*J.I.A.* **121**, I, 119–134

# ON THE CHOICE OF BANDWIDTH FOR KERNEL GRADUATION

BY J. B. GAVIN, M.Sc., S. HABERMAN, M.A., Ph.D., F.I.A., A.S.A., F.S.S., A.F.I.M.A., F.R.S.A. AND R. J. VERRALL, M.A., M.Sc., Ph.D.

## ABSTRACT

This paper considers cross-validation as an objective and risk-based method for selecting the smoothing parameter in a non-parametric graduation. In addition, the relative merits of two kernel estimators are compared in the context of mortality graduation. Finally, it is well known in the statistical literature that the use of theoretically superior kernels is not as important as the choice of bandwidth. Our results support this conclusion, suggesting that the focus on such weights is misguided in the actuarial textbooks on moving weighted averages.

## KEYWORDS

Cross-Validation; Graduation; Kernel Estimation; Optimal Smoothing Kernel

## 1. INTRODUCTION

AN alternative approach to the theory of moving weighted average graduation (MWA) was described by Gavin, Haberman & Verrall (1993). This method relies on the use of kernel estimation techniques which were first used for graduation by Copas & Haberman (1983) and Ramlau-Hansen (1983). These papers describe two forms of estimators for the initial rate of mortality, $q_x$, and also suggest various kernels which may be used. Copas & Haberman (1983) use the normal and Laplace kernels with an estimator that we denote by $\hat{q}_x^{CH}$ while Ramlau-Hansen (1983) discusses the optimality properties of the Nadaraya–Watson estimator, $\hat{q}_x^{NW}$ (Nadaraya, 1964; Watson, 1964). We consider the relative merits of both estimators.

Both kernel estimators contain a bandwidth which governs the amount of smoothing that is applied in the graduation process. In a similar way to other non-parametric graduation methods, the amount of smoothing can be varied, over a continuous range, by the choice of bandwidth. The similarity with Whittaker–Henderson graduation is clear. This is often cited as an advantage over parametric techniques, in which the amount of smoothing can only be varied over a discrete range, by changing the number of parameters. While it is sometimes the case that the amount of smoothing that is appropriate can be decided by studying the resulting graduations, it is desirable to have an objective, data-dependent technique for choosing the bandwidth. In particular, we consider the use of cross-validation for choosing the bandwidth and hence the amount of smoothing. This method can be compared with the more traditional actuarial approach to this problem adopted by Bloomfield & Haberman (1987), of using tests of goodness of fit to determine the bandwidth. There is a

considerable body of work in the statistical literature on the choice of bandwidth for both density estimation (Silverman, 1986; Sheather & Jones, 1991; Hall, Sheather, Jones & Marron, 1991; Jones, Marron & Sheather, 1993) and kernel regression (Härdle, Hall & Marron, 1988; Hastie & Tibshirani, 1990; Scott, 1992b; Hall & Johnstone, 1992). In the area of graduation, Brooks, Stone, Chan & Chan (1988) also use cross-validation, although it is in conjunction with Whittaker–Henderson graduation.

Historically the MWA literature has made use of quite complicated weights to smooth mortality data based on desirable theoretical properties. These formulae can also be derived in the more general context of kernel functions. One such kernel is derived and compared with a standard kernel function.

### 1.1 *Background*

Kernel estimation methods are applied to a probability density function as follows. Suppose we wish to estimate the probability function for a random variable $X$, and we have observations $\{x_i : i = 1, 2, \ldots, n\}$. The kernel estimate of the density at $x$ is estimated by:

$$\hat{f}(x) = (nb)^{-1} \sum_{i=1}^{n} K_b(x - x_i)$$

where $K_b(x) \equiv K(x/b)$ is a kernel function which satisfies:

$$\int_{-\infty}^{\infty} K(x)dx = 1$$

and $b$ is the bandwidth or smoothing parameter.

The bandwidth governs the amount of smoothing which is applied. The larger the value of $b$ is, the more smooth is the resulting estimate. In effect, a kernel density estimator is formed by placing a kernel function at each data point and then summing these functions to form the density estimate. This can be seen in Figure 1, which also shows how the density estimate becomes smoother as $b$ increases, changing from a bimodal to a unimodal density. In this example a scale factor has been omitted from the kernel density and only five observations are used, for clarity. In practice, a larger sample would be required in order to calculate a kernel estimate. A more complete discussion of kernel density estimation is given in Silverman (1986) and Scott (1992b).

Kernel estimation can also be used in a regression context. Suppose we wish to smooth a bivariate scatterplot where the data are $\{(x_i, y_i): i = 1, \ldots, n\}$. The Nadaraya–Watson estimator of the smooth curve is:

$$\hat{f}(x) = \sum_{i=1}^{n} y_i K_b(x - x_i) \Big/ \sum_{i=1}^{n} K_b(x - x_i). \tag{1}$$

This estimator fits a constant to the data which is local to the point of interest, $x$.

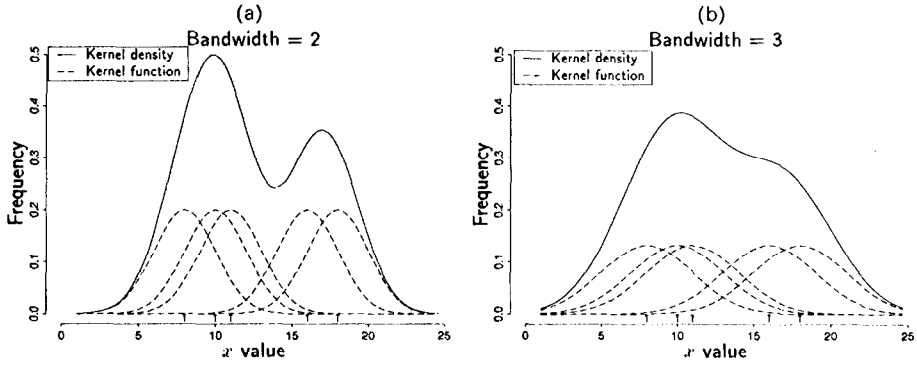


Figure 1. Plot (a) shows five data points, indicated by arrows, ↑. A normal kernel function has been centred around each data point. The kernel estimate of the density is proportional to the sum of the kernel functions. Plot (b) shows the results for a larger bandwidth. The estimate of the density has changed from being bimodal to a unimodal density.

