MULTIDIMENSIONAL SMOOTHING BY ADAPTIVE LOCAL KERNEL-WEIGHTED LOG-LIKELIHOOD: APPLICATION TO LONG-TERM CARE INSURANCE

Julien TOMAS [α] [*] Frédéric PLANCHET [α] [†]

[α] *ISFA - Laboratoire SAF* [‡]

April 26, 2016

Abstract

We are interested in modeling the mortality of Long-Term Care (LTC) claimants having the same level of severeness (heavy claimant). Practitioners often use empirical methods that rely heavily on experts opinion. We propose approaches not depending on experts advice. We analyze the mortality as a function of both the age of occurrence of the claim and the duration of the care. LTC-claimants are marked by a relatively complex mortality pattern. Hence rather than using parametric approaches or models with expert opinions, adaptive local likelihood methods allow to extract pertinently the information from the data. We distinguish a locally adaptive smoothing pointwise method using the intersection of confidence intervals rule, as well as a global method using local bandwidth correction factors. The latest is an extension of the adaptive kernel method proposed by Gavin *et al.* (1995) to likelihoods techniques. We vary the amount of smoothing in a location dependent manner and allow adjustments based on the reliability of the data. Tests, and single indices summarizing the lifetime probability distribution are used to compare the graduated series obtained by adaptive local kernel-weighted log-likelihoods to p-splines and local likelihood models.

Keywords. Adaptive parameters choice, Local likelihood, p-splines, Long-term care insurance, Graduation.

Résumé

Nous nous intéressons à la construction de la loi de survie d'individus dépendants ayant le même niveau de sévérité (dépendance lourde). En pratique, les actuaires utilisent souvent des méthodes s'appuyant fortement sur l'opinion d'experts. Nous proposons des approches ne dépendant pas d'avis d'experts. La mortalité est analysée en fonction de l'âge à la survenance de la pathologie et l'ancienneté. La mortalité des dépendants est caractérisée par une structure relativement complexe. Plutôt que d'utiliser des approches paramétriques ou des modèles avec avis d'experts, les méthodes adaptatives de vraisemblance locale permettent d'extraire de façon pertinente l'information contenue dans les données en variant les paramètres du lissage selon l'âge à la survenance et l'ancienneté. Nous distinguons une méthode ponctuelle de vraisemblance locale utilisant la règle de l'intersection des intervalles de confiance et un modèle global avec des facteurs d'ajustement local de la fenêtre d'observations. La dernière est une extension de la méthode adaptative à noyaux proposée par Gavin *et al.* (1995) aux techniques de vraisemblance. Nous modifions le niveau de lissage en fonction de l'emplacement et nous permettons des ajustements de la fenêtre d'observations en fonction de la fiabilité des données. Des tests et marqueurs résumant les distributions de survie sont utilisés pour comparer les séries graduées obtenues par les méthodes adaptatives de vraisemblance locale aux modèles de p-splines et vraisemblance locale.

Mots-clés. Choix adaptif des paramètres, Vraisemblance locale, p-splines, Assurance dépendance, Graduation.

[*]Contact: julien.tomas@univ-lyon.fr. All computer programs are available on request.

[†]Contact: frederic.planchet@univ-lyon1.fr.

[‡]Institut de Science Financière et d'Assurances - Université Claude Bernard Lyon 1 - 50 Avenue Tony Garnier - 69366 Lyon - France

Contents

1 Introduction

In this article, we are interested in modeling the mortality of Long-Term Care (LTC) claimants. LTC is a mix of social and health care provided on a daily basis, formally or informally, at home or in institutions, to people suffering from a loss of mobility and autonomy in their activity of daily living. Although loss of autonomy may occur at any age, its frequency rises with age. LTC insurance contracts are individual or collective and guarantee the payment of a fixed allowance, in the form of monthly cash benefit, possibly proportional to the degree of dependency, see Kessler (2008) and Courbage and Roudaut (2011) for studies on the French LTC insurance market.

Most of the actuarial publications on this topic focus on the construction of models of projected benefits, see Deléglise *et al.* (2009) and modeling the life-history of LTC-patients using Markovian multi-state models. Gauzère *et al.* (1999) model the progression of the pathologies. They use a non-parametric approach to derive smoothed estimates for the different transition intensities with a multi-state model. This approach has the advantage of avoiding unreal assumptions, such as constant intensities resulting from the homogeneous Markov model. However, they assumes that the transition probabilities depend only on the age of the LTC-patients. This assumption is inadequate when modeling heavy LTC-claimants mortality. Czado and Rudolph (2002) assess the influence of factors like severeness of a pathology, gender and type of care on the survival curve of the observed claims. Using the estimated hazard rates of the Cox (1972) proportional hazard model as transition intensities between level of severity in a multiple state Markow model, they are able to fit a multiple state insurance model. In contrast to Cox's proportional hazard model where the transition probabilities are calculated from the transition intensities, Helms *et al.* (2005) estimate the transition probabilities directly which are then used to compute actuarial values of a given LTC-plan and the required premiums. Recently, the risk coming from uncertainty in future demographical trends and the relevant impact on an LTC portfolio has been addressed by Levantesi and Menzietti (2012). Given a Markovian Multi-state mode, they define the benefits payable to the policyholder and the premiums payable to the annuity provider, when the insured is in a specific state. They propose then a stochastic model to assess mortality and disability risk in life annuities with LTC benefits and measure the risk evolution.
In health insurance, the effect of the age of the policyholder is often of unknown nonlinear form. In addition, neglecting the effect of calendar time of claims and district where the policyholder lives in modeling the claims process lead to biased fits with corresponding consequences for risk premium calculation. Lang *et al.* (2002) present a space-time analyses of insurance data and allow to explore temporal and spacial effect simultaneously with the impact of other covariates. They apply a semi-parametric Bayesian approach for unified treatment of such effects within a joint model, developed in the context of generalized additive mixed models. Lang and Umlauf (2010) have extended a hierarchical version of regression models with structured additive predictor allowing nonlinear covariate terms in every level of the hierarchy. Their approach can deal simultaneously with nonlinear covariate effects and time trends, unit- or cluster specific heterogeneity, spatial heterogeneity and complex interactions between covariates of different type.

In contrast with the approaches exposed previously, we have no exogenous information about the LTC-claimants, neither the gender, place or level of care. We observe only the aggregated exposition and number of deaths over two dimensions. These are the age of occurrence of the claim and the duration of the care. Here, LTC-claimants belong only to one state of severeness (heavy claimants) and we are concerned about the construction of the survival distribution.

