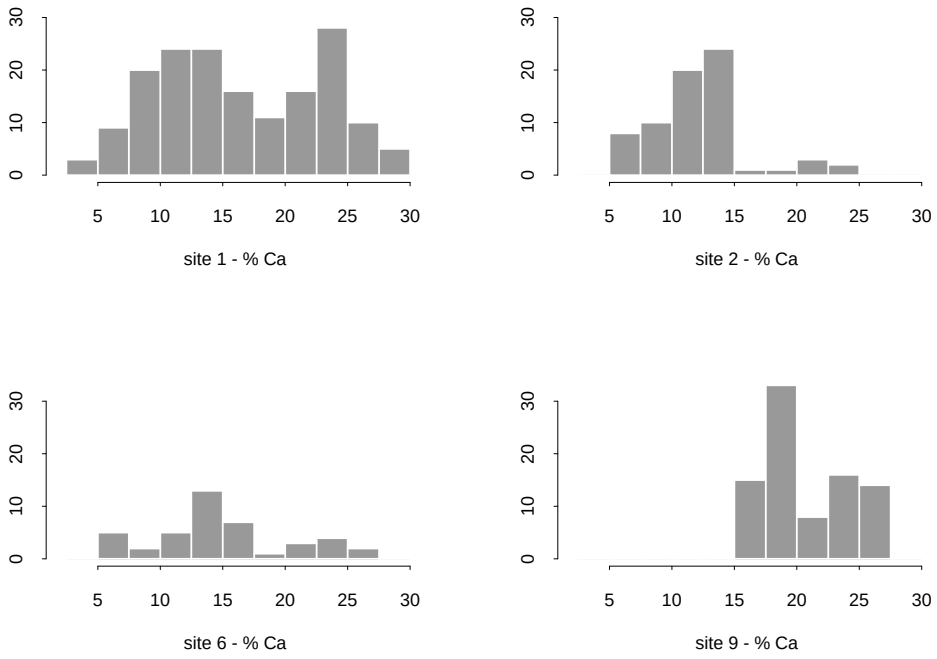

Figure 3: Histograms showing the distribution of calcium in specimens of glass from four French sites. Labelling corresponds to that in Barrera and Velde (1989) from which the data are taken.

Figure 4 where STE density estimates for the four histograms are superimposed (a relative frequency scale is also used). Comparisons are now much more direct and easy to make. Multi-modality is evident for some of the sites and the generally higher content of calcium in specimens from site 9 is obvious. The wider spread of values from site 1, which is also bi-modal, can be seen.

There is, of course, a limit to the number of estimates that can be usefully superimposed before the graph becomes too crowded. For exploratory work the colours available with MATLAB mean that more estimates can be superimposed and compared than one might wish to publish when restricted to black-and-white. We note also, in passing, that box-and-whisker plots are often used to effect comparisons between distributions (Cleveland, 1993). They are not suitable in the present case because not all distributions are uni-modal.

Finally, a confession. It was the original intention to use the Neolithic pot data from Section 2 as an illustrative example, but we are uncertain about how our results are to be interpreted. There is some suggestion from Figure 1 that these data are bi-modal. Figure 5 shows KDEs for these data for various choices of h. Except for the smallest value, which shows two modes and a number of bumps, a smooth, skewed, unimodal distribution is obtained. The various DPI and STE estimates give an h close to 1.4; that is, they suggest that the distribution is uni-modal. Either the apparent appearance of modes is spurious or else the automatic selection methods are over-smoothing. Assessing whether apparent modes are genuine is a separate problem to which kernel density estimation can be applied (Silverman, 1986, pp. 137-41) that we

