TRANSACTIONS OF SOCIETY OF ACTUARIES
1995 VOL. 47

GRADUATION BY KERNEL AND ADAPTIVE KERNEL METHODS WITH A BOUNDARY CORRECTION*

JOHN GAVIN†, STEVEN HABERMAN, AND RICHARD VERRALL‡

ABSTRACT

This paper explores the flexibility of kernel estimation as a means of nonparametric graduation and relates it to moving-weighted-average graduation. Our primary objective is to focus attention on a model that makes explicit allowance for the variation in exposure over age. We also consider various transformations of the data, cross-validation as an objective method for choosing the smoothing parameter, and diagnostic methods for checking assumptions. A kernel function for improving the estimate at a boundary is discussed, and the results are applied to two mortality tables.

1. INTRODUCTION

Sets of mortality rates, in the form of mortality tables, are widely used by actuaries to calculate life insurance premiums, annuities, reserves, and so on. Producing these tables from a suitable set of crude (or raw) mortality rates is called graduation, and this subject has been extensively discussed in the actuarial literature. To be specific, given a set of crude mortality rates, $\dot{q}_i$, for each age x_i, we wish to systematically revise these initial estimates to produce smoother estimates, $\hat{q}_i$, of the true but unknown mortality rates, q_i, where $i=1, \ldots, n$. The crude rate at age x_i is typically based on the number of deaths recorded, d_i, relative to the number of policy-years or person-years initially exposed to the risk of death, e_i, for a homogeneous cohort over a certain time interval. By reducing the unit of time from a year, we could alternatively consider the instantaneous rate of mortality, which is called the force of mortality, also known as the hazard or intensity rate in survival analysis. Intuition and practical convenience lead us to believe that a smooth sequence of graduated rates will more closely reflect the variation

*Supported by the U.K. Engineering and Physical Sciences Research Council.

†John Gavin, not a member of the Society, is a research student, School of Mathematics, University of Bath, England.

‡Richard Verrall, not a member of the Society, is a Senior Lecturer in the Department of Actuarial Science and Statistics at City University, London, England.

due to age in the unknown, true rates of mortality compared to the crude rates. Some nonparametric models reflect this belief by allowing the amount of smoothing to vary over a continuous range.

There are a variety of possible uses of nonparametric methods. A larger class of possible regression surfaces can be considered, reflecting the possibility that the graduated rates may not follow a neat parametric formula. We can use a nonparametric approach to choose the simplest suitable parametric model, to provide a diagnostic check of a parametric model, or to simply explore the data. The power of modern computers and software has made nonparametric smoothing more feasible and consequently more popular. Some of the more popular statistical methods are nearest-neighbor smoothing [12], spline-smoothing [26], [56], and kernel-smoothing, which is discussed in this paper. Kernel-smoothing is not new. For a scatter plot of bivariate data, X and Y, Watson [59] suggests estimating the conditional mean $E(Y|X)$ from nonparametric kernel estimates of the joint density of X and Y and the marginal density of X. Kernel smoothers are also suggested by Nadaraya [47]. Scott [54] offers an introduction to kernel density estimation and regression.

A mortality table can be viewed as a bivariate scatter plot of mortality against age, in which the true mortality rates can be estimated from the mean regression function by using kernel estimators. Although age is a continuous variable, it is typically truncated in some way, such as age last birthday. Thus, the data consist of e_i observations at age x_i, of which d_i die and $e_i - d_i$ survive. Given the discretized nature of a mortality table, it is natural to pool the data by using the average d_i/e_i at each age. This reduces the computational burden and leads to a fixed design model, in which we have a single observed mortality rate at equally spaced ages. In later sections of this paper, we consider how to adjust the model to reflect the amount of exposure at each age. Because the data may not have a constant variance, we may need to consider transforming it to satisfy the model, which is

$$\dot{q}_i^t = q_i^t + r_i, \qquad \text{for } i = 1, \ldots, n, \tag{1}$$

where t denotes some transformation and the residuals r_i are assumed to be independently, identically distributed random variables, with zero mean and a constant, finite variance. We need to ensure that these assumptions are reflected in the data and, if not, to make appropriate adjustments. Although $\dot{q}_i^t$ is treated as a random variable, we adopt the standard actuarial notation for mortality by using a lowercase letter. Once the graduation process is

complete, the transformation is reversed to obtain the graduated rates on the original scale.

The Nadaraya-Watson kernel estimator of the true mortality rate is

$$\hat{q}_i' = \frac{\sum_{j=1}^{n} q_j' K_b(x_i - x_j)}{\sum_{j=1}^{n} K_b(x_i - x_j)}, \qquad \text{for } i = 1, \ldots, n. \tag{2}$$

For convenience, $K_b(x) \equiv b^{-1}K(x/b)$ is used throughout. The function K, called a kernel function, is any function for which $\int_{-\infty}^{\infty} K(x)dx = 1$; thus, any probability density function is a kernel function. Frequently, but not always, kernel functions are non-negative, $K(x) \geq 0$. A common example is the standardized normal or Gaussian kernel

$$K^N(x) = (2\pi)^{-1/2} \exp\{-x^2/2\} = \phi(x), \qquad \text{for } -\infty < x < \infty. \tag{3}$$

The bandwidth b acts as a smoothing parameter. Choosing a small bandwidth means that only nearby points are influential; choosing a large bandwidth means that information is averaged over a larger region, and consequently individual points have less influence on our estimate. At the point at which estimation is to take place, x_i, we first use the kernel function, K, and the bandwidth, b, to decide which of our n observations lie nearby; then we fit a constant to these points by averaging. This is our estimate of the curve at x_i.

Ramlau-Hansen [49] discusses the motivation for using this estimator to calculate mortality rates, and its advantages over a related estimator used by Copas and Haberman [16] and by Bloomfield and Haberman [4] are discussed by Gavin, Haberman, and Verrall [25]. It is also the estimator used in this paper, but there are other well-known kernel regression estimators in the statistical literature [43]. A well-known rival to the Nadaraya-Watson estimator is an integral-based estimator due to Gasser and Müller [23]. All kernel estimators have their relative advantages, and which is more suitable for graduation is currently an open question. The recent paper by Chu and Marron [11] comparing the Nadaraya-Watson and Gasser-Müller estimators is highly recommended. We provide a detailed list of references with the aim of generating a greater interest in the application of kernel estimation, and nonparametric methods in general, in actuarial science. However, this has been an active area of research in recent years, so our bibliography is not complete.

Kernel graduation is very similar to moving- (or local) weighted-average graduation (MWA), which is applied to equally spaced observations such as mortality rates or time series. The traditional problem with MWA is that it does not produce smoothed values at the ends of the table. However, in recent years this problem has been addressed in a series of papers by Greville [28], [29], [30] and also by Hoem and Linnemann [39]. In addition, London [45] and Ramsay [50] consider relaxing the assumption of constant variance in an MWA graduation. The kernel estimator in Equation (2) can be viewed as a continuous form of MWA graduation by expressing it as

$$\hat{q}'_i = \sum_{j=1}^{n} S_{ij}\, q'_j, \qquad \text{where} \qquad S_{ij} = \frac{K_b(x_i - x_j)}{\sum_{j=1}^{n} K_b(x_i - x_j)}, \tag{4}$$

so that $\Sigma_{j=1}^{n} S_{ij}=1$. This suggests that kernel graduation is very similar to MWA graduation. Although Equation (4) produces graduated rates at the ends of the table, we need to consider the properties of the estimator in those regions. Also, kernel graduation is not restricted to equally spaced crude rates, and it can be used to interpolate the mortality rate between the ages for which we have crude rates. In this way, it reflects the fact that age is a continuous variable, whereas MWA treats age as being discrete. However, unequally spaced observations may result in an increase in bias in the Nadaraya-Watson estimator, depending on the distribution of the observed data and on the curvature of the true mortality curve (Chu and Marron [11], Gavin et al [25]). Another contrast with MWA is that the bandwidth parameter in kernel graduation can be varied continuously, whereas the range of a MWA graduation is varied discretely.

It can be shown that the Nadaraya-Watson estimator leads to biased estimates of the true mortality rate [54]. However, the increase in bias leads to a reduction in variance, and so a trade-off between the two can be made through the bandwidth. This governs the amount of smoothing in the graduation process, in a continuous manner. The value for b may be a subjective choice, or it may be chosen as a function of the data. Too large a value for the bandwidth produces a smooth set of graduated rates at a cost of a lack of fidelity to the data. If the bandwidth is too small, then the converse is true. Cross-validation is one method for choosing an objective value for the bandwidth, and it is defined in Section 2.3. In the related problem of density estimation, Scott [54] discusses various other ways of selecting a value for the bandwidth. One disadvantage of Equation (2) is that the bias increases

near the ends of the mortality table. This problem is addressed in Section 3. Notice that the bandwidth in Equation (2) is fixed across the entire age range. A more general approach is to allow the bandwidth to vary with age. For example, the bandwidth could be inversely related to the sample size for that age; this leads to a variable or adaptive kernel estimator, which is the main topic of this paper.

The paper is set out as follows: Section 2 discusses the basic ideas needed for kernel graduation; Section 3 describes a method for improving the Nadaraya-Watson estimator near a boundary; Section 4 defines and discusses more general adaptive kernel estimators that allow the bandwidth to vary with age; two mortality tables are considered in Section 5; and finally we summarize our conclusions in Section 6.

2. KERNEL GRADUATION

We start by considering transformations of the data. Thereafter the graduation process is carried out on the transformed scale before back-transforming is used to obtain the graduated rates. In Section 2.2, we discuss the graduation process in detail and in Section 2.4 provide an illustration. Also in this section, we briefly consider the use of diagnostic tests, standard tables, duplicate policies, and measures of smoothness.

2.1 Transforming Mortality Data

Before the model is applied, a key part of any data analysis is to consider transforming the data into a more tractable form that reflects the strengths of the model or that more clearly reveals the structure of the data. In parametric graduation, for example, it may be easier to transform the data and work with a linear model than to graduate the raw rates using a more mathematically demanding nonlinear model. The same philosophy applies in nonparametric graduation. In this section, we consider transforming the crude rates before graduating and then back-transforming to obtain our estimate of the true rates.

Several transformations were considered, such as taking logs of the mortality rates and ages separately and combined and using the logit, Weibull, Gompertz, and $\sin^{-1}(\sqrt{\hat{q}_i})$ transformations. For example, if the transformed crude rates broadly follow a straight line, then this may lead to reduced bias over much of the age range, if the data are also evenly spaced. We consider this effect in more detail in Section 3. Because of the relatively large differences in mortality rates across the age range, the transformed data also

result in a more evenly spread scatter plot. In this case, we are aiming to ensure that the residuals in Equation (1) have a constant variance. Nielsen [48] offers a decision theoretic approach to bias reduction via transformations.

