

Non-parametric calibration for age estimation

D. Lucy¹, R.G. Aykroyd² and A.M. Pollard³

^{1,3} Department of Archaeological Sciences, University of Bradford, Bradford, BD7 1DP

² Department of Statistics, University of Leeds, Leeds, LS2 9JT

Abstract

A method is proposed for the calibration of a continuous random variable when the dependent variables are a combination of continuous and categorical, and the model between the controlling variables and calibrated variable is empirically derived. The various probability distributions are estimated from training data using kernel density procedures with bivariate normal kernels for continuous variables and uniform smoothing for discrete variables. Bayes theorem is then used to produce the posterior distribution from which point estimates and estimates of confidence may be made. Individual posterior densities allow each case to be considered separately and cases with conflicting evidence can easily be identified for further investigation. This approach is illustrated using part of a data set of human adult teeth from individuals of known age. Estimates from the proposed method show less bias than those from the widely used multiple regression. This allows more accurate reconstruction of the age distributions of ancient populations. In particular bias reduction is most notable at the extreme ages, which also tend to be the least frequent, thereby widening the age distribution. This will allow more reliable consideration of archaeological and anthropological questions relating to, for example, maximum life span, age related social structure and the development of age related disease.

KEYWORDS: Bayesian estimation; empirical Bayes; kernel density; regression; smoothing.

1 Introduction

Human age estimation has been a long standing area of interest in anthropology. Early uses date from as far back as the 1840's where, following the factories act, a reliable method for

¹Now at the Centre for Forensic Statistics and Legal Reasoning, Department of Mathematics and Statistics, University of Edinburgh

finding whether young adults were old enough to work was required. More recently forensic scientists use skeletal age estimation techniques as an aid to the identification of human remains, and human biologists to chart the development of the human lifespan in relation to the living environment.

Only in modern populations are accurate records of births and deaths available and hence in general age cannot be measured directly, but instead a process of calibration must be used. Calibration is the general name given to any process by which the value of one variable, which cannot be easily observed, can be estimated through knowledge of the value of a related variable which can be observed easily. For example, forensic scientists have used calibration to estimate the age at death for human remains from a whole range of age related changes in the hard tissue of the skeleton and teeth (for a general introduction to human skeletal age markers see Iscan, 1989).

Previous approaches to age estimation have either treated all variables as continuous and employed linear regression or treated all as discrete and used histogram based approaches. The technique introduced here is a development of one proposed by Lucy *et al.* (1996) where both age and age indicator were treated as ordinal variables. The proposed method treats age as a continuous variable and is capable of using continuous, ordinal, and even nominal age indicator variables. The overall approach is nonparametric and can be seen as part of a wider movement to overcome problems of non-normality and non-linearity inherent in applied sciences by the use of kernel density methods (eg Baxter and Beardah, 1997).

2 Age-related variables

Age is easy to estimate for children and adolescents as the skeleton and teeth are undergoing rapid genetically determined development. For example, by observing the eruption sequence of teeth, relative to the expected dental development, age can be reliably estimated to within a year. Adults, however, are more problematic. After bone and teeth are fully developed time related change slows down. Change in the skeleton tends to be most obvious in areas which remodel through some physical disorder, such as in a degenerative joint disease like osteoarthritis. Some, such as the combined pelvis and hip bones, undergo structural change without being subject to non age-related forces, and these have proven to be more reliable indicators of adult age (Iscan, 1989).

Human teeth also change with age, and these changes are often more reliable than skeletal age changes. Adult teeth are largely ‘dead’ and not subject to the intense remodelling due to use and physical illness. There are three main ways in which the fully developed adult human tooth can change structurally: general wear, changes to the dentine and changes

to the root due to the continuous repositioning of the tooth (see Kilian in Iscan 1989, and Figure 1).

Most of the skeletal and dental changes in the adult human have an underlying structure which is continuous. However, for many of them it is difficult, or even impossible, to make observations which can be quantified on a continuous scale. For example, one of the joints of the pelvis undergoes a complex series of continuous surface texture and surface shape changes. Although it is possible to arrange a series of these surfaces in an age sequence, it is not possible to assign any meaningful description to a single joint surface seen individually. Biological anthropologists have traditionally approached this problem by using a limited number of broad groupings, usually four or five, into which any observation is placed.

A more internally consistent scheme was devised by Gustafson (1950) for application to the problem of dental age changes. An age indicator variable such as wear was classified into four stages giving easily identified numerical classes. Categories were defined when no wears was present, some wear of the enamel was evident, if the enamel had been worn through and finally if the pulpal chamber had been penetrated. This approach was also applied to five other age related changes.