The kernel function and, more importantly, the bandwidth are used to decide which observations are local. For example, Figure 2 shows the Nadaraya-Watson estimator for different bandwidths where $Y$ is binary (survived/died) and we want to estimate the expected value of $Y$ given $X$. The simulated data are denoted by line segments at $Y=0$ (survived) and $Y=1$ (died). As the bandwidth increases the curve becomes more smooth. The weights used to estimate a point on the curve, $X=0{\cdot}5$, are superimposed at the bottom, for the case $b=0{\cdot}1$. Hastie & Tibshirani (1990) offers an excellent introduction to non-parameteric smoothers.

The paper is set out as follows; Section 2 outlines kernel graduation, Section 3 refers to choosing a bandwidth, Section 4 gives some examples and in Section 5, some conclusions are drawn.

## 2. Kernel Graduation

If we denote the random event of a life being alive or dead by $E$, where $E=d$ indicates dead, then we require an estimate of $q_x=\Pr(E=d|x)$. Here $x$ is the age of the life and we require estimates for a range of values of $x$. The application of Bayes theorem results in three probability functions to be estimated, namely:

$$q_x=\Pr(E=d|X=x)=\Pr(X=x|E=d)\ \Pr(E=d)/\Pr(X=x). \qquad (2)$$

Now $\Pr(X=x|E=d)$ and $\Pr(X=x)$ can be estimated using kernel functions and a simple estimate of $\Pr(E=d)$ can be used. Before defining these estimates we give some notation.

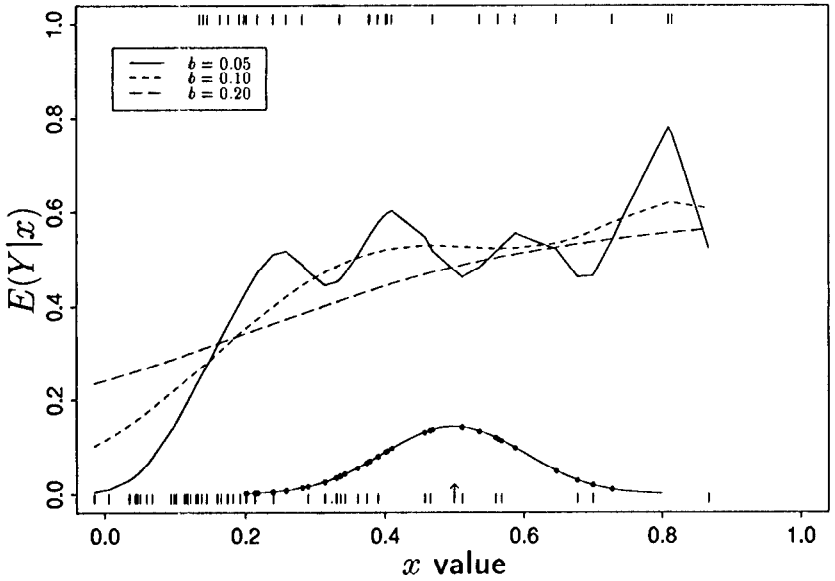


Figure 2. This figure shows a plot of a binary variable $Y$ against $X$, for varying values of the bandwidth parameter, $b$. The line segments at zero (alive) and one (dead) represent the observed data. The Nadaraya-Watson estimator with a normal kernel and different values of the bandwidth, $b$, provides an estimate of the proportion of lives in the whole population who died at a specific $X$ value, $X=x$. The curve becomes more smooth as $b$ increases. The non-zero weights, denoted by ●, used to estimate the curve at $X=0{\cdot}5$ are superimposed at the bottom, for the case $b=0{\cdot}1$. Using these weights, a constant is fitted to the data to obtain the estimate of the curve at $X=0{\cdot}5$.

### 2.1 *The Data*

The estimates, $\hat{q}_x$, are based on crude data for a set of ages, $C=\{x_1, x_2, \ldots, x_n\}$. For each age, $x_i$, we are given a measure of exposure, $e_i$, and the corresponding number of deaths, $d_i$, where $i=1, 2, \ldots, n$. The crude estimate, $\hat{q}_x$, of the true mortality rate, $q_x$, at the $i$th age is denoted by $\hat{q}_i$, where $\hat{q}_i=d_i/e_i$. For convenience, let $q_i \equiv q_{x_i}$. The age of a life, which is regarded as a random variable, is denoted by $X$ and its realised value by $x$. Note that $x$ does not have to be one of the crude ages in the set $C$. If we assume that the observed lives are independent, then, for a given age, the number of deaths, $d_i$, is binomially distributed with index, $e_i$, and probability, $q_i$. This assumption is invalidated by any migration of lives between ages, during the period of exposure, and the presence of multiple policies for individual lives, in the case of insurance-based data.

