Assurances et gestion des risques
Insurance and Risk Management

Essays on Boundaries Effects and Practical Considerations for Univariate Graduation of Mortality by Local Likelihood Models

Julien Tomas

Volume 80, Number 2, 2012

URI: https://id.erudit.org/iderudit/1091845ar
DOI: https://doi.org/10.7202/1091845ar

See table of contents

Publisher(s)

Faculté des sciences de l'administration, Université Laval

ISSN

1705-7299 (print)
2371-4913 (digital)

Explore this journal

Cite this article

Tomas, J. (2012). Essays on Boundaries Effects and Practical Considerations for Univariate Graduation of Mortality by Local Likelihood Models. *Assurances et gestion des risques / Insurance and Risk Management*, *80*(2), 203–261. https://doi.org/10.7202/1091845ar

Article abstract

Local regression is a popular form of non-parametric regression, combining excellent theoretical properties with conceptual simplicity and flexibility to find structure in many datasets. Local regression smoothers fit low-order polynomials locally in the points surrounding a target point. The estimate, at each target point, is a weighted mean taken from the polynomial with observations close to the target point receiving the largest weights. Unfortunately this simplicity has flaws. At the boundary, the weight function is asymmetric and the estimate may have substantial bias. Bias can be a problem if the regression function has relatively high curvature in the boundary. It leads to a disturbing nuisance affecting applications as well as global measures of performance of the estimators like mean squared error or deviations between the true curve and estimated curve.

In this article, we consider the alleviation of this boundary problem for the context of univariate graduation of mortality by local likelihood models. We consider three specifc treatments to reduce the impact of these boundary effects including symmetric and asymmetric weight systems. We analyze local statistical properties of smoothers subject to an a priori fixed bandwidth restriction. The weighting systems of these estimators depend on smoothing parameters that traditionally are estimated by means of data dependent optimization criteria. However by imposing to all of them the condition of a fixed bandwidth, we can measure the performance of each smoother and study where the contribution to these criteria are coming from the design space. Apart from statistical considerations, the choice of the parameters could be refined by taking into account the nature of the risk considered. The results are compared to the Whittaker-Henderson model for which it is not necessary to give specific treatment at the boundary.

Assurances et gestion des risques, vol. 80(2), octobre 2012, 203-261
Insurance and Risk Management, vol. 80(2), October 2012, 203-261

Essays on Boundaries Effects and Practical Considerations for Univariate Graduation of Mortality by Local Likelihood Models

by Julien Tomas

ABSTRACT

Local regression is a popular form of non-parametric regression, combining excellent theoretical properties with conceptual simplicity and flexibility to find structure in many datasets. Local regression smoothers fit low-order polynomials locally in the points surrounding a target point. The estimate, at each target point, is a weighted mean taken from the polynomial with observations close to the target point receiving the largest weights. Unfortunately this simplicity has flaws. At the boundary, the weight function is asymmetric and the estimate may have substantial bias. Bias can be a problem if the regression function has relatively high curvature in the boundary. It leads to a disturbing nuisance affecting applications as well as global measures of performance of the estimators like mean squared error or deviations between the true curve and estimated curve.

In this article, we consider the alleviation of this boundary problem for the context of univariate graduation of mortality by local likelihood models. We consider three specific treatments to reduce the impact of these boundary effects including symmetric and asymmetric weight systems. We analyze local statistical properties of smoothers subject to an a priori fixed bandwidth restriction. The weighting systems of these estimators depend on smoothing parameters that traditionally are estimated by means of data dependent optimization criteria. However by imposing

Acknowledgment

We wish to thank Professor F. Planchet for helpful and constructive suggestions which he provided in relation to earlier draft of this article. We are also grateful to a referee for a careful reading of the manuscript and comments which lead to an improved version of the paper.

The author:

RESAM, Department of Quantitative Economics, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands. Contact: j.tomas@uva.nl

to all of them the condition of a fixed bandwidth, we can measure the performance of each smoother and study where the contribution to these criteria are coming from the design space. Apart from statistical considerations, the choice of the parameters could be refined by taking into account the nature of the risk considered. The results are compared to the Whittaker-Henderson model for which it is not necessary to give specific treatment at the boundary.

Keywords: Boundary effect, Local likelihood, Life insurance, Graduation.

JEL - Code: C14, G22.

RÉSUMÉ

La régression locale est une forme populaire de régression non-paramétrique, combinant d'excellentes propriétés théoriques avec une simplicité conceptuelle et une flexibilité capable de trouver la structure dans de nombreux ensemble de données. Les lisseurs issus des régressions locales ajustent localement des polynômes aux points aux alentours d'un point cible. L'estimation, à chaque point cible, est une moyenne pondérée du polynôme où les observations proches du point cible reçoivent les poids les plus élevés. Malheureusement, cette simplicité a des failles. Aux bordures, la fonction de poids est asymétrique et l'estimation peut engendrer un biais important. Ce biais peut être un problème si la fonction de régression a une courbure relativement élevée à la bordure. Cela entraîne une nuisance qui affecte les applications ainsi que les mesures globales de performance des estimateurs comme la moyenne des carrés des résidus ou les écarts entre la courbe réelle et la courbe estimée.

Dans cet article, nous considérons la réduction des effets de bordures pour le lissage des données d'expérience provenant de l'assurance vie. Nous considérons trois traitements spécifiques pour réduire l'impact des effets de bordures, incluant des systèmes de poids symétriques et asymétriques. Nous analysons les propriétés statistiques locales des lisseurs sujets à une restriction a priori de la fenêtre d'observations. Les systèmes de pondération de ces estimateurs dépendent des paramètres de lissage qui sont traditionnellement estimés au moyen de critères d'optimisation dépendants des données. Cependant, en imposant à chacun d'eux la condition d'une fenêtre d'observations fixe, on peut mesurer la performance de chaque lisseur et étudier la provenance des contributions à ces critères. Mis à part les considérations statistiques, le choix des paramètres pourrait être affiné en tenant compte de la nature du risque. Les résultats sont comparés au modèle de Whittaker-Henderson pour lequel il n'est pas nécessaire de donner un traitement spécifique aux bordures.

Mots clés : Effet de bordures, modèle de vraissamblances, assurance-vie, lignage des données.

Code JEL : C14, G22.

1. INTRODUCTION

Life tables are used to describe the one-year probability of death within a well defined population as a function of attained age. These probabilities play an important role in the determination of premium

rates and reserves in life insurance. The crude estimates on which life tables are based might be considered as a sample from larger population and are, as a result, subject to random fluctuation. However, the actuary wishes most of the time to smooth these quantities to enlighten the characteristics of the mortality of the group considered which he thinks to be relatively regular.

Assume that we are given the number of deaths recorded, d_i, and the number of individuals initially exposed to the risk of death, l_i, all aged x_i last birthday, and that our experience is limited to this single age x_i where $i = 1, 2, ..., n$. The crude estimate of the observed mortality rate, q_i, is denoted by $\overset{\circ}{q}_i$,

$$\overset{\circ}{q}_i = \frac{d_i}{l_i}. \tag{1}$$

Then $\overset{\circ}{q}_i$ represents the one-year observed probability of death for a particular population at age x_i which lies above or below the true underlying value.

A common prior opinion about the form of the true rates is that each true rate of mortality is closely related to its neighbors, that is the observations $\overset{\circ}{q}_j$ near $\overset{\circ}{q}_i$ should contain information about the value of the unknown response function ψ at x_i. This relationship is expressed, recall Gavin *et al* (1993), by the belief that the true rates progress smoothly from one age to the next. It follows that the data for several ages x_j on either side of age x_i can be used to augment the basic information we have at age x_i, and an improved estimate of q_i can be obtained by smoothing the individual estimates. In two recent studies, Tomas (2012) and Tomas (2011) show the applicability of local regressions to model the relation between the crude death rates and attained age.

Local likelihood is introduced as a method of smoothing by local polynomial in non-Gaussian regression models. A local Binomial likelihood model is proposed when the number of initial policyholders exposed to risk is available. Let suppose that L_j persons come under observation at age x_j and continue under observation until they survive to $x_j + 1$ or die before. In this case we denote initial exposed to risk as L. Moreover, let suppose that the probability of death during the year for each one of them is q_j, and that the death or survival of one is independent of the death or survival of the others. If we call D_j the random variable which represents the number of deaths that occur in the year, we will use the usual model for the number of deaths,

$D_j \sim \text{Binomial}(l_j, q_j)$,

and the observed death rate, which is the maximum likelihood estimate of q_j, is denoted as $\overset{\circ}{q}_j = d_j / l_j$.

Let now suppose that L_j persons enter observation under hypothesis that the force of mortality (instantaneous mortality rate) is a constant during the period of observation and that the death or survival of each one is independent. In this case L_j represents those central exposed to risk, whereas in the previous section L_j denoted initial exposures. Hence the force of mortality, φ_j, is the average risk to which population is subjected during its passage through the year of age $x_j + 1$ and is a different concept from q_j, which represents the total effect of mortality in terms on proportion who fail to survive the whole year of age $x_j + 1$ without reference to the variation of mortality risk over the course of that year. According to Brillinger (1986, p. 697), the number of deaths which occur in the period of observation, D_j, will have a Poisson distribution with average and variance equal to μ_j. We would consider the graduation of μ_j / L_j, with

$$D_j \sim \text{Poisson}(\mu_j).$$

The local Binomial and Poisson models, presented above, apply the local fitting technic to data of which the relationship can be expressed through a likelihood function. Suppose that we have n independent realizations $y_1, y_2, ..., y_n$ of the random variable Y with

$$Y_i \sim f(Y, \theta), \text{ for } i = 1, 2, ..., n$$

where θ_i is an unspecified smooth function $\psi(x_i)$. To estimate $\psi(x_i)$, suppose that the function ψ has a $(p + 1)$th continuous derivative at the point x_i. For data point x_j in a neighborhood of x_i we approximate $\psi(x_j)$ via a Taylor expansion by a polynomial of degree p as:

$$\begin{aligned}\psi(x_j) &\approx \sum_{p=0}^{P} (f^{(p)}(x_i) / p!)(x_j - x_i)^p \\ &= \sum_{p=0}^{P} \beta_p(x_i)(x_j - x_i)^p.\end{aligned}$$

The local log-likelihood, or *local kernel-weighted log-likelihood* as named by Fan *et al.* (1998) is given by

$$L(\beta|\lambda, x_i) = \sum_{j=1}^{n} l\left(y_j, \boldsymbol{x}^T\beta\right) w_j, \tag{2}$$

where $\boldsymbol{x} = (1, x_j - x_i, ..., (x_j - x_i)^p)^T$, $\beta = (\beta_0, ..., \beta_p)^T$, with $\beta_p = \psi^{(p)}(x_i) / p!$, $p = 0, 1, ..., P$ and w_j denotes a non-negative weight function depending

on the target value x_i and the measurement pointsj, and in addition, it contains a smoothing parameter $h = (\lambda - 1) / 2$ which determines the sizes of the neighborhood of x_i. Maximizing the local log-likelihood (2) with respect to β gives the vector of estimators $\widehat{\beta} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)^T$. Estimators $\psi^{(p)}(x_i)$, $p = 0, 1, ..., P$, are given by $\widehat{\psi}^{(p)}(x_i) = p!\widehat{\beta}_p$.

The weighted average produces a smooth estimate of $\psi(x_i)$ resulting from the weighted linear combination of $2h + 1$ observations surrounding $\psi(x_i)$, including itself. The smooth estimates for the whole series are obtained by applying the expression (2) in a moving manner.

Unfortunately this simplicity has flaws. At the boundary, the smoothing weights function is asymmetric and the estimate may have substantial bias. Bias can be a problem if the regression function has relatively high curvature in the boundary. It may force the criteria to select a smaller bandwidth at the boundary to reduce the bias, but this may lead to under-smoothing in the middle of the table.

In this article, we consider the alleviation of this boundary problem in the context of graduation by local likelihood models of experience data originating from life insurance.

We analyze local statistical properties of the non-parametric smoothers, used in the two models presented above, which are subjected to an a priori fixed bandwidth restriction.

We study three specific treatments to reduce the impact of these boundary effects including symmetric and asymmetric weight systems. The weighting systems of these estimators depend on smoothing parameters that traditionally are estimated by means of data dependent optimization criteria.

Here and throughout we impose the arbitrary condition of a fixed bandwidth, $\lambda = 19$, which is the optimal theoretical bandwidth founded in our previous study, see Tomas (2011). In consequence, we can measure the performance of each smoother and study where the contributions to these criteria are coming from the design space. Apart from statistical considerations, the choice of the parameters can be refined by taking into account the nature of the risk considered. The results are compared to the Whittaker-Henderson model which does not need any treatment in the boundary.

This article begins by presenting in Section 2 a review of the development of smoothing approaches. We introduce briefly the non-parametric smoothers in Section 3, and their weighting system in Section 4. Section 5 discusses the smoothing properties based on local measures such as fitted degrees of freedom and influence values. All

these measures are computed for symmetric and asymmetric (point in the boundary) smoothers with the boundaries corrections considered. Section 6 presents briefly the criteria used for models selection and studies where the contributions to these criteria are coming from the design space. Finally, Section 7 summarizes the conclusions drawn in the paper.

2. FROM AN HISTORICAL PERSPECTIVE: HISTORICAL REVIEW OF THE DEVELOPMENT OF SMOOTHING APPROACHES

The problem of smoothing sequences of observations is relevant in many branches of sciences. This section reviews the development of smoothing methods starting in the late eighteenth to the early twenty first centuries, leading up to the development of the use of local polynomial regression and afterward local likelihood methods.

2.1 Early work

Local regression is a natural extension of parametric fitting, so natural that local regression arose independently at different points in time and in different countries. The setting for this early work was univariate and equally spaced x_i. It was simple enough that good-performing smoothers could be developed and were computationally feasible by hand calculation. Also, most of the early work arose in actuarial studies remark Cleveland and Loader (1996). Mortality and sickness rates were smoothed as a function of age.

Haberman (1996, p. 40) reports that smoothing was used as early as 1765 by the Swiss mathematician and physicist Johann Lambert Daw (1980, p. 357) explains in his 1765's work (volume 1), he graduated the value l_i, at decennial ages, which he had calculated from the deaths recorded in the London Bills of Mortality for 1753-1758. He does not read off the graduated values of l_i at all ages from his graph, but gives two methods of graduation and/or interpolation. The first was a graphical method for introduction *osculating parabolas* between two points. The second was a method of fitting a polynomial of fifth degree to represent a section of the curve which was then able to *hang together* with the corresponding polynomials for the immediately preceding and succeeding sections of the curve. This methodology is effectively what come to be known as *osculatory interpolation*, and was re-invented more than 100 years later by Thomas Sprague.

John Finlaison, subsequently first president of the Institute of Actuaries in January 1823, started preparing the mortality data that were to provide the first life table consisting of graduated observations at individual ages. His 1829's work is described by Seal (1982 p. 89), where his formula is based on overlapping piecewise linear arcs extending over nine successive values, with eight of the nine being used in the next arc, and thus represents the first published example of a graduation by the adjusted-average method.