3 Current calibration methods in age estimation

Many approaches to calibration have been used in forensic science. Since most age indicators are classified in categories a form of nearest neighbour analysis has been commonly used. The estimate of age is given by the age of the closest match in a training dataset, usually following some cross-tabulation procedure. In this case the estimate is taken to be the mean age for comparable values, and the error term derived from the variance of the same subset (Ferembach *et al.* 1980). This type of nearest neighbour analysis is in principle unproblematic as it can be viewed as a type of multiple classical calibration. However, practical problems can occur as training sets tend to be small, and if more than a couple of age indicators are used many cells in the cross-tabulation have very few members. When these methods are used with datasets with known ages the mean deviation is of the order of fifteen years (Lucy and Pollard, 1995).

Inevitably regression methods have been applied to age estimation. Gustafson (1950) summed six age indicators, each scored in four categories, which was then treated as the explanatory variable in an inverse calibration. Bang and Ramm (1970) used non-linear inverse calibration to model the relationship between root dentine translucency and age, and Aiello and Molleson (1993) used linear inverse calibration to model bone osteon counts against age. Later Johanson (1971) employed multiple regression using refinements to Gustafson's six age

indicators.

Such regression methods, however, have a set of properties which are undesirable. Foremost is the fact that almost universally inverse calibration is used which treats age as the response variable. The nature of the problem, however, means that changing age causes change in the indicator variables and hence age must be the explanatory variable. Unfortunately inverse calibration inevitably leads to substantially biased estimates for age (Eisenhart, 1939; Aykroyd *et al.*, 1997) as the correlation between age and indicator is usually quite low ($r < 0.70$). This bias means that the age of young individuals is overestimated and the age of old individuals is underestimated. Reconstructed age distributions will be too narrow and conclusions involving the age extremes, such as maximum lifespan, will be biased. To remove this bias Lucy and Pollard (1995) recommended the use of classical calibration where age is correctly treated as the explanatory variable. However, it is more difficult to adapt this approach to situations where the indicators are multivariate, and are mixed continuous and ordinal.

4 A framework for nonparametric calibration

4.1 Bayes and empirical Bayes estimation

Suppose interest is in estimating the unknown values of a single continuous variable X (age in this example) given the values of m indicator variables $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_m\}$. In age estimation these will typically be a mixture of continuous and discrete. If a Bayesian approach is adopted, to allow prior knowledge of X to be incorporated, then the posterior distribution, $f(x|\mathbf{y})$, of X given all m indicator variables is found from Bayes' theorem

$$f(x|\mathbf{y}) = \frac{f(\mathbf{y}|x)f(x)}{f(\mathbf{y})}, \quad (1)$$

where $f(\mathbf{y}|x)$ is the m -dimensional joint likelihood of the indicators given X , and $f(x)$ is the prior distribution of X . The probability function $f(\mathbf{y})$, a mix of continuous and discrete variables, ensures proper normalisation and can be found, from the numerator, by integrating over the continuous variables and summing over the discrete variables. In practice this is usually done by numerical approximation. A standard approach would now involve assuming parametric forms for each of the distributions in Equation (1) such as multivariate normal distributions. In this paper, however, an empirical and nonparametric approach is proposed.

The selection of a suitable prior distribution can be both difficult and controversial. It is usual to base the choice of prior on some *subjective* expert opinion, or by selecting a conjugate prior for computational simplicity. The alternative approach of *empirical Bayes*, uses observed data values to determine the prior. This may be in the form of previous

experiments, or sometimes may involve using data twice in the likelihood and the prior. The term “empirical Bayes” is not universally accepted. Some believe using “empirical” is misleading (eg Gelman *et al.*, 1995, p123) in that it suggests other Bayesian approaches are not “empirical”, whilst others believe that estimation of the prior is against “Bayesian philosophy” (Carlin and Louis, 1996, p. 38). For discussions on empirical Bayes see, for example, Carlin and Louis (1996) and Maritz and Lwin (1989).

For forensic age estimation one could select a prior that reflects the whole current population which can be taken from mortality data. If other information such as sex can be determined then the distribution can be taken from the age distribution for that particular sex. For individuals from past populations prior selection can be more difficult as it is expected that a historical age structure will not reflect a modern one. Hence a non-informative prior might be better for archaeological applications. In Section 5 two approaches will be considered: (i) a non-informative prior and (ii) an empirical prior distribution derived from training data. It should be noted, however, that many other choices of prior are possible and can easily be incorporated into the proposed framework.

The full likelihood in Equation (1) requires knowledge of an m dimensional conditional distribution. Unless m is small, eg 2 or 3, then this is impractical and some simplifying assumption is necessary. In particular if estimating a standard multivariate normal using a kernel density approach with a normal kernel, then a sample of size 67 is required in 3 dimensions; this rises to 223 in 4 dimensions (Silverman, 1986, reproduced in Webb, 1999, Tables 3.4).

If it is reasonable to assume conditional independence of the Y_i 's given X then the joint likelihood in Equation (1) can be written in the form

$$f(\mathbf{y}|x) = \prod_{i=1}^m f(y_i|x) \quad (2)$$

where $f(y_i|x)$ is the individual univariate conditional distribution of indicator Y_i given X . Details of this, *naive or independence Bayes* approach, are given in Webb (1999, Section 3.2). Adopting a histogram approach then reduces the number of cells from k^m (where k is the number of cells per variable) to mk .