### 2.2 *Kernel Estimators for Graduation*

We consider two possibilities for estimating $q_x$ by kernel methods, the Nadaraya-

Watson and Copas–Haberman estimators. Since $\hat{q}_x^{CH}$ is, in a sense, more fundamental than $\hat{q}_x^{NW}$ we give that first. The $\hat{q}_x^{CH}$ estimator is obtained by using kernel estimates of the probability function $\Pr(X=x)$ and $\Pr(X=x|E=d)$ in (2). The simple estimate of:

$$\Pr(E=d) \text{ is } \sum_{i=1}^{n} d_i \Big/ \sum_{i=1}^{n} e_i.$$

After some cancellation (Copas & Haberman, 1983; Bloomfield & Haberman, 1987), this gives:

$$\hat{q}_x^{CH} = \sum_{i=1}^{n} d_i K_b(x-x_i) \Big/ \sum_{i=1}^{n} e_i K_b(x-x_i)$$

using the notation in §2.1. A referee has pointed out that if we adopt a kernel-weighted likelihood approach by choosing the estimator $\hat{q}$, that minimises the local binomial log likelihood:

$$\sum_{i=1}^{n} K_b(x-x_i)\{d_i \log(\hat{q}) + (e_i - d_i)\log(1-\hat{q})\}$$

then we get $\hat{q}=\hat{q}_x^{CH}$ (Staniswalis, 1989). See Copas (1983) for a discussion of the Bernoulli case.

If the binomiality inherent in $\hat{q}_x^{CH}$ is ignored, then the $d_i$ deaths out of an exposure of $e_i$ lives becomes a single observation $d_i/e_i$, at each age $x_i$. So the data are condensed to $n$ equally spaced observations and our estimator becomes the Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964). So equation (1) becomes:

$$\hat{q}_x^{NW} = \sum_{i=1}^{n} \left(\frac{d_i}{e_i}\right) K_b(x-x_i) \Big/ \sum_{i=1}^{n} K_b(x-x_i) = \sum_{i=1}^{n} \mathring{q}_i K_b(x-x_i) \Big/ \sum_{i=1}^{n} K_b(x-x_i). \quad (3)$$

This estimator minimises:

$$n^{-1} \sum_{i=1}^{n} K_b(x-x_i)(\mathring{q}-\hat{q})^2$$

and can be viewed as a continuous analogue to moving weighted average graduation (Gavin, Haberman & Verrall, 1993).

In kernel hazard estimation the difference between $\hat{q}_x^{NW}$ and $\hat{q}_x^{CH}$ corresponds to the difference between Ramlau-Hansen's estimator and the more recent local likelihood estimator of Hjort (1994).

### 2.3 *Some Statistical Properties*

Both $\hat{q}_x^{NW}$ and $\hat{q}_x^{CH}$ are intuitive and simple estimators that remove random

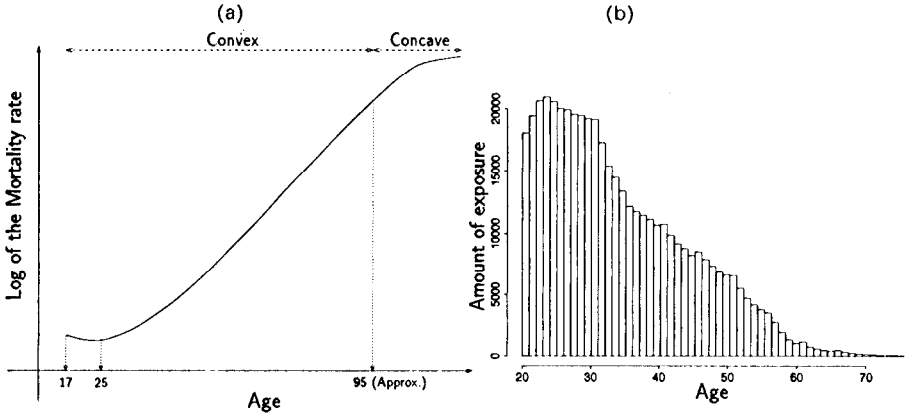


Figure 3. Plot (a) shows a rough outline of male adult mortality rates on a log scale. The curve is convex from about ages 17 to 95 and concave thereafter. The force of mortality is approximately linear from age 25 to age 95, on a log scale. A typical example of how exposure decreases with increasing age is shown in plot (b). The data are taken from duration 1 of The Female Assured Lives 1975–78 (Continuous Mortality Investigation Bureau, 1983).

fluctuations by smoothing the data. In doing so, there is the usual risk that bias may be induced in the resulting estimates. We now consider two features commonly found in mortality data that have an influence on the bias of each estimator.

(1) Generally speaking, mortality rises exponentially. The exceptions are: the first year of life, when the force of mortality (hazard rate) drops sharply; males in their early twenties; and older ages (95+), where the mortality curve for $q_x$ may level off (though data are scanty in this region). The approximate shape of the mortality curve on a log scale from ages 20 to about 100 is shown in Figure 3(a). Note that away from the boundaries, the curve is linear on a log scale. This age range is typical of many insurance-based mortality datasets.

(2) It is often the case with insurance data that the number of lives, exposed to the risk of dying at a given age, decreases with increasing age over the ages 25 to 95. Figure 3(b) shows the amount of exposure from the duration 1, Female Assured Lives 1975–78 experience which reflects this feature. The exception is at the youngest ages, where the exposure is increasing with increasing age, presumably because less insurance is sold to those below twenty compared to those in their twenties. In any case, estimates in this region are likely to be influenced by the bias that arises near the boundary.

For population data, the number of lives exposed to the risk of dying depends on past fertility, migration and mortality levels. In some cases the sharp variation in exposure with age may be less apparent than in Figure 3(b), with the result that the bias in $\hat{q}_x^{CH}$ is less serious.