This piecewise approach to smoothing was extended in 1866 by the Italian meteorologist and astronomer Giovanni Schiaparelli who assumed a cubic polynomial to extend to a stretch of consecutive observed values.

In the same year (1866) that Schiaparelli wrote, Wesley Woolhouse presented a detailed exposition of graduation of mortality rates using summation formulae, stressing the conceptual differences between graduation and interpolation. He considered the case where the fourth differences of the corrections $v_i = \overset{\circ}{q}_i - q_i$ to an observed series of rates had small values and proposed to minimize $\sum v_i^2$ in terms of $\Delta^4 v_i$ and thus obtain estimates of v_i and hence q_i. (Seal, 1982, p. 93) demonstrates that the equations for $\hat{q}_i$ are equivalent to those which arise from fitting a piecewise cubic polynomials by least squares to equidistant observations.

The use of symmetrical moving weighted average formulae to smooth equally spaced observations of a function of one variable which generalized Woolhouse's summation formulae, was systematically investigated in a series of papers by the American statistician Erastus De Forest reports Haberman (1996, p. 41). De Forest's principal innovation was to introduce optimality criteria into the problem of estimating the coefficients.

In 1887, occurred Thomas Sprague's paper on the graphic method of graduation. Sprague's paper of 1887 rediscovered (following Lambert) osculatory interpolation showing how formulae could be devised to ensure continuity of the first derivatives of overlapping interpolation curves. Osculatory interpolation was used as a method of graduation for the English life table in the early nineteenth century.

A new style of summation graduation and its testing had started with Spencer, in 1904 and 1907, and had blossomed in Vaughan's 1933, 1934 and 1935 articles. The method developed by Spencer in his 1904's article had became popular because it was computationally efficient and had good performance. We note three crucial properties. First, the smoother exactly reproduces cubic polynomials as explained in Cleveland and Loader (1996). Second, the smoothing coefficients are

a smooth function of length of the bandwidth and decay smoothly to zero at the ends. Third, the smoothing can be carried out by applying a sequence of smoothers each of which is simple; this was done to facilitate hand computation. Achieving all three of these properties is remarkable.

Whittaker (1923) suggested an alternative method of graduation. This can be regarded as what would now be called a Bayesian approach to graduation, see Taylor (1992). It results in the minimization of the combination of a measure of goodness of fit of the graduation to the observation and a measure of smoothness.

2.2 Modern work

We have seen that the methods presented in the introduction are inherited from a long actuarial tradition. However local regression methods received little attention in the statistical literature until the late 1970's.

For Cleveland and Loader (1996), the modern view of smoothing by local regression has origins in the 1950's and 1960's, with kernel methods introduced in the density estimation setting (Rosenblatt, 1956; Parzen, 1962) and the regression setting (Watson, 1964). Kernel methods are a special case of local regression; it amounts to choosing the parametric family to consist of constant functions. Kernel methods have been brought up to actuarial application by Copas and Haberman (1983) and followed by Gavin *et al*. (1993) and Gavin *et al*. (1995).

However, recognizing the weaknesses of a local constant approximation, the more general local regression enjoyed a reincarnation beginning in the late 1970's. It includes the mathematical development of Stone (1977), Stone (1980), and the *lowess* procedure of Cleveland (1979). It provides a number of important insights about the choices of the smoothing parameters. For example it was nearly a given that for most applications the weight function needed to be smooth, that local constant fitting was inadequate, and that smoothers needed to reproduce exactly (and not just asymptotically) at least a quadratic.

Among other features, the local regression method and linear estimation theory trivialize problems that have proven to be major stumbling blocks for more widely studied kernel methods. The kernel estimation literature contains extensive work on bias correction methods: finding modifications that *asymptotically* remove dependence of the bias on the slope, curvature, and so on. Examples include boundary kernels Müller (1987), higher order kernels Gaser *et al*. (1985) and Schucany (1989). Local regression method can then be viewed as an

extension of kernel methods and attempt to extend the theory of kernel methods. This treatment has become popular in the 1990s, for example Hastie and Loader (1993) and to some extend in Loader (1999). The approach has it uses: Small bandwidth asymptotic properties of local regression, such as rates of convergence and optimality theory, rely heavily on results for kernel methods. But for practical purposes, the kernel theory is of limited use, since it often provides poor approximations and requires restrictive conditions.

Furthermore, while the early smoothing work was based on an assumption of a near-Gaussian distribution, the modern view extended smoothing to other distributions. Cleveland (1979) developed robust smoothers. Later, Tibshirani and Hastie (1987) took local fitting one steps further; in any situation where a dependent variable depends on independents variables, a local likelihood procedure can be carried out. Hence they substantially extended the domain of smoothing to many distributional settings such as logistic regression and developed general fitting algorithms. The extension to new settings has continued in the 1990's with Loader (1996) and Fan *et al*. (1998).

We have seen a substantial number of smoothing methods that have been suggested in the literature through the history. Most of these methods involve a trade off between goodness of fit (closeness of the $\psi(x_i)$ to the $\overset{\circ}{q}_i$) and the smoothness of the smoothed values $\psi(x_i)$. The motivation for local regression is that it is easy to understand and to interpret; because of its simplicity it can be tailored to work for many different distributional assumptions; it adapts well to bias problem at boundaries and in regions of high curvature; it does not require smoothness and regularity conditions required by other methods such as boundary kernels; and so on, see Hastie and Loader (1993) for a detailed presentation of the strengths of local regression. Singly, none of these provides a strong reason to favor local regression over other smoothing methods such as smoothing splines, regression splines with knot selection, wavelets, and various modified kernel methods. Rather, it is the combination of these issues that combine to make local regression attractive.

3. THE NON-PARAMETRIC ESTIMATORS

The non-parametric estimators to be discussed are based on different assumptions of smoother building. The local Binomial and Poisson kernel-weighted log-likelihood assume different expectation-variance structures, while the Whittaker-Henderson model is derived from the graduation theory and does not take the nature of the data into

account. The parameters are estimated by means of data dependent optimization criteria which search for an optimal solution between both fitting and smoothing of the data.

Next, we discuss briefly each non-parametric smoother and refer the reader to Tomas (2012) and Tomas (2011) for more details.

3.1 Local likelihood models

A special case of model (2) occurs when the conditional density of Y given X belongs to the exponential dispersion family with a probability mass function which can be written in the form:

$$f_Y(y_j;\theta_j,\phi) = \exp\left\{\frac{y_j\theta_j - b(\theta_j,m_j)}{a_j(\phi)} + c(y_j,\phi)\right\},$$

for specific functions a(.), b(.) and c(.). ϕ is called the dispersion parameter. It is a nuisance parameter not depending on x_j. The function a and c are such that $a_j(\phi) = \phi / m_j$ and $c = c(y_j, \phi / m_j)$, where m_j is a known weight for each observation x_j. Two examples are presented in Table 1.

In the context of graduation of mortality, a local binomial likelihood model is used when the number of initial policyholders exposed to risk is available, and hence the graduated probability of death are given by $\hat{\eta}(x_i)$, the linear predictor in the Generalized Linear Models (GLMs) framework; while for those central exposed to risk, a local Poisson model is used and the graduated forces of mortality are derived by $\hat{\mu}(x_i / L_i$.

The unknown function,

$$\mu(x_i) = \mathbb{E}\left[Y \middle| X = x_i\right],$$

TABLE 1
DISTRIBUTIONS OF INTEREST BELONGINGS TO THE EXPONENTIAL DISPERSION FAMILY

Distribution of y_j	θ_j	m_j	$a_j(\phi)$	$b(\theta_j, m_j)$	$c(y_j, \phi)$
Poisson(μ_j)	$\log(\mu_j)$	1	1	$\exp(\theta_j)$	$-\log y_j!$
Binomial(L_j; q_j)	$\log\left(\frac{q_j}{1-q_j}\right)$	L_j	$\left(\frac{1}{L_j}\right)$	$L_j \log(1 + \exp\theta_j)$	$\log\binom{L_j}{L_j d_j}$

is modeled in X by a link function $g(.)$ such as

$$g\left(\frac{\mu(x_i)}{m_i}\right) = \eta(x_i).$$

$\mathbb{E}[Y_i]$ is tied to a linear combination $\sum_{j=1}^{n} w_j m_j \sum_{p=0}^{n} \beta_p (x_j - x_i)^p$ of the parameters β by a monotonous and differentiable function $g(.)$, not necessarily the identity. In consequence, the role of GLMs is that of a background model which is fitted locally. We proceed by forming the local likelihood as in (2) and estimate the coefficients $\beta = \beta_0, ..., \beta_p$ based on data in the neighborhood of the target point x_i. It consists of maximizing the local log-likelihood

$$L\left(\beta \middle| \boldsymbol{y}, w_j, \varphi, m_j\right) = \sum_{j=1}^{n} w_j \frac{y_j\, \theta_j - b(\theta_i, m_j)}{\varphi / m_j} + \sum_{j=1}^{n} w_j\, c(y_j, \varphi / m_j), \quad (3)$$

where $\mathbb{E}[Y_j] = b'(\theta_j, m_j) = \mu_j$ and $g(\mu_j / m_j) = \sum_{p=0}^{n} \beta_p (x_j - x_i)^p = \eta_j$, with $g(.)$ denoting the link function. Since we want to maximize the log likelihood for $\beta_0, \beta_1, ..., \beta_p$ we look for a solution of the set of normal equations to be fulfilled by the maximum likelihood parameter estimates β:

$$\frac{\partial L\left(\beta_v \middle| \boldsymbol{y}, w_j, \phi, m_j\right)}{\partial \beta_v} = 0 \Leftrightarrow \frac{1}{\phi} \sum_{j=1}^{n} w_j\, m_j\, (y_j - \mu_j) \frac{(x_j - x_i)^v}{b''(\theta_j, m_j) g'(\mu_j / m_j)} = 0$$

for $v = 0, 1, \ldots, p$.

These equations are usually non-linear, and so the solution must be obtained through iterative methods. One way to solve those is to use Fisher's scoring method. After some computations, see Tomas (2011), the Fisher scoring iterative equation are given by, in matrix notation,

$$\mathcal{I}\, \beta^* = \frac{1}{\phi} \boldsymbol{X}^T \boldsymbol{W}\, \Omega \boldsymbol{z},$$

where $\mathcal{I}$ denotes the Fisher information matrix, $\boldsymbol{X}$ is the design matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 - x_i & (x_1 - x_i)^2 & \dots & (x_1 - x_i)^P \\ 1 & x_2 - x_i & (x_2 - x_i)^2 & \dots & (x_2 - x_i)^P \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x_i & (x_n - x_i)^2 & \dots & (x_n - x_i)^P \end{bmatrix} \tag{4}$$

and $\boldsymbol{W}$ is a diagonal matrix, with entries $\{w_j\}_{j=1}^n$, such that

$$w_j = \begin{cases} W(\left|x_j - x_i\right| / h) & \text{if } \left|x_j - x_i\right| / h \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

$W()$ denotes a non-negative weight function depending on the target value x_i and the measurement points x_j, and in addition, it contains a smoothing parameter $h = (\lambda - 1) / 2$ which determines the sizes of the neighborhood of x_i. Ω is a diagonal matrix with elements

$$\omega_{jj} = \frac{m_j^2}{b''(\theta_j, m_j)} \left(\frac{\partial \mu_j}{\partial \eta_j} \right)^2, \tag{6}$$

depending on the variance and link function, since $\eta_j = g(\mu_j / m_j)$, we have $\partial \eta_j / \partial \mu_j = g'(\mu_j / m_j)$. Finally, $\boldsymbol{z}$ is the vector of the working dependent variables with elements

$$z_j = \hat{\eta}_j + \frac{y_j - \hat{\mu}_j}{m_j} g'(\hat{\mu}_j / m_j). \tag{7}$$

Hence, a maximum likelihood estimate of β is found by the following iterative process:

- Repeat $\beta^* := (\boldsymbol{X}^T \boldsymbol{W} \Omega \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \Omega \boldsymbol{z}$;
- using β^*, update the working weight Ω, as well as the working dependent variable $\boldsymbol{z}$ until convergence.

Estimation of β is performed using a Fisher's scoring method search in each neighborhood, going in order as i runs from 1 to n.

3.2 The Whittaker-Henderson model

The Whittaker-Henderson model is non-parametric and forms a relatively simple and natural version of Bayesian smoothing. The

method relies on the combination of a fit and smoothness measure. The chosen parameters minimize a linear combination of these two criteria,

$$M = F + h \times S,$$

where F and S denote the fit and smoothness measures respectively and h a parameter allowing more emphasis on the smoothness criterion. The fit and smoothness measures are

$$F = \sum_{i=1}^{n} v_i (y_i - \widehat{y}_i)^2 \quad \text{and} \quad S = \sum_{i=1}^{n-z} (\ {}^{z}y_i)^2,$$

where v_i represents the weight for observation i, taken generally as the ratio l_i / max(l_i); and z being an other parameter representing the polynomial degree. For this optimization problem, we solve the n equations given by the partial derivatives of M with respect to each of the y_i such that,

$$\frac{\partial M}{\partial y_i} = 0, \quad i = 1, \ldots, n.$$

With $\boldsymbol{y} = (y_i)_{1 \le i \le n}$, $\widehat{\boldsymbol{y}} = (\widehat{y}_i)_{1 \le i \le n}$ and $\boldsymbol{V}$ = diag$(v_i)_{1 \le i \le n}$, F can be written in matrix notation as

$$F = (\boldsymbol{y} - \widehat{\boldsymbol{y}})^T \boldsymbol{V} (\boldsymbol{y} - \widehat{\boldsymbol{y}}).$$

For the smoothness criterion, writing $\Delta^z \boldsymbol{y} = (\Delta^z y_i)_{1 \le i \le n-z}$, yields to $S = (\Delta^z \boldsymbol{y})^T \Delta^z \boldsymbol{y}$. To find $\Delta^z \boldsymbol{y}$, we introduce a matrix denoted K_z, of dimension $(n - z) \times z$, where the terms are binomial coefficients of order z and where the sign of the coefficients alternates and starts positively for z even, $\Delta^z \boldsymbol{y} = K_z \times y$.

The M criterion can finally be written as

$$\begin{aligned} M &= (\boldsymbol{y} - \widehat{\boldsymbol{y}})^T \boldsymbol{V} (\boldsymbol{y} - \widehat{\boldsymbol{y}}) + h \boldsymbol{y}^T K_z^T K_z \boldsymbol{y} \\ &= \boldsymbol{y}^T \boldsymbol{V} \boldsymbol{y} - 2 \boldsymbol{y}^T \boldsymbol{V} \widehat{\boldsymbol{y}} + \widehat{\boldsymbol{y}}^T \boldsymbol{V} \widehat{\boldsymbol{y}} + h \boldsymbol{y}^T K_z^T K_z \boldsymbol{y}. \end{aligned}$$

It leads to $\dfrac{\partial M}{\partial \boldsymbol{y}} = 2\boldsymbol{V}\boldsymbol{y} - 2\boldsymbol{V}\widehat{\boldsymbol{y}} + 2hK_z^T K_z \boldsymbol{y}$. Solving $\partial M / \partial \boldsymbol{y} = 0$ leads to the expression:

$$\widehat{\boldsymbol{y}} = (\boldsymbol{V} + hK_z^T K_z)^{-1} \boldsymbol{V} \boldsymbol{y}. \tag{8}$$

The form of the estimate is simple in that it is linear in the y_i. In consequence, we can apply the so-called *classical criteria* to find the optimal value of parameters h and z.