If full conditional independence is not a reasonable assumption then the partial independence method of Chow and Liu (1968) using *maximum weight dependence trees* can be used to capture some of the dependency without resort to the full joint distribution (see Webb 1999, Section 3.2). The joint likelihood in Equation (1) is then written as

$$f(\mathbf{y}|x) = \prod_{i=1}^m f(y_i|y_{j(i)}, x) \quad (3)$$

where $y_{j(i)}$ is the *parent* of y_i in a conditional independence tree whose root y_l is chosen

arbitrarily and $f(y_l|y_{j(l)}, x) = f(y_l|x)$. This representation has a total of $k(k-1)(m-1)+k-1$ (Webb 1999, p. 60) parameters, that is of the order mk^2 .

The tree is constructed by calculating the *branch weights* (eg, correlation or information) for all pairs of variables. Then starting with *nodes* only, add the branch with the greatest (absolute) weight. Continue adding branches using the next greatest weight, but discarding cases which create cycles. Once all branches have been considered, stop and choose an arbitrary node as the root. See Section 5 for an example.

4.2 Non-parametric density estimation

Smoothing of continuous variables

In order to perform the above estimation it is necessary to have estimates of the various conditional and marginal densities. We have chosen to use an empirical and nonparametric approach using training data $\{(x_i, y_i) : i = 1, \dots, n\}$. The simplest approach is to estimate a distribution by the appropriate histogram (Lucy *et al.*, 1996), but this is prone to instability particularly for small training sets. To reduce the effects of noise on these estimates smoothing using kernel density methods is proposed.

Starting with the joint distribution, $f(x, y)$ with both X and Y continuous, the kernel density estimate is given by:

$$\hat{f}(x, y; H) = \frac{1}{n} \sum_{i=1}^n K_H(x - x_i, y - y_i) \quad (4)$$

where K is a bivariate kernel function and H a symmetric positive definite 2×2 *bandwidth* matrix. Here the kernel chosen is the bivariate Gaussian,

$$K_H(x, y) = \frac{1}{2\pi|H|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} [x, y] H^{-1} \begin{bmatrix} x \\ y \end{bmatrix} \right\}. \quad (5)$$

The general bandwidth matrix has three independent parameters, h_x^2 , h_y^2 and h_{xy} , defining the variances of X and Y , and the correlation between X and Y . This means that the kernel will have elliptical contours with alignment determined by h_{xy} .

An obvious simplification is to have H a diagonal 2×2 matrix, $H = \text{diag}(h_x^2, h_y^2)$. This will constrain the kernel to have contours aligned with the coordinate directions. The final simplification making both variances equal, $H = h^2I$, is unreasonable for this type of example where the variables are measuring different quantities and therefore have quite different scales. The choice of which type of bandwidth matrix to use comes down to a balance between flexibility and complexity, and will be very problem dependent.

Next consider the joint distribution, $f(x, y)$ with X continuous and Y discrete. In this section only smoothing of the continuous variable will be considered; see the next section for

discrete smoothing. So the appropriate kernel density estimate is given by:

$$\hat{f}(x, y; H) = \frac{1}{n(y)} \sum_{i=1}^n K_{h_x}(x - x_i) \delta(y - y_i) \quad (6)$$

where K is a univariate kernel function with single *bandwidth* parameter h_x^2 , $\delta(u)$ is an indicator function with $\delta(u) = 1$ if $u = 0$ and $\delta(u) = 0$ if $u \neq 0$, and $n(y) = \sum_{i=1}^n \delta(y - y_i)$ the number of observations with value y . Again a Gaussian kernel is chosen:

$$K_{h_x}(x) = \frac{1}{\sqrt{2\pi h_x^2}} \exp\left\{-\frac{1}{2} \frac{x^2}{h_x^2}\right\}. \quad (7)$$

Finally, the marginal distribution, $f(x)$, of continuous variable X is simply estimated by the univariate kernel density estimate:

$$\hat{f}(x; H) = \frac{1}{n} \sum_{i=1}^n K_{h_x}(x - x_i) \quad (8)$$

where K and h_x^2 are as above.

An appropriate choice of the bandwidth matrix H must now be considered. It can be shown that if the variables form a bivariate normal distribution then the optimal H , based on asymptotic mean integrated squared error (AMISE), is proportional to the population covariance matrix with constant of proportionality which only depends on the sample size (p. 106, Wand and Jones, 1995). Hence a sensible approach, called *sphering*, is to estimate H using the sample covariance matrix, S . Here a rule based on the normal reference rule (p. 152, Scott, 1992) is proposed with $H = (4/(d+2))^{1/(d+4)} S n^{-1/(d+4)}$, where d is the dimension (here $d = 2$), S is the sample covariance matrix and n is the sample size. For details of other ways to estimate the bandwidth matrix, such as cross-validation, see, for example, Scott (1992), Silverman (1986) or Wand and Jones (1995).