From Equation (1), $E(\dot{q}_i|x_i)$ is the expected proportion of lives aged x_i who died during the period of investigation. A commonly used transformation, t, in binary analysis is the logit (or log-odds) transformation. For our application, we have

$$\dot{q}_i^t = \ln \frac{\dot{q}_i}{(1 - \dot{q}_i)}$$

with back-transform

$$\hat{q}_i = \frac{\exp\left\{\sum_{j=1}^{n} S_{ij}\, \dot{q}_j^t\right\}}{1 + \exp\left\{\sum_{j=1}^{n} S_{ij}\, \dot{q}_j^t\right\}},$$

for $i=1, \ldots, n$. By smoothing on a logistic scale and then back-transforming, we are guaranteed that $0 \leq \hat{q}_i \leq 1$. This transformation also reflects the fact that small changes when the mortality rate is near zero are as important as larger changes when the mortality rate is much higher. Renshaw [51] provides further motivation for this transformation, based on the theory of generalized linear models. Note that binary data are often assumed to be independent, but this may not be the case for mortality data due to migration between ages during the period of investigation. This leads us to look for smooth relations between neighboring rates by merging information from individuals with similar ages.

For the Gompertz transformation, we fit $\ln(-\ln(1-\dot{q}_i))$ to x_i, and for the Weibull, we fit $\ln(-\ln(1-\dot{q}_i))$ to $\ln(x_i)$, where $i=1, \ldots, n$. Now the x-axis no longer has evenly spaced observations, but this does not present any computational problem for the kernel method, unlike the related MWA graduation. However, this transformation will induce some bias when we fit a local constant because more of the observations will now lie in the interval (x_i, x_i+b) than in (x_i-b, x_i) [25].

Many other transformations are possible ([9], [17], [18]), but their relative merits are beyond the scope of this paper. Overall, the choice of transformation remains subjective, and the relative success of a particular

transformation seems to depend on the data set. For the examples in Section 5, we have chosen the logit transformation.

2.1.1 Crude Rates with No Deaths

One potential problem with transformations that involve taking logs is that the transformed crude rate is not defined for ages at which no deaths are recorded, $d_i=0$. This often happens at older ages, with small data sets. A solution applicable to any transformation is to group together ages for which there are relatively few deaths. The cumulative number of deaths and amount of exposure for the group could be attributed to the midpoint of the group [4].

2.2 Building a Smoother

One way of implementing the Nadaraya-Watson estimator, given in Equation (2), is to place a kernel function at the point for which we wish to estimate the true rate of mortality and then form a weighted average over all the crude rates, where the weight attached to each crude rate is the value of the kernel function at that age.

The kernel has the same basic shape at each age x_i. Let the weight attached to the point x_j to estimate the true curve at x_i be denoted by $S_{ij}=cK_b(x_i-x_j)$, where $c^{-1}=\Sigma_{j=1}^{n} K_b(x_i-x_j)$ is a normalizing constant. This gives the $1\times n$ matrix of weights needed to estimate the true value of the curve at x_i. By sliding the kernel function along the x-axis and centering it at every point for which we wish to estimate the mortality curve, we can build up a matrix $\mathbf{S}=\{S_{ij};j=1, \ldots, n\}$. The i-th row of the matrix contains the n weights allocated to the transformed crude rates, to estimate the true mortality rate at that age. The matrix has a row for every point at which we wish to estimate the true curve. Without loss of generality, we constrain the set of estimated ages to be the same as the set of observed ages, because this is often the case for mortality data. This gives an $n\times n$ matrix of weights that we call a *smoother* matrix (or a *hat* matrix). To help produce smooth graduated rates, we use weights that decrease smoothly towards zero as $|x_i-x_j|$ increases. So if we let $\dot{\mathbf{q}}'$ be the n-dimensional vector of transformed crude rates and $\hat{\mathbf{q}}'$ be the vector of transformed graduated rates, then the smoother $\mathbf{S}$ defines the relationship between them as

$$\hat{\mathbf{q}}' = \mathbf{S}\dot{\mathbf{q}}',$$

where $\mathbf{S}$ has been renormalized as in Equation (4), so that $\Sigma_{j=1}^{n} S_{ij}=1$, for $i=1, \ldots, n$. Thus, the transformation from the crude to the estimated rates

is achieved by filtering the crude rates through the smoother. This equation succinctly summarizes kernel graduation. In particular, for the i-th element of $\hat{\mathbf{q}}'$, we get Equation (2). Notice that the kernel smoother is linear (or distributive); that is, $\mathbf{S}(a\mathbf{v}_1+b\mathbf{v}_2)=a\mathbf{S}\mathbf{v}_1+b\mathbf{S}\mathbf{v}_2$, for constants a and b and vectors $\mathbf{v}_1$ and $\mathbf{v}_2$. So from Equation (1), if we believe that the transformed crude rates consist of the transformed, unknown true rates $\mathbf{q}$ plus a vector of residuals $\mathbf{r}$, we arrive at $\hat{\mathbf{q}}'=\mathbf{S}\hat{\mathbf{q}}'=\mathbf{S}(\mathbf{q}'+\mathbf{r})=\mathbf{S}\mathbf{q}'+\mathbf{S}\mathbf{r}$. We believe that by graduating the error term $\mathbf{S}\mathbf{r}$, we reduce it in a way that more than compensates for any induced bias, which we define as the difference between the true and estimated mortality rates on the transformed scale.

Many other nonparametric smoothers are also linear such as the running-mean, running-line, cubic smoothing spline (Whittaker graduation), regression spline, and locally weighted running line, but there are also nonlinear smoothers such as the running median smoother. Hastie and Tibshirani [38, chapters 2 and 3] offers an excellent introduction to nonparametric smoothers, drawing out the similarity between these methods. Verrall [58] views Whittaker graduation as a dynamic generalized linear model.

2.2.1 Choice of Kernel Function

Some kernel functions such as the Epanechnikov kernel [19],

$$K(x) = \begin{cases} 3(1 - x^2)/4, & \text{for } |x| \leq 1; \\ 0 & \text{otherwise,} \end{cases}$$

have greater theoretical justification than others. This particular kernel minimizes the mean squared error asymptotically. Another potential kernel is one that minimizes the variance of the estimated curve, in some sense, and one such kernel is explored in Gavin, Haberman, and Verrall [24]. The current literature indicates that the choice of kernel function is not as influential as the value of the bandwidth. So for convenience, we use the standardized normal kernel defined in Equation (3) throughout this paper. In general, it would be computationally cheaper to use a truncated kernel such as the Epanechnikov kernel.

2.3 Bandwidth Selection

The choice of bandwidth in Equation (2) is important. Although it is informative to choose the bandwidth by trial and error, it is also convenient to have an objective, risk-based method for selecting the best value for b. The literature on data-driven methods for selecting the optimal bandwidth

is vast and continues to grow. Cross-validation [57] is just one such method that is commonly used and simple to understand. This technique has been used by Brooks, Stone, Chan, and Chan [7] to smooth some mortality tables using Whittaker graduation, and Gregoire [27] offers a more rigorous approach.

Working on the transformed scale, cross-validation simultaneously fits and smooths the data by removing one data point at a time, estimating the value of the curve at that missing point, and then comparing the estimate to the omitted, observed value. So our cross-validation statistic or score, $CV(b)$, is

$$CV(b) = n^{-1}(\dot{\mathbf{q}}^t - (\hat{\mathbf{q}}^t)^{(-i)})^T (\dot{\mathbf{q}}^t - (\hat{\mathbf{q}}^t)^{(-i)}) = n^{-1} \sum_{i=1}^{n} (\dot{q}_i^t - (\hat{q}_i^t)^{(-i)})^2, \quad (5)$$

where $(\hat{q}_i^t)^{(-i)}$ is the estimated value at age x_i computed by removing the crude rate at that age on the transformed scale. It is sometimes called the jackknifed fit at x_i. It is easy to calculate $(\hat{q}_i^t)^{(-i)}$: set the i-th weight in the i-th row of $\mathbf{S}$ to zero and renormalize the weights. That is,

$$(\hat{q}_i^t)^{(-i)} = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} S_{ij}\, \dot{q}_j^t}{(1 - S_{ii})} = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} \dot{q}_j^t\, K_b(x_i - x_j)}{\sum_{\substack{j=1 \\ j \neq i}}^{n} K_b(x_i - x_j).} \quad (6)$$

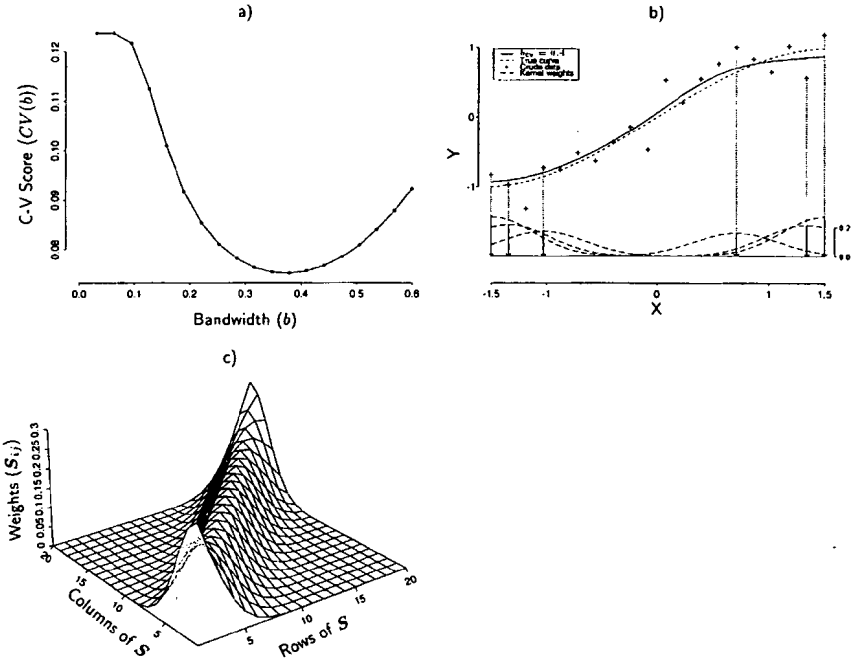
To further speed computation, we can use the relation $\dot{q}_i^t - (\hat{q}_i^t)^{(-i)} = (\dot{q}_i^t - \hat{q}_i^t)/(1 - S_{ii})$. The bandwidth that minimizes $CV(b)$ is referred to as the cross-validation bandwidth, b_{cv}, and we find it by systematically searching across a suitable bandwidth region. So we need to balance the benefit of getting close to the optimal bandwidth against the cost of a detailed search. Scott [54] suggests that getting to within ± 15 percent often suffices. For convenience, the bandwidth is always selected by cross-validation in this paper, although some evidence suggests that it may undersmooth the data ([32], [34], [53], [54]).

2.4 Example 1

The techniques outlined in Sections 2.2 and 2.3 are illustrated in Figure 1b, which shows a scatter plot of 20 evenly spaced points where $Y \sim \sin(X) + N(0, 0.05)$. No transformation is needed in this simple example.

FIGURE 1

A CURVE IS FITTED TO SOME RAW DATA, USING CROSS-VALIDATION TO SELECT THE SMOOTHING PARAMETER. THE CROSS-VALIDATION SCORE IS SHOWN IN PLOT A). PLOT B) SHOWS THE RAW DATA, THE TRUE CURVE AND THE FITTED CURVE. THE WEIGHTS IN THE SMOOTHER MATRIX, S, ARE SHOWN IN PLOT C).