Local polynomials fitting have a long history in the smoothing of data, Henderson (1916) being one of the earliest classical references. Henderson was concerned with the smoothing properties of linear estimators, being the first to show that the smoothing power of a linear smoother depends on the shape of its weighting system.

4. WEIGHTING SYSTEMS

4.1 The smooth weight diagram

We have seen that, for the Whittaker-Henderson model the form of the estimate is simple in that it is linear in the responses y_i. That is, for each x_i there exists some smoothing weights $u_1(x_i)$, $u_2(x_i)$, ..., $u_n(x_i)$ such that

$$\widehat{y}_i = \sum_{j=1}^{n} u_j(x_i) y_j. \tag{9}$$

Likewise in a matrix form,

$$\begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_n \end{bmatrix} = \boldsymbol{U} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

where $\boldsymbol{U}$ denotes the smooth weight diagram for the Whittaker-Henderson model, an $n \times n$ matrix

$$\boldsymbol{U} = \begin{bmatrix} u_1(x_1) & u_2(x_1) & \dots & u_n(x_1) \\ u_1(x_2) & u_2(x_2) & \dots & u_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ u_1(x_n) & u_2(x_n) & \dots & u_n(x_n) \end{bmatrix},$$

with rows $\boldsymbol{u}(x_i)^T = \left(u_1(x_i), u_2(x_i), \ldots, u_n(x_i)\right) = \boldsymbol{e}_1^T(\boldsymbol{V} + hK_z^T K_z)^{-1}\boldsymbol{V}$, see expression (8). Here and throughout, we let $\boldsymbol{e}_v$ denote a column vector having 1 as its *v*th entry and all other entries equal to zero. The length of $\boldsymbol{e}_v$ will be clear from the context.

The weighting system of the Whittaker-Henderson model, $\boldsymbol{u}(x_i)$, depends on the polynomial degree and on the emphasis given to the smoothness criterion. These parameters are estimated by means of data dependent optimization criteria, such as *classical criteria*, see Section 4 in Tomas (2012).

Since the local likelihood estimate does not have an explicit representation, the smooth weight system can not be derived as in the previous case. However, we can provide an illustration of the smooth weight function associated with the *i*-th point at the last iteration. The weight function associated with the *i*-th point is used to compute the weights in the *i*-th row of the $n \times n$ matrix denoted by $\boldsymbol{S}$ with rows,

$$\boldsymbol{s}(x_i)^T = \boldsymbol{e}_1^T(\boldsymbol{X}^T\boldsymbol{W}\Omega\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\Omega. \tag{10}$$

with $\boldsymbol{X}$ denoting the design matrix such as expression (4) and $\boldsymbol{W}$ and Ω defined as in (5) and (6) respectively. The weighting system, $\boldsymbol{s}(x_i)$, depends on the shape of the weight function, the window width and the order of the polynomial. Moreover, it depends on the variance function and link function through the diagonal matrix Ω.

Note that by assuming Gaussian errors, Ω becomes an diagonal matrix with entries 1. In consequence the model reduces to a weighted least squares problem and the estimate is linear in the responses. The rows of the smooth weight diagram reduces to $\boldsymbol{e}_1^T(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}$.

4.2 The weighting system shape

The weighting system of local likelihood regression, depends on the constellation of smoothing parameters formed by the weight function, the bandwidth and the degree of the polynomial. In addition, it depends on the variance function and on the link function. The first three choices depend on assumptions we make about the behavior of the true curve. The two latest choices depend on the assumptions we make about other aspects of the distribution of the y_i.

The choice of link can be driven by convenience as, with local regression models, we do not assume the model to be globally correct. The variance function depends on the nature of the data considered. In consequence only the constellation formed by the weight function,

the bandwidth and the degree of the polynomial is estimated by means of data dependent optimization criteria, solving the bias and variance trade-off. However, by imposing the condition of a fixed bandwidth, we can study the weighting system of the smoother.

It is well know that between the three smoothing parameters, the weight function has much less influence on the bias and variance trade-off. The choice is not too crucial, at best it changes the visual quality of the regression curve. We consider a weight function $W(a)$ that is

- $W(a) > 0$ for $|a| < 1$;
- $W(-a) = W(a)$;
- $W(a)$ is a non increasing function for $a \geq 0$;

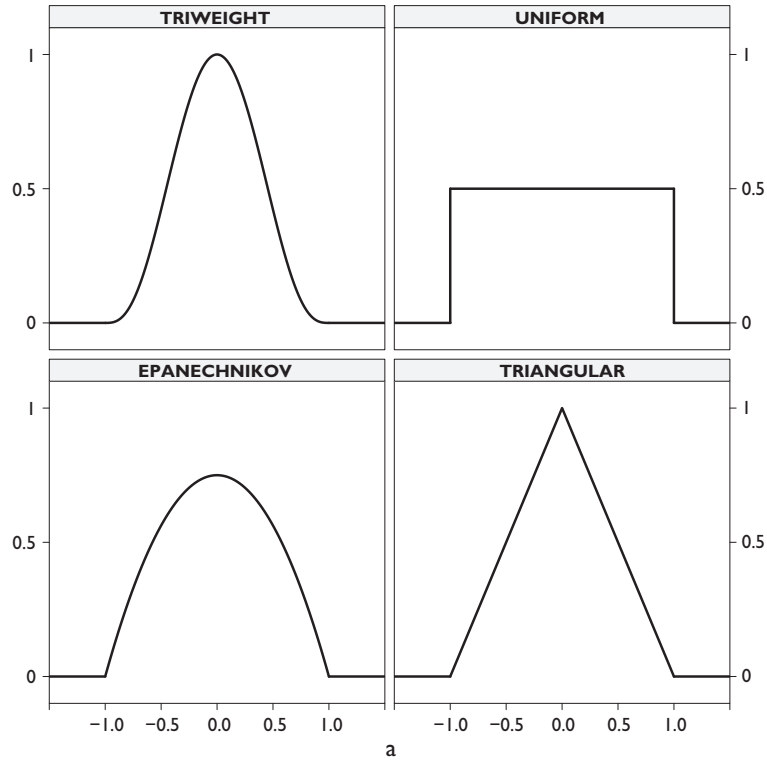
$W(a)$ is some weight function like those given in Table 2, below.

The requirements for $W(a)$ described above are needed for the following reasons: (i) is necessary, of course, since negative weights do not make sense; (ii) is required since there is no reason to treat points to the left of x_i differently from those to the right; (iii) is required for it seems unreasonable to allow a particular point to have less weight than one that is further from x_i. Figure 1 displays some of the weight functions presented above.

TABLE 2
EXAMPLE OF WEIGHT FUNCTIONS WITH $a = |x_j - x_i| / h$

Weight function	$W(a)$
Uniform or Rectangular	$\frac{1}{2} I(\lvert a \rvert \leq 1)$
Triangular	$(1-\lvert a \rvert) I(\lvert a \rvert \leq 1)$
Epanechnikov	$\frac{3}{4}(1-a^2) I(\lvert a \rvert \leq 1)$
Quartic (Biweight)	$\frac{15}{16}(1-a^2)^2 I(\lvert a \rvert \leq 1)$
Triweight	$\frac{35}{32}(1-a^2)^3 I(\lvert a \rvert \leq 1)$
Tricube	$(1-\lvert a^3 \rvert)^3 I(\lvert a \rvert \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{2} a^2\right)$

FIGURE 1
WEIGHTING SYSTEM SHAPE OF SOME WEIGHT FUNCTIONS

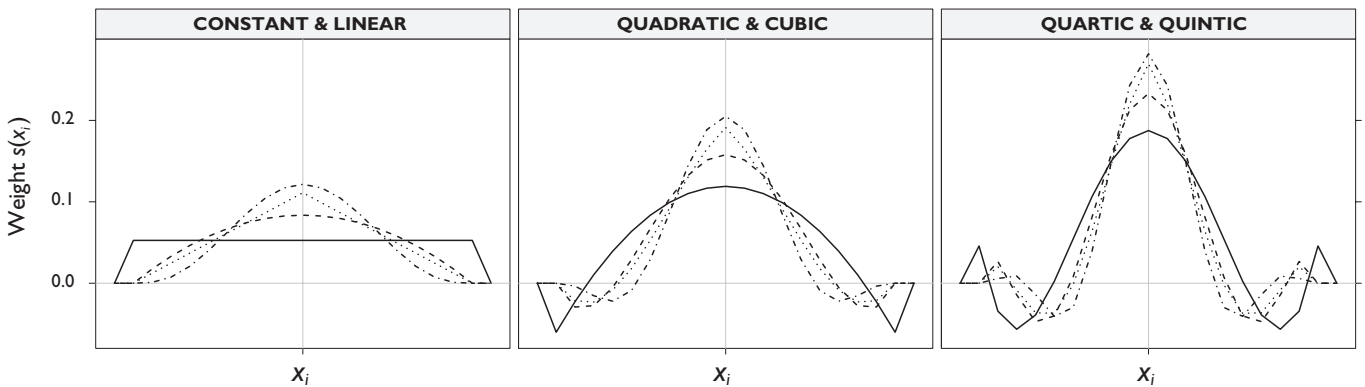


For a weight function $W(a)$, the weights decrease with increasing distance $|x_j - x_i|$. The window-width or bandwidth λ determines how fast the weights decrease. For small λ, only values in the immediate neighborhood of x_i will be influential; for large λ, values more distant from x_i may also influence the estimate. Such a weight function produces smoothed points that have a smooth appearance.

Figure 2 presents the smooth weights, $s(x_i)$, for the local Poisson log-likelihood model with the corresponding variance stabilizing link, according to the order of polynomial, for the four weighting system shapes drawn in Figure 1.

It is obvious that the triweight weight function has the smallest dispersion around the target point x_i while the rectangular weight function implies more smoothing. Note that the fit to a polynomial of even

FIGURE 2
SMOOTH WEIGHTS $s(x_i)$, FOR OBSERVATION i IN THE CENTRAL REGION, COMPUTED WITH $\lambda = 19$ FOR RECTANGULAR (SOLID LINE), TRIANGULAR (DOTTED LINE), EPANECHNIKOV (DASHED LINE) AND TRIWEIGHT (DOTDASHED LINE) WEIGHT FUNCTIONS



degree gives the same result as that of the next odd degree for values at the central region. It is due to the use of the variance stabilizing link reducing entries w_{jj} in (6) to a constant, see Section 6 in Tomas (2011).

This has been discussed for the least-squares fitting case by Fan and Gijbels (1995a, p. 215-218) and Ruppert and Wand (1994). It leads as well to symmetric smooth weights while for the Binomial model, by the use of the variance stabilizing link, the number of exposures appears in w_{jj}, see Section 5 in Tomas (2011), leading to asymmetric smooth weights. In addition the smooth weights obtained by a fit to a polynomial of even degree are not identical anymore as that of the next odd degree for a local likelihood Binomial model.

Figure 3 provides an illustration of the smooth weight diagram $\boldsymbol{S}$. The weight function associated with the i-th point is used to compute the weights in the i-th row, $B(x_i)$. $\boldsymbol{S}$ in Figure 3 has been computed with $\lambda = 19$, a polynomial of degree 3 and a triweight weight function with boundary correction *type 1*, see the Section 4.3.

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 triweight kernel such as Figure 2, center panel. But as the point, at which we are estimating the true curve, moves towards the boundaries, the kernel overlaps the boundary, becomes asymmetric and some weights are negative. Moreover, the height of the kernel increases because fewer observations are available.

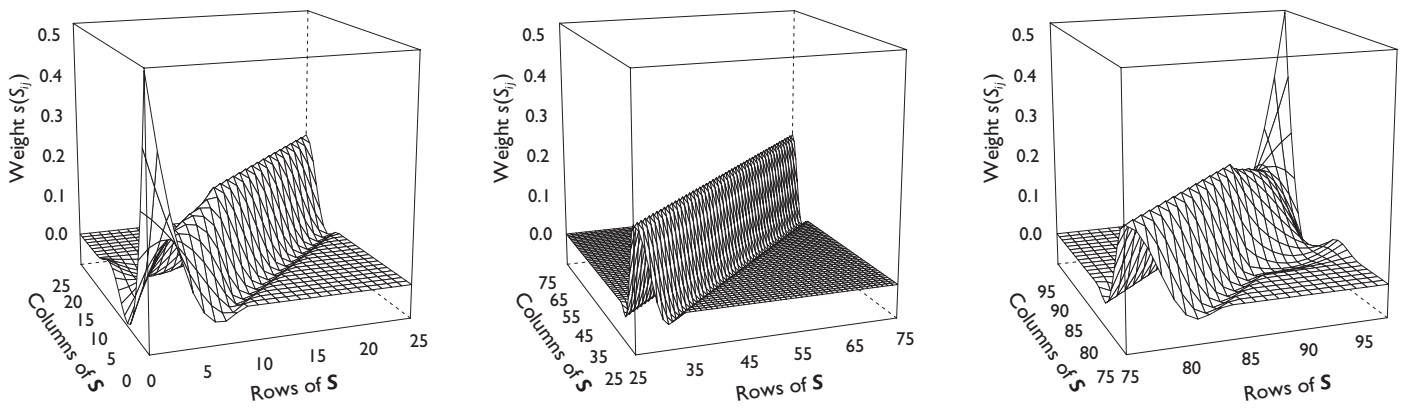
By fitting local polynomials models to series originating from life insurance, we observe a relatively high curvature in the boundaries. In consequence, the selection of the constellation of the smoothing parameters may be mainly driven by minimizing the criteria in the boundaries rather than for the whole data points. It may force the criteria to select a smaller bandwidth at the boundary to reduce the bias, but this may lead to under-smoothing in the middle of the table.

4.3 Specific treatments for the boundaries

To understand the boundary problem in the context of graduation, we study three specific treatments including symmetric and asymmetric weight systems.

- *type 1* uses an asymmetric weighting system. It always uses λ observations whatever the target point is. It means, for instance, for a target point at the left boundary, it uses all the observations κ available at the left side, and $\lambda - 1 - \kappa$ at the right side. Reciprocally for the right boundary. This type of

FIGURE 3
SMOOTHER S_{ij}: LEFT PANEL: *i*, *j* = 0, ..., 25, CENTER PANEL: *i*, *j* = 25, ..., 75, AND RIGHT PANEL: *i*, *j* = 75, ..., 98, COMPUTED WITH λ = 19, A POLYNOMIAL OF DEGREE 3, A TRIWEIGHT WEIGHT FUNCTION AND BOUNDARY CORRECTION *TYPE 1*



correction is found in most smoothing software such as the loess() or locfit() functions in R, R Development Core Team (2011).