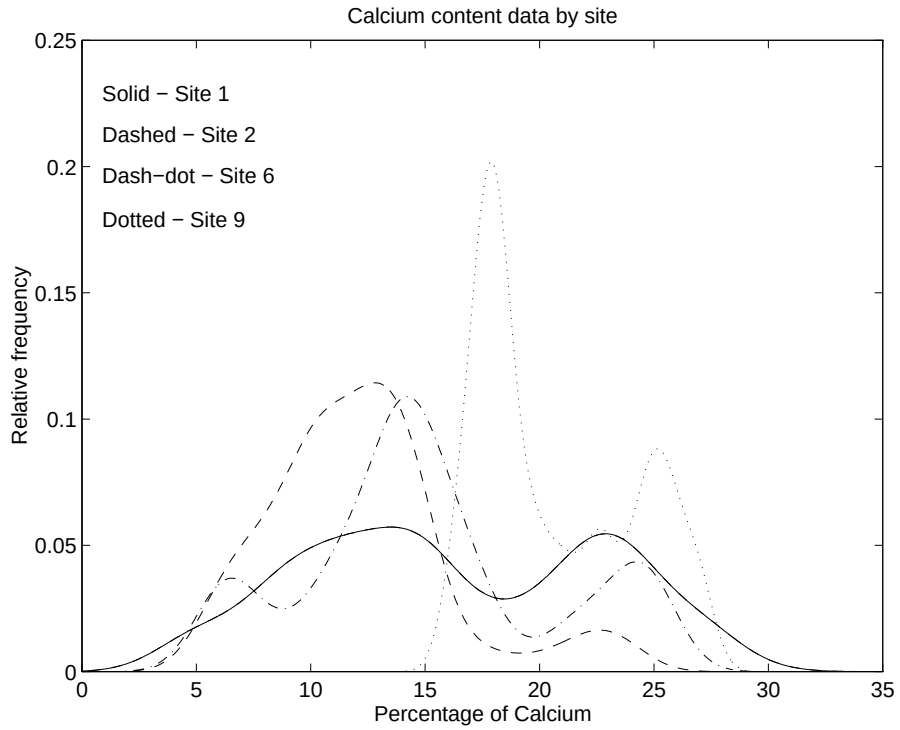

Figure 4: KDEs showing the distribution of calcium in specimens of glass from four French sites.

shall not pursue here.

5 Extensions

This paper has deliberately concentrated on univariate examples, which commonly arise in archaeology and where the advantages of kernel density estimates are, we hope, obvious. The MATLAB routines that have been developed can also handle bounded data where, for example, data are non-negative so that the KDE should be zero for negative values, and adaptive estimation (analogous to the use of variable bin-widths) where h can vary and is typically greater in less dense areas of the data space.

The most productive extension of the univariate KDE for archaeologists is likely to be to the bivariate case for which histograms, though occasionally presented, are unwieldy and difficult to interpret. The mathematical development is straightforward, although theory for the optimal choice of window-widths is less advanced than that for the univariate case. Baxter and Beardah (1995) have presented a successful application, based on bivariate plotting of the first two principal components from an analysis of glass compositions, in which the existence of three groups was clearly evident. The methodology is most useful for large data sets, where conventional two-dimensional plots are too dense for any patterns to be easily seen. Another potential area of application would be to the analysis of co-ordinates of finds, of the kind that are used in spatial k-means clustering for example (Baxter, 1994, pp. 148-9).