The bias in both the $\hat{q}_x^{NW}$ and $\hat{q}_x^{CH}$ estimators is well known. In order to define this, we view the observed ages as a random variable, $X$, with probability density, $f$, say. In addition, denote derivatives with respect to $x$ by $q'$. With this notation, the bias for both $\hat{q}_x^{NW}$ and $\hat{q}_x^{CH}$ is proportional to:

$$q_x''+2f'(x)q_x'/f(x)+R \tag{4}$$

by a Taylor series expansion, where $R$ is a remainder term consisting of higher order derivatives. So the bias inherent in the two estimators depends on the distribution of the data over the age range, $f$, and on the curvature of the mortality function, $q_x''$. There is also additional bias near the boundaries, but we are mainly concerned with the ages in the interior.

Applying this formula to the mortality data described in § 2.1, the bias for $\hat{q}_x^{NW}$ becomes:

$$\frac{\sum_{i=1}^{n}(x_i-x)K(x-x_i)}{\sum_{i=1}^{n}K(x-x_i)}q_x'+\frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-x)^2K(x-x_i)}{\sum_{i=1}^{n}K(x-x_i)}q_x''+R. \tag{5}$$

The remainder term, $R$, is small because of the locality of the kernel function and so is ignored. In the interior of the age range, if we have a crude mortality rate for every age, then the data are symmetric around $x$, so the coefficient of $q_x'$ is zero in (5). An alternative view is to say that the data are uniformly distributed across the age range, so $f'(x)=0$ in (4). This means that the bias in the Nadaraya–Watson estimator depends only on the curvature of the mortality curve. This is known as the fixed design case in the statistical literature.

Mortality rates for ages 25 to about 80 are financially significant for insurance purposes. If the data extend a little above and below these limits, then this region can be regarded as being in the interior. Therefore the estimates for these ages will not be heavily influenced by boundary effects. For these ages, the mortality curve is approximately exponential in shape. So, if we transform the crude rates by taking logs, then the mortality curve approximates a straight line over that region which implies zero curvature, as shown in Figure 3(a). Therefore the Nadaraya–Watson estimator is expected to give an unbiased estimate of the true mortality rate for this region. Without transformation there is a positive bias.

If the crude rates are not evenly spaced there may be considerable bias. This feature is shown in Figure 2. The kernel weights superimposed at the bottom of the graph show that the data used to estimate the curve at $X=0{\cdot}5$ lie mainly to the left of 0·5, causing a negative bias in a generally increasing curve.

The bias in the $\hat{q}_x^{CH}$ estimator (Copas & Haberman, 1983) is:

$$\frac{\sum_{i=1}^{n}(x_i - x)e_i K_b(x - x_i)}{\sum_{i=1}^{n} e_i K_b(x_i - x)} q'_x + \frac{1}{2}\frac{\sum_{i=1}^{n}(x_i - x)^2 e_i K_b(x - x_i)}{\sum_{i=1}^{n} e_i K_b(x_i - x)} q''_x + R. \qquad (6)$$

If the data were symmetrically placed in the neighbourhood of $x$, then the bias for $\hat{q}_x^{CH}$ would be the same as for $\hat{q}_x^{NW}$. However this is not the case, as the number of lives exposed to the risk of dying tends to decrease with increasing age, as is shown in Figure 3(b). The distribution of the ages, $f$, would have a similar shape. So, for the $\hat{q}_x^{CH}$ estimator, $f'(x) < 0$ in (4), for ages in the middle of the table. The asymmetry suggests that the coefficient of the $q'_x$ term is negative, giving $\hat{q}_x^{CH}$ a negative bias for most ages. This is referred to as the random design case in the statistical literature.

If the graduation is carried out on the original exponential scale, then the $q''$ term is positive in both estimators, over the ages 25 to 95. So $\hat{q}_x^{NW}$ has a positive bias while $\hat{q}_x^{CH}$ has a negative bias from the coefficient of the $q'_x$ term, offset to some extent by the positive bias of the coefficient of $q''_x$. Overall we might expect the $\hat{q}_x^{CH}$ estimator to lie below $\hat{q}_x^{NW}$, over this region. Recently Nielsen (1992) has considered transformations to reduce bias.

Several methods have been proposed for dealing with the increased bias that can arise at the boundary. Rice (1984) and Jones (1993) advocate an extrapolation method that merges two different kernels with two different bandwidths to remove the $q'_x$ term in the bias. Alternatively a reflection approach, due to Hall & Wehrly (1991), generates pseudo-data which effectively extend the boundaries so that the original data are now in the interior and so are not subject to boundary effects. Some kind of boundary correction is essential if the graduated rates are to extend across the entire age range. This complication is not considered further here, but Gavin, Haberman & Verrall (1995) have found the extrapolation method to be quite effective in an adaptive kernel model. Hoem & Linnemann (1988) provide a rigorous mathematical approach for dealing with the ends of the table, in the context of moving weighted averages, and it seems plausible that their theory could be incorporated within kernel graduation. More recently attention has focused on a method related to the Nadaraya-Watson estimator which fits higher order polynomials locally. This method automatically adjusts to allow for boundary effects and is discussed further in Section 5. In summary, we expect $\hat{q}_x^{NW}$ to perform better than $\hat{q}_x^{CH}$ in the centre of the age range.

### 2.4 *Choosing a Kernel*

A standard kernel in the statistical literature is the normal kernel, $K^N$, defined as:

$$K^N(x) = \frac{\exp\{-(x/2b)^2\}}{\sqrt{2\pi b^2}} \quad \text{where} \quad -\infty < x < +\infty. \qquad (7)$$

This kernel is used to graduate some mortality data in Section 4.