The pricing and reserving as well as the management of LTC portfolios are very sensitive to the choice of the mortality table adopted. In addition, the construction of such table is a difficult exercise due to the following features:

i. the mortality law consists of a mixture of pathologies and non-monotonic phenomenons appear;

ii. French LTC portfolios are relatively small and the estimation of crude death rates is very volatile;

iii. because of the strong link between the age of occurrence of the claim and the related pathology, it is usual to construct a mortality table based on both age of occurrence of the claim, *which is an explanatory variable*, and duration of the care (or seniority), *which is the duration variable*. Hence, it is necessary to construct a mortality surface;

iv. mortality rates decrease very rapidly with the duration of the care. In consequence, the first year is often difficult to integrate, disqualifying the usual parametric approaches.

Thus practitioners often use empirical methods that rely heavily on experts opinion. We therefore propose, in this article, methods not depending on experts advice and allowing to extract more pertinently the information from the data. Unlike the problems presented in the literature above, this issue has not been addressed extensively.

We analyze the survival law as a function of both the age of occurrence of the claim and the duration of the care. The life table values computed are estimates of the true parameters, based on the finite amount of data available. The data examined should be regarded as a sample. Estimates based on the data will be subject to sampling errors and the smaller the group is, the greater will be the relative random errors in the number of deaths and the less reliable will be the resulting estimates. However, we wish to smooth these quantities to enlighten the characteristics of the mortality of LTC-claimants which we think to be relatively regular.

The article is organized as follow. Section 2 has still an introductory purpose. It makes precise the notation used in the following and introduce the ideas behind local likelihood fitting. Section 3 present the non-parametric estimators and covers model selection issues. The adaptive methods are introduced in Section 4. We distinguish the intersection of confidence intervals rule and local bandwidth correction factors. Section 5 discusses the application on LTC-claimants. Tests and single indices summarizing the probability lifetime distribution are used to compare the graduated series with those obtained from global non-parametric approaches, p-splines and local likelihood, in Section 6. Finally, some remarks in Section 7 conclude the paper.

2 Notation, assumptions and Premises of local likelihood fitting

2.1 Notation and assumptions

We analyze the changes in mortality of individuals subscribing LTC insurance policies as a function of both the duration of the care and the age of occurrence of the claim. Let $T_u(v)$ be the remaining lifetime of an individual when the pathology occurred at age v, for the duration of the care u, with v and u being integers. We are working with two temporal dimensions u and v, however, they do not have the same status: v is a variable denoting the heterogeneity while u represents the variable linked with the duration. The distribution function of $T_u(v)$ is denoted as ${}_\tau q_u(v) = \Pr\big[T_u(v) \leq \tau\big] = 1 - {}_\tau p_u(v)$. The force of mortality at duration $u+\tau$ for the age of occurrence v, denoted by $\varphi_{u+\tau}(v)$ is defined by

$$\varphi_{u+\tau}(v) = \lim_{\Delta\tau\to 0^+} \frac{\Pr\big[\tau < T_u(v) \leq \tau + \Delta\tau | T_u(v) > \tau\big]}{\Delta\tau} = \frac{1}{{}_\tau p_u(v)} \frac{\partial}{\partial\tau} {}_\tau q_u(v),$$

$$\text{and,}\quad {}_\tau p_u(v) = \exp\left(-\int_0^\tau \varphi_{u+\xi}(v+\xi)\, d\xi\right).$$

We assume that the duration-specific forces of mortality are piecewise constant in each unit square, but allowed to vary from one unit square to the next, $\varphi_{u+\tau}(v+\xi) = \varphi_u(v)$ for $0 \leq \tau < 1$ and $0 \leq \xi < 1$. Under this assumption, $p_u(v) = \exp\big(-\varphi_u(v)\big) \Leftrightarrow \varphi_u(v) = -\log\big(p_u(v)\big)$.

We define the exposure-to-risk $(E_{u,v})$, measuring the time during which individuals are exposed to the risk of dying. It is the total time lived by these individuals. Assume that we have $L_{u,v}$ individuals at duration u and age of occurrence v. Using the notation of Gschlössl *et al.* (2011), we associate to each of these $L_{u,v}$ individuals the dummy variable

$$\delta_i = \begin{cases} 1 & \text{if individual } i \text{ dies,} \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, 2, \ldots, L_{u,v}$. We define the time lived by individual i before $(u+1)$st duration when the pathology occurred at age v by τ_i. We assume that we have at our disposal iid observations (δ_i, τ_i) for each of the $L_{u,v}$ individuals. The contribution of individual i to the likelihood equals $\exp\big(-\tau_i\varphi_u(v)\big)\big(\varphi_u(v)\big)^{\delta_i}$. Finally we define

$$\sum_{i=1}^{L_{u,v}} \tau_i = E_{u,v} \quad \text{and} \quad \sum_{i=1}^{L_{u,v}} \delta_i = D_{u,v}.$$

Under these assumptions, the likelihood becomes

$$\mathcal{L}(\varphi_u(v)) = \prod_{i=1}^{L_{u,v}} \exp\big(-\tau_i\varphi_u(v)\big)\big(\varphi_u(v)\big)^{\delta_i} = \exp\big(-E_{u,v}\ \varphi_u(v)\big)\big(\varphi_u(v)\big)^{D_{u,v}}.$$

The associated log-likelihood is $\ell\big(\varphi_u(v)\big) = \log\mathcal{L}\big(\varphi_u(v)\big) = -E_{u,v}\,\varphi_u(v) + D_{u,v}\,\log\varphi_u(v)$. Maximizing the log-likelihood $\ell\big(\varphi_u(v)\big)$ gives $\widehat{\varphi}_u(v) = D_{u,v}/E_{u,v}$ which coincides with the central death rates $\widehat{m}_u(v)$. Then it is apparent that the likelihood $\ell\big(\varphi_u(v)\big)$ is proportional to the Poisson likelihood based on $D_{u,v} \sim \text{Poisson}\big(E_{u,v}\,\varphi_u(v)\big)$ and it is equivalent to work on the basis of the *true* likelihood or on the basis of the Poisson likelihood, as recalled in Gschlössl *et al.* (2011). Thus, under the assumption of constant forces of mortality between non-integer values of u and v, we consider

$$D_{u,v} \sim \text{Poisson}\big(E_{u,v}\,\varphi_u(v)\big), \tag{1}$$

to take advantage of the Generalized Linear Models (GLMs) framework for statistical inference.

2.2 Premises of local likelihood fitting

A common prior opinion about the form of the forces of mortality is that each force of mortality is closely related to its neighbors. This relationship is expressed, recall Gavin *et al.* (1993), by the belief that the forces of mortality progress smoothly from one observation to the next. It follows that the data for several observations on either side of a point x_i can be used to augment the basic information we have at x_i, and an improved estimate can be obtained by smoothing the individual estimates.
Following the approach taken by Tibshirani and Hastie (1987), local likelihood models apply the local fitting technic to data of which the relationship can be expressed through a likelihood function. Suppose we have n independent realizations $y_1, y_2, \ldots, y_n$ of the random variable Y with

$$Y_i \sim f\big(Y|\theta(x_i)\big), \quad \text{for } i = 1, 2, \ldots, n,$$

where $f\big(\cdot|\theta(x_i)\big)$ is a probability mass/density function in the exponential dispersion family and $\theta(x_i)$, the natural parameter in the GLMs framework, is an unspecified smooth function $\psi(x_i)$. For simplicity, we use $x_i = (u_i, v_i)$ to denote the vector of the predictor variables.