Baxter and Beardah (1995) also exploit the use of contouring, based on work by

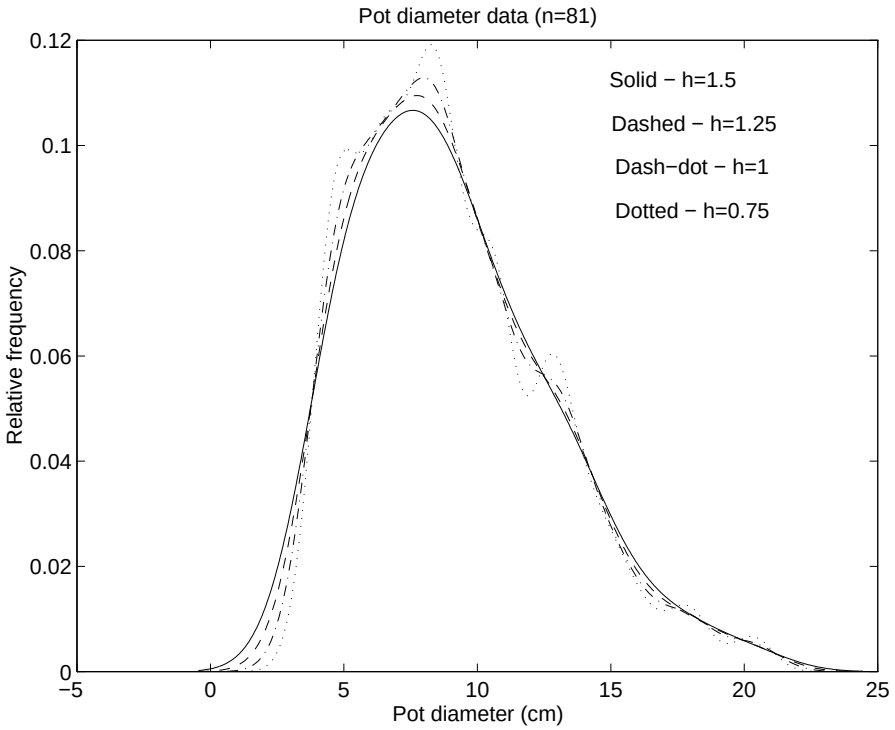

Figure 5: Four different values of h generating KDEs based upon data representing the diameters of 81 Danish Neolithic pots.

Bowman and Foster (1993). Bivariate kernel density estimates lend themselves naturally to contouring so that, for example, it is possible to highlight - for subsequent interpretation - the densest parts of a plot. If a data set can be divided into sub-groups, by context or period for example, it is possible to plot selected contours for each sub-group separately in order to examine their similarities and differences. An assessment of the potential of these possibilities for archaeology awaits the wider exploitation of the methodologies that we have outlined.

Bibliography

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

BAXTER, M.J. & C.C. BEARDAH 1995. 'Graphical Presentation of Results from Principal Components Analysis', in Huggett, J. & Ryan, N., (eds.), Computer Applications and Quantitive Methods in Archaeology 1994, International Series 600, Oxford. British Archaeological Reports: 63-67.

BEARDAH C.C. & M.J. BAXTER 1995. 'MATLAB routines for kernel density estimation and the graphical representation of archaeological data'. Nottingham Trent University Department of Mathematics Research report 2/95.

BOWMAN, A. & P. FOSTER 1993. 'Density Based Exploration of Bivariate Data'. *Statistics and Computing*, 3: 171-7.

CLEVELAND, W.S. 1993. *Visualising Data.* Hobart Press, New Jersey.

LUKESH, S.S. & S. HOWE 1978. 'Protoapennine vs. Subapennine: Mathematical Distinction Between Two Ceramic Phases'. *Journal of Field Archaeology*, 5: 339-47.
MADSEN, T, 1988. 'Multivariate statistics and archaeology', in Madsen, T. (ed.), *Multivariate Archaeology*: 7-27. Aarhus University Press, Aarhus.
ORTON, C.R. 1988. 'Review of Quantitative Research in Archaeology', Aldenderfer, M.S., (ed.), *Antiquity*, 62: 597-98.
SCOTT, D.W. 1992. *Multivariate Density Estimation.* Wiley, New York.
SCOTT, D.W., TAPIA, R.A. & THOMPSON, J.R. 1977. 'Kernel Density Estimation Revisited'. *Nonlinear Anal. Theory Meth. Applic.*, 1: 339-72
SILVERMAN, B. 1986. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.
VENABLES, W.N. & B.D. RIPLEY 1994. *Modern Applied Statistics with S-Plus.* Springer-Verlag, New York.
WAND, M.P. & M.C. JONES 1995. *Kernel Smoothing.* Chapman and Hall, London.
WHALLON, R. 1987. 'Simple Statistics', in Aldenderfer, M.S., (ed.), *Quantitative Research in Archaeology: Progress and Prospects*: 135-50. Sage, London.

6 Technical appendix

Several methods of objectively estimating h are discussed by Wand and Jones (1995) and are implemented in our MATLAB routines. A brief technical account of some of these methods, not all of which are illustrated in the main text, follows.