The bandwidth used in Figures 1b and 1c is chosen by cross-validation. Figure 1a shows the cross-validation scores for about 20 evenly spaced bandwidth values. The optimal cross-validation bandwidth is about b_{cv}=0.4. In Figure 1b, the true curve and the best fit using a normal kernel and the best cross-validation bandwidth are shown. Six normal kernel functions have been superimposed on the bottom of this plot to show the relative weights attached to each of the observed values when estimating the true curve at the six points indicated by arrows. Each arrow is connected to its corresponding kernel and observed data point. At both ends, the normal kernel overlaps the boundary, but the denominator in Equation (2) is now summed over fewer data points, forcing the kernel to rise slightly. This reflects the fact that we have less information at the boundaries.

The kernel function associated with the i-th point is used to calculate the weights in the i-th row of the 20×20 smoother and is shown in Figure 1c. The weights are shown as the height along the i-th row of the surface. For values in the central region the weights form a normal kernel, but as the point at which we are estimating the true curve moves towards the boundaries the kernel overlaps the boundary. This causes the height of the kernel to increase because fewer observations are available. Notice that all the weights in the smoother are non-negative. For evenly spaced data and a kernel with bounded support, there can be computational savings when the data in the center of the table are estimated, because the denominator in Equation (2) is constant.

2.5 Diagnostic Checks

Having produced graduated rates on the transformed scale, we now consider diagnostic plots of the results to help confirm that the assumptions made by the model, in Equation (1), are valid.

We need to check that the estimator is unbiased with a constant variance. The former assumption means that we require the residuals to have a mean of zero. Plotting the residuals from Equation (1) against the estimated mortality rates on the transformed scale and against age should reveal no clear pattern. One way to check this is to smooth the residuals and get a fairly flat line about zero.

Alternatively, after graduating the crude rates and back-transforming, we can use the mean and variance of the binomial distribution to calculate the standardized deviation between actual and expected deaths,

$$\frac{(d_i - e_i\hat{q}_i)}{\sqrt{e_i\hat{q}_i(1 - \hat{q}_i)}}, \qquad \text{for } i = 1, \ldots, n, \tag{7}$$

on the grounds that most of the samples at each age are large. We expect this statistic to have a mean of zero and most of the values to be less than two. Note that the distribution may not be normal. If a suitable standard mortality table is available, then we might use that in the denominator of (7).

We also require independent crude rates. One diagnostic check is to examine plots of the estimated autocorrelation of the residuals. To do this, we need equally spaced residuals, so the transformation of the data must be restricted to the mortality rate and not age as well.

Several other diagnostic plots and nonparametric tests could be considered ([2], [13], [21]).

2.6 Reference to a Standard Mortality Table

We may wish to standardize the data relative to a suitable graduated mortality table so that the standard table acts as a prior assumption. This information can be incorporated into a Whittaker or a Bayesian graduation ([44], [46]). One simple way to use this prior knowledge in a kernel graduation is to subtract the crude rates from the standard rates, smooth the residuals, and add the smoothed residuals to the standard table to get the kernel graduated rates, all on the transformed scale ([4], [16]). Subtracting the standard table rates from the crude rates may filter out much of the curvature in the true rates, assuming that the standard table rates are similar in shape to the true rates. This may mean that the residuals are scattered about a simple curve, such as a constant or a straight line. When we investigate the bias of the Nadaraya-Watson estimator, in Section 3, we see that it has a relatively small bias in such situations.

Using a standard table is one way of ensuring that the graduated results reflect known theoretical or empirical models. For example, a small company might want to adjust a standard table to reflect the company's own particular circumstances, such as underwriting practices or geographical location. It is also possible to ensure that monotonicity in the standard table is reflected in the graduated rates by choosing a large enough bandwidth. Unfortunately this is a rather trivial case, because we would simply be adding a constant to the standard table. However, imposing a monotonicity constraint on a relatively simple nonparametric method and expecting good results is being rather optimistic. In Section 5, we consider using a standard table when measuring the relative difference between select and ultimate mortality rates.

2.7 Duplicate Policies

For duplicate policies, an additional complication may arise if the data are based on policy-years rather than person-years. This occurs when a policyholder buys multiple policies, perhaps from different life offices at different times, and consequently is counted more than once in the investigation. As a result, the residuals in Equation (1) may not be independent. This area presents considerable difficulty, because there is little information available that can be justifiably used to filter this undesirable effect from the data. For ultimate data, it could have a potentially significant influence on the number of observed deaths. One possible approach ([14], [40]) is to adjust the data by age, using a variance ratio to reduce the amount of exposure, and Renshaw [51] provides a more recent discussion of this topic. Another

possible problem is that correlated observations can affect the cross-validation score. Hart and Wehrly [36] and Altman [1] offer some adjustments to the score statistic for resolving this problem. The issue is not pursued further in this paper, partly for simplicity but also because the adaptive bandwidth, used in Section 4, does not depend heavily on the choice of global bandwidth. However, in Section 4, we briefly mention an adjustment that might be made to one of the adaptive kernel models to help compensate for duplicate policies.

2.8 Choice of Smoothness Criterion

Smooth graduated rates are a primary objective. There are various ways of measuring this criterion, but ultimately it is a subjective choice that depends on the context in which the results are to be used. With the original scale, a traditional actuarial approach is to repeatedly calculate differences of the graduated rates and confirm that the third or fourth differences are random and small by using standard statistical tests. Bloomfield and Haberman [4] define a relative measure of smoothness, which expresses the k-th difference of the graduated rates relative to the graduated rates, as $D^k = (\hat{q}_i/|\Delta^k \hat{q}_i|)^{1/k}$, where Δ^k is the usual forward differencing operator applied repeatedly k times. Other measures of smoothness ([5], [8]) require monotonically increasing or increasing-convex rates over some region of the age range. The former measure requires that

$$\{\hat{q}_i \leq \hat{q}_{i+1} \text{ for } i = 1, \ldots, n-1\},$$

and the more stringent, latter measure requires graduated rates that satisfy

$$\hat{q}_i - \hat{q}_{i-1} \leq \hat{q}_{i+1} - \hat{q}_i \qquad \text{for } i = 2, \ldots, n-1, \tag{8}$$

excluding the first year of life and males in their 20s. In general, a kernel graduation cannot be guaranteed to preserve monotonicity, unless this prior information is built into the kernel model. Referring to a standard mortality table may be one way of doing this.

3. EXPLICIT ALLOWANCE FOR THE BOUNDARIES

3.1 A Boundary-Correcting Kernel

Figure 1b shows that values in the middle of the age range enjoy full support, for all practical calculations. However, as the normal kernel slides towards young or old ages, it increasingly overlaps the ends of the table and

the resulting truncated kernel leads to an increase in bias. For example, at the youngest and oldest ages, half the kernel function will extend beyond the ends of the table.

To consider this problem further, we need to calculate the bias of our estimator. We start with a Taylor series expansion of the Nadaraya-Watson estimator in Equation (2),

$$E(\hat{q}_i^t) = q_i^t + \frac{\sum_{j=1}^n (x_j - x_i) K_b(x_j - x_i)}{\sum_{j=1}^n K_b(x_j - x_i)} (q_i^t)' + R, \tag{9}$$

where $(q_i^t)'$ denotes the slope of the transformed true curve at age x_i and R is a remainder term consisting of higher-order derivatives. For an age x_i in the middle of the table, the coefficient of the $(q_i^t)'$ term is zero if the crude rates are evenly spaced. However, when estimating rates at the youngest and oldest ages, all the other crude rates will lie to the right and to the left, respectively. As a result, the $x_j - x_i$ term in the coefficient of $(q_i^t)'$ has the same sign for $j=1, \ldots, n$, so that this coefficient is non-zero. This means that there is increased bias near the ends of the table. A comparison between the bias in the Nadaraya-Watson and the related Copas-Haberman kernel estimator is considered in Gavin, Haberman, and Verrall [25].

To improve the estimate at the boundaries, Hall and Wehrly [33] suggest reflecting the data so that the original data lie in the interior of an enlarged data set. In this way, the original data are less influenced by boundary effects.

We use an alternative method suggested by Rice [52]. Rice's extrapolation method is based on a linear combination of two different kernels with different bandwidths to eliminate the first-order bias. Suppose our two estimates are $\hat{q}_i^1$ and $\hat{q}_i^2$; then from Equation (9), we get

$$E(\hat{q}_i^1) = q_i + C_1 q_i' + R_1$$

$$E(\hat{q}_i^2) = q_i + C_2 q_i' + R_2,$$

where C_1 and C_2 are the coefficients of the first-order terms and R_1 and R_2 are the remainder terms for $\hat{q}_i^1$ and $\hat{q}_i^2$, respectively. Notice that C_1 and C_2 depend on age and not on the true mortality curve, so by a suitable linear combination of the two estimates, we can eliminate the q_i' term. This means that the bias of our estimator at the boundary does not depend on the slope of the mortality curve but only on higher-order terms such as curvature. In the same spirit but in the context of density estimation, Jones [42] suggests redefining the kernel function to be a linear combination of $K(x)$ and $xK(x)$. This leads to a kernel function for the right-hand boundary

$$K_b^R(x) = \frac{[a_2(p) - a_1(p)x]K_b(x)}{a_0(p)a_2(p) - a_1^2(p)}, \tag{10}$$

where $a_l(p)=\int_{-\infty}^{p} u^l K_b(u)du$ and $p=x/b$, which can be used to reduce the bias near the upper boundary. The variable p measures the distance from the point at which we are calculating the mortality rate to the right-hand boundary in units of bandwidth. The variable x measures the distance from the point at which we are calculating the mortality rate to each of the n crude rates, again in units of bandwidth. Substituting the normal kernel from Equation (3) for K_b in Equation (10), we get

$$K_b^R(x) = \frac{[\Phi(p) + (x - p)\phi(p)]\phi(x)}{\Phi(p)[\Phi(p) - p\phi(p)] - \phi^2(p)}, \tag{11}$$

where $\Phi(x)=\int_{-\infty}^{x}\phi(y)dy$ and $\phi(y)$ is as defined in Equation (3). For ages that are closer to the left-hand boundary, that is, younger ages, some obvious adjustments to the formula yield

$$K_b^L(x) = \frac{\{[1 - \Phi(p)] + (p - x)\phi(p)\}\phi(x)}{[1 - \Phi(p)]\{[1 - \Phi(p)] + p\phi(p)\} - \phi^2(p)}. \tag{12}$$

The transformed kernel functions, K_b^L and K_b^R, both behave like the standard Gaussian kernel in the middle of the age range. That is, if $p>2$, then $a_0(p)\rightarrow 1$ and $a_1(p)\rightarrow 0$, so $K_b^L \approx K_b^N$ and $K_b^R \approx K_b^N$. As the age at which we are estimating the curve moves closer to the boundary, the weights change shape, becoming asymmetric and negative over some regions, as is shown in Example 2.

3.2 Implementing the Boundary-Correcting Kernel

It is possible to combine the two kernel functions, K_b^L and K_b^R, into a single smoother by first deciding which boundary is closest to the point at which we are estimating the curve. This approach worked satisfactorily for several data sets. However, if the distance from the center of the data to the boundaries is roughly two bandwidths or less, then a kink develops in the graduated rates as the smoother switches from using K_b^L to K_b^R, moving from left to right across the age range. Example 3 in Section 5.1 has a small age range, and cross-validation chooses a large global bandwidth. This results in a noticeable jump in the graduated rates at the central ages.