- *type 2* uses a different asymmetric weighting system. For instance at the left boundary, it uses all observations available at the left side, and $(\lambda - 1) / 2$ observations at the right side. Reciprocally at the right boundary.
- *type 3* is a combination of observed rates and an adaptive symmetric weighting system. This correction is only applied to the left boundary. From age 0 to $v_{p,W}$ the mortality rates equal the observed ones. $v_{p,W}$ depends on the polynomial degree p and on the weight function $W(.)$ to ensure sufficiently observations to fit a polynomial of degree p. Then from $v_{p,W} + 1$ to $(\lambda - 1) / 2$, we use an adaptive symmetric window width with $2 \times (x_i - 1) + 1$ observations, where x_i is the target point. This correction is based on an idea presented by the Dutch Actuarieel Genootschap (the Dutch Actuarial Society), see Donselaar *et al*. (2007).

We apply these corrections to the smoothers, presented in Section 3, of degree 0 to 4 with four weighting system shapes. Figure 4 shows the symmetric and asymmetric weighting system $\boldsymbol{s}(x_i)$ for $i = 5$ (left boundary) of the corrections mentioned above with $\lambda = 19$.

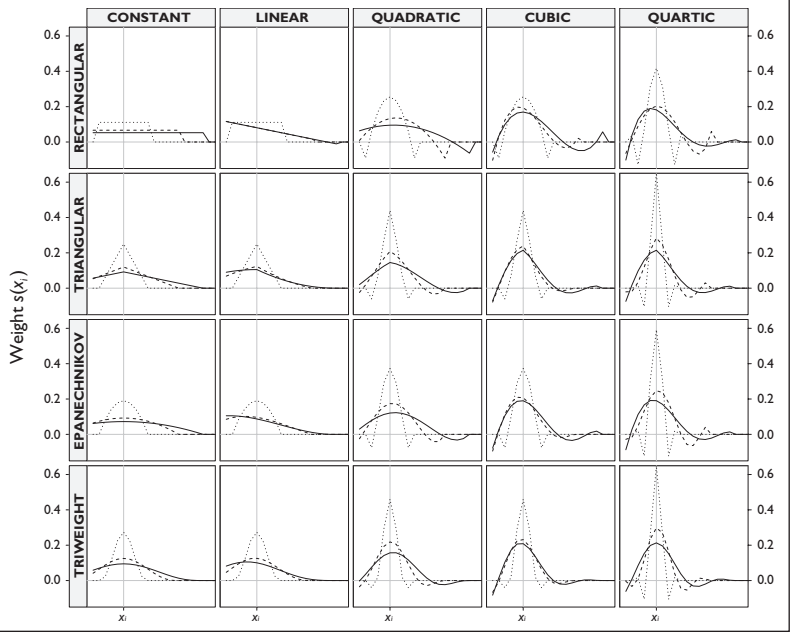
It is apparent that the symmetric weights of correction *type 3* have the smallest dispersion around the central value while correction *type 1* implies more smoothing.

5. THE SMOOTHING PROPERTIES

In this Section, we use local statistical measures based on the weighting system shapes and weight values to analyze the smoothing properties of the smoothers with the boundaries corrections introduced in the previous section. The window width, λ, is fixed to 19 observations and quantities such the fitted degrees of freedom and the influence values are used to measure the amount of smoothing applied. For ease of comparisons with the Whittaker-Henderson model smoother $\boldsymbol{U}$, h is fixed to 5 in expression (8) leading approximatively to 19 observations participating non negligibly to the estimation, having weights higher than $1e^{-2}$.

The computations are carried out with the help of the software R, R Development Core Team (2011). The scripts are available on request.

FIGURE 4
SMOOTH WEIGHTS $s(x_i)$ FOR i = 5 (LEFT BOUNDARY) WITH λ = 19 FOR CORRECTION *TYPE 1* (SOLID LINE), *TYPE 2* (DASHED LINE) AND *TYPE 3* (DOTTED LINE)



5.1 Fitted degrees of freedom

Although the notion of degrees of freedom (DF) does not really apply to smoothers, the usefulness is in providing a measure of the amount of smoothing that is comparable between different estimates applied to the same dataset.

For linear smoothers, among the several possible definitions, we define υ_2, the equivalent degrees of freedom, by $\upsilon_2 = \sum_{i=1}^{n} \left\| \boldsymbol{u}(x_i) \right\|^2 = \mathrm{tr}(\boldsymbol{U}\boldsymbol{U}^T)$. While for local likelihood models, the fitted DF are defined as $\upsilon_2 = \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}\left[\hat{\eta}(x_i)\right] \omega_{ii}$, where w_{ii} is defined as in (6).

As we face a fixed design model, in which we have a single observed mortality rate at equally spaced ages, the amount of smoothing applied by the local Poisson kernel-weighted log-likelihood is identical in the left and right boundary. Oppositely for the local Binomial and Whittaker-Henderson models for which the smooths weights depend on the sample size l_i. Hence the amount of smoothing is lower in the left boundary than to the right as the number of exposures is larger.

Table 3 presents the fitted DF for smoother $\boldsymbol{S}$ in the left boundary, that is for observations x_i for $i = 1, ..., 10$. (Recall $\lambda = 19$).

The corresponding fitted DF for the Whittaker-Henderson model in the left boundary are presented in Table 4.

The fitted DF aid interpretation in providing a measure of the amount of smoothing applied. For instance, 1 DF represents a smooth model with very little flexibility while 7 DF represents a noisy model showing many features.

It is obvious that the amount of smoothing decreases when increasing the degree of polynomial. In addition we observe that the amount of smoothing applied is higher when the weight function has a high dispersion around the central value. A rectangular weighting system shape implies very little flexibility. At the opposite, a triweight weighting shape shows more features.

Note again that a least-squares fit to a polynomial of even degree gives the same result as that of the next odd degree for symmetric weight function. For the local Binomial model, the difference is at three decimals and can not be seen due to rounding.

Finally, it is apparent that boundary correction *type 1* induces more smoothing in the boundaries than *type 2* and *type 3*. Correction *type 3*, having smooth weights showing the smallest dispersion, has the property of showing more features.

The amount of smoothing in the left boundary implied by the Whittaker-Henderson model is lying between corrections *type 1* and *type 2*. Hence the model has the ability to be slightly more flexible at the left boundary than applying correction *type 1*.

5.2 Influence values

The influence or leverage values, denoted $\text{infl}(x_i)$, are the diagonal elements $u_i(x_i)$ or equivalently $s_i(x_i)$ of the smooth weight diagram. These measure the sensitivity of the fitted curve to the individual data points. For local likelihood models, we define the influence function at x_i by

$$\text{infl}(x_i) = e_1^T (\boldsymbol{X}^T \boldsymbol{W} \Omega \boldsymbol{X})^{-1} e_1,$$

with $\boldsymbol{X}$, $\boldsymbol{W}$ and Ω defined as in expressions (4), (5) and (6) respectively. The property of *influence* relates to the fact that as $\text{infl}(x_i)$ approaches one, the corresponding residual approaches zero.

TABLE 3
FITTED DF FOR LOCAL POISSON AND BINOMIAL MODELS IN THE LEFT BOUNDARY

		Local Poisson model					Local Binomial model				
	Weight fct.	$p = 0$	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 0$	$p = 1$	$p = 2$	$p = 3$	$p = 4$
Corr. *type 1*	Rectangular	0.40	1.03	1.47	2.04	2.51	0.40	1.03	1.48	2.05	2.52
	Triangular	0.65	1.17	1.73	2.24	2.80	0.65	1.17	1.74	2.25	2.81
	Epanechnikov	0.56	1.11	1.65	2.17	2.72	0.57	1.12	1.66	2.18	2.73
	Triweight	0.78	1.27	1.91	2.39	2.99	0.79	1.27	1.92	2.40	2.99
Corr. *type 2*	Rectangular	0.66	1.33	1.94	2.61	3.21	0.66	1.33	1.95	2.62	3.22
	Triangular	1.03	1.95	2.87	3.67	4.37	1.03	1.96	2.88	3.67	4.38
	Epanechnikov	0.92	1.81	2.72	3.55	4.25	0.92	1.82	2.73	3.55	4.26
	Triweight	1.25	2.23	3.15	3.93	4.59	1.25	2.24	3.17	3.94	4.60
Corr. *type 3*	Rectangular	3.11		4.77		6.08	3.11	3.11	4.78	4.78	6.08
	Triangular	4.22		5.75		6.94	4.23	4.23	5.77	5.77	6.95
	Epanechnikov	4.11		5.66		6.88	4.12	4.12	5.68	5.68	6.89
	Triweight	4.40		5.90		7.02	4.41	4.41	5.92	5.92	7.03

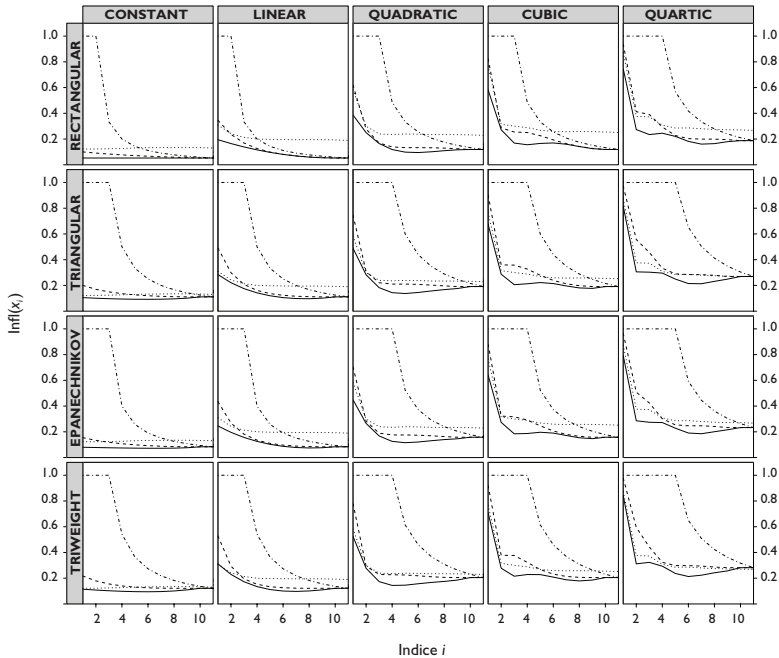
TABLE 4
FITTED DF IN THE LEFT BOUNDARY FOR THE WHITTAKER-HENDERSON MODEL, WITH $h = 5$

Whittaker-Henderson smoother				
$z = 0$	$z = 1$	$z = 2$	$z = 3$	$z = 4$
0.17	1.23	2.17	2.79	3.26

Figure 5 displays the influence values of the smoothers implied by the local Poisson and Whittaker-Henderson models.

We are not presenting the values for the local Binomial model as we have seen there is not much difference in the amount of smoothing applied between the two, and remarks drawn on the Poisson model apply to the Binomial as well.

FIGURE 5
INFLUENCE VALUES IN THE LEFT BOUNDARY FOR CORRECTION *TYPE 1* (SOLID BLACK LINE), *TYPE 2* (DASHED LINE), *TYPE 3* (DOTTED-DASHED LINE) AND WHITTAKER-HENDERSON MODEL (DOTTED LINE)



For correction *type 3*, from x_1 to $\upsilon_{p,W}$, the smoothed mortality rates equal the observed ones. In consequence, the corresponding influence values equal 1. $\upsilon_{p,W}$ depends on the degree of polynomial and on the weighting system to ensure that a sufficient number of observations is used to fit the corresponding polynomial.

By using a rectangular weighting system, corrections *type 1* and *type 2* gives similar results. Then, by using a weighting system shape inducing less dispersion around the central value, the differences become more apparent. The shape of the influence functions drawn by a triangular, Epanechnikov or Triweight weight function is relatively similar. In consequence, without loss of generality, we will not distinguish the weighting system shapes in the following sections.

Note that the influence values of the Whittaker-Henderson model are lying mostly between corrections *type 1* and *type 2*.

By a constant fit, the influence values for corrections *type 1*, *type 2* and Whittaker-Henderson model are approximatively equal to 0.1 indicating that y_i constitutes about 10% of the fitted value. But the main feature is the boundary effect where the influence function shows a huge increase. This reflects the difficulty of fitting a polynomial at boundary regions. Note also that the effect is more pronounced as we increase the order of polynomial. This shows that boundaries are a main concern when choosing the order of approximation and, more largely, the constellation of smoothing parameters.

6. CRITERIA AND CONTRIBUTION FROM THE DESIGN SPACE

The constellation of smoothing parameters are traditionally estimated by means of data dependent optimization criteria. We consider two class of criteria. The so-called *classical criteria* and the *plug-in* methodology introduced by Fan and Gijbels (1995b, p. 376-378) for local polynomial regression and extended for local likelihood inFan *et al*. (1998, p. 594-597).

We start by presenting the mortality data and give a brief description of the criteria. For a more detailed discussion, we refer the reader to Tomas (2012) and Tomas (2011). We analyze where the contributions to these criteria are coming from the design space. Finally we show that apart from statistical consideration, the choice of the parameters could be refined by taking into account the nature of the risk considered.

6.1 The data

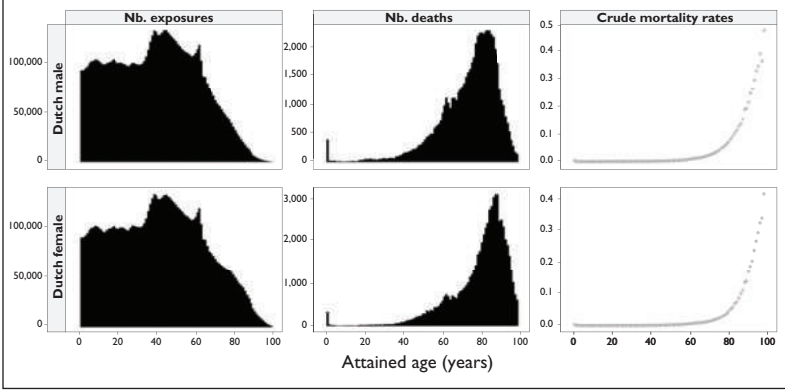
In the context of graduating mortality data, although age is a continuous variable, it is typically truncated in some way, such as age last birthday. Thus, the data consist of l_i observations at age x_i, of which d_i die and $l_i - d_i$ survive.

Given the discretized nature of a mortality table, it is natural to pool the data by using the average at each age, such as expression (1). This leads to a fixed design model, in which we have a single observed mortality rate at equally spaced ages.

For these applications, we focus on the measurements of the one-year probability of death for the Dutch Male and Female population for the year 2008 at age x_i, with $i = 0, ..., 98$. Figure 6 displays the observed statistics of the two datasets.

The data are reported by the Human Mortality Database (2011). The Human Mortality Database (HMD) has been initiated by the Department of Demography at the University of California Berkeley, USA, and the Max Planck Institute for Demographic Research, Rostock, Germany. This international project provides detailed mortality and population data which can be accessed online for research purposes.