The bivariate local likelihood fits a polynomial model locally within a bivariate smoothing window. Suppose that the function ψ has a $(p+1)$st continuous derivative at the point $x_i = (u_i, v_i)$. For data point $x_j = (u_j, v_j)$ in a neighborhood of $x_i = (u_i, v_i)$ we approximate $\psi(x_j)$ via a Taylor expansion by a polynomial of degree p.

If locally linear fitting is used, the fitting variables are just the independent variables. If locally quadratic fitting is used, the fitting variables are the independent variables, their squares and their cross-products. For example, a local quadratic approximation is:

$$\begin{aligned}\psi(x_j) =& \psi(u_j, v_j) \approx \beta_0(x_i) + \beta_1(x_i)\,(u_j - u_i) + \beta_2(x_i)\,(v_j - v_i) \\ &+ \frac{1}{2}\beta_3(x_i)(u_j - u_i)^2 + \beta_4(x_i)(u_j - u_i)(v_j - v_i) + \frac{1}{2}\beta_5(x_i)(v_j - v_i)^2.\end{aligned}$$

The local log-likelihood can be written as

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

where, in the case of locally quadratic fitting, $\boldsymbol{x} = \left(1, u_j - u_i, v_j - v_i, (u_j - u_i)^2, (v_j - v_i)(u_j - u_i), (v_j - v_i)^2\right)^T$, and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_5)^T$.
The weights are defined on the bivariate space. The non-negative weight function, $w_j = w_j(x_i)$, depends on the distance $\rho(x_i, x_j)$ between the observations $x_j = (u_j, v_j)$ and the fitting point $x_i = (u_i, v_i)$ and in addition, it contains a smoothing parameter $h = (\lambda - 1)/2$ which determines the radius of the neighborhood of x_i.
Maximizing the local log-likelihood (2) with respect to $\boldsymbol{\beta}$ gives the vector of estimators $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_5)^T$. Estimator $\psi(x_i)$ is given by $\widehat{\psi}(x_i) = \widehat{\beta}_0$.

Local fitting technics combines excellent theoretical properties with conceptual simplicity and flexibility. It is very adaptable, and it is also convenient statistically, see Loader (1999) for an extensive discussion of the strengths of local modeling.

In a recent study, Tomas (2011) shows the applicability of local kernel-weighted log-likelihoods to model the relation between the crude death rates and forces of mortality with attained age.

Unfortunately, as we face situation where data have a large amount of structure, this simplicity has flaws. For instance, 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, see Tomas (2012).

In consequence, in some cases no global smoothing parameter provides an adequate fit to the data. Rather than restricting the smoothing parameters to a fixed value, a more flexible approach is to allow the constellation of smoothing parameters to vary across the data.

We distinguish locally adaptive smoothing pointwise method using the intersection of confidence intervals rule as well as global method local bandwidth correction factors. The latest is an extension of the adaptive kernel method proposed by Gavin *et al.* (1995) to likelihoods techniques. We vary the amount of smoothing in a location dependent manner and allow adjustments based on the reliability of the data.

3 The non-parametric global estimators

3.1 Local likelihood Poisson model

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:

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

for specific functions $a(.)$, $b(.)$ and $c(.)$ and where ϕ is called the dispersion parameter. It is a nuisance parameter not depending on x_j. The function a and c are such that $a_j(\phi) = \phi$ and $c = c(y_i,\phi)$. The Poisson example, of interest, is presented in Table 1.

Distribution of y_j	θ_j	$a_j(\phi)$	$b(\theta_j)$	$c(y_j,\phi)$
Poisson(μ_j)	$\log(\mu_j)$	1	$\exp(\theta_j)$	$-\log y_j!$

Table 1: Poisson distribution belonging to the Exponential Dispersion Family.

The unknown function $\mu(x_j) = \mathbb{E}\big[Y|X = x_j\big]$ is modeled in X by a link function $g(.)$ such as $g\big(\mu(x_j)\big) = \eta(x_j)$. $\mathbb{E}\big[Y_i\big]$ is tied to a linear combination of the parameters β by a monotonous and differentiable function $g(.)$. Here, the role of Generalized Linear Models (Nelder and Wedderburn (1972) and McCullagh and Nelder (1989)) is that of a background model which is fitted locally. In a parametric GLM, $\eta(x_i) = \beta_0 + \beta_1 x_i$ for some unknown parameter β_0 and β_1. In a non-parametric setting, there is no model assumption about $\eta(x_i)$. The primary goal is to estimate $\mu(x_i)$ or equivalently $\eta(x_i)$, non-parametrically, that is, $\beta_0 + \beta_1 x_i$ is generalized to $\psi(x_i)$. The obvious extension of this idea is to suppose that $\eta(x_i)$ is a pth degree polynomial in x_j, with x_j being an element of the neighborhood of x_i. Therefore, fitting procedures that are familiar from GLMs are needed, but, of course, the modeling itself is smooth and no longer parametric.

Extensive experience graduation using GLMs have been built up in the actuarial literature with Renshaw (1991) and reviewed by Haberman and Renshaw (1996). Local likelihood methods to graduate of mortality tables have been applied in Delwarde *et al.* (2004), Debón *et al.* (2006) and more recently in Gschlössl *et al.* (2011) and Tomas (2011). However their works only cover smoothing in one dimension.

We proceed by forming the local likelihood as in (2) and estimate the coefficients $\boldsymbol{\beta}$ based on data in the neighborhood $x_j = (u_j, v_j)$ of the target point $x_i = (u_i, v_i)$. It consists of maximizing the local log-likelihood

$$L\big(\boldsymbol{\beta}|\boldsymbol{y}, w_j, \phi\big) = \sum_{j=1}^{n} w_j \frac{y_j\ \theta_j - b(\theta_i)}{\phi} + \sum_{j=1}^{n} w_j\ c(y_j,\phi),$$

where $\mathbb{E}\big[Y_j\big] = b'(\theta_j) = \mu_j$ and $g(\mu_j) = \eta_j$, with $g(.)$ denoting the link function. For our Poisson case, the probability distribution function can be formulated as $f_D(d_j;\mu_j) = e^{-\mu_j}\mu^{d_j}/d_j!$ and in exponential family form as

$$f_D(d_j;\mu_j) = \exp\big\{d_j\log(\mu_j) - \mu_j - \log\ d_j!\big\}.$$

The local log-likelihood function at x_i can be abstracted from the exponential form of the distribution,

$$L(\mu_i) = \sum_{j=1}^{n} w_j \log f_D(d_j;\mu_j) = \sum_{j=1}^{n} w_j\ \{d_j\log(\mu_j) - \mu_j - \log\ d_j!\}\,.$$