An ad hoc solution to this large boundary problem is to smooth the data using the left-hand and right-hand kernels, K_b^L and K_b^R, separately. This gives

two sets of graduated rates that are blended linearly. So at the left-hand boundary, weights of 1 and 0 are given to the left-hand and right-hand boundaries, respectively. The weights change linearly across the age range to become 0 and 1, respectively, at the right-hand boundary. Benjamin and Pollard [3] mention other ways of blending the data, but this simple linear approach is adequate for our purposes. However, a referee has drawn our attention to a recent paper by Hart and Wehrly [36], which describes kernels that deal with large boundary regions, and these models may be more suitable in this context.

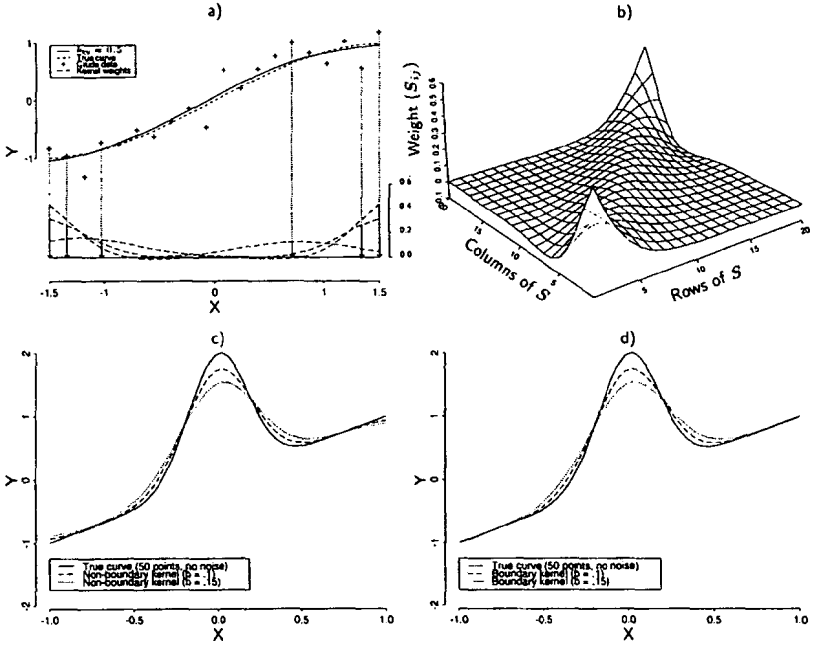
3.3 Example 2

Figure 2a shows the data from Figure 1b fitted by using a linear combination of K_b^L and K_b^R, instead of just K_b^N. The cross-validation curve arising from the boundary-correcting smoother is similar in shape to that given in Figure 1a, so it is not shown. It results in b_{cv}=0.5. However, the fitted values and the kernel functions superimposed on the bottom of Figure 2a are quite different from those in Figure 1. The kernel functions shown are those used to estimate the curve at the same six points as in example 1. For x_{15}, the kernel is almost the same as that of a normal kernel, but at x_2, x_3 and x_{19} the function becomes more truncated and its mode increases. When estimating the curve at both boundaries, x_1 and x_{20}, and moving towards the interior, we see that the weights attached to the other observations decrease rapidly, becoming negative and then gradually increasing back towards zero. So the smoother can take negative values.

If we consider the value of the weights in the smoother matrix **S** to be the height of a surface above a plane, then we can plot the surface using a mesh, and this is shown in Figure 2b. The view in Figure 2b has S_{11} as the closest point on the surface and S_{nn} as the furthest away point. The weights needed to estimate the i-th age come from the i-th row of the smoother. For example, the last row of the matrix in Figure 2b contains the weights needed to estimate x_{20}, and these weights are drawn in the bottom-right corner of Figure 2a. For values in the central region, the weights form a normal kernel, but as the point at which we are estimating the true curve moves towards the boundaries, the kernel becomes asymmetric and some of the weights are negative. From Equation (10), we can see that p measures the distance between a given point and the boundary in units of bandwidth, so ages for which $p \geq 2$ have kernel functions that are almost the same as a standardized normal, with asymptotic equality as $p \rightarrow \infty$.

FIGURE 2

PLOT A) HAS THE SAME DATA AS IN FIGURE 1B), BUT HERE THE DATA ARE SMOOTHED USING A LINEAR COMBINATION OF K_b^L AND K_b^R. PLOT B) IS A SURFACE PLOT OF THE BOUNDARY-CORRECTING SMOOTHER. PLOTS C) AND D) SHOW THE BENEFITS OF REDUCING FIRST-ORDER BIAS IN THE NADARAYA-WATSON ESTIMATOR BY USING A BOUNDARY ADJUSTMENT.



The benefit of allowing the kernel to take negative values at the boundary is that we can reduce the first-order bias term in Equation (9) by building a kernel from Equations (11) and (12). So if the transformed, true curve is approximately a straight line, we can produce better estimates even at the boundary. Figures 2c and 2d show a true curve, which is a straight line at the boundaries. The data are observed without noise, so any error in estimation is due to bias. Fifty evenly spaced observations (not shown) are used, so we might expect any reasonable estimator to do well under these ideal conditions. Both estimated curves use the Nadaraya-Watson estimator from Equation (2). Without adjustment the Nadaraya-Watson estimator is noticeably biased at the boundary in Figure 2c, and this bias increases as the

bandwidth increases. The boundary-correcting kernel in Figure 2d can correctly estimate the straight line part of the true curve, even at the boundary. From Equation (9), we know that both estimators have second-order bias terms, so they both incorrectly estimate the curvature in the middle of the graph, underestimating peaks and overestimating troughs.

If we are interested only in graduating the interior of the age range and the bandwidth is small, then the bias caused by the boundary may be negligible. If this is not the case, then the effort required to reduce the extra bias at the boundary complicates the Nadaraya-Watson estimator. This is analogous to the complications that arise when MWA is adjusted to produce graduated rates at the ends of the table ([28], [29], [30], [39]). Example 3 in Section 5.1 suggests that if the age range is small, then the extra bias due to the boundaries may be serious. In such cases, using one of the techniques mentioned above to reduce this bias may be well rewarded. Some authors have argued that other kernel estimators can be adjusted more easily than the Nadaraya-Watson to allow for boundary problems. Chu and Marron [11] offer a very readable comparison between the two most popular kernel estimators, namely, the Nadaraya-Watson and the Gasser-Müller estimators.

Another possible complication is that the boundary problem may force cross-validation to select a smaller bandwidth at the boundary to reduce the bias, but this may lead to undersmoothing in the middle of the table. Using an adaptive kernel estimator allows the bandwidth to vary across the table, so it may help to alleviate this problem.

4. AN ADAPTIVE KERNEL ESTIMATOR

In previous sections, the kernel functions have always had a fixed or global bandwidth, so once b is chosen, it remains constant. Rather than restricting the bandwidth to a fixed value, a more flexible approach is to allow the bandwidth to vary according to the reliability of the data. Thus, for regions in which the amount of exposure (sample size) is large, a low value for b results in an estimate that more closely reflects the crude rates. For regions in which the exposure is small, such as at old ages, a higher value for the bandwidth allows the estimate of the true rates of mortality to progress more smoothly. This means that at older ages we are calculating local averages over a greater number of observations, which reduces the variance of the graduated rates but at a cost of potentially greater bias. This technique is often referred to as a variable or adaptive kernel estimator.

4.1 Some Adaptive Models

We can build our knowledge of the amount of exposure into the basic model in Equation (2) in a number of ways:

- We can calculate a different bandwidth for each age at which the curve is to be estimated. Using that bandwidth, we then measure the distance from the age at which the curve is to be estimated to each of the observed ages. For example, assuming that the age to be estimated is x_i, we measure the distance from x_i to x_j using b_i, for $j=1, \ldots, n$. So the model is

$$\hat{q}_i^l = \sum_{j=1}^{n} S_{ij}\, \dot{q}_j^l, \qquad \text{where} \qquad S_{ij} = \frac{K_{b_i}(x_i - x_j)}{\sum_{j=1}^{n} K_{b_i}(x_i - x_j)} \tag{13}$$

  for $i=1, \ldots, n$. If the age to be estimated is not one of the observed ages, then we could smooth the empirical probability density estimate of age,

$$\hat{f}_i = \frac{e_i}{\sum_{j=1}^{n} e_j} \qquad \text{for } i = 1, \ldots, n. \tag{14}$$

- Alternatively, we can calculate a different bandwidth, b_j, for each observed age x_j, for $j=1, \ldots, n$. Then for each observed age, use the corresponding bandwidth to measure the distance from that observed age to the age at which the curve is to be estimated. For example, assuming that the age to be estimated is x_i, we measure the distance from x_i to x_j using b_j, for $j=1, \ldots, n$. This results in a new smoother

$$\hat{q}_i^l = \sum_{j=1}^{n} S_{ij}\, \dot{q}_j^l \qquad \text{where} \qquad S_{ij} = \frac{K_{b_j}(x_i - x_j)}{\sum_{j=1}^{n} K_{b_j}(x_i - x_j)} \tag{15}$$

  for $i=1, \ldots, n$.

The local bandwidth at each age is simply the global bandwidth multiplied by a local bandwidth factor, $b_i = b l_i^s$ for $i=1, \ldots, n$. The variation in exposure between different tables and between young and old ages within a table can be enormous. To dampen the effect of this variation, we have chosen

$$l_i^s \propto \hat{f}_i^{-s} \qquad \text{for } i = 1, \ldots, n \quad \text{and} \quad 0 \le s \le 1, \tag{16}$$

where s is a sensitivity parameter. Choosing $s=0$ reduces both models to the fixed bandwidth case, while $s=1$ may result in very large bandwidth variation, depending on the particular table. For convenience, we have chosen the inverse of $\max\{f_i^{-s}: i=1, \ldots, n\}$ as the constant of proportionality in Equation (16), so that $0 < l_i^s \le 1$, for $i=1, \ldots, n$. If there is a small amount of

exposure at age x_i, then l_i^s is large. This increases the size of the effective bandwidth, which in turn reduces the weight attached to the crude rate for that age. This allows us to apply more smoothing at those ages. The converse is true if the amount of exposure is large. The first example in Section 5 uses the model defined in Equation (13), and the second uses Equation (15) to graduate some mortality tables.

Once the local bandwidth factors are chosen, they remain fixed in both models, regardless of the location of the age that we are trying to estimate. So another possibility is to choose

$$l_{ij}^s = (e_j/e_i)^s, \qquad \text{for } i, j = 1, \ldots, n. \tag{17}$$

The sensitivity parameter is still necessary to dampen the extreme variations that can arise. In this case, the relative exposure is used to adjust the global bandwidth when a weight is attached to the j-th crude rate to estimate the true rate at the i-th age. This leads to

$$\hat{q}_i' = \sum_{j=1}^{n} S_{ij}\, \dot{q}_j', \qquad \text{where} \qquad S_{ij} = \frac{K_{b_{ij}}(x_i - x_j)}{\sum_{j=1}^{n} K_{b_{ij}}(x_i - x_j)}, \tag{18}$$

where $b_{ij}=bl_{ij}^s$ and $i, j=1, \ldots, n$. This model also offers the possibility of building in a variance ratio to allow for duplicate policies [15].