Some of the statistical literature has focused on allowing the kernel functions to take negative values as a possible bias reduction technique. If the curvature of the true curve is constant, then these higher-order kernels may offer a reduction in bias, but at a cost of increased variance (Hastie & Loader, 1993). Although negative weights have been used with MWA in the actuarial literature, kernels that take negative values are not popular in the statistical literature, partly because they are difficult to interpret. For example, if we minimise the asymptotic variance of the estimator subject to:

$$\int_{-\infty}^{\infty} K(x)dx = 1 \qquad \text{and} \qquad \int_{-\infty}^{\infty} x^2K(x)dx = 0 \tag{8}$$

and $K$ being bounded, then we get the kernel:

$$K(x) = \begin{cases} 3(3 - 5x^2)/8, & |x| < 1 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

(Silverman, 1986, § 3.6.2). In the context of moving weighted averages, actuaries have historically given a lot of thought to the best theoretical weights to use in a moving weighted average. A traditional way of deciding if a set of graduated rates is smooth is to calculate second or third differences of the graduated rates and check if they are small and random, via a set of statistical tests. With this in mind, attention has focused on choosing weights that minimise the variance of the $k$th differences of the graduated rates relative to the variance of the $k$th differences of the crude rates (London, 1985, chapter 3; Benjamin & Pollard, 1980, (13.16); Ramsay, 1993). Benjamin & Pollard (1980) refers to such weights as 'optimal smoothing weights'. If we repeat this approach in the context of kernel estimation we get the kernels:

$$K(x) = \begin{cases} 3(3b^2 - 5x^2)/8b^3, & |x| \leqslant b \\ 0 & |x| > b \end{cases} \tag{10}$$

and

$$K(x) = \begin{cases} 15(b^2 - x^2)(3b^2 - 7x^2)/32b^5, & |x| \leqslant b \\ 0 & |x| > b \end{cases} \tag{11}$$

for $k = 0$ and $k = 1$ respectively. The shape of these kernels is shown in Figure 4. The kernel for the case $k = 0$ is formed by minimising the asymptotic variance of the graduated rates relative to the variance of the crude rates, subject to the conditions in equation (8). It is the same as equation (9), but, because it suffers from being discontinuous at $\pm b$, we will not consider it further. The procedure for the case $k = 1$ is similar, but is applied to the first-differences of the graduated and crude rates. The resulting kernel is continuous with support over the interval $(-b,b)$.

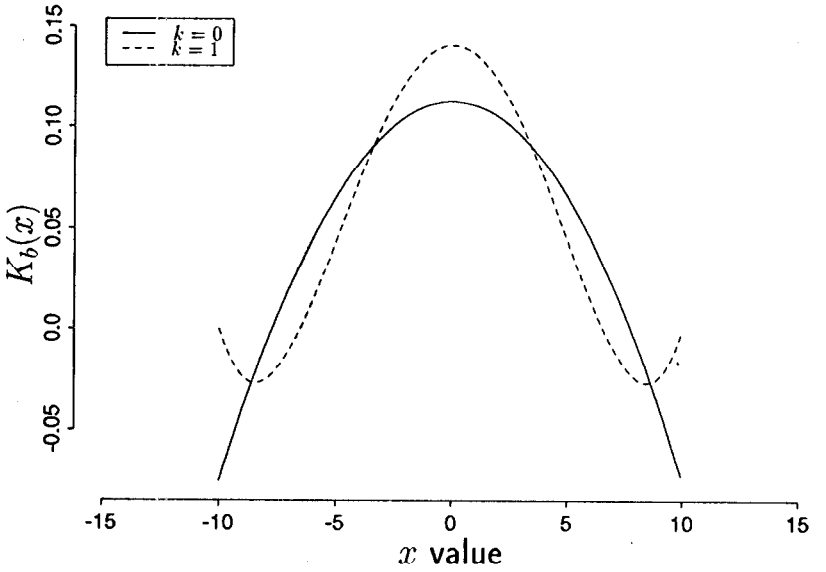


Figure 4. This figure shows a plot of the kernel functions given in equations (10) and (11) for $b = 10$. The case $k = 0$ is formed by minimising the asymptotic variance of the graduated rates relative to the variance of the crude rates. The case $k = 1$ is similar, but is applied to the first-differences of the graduated and crude rates.

For consistency with Benjamin & Pollard (1980), we refer to the kernels in (10) and (11) as optimal smoothing kernels, $K^{OSK}$, though this phrase is not standard. The bandwidth in the optimal smoothing kernel can be related to the range of a moving weighted average graduation. For example, Spencer's 21-term formula can be approximated using the optimal smoothing kernel with $k = 3$ and a bandwidth of $b = 10$. Both London (1985) and Benjamin & Pollard (1980) discuss the case $k = 3$ for MWA, but the formulae become more awkward as $k$ increases. Therefore we will not consider the cases where $k > 1$, but further details are available in Gavin, Haberman & Verrall (1993).

The exponential growth inherent in mortality data over much of the age range results in a changing rate of curvature. This suggests that higher-order kernels and the moving average weights suggested by Benjamin & Pollard (1980, (13.16)) and London (1985, chapter 3) do not offer any savings in bias over simpler, positive kernels. Experiments with such kernels tend to support these arguments. In Section 4, we compare the two estimators, $\hat{q}_x^{NW}$ and $\hat{q}_x^{CH}$, and the two kernels, the normal and optimal smoothing kernel for the case $k = 1$.

## 3. Choosing a Bandwidth

A standard technique in actuarial science is to first choose a model that best fits the data and then test it for smoothness. A more modern statistical approach is to combine both of these steps by using a risk-based method to choose the bandwidth, simultaneously striking a balance between variance and bias. For non-parametric regression, this can be achieved through a suitable choice for the bandwidth parameter. We consider one method for achieving this, called cross-validation. This method can be compared with that due to Bloomfield & Haberman (1987), which fitted a curve to the data and then separately tested the graduated rates for smoothness using standard actuarial tests of fit.