6.1 Normal scale selection of h

The asymptotic mean integrated square error (AMISE) in equation (1) is given more fully by

$$AMISE(\hat{f}) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f''),$$

where

$$R(K) = \int_{x\in\Re} K(x)^2\,dx = 1$$

and

$$\mu_2(K) = \int_{x\in\Re} x^2K(x)\,dx = 1/(2\sqrt{\pi})$$

for the normal kernel.

The AMISE is a large sample approximation to the less easily manipulated mean integrated square error (MISE),

$$MISE(\hat{f}) = E\int(\hat{f}(x) - f(x))^2\,dx. \tag{3}$$

The minimising value of equation (2) is written more fully as

$$h_{AMISE} = \left[\frac{R(K)}{\mu_2(K)^2R(f'')n}\right]^{1/5}.$$

Assuming that the true density f is normal with variance σ^2 leads to the *normal scale* estimate of h_{AMISE}, given by

$$\hat{h}_{NS} = \left[\frac{8\sqrt{\pi}R(K)}{3\mu_2(K)^2n}\right]^{1/5}\hat{\sigma},$$

where $\hat{\sigma}$ is an estimate of σ, the unknown standard deviation of f. If the normal kernel is used this reduces to the simple formula

$$\hat{h}_{NS} = 1.06n^{-1/5}\hat{\sigma}.$$

6.2 Plug-in selection of h

Plug-in methods represent an improvement over the normal scale rule since they estimate $R(f'')$ for the true density rather than assuming that the true density is normal. To see how this is done we note that $R(f'')$ can be written as

$$R(f'') = \psi_4 = \int_{x\in\Re} f^{(4)}(x)f(x)\,dx$$

where $f^{(4)}(x)$ is the fourth derivative of $f(x)$.

This means that equation (2) can be written as

$$h_{AMISE} = \left[\frac{R(K)}{\mu_2(K)^2 \psi_4 n}\right]^{1/5} \tag{4}$$

and hence that h_{AMISE} can be approximated accurately if a good estimate of ψ_4 is available. Since $\psi_4 = E(f^{(4)}(X))$, a natural estimator for ψ_4 is

$$\begin{aligned} \hat{\psi}_4(g) &= n^{-1} \sum_{i=1}^{n} \hat{f}^{(4)}(X_i), \\ &= n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{g} K^{(4)} \left(\frac{X_i - X_j}{g}\right), \end{aligned} \tag{5}$$

where $\hat{f}$ is a KDE based upon a smoothing parameter g. In addition to providing more detail on the above, Wand and Jones (1995) show that the optimal value of g (minimising the mean square error) is given by

$$g_{AMSE} = \left[\frac{2K^{(4)}(0)}{-\mu_2(K)\psi_6 n}\right]^{1/7}. \tag{6}$$

Unfortunately, this expression for g depends upon

$$\psi_6 = \int_{x \in \Re} f^{(6)}(x) f(x)\, dx.$$

which in turn depends upon the unknown density f. Furthermore, estimating ψ_6 by $\hat{\psi}_6(g)$ (defined analogously to (5)) will not help as its optimal band-width g (defined in a manner similar to (6)) depends upon ψ_8, and so on.

We can get around this problem by initially using a quick and simple h selection strategy such as the normal scale rule to estimate g and hence calculate ψ_6. By assuming that f is normal with variance σ^2, it easily shown that

$$\psi_r = \frac{(-1)^{r/2} r!}{(2\sigma)^{r+1} (r/2)! \sqrt{\pi}}. \tag{7}$$

where r is even. Therefore, the so-called 1-stage direct plug-in (DPI) band-width selector consists of

1. Estimating ψ_6 via (7) with a suitable estimate, $\hat{\sigma}$ of σ.
2. Evaluating g_{AMSE} via (6).
3. Estimating ψ_4 via (5), where $g = g_{AMSE}$.