Clearly there is room for other models to be developed. In theory, we could try taking account of the shape of the true curve by using

$$l_i^s = (|(q_i')''|f_i)^{-s}. \tag{19}$$

Consider the true curve in Figures 2c and 2d to provide some motivation for this model. A formula such as $l_i^s=(|(q_i')''|f_i)^{-s}$ is saying that to improve the estimate in the center of Figure 2c, we should decrease the bandwidth as the amount of curvature in the true mortality curve increases, provided that the crude rates in that region are reliable. The true mortality rate, q_i, is unknown, so an initial estimate is required. We can use $\hat{f}_i$, as defined in Equation (14), as an estimate of f_i. Second differences could be used to approximate curvature, and we do not distinguish between positive or negative curvature.

We expect explicit allowance for exposure to be a beneficial feature in the models, because this factor directly influences the variability of the crude rates and exposure may vary enormously across the age range. It is worth asking whether models that allow for the shape of the true mortality curve as well as the amount of exposure are worthwhile. This would seem to

depend on the purpose of the graduation. If we are merely exploring the data, then the additional information derived might not justify the effort. However, if we wish to use a kernel estimator to check on a parametric graduation [2], then a more detailed model may be worth the effort. This is especially so for large tables where considerable time and effort have been invested in gathering and validating the data.

We do not consider models like Equation (19) further in this paper. Nor do we derive the properties of an adaptive kernel estimator that are more complicated than those of the fixed-bandwidth estimator [31]. Jones [41] considers an alternative approach using a model of the form

$$\hat{q}_i' = \frac{\sum_{j=1}^{n} w_j \, \hat{q}_j' \, K_b(x_i - x_j)}{\sum_{j=1}^{n} w_j \, K_b(x_i - x_j)},$$

where w_j are weights that could depend on the amount of exposure.

4.2 Choice of Parameter Values

For each of the models in the previous section, two parameters need to be considered: sensitivity, s, and global bandwidth, b.

The sensitivity parameter could be chosen by cross-validation. However, as s increases from zero, the adaptive kernel becomes more sensitive to the variation in exposure. The amount of variation in exposure can be very large for some mortality data sets, ranging from thousands of person-years at younger ages down to single figures at the oldest ages. In such cases, a large value for s may be unreasonable, because it might result in bandwidths for some ages being several times the age interval covered by the data. Therefore, this parameter is chosen subjectively. Once s has been chosen, cross-validation is still used to choose b.

5. SOME PRACTICAL EXAMPLES

In this section we illustrate how the adaptive kernel model might be used to graduate two mortality tables. These two tables were chosen because the first has relatively little variation in exposure over the age range, while the second has a much greater variation. For both tables, we consider letting the bandwidth vary across the age range. Trial and error indicates that a doubling

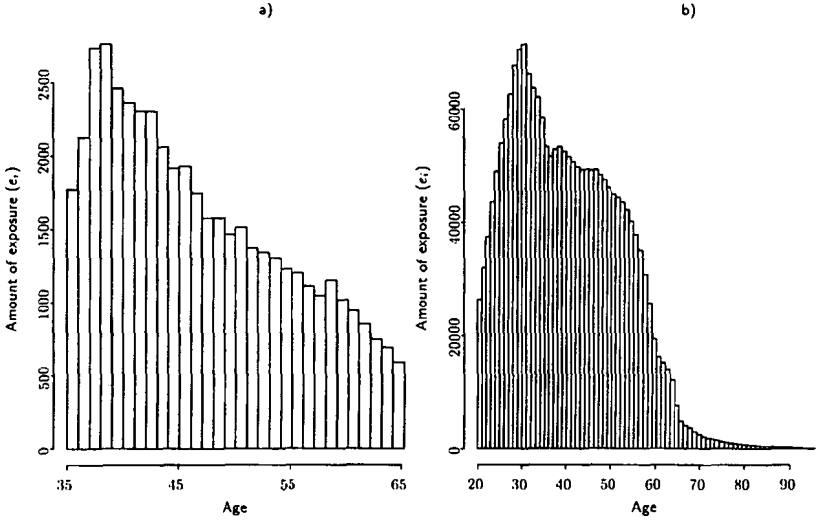
of the bandwidth, from a minimum at ages with high exposure to a maximum for ages with the lowest exposure, gives reasonable results.

5.1 Example 3

Figure 3a shows a bar plot of the amount of exposure for the crude mortality rates taken from Broffitt [6]. Broffitt adopts a Bayesian approach to graduation. The same data set has subsequently been considered by Carlin [8] using the Gibbs sampler to implement a Bayesian model. The data cover only a small age range, for which it might be expected that the true mortality rates are monotonically increasing. The decrease in exposure with age is typical of mortality tables reflecting the fact that there are relatively fewer older people and that whole-of-life and endowment policies are less likely to be sold to older people, in the case of the females table (Figure 3b). In comparison to the second table, the first has a relatively small amount of exposure and the variation in exposure over age is relatively small. The boundary-correcting kernel discussed in Section 3 along with the adaptive kernel defined in Equation (13) are used to graduate this table.

FIGURE 3

BAR PLOTS OF A) THE AMOUNT OF EXPOSURE FOR THE DURATION SIXTEEN-OR-MORE, MALE ULTIMATE DATA TAKEN FROM BROFFITT [6] AND B) THE AMOUNT OF EXPOSURE FOR THE DURATION TWO-OR-MORE, FEMALE ASSURED LIVES 1975–78 TABLE [14].



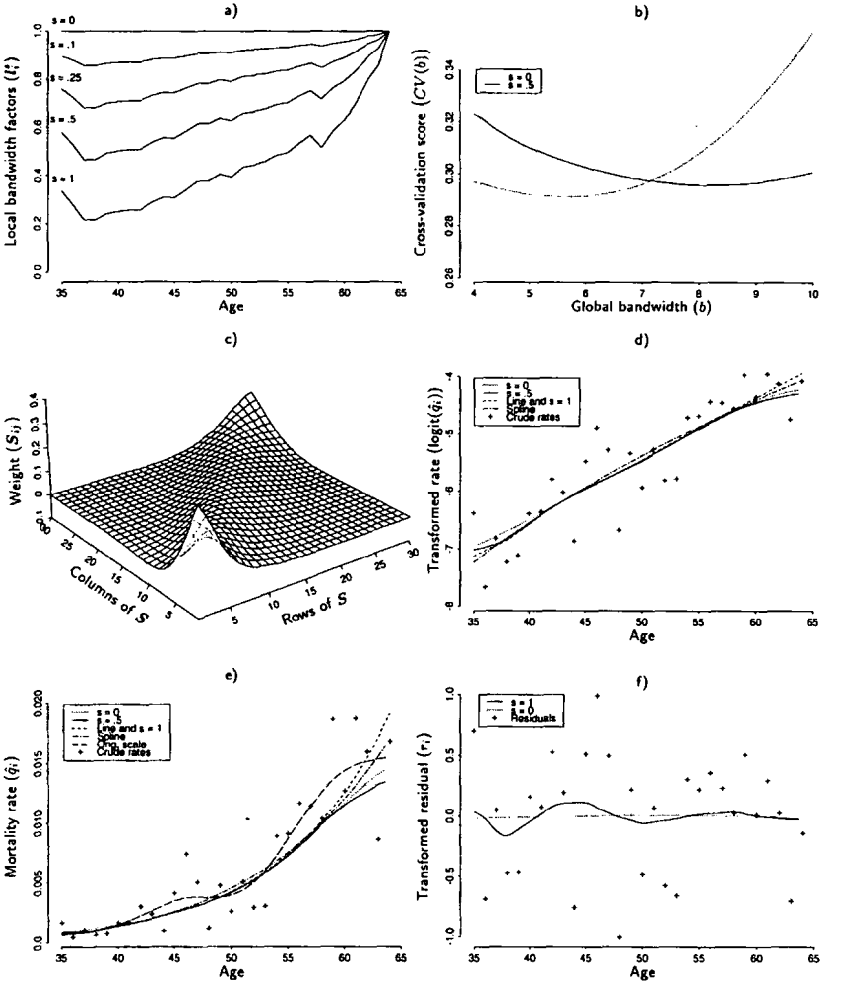
The exposure in Figure 3a decreases with age, though in a less dramatic fashion than is often the case with mortality tables. The resulting local bandwidth parameter values, l_i^s, for various values of the sensitivity parameter, s, are shown in Figure 4a. The observed exposures decide the shape of the local bandwidth curves, but the sensitivity parameter, s, determines the magnification of that shape, becoming more pronounced as $s \rightarrow 1$. Notice that for $s=0$ the local bandwidth curve has a constant value of 1. In this case we are ignoring the variation in exposure, which gives a fixed-width estimator. From Figure 4a, where $s=0.5$, the minimum local bandwidth factor is about 0.5, at age 40. This means that the bandwidth at the oldest ages is about double that at the younger ages. A bandwidth that approximately doubles across the age range produces reasonable results.

After the data have been transformed using the logit transformation, discussed in Section 2.1, cross-validation is then applied to select the optimal bandwidth, b_{cv}. For $s=0$ and $s=0.5$, the cross-validation score, from Equation (5), is calculated for a range of global bandwidths, using the smoother defined in Equation (13). The results are shown in Figure 4b. The cross-validation curve for $s=0$ suggests that there is little to choose between bandwidth values up to about 6. This provides some support for using a subjective choice in or about this value, if desired. By using $s=0.5$, the cross-validation bandwidth is larger. Because we have already obtained the shape and magnification of the local bandwidth factors, this process of cross-validation decides the global value at which the bandwidth curve, from Figure 4a, is located.

The smoother for the case $s=0.5$ is shown in Figure 4c. The basic shape is the same as that of Figure 2. However, row 1 has a smaller effective bandwidth than row 30. So in row 1 the weights decrease rapidly to zero but in row 30 the weights decrease more slowly in order to smooth the older ages more. Rows in the middle of the smoother correspond to ages in the middle of the table and are approximately normal in shape. For example, for element S_{11} of the smoother, we have $b_{cv} l_{35}^{s=.5} \approx 8 \times 0.6$, but for $S_{30\,30}$, we have $b_{cv} l_{64}^{s=.5} \approx 8 \times 1$, giving weights of 0.34 and 0.19, respectively. So the weights in row 1 decrease more slowly than the weights in row 30, because each row is standardized to sum to one. As a second example, consider the center of the surface where $S_{15\,15} \approx 0.08$. This is the weight attached to the crude rate at age 50 when the true rate at age 50 is estimated. It can be calculated from a global bandwidth of about 8 and a local bandwidth factor of about 0.6 for age 50, giving a weight equal to a normal density with mean 0 and a standard deviation of 8×0.6 evaluated at 0.