FIGURE 6
OBSERVED STATISTICS FOR DUTCH MALE AND FEMALE POPULATION, 2008



Source: HMD.

6.2 The classical criteria

6.2.1 For linear smoothers

Classical methods are more or less natural extensions of methods used in parametric modeling. We consider the *AIC*, *GCV*, Rice's *T* and the *AICC* criterion.

The generalized cross-validation (*GCV*) score, as introduced by Craven and Wahba (1979) is

$$GCV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{f}(x_i)}{1-\frac{1}{n}\sum_j s_j(x_j)}\right)^2 = \frac{n}{(n-\upsilon_1)^2}\sum_{i=1}^{n}\left(y_i - \widehat{f}(x_i)\right)^2.$$

It is the sum of the single average squared error corrected by a factor. In this form, the criterion is very sensitive to the design space. Table 5 presents the proportion of the residuals sum of squares (*RSS*) in the boundaries given by the Whittaker-Henderson model targeting either the number of death d_i or the mortality rates q_i on the original scale.

The proportion of the *RSS* varies with the underlying structure of the data as well as the degree of polynomial *z* chosen. A quadratic fit leads to the highest disturbing nuisance while a cubic and quadratic fit perform better in the boundaries.

TABLE 5
PROPORTION OF THE *RSS* IN THE BOUNDARIES (IN %) BY THE WHITTAKER-HENDERSON MODEL, FOR THE DUTCH MALE AND FEMALE POPULATION, 2008

		Male population		Female population	
Target	**z**	**Left**	**Right**	**Left**	**Right**
d_i	2	10.07	52.63	6.63	41.13
	3	10.38	35.34	6.70	13.36
	4	7.11	31.52	4.04	16.60
q_i	2	0.01	97.24	0.12	95.14
	3	0.01	94.51	0.20	87.72
	4	0.01	94.66	0.11	86.93

Source: HMD.

By targeting the number of death d_i, we observe that most of the curvature appears in the central region, see Figure 6. In consequence, the selection of the parameters is less influenced by the boundaries than the model targeting the mortality rates on the original scale. However, the proportion of the *RSS* is still high with at most 62.70% for the 18 observations being in the boundaries (adding the left and right boundaries).

By targeting the mortality rates q_i, most of the curvature appears in the right boundary. The proportion of the *RSS*, in the right boundary, represents at least 86.93%. It is apparent that the selection of the parameters is driven by minimizing the *RSS* in the right boundary rather than the whole data.

However, the *GCV* can be seen as a special case of minimizing

$$\log(\widehat{\sigma}^2) + \psi(\boldsymbol{U}),$$

where $\psi(.)$ is a penalty function that decreases with increasing smoothness and $\widehat{\sigma}^2 = (1/n)\sum_i (y_i - \widehat{y})^2$ is the average squared residuals, see Hurvich *et al.* (1998, p. 273).

Table 6 presents the proportion of the natural logarithm of *RSS* in the boundaries given by the Whittaker-Henderson model targeting either the number of death d_i or the mortality rates q_i on the original scale.

TABLE 6
PROPORTION OF THE LOG(*RSS*) IN THE BOUNDARIES (IN %) BY THE WHITTAKER-HENDERSON MODEL, FOR THE DUTCH MALE AND FEMALE POPULATION, 2008

		Male population		Female population	
Target	**z**	**Left**	**Right**	**Left**	**Right**
d_i	2	11.05	17.62	10.62	17.83
	3	10.63	15.68	11.17	15.43
	4	10.00	13.49	8.85	15.46
q_i	2	8.91	4.99	8.66	5.56
	3	8.90	5.00	8.71	5.78
	4	9.20	5.04	9.24	6.02

Source: HMD.

By taking the natural logarithm of the average square errors, the variability is reduced and the criterion less affected by the boundaries.

The choice $\psi(\boldsymbol{U}) = -2\log(1 - \mathrm{tr}(\boldsymbol{U}/n))$ yields the *GCV* criterion, while $\psi(\boldsymbol{U}) = 2\,\mathrm{tr}(\boldsymbol{U}/n)$ yields the *AIC* criterion

$$\log(\widehat{\sigma}^2) + 2\,\mathrm{tr}(\boldsymbol{U})/n\ .$$

If $\psi(\boldsymbol{S}) = -\log\{1 - 2\mathrm{tr}(\boldsymbol{U}/n)\}$ is chosen, we obtain the criterion suggested by Rice (1984). A last alternative can be mentioned. Hurvich *et al.* (1998, p. 277) propose to use the criterion *AICC*, a corrected version of the *AIC*,

$$AICC = \log(\widehat{\sigma}^2) + 1 + \frac{2\left(\mathrm{tr}(\boldsymbol{U}) + 1\right)}{n - \mathrm{tr}(\boldsymbol{U}) - 2}.$$

The first term measures the quality of the adjustment while the second term evaluate the model complexity.

Figures 7, 9, 11, and 13 present the pointwise contribution to the criteria mentioned for the Whittaker-Henderson models, for the Dutch male and female population. They illustrate the homogenization of the pointwise contribution and the loss of influence of the boundaries.

Having in common the logarithm of the *RSS* as stochastic component, the criteria mentioned only differ by the penalty functions $\psi(\boldsymbol{U})$. Hence the pointwise contributions to the criteria, in Figures 7, 9, 11, and 13, display a relatively similar shape.

6.2.2 For likelihood models

In case of local likelihood models, it is natural to consider diagnostics based on the ratio $y_i / \widehat{\mu}(x_i)$. One possible loss function is the deviance (or *scaled deviance*) for a single observation (x_i, y_i), defined by

$$D\left(y_i, \widehat{\theta}(x_i)\right) = 2/\phi\, m_i \left(y_i\left(\theta(y_i) - \theta(\widehat{\mu}_i)\right) - b\left\{\theta(y_i)\right\} + b\left\{\theta(\widehat{\mu}_i)\right\}\right).$$

Examples of the form of deviances are given in Table 7.

The total deviance is defined as $\sum_{i=1}^{n} D\left(y_i, \widehat{\theta}(x_i)\right)$. It leads to a generalization of the Akaike information criterion to local likelihood models defined in Loader (1999, p. 69) as

TABLE 7
EXAMPLES OF FORMS OF SCALED DEVIANCE

GLM	Scaled Deviance
Poisson	$2/\phi \sum_i m_i \left(y_i \log(y_i / \widehat{\mu}_i) - (y_i - \widehat{\mu}_i) \right)$
Binomial	$2/\phi \sum_i m_i \left(y_i \log(y_i / \widehat{\mu}_i) + (n_i - yi) \log\left((n_i - y_i)/(n_i - \widehat{\mu}_i) \right) \right)$

$$AIC\left(\theta(\widehat{\mu}_i)\right) = \sum_{i=1}^{n} D\left(y_i, \left(\theta(\widehat{\mu}_i)\right)\right) + 2\,\upsilon_1,$$

where υ_1 is the degrees of freedom for the local likelihood fit, defined by $\sum_{i=1}^{n} \text{infl}(x_i)\,\omega_{ii}$, where w_{ii} is defined as in (6). Table 8 presents the proportion of the contribution to the *AIC* criterion in the boundaries given by the local likelihood models.

The contribution varies with the underlying structure of the data. The females mortality patterns are less pronounced than the males, and, thus the resulting contribution to the criterion is smaller. It is apparent that the local Poisson model is less influenced by the boundaries than the local Binomial model as most of the curvature appears in the central region.

Correction *type 1* leads to the highest contributions to the *AIC*. This treatment induces the highest amount of smoothing in the boundaries and thus leads to the highest disturbing nuisance when choosing the constellation of smoothing parameters. With at least 52.02% and 57.51% to 62.03% and 81.01% of the *AIC* in the boundaries (adding the left and right boundaries)respectively for the local Poisson and Binomial models, it is obvious that the selection of the smoothing parameter is driven by minimizing the criterion in the boundaries rather than for the whole data points.

The disturbing nuisance has reduced when treatment *type 2* is used. However the contribution to the *AIC* is still relatively high with at most 51.74% and 32.94%, for the local Poisson and Binomial models respectively, for the 18 observations in the boundaries.

Correction *type 3* implies smooth weights having the smallest dispersion around the central value. In consequence, it leads to the smallest disturbing nuisance. The contribution to the criterion for

TABLE 8
CONTRIBUTION TO THE *AIC* IN THE BOUNDARIES (IN %), FOR THE DUTCH MALE AND FEMALE POPULATION, 2008

		Local Poisson model				Local Binomial model			
		Male population		Female population		Male population		Female population	
Treatment	*p*	Left	Right	Left	Right	Left	Right	Left	Right
Type 1	2	49.21	12.82	45.07	14.64	78.48	2.53	73.20	4.52
	3	45.22	12.91	32.99	20.77	71.87	2.95	60.59	7.20
	4	38.42	15.15	26.54	25.48	63.31	3.41	48.74	8.77
Type 2	2	35.45	16.29	29.22	18.86	23.99	8.95	16.67	14.01
	3	27.86	17.00	19.37	24.99	14.33	8.99	8.80	16.65
	4	19.00	19.94	12.40	30.38	8.90	8.47	4.66	16.31
Type 3	2	1.14	25.92	0.94	26.37	4.06	11.32	2.76	16.46
	3	1.42	23.27	0.98	30.72	4.10	10.09	2.62	17.81
	4	1.56	24.31	1.14	34.30	4.17	8.93	2.61	16.67

Source: HMD.

observations in the left boundary has strongly reduced while the contribution in the right boundary has inflated. This type of treatment leads to under-smoothed figures in the left boundary and his relative merit would depend on the underlying smoothness of the data.

The pointwise contribution to the criteria is displayed in Figures 8 and 12 for the local Poisson model and in Figures 10 and 14 for the local Binomial model, left column, respectively for the Dutch male and female population, with the three specific treatments considered in section 4.3.

It is apparent that corrections *type 1* and *type 2* yield the highest contribution in the boundaries while correction *type 3* is less suffering from their influences. However, this treatment induces that observations in the right side receive more weight in the selection of the smoothing parameters than observations in the left side. For local Poisson model, we observe an important contribution to the criterion for observations around age 60, that is where the number of death shows a little hump due to a cohort effect, see Figure 6. This contribution is larger for the male than the female population, as the mortality patterns of the male population are more pronounced. This is true as well for the local Binomial model, where we observe, for the male population, an important hump around age 18 corresponding to the fit of the *accident hump*. This contribution is smaller for the female population as the *accident hump* is less accentuated.

It shows the resulting difficulty of applying a global smoothing approach when the true curve presents a clear structure and rapid changes in the curvature.

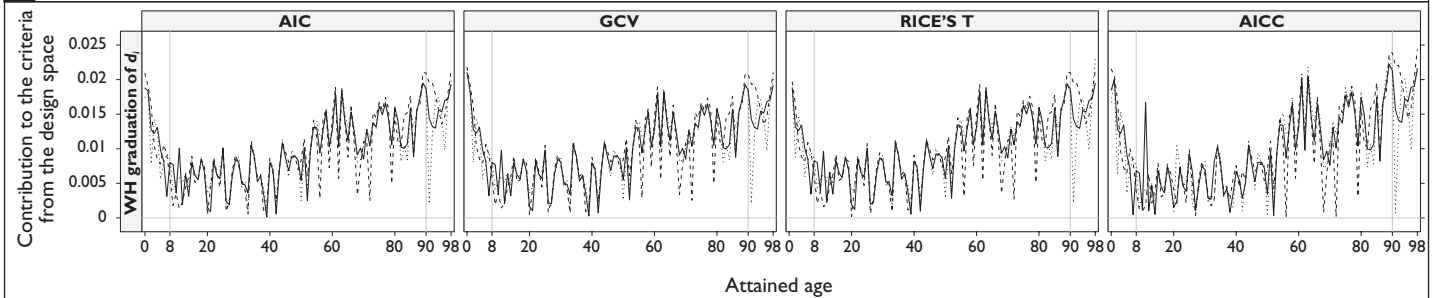
One solution for an homogeneous contribution of the design space to the criterion would be to modify the *AIC* by taking the logarithm of the deviance. It would lead, as for the criteria for linear smoothers, to a reduction in the variability and the criterion would be less affected by the boundaries.

6.3 The Plug-in methodology

Plug-in methods rely on an approximation of the bias via Taylor series expansions. The bias of the estimate is written as a function of the unknown ψ, and is approximated through Taylor series expansions. A pilot estimate of ψ is then plugged in to derive an estimate of the bias and hence an estimate of the mean squared error. The optimal bandwidth minimizes this estimated measure of fit.

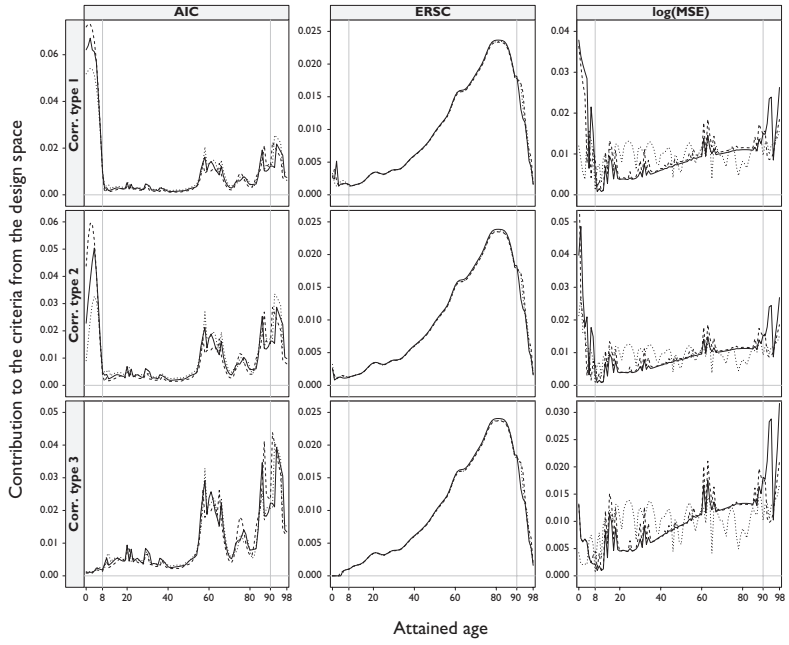
$$\widehat{MSE}_{p,v}(x_i,h) = (v!)^2\left(\widehat{\text{bias}}^2_{p,v}(x_i) + \widehat{\text{var}}_{p,v}(x_i)\right). \tag{11}$$

FIGURE 7
POINTWISE CONTRIBUTION TO THE CLASSICAL CRITERIA FOR THE WHITTAKER-HENDERSON MODEL TARGETING THE NUMBER OF DEATH, D_j, DUTCH MALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE)



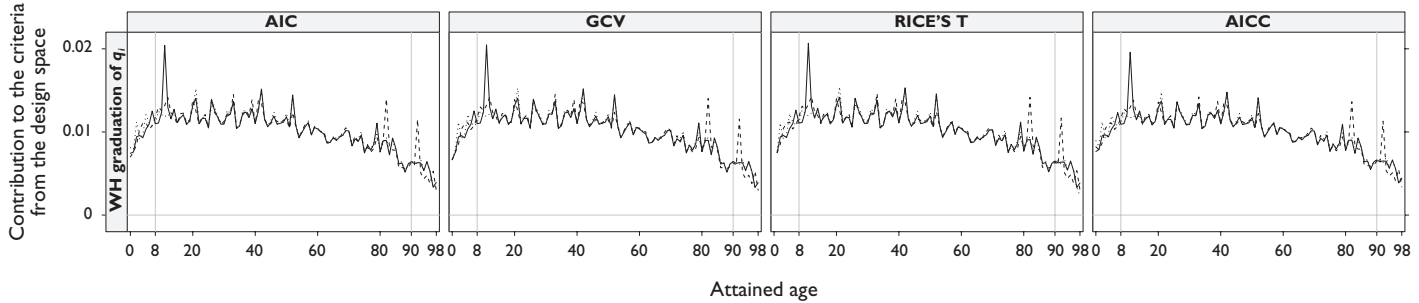
Source: HMD.