### 3.1 *Cross-Validation*

This common technique has been examined in the statistical literature by Stone (1974) and more recently by Gregoire (1993). Brooks, Stone, Chan & Chan (1988) have considered cross-validation in the context of a Whittaker–Henderson graduation. It is an automatic and simple method for selecting a bandwidth that reflects the data, but which also considers smoothness. Given any estimator, $\hat{q}_x$, of the true rate of mortality, $q_x$, we choose the bandwidth which minimises the function $CV(b)$ where:

$$CV(b) = n^{-1}\sum_{j=1}^{n}(\hat{q}_j - \hat{q}_j^{(-j)})^2 \tag{12}$$

$$\text{where: } \hat{q}_j^{(-j)} = \begin{cases} \sum_{\substack{i=1\\ i\neq j}}^{n} d_i K_b(x_j - x_i) \Big/ \sum_{\substack{i=1\\ i\neq j}}^{n} e_i K_b(x_j - x_i) & \text{for } \hat{q}_x^{\mathrm{CH}} \\ \sum_{\substack{i=1\\ i\neq j}}^{n} \hat{q}_i K_b(x_j - x_i) \Big/ \sum_{\substack{i=1\\ i\neq j}}^{n} K_b(x_j - x_i) & \text{for } \hat{q}_x^{\mathrm{NW}} \end{cases}$$

depending on which estimator is being used. For a fixed bandwidth, $\hat{q}_j^{(-j)}$ is the estimate of the rate of mortality using all the crude rates except the one where $i=j$. Having calculated $\hat{q}_j^{(-j)}$ using this 'leave one out' approach, for $i=1, 2, \ldots, n$, we then compare these values to the crude rates by calculating the average of the squared differences to get the cross-validation score, $CV(b)$.

Cross-validation says that we should choose the bandwidth which minimises $CV(b)$. Theoretically, minimising $CV(b)$ is approximately equivalent to minimising the mean integrated squared error. The mean integrated squared error ($MISE$) of $\hat{q}$ as an estimator of $q$ is defined as:

$$\begin{aligned} MISE(\hat{q}) &= E\int\{\hat{q}_x - q_x\}^2 dx \\ &= \int\{E\hat{q}_x - q_x\}^2 dx + \int V\{\hat{q}_x\} dx \\ &= \text{Integrated squared bias} + \text{Integrated variance.} \end{aligned}$$

Thus it allows us to balance objectively bias and variance. We refer to the bandwidth that minimises (12) as $b_{CV}$. It is found by using a grid search. Numerical instability may arise for $b_{CV} \approx 0$, but such graduations are not smooth and so are ignored.

Cross-validation is just one data-driven method for selecting the smoothing parameter. Despite its simplicity and intuitive appeal it can produce variable results, and some evidence suggests it under-smooths the data (Hall & Johnstone, 1992; Scott, 1992a; Jones, Marron & Sheather, 1993).

## 4. An Example: The Female Assured Lives 1975–78

This section presents some results from graduating a large, standard mortality table using the estimators $\hat{q}_x^{\mathrm{NW}}$ and $\hat{q}_x^{\mathrm{CH}}$, the two kernel functions $K^N$ and $K^{OSK}$ and cross-validation. The chosen table is the Female Assured Lives 1975–78. This mortality table has a select period of two years, but the results for each duration were broadly similar, so only duration one is presented. The age range for this table is 20·5 to 74·5 and the total exposure is 459,068 policy years for which there were 334 recorded deaths.

The cross-validation scores for a range of bandwidths are shown in Figure 5(a), using $K^N$. The results show a clear minimum for $\hat{q}_x^{\mathrm{NW}}$ with $b_{CV}=4$ and $b_{CV}=2{\cdot}75$ for $\hat{q}_x^{\mathrm{CH}}$. The graduations for these bandwidths are shown in 5(b). From §2.3, the negative bias in $\hat{q}_x^{\mathrm{CH}}$ relative to $\hat{q}_x^{\mathrm{NW}}$ is clear, even though the graduations are carried out on the untransformed crude rates. Both estimators show a positive bias relative to the published tables, in the middle of the table. The published rates at durations zero and one are based on an adjustment to the duration two-plus rates (Continuous Mortality Investigation Bureau, 1983) and not the crude rates shown in Figure 5(b). The negative bias at old ages and the positive bias at young ages is attributed to the boundary effects. Some solutions to this problem are mentioned in §2.3.

Figures 5(c) and 5(d) show the corresponding results for $K^{OSK}$. The choice of $b_{OSK}$ is not as clear as for the normal kernel. A bandwidth of 12 is chosen for both $\hat{q}_x^{\mathrm{NW}}$ and $\hat{q}_x^{\mathrm{CH}}$, but this seems to be excessively large for practical purposes. From the comments at the end of Section 2, we might expect $K^{OSK}$ to show less bias, but at a cost of greater variability. The two graduations are in closer agreement, but the lack of smoothness is disappointing, especially as the kernel spans almost half the age range. The kernel for the case $k=3$ gave smoother results, but the complicated formula was not considered very practical and so is not shown. This suggests that the attention paid in actuarial textbooks to deriving superior weights for moving weighed averages is misguided.