FIGURE 4

A MORTALITY TABLE TAKEN FROM BROFFITT [6] IS GRADUATED USING EQUATION (13). PLOT A) SHOWS THE LOCAL BANDWIDTH FACTORS l_i^s FOR DIFFERENT VALUES OF THE SENSITIVITY PARAMETER. PLOT B) SHOWS CROSS-VALIDATION SCORES FOR EACH OF THE CASES $s=0$ AND $s = 0.5$. THE SMOOTHER FOR THE CASE $s=0.5$ IS SHOWN IN PLOT C). PLOTS D) AND E) SHOW THREE SETS OF KERNEL GRADUATED RATES ON THE LOGIT AND ORIGINAL SCALES, RESPECTIVELY. FINALLY, PLOT F) SHOWS THE RESIDUALS FROM FITTING A STRAIGHT LINE TO THE TRANSFORMED CRUDE RATES IN D) AND SMOOTHED RESIDUALS USING EQUATION (13).



This data set is unusual in that b_{cv} is relatively large compared to other data sets that were tested, especially considering that the age range is quite small. As a result, the distance from age 49 to the left-hand boundary and from age 50 to the right-hand boundary is approximately twice b_{cv}, when $s=0.5$. Consequently, many of the kernels in the center of the table are not quite normal in shape. This gives a notable discontinuity in the fitted values unless the graduated rates are blended in some way. An ad hoc solution is explained in Section 3.2.

The results of the two graduations, using $s=0$ and $s=0.5$, are shown on the transformed scale (logit) in Figure 4d and on the original scale in Figure 4e. For comparison, a natural, cubic, smoothing spline graduation is shown. It also has a smoothing parameter chosen by cross-validation. The curve labeled "orig. scale" in Figure 4e is from the adaptive kernel model but fitted *without* first transforming the data. With $s=0$, the graduated rates are smooth, meeting the increasing convex condition in Equation (8), except at the oldest ages. For $s=0.5$, the graduated rates are lower than those for $s=0$, at the youngest ages. This appears to be due to greater weight being attached to the crude rates for ages 37 to 39, where the exposure is greatest. At the oldest ages, the graduated rates for $s=0.5$ lie below those for $s=0$, due to the larger bandwidth under $s=0.5$ at those ages.

The possibility of building in prior knowledge is discussed in Section 2.6. In the absence of a suitable prior table, we have fitted a straight line by least squares to the crude rates, on a logit scale. This requires the additional assumption of normally distributed residuals. In Figure 4f, the residuals from fitting the straight line are smoothed by using Equation (13) with $s=0$ and $s=1$. The smoothed residuals for $s=1$ are then added to the straight line to get the graduated rates labeled "line and $s=1$" in Figures 4d and 4e. Diagnostic plots of the residuals are satisfactory except that the quantile-quantile plot [10] suggests that the residuals are too scattered in the middle of the table. This might indicate the need for further investigation of the normality assumption. Otherwise, we might conclude that fitting a straight line by least squares on the logit scale gives a satisfactory graduation without any kernel adjustment, because the smoothed residuals in Figure 4e are almost zero at all ages. Azzalini and Bowman [2] offer a more formal approach to this problem by using a ratio test to measure the distance between a parametric and a nonparametric model. In this case, we have used a kernel smoother simply as a way of exploring the data before using a parametric model to estimate the mortality rates.

For comparison with the kernel graduations, another nonparametric graduation is also shown in Figures 4d and 4e. This curve is fitted using a well-known statistical method called natural cubic smoothing splines ([26], [56]). It produces results similar to Whittaker graduation [44, chapter 5], but it uses a slightly different smoothness penalty. We refer to this method as the spline graduation, and it is the set of graduated rates that minimizes the function

$$\sum_{i=1}^{n} (\hat{q}_i^t - q_i^t)^2 + b \int_{x_1}^{x_n} ((\hat{q}_x^t)'')^2 \, dx, \tag{20}$$

where $(\hat{q}_x^t)''$ is the second derivative of the graduated, transformed rates and b is again chosen by cross-validation. So as $b \to \infty$, we fit a straight line by least squares and as $b \to 0$, we fit an interpolating, twice differentiable function. A spline graduation has been chosen for comparison, because Silverman [55] calculates an asymptotically equivalent kernel for this smoother, and he also shows that it is an adaptive as opposed to a fixed-width smoother, so the two methods are consistent in this respect. As can be seen from Figure 4d, the spline graduation is very smooth. After Figure 4e has been back-transformed, this graduation is increasingly convex at all ages.

In plot Figure 4e, the curve labeled "orig. scale" is from the adaptive kernel model fitted *without* first transforming the data. The spline graduation fitted without transforming the data produces a similar result (not shown). This illustrates the importance of a good transformation before a nonparametric method is applied.

5.2 Example 4

The data are taken from a report by the Continuous Mortality Investigation Bureau [14], which contains the crude rates for all causes of death for durations 0, 1, and 2 or more of the Female Assured Lives 1975–78 Table. This mortality table arose from the experience of contributing U.K. life offices from whole-of-life and endowment policies on female lives during the years 1975 to 1978. The boundary-correcting kernel discussed in Section 3 along with the adaptive kernel defined in Equation (15) are used to graduate this table. The results for the adaptive kernel defined in Equation (13) were similar.

Figure 3b shows the exposure for this data set. The overall shape is similar to that of Example 3 in Section 5.1, but the range of exposures is much greater. The small exposures for the oldest ages are likely to result in large

variations in the crude rates. The model incorporates these variations using the local bandwidth factors that are shown in Figure 5a, for various values of the sensitivity parameter. The variation in exposures is dampened by reducing the sensitivity parameter to $s=0.1$ in this case. The minimum local bandwidth factor at $s=0.1$ is about 0.5, so the bandwidth at age 94 is about double that at age 30. Thus, the variation in bandwidth is similar to that of the previous example.

After transformation, cross-validation results for a range of global bandwidth values are shown in Figure 5b. For clarity and comparison with Figure 4c, we only show a surface plot of the last 30 ages in Figure 5c. Despite appearances, the maximum weight over this part of the smoother is $S_{64\,64}\approx 0.2$, which is the weight attached to the crude rate at age 64 when the true rate at age 64 is estimated. $S_{64\,64}$ is the point on the surface that is closest to the viewer. This can be calculated from a global bandwidth of 3.3 and a local bandwidth factor of about 0.6 for age 64, giving a weight equal to a normal density with mean 0 and a standard deviation of 3.5×0.6 evaluated at 0, which is approximately 0.2.

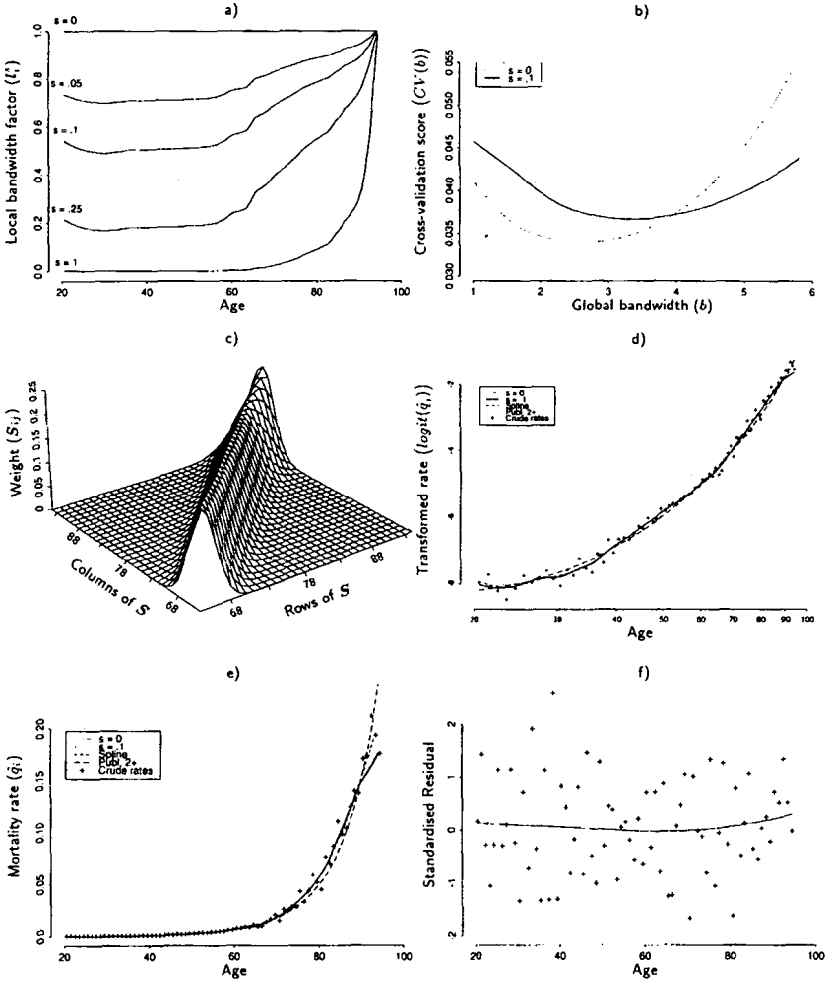
Because we are now using Equation (15), the smoother in Figure 5c has a different shape from that in the previous example. To estimate $\hat{q}_i^t$, the i-th row of the smoother shows the weights attached to the crude rates, where each weight is a function of the bandwidth b_j associated with that crude rate $\dot{q}_j^t$ for $j=1, \ldots, n$. One effect of this is that none of the weights are negative.

The calculations for Figure 5d are carried out after the logit transformation has been used. However, so we can see the graduated rates at the younger ages in greater detail, the results are presented by using a log transformation of the x-axis. Using a logit transformation in Figure 5d, a sensitivity value of $s=0.1$ allows the graduated rates to follow the crude rates more closely at younger ages while smoothing more heavily over the older ages, relative to the fixed-bandwidth graduation. Again, for comparison, a natural, cubic, smoothing spline graduation is fitted with the cross-validation used to choose the smoothing parameter. The spline and the $s=0.1$ graduations are both very similar at the youngest ages, but both kernel graduations are less smooth than the spline graduation at the very oldest ages. The published rates, which were produced using a parametric graduation, are also shown.

An interesting aside is that both the adaptive kernel with $s=0.1$ and the spline graduation indicate a fall in mortality rates with increasing age, for females in their 20s (see Figure 5d and 6e). This suggests that like males, females also suffer from an "accidental hump" but at later ages and to a

FIGURE 5

DATA FROM THE FEMALE ASSURED LIVES 1975–78 TABLE ARE ANALYZED IN A MANNER SIMILAR TO THAT SHOWN IN FIGURE 4, BUT USING THE MODEL DEFINED IN EQUATION (15) WITH $s=0$ AND $s=0.1$. PLOT B) SHOWS THE CROSS-VALIDATION SCORES FOR $s=0$ AND $s=0.1$. PLOT C) SHOWS THE SMOOTHER FOR $s=0.1$ FOR THE OLDEST THIRTY AGES, 64–94. THE FITS ON THE TRANSFORMED SCALE ARE SHOWN IN PLOT D) ALONG WITH A SPLINE GRADUATION AND THE PUBLISHED RATES. PLOT E) SHOWS THE RESULTS FROM PLOT D) AFTER BACK-TRANSFORMING. THE STANDARDIZED RESIDUALS FOR $s = 0.1$ ARE SHOWN IN PLOT F).



much lesser extent [4]. This feature is not present in the published tables, which are fitted by using a parametric method [14].