FIGURE 8
POINTWISE CONTRIBUTION TO THE CRITERIA FOR THE LOCAL POISSON MODEL TARGETING THE NUMBER OF DEATH, D_l, DUTCH MALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE)



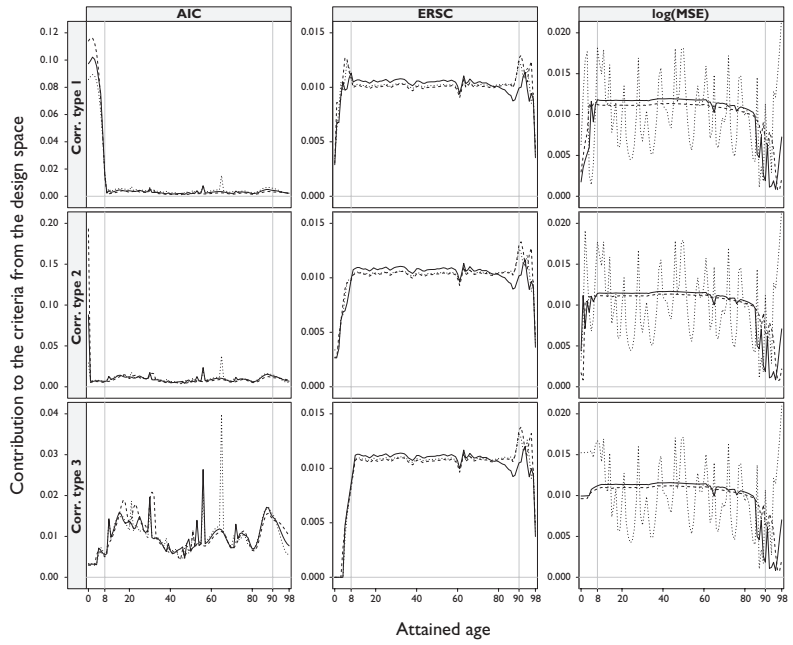
Source: HMD.

FIGURE 9
POINTWISE CONTRIBUTION TO THE CLASSICAL CRITERIA FOR THE WHITTAKER-HENDERSON MODEL TARGETING THE MORTALITY RATE, Q_j, DUTCH MALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE)



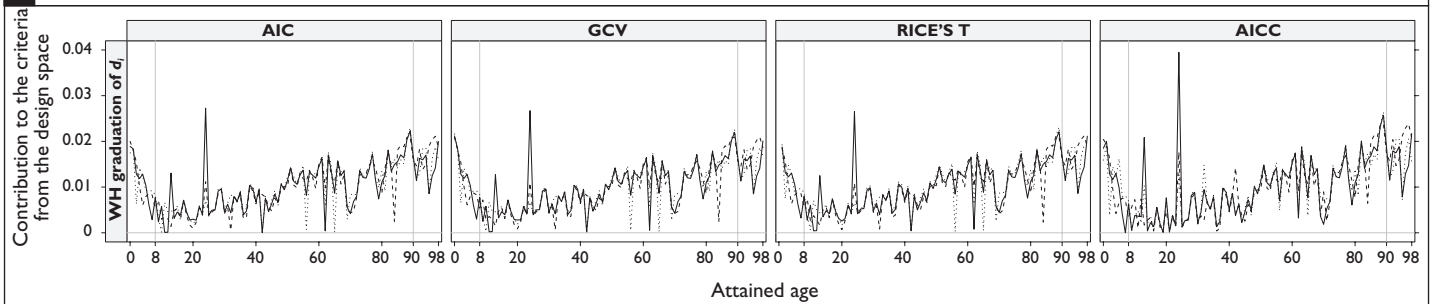
Source: HMD.

FIGURE 10
POINTWISE CONTRIBUTION TO THE CRITERIA FOR THE LOCAL BINOMIAL MODEL TARGETING THE MORTALITY RATE, Q_j, DUTCH MALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE)



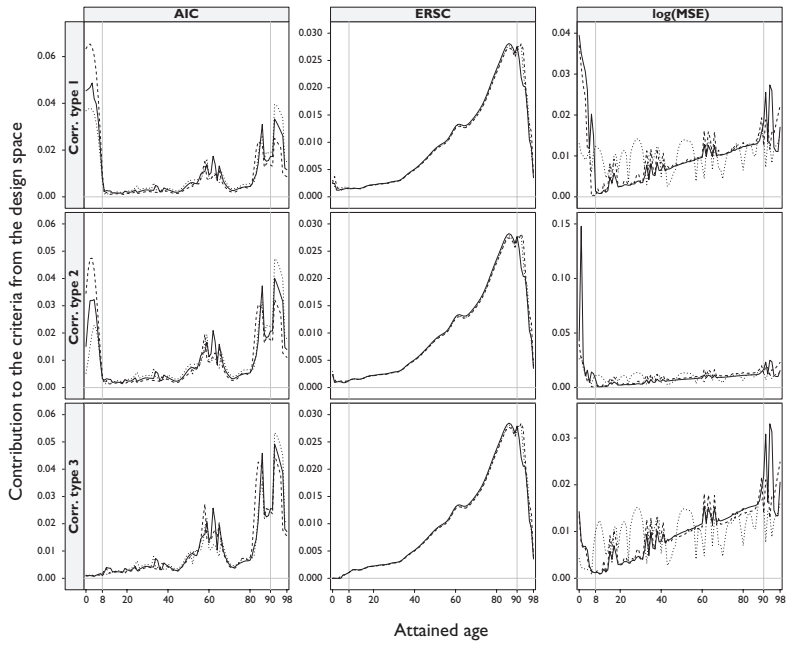
Source: HMD.

FIGURE 11
POINTWISE CONTRIBUTION TO THE CLASSICAL CRITERIA FOR THE WHITTAKER-HENDERSON MODEL TARGETING THE NUMBER OF DEATH, D_l, DUTCH FEMALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE



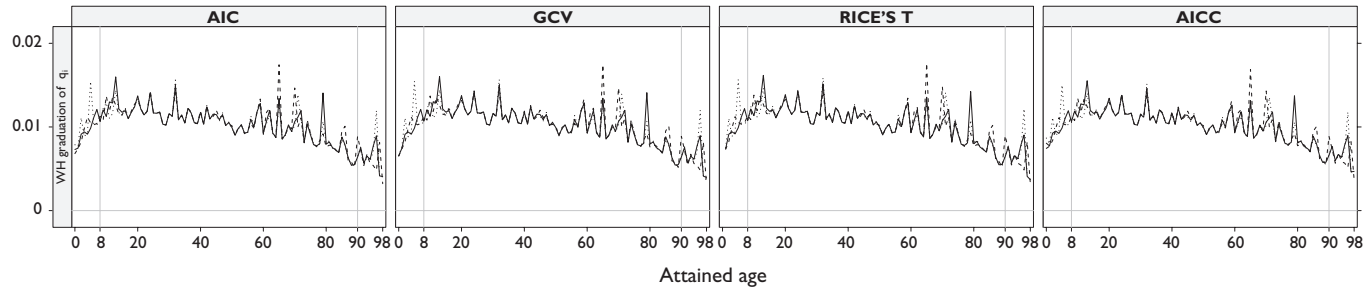
Source: HMD.

FIGURE 12
POINTWISE CONTRIBUTION TO THE CRITERIA FOR THE LOCAL POISSON MODEL TARGETING THE NUMBER OF DEATH, D_l, DUTCH FEMALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE)



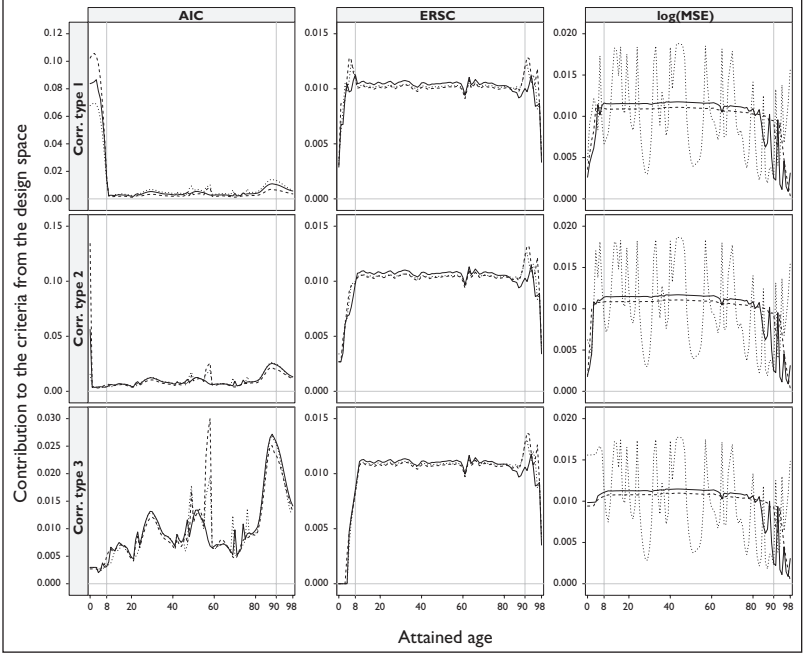
Source: HMD.

FIGURE 13
POINTWISE CONTRIBUTION TO THE CLASSICAL CRITERIA FOR THE WHITTAKER-HENDERSON MODEL TARGETING THE MORTALITY RATE, Q_j, DUTCH FEMALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE)



Source: HMD.

FIGURE 14
POINTWISE CONTRIBUTION TO THE CRITERIA FOR THE LOCAL BINOMIAL MODEL TARGETING THE MORTALITY RATE, Q_l, DUTCH FEMALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE)



Source: HMD.

With the estimated *MSE*, Fan *et al.* (1998, p. 599-600) formulate a bandwidth selection rule as follows: Fit a polynomial of order $p + a$ (choosing $a = 2$) and find the *pilot* bandwidth $h°$ that minimizes the integrated extended residual squares criterion,

$$IERSC(h) = \int ERSC(t,h)dt,$$

with the *ERSC* defined as

$$ERSC(x_i,h) = \widehat{\sigma}^2_{\circ}(x_i)\big(1+(p+1)/N\big),$$

where N^{-1} is the first diagonal element of the matrix $(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}^2 \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1}$ and $\widehat{\sigma}^2_{\circ}(x_i)$ is the normalized weighted residual sum of squares using the working dependent variable z defined as expression (7) after fitting locally a $(p + a)$th order polynomial. The intuition behind the *ERSC* criterion is that when the local polynomial does not fit well (the bandwidth is too large), the bias is large and hence also the residual sum of squares $\widehat{\sigma}^2_{\circ}(x_i)$. When the bandwidth is to small, the variance term N tends to be larger. So the *ERSC* quantity protects against both extreme choices.

Thus, having the optimal bandwidth $h°$ for estimating β_{p+1}, obtain estimates $\widehat{\beta}^{\circ}_{p+1}(x_i)$, $\widehat{\beta}^{\circ}_{p+2}(x_i)$ and $\widehat{\sigma}^2_{\circ}(x_i)$. With these estimated parameters, compute the estimated bias $\widehat{bias}_{p,v}(x_i)$ and variance $\widehat{\mathrm{var}}_{p,v}(x_i)$ of $\widehat{\beta}_v$. Combining these estimates yield to the estimated *MSE* (11). This leads to the bandwidth selector

$$\widehat{h}_{p,v} = \arg\min_h \left\{ \int \widehat{MSE}_{p,v}(t,h)dt \right\}.$$

The approach developed by Fan *et al.* (1998) makes it possible to assess the bias without going into deep asymptotics. It differs from the usual plug-in procedure (see for instance Park and Marron (1990), Sheather and Jones (1991), and Gasser *et al.* (1991) in the sense that the elements of the smooth weight diagram $\boldsymbol{S}$ are not further replaced by their asymptotics counterparts. These quantities are already known, andFan and Gijbels (1995b, p. 377) argue that replacing them by their corresponding asymptotic quantities introduces not only some extra approximation but also extra unknown parameters.

Figures 8, 10 and 12, 14, center column, show the pointwise contribution to the *ERSC* for the local Poisson model and the local

Binomial model respectively, for the Dutch male and female population, with the three specific treatments considered in section 4.3.

We observe that the contribution of the observations to the *ERSC* depends on the data on which the criterion is applied. For the local Poisson model, the contributions follow broadly the distribution of the number of death, while for the local Binomial model the contributions are more uniform, raising with the increasing curvature of the mortality rates when approaching the right boundary. Moreover, the criterion is not suffering from the boundary effects as the *RSS* component is weighted by the variance term *N*. Because the variance is larger at the boundaries, the resulting contributions of the observations are lower.

Table 9 presents the contribution (in %) to the log(*MSE*) in the boundaries given by the local likelihood models. We chose to use the natural logarithm of the (*MSE*) rather than the original scale because the selection of the smoothing parameter on the original scale is driven by minimizing the criterion in the boundaries at 99.99%. In consequence, by taking the natural logarithm, the variability is reduced and the choice of the smoothing parameters is not entirely dictated by the fit in the boundaries as shown in the table below.

As remarked previously, the contribution varies with the underlying structure of the data, however, the relation is reversed. The local Poisson model targeting the number of death is more influenced by the boundaries than the local Binomial model due to the log transform of the criterion.

Corrections *type 1* and *type 2* lead to similar results showing the highest disturbing nuisance. With at most 37.47% and 42.79% of the log(*MSE*) in the boundaries for the local Poisson, respectively for corrections *type 1* and *type 2*, the weight given to the 18 observations in the boundaries in selecting of the smoothing parameters is still relatively large.

Correction *type 3* leads to the smallest disturbing nuisance for the local Poisson model. The contribution for observations in the left boundary has strongly reduced while the contribution in the right boundary has inflated, being similar to the two other treatments. For the local Binomial model, the three specific treatments give relatively similar results. The benefit of using correction *type 3* is lost by using the log transform of the criterion.

The pointwise contribution to the log(*MSE*) is presented in Figures 8 and 12 for the local Poisson model and in Figures 10 and 14 for the local Binomial model, right column, respectively for the Dutch male and female population, with the three specific treatments considered in section 4.3.