For comparison, results were also produced using the method in Bloomfield & Haberman (1987) for choosing the bandwidth. This method has been used on standard mortality tables such as the Assured Lives 1967–70 Table. The method tests fidelity to the data by using three tests: the chi-squared test, the runs test and the serial-correlation test. The choice of tests of fit are subjective and rely on

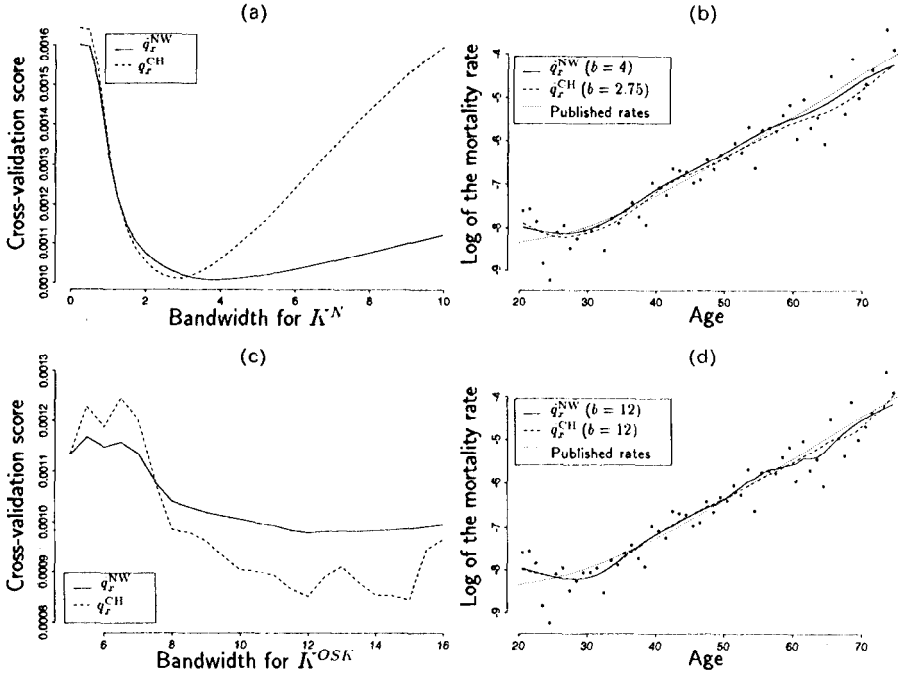


Figure 5. The results from graduating duration one of the Female Assured Lives 1975–78 are shown. Plot (a) shows the cross-validation scores for a range of bandwidths, using the normal kernel. Both curves produce a clear minimum, suggesting $b_{CV}=4$ for $\hat{q}_x^{NW}$ and $b_{CV}=2{\cdot}75$ for $\hat{q}_x^{CH}$. Note that $\hat{q}_x^{CH}$ tends to lie below $\hat{q}_x^{NW}$ for most of the ages. The graduated rates in plot (b) show clearly the negative bias in $\hat{q}_x^{CH}$, relative to $\hat{q}_x^{NW}$. Plots (c) and (d) show the corresponding figures using $K^{OSK}$ instead of $K^N$. Both the cross-validation scores and the graduations are more erratic.

asymptotic arguments. In practice, some of the tests require a minimum of about 50 crude rates in order to be valid, and overall more work is required in order to choose a bandwidth. However, both methods produce broadly similar choices for the best bandwidth, but cross-validation is clearly more intuitive and theoretically more sound.

## 5. Conclusions

Cross-validation provides an intuitive, data-driven, risk-based method for selecting the bandwidth and it is easy to implement. It simultaneously provides

an objective, smooth estimate that fits the data, and so has much to offer over the more traditional actuarial approach.

Overall the $\hat{q}_x^{NW}$ estimator is more successful than the $\hat{q}_x^{CH}$ estimator, at least for mortality data. The main difference between the two estimators is that $\hat{q}_x^{NW}$ combines all of the data at each age into a single observation $d_i/e_i$, so that this estimator is based on evenly spaced data. Although the $\hat{q}_x^{CH}$ estimator makes explicit use of the number of deaths and the amount of exposure at each age, the interval $(x-b, x)$ almost always contains more data than the interval $(x, x+b)$ and this causes a systematic bias. However, for some population tables this bias may be small, for ages in the middle of the table.

The optimal smoothing kernel is disappointing as it does not perform as well as the normal kernel. This suggests that the use of theoretical weights, as presented in the standard actuarial textbooks on MWA, do not produce superior results to much simpler weights. At least in the context of kernel graduation, the choice of bandwidth is dominant, so attention should focus on ways of choosing this parameter, such as cross-validation. In addition, kernel graduation offers a more flexible approach to graduation than MWA.

As mentioned in Section 1, kernel regression fits a local constant to the data. The next logical step is to consider the closely related problem of fitting higher order functions locally, such as a straight line or a quadratic. This technique was popularised by Cleveland (1979). As before, the kernel function and the bandwidth decide which observations lie near the point we wish to estimate, but we now fit a line to these local rates using least squares. Higher order models allow the bias associated with the first and second derivatives of $q_x$ to be eliminated, but without a substantial increase in variance. In addition, automatic adjustment is made for the increased bias at the boundary. There has been renewed interest in the statistical literature in kernel-weighted local linear regression recently (Fan, 1992; Fan & Gijbels, 1992; Jones, Davies & Park, 1993). Hastie & Loader (1993) review the recent literature on this topic. For a more global view of the subject, Hastie & Tibshirani (1994) explore a class of generalised regression models, called varying-coefficient models, which includes local linear regression amongst others.