The DPI technique can be extended to 2 stages by initially estimating ψ_8 with a normal scale rule (i.e. using equation (7)), followed by estimation of ψ_6 and finally ψ_4 (see Wand and Jones (1995), section 3.6 for details). The extension to a still higher number of stages is obvious. Also, note that the 0-stage DPI method (where (7) is used to directly estimate ψ_4 and (4) to estimate h) is equivalent to the normal scale rule.

6.3 Solve the equation methods for the selection of h

These methods are similar to the DPI approach, but take one step further by creating an iterative process with the 'optimal' h value as the solution of the iteration. The simplest such method is that of Scott, Tapia and Thompson (1977) which again takes equation (2) as its starting point. Specifically, equation (3) can be written as

$$h = \alpha(K)\beta(f)n^{-1/5}.$$

In other words a product of two terms, the first of which is a function of the kernel only and a second term which depends upon (the second derivative of) f. If we guess a value, h_0, for the band-width and use it to form a KDE, $\hat{f}_0$, then we can calculate a new, hopefully better, value h_1 via

$$h_1 = \alpha(K)\beta(\hat{f}_0)n^{-1/5}.$$

It is clear that this idea can be extended to an *iteration* whereby

$$h_{i+1} = \alpha(K)\beta(\hat{f}_i)n^{-1/5}$$

for $i = 0, 1, \ldots$ and h_0 is an initial guess for the band-width. Each KDE, $\hat{f}_i$, is calculated using band-width h_i.

More sophisticated *solve the equation* (STE) rules are explained in Wand and Jones. Essentially, such methods are based upon an iteration of the form

$$h_{i+1} = \left[\frac{R(K)}{\mu_2(K)^2\hat{\psi}_4(g(h_i))n}\right]^{1/5}. \tag{8}$$

where the band-width g used to calculate $\hat{\psi}_4$ is a function of h_i. Put simply, these methods take a value of h_i (initially a guess h_0), use h_i to calculate a value of g which is used to calculate a new, hopefully better, value of h, h_{i+1} via (8). The process is repeated *to convergence*, that is until h_i and h_{i+1} are sufficiently close together.

6.4 Cross validation methods for the selection of h

These methods attempt to minimise an estimate of the MISE of $\hat{f}$ (equation (3)). Equation (3) can be expanded as

$$MISE(\hat{f}) = E\int \hat{f}(x)^2\,dx - 2E\int \hat{f}(x)f(x)\,dx + \int f(x)^2\,dx. \tag{9}$$

The third term in (9) does not depend upon h, so that minimising the MISE can be acheived by minimising

$$E\int \hat{f}(x)^2\,dx - 2E\int \hat{f}(x)f(x)\,dx.$$

The following can be shown to be a good estimator of this quantity:

$$L(h) = \int \hat{f}(x)^2\,dx - 2n^{-1}\sum_{i=1}^{n}\hat{f}_{-i}(X_i), \tag{10}$$

where

$$\hat{f}_{-i}(X_i) = (n-1)^{-1} \sum_{j \neq i}^{n} \frac{1}{h} K\left(\frac{X_i - X_j}{h}\right).$$

This latter quantity is an estimate of the density at the data point X_i, where the density estimate is based upon the whole data set *except* X_i.

This so-called *least squares cross validation* method proceeds by finding the value of h which minimises the *objective function* L defined in (10). Since each evaluation of the function L involves heavy computational expense this method has to be carefully implemented. Other cross validation methods have been proposed and are discussed and compared in Wand and Jones. The main difference between such methods is the form and ease of calculation of the objective function.

6.5 Conclusions

We have seen that the DPI methods essentially provide an explicit formula for an 'optimal' band-width value. Solve the equation methods find h as the solution of an iteration, while cross-validation (CV) techniques rely upon minimising an often expensive to calculate objective function. Therefore it is no surprise that in terms of ease of implementation, the DPI and STE methods are rather more convenient than CV methods. In terms of performance Wand and Jones recommend 2-stage DPI or STE methods as generally better than CV variants, with the exception of the *smoothed* cross validation method (which is in some ways a hybrid between DPI and CV). All of these techniques have been implemented in our MATLAB routines so direct comparisons can be made.