The link and the cumulant function are then derived $\theta_j = \log(\mu_j)$ and $b(\theta_j, m_j) = \mu_j$. The mean and variance functions are calculated as the first and second derivative with respect to θ_j.

$$b'(\theta_j) = \frac{\partial b}{\partial \mu_j}\frac{\partial \mu_j}{\partial \theta_j} = \mu_j \quad \text{and} \quad b''(\theta_i) = \frac{\partial^2 b}{\partial \mu_j^2}\left(\frac{\partial \mu_j}{\partial \theta_j}\right)^2 + \frac{\partial b}{\partial \mu_j}\frac{\partial^2 \mu_j}{\partial \theta^2} = \mu_j.$$

The dependence of μ_j on the covariate vector $\boldsymbol{X}$ is specified by the link function. Using the canonical link, i.e. the log link, we have

$$g(\mu_j) = \eta_j = \log(\mu_j) \quad \text{and} \quad \mu_j = g^{-1}(\mu_j) = \exp(\eta_j)$$

Since we want to maximize the log likelihood, we look for a solution of the set of normal equations to be fulfilled by the maximum likelihood parameter estimates $\boldsymbol{\beta}$. In case of locally quadratic fitting,

$$\frac{\partial L\big(\beta_v|\boldsymbol{y}, w_j, \phi\big)}{\partial \beta_v} = 0 \quad \text{for } v = 0, 1, \ldots, 5.$$

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.
The derivatives of the local Poisson log-likelihood function with respect to $\boldsymbol{\beta}$ are

$$\begin{aligned}
&\frac{\partial}{\partial \beta_0}L = \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j}\frac{\partial \mu_j}{\partial \eta_j}; && \frac{\partial}{\partial \beta_1}L = \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j}\frac{\partial \mu_j}{\partial \eta_j}(u_j - u_i);\\
&\frac{\partial}{\partial \beta_2}L = \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j}\frac{\partial \mu_j}{\partial \eta_j}(v_j - v_i); && \frac{\partial}{\partial \beta_3}L = \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j}\frac{\partial \mu_j}{\partial \eta_j}(u_j - u_i)^2;\\
&\frac{\partial}{\partial \beta_4}L = \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j}\frac{\partial \mu_j}{\partial \eta_j}(v_j - v_i)(u_j - u_i); && \frac{\partial}{\partial \beta_5}L = \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j}\frac{\partial \mu_j}{\partial \eta_j}(v_j - v_i)^2.
\end{aligned} \tag{3}$$

The derivative of the link function is calculated as

$$g'(\mu_j) = g' = \frac{\partial \eta_j}{\partial \mu_j} = \frac{\partial}{\partial \mu_j}\log(\mu_j) = \mu_j^{-1}.$$

The fisher information for $\boldsymbol{\beta}$ is given, in matrix notation, by

$$\mathcal{I}_{vk} = \Big\{\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{X}\Big\}_{vk},$$

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

$$\boldsymbol{X} = \begin{bmatrix}
1 & u_1 - u_i & v_1 - v_i & (u_1 - u_i)^2 & (u_1 - u_i)(v_1 - v_i) & (v_1 - v_i)^2\\
1 & u_2 - u_i & v_2 - v_i & (u_2 - u_i)^2 & (u_1 - u_i)(v_1 - v_i) & (v_1 - v_i)^2\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\
1 & u_n - u_i & v_n - v_i & (u_n - u_i)^2 & (u_1 - u_i)(v_1 - v_i) & (v_1 - v_i)^2
\end{bmatrix}$$

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

$$w_j = \begin{cases} W(\rho(x_i, x_j)/h) & \text{if } \rho(x_i, x_j)/h \leq 1,\\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

$W(.)$ denotes a non-negative weight function depending on the distance $\rho(x_i, x_j)$. A common choice is the Euclidean distance,

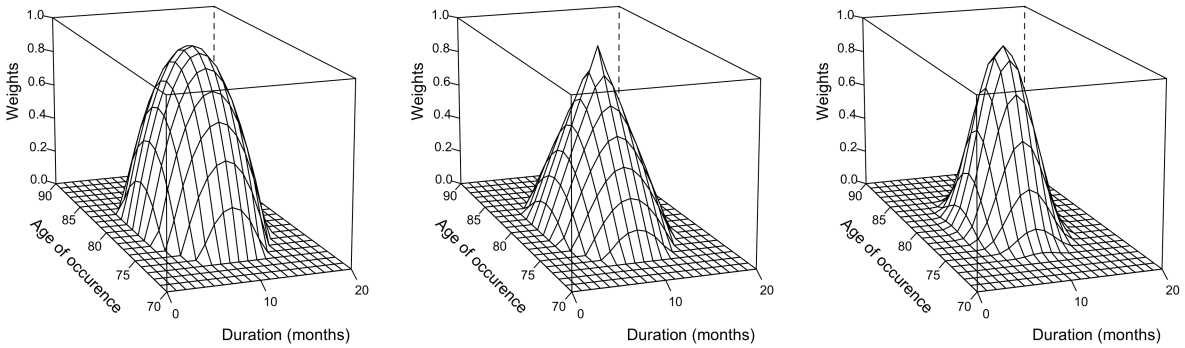
$$\rho(x_i, x_j) = \sqrt{(u_j - u_i)^2 + (v_j - v_i)^2}.$$

In addition, it contains a smoothing parameter $h = (\lambda - 1)/2$ which determines the radius of the neighborhood of x_i. The two components of the Euclidean distance can be scaled in order to apply more smoothing in one direction than the other. $W(.)$ is some weight function like those given in Table 2, below.

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

Table 2: Example of weight functions with $a = \rho(x_i, x_j)/h$.

Figure 1 displays some of the weight functions presented. For a weight function $W(a)$, the weights decrease with increasing distance $\rho(x_i, x_j)$. 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 and it is widely appreciated in the literature that a smooth weight function results in a smoother estimate, see Cleveland and Loader (1996, p.10-11).



(a) Epanechnikov (b) Triangular (c) Triweight

Figure 1: Weighting system shape of some weight functions, with, the radius, $h = 7$.

$$\omega_{jj} = \frac{1}{b''(\theta_j)} \left(\frac{\partial \mu_j}{\partial \eta_j} \right)^2, \tag{5}$$

depending on the variance and link function. Since $\eta_j = g(\mu_j)$, we have $\partial \eta_j / \partial \mu_j = g'(\mu_j)$, hence in using the canonical link, ω_{jj} reduces to μ_j.