The lower costs and greater speed of modern computing technology have popularised non-parametric modelling. This trend is likely to accelerate in the future. By carefully choosing appropriate models, the non-parametric approach allows the detailed structure of the data to be explored. It does not require the estimation of an unwieldy number of parameters, which can sometimes arise in parametric graduation. We are not advocating that a non-parametric model should always be used instead of a parametric one, but if the shape of the curve is not known in advance then this method can be used to interrogate the data, initially. It can, therefore, be viewed as an explanatory step towards the final model choice which may be parametric because of its inherent smoothness. Differences between the best parametric and non-parametric graduations will highlight the extent of the actuary's desire for smoothness at a cost of lack of fit to the data.

ACKNOWLEDGEMENT

The authors would like to thank two anonymous referees for their helpful comments which led to substantial improvements. We are also grateful to M. C. Jones and D. W. Scott for supplying preprints of their work.

REFERENCES

BENJAMIN, B. & POLLARD, J. H. (1980). *The Analysis of Mortality and Other Actuarial Statistics.* London: Heinemann.

BLOOMFIELD, D. S. F. & HABERMAN, S. (1987). Graduation: Some experiments with kernel methods. *J.I.A.* **114,** 339–369.

BROOKS, R. J., STONE, M., CHAN, F. Y. & CHAN, L. Y. (1988). Cross validatory graduation. *Insurance: Mathematics and Economics* **7,** 59–66.

CLEVELAND, W. S. (1979). Robust locally-weighted regression and smoothing scatterplots. *Jour. Amer. Stat. Assoc.* **74,** 829–836.

CONTINUOUS MORTALITY INVESTIGATION BUREAU (1983). Graduation of the mortality experience of female assured lives: 1975–78. **Report Number 6,** Institute and Faculty of Actuaries.

COPAS, J. B. (1983). Plotting p against x. *Applied Statistics* **32,** 25–31.

COPAS, J. B. & HABERMAN, S. (1983). Non-parametric graduation using kernel methods. *J.I.A.* **110,** 135–156.

FAN, J. (1992). Design-adaptive nonparametric regression. *Jour. Amer. Stat. Assoc.* **87,** 998–1004.

FAN, J. & GIJBELS, I. (1992). Variable bandwidth and local linear regression smoothing. *Annals of Statistics* **20(4),** 2008–2031.

GAVIN, J. B., HABERMAN, S. & VERRALL, R. J. (1993). Moving weighted average graduation using kernel estimation. *Insurance: Mathematics and Economics* **12,** 113–126.

GAVIN, J. B., HABERMAN, S. & VERRALL, R. J. (1995). Graduation by kernel and adaptive kernal methods with a boundary correction. To appear in *Trans. Soc. Actuaries.*

GREGOIRE, G. (1993). Least square cross validation for counting processes intensities. To appear in *Scandinavian Journal of Statistics.*

HALL, P. & JOHNSTONE, I. M. (1992). Empirical functionals and efficient smoothing parameter selection (with discussion). *Jour. Royal Statist. Soc. B* **(54),** 475–530.

HALL, P., SHEATHER, S. J., JONES, M. C. & MARRON, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78,** 263–270.

HALL, P. & WEHRLY, T. E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimates. *Jour. Amer. Stat. Assoc.* **86,** 665–672.

HÄRDLE, W., HALL, P. & MARRON, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? (with comments). *Jour. Amer. Stat. Assoc.* **83,** 86–101.

HASTIE, T. & LOADER, C. (1993). Local regression: automatic kernel carpentry (with comments). *Statistical Science* **8(2),** 120–143.

HASTIE, T. & TIBSHIRANI, R.J. (1993). Varying-coefficient models (with discussion). *Jour. Royal Statist Soc. B.* **55,** 757–796.

HASTIE, T. J. & TIBSHIRANI, R. J. (1990). *Generalized Additive Models.* London: Chapman and Hall.

HJORT, N. L. (1994). Dynamic likelihood hazard rate estimation. To appear in *Biometrika.*

HOEM, J. M. & LINNEMANN, P. (1988) The tails in moving average graduation. *Scand. Actuarial Jour.*, 193–229.

JONES, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and Computing,* 135–146.

JONES, M. C., DAVIES, S. J. & PARK, B. U. (1993). Versions of kernel-type regression estimators. To appear in *Jour. Amer. Stat. Assoc.*

JONES, M. C., MARRON, J. S. & SHEATHER, S. J. (1993). Progress in data-based bandwidth selection for kernel density estimation. Private communication, submitted for publication.

LONDON, D. (1985). *Graduation—The Revision of Estimates.* Winsted and Abington, Connecticut, USA: ACTEX Publications.

NADARAYA, E. A. (1964). On estimating regression. *Theor. Prob. Appl.* **9,** 141–142.

NIELSEN, J. P. (1992). A transformation approach to bias correction in kernel hazard estimation. Research Report Number 115, Laboratory of Actuarial Mathematics, University of Copenhagen.

RAMLAU-HANSEN, H. (1983). The choice of a kernel function in the graduation of counting process intensities. *Scand. Actuarial Jour.*, 165–182.

RAMSAY, C. M. (1993). Minimum variance moving-weighted-average graduation. *Trans. Soc. Actuaries*, **43**.

RICE, J. A. (1984). Boundary modification for kernel regression. *Communications in statistics—theory and methods* **13,** 893–900.

SCOTT, D. W. (1992a). Constrained oversmoothing and upper bounds on smoothing parameters in regression and density estimation. Technical Report 92-8, Department of Statistics, Rice University.

SCOTT, D. W. (1992b). *Multivariate Density Estimation; Theory, Practice and Visualisation.* John Wiley & Sons.

SHEATHER, S. J. & JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Jour. Royal Statist. Soc. B*(**53**), 683–690.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

STANISWALIS, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *Jour. Amer. Stat. Assoc.* **84,** 276–283.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Jour. Royal Statist. Soc. B*(**36**), 111–147.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhya A*(**26**), 359–372.