TABLE 9
CONTRIBUTION TO THE LOG(*MSE*) IN THE BOUNDARIES (IN %), FOR THE DUTCH MALE AND FEMALE POPULATION, 2008

		Local Poisson model				Local Binomial model			
		Male population		Female population		Male population		Female population	
Treatment	*p*	Left	Right	Left	Right	Left	Right	Left	Right
Type 1	2	16.72	11.49	14.49	15.17	7.81	4.03	7.94	4.87
	3	21.07	16.40	21.24	16.17	6.46	2.83	7.07	2.73
	4	8.35	11.23	9.34	10.88	8.93	10.67	8.91	8.13
Type 2	2	16.21	11.56	11.49	15.70	7.88	4.03	8.15	4.86
	3	19.37	16.75	28.01	14.78	8.61	2.77	7.56	2.72
	4	12.10	10.77	12.39	10.51	10.03	10.54	9.69	8.06
Type 3	2	5.30	13.07	4.40	16.97	9.12	3.98	8.92	4.82
	3	4.94	19.75	4.86	19.52	9.37	2.74	9.27	2.67
	4	3.29	11.87	3.28	11.61	13.94	10.07	13.86	7.67

Source: HMD.

It is obvious that, in case of local Poisson model, the treatments *type 1* and *type 2* give the highest contribution in the boundaries while correction *type 3* is less suffering from their influence. We observe that observations in the right side, that is where the number of death is higher, contribute more in the criterion and thus receive more weight in the choice of the smoothing parameters.

In case of local Binomial model, we observe that observations in the central region contribute more to the criterion than observations in the boundaries. However, taking the log transform of the criterion does not only reduced the influence of the boundaries regions, but reduce the pointwise contribution with the increasing curvature of the observed mortality rates. In consequence, it may force the criterion to select a larger bandwidth. This may lead to over-smoothing in the end of the table and thus underestimating the mortality rates at the oldest ages.

Further, we consider restricting the computation of the criteria to observations in the central region and study where the contribution to these criteria are coming from the design space. Then, apart from statistical considerations, the choice of the smoothing parameters can be refined by taking into account the nature of the risk considered.

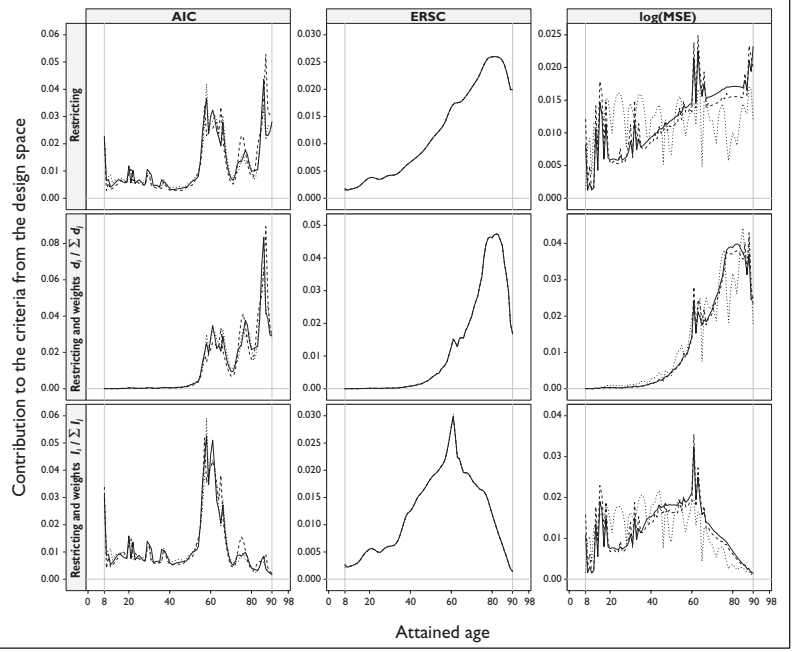
6.4 Practical considerations

Restricting the observations participating in the computation of the criteria helps to reduce the boundary effects argue Fan *et al.* (1998). At the boundaries, the pointwise contribution are too large because of numerical instabilities, underlying structure and scarcity of the data. Figures 15, 17 and 19, 21, first row, show the pointwise contribution to the criteria for the local Poisson model and the local Binomial model respectively, for the Dutch male and female population, when restricting the contribution to observations in the central region.

The pointwise contributions to the criteria differ due to the underlying structure of the data as the mortality patterns are more pronounced for the male than the female population. We observe that observations around age 18, corresponding to the *accident hump*, as well as observations around 60, corresponding to a cohort effect, contribute more to the criteria for the male population when fitting both of the local likelihood models.

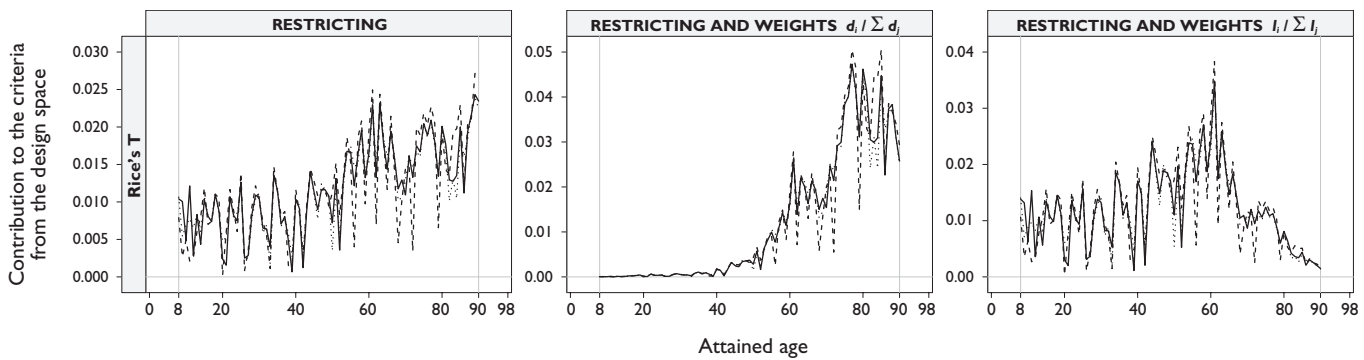
By fitting the local Poisson model, we notice an increase of the pointwise contribution with the number of death. This is particularly visible for the *ERSC* and $\log(MSE)$. On the other hand, in case of local Binomial model, the pointwise contribution to the *ERSC* and $\log(MSE)$ tends to decrease as the curvature of the observed mortality rates increases.

FIGURE 15
POINTWISE CONTRIBUTION TO THE CRITERIA WHEN RESTRICTING AND WEIGHTING THE OBSERVATIONS FOR THE LOCAL POISSON MODEL TARGETING THE NUMBER OF DEATH, D_l, DUTCH MALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE)



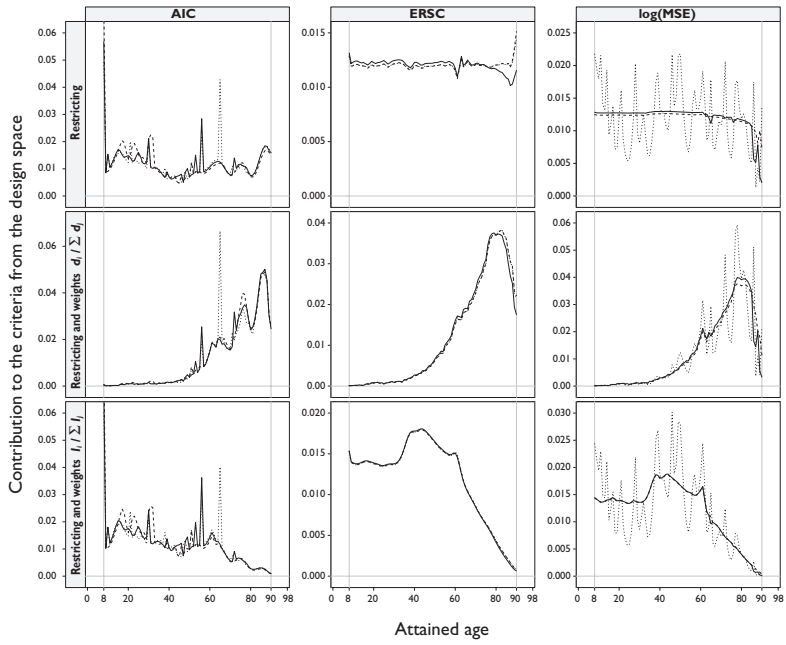
Source: HMD.

FIGURE 16
POINTWISE CONTRIBUTION TO THE RICE'S *T* CRITERION WHEN RESTRICTING AND WEIGHTING THE OBSERVATIONS FOR THE WHITTAKER-HENDERSON MODEL TARGETING THE NUMBER OF DEATH, D_j, DUTCH MALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE)



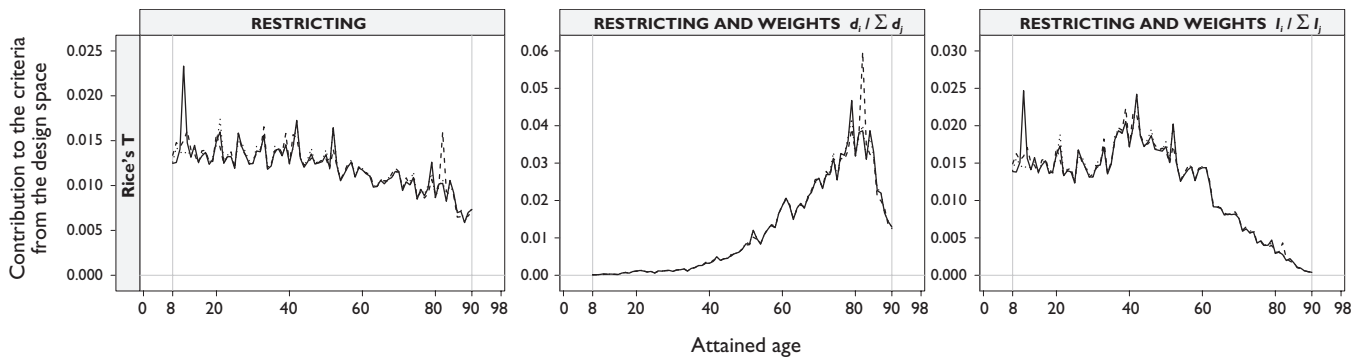
Source: HMD.

FIGURE 17
POINTWISE CONTRIBUTION TO THE CRITERIA WHEN RESTRICTING AND WEIGHTING THE OBSERVATIONS FOR THE LOCAL BINOMIAL MODEL TARGETING THE MORTALITY RATE, Q_t, DUTCH MALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE)



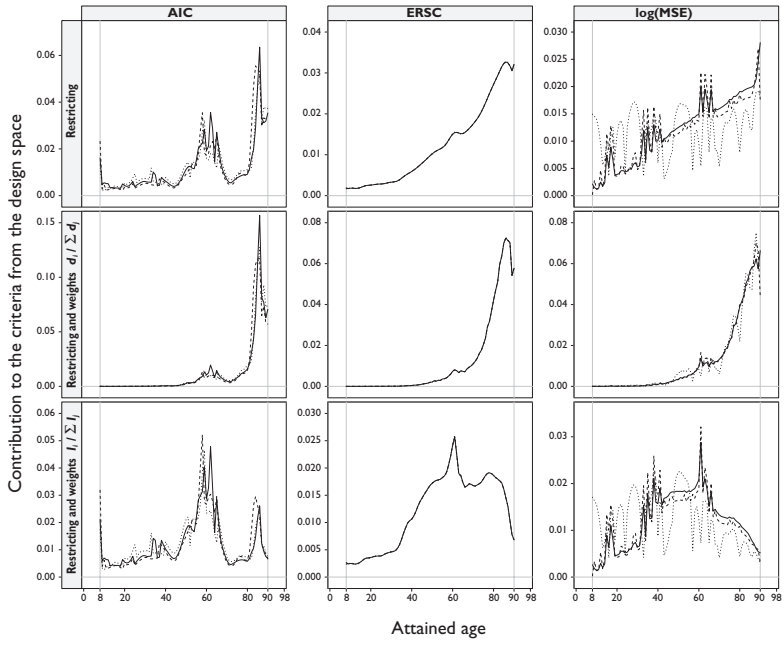
Source: HMD.

FIGURE 18
POINTWISE CONTRIBUTION TO THE RICE'S *T* CRITERION WHEN RESTRICTING AND WEIGHTING THE OBSERVATIONS FOR THE WHITTAKER-HENDERSON MODEL TARGETING THE MORTALITY RATE, Q_j, DUTCH MALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE)



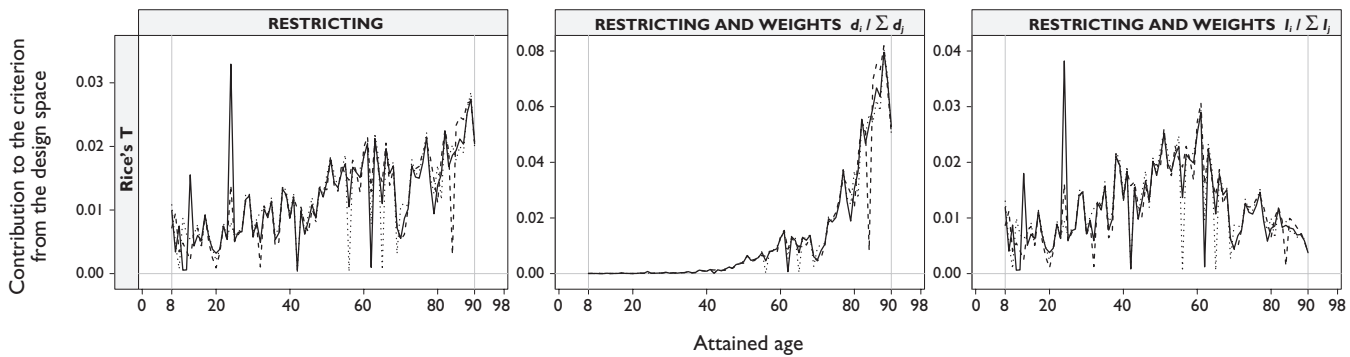
Source: HMD.