Following the general Fisher scoring procedure, see Tomas (2011, p.8-10), we obtain the estimates. Given initial estimates $\widehat{\boldsymbol{\beta}}^*$, we may compute the vector $\widehat{\boldsymbol{M}}^*$ and $\widehat{\boldsymbol{\eta}}^*$. Using these values, we define the adjusted dependent variable $\boldsymbol{z}$ with components

$$z_j = \widehat{\eta}_j + (y_j - \widehat{\mu}_j)\ g'(\mu_j) = \widehat{\eta}_j + \frac{(d_j - \widehat{\mu}_j)}{\mu_j},$$

all quantities being computed at the initial estimate $\widehat{\boldsymbol{\eta}}^*$.
Maximum likelihood estimates satisfy the equations

$$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{z}, \tag{6}$$

which are solved iteratively. Hence, a maximum likelihood estimate of $\boldsymbol{\beta}$ is found by a localized version of the iteratively reweighted least squares (IRWLS) algorithm for GLMs. It is the following iterative process:

> Repeat $\boldsymbol{\beta}^* := \left(\boldsymbol{X}^T\boldsymbol{W}\ \boldsymbol{\Omega}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\ \boldsymbol{\Omega}\ \boldsymbol{z}$;
>
> using $\boldsymbol{\beta}^*$, update the working weights $\boldsymbol{\Omega}$, as well as the working dependent variables $\boldsymbol{z}$ until convergence.

Estimation of $\boldsymbol{\beta}$ is performed using a Fisher's scoring method search in each neighborhood, going in order as i runs from 1 to n.

When modeling experience data from life-insurance, we wish generally to take in account the exposure in the setting. Specifically, we are looking for a smooth estimate of the observed forces of mortality and from equation (1) the linear predictor η_j can be written as

$$\eta_j = \log\left(\mathbb{E}\big[Y|X = x_j\big]\right) = \log(\mu_j) = \log(E_j\varphi_j) = \log(E_j) + \log(\varphi_j)$$

The term E_j called the offset can be easily incorporated in the regression system (6).

3.2 *P*-splines framework for count data

As an alternative to local likelihood modeling, we compare the p-splines method. In this section, we present the essential background material on p-splines methodology for count data. Descriptions of the p-splines method can be found in the seminal paper of Eilers and Marx (1996), as well as in Marx and Eilers (1998), Eilers and Marx (2002), and in Currie and Durbán (2002). Currie *et al.* (2006) present a comprehensive study of the methodology. Applications covering mortality can be found in Currie *et al.* (2004), Richards *et al.* (2006), Kirkby and Currie (2010) and in the Ph.D. thesis of Camarda (2008). Planchet and Winter (2007) use the same framework to discuss an application concerning sick leave retentions.

Again, we suppose that the data can be arranged as a column vector, $\boldsymbol{y} = \text{vec}(\boldsymbol{Y}) = (y_1, y_2, \ldots, y_n)^T$. Let $\boldsymbol{B}_u = \boldsymbol{B}(\boldsymbol{u})$ and $\boldsymbol{B}_v = \boldsymbol{B}(\boldsymbol{v})$, be regression matrices, of dimensions $n_u \times k_u$ and $n_v \times k_v$, of B-splines based on the duration $\boldsymbol{u}$ and age of occurrence $\boldsymbol{v}$, respectively, with k denoting the number of internal knots.
Specifically, B-splines are bell-shaped curves composed of smoothly joint polynomial pieces. Polynomials of degree 3 are used in the following. The positions on the horizontal axis where the pieces come together are called knots. We use equally spaced knots. The numbers of columns of $\boldsymbol{B}_u$ and $\boldsymbol{B}_v$ are related to the number of knots chosen for the B-splines. Details on B-splines can be found in de Boor (2001).
The regression matrix for our two dimensional model is the Kronecker product

$$\boldsymbol{B} = \boldsymbol{B}_u \otimes \boldsymbol{B}_v.$$

The matrix $\boldsymbol{B}$ has an associated vector of regression coefficients $\boldsymbol{a}$ of length $k_u\ k_v$. As in the GLM framework, the linear predictors $\boldsymbol{\eta}$ is linked to the expectation of $\boldsymbol{y}$ by a link function $g(.)$.

$$\boldsymbol{\eta} = g\big(\mathbb{E}\big[\boldsymbol{y}\big]\big) = \log(\boldsymbol{\mu}) = \boldsymbol{B}\ \boldsymbol{a} = \big(\boldsymbol{B}_u \otimes \boldsymbol{B}_v\big)\ \boldsymbol{a}, \tag{7}$$

The elements of $\boldsymbol{a}$ can be arranged in a $k_u \times k_v$ matrix $\boldsymbol{A}$, where $\boldsymbol{a} = \text{vec}(\boldsymbol{A})$. The columns and rows of $\boldsymbol{A}$ are then given by $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_u)$ and $\boldsymbol{A}^T = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_v)$. Then instead of computing equation (7) as a vector, it can be written as

$$\log\big(\mathbb{E}\big[\boldsymbol{y}\big]\big) = \log\big(\boldsymbol{M}\big) = \boldsymbol{B}_u\ \boldsymbol{A}\ \boldsymbol{B}_v^T. \tag{8}$$

From the definition of the Kronecker product, the linear predictor of the columns of $\boldsymbol{Y}$ can be written as linear combinations of k_v smooths in the duration u. The linear predictors corresponding to the jth column of $\boldsymbol{Y}$ can be expressed as

$$\sum_{k=1}^{k_v} b_{jk}^v\ \boldsymbol{B}_u\ \boldsymbol{a}_k,$$

where $\boldsymbol{B}_v = b_{ij}^v$. We apply a roughness penalty to each of the columns of $\boldsymbol{A}$. The penalty is given by

$$\sum_{j=1}^{k_v} \boldsymbol{a}_j^T\ \boldsymbol{D}_u^T\ \boldsymbol{D}_u\ \boldsymbol{a}_j = \boldsymbol{a}^T \left(\boldsymbol{I}_{k_v} \otimes \boldsymbol{D}_u^T\ \boldsymbol{D}_u\right)\ \boldsymbol{a},$$

where $\boldsymbol{D}_u$ is the second order difference matrix acting on the columns of $\boldsymbol{A}$. Similarly by considering the linear predictor corresponding to the ith row of $\boldsymbol{Y}$,

$$\sum_{i=1}^{k_u} \boldsymbol{a}_i^T\ \boldsymbol{D}_v^T\ \boldsymbol{D}_v\ \boldsymbol{a}_i = \boldsymbol{a}^T \left(\boldsymbol{D}_v^T\ \boldsymbol{D}_v \otimes \boldsymbol{I}_{k_u}\right)\ \boldsymbol{a},$$