Much of this detail is lost when the rates are redrawn on the original scale in Figure 5e, which is one of the reasons for transforming the crude rates. On the original scale, the differences between the spline graduation and the crude rates at the oldest ages are magnified. Figure 5f shows that the standardized residuals, defined in Equation (7), for $s=0.1$ are well scattered. The residuals can also be smoothed by using a kernel approach, in which case we expect to see a fairly flat line about zero.

5.3 Example 5

Next we consider using the published table for duration two-or-more to graduate the crude rates for duration 1.

In effect, the duration 2 or more table acts as a prior assumption and thus influences the shape and level of the graduated kernel rates for duration 1. This approach is motivated by the fact that both tables are based on the same population, but the duration two-or-more table has a total of 4,616 deaths out of a total of 2,042,853 policyholders exposed to risk during the period of investigation. The corresponding figures for duration 1 are much less at 334 and 459,068, respectively. So having graduated the larger table, we might want to incorporate that knowledge into the graduation of the smaller table.

The procedure is the same as in Figures 4d–4f: subtract the crude rates from the standard table, smooth the residuals using the smoother defined in Equation (13), and then add the smoothed residuals to the standard table. By approaching the problem in this way, we are emphasizing the relative differences in mortality rates among the durations rather than the absolute mortality rates.

Figure 6a shows the crude rates for duration 1 and the published rates for durations two-or-more. The residuals, shown on a logit scale in Figure 6b, are the differences between these two sets of rates. The residuals are smoothed in a similar manner to that in Example 3. In the middle of the age range, both kernels in Figure 6b are fairly constant. This suggests that mortality rates are consistently lower for the lower-duration table, in that part of the table. At both ends of the table, the fixed-bandwidth kernel is affected by the high residuals. However, the adaptive kernel ignores the high residuals at the oldest ages because the exposure is low, but the relatively high exposures at the youngest ages suggest that the upward trend is real,

at that end of the table. For example, the exposures at the two youngest ages are 18,018 and 19,408 and the exposures at the two oldest ages are 61 and 48. To ensure a monotonically increasing table, the actuary might decide to ignore this feature. In fact, the published rates for durations 0 and 1 are based on an adjustment to the published rates for durations two-or-more. Adding the smoothed kernels to the durations two-or-more published rates gives the two kernel graduations shown in Figure 6a. For $s=0.3$, the smoothness of the graduated rates at the oldest ages is partly due to the large bandwidth, but it is also due to the smoothness of the standard table. The published rates for duration 1 are also shown in Figure 6a.

This example shows how a nonparametric approach to graduation can provide qualitative information about the bias present in subsequent parametric graduations.

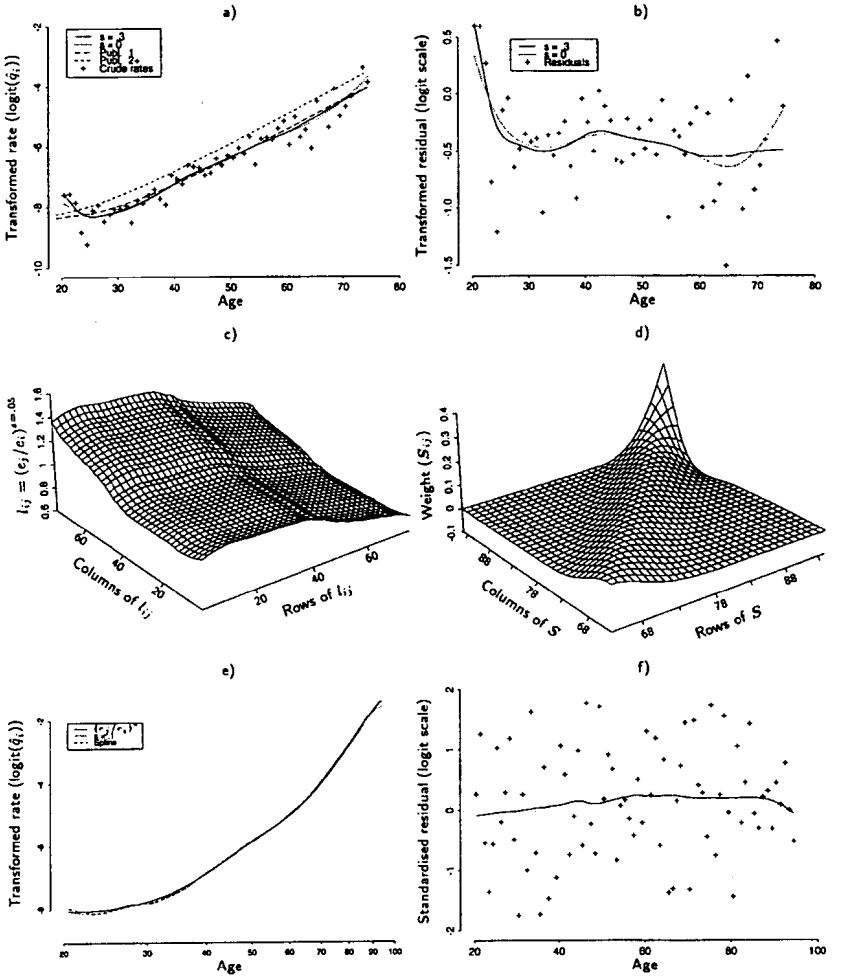
5.4 Example 6

As a final application, we consider using the model defined by Equation (18) to smooth the crude rates of the duration two-or-more table.

For this model, we have a vector of local bandwidth factors for each crude age, which results in a matrix, $\{l_{ij}\}$ where $l_{ij}=e_j/e_i$, for $i, j=1, \ldots, n$. Figure 6c shows this matrix as a surface plot. For $s=0.05$, the local bandwidth factors vary from 0.7 to 1.42. This results in a doubling of the bandwidth across the age range, which like the previous examples gives reasonable results. The diagonal from the nearest to the furthest point in Figure 6c has $l_{ii}=1$, for $i=1, \ldots, n$, and the shape of any row is similar to each of the lines in Figure 5a. The last 30 rows and columns of the resulting smoother are shown in Figure 6d. The shape of this smoother is similar to that in Figure 4c. Notice that the far corner of the smoother is more peaked; this results in graduated rates that rise more sharply at the oldest ages. Figure 6e also shows the results from example 4 in Section 5.2 based on Equation (15), for $s=0.1$. For clarity, the crude and published rates are not shown. At the youngest ages, the graduated rates from Equation (18) are as smooth as the results from Equation (15) with $s=0$, while at the oldest ages the graduated rates are similar to those of the spline graduation. For the duration two-or-more mortality table, the graduated rates from Equation (18) give the smoothest of the three kernel models that we consider, but further work is needed to test these results on other tables. The standardized residuals, in Figure 6f, show no clear pattern.

FIGURE 6

PLOT A) SHOWS THE CRUDE AND PUBLISHED RATES FOR DURATION ONE, THE PUBLISHED RATES FOR DURATION TWO-OR-MORE AND TWO KERNEL GRADUATIONS OF THE DURATION-ONE CRUDE RATES. THE RESIDUALS IN PLOT B) ARE THE DIFFERENCE BETWEEN THE CRUDE RATES FOR DURATION ONE AND THE PUBLISHED RATES FOR DURATION TWO-OR-MORE. THE SMOOTHED RESIDUALS USING EQUATION (13), WITH $s=0$ AND $s=0.3$, ARE ALSO SHOWN. PLOT C) SHOWS THE SURFACE GENERATED FROM EQUATION (17) FOR $s=0.05$ AND THE LAST 30 ROWS AND COLUMNS OF THE CORRESPONDING SMOOTHER FROM EQUATION (18) ARE SHOWN IN PLOT D). PLOT E) SHOWS THE RESULTS FROM EXAMPLE 4 IN SECTION 5.2 BASED ON EQUATION (15), FOR $s=0.1$. THE STANDARDIZED RESIDUALS ARE SMOOTHED IN PLOT F).



6. DISCUSSION

We start with a bivariate scatterplot of age against mortality, but as the data are grouped by age, we simply average within each group to produce a set of equally spaced observations. This eliminates the first-order bias of the Nadaraya-Watson estimator in the interior of the table. Transforming the data helps to stabilize the variance and to reduce the curvature. This means that the second-order bias, due to curvature, is reduced. The second-order bias may be further reduced by using a suitable, standard mortality table to filter out some of the curvature of the true mortality rates. Also, a boundary correction helps to reduce the extra bias encountered at the ends of the table. Thus, pooling and transforming the data, using prior knowledge of the shape of the curve and an adjustment at the extreme ages all help to validate the assumptions of residuals with zero mean and constant variance in the model. The requirement of independent residuals is more difficult to achieve, but Equation (18) combined with possible adjustments [15] may help to alleviate this problem. Finally, some diagnostic plots, discussed in Section 2.5, offer an easy means of assessing the validity of the assumptions made.

A kernel function that makes explicit allowance for the boundary is defined and illustrated in Section 3. Complications such as the extrapolation method ([42], [52]) applied to the Nadaraya-Watson estimator or the large boundary adjustment [36] applied to the Gasser-Müller estimator subtract from the intuitive appeal of kernel models. However, the extra effort required to specify the kernel does improve the results in the examples shown and for other mortality tables not reported here. This is because the variable bandwidth in our estimators usually increases for older ages because of lower exposures at those ages.

The adaptive kernel model in Section 4 allows the estimated rates of mortality to include explicitly the extra information provided by the changing amounts of exposure, in addition to the information from the crude rates themselves. A sensitivity parameter allows the user to control the degree of emphasis placed on the changing exposures through the local bandwidth factors. The global bandwidth parameter is used to control the absolute level of the bandwidth curve. If desired, its value can be chosen objectively using cross-validation or some equivalent method.

Although our applications have been restricted to the more traditional application of constructing a single-decrement life table, other applications are conceivable, such as transition intensities or probabilities in multiple-decrement and multiple-state models.

Throughout this paper, we have adopted a heuristic approach to kernel graduation, but a more theoretical perspective may offer deeper insight into the connection between this method and MWA graduation. Because of the adaptive bandwidth, both the number of crude rates and the corresponding weights vary in the estimation of the transformed, true rate at each age. In contrast, the range of a MWA is often kept fixed and only the weights are allowed to vary. In this respect, kernel models may offer a more flexible approach to local smoothing. In addition, this paper has concentrated on the Nadaraya-Watson kernel estimator only, but there are many others [43]. Further work is needed to assess their relative merits for graduation.

An alternative model that is closely related to kernel-smoothing is to fit low-order polynomials locally. So instead of fitting a constant, we now fit a straight line or a quadratic using least squares. This approach was popularized by Cleveland [12]. Hastie and Loader [37] review the recent statistical literature on this subject and argue that higher-order models result in a lower order of bias without a corresponding increase in variance. Fan and Marron [22] have pointed out that fast implementations of kernel and local polynomial methods have recently emerged, and they claim speeds comparable to those of smoothing splines. For mortality applications, this model has the advantages of automatically adjusting at the boundaries to reduce the bias. It also provides more reliable estimates of the derivatives of the mortality curve than the Nadaraya-Watson estimator. In the actuarial literature, Renshaw [51] considers generalized linear and nonlinear graduation.