Smoothing of discrete variables

When dealing with categorical variables the idea of smoothing between values is not at all meaningful. Even for ordinal variables it may not always be appropriate particularly if the values represent very irregular (and unquantified) differences. Hence the methods described in the previous section will often be inappropriate and so an alternative approach is considered.

In the univariate case let $h(x_i) : i = 1, \dots, k$ be the histogram frequency data with $\sum h(x_i) = N$ the total number of observations. Instead of the interpolating-type kernel approach, mixing with a discrete uniform distribution (Titterton, 1980) will be considered with form

$$\hat{f}(x_i) = \omega \frac{h(x_i)}{N} + (1 - \omega) \frac{1}{k} \quad 0 < \omega < 1, \quad (9)$$

where ω is a mixing proportion measuring the belief in the data. With ω close to one the prior belief of uniformity is virtually abandoned in favour of the data. Whereas for ω close to zero the evidence from the data is ignored giving full weight to the uniform distribution.

An alternative form of smoothing can be thought of as the result of adding additional observations divided equally between the categories in the histogram which leads to the form

$$\hat{f}(x_i) = \frac{h(x_i) + \alpha/k}{N + \alpha} \quad \alpha > 0 \quad (10)$$

where α is the smoothing constant. With no added observations ($\alpha = 0$) the unadjusted data histogram is obtained, but as more and more *phantom* observations are added equally, the histogram becomes more and more uniform. With $\alpha = 1$ this approach was proposed by Titterington *et al.* (1981), however with the above generalisation fine-tuning of the smoothing can be achieved. It is easy to show that the above forms are in fact equivalent with $\alpha = (1 - \omega)/\omega$.

Using a minimum mean-squared error approach Titterington (1980) shows that values for ω can be found by:

$$\omega = \frac{C}{\{C + N \sum (h(x_i)/N - 1/k)^2\}} \quad (11)$$

where $C = 1 - \sum (h(x_i)/N)^2$.

In the bivariate case let $h(x_i, y_j) : i = 1, \dots, k_x; j = 1, \dots, k_y$ be the histogram frequency data with $\sum h(x_i, y_j) = N$ the total number of observations. When both variables are discrete then obvious generalisations of the univariate smoothers above are

$$\hat{f}(x_i, y_j) = \begin{cases} \omega \frac{h(x_i, y_j)}{N} + (1 - \omega) \frac{1}{k_x k_y} & 0 < \omega < 1 \\ \frac{h(x_i, y_j) + \alpha/(k_x k_y)}{N + \alpha} & \alpha > 0 \end{cases} \quad (12)$$

where ω and α are smoothing constants as above. Again with $\alpha = 1$ the second of these was suggested by Titterington *et al.* (1981).

If a pair of continuous and discrete variables are used then Equation (12) cannot be applied. Suppose instead that smoothing has already been performed using a univariate kernel density approach for the continuous variable (X say) separately for each value of the discrete variable producing the smoothed estimates $\hat{g}(x_i, y_j)$, then the appropriate forms for smoothing in the discrete variable are

$$\hat{f}(x_i, y_j) = \begin{cases} \omega \hat{g}(x_i, y_j) + (1 - \omega) \frac{\hat{g}(x_i)}{k_y} & 0 < \omega < 1 \\ \frac{\hat{g}(x_i, y_j) + \alpha/k_y}{1 + \alpha/\hat{g}(x_i)} & \alpha > 0 \end{cases} \quad (13)$$

where $\hat{g}(x_i) = \sum_{j=1}^{k_y} \hat{g}(x_i, y_j)$ is the marginal density of X calculated from the individually smoothed profiles through the joint distribution. This will, in general, be different to the smoothed marginal distribution.

5 An example of age estimation

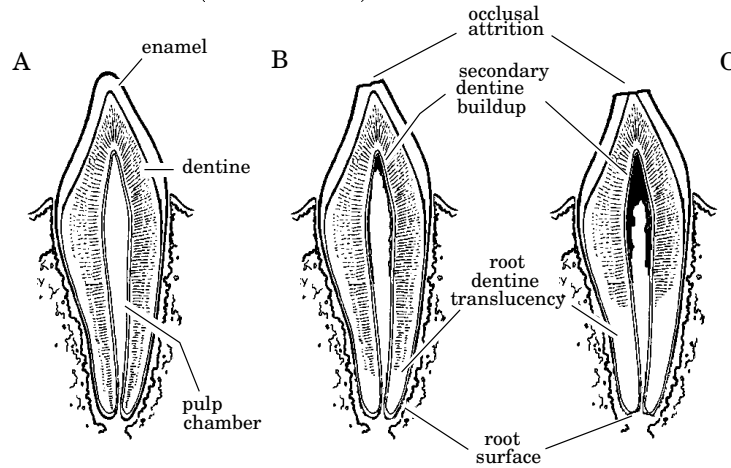
5.1 Background

As an illustration of the use of the calibration outlined above, forensic age data was used from 76 central incisors from the lower jaw (Solheim, 1989). The use of a single type of tooth is necessary (Solheim and Kvaal, 1993) because of the effect of different maturation times. The following four variables, from which to make estimates of age, were considered:

1. Attrition; the gradual wearing down of the tooth surface due to abrasion. This continuous quantity is recorded on an integer scale of 0 to 6. There is a correlation of 0.55 with age.
2. Root surface estimate; a measure of the roughness of the tooth surface. Changes are caused by the continual repositioning of teeth in the mouth throughout life. It takes the form of pitting caused by changes in the attachment apparatus of the tooth, from a smooth surface when the tooth is newly formed getting more and more pitted with age (Solheim and Kvaal, 1993). Scored on an integer scale of 0 to 6 though the structure is again continuous, and with a correlation with age of 0.66.
3. Secondary dentine formation. After the tooth comes to maturity, dentine continues to form on the inside walls of the pulp chamber. Again a continuous quantity scored on an integer scale of 0 to 6, with a correlation with age of 0.60.
4. Root dentine translucency; measures chemical changes in the root. This is a continuous quantity measured on a continuous scale. It is considered to be one of the most reliable age indicators (Solheim, 1989) and has a correlation with age of 0.72.

The changes in these variables are illustrated in Figure (1). (A) is from a young person (18-25), and shows each of the age related traits in an initial state. Indicated are the three main physiological structures of the tooth; the enamel, dentine, and pulp chamber. The attachment apparatus consists of a layer of calcified cellular tissue called cementum, and some calcified fibres called the periodontal ligament - for clarity these structures are not explicitly depicted. (B) is from a slightly older individual (35-45) and has moderate age

Figure 1: Schematic section through a tooth at various stages in its mature development: (A) is from a young person (18-25 years), (B) is from a slightly older individual (35-45 years) and (C) is from an older person (50-60 years)



related changes. (C) is from an older person (50-60) and depicts marked developments in all four areas including the exposure of the dentine through tooth surface wear.

5.2 Estimation of the likelihood function from training data

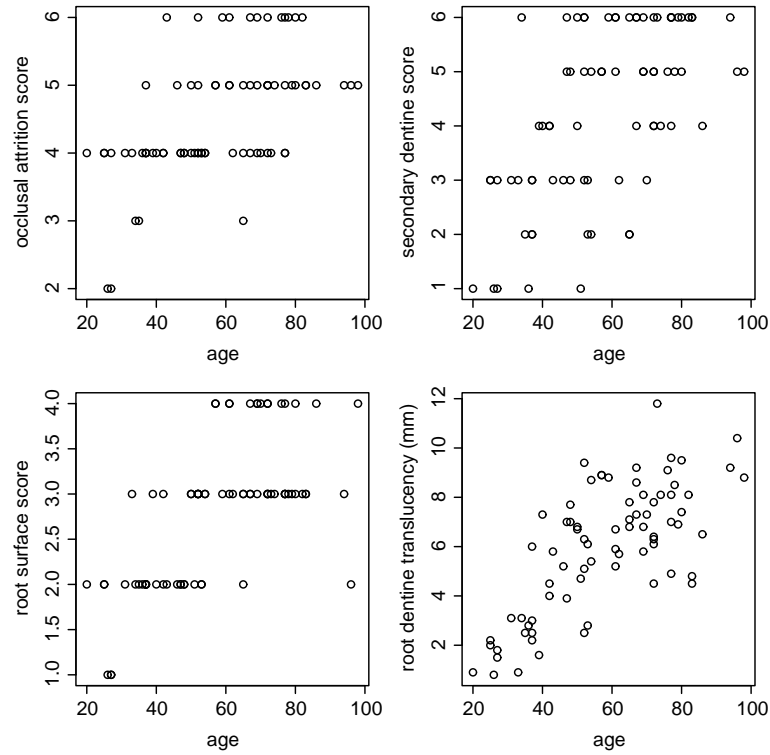
In this section the training data will be described and used to estimate the likelihood function. Plots of each age related variable are shown in Figure 2. The small number of distinct levels of the indicator can clearly be seen for attrition, secondary dentine, and root surface. Whereas root dentine translucency is seen to vary continuously with age. The poor correlation with age evident is typical for all biological indicators of age.

To examine the dependency structure between the indicator variables a series of partial correlations, controlling for age, can be calculated. The results are summarised in Table 1. Secondary dentine has significant correlations with all three of the other variables. The next highest partial correlation is between attrition and the root surface estimate.