where $\boldsymbol{D}_v$ is the second order difference matrix acting on the rows of $\boldsymbol{A}$.
The penalized log-likelihood to be maximized can be written as

$$\ell^* = \ell\big(\boldsymbol{a}; \boldsymbol{B}, \boldsymbol{y}\big) - \frac{1}{2}\boldsymbol{a}^T\boldsymbol{P}\boldsymbol{a}. \tag{9}$$

where $\ell(\boldsymbol{a}; \boldsymbol{B}, \boldsymbol{y})$ is the usual log-likelihood for a GLM and the penalty term $\boldsymbol{P}$ is given by

$$\boldsymbol{P} = \lambda_u \left(\boldsymbol{I}_{k_v} \otimes \boldsymbol{D}_u^T\ \boldsymbol{D}_u\right) + \lambda_v \left(\boldsymbol{D}_v^T\ \boldsymbol{D}_v \otimes \boldsymbol{I}_{k_u}\right),$$

where λ_u and λ_v are the smoothing parameters used for the duration and the age of occurrence respectively, $\boldsymbol{I}_{k_u}$ and $\boldsymbol{I}_{k_v}$ being identity matrices of dimension k_u and k_v respectively. More details can be found in Currie *et al.* (2004). Then maximizing equation (9) gives the penalized likelihood equations

$$\boldsymbol{B}^T\big(\boldsymbol{y} - \boldsymbol{M}\big) = \boldsymbol{P}\boldsymbol{a},$$

which can be solved by a penalized version of the IRWLS algorithm,

$$\left(\boldsymbol{B}^T\boldsymbol{\Omega}\boldsymbol{B} + \boldsymbol{P}\right)\boldsymbol{a} = \boldsymbol{B}^T\boldsymbol{\Omega}\boldsymbol{z}, \tag{10}$$

where $\boldsymbol{\Omega}$ is the matrix of the working weights similar to (5). Again in case of Poisson errors, $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\mu})$. The working dependent variable $\boldsymbol{z}$ is defined by

$$\boldsymbol{z} = \boldsymbol{B}\ \boldsymbol{a} + \frac{\boldsymbol{y} - \boldsymbol{\mu}}{\boldsymbol{\mu}}.$$

Hence, a maximum likelihood estimate of $\boldsymbol{a}$ is found by a penalized version of IRWLS algorithm:

Repeat $\boldsymbol{a}^* := \boldsymbol{B}\big(\boldsymbol{B}^T\boldsymbol{\Omega}\,\boldsymbol{B}+\boldsymbol{P}\big)^{-1}\boldsymbol{B}^T\boldsymbol{\Omega}\,\boldsymbol{z}$;

using $\boldsymbol{a}^*$, update the working weights $\boldsymbol{\Omega}$, as well as the working dependent variables $\boldsymbol{z}$ until convergence.

Again when modeling mortality data, we may take into account the exposure in the setting. The linear predictor $\boldsymbol{\eta}$ can be written as

$$\boldsymbol{\eta} = g\big(\mathbb{E}\big[\boldsymbol{y}\big]\big) = \log(\boldsymbol{\mu}) = \log(\boldsymbol{e}) + \log(\boldsymbol{\varphi}) = \log(\boldsymbol{e}) + \boldsymbol{B}\,\boldsymbol{a} = \log(\boldsymbol{e}) + \big(\boldsymbol{B}_u \otimes \boldsymbol{B}_v\big)\,\boldsymbol{a},$$

where $\boldsymbol{e}$ denotes the vector of exposure. Similarly to Section 3.1, the offset can be easily incorporated in the regression system (10).

The penalized IRWLS would be efficient only in moderate-sized problems. For our application, the parameter vector $\boldsymbol{a}$ has length 2520 and this required the usage of 2520×2520 matrices. The size is moderate, but for larger dimensional matrices the penalized IRWLS algorithm can run into storage and computational difficulties. Currie *et al.* (2006) and Eilers *et al.* (2006) proposed an algorithm that takes advantage of the special structure of both the data as a rectangular array and the model matrix as a tensor product.

3.3 Effective dimension of a smoother

The effective dimension of the fitted model is an important concept in modeling. For linear models, this concept is clear and intuitive. The number of parameters used in the model determines its dimension. In non-parametric settings, a different definition is needed.

In linear models, the hat matrix, $\boldsymbol{H}$, is idempotent, $\mathrm{tr}\big(\boldsymbol{H}\;\boldsymbol{H}^T\big) = \mathrm{tr}\big(\boldsymbol{H}\big) = \mathrm{rank}\big(\boldsymbol{H}\big)$. Hence the crossproduct of the trace of the hat matrix is equal to the number of parameters in the fitted model. Given this feature of classic linear models, the trace of the hat matrix can be used to assess the fitted degrees of freedom and hence the effective dimension of a smoother.

Since the local likelihood estimate does not have an explicit representation, the hat matrix (or smooth weight diagram, named in this case) can not be derived as in the local regression case, Tomas (2011).

However, we can provide an estimation of the weight function associated with the i-th point at the last iteration after the convergence of the localized version of the IRWLS algorithm.
The weight function associated with the i-th point is used to compute the weights in the i-th row, $\boldsymbol{s}(x_i)^T$, of the $n \times n$ (where $n = 2520$ in our case) smooth weight diagram $\boldsymbol{S}$,

$$\boldsymbol{S} = \begin{bmatrix} s_1(x_1) & s_2(x_1) & \dots & s_n(x_1) \\ s_1(x_2) & s_2(x_2) & \dots & s_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ s_1(x_n) & s_2(x_n) & \dots & s_n(x_n) \end{bmatrix},$$

with rows

$$\boldsymbol{s}(x_i)^T = \big(s_1(x_i), s_2(x_i), \dots, s_n(x_i)\big) = \boldsymbol{e}_1^T\big(\boldsymbol{X}^T\boldsymbol{W}\,\boldsymbol{\Omega}\,\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\boldsymbol{W}\,\boldsymbol{\Omega}, \tag{11}$$

where $\boldsymbol{e}_1^T$ denotes a column of the same length of $\boldsymbol{\beta}$, having 1 at its first entry and all other entries equal to zero. For instance, in locally quadratic fitting, $\boldsymbol{e}_1^T$ is of length 5.