Another area of future interest is robustness. Some of the examples in Section 5 appear to be influenced by outliers, because the Nadaraya-Watson estimator offers no explicit resistance to unusual observations. An influential point may also affect the choice of bandwidth when an automatic selection method is used, such as cross-validation. Cleveland [12] extends his model to include robust iterative estimation. In fact, any smoother can be made robust by using more resistant local averaging, such as the mode or median [54].

The potential uses of a nonparametric approach, listed in the introduction, suggest that they have much to offer as part of the actuarial toolkit. Note that we are not advocating that a nonparametric model should always be used instead of a parametric one. A nonparametric model should be viewed as an exploratory step towards the final model choice, which may be parametric because of its inherent smoothness. Differences between the best parametric and nonparametric graduations will highlight the extent of the actuary's desire for smoothness, at a cost of lack of fit to the data.

ACKNOWLEDGMENT

We thank the panel of referees for their many helpful comments, which led to a much improved paper.

We are grateful to Chris Jones, William Schucany, and David Scott for making available preprints of their work. We would also like to thank Glenn Stone for the use of his dynamic spline fitting program. It was used to produce some of the results in Section 5.

REFERENCES

1. ALTMAN, N.S. "Kernel Smoothing of Data with Correlated Errors," *Journal of the American Statistical Association* 85 (1990): 749–59.
2. AZZALINI, A., AND BOWMAN, A. "On the Use of Nonparametric Regression for Checking Linear Relationships," *Journal of the Royal Statistical Society B* 55, (1993): 549–57.
3. BENJAMIN, B., AND POLLARD, J.H. *The Analysis of Mortality and Other Actuarial Statistics*. London: Heinemann, 1980.
4. BLOOMFIELD, D.S.F., AND HABERMAN, S. "Graduation: Some Experiments with Kernel Methods," *Journal of the Institute of Actuaries* 114 (1987): 339–69.
5. BROCKETT, P.L. "Information Theoretic Approach to Actuarial Science: A Unification and Extension of Relevant Theory and Applications," *TSA* 43 (1991): 73–135.
6. BROFFITT, J.D. "Increasing and Increasing Convex Bayesian Graduation," *TSA* 40 (1988): 115–48.
7. BROOKS, R.J., STONE, M., CHAN, F.Y., AND CHAN, L.Y. "Cross Validatory Graduation," *Insurance: Mathematics and Economics* 7 (1988): 59–66.
8. CARLIN, B.P. "A Simple Monte Carlo Based Approach to Bayesian Graduation," *TSA* 44 (1992): 1–22.
9. CARROLL, R.J., AND RUPPERT, D. *Transformation and Weighting in Regression*. New York: Chapman and Hall, 1988.
10. CHAMBERS, J.M., CLEVELAND, W.S., KLEINER, B., AND TUKEY, P.A. *Graphical Methods for Data Analysis*. Belmont, Calif: Wadsworth, 1983.
11. CHU, C.K., AND MARRON, J.S. "Choosing a Kernel Regression Estimator," *Statistical Science* 6, no. 4 (1991): 404–36.
12. CLEVELAND, W.S. "Robust Locally-Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association* 74 (1979): 829–36.
13. CLEVELAND, W.S., DEVLIN, S.J., AND GROOSE, E. "Regression by Local Fitting," *Journal of Econometrics* 37 (1988): 87–114.
14. CONTINUOUS MORTALITY INVESTIGATION BUREAU. "Graduation of the Mortality Experience of Female Assured Lives: 1975–78," *Report Number 6*. London: Institute and Faculty of Actuaries, 1983.

15. CONTINUOUS MORTALITY INVESTIGATION BUREAU. "An Investigation into the Distribution of Policies Per Life Assured in the Cause of Death Investigation Data," *Report Number 8*. London: Institute and Faculty of Actuaries, 1986.
16. COPAS, J.B., AND HABERMAN, S. "Non-parametric Graduation Using Kernel Methods," *Journal of the Institute of Actuaries* 110 (1983): 135–56.
17. COX, D.R., AND SNELL, E.J. *Analysis of Binary Data*. 2nd ed. New York: Chapman and Hall, 1989.
18. ELANDT-JOHNSON, R.C., AND JOHNSON, N.L. *Survival Models and Data Analysis*. New York: John Wiley & Sons, 1980.
19. EPANECHNIKOV, V.A. "Non-parametric Estimation of a Multivariate Probability Density," *Theory of Probability and Its Applications* 14 (1969): 153–58.
20. SCHUCANY, W.R. "Adaptive Bandwidth Choice for Kernel Regression," *Journal of the American Statistical Association* 90, no. 430 (1995): 535.
21. EUBANK, R.L., AND THOMAS, W. "Detecting Heteroscedasticity in Nonparametric Regression," *Journal of the Royal Statistical Society B* 55 (1993): 145–55.
22. FAN, J., AND MARRON, J.S. "Comment on Hastie & Loader's Paper—Local Regression: Automatic Kernel Carpentry," *Statistical Science* 8, no. 2 (1993): 129–34.
23. GASSER, T., AND MÜLLER, H.G. "Estimating Regression Functions and Their Derivatives by the Kernel Method," *Scandinavian Journal of Statistics* 11 (1984): 171–85.
24. GAVIN, J.B., HABERMAN, S., AND VERRALL, R.J. "Moving Weighted Average Graduation Using Kernel Estimation," *Insurance: Mathematics and Economics* 12 (1993): 113–26.
25. GAVIN, J.B., HABERMAN, S., AND VERRALL, R.J. "On the Choice of Bandwidth for Kernel Graduation," *Journal of the Institute of Actuaries* 121 (1994): 119–34.
26. GREEN, P.J., AND SILVERMAN, B.W. *Nonparametric Regression and Generalised Linear Models*. London: Chapman and Hall, 1994.
27. GREGOIRE, G. "Least Square Cross Validation for Counting Processes Intensities," *Scandinavian Journal of Statistics* 20, no. 4 (1993): 343–60.
28. GREVILLE, T.N.E. "Moving-Weighted-Average Smoothing Extended to the Extremities of the Data. I. Theory," *Scandinavian Actuarial Journal* (1981): 38–55.
29. GREVILLE, T.N.E. "Moving-Weighted-Average Smoothing Extended to the Extremities of the Data. II. Methods," *Scandinavian Actuarial Journal* (1981): 65–81.
30. GREVILLE, T.N.E. "Moving-Weighted-Average Smoothing Extended to the Extremities of the Data. III. Stability and Optimal Properties," *Journal of Approximation Theory* 33 (1981): 43–58.
31. HALL, P. "On the Bias of Variable Bandwidth Curve Estimators," *Biometrika* 77, no. 3 (1990): 529–35.
32. HALL, P., AND JOHNSTONE, I.M. "Empirical Functionals and Efficient Smoothing Parameter Selection," *Journal of the Royal Statistical Society B* 54 (1992): 475–530.

33. HALL, P., AND WEHRLY, T.E. "A Geometrical Method for Removing Edge Effects from Kernel-Type Nonparametric Regression Estimates," *Journal of the American Statistical Association* 86 (1991): 665–72.
34. HÄRDLE, W., HALL, P., AND MARRON, J.S. "How Far Are Automatically Chosen Regression Smoothing Parameters from Their Optimum?" *Journal of the American Statistical Association* 83 (1988): 86–101.
35. HART, J.D., AND WEHRLY, T.E. "Kernel Regression Estimation Using Repeated Measurement Data," *Journal of the American Statistical Association* 81 (1986): 1080–88.
36. HART, J.D., AND WEHRLY, T.E. "Kernel Regression When the Boundary Region is Large, with an Application to Testing the Adequacy of Polynomial Models," *Journal of the American Statistical Association* 87 (1992): 1018–24.
37. HASTIE, T., AND LOADER, C. "Local Regression: Automatic Kernel Carpentry," *Statistical Science* 8, no. 2 (1993): 120–43.
38. HASTIE, T.J., AND TIBSHIRANI, R.J. *Generalized Additive Models.* London: Chapman and Hall, 1990.
39. HOEM, J.M., AND LINNEMANN, P. "The Tails in Moving Average Graduation," *Scandinavian Actuarial Journal* (1988): 193–229.
40. JOINT MORTALITY INVESTIGATION COMMITTEE. "Continuous Investigation into the Mortality of Assured Lives: Memorandum on a Special Inquiry into the Distribution of Duplicate Policies," *Journal of the Institute of Actuaries* 83 (1957): 34–36.
41. JONES, M.C. "Do Not Weight for Heteroscedasticity in Nonparametric Regression," *Australian Journal of Statistics* 35, no. 1 (1993): 89–92.
42. JONES, M.C. "Simple Boundary Correction for Kernel Density Estimation," *Statistics and Computing* (1993): 135–46.
43. JONES, M.C., DAVIES, S.J., AND PARK, B.U. "Versions of Kernel-Type Regression Estimators," *Journal of the American Statistical Association* 89 (1994): 825–32.
44. LONDON, D. *Graduation: The Revision of Estimates.* Winsted and Abington, Conn.: ACTEX Publications, 1985.
45. LONDON, R.L. "In Defence of Minimum-r_0 Linear Compound Graduation and a Simple Modification for Its Improvement," *ARCH* (1981.2): 75–78.
46. LOWRIE, W.B. "An Extension of the Whittaker-Henderson Method of Graduation," *TSA* 34 (1982): 329–72.
47. NADARAYA, E.A. "On Estimating Regression," *Theory of Probability and Its Applications* 9 (1964): 141–42.
48. NIELSEN, J.P. "A Transformation Approach to Bias Correction in Kernel Hazard Estimation," *Research Report Number 115.* University of Copenhagen: Laboratory of Actuarial Mathematics, 1992.
49. RAMLAU-HANSEN, H. "The Choice of a Kernel Function in the Graduation of Counting Process Intensities," *Scandinavian Actuarial Journal* (1983): 165–82.
50. RAMSAY, C.M. "Minimum Variance Moving-Weighted-Average Graduation," *TSA* 43 (1991): 305–33.

51. RENSHAW, A.E. "Actuarial Graduation Practice and Generalised Linear and Non-linear Models," *Journal of the Institute of Actuaries* 118 (1991): 295.
52. RICE, J.A. "Boundary Modification for Kernel Regression," *Communications in Statistics. Theory and Methods* 13 (1984): 893–900.
53. SCOTT, D.W. "Constrained Oversmoothing and Upper Bounds on Smoothing Parameters in Regression and Density Estimation," *Technical Report 92-8*. Rice University: Department of Statistics, 1992.
54. SCOTT, D.W. *Multivariate Density Estimation; Theory, Practice and Visualisation*. New York: John Wiley & Sons, 1992.
55. SILVERMAN, B.W. "Spline Smoothing: The Equivalent Variable Kernel Method," *Annals of Statistics* 12 (1984): 898–916.
56. SILVERMAN, B.W. "Some Aspects of Spline Smoothing Approaches to Non-parametric Regression Curve Fitting," *Journal of the Royal Statistical Society B* 47 (1985): 1–52.
57. STONE, M. "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society B* 36 (1974): 111–47.
58. VERRALL, R.J. "Whittaker Graduation Viewed as Dynamic Generalised Linear Models," *Insurance: Mathematics and Economics* 13 (1994): 7–14.
59. WATSON, G.S. "Smooth Regression Analysis," *Sankhya A* 26 (1964): 359–72.