Table 1: Partial correlations between variables conditioning on age

Variable pair	Attrition	Secondary dentine	Root surface estimate	Root dentine translucency
Attrition (OA)	1.00	0.42	0.27	0.07
Secondary dentine (SD)		1.00	0.24	0.26
Root surface estimate (RSE)			1.00	0.09
Root dentine translucency (RDT)				1.00

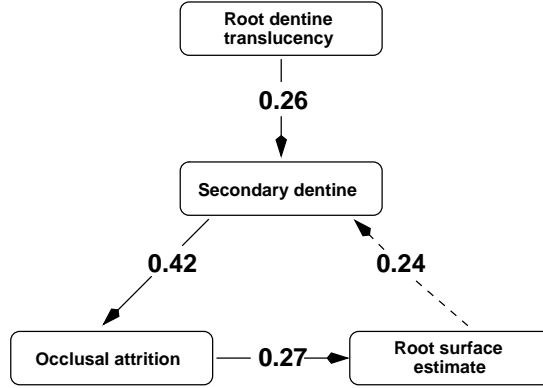
Figure 2: Plot of each age related indicator variable against age. The small number of distinct levels of the indicator can clearly be seen for attrition, secondary dentine, and root surface. Whereas root dentine translucency is seen to vary continuously with age. The poor correlation with age evident is typical for all biological indicators of age



As there are only 76 cases the maximum number of dimensions which can be considered at any one time is three (Silverman, 1986, reproduced in Webb, 1999). Hence it is not possible to consider estimation using the full multivariate form of the posterior distribution (Equation 1) which requires estimation of 5-dimensional distributions.

To use the partial independence approach, based on pair-wise conditional distributions, a conditional independence tree must first be constructed (see Figure 3). As described in Section 4.1, first find the largest partial correlation in Table 1. That is 0.42 between secondary dentine and attrition and so add a branch between these two variables. The next largest is 0.27 between root surface estimate and attrition. Value 0.26 is next largest and so a branch is added between root dentine translucency and secondary dentine. All variables are now linked together and any further branch would create a cycle. Note, however, that the next largest correlation, between root dentine translucency and secondary dentine takes value 0.24 whereas all remaining values are very small. It is clear that small changes in the data could easily produce a different tree with a branch between root surface estimate and secondary dentine instead of between root surface estimate and attrition.

Figure 3: Conditional independence tree based on the partial correlations in Table 1. The link between root surface estimate and secondary dentine has high correlation but is excluded as it would create a cycle in the tree



This conditional independence tree indicates that the appropriate form for Equation (3) is

$$f(\text{All indicators}|\text{Age}) = f(\text{RSE}|\text{OA}, \text{Age})f(\text{OA}|\text{SD}, \text{Age})f(\text{SD}|\text{RDT}, \text{Age})f(\text{RDT}|\text{Age}) \quad (14)$$

where the root has been chosen arbitrarily as root dentine translucency. These distributions can now be calculated from the training data.

5.3 Estimation and performance measures

Age estimates made by the nonparametric calibration approach will be compared to estimates from multiple linear regression. Details of multiple linear regression can be found in most intermediate texts on applied statistics. For the nonparametric approach results using the different priors and the different smoothing will be compared

For all methods point estimates and confidence intervals were derived. For multiple regression conventional estimates were made for both. For the nonparametric approach the median of the posterior density function was used for the point estimate and the confidence interval is given simply as the interval between the 2.5th and 97.5th percentage points of the posterior distribution.

The performance measures used to evaluate the approaches are: (i) the mean absolute deviation (MAD),

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - \hat{x}_i|$$

(ii) the slope (β) of the regression of the residuals against known age (which is a measure of systematic bias),

$$\beta = \frac{s_{x\hat{x}}}{s_x^2}$$

where $s_{x\hat{x}}$ is the sample covariance of x and \hat{x} , and s_x^2 the sample variance of x ; and (iii) the mean width of the 95% confidence intervals. These three measures will allow us to quantify and hence compare the accuracy and precision of the estimation approaches.

In many archaeological studies of human populations a key factor is the maximum human lifespan, and hence the most important of these measures is bias. Many approaches have a tendency to under estimate the ages of the oldest individuals and over estimate the youngest. This will have a reduced effect on the usual measures of accuracy but looking at the systematic pattern will more appropriately reflect the requirements of the application.

Given the relatively small data set ($n = 76$) it was felt that using all the data to fit the models and then again to estimate the ages would lead to unrealistically favourable results. Also the data set is not sufficiently large to divide into separate training and test sets. Hence all estimates were made using a jackknife re-sampling strategy (Efron, 1982). The recorded performance is believed to better reflect a real applied situation where true ages are not known.

5.4 Results

Performance results are presented in Table 2. Ideally any calibration method should have a low mean absolute deviation, low bias, and narrow confidence intervals. As discussed in the previous section the relative importance depends on the application considered. Here primary interest is in the absence of systematic bias, so that the extreme ages are not under-represented.