The same idea applies to the p-spline models, and the hat matrix is given by

$$\boldsymbol{H} = \boldsymbol{B}\big(\boldsymbol{B}^T\boldsymbol{\Omega}\,\boldsymbol{B} + \boldsymbol{P}\big)^{-1}\boldsymbol{B}^T\boldsymbol{\Omega} \tag{12}$$

The usefulness of the fitted degree of freedom is in providing a measure of the amount of smoothing that is comparable between different estimates applied to the same dataset. For local likelihood models, we define υ as

$$\upsilon = \mathrm{tr}\big(S\big) = \sum_{i=1}^{n} \widehat{\mathrm{infl}(x_i)}\;\omega_{ii}, \tag{13}$$

where $\mathrm{infl}(x_i) = \boldsymbol{e}_1^T\big(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{X}\big)^{-1}\boldsymbol{e}_1$ denotes the influence values and ω_{ii} is defined as in (5). The influence values measure the sensitivity of the fitted curve to the individual data points. The property of influence relates to the fact that as $\mathrm{infl}(x_i)$ approaches one, the corresponding residual approaches zero. Within the p-splines framework, the fitted degrees of freedom are the sum of the eigenvalues of $\boldsymbol{H}$, the trace of the hat matrix.
The fitted degrees of freedom aid interpretation. For instance, 10 fitted DF represents a smooth model with very little flexibility while 30 fitted DF represents a noisy model showing many features.

3.4 Parameters selection

One has to be keep in mind that graduation, and hence model selection issue, is a very effective compromise between two objectives, the elimination of irregularities and the achievement of a desired mathematical shape to the progression of the mortality rates. This underlines the importance of experience, and above all, of thorough investigation of data as the prerequisites of reliable judgment, as we must first inspect the data and take the decision as the type of irregularity we wish to retain. To quote Hickman and Miller (1977, p15), without prior information, smoothing is an unjustified process. Therefore, in practice, one needs to choose λ and the fitting variable to balance the trade-off between bias and variance.

To find the constellation of smoothing parameters, the strategy is to compute a number of candidate fits and use criteria to select, among the fits, the one with the lowest score. The predictor of a future observation at a point x_i is $g^{-1}(\widehat{\eta}(x_i))$ where $g(.)$ is the link function. One possible loss function is the deviance for a single observation (x_i, y_i), defined by

$$\begin{aligned} D(y_i, \widehat{\theta}(x_i)) &= 2\,\left(\sup_{\theta} l\big(y_i, \theta(y_i)\big) - l\big(y_i, \theta(\widehat{\mu}_i)\big)\right) \\ &= 2\,\left(y_i\big(\theta(y_i) - \theta(\widehat{\mu}_i)\big) - b\{\theta(y_i)\} + b\{\theta(\widehat{\mu}_i)\}\right). \end{aligned}$$

It is easily seen that $D\big(y_i, \theta(\widehat{\mu}_i)\big) \geq 0$, and $D\big(y_i, \theta(\widehat{\mu}_i)\big) = 0$ if $y_i = g^{-1}(\widehat{\eta}_i)$.

Since it is based on the likelihood, the deviance provides a measure of the evidence an observation y_i provides against $\widehat{\eta}(x_i)$ being the true value of $\eta(x_i)$.
The total deviance is defined as

$$\sum_{i=1}^{n} D\big(y_i, \widehat{\theta}(x_i)\big), \tag{14}$$

where the deviance for our Poisson case is given by

$$D\big(y_i, \widehat{\theta}(x_i)\big) = 2\,\big(y_i \log(y_i/\widehat{\mu}_i) - (y_i - \widehat{\mu}_i)\big).$$

When the response is zero, the individual deviance reduces to $D\big(y_i, \widehat{\theta}(x_i)\big) = 2\,\mu_i$.

It leads to a generalization of the Akaike information criterion (AIC) based directly on the deviance function

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

where υ is the fitted degrees of freedom, defined in (13).

Alternatively, for p-splines methods, we use the Bayesian information criterion (BIC) which penalizes more heavily the model complexity particularly when n is large,

$$BIC = \sum_{i=1}^{n} D\big(y_i, \theta(\widehat{\mu}_i)\big) + \log(n)\,\upsilon.$$

However as Cleveland and Devlin (1988) strongly argue, it is important to note that exclusive confidence in practice on a global criterion is unwise because a global criterion does not provide information about where the contribution to bias and variance are coming from the design space.

In conjunction one always has to look at residual plots. Residual analysis and goodness of fit diagnostics are just as important for non-parametric procedures as they are for parametric models. In the case of generalized linear models, we denote:

i. The response residual: $r_i = y_i - \widehat{\mu}_i$;

ii. The Pearson residual: $r_i = (y_i - \widehat{\mu}_i) \,/\, \sqrt{\mathbb{V}\mathrm{ar}\big[\widehat{\mu}_i\big]}$;

iii. The deviance residual: $r_i = \mathrm{sign}(y_i - \widehat{\mu}_i) D\big(y_i, \theta(\widehat{\mu}_i)\big)^{1/2}$.

Such residual plots provide a powerful diagnostic that nicely complements the selection criteria. The object is to determine whether large residuals correspond to features in the data that have been inadequately modeled.
The purpose of the plots can be related to the bias-variance traded-off. Plotting only the fits gives an one-sided view of the bias-variance trade-off, seeing the variance but not the bias.
Therefore, the diagnostic plots can show lack of fit locally and we have the opportunity to judge the lack of fit based on our knowledge of both the mechanism generating the data and of the performance of the smoothers used in the fitting.

In consequence, if a model correctly models a dataset, no strong patterns should appear in the response and Pearson residuals. Moreover, if the deviance residuals exhibit several successive residuals have the same sign, this indicates that the data are over-smoothed.

3.5 Confidence intervals

The confidence intervals should ideally take into account the underlying family of distribution. However, the theory for deriving such intervals seems quite intractable following Loader (1999, p.171). Therefore, we must rely on methods based on normal assumption, using the approximate variance. The local maximum likelihood estimator is usually asymptotically normal. This has been shown by Fan *et al.* (1995, p.143-145) in the context of generalized linear models

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \to N\big(0, \mathbb{V}\mathrm{ar}\big[\widehat{\boldsymbol{\beta}}\big]^{1/2}\big).$$

In case of local likelihood models, the variance approximation of the linear predictor reduces to

$$\mathbb{V}\mathrm{ar}\big[\widehat{\boldsymbol{\eta}}\big] \approx \boldsymbol{\Omega}^{-1}\boldsymbol{S}\boldsymbol{S}^T,$$

where $\boldsymbol{S}$ is the smooth weight diagram as in equation (11). Hence for a single observations, we have the following compact forms,

$$\begin{aligned}\mathbb{V}\mathrm{ar}\big[\widehat{\eta}(x_i)\big] &= (\omega_{ii})^{-1}\sum_{j=1}^{n} s_j^2(x_i) \\ &= b''(\theta_i)\big(g'(\widehat{\mu}_i)\big)^2\|\boldsymbol{s}(x_i)\|^2.\end{aligned}$$

With the use of the canonical link and by the delta method, the variance approximation of μ_i for local likelihood models is given by

$$\begin{aligned}\mathbb{V}\mathrm{ar}\big[\widehat{\mu}_i\big] &= \mathbb{V}\mathrm{ar}\big[g^{-1}\big(\widehat{\eta}(x_i)\big)\big] \\ &\approx \left(\frac{\partial}{\partial\eta_i}g^{-1}\big(\widehat{\eta}(x_i)\big)\right)^2 \omega_{ii}^{-1}\|\boldsymbol{s}(x_i)\|^2 \\ &= \widehat{\mu}_i\|\boldsymbol{s}(x_i)\|^2.\end{aligned}$$