FIGURE 19
POINTWISE CONTRIBUTION TO THE CRITERIA WHEN RESTRICTING AND WEIGHTING THE OBSERVATIONS FOR THE LOCAL POISSON MODEL TARGETING THE NUMBER OF DEATH, D_l, DUTCH FEMALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE)



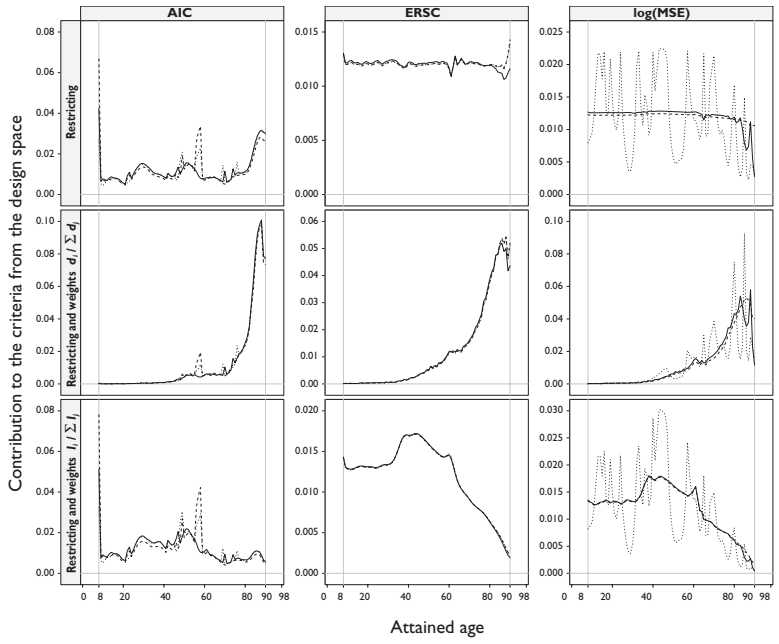
Source: HMD.

FIGURE 20
POINTWISE CONTRIBUTION TO THE RICE'S *T* CRITERION WHEN RESTRICTING AND WEIGHTING THE OBSERVATIONS FOR THE WHITTAKER-HENDERSON MODEL TARGETING THE NUMBER OF DEATH, D_l, DUTCH FEMALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE)



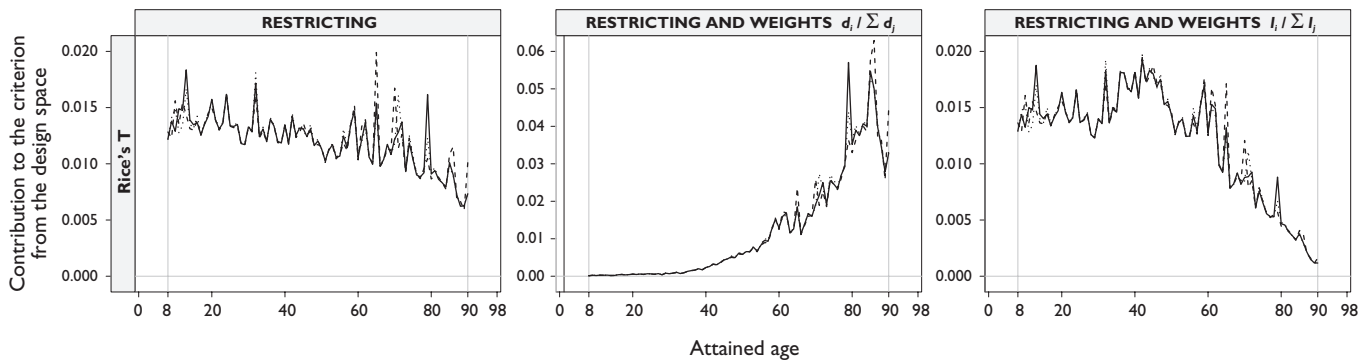
Source: HMD.

FIGURE 21
POINTWISE CONTRIBUTION TO THE CRITERIA WHEN RESTRICTING AND WEIGHTING THE OBSERVATIONS FOR THE LOCAL BINOMIAL MODEL TARGETING THE MORTALITY RATE, Q_t, DUTCH FEMALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE)



Source: HMD.

FIGURE 22
POINTWISE CONTRIBUTION TO THE RICE'S *T* CRITERION WHEN RESTRICTING AND WEIGHTING THE OBSERVATIONS FOR THE WHITTAKER-HENDERSON MODEL TARGETING THE MORTALITY RATE, Q_j, DUTCH FEMALE POPULATION, 2008. QUADRATIC FIT (DASHED LINE), CUBIC FIT (FULL LINE) AND QUARTIC FIT (DOTTED LINE)



Source: HMD.

These features can also be seen in the pointwise contribution to the Rice's T criterion used for linear smoother, shown in Figures 16, 18 and 20, 22, for the Whittaker-Henderson model targeting the number of death and the mortality rates on the original scale, for the Dutch male and female population respectively.

However, in graduating the mortality rates, the diminution of the pointwise contribution with the increasing curvature can be problem. It may force the criterion to select a larger bandwidth and this may lead to over-smoothing in the end of the table. It results in underestimating the mortality rates and in missing the mortality pattern of the oldest ages.

In practice, the search for an optimal criterion depends not only on statistical considerations but also on the nature of the risk considered. A smoothing method well suited for the annuities may be not the case for the death benefits. In the first case, we have to represent effectively the remaining life expectancy in the regions where the amount of exposure is high. In the second case, we have to well represent the observed deaths where the number of death is large and these regions may not necessarily those where there are more exposures, such as the female population.

By weighting the criteria according to the nature of the risk considered, we can take these practical considerations into account. The choice of the constellation of the smoothing parameters can be refined by weighting the criteria according the reliability of the data,

- by $l_i \Big/ \sum_j (l_j)$ in case of annuities, and

- by $d_i \Big/ \sum_j (d_j)$ in case of death benefits.

Table 10 presents the contribution to the criteria for observations in the age range representing 80% of the exposures and number of death for the Dutch male and female population after weighting the criteria according to the nature of the risk considered.

Figures 15, 17 and 19, 21, second and third row, show the pointwise contribution to the criteria for the local Poisson model and the local Binomial model respectively, for the Dutch male and female population, when restricting the contribution to observations in the central region and weighting according the reliability of the data.

For the male population, 80% of the exposures appears in the age range 8-67. For the female population the age range corresponds to 8-70. In case of annuities, by weighting by l_j / $\Sigma_j(l_j)$ most of the

TABLE 10
CONTRIBUTION TO THE CRITERIA (IN %) FOR OBSERVATIONS IN THE AGE RANGE REPRESENTING 80% OF THE EXPOSURES AND NUMBER OF DEATH FOR THE DUTCH MALE AND FEMALE POPULATION, 2008. COMPUTED WITH A CUBIC FIT AND A TRIWEIGHT WEIGHT FUNCTION

				Local Poisson			Local Binomial		
Population	Age range	l_i	Rice's *T*	*AIC*	*ERSC*	log(*MSE*)	*AIC*	*ERSC*	log(*MSE*)
Male	8-67	80	91.27	85.85	72.87	82.92	90.06	89.20	89.84
Female	8-70	80	90.38	81.86	67.98	79.54	84.74	88.36	86.40
Population	Age range	d_i	Rice's *T*	*AIC*	*ERSC*	log(*MSE*)	*AIC*	*ERSC*	log(*MSE*)
Male	59-90	80	89.76	90.07	93.09	89.91	86.83	85.05	83.49
Female	46-90	80	98.53	99.26	99.34	98.52	98.21	96.34	96.28

Source: HMD

criteria applied to the local Poisson model (force of mortality) and to the Binomial model (probability of death) provides a good representation. The contribution to these criteria, for observations in the age range considered, are mostly above 80%. Only the *ERSC* provides a poor representation when fitting the local Poisson model due to the distribution of the criterion following broadly the observed number of death.

80% of the deaths appears in the age ranges 59-90 and 46-90 respectively for the male and female population. In case of death benefits, by weighting the criteria by $d_i / \sum_j (d_j)$, the proportion of the contributions from observations in the age range are above 80% showing a good representation of the risk considered.

For linear smoothers, the representation of the risk given by the Rice's T and variations of the classical criteria is satisfactory. In consequence, weighting the criteria by the reliability of the data leads to a better representation of the nature of the risk considered whatever the model used, graduating the force of mortality or the probability of death.

7. CONCLUSIONS AND FURTHER RESEARCH

Local regression combines excellent theoretical properties with conceptual simplicity and flexibility. It is very adaptable, and it is also convenient statistically.

However, for the purpose of graduating series originating from life insurance, the boundaries effects are real problems. For graduating the force of mortality by a local Poisson model targeting the number of death, or graduating the probability of death by a local Binomial model targeting the mortality rates, the selection of the smoothing parameters by classical criteria is driven by minimizing the criteria in the left boundary rather than the whole data. In consequence it forces the criteria to select a smaller bandwidth at the boundary to reduce the bias, leading to under-smoothed figures in the middle of the table.

We studied three specific treatments to reduce these boundaries effects including symmetric and asymmetric weighting systems. Between the treatments considered, correction *type 1* leads to the worst results. This correction leads to the highest amount of smoothing applied in the boundaries. In consequence, by minimizing the deviance in the boundaries, the criteria select a smaller bandwidth to reduce the bias in the boundaries leading to under-smoothed figures in the central region. Note that this type of correction is found in most

in smoothing softwares. It implies that a lot of care should be taken when selecting the constellation of smoothing parameters and exclusive reliance in practice on a global criterion is unwise because a global criterion does not provide information about where the contributions are coming from the design space.

One solution leading to homogeneous contributions for likelihood models, would be to take the log transform of the criteria in a similar manner than the criteria for linear smoothers. It would reduce the variability and the criteria would be less affected by the boundaries effects.

However, such transformation does not solve all the issues. By using a plug-in method for local Binomial model or graduating the mortality rates by the Whittaker-Henderson model, the contribution of the observations to the estimated log(*MSE*) or respectively by the classical criteria decreases with the increasing curvature of the observed mortality rates. It may force the criterion to select a larger bandwidth and this may lead to over-smoothing in the end of the table, resulting in underestimating the mortality rates and missing the mortality pattern of the oldest ages.

Finally, we restricted the observations contributing to the criteria to the central region and applied weights according to the reliability of the data. These practical considerations enhance clearly the optimization criteria and the choice of the constellation of the smoothing parameters is refined, leading to a good representation of the risk considered.

These are illustrations of the weakness of a global bandwidth and call for an adaptive smoothing procedure. Rather than restricting the observations to the central region, we would use a more flexible approach. It would be to allow the constellation of smoothing parameters to vary across the age range to vary the amount of smoothing in a location dependent manner or to allow adjustment based on the reliability of the data and on the nature of the risk considered.

References

Brillinger, D. R. (1986), The natural variability of vital rates and associated statistics. *Biometrics*, 42(4), 693-734.

Cleveland, W. S. (1979), Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368), 829-836.

Cleveland, W. S. and Loader, C. R. (1996), Smoothing by local regression: principles and methods. In *Statistical Theory and Computational Aspects of Smoothing*, pages 10-49. W. Hardle and M. G. Schimek, eds.

Copas, J. B. and Haberman, S. (1983), Non-parametric graduation using kernel methods. *Journal of the Institute of Actuaries*, 110, 135-156.

Craven, P. and Wahba, G. (1979), Smoothing noisy data with spline functions. *Numerische Mathematik*, 31, 377- 403.

Daw, R. H. (1980), Johann Heinrich Lambert (1728-1777). *Journal of Institute of Actuaries*, 107, 345-363.

Donselaar, J., Attema, J., Broekhoven, H., Roodenburg-Berkhout, L., Willemse, W., and Zijp, P. (2007), On mortality and life expectancy. *Dutch Actuarial Association (Actuarieel Genootschap)*.

Fan, J. and Gijbels, I. (1995a), Adaptive order polynomial fitting: bandwidth robustification and bias reduction. *Journal of Computational and Graphical Statistics*, 4(3), 213-227.

Fan, J. and Gijbels, I. (1995b), Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society*, 57(2), 371-394.

Fan, J., Farmen, M., and Gijbels, I. (1998), Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society*, 60(3), 591-608.

Gasser, T., Müller, H.-G., and Mammitzsch, V. (1985), Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society*, 47(2), 238-252.

Gasser, T., Kneip, A., and Kohler, W. (1991), A flexible and fast method for automatic smoothing. *Journal of the American Statistical Association*, 86(415), 643-652.

Gavin, J. B., Haberman, S., and Verrall, R. J. (1993), Moving weighted graduation using kernel estimation. *Insurance: Mathematics & Economics*, 12(2), 113-126.

Gavin, J. B., Haberman, S., and Verrall, R. J. (1995), Graduation by kernel and adaptive kernel methods with a boundary correction. *Transactions of the Society of Actuaries*, XLVII, 173-209.

Haberman, S. (1996), Landmarks in the history of actuarial science (up to 1919). *Actuarial Research Paper No. 84, Dept. of Actuarial Science and Statistics*, City University, London.

Hastie, T. and Loader, C. R. (1993), Local regression: automatic kernel carpentry (with discussion). *Statistical Science*, 2, 120-143.

Henderson, R. (1916), Note on graduation by adjusted average. *Transactions of the Actuarial society*, 17, 43-48.

Human Mortality Database (2011), University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on June 2011).

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998), Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society*, 60(2), 271-293.

Loader, C. R. (1996), Local likelihood density estimation. *The Annals of Statistics*, 24(4), 1602-1618.

Loader, C. R. (1999), *Local Regression and Likelihood*. Statistics and Computing Series. New York: Springer Verlag.

Müller, H. G. (1987), Weighted local regression and kernel method for nonparametric curve fitting. *Journal of the American Statistical Association*, 82, 231-238.

Park, B. U. and Marron, J. S. (1990), Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, 85(409), 66-72.

Parzen, E. (1962), On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3), 1065-1076.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Rice, J. A. (1984), Bandwidth choice for non-parametric regression. *Annals of Statistics*, 12(4), 1215-1230.

Rosenblatt, M. (1956), Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27(3), 832-837.

Ruppert, D. and Wand, M. P. (1994), Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22(3), 1346-1370.

Schucany, W. R. (1989), On nonparametric regression with high-order kernels. *Journal of Statistical Planning and Inference*, 23, 141-151.

Seal, H. L. (1982), Graduation by piecewise cubic polynomials: a historical review. *Blatter der Deutschen Gesellschaft fur Versicherungsmathematik*, 15, 89-114.

Sheather, S. J. and Jones, M. C. (1991), A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, 53(3), 683-690.

Stone, C. J. (1977), Consistent nonparametric regression (with discussion). *Annals of Statistics*, 5(4), 595-645.

Stone, C. J. (1980), Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*, 8(6), 1348-1360.

Taylor, G. (1992), A bayesian interpretation of whittaker-henderson graduation. *Insurance: Mathematics and Economics*, 11(1), 7-16.

Tibshirani, R. J. and Hastie, T. J. (1987), Local likelihood estimation. *Journal of the American Statistical Association*, 82(398), 559-567.

Tomas, J. (2011), A local likelihood approach to univariate graduation of mortality. *Bulletin Français d'Actuariat*, 11(22), 105-153.

Tomas, J. (2012), Univariate graduation of mortality by local polynomial regression, pages 1-38. Accepted for publication – *Bulletin Français d'Actuariat*.

Watson, G. S. (1964), Smooth regression analysis. *Sankhya: The Indian Journal of Statistics*, 26(4), 359-372.

Whittaker, E. T. (1923), On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41, 62-75.

Notes

1. ASE-RI - University of Amsterdam, Roetersstraat 11– 1018 WB Amsterdam – The Netherlands.

2. Université de Lyon – Université Claude bernard Lyon 1 – 50 Avenue Tony Garnier – 69366 Lyon Cedex 07 – France.