The multiple regression results give a *base-line* against which to judge the nonparametric estimation procedures. This has a bias of 0.36 which corresponds to an underestimate of about 15 years for the oldest individuals. The majority of the nonparametric estimates have a bias which is substantially less. The better procedures have bias around 0.25 corresponding to an extreme underestimate of less than 10 years.

It should be noted that multiple regression has the smallest mean absolute error of about 9.5 years, but this only goes up to about 11 years for the worst of the nonparametric methods. This average error over the full age range is far less important than the errors at the extremes.

The confidence interval widths for the nonparametric methods vary around that for multiple regression. The widest being 9 years wider and the narrowest 15 years narrower. The reduction in confidence interval width may sound advantageous, however, in the extreme cases there is also a reduction in coverage beyond that expected by chance. Overall, the nonparametric procedures perform better than multiple regression. The nonparametric approaches will now be considered in more detail.

Three factors must be considered: Full independence against partial independence; prod-

Table 2: Summary of jackknife results for multiple regression and nonparametric estimation

Calibration method	Mean absolute deviation		Bias		Mean Confidence interval width		
Multiple regression	9.5		0.36		45.3		
	Prior	Uniform	Data	Uniform	Data	Uniform	Data
Full independence, product kernel							
- unsmoothed discrete		10.2	10.0	0.24	0.37	39.8	36.7
- smoothed discrete		10.1	10.5	0.45	0.54	54.3	47.3
Full independence, full kernel							
- unsmoothed discrete		10.0	9.7	0.26	0.35	37.0	34.6
- smoothed discrete		10.1	10.2	0.41	0.49	48.1	43.2
Partial independence, product kernel							
- unsmoothed discrete		11.1	10.4	0.24	0.35	32.1	30.0
- smoothed discrete		10.1	9.7	0.29	0.40	47.0	41.2
Partial independence, full kernel							
- unsmoothed discrete		11.2	10.7	0.25	0.33	39.5	28.1
- smoothed discrete		10.3	9.9	0.26	0.36	41.7	37.9

uct kernel versus the full three-parameter kernel, and the use of discrete smoothing. Unfortunately the effects of these are not independent. In all cases the inclusion of discrete smoothing makes the bias and the confidence interval width larger. However, in some cases the difference in bias is marginal. Also without discrete smoothing the confidence intervals have coverage less than the nominal 95% whereas those with discrete smoothing are all close to 95%. In most cases the discrete smoothing makes the mean absolute deviation smaller. Overall there is little to choose between using a product or full kernel for the bivariate continuous smoothing with no clear pattern in the differences. Finally, using a partial independence method works well in conjunction with smoothing between discrete categories.

Generally, the use of the non-informative prior leads to better results than using the prior derived from the training data. This is a somewhat surprising results, but is likely to be caused by the use of jack-knifing and the small dataset. In that each time an individual of extreme age is considered the most informative data point is removed. Given that we are particularly interested in estimating extremes of age, the effect is more noticeable. This type of problem is described by Berger (1980, Section 3.2) who suggests determining the central and tail parts of the prior separately.

In this instance it would appear that taking a uniform prior, estimating the likelihoods using pair-wise conditional independence, the three-parameter kernel, and smoothing between discrete categories is the best combination.

6 Discussion

In forensic applications, where the focus is on identification of an individual based on incomplete skeletal remains, high accuracy and narrow confidence intervals are key parameters, with bias being of secondary importance. Describing an adult as aged between 25 and 65 years describes a large proportion of the population and would be little help in identifying an individual aged about 45. A description of between 35 and 60 being much better even though the true age is not at the centre of the interval. From our results multiple regression meets the forensic requirements of accuracy, although the confidence intervals are rather wide. Some of the nonparametric methods offer similar accuracy, but have slightly narrower confidence intervals, particularly the partial independence approaches with discrete smoothing and full kernel bandwidth matrix.

In contrast, when trying to reconstruct the age structure of a historical population from a series of point estimates, bias is the most important performance parameter. Boddington (1987) compared samples for which the true ages had been recorded, with samples which had been estimated from skeletal data. Attention was drawn to the fact that the mean age at death was higher for the recorded samples by as much as 15 years. Although there is evidence that people in the past occasionally *rounded* their ages to the nearest five years, it is also thought to be unlikely that they grossly exaggerated their ages. It should be noted that this bias is the same magnitude as that obtained when using the multiple regression on the reliable modern dataset used in our paper. A partial explanation was found in the calibration method used by archaeologists to assess age at death (Aykroyd *et al.* 1997, 1999). The archaeologists and anthropologists who had held an *a priori* view of short life expectancy for past peoples found that their opinion was reinforced by inappropriate use of calibration. Thus the methods used, and the ages obtained, did not get sufficient critical review. Had appropriate calibration methods been used from the outset the full range of adult ages for past populations would have been readily apparent. This is an example of where the misapplication of statistical techniques has had a far reaching and profound effect upon the core theories of an important area of biological anthropology and human demography. For population demography, multiple regression fails due to the high systematic bias, whereas the smoothed empirical Bayesian calibration method can maintain similar accuracy and precision, and produce a substantial reductions in systematic bias at the extremes of the age

distribution.

A drawback, however, is the relatively large amount of computational work involved, particularly when more than one continuous indicator variable is used, or when Jackknife re-sampling or cross-validation is used. For instance, the most involved of the kernel-based methods takes about 20 minutes on a 300MHz Pentium II, by contrast the multiple regression takes about 10 seconds. The amount of computational work may be dramatically cut by employing a Fourier transform approach to construct the large joint distributions (for further details see Wand and Jones, 1995: Appendix D).

Acknowledgements

The authors acknowledge the financial support of the Natural Environment Research Council (NERC GR3/11395), and thank Professor T. Solheim, Oslo, Norway, for providing the data used. The authors also thank three anonymous referees for their helpful and constructive comments on an earlier draft of this paper.

References

- Aiello, L.C. and Molleson, T. (1993). Are microscopic ageing techniques more accurate than macroscopic ageing techniques. *Journal of Archaeological Sciences*, **20**:689-704.
- Aykroyd, R.G., Lucy, D., Pollard, A.M. and Solheim, T. (1997). Regression analysis in adult age estimation. *American Journal of Physical Anthropology*, **104**:259-265.
- Aykroyd, R.G., Lucy, D., Pollard, A.M. and Roberts, C.A. (1999). Nasty, brutish, but not necessarily short: a reconsideration of the statistical methods used to calculate age from adult human skeletal and dental age indicators. *American Antiquity*, **64**:55-70.
- Bang, G. and Ramm, E. (1970). Determination of age in humans from root dentine transparency. *Acta Odontologica Scandinavica*, **28**:3-35.
- Baxter, M.J. and Beardah, C.C. (1997). Some archaeological applications of kernel density estimates. *Journal of Archaeological Sciences*, **24**:347-354.
- Berger, J.O. (1985). *Statistical Decision Theory*. Second Edition. Springer-Verlag, New York.
- Boddington, A. (1987). From bones to population: the problem of numbers. In A Boddington, A Garland and R Janaway (eds.): *Death, Decay and Reconstruction: Approaches to Archaeology and Forensic Science*. Manchester: Manchester University Press, pp. 179-197.

- Carlin, B.P. and Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.
- Chow, C.K. and Liu, C.N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, **14**: 462-467.
- Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. The Society for industrial and applied mathematics. Philadelphia.
- Eisenhart, C. (1939). The interpretation of certain regression methods and their use in biological and industrial research. *Annals of Mathematical Statistics*, **10**:162-186.
- Ferembach, D., Schwidetzky, I. and Stoukal, M. (1980). Recommendations for age and sex diagnosis of skeletons. *Journal of Human Evolution*, **9**:517-549.
- Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman & Hall, London.
- Gustafson, G. (1950). Age determination on teeth. *Journal of the American Dental Association*, **41**:45-54.
- Iscan, M.Y. (1989). *Age Markers in the Human Skeleton*. Springfield Illinois, U.S.A: Charles C. Thomas.
- Johanson, G. (1971). Age Determinations from Human Teeth. *Odontologisk Revy* 22 Supplement 1.
- Hopkins, K. (1966). On the probable age structure of the Roman population. *Population Studies*, **20**: 245-264.
- Lucy, D., Aykroyd, R.G., Pollard, A.M. and Solheim, T. (1996). A Bayesian approach to adult human age estimation from dental observations by Johanson's age changes. *Journal of Forensic Sciences*, **41**:189-194.
- Lucy, D. and Pollard, A.M. (1995). Further comments on the estimation of error associated with the Gustafson dental age estimation method. *Journal of Forensic Sciences*, **40**:222-227.
- Maritz, J.S. and Lwin, T. (1989). *Empirical Bayes Methods*. Second Edition. Chapman & Hall, London.
- Scott, D.W. (1992). *Multivariate Density Estimation*. Wiley. New York.
- Silverman, B.B (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall. London.
- Solheim, T. (1989). Dental root transparency as an indication of age. *Journal of Dental Research*, **97**:189-197.

- Solheim, T. and Kvaal, S. (1993). Dental root surface structure as an indicator of age. *Journal of Forensic Odonto-stomatology*, **11**:9-21.
- Titterington, D.M. (1980). A comparative study of kernel-based density estimates for categorical data. *Technometrics*, **22**:259-268.
- Titterington, D.M., Murray, G.D., Murray, L.S., Spiegelhalter, D.J., Skene, A.M., Habbema, J.D.F. and Gelpke, G.J. (1981). Comparison of discrimination techniques applies to a complex data set of head injured patients. *J. R. Statist. Soc. A*, **144**: 145–175.
- Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. Chapman and Hall. London.
- Webb, A. (1999). *Statistical Pattern Recognition*. Arnold. London.