The unknown function $\mu(x_i)$ falls in the random interval with approximately $(1-\alpha)$ coverage probability,

$$\widehat{\mu}(x_i) \pm c\,\widehat{\mu}_i^{1/2}\|\boldsymbol{s}(x_i)\|,$$

and a confidence interval for the forces of mortality is given by

$$\widehat{\psi}(x_i) \pm c\,\widehat{\psi}_i^{1/2}E_i^{-1}\|\boldsymbol{s}(x_i)\|,$$

where c is chosen as the $(1-\alpha/2)$ quantile of the standard normal distribution. See Tomas (2011) for the complete derivation.
For p-spline models, the variance approximation of the linear predictor is given by

$$\mathbb{V}\mathrm{ar}\big[\widehat{\boldsymbol{\eta}}\big] \approx \boldsymbol{B}\big(\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}+\boldsymbol{P}\big)^{-1}\boldsymbol{B}^T.$$

Similarly than the local likelihood model, we obtain the pointwise confidence intervals with approximately $1-\alpha$ coverage probability.

4 Adaptive local likelihood Methods

This section presents the adaptive methods and covers model selection issues. We treat the choices of bandwidth, polynomial degree and weight function as modeling the data and choose the constellation of smoothing parameters to balance the trade-off between bias and variance.

We distinguish a locally adaptive pointwise smoothing method using the intersection of confidence intervals rule and a global method using local bandwidth correction factors. The latest is an extension of the adaptive kernel method proposed by Gavin *et al.* (1995) to likelihoods techniques. We vary the amount of smoothing in a location dependent manner and allow adjustments based on the reliability of the data.

It is well known that of the smoothing parameters, the weight function has much less influence on the bias and variance trade-off than the bandwidth or the order of approximation. The choice is not too crucial, at best it changes the visual quality of the regression curve.

For convenience, we use the Epanechnikov weight function in expression (4) throughout this article, as it is computationally cheaper to use a truncated kernel. Moreover, it has been shown that the Epanechnikov kernel is optimal in minimizing the mean squared errors for local polynomial regression, see Fan *et al.* (1997). The biweight and triweight kernel, which behave very similarly, could have also been chosen. The choice remains subjective.

4.1 Intersection of confidence intervals

The intersection of confidence intervals was introduced by Goldenshulger and Nemirovski (1997) and further developed by Katkovnik (1999). Application of the ICI rule in case of Poisson local likelihood for adaptive scale image restoration has been studied in Katkovnik *et al.* (2005). Chichignoud (2010) in his Ph.D. thesis presents a comprehensive illustration of the method. The intersection of confidence intervals (ICI) provides an alternative method of assessing local goodness of fit.

We start by defining a finite set of window sizes

$$\Lambda = \{\lambda_1 < \lambda_2 < \ldots < \lambda_K\},$$

and determines the optimal bandwidth by evaluating the fitting results.

Let $\widehat{\psi}(x_i, \lambda_k)$ be the estimate at x_i for the window λ_k. To select the optimal bandwidth, the ICI rule examines a sequence of confidence intervals of the estimates $\widehat{\psi}(x_i, \lambda_k)$:

$$\begin{aligned}
\widehat{I}(x_i, \lambda_k) &= \left[\widehat{L}(x_i, \lambda_k), \widehat{U}(x_i, \lambda_k)\right], \\
\widehat{U}(x_i, \lambda_k) &= \widehat{\psi}(x_i, \lambda_k) + c\,\widehat{\sigma}(x_i)\|\boldsymbol{s}(x_i, \lambda_k)\|, \\
\widehat{L}(x_i, \lambda_k) &= \widehat{\psi}(x_i, \lambda_k) - c\,\widehat{\sigma}(x_i)\|\boldsymbol{s}(x_i, \lambda_k)\|,
\end{aligned}$$

where c is a threshold parameter of the confidence interval. Then, from the confidence intervals, we define

$$\begin{aligned}
&\overline{\widehat{L}}(x_i, \lambda_k) = \max\left[\widehat{L}(x_i, \lambda_{k-1}), \widehat{L}(x_i, \lambda_k)\right], \\
&\underline{\widehat{U}}(x_i, \lambda_k) = \min\left[\widehat{U}(x_i, \lambda_{k-1}), \widehat{U}(x_i, \lambda_k)\right], \\
&k = 1, 2, \ldots, K \quad \text{and} \quad \overline{\widehat{L}}(x_i, \lambda_0) = \underline{\widehat{U}}(x_i, \lambda_0) = 0.
\end{aligned}$$

The largest value for these k for which $\underline{\widehat{U}}(x_i, \lambda_k) \geq \overline{\widehat{L}}(x_i, \lambda_k)$ gives k^*, and it yields a bandwidth λ_k^*, that is the required optimal ICI bandwidth.

In other words, denoting $\mathcal{I}_j = \bigcap_{j=k}^{K} \widehat{I}(x_i, \lambda_j)$ for $k = 1, 2, \ldots, K$, we choose k^* such that

$$\begin{cases}
\mathcal{I}_j \neq \emptyset, & \forall j \geq k^*, \\
\mathcal{I}_{k^*-1} = \emptyset.
\end{cases}$$

As the bandwidth λ_k is increased, the standard deviation of $\widehat{\psi}(x_i, \lambda_k)$, and hence $\|\boldsymbol{s}(x_i, \lambda_k)\|$, decreases. The confidence intervals become narrower. If λ_k is increased too far, the estimate $\widehat{\psi}(x_i, \lambda_{k)}$ will become heavily biased, and the confidence intervals will become inconsistent in the sense that the intervals constructed at different bandwidths have no common intersection. The optimal bandwidth λ_{k^*} is the largest k when $\underline{\widehat{U}}(x_i, \lambda_k) \geq \overline{\widehat{L}}(x_i, \lambda_k)$ is still satisfied, i.e. when $\mathcal{I}_j \neq \emptyset$.

Because the optimal bandwidth is decided by c, this parameter plays a crucial part in the performance of the algorithm. When c is large, the segment $\widehat{I}(x_i, \lambda_k)$ becomes wide and it leads to a larger value of λ_k^*. This results in over-smoothing. On the contrary, when c is small, the segment $\widehat{I}(x_i, \lambda_k)$ would become narrow and it leads to a small value of λ_k^* so that the volatility can not be removed effectively. In theory, we could apply the criteria presented in Section 3.4 to determine a reasonable value c. However, because of practical constraints, the choice of c is done subjectively.

4.2 Local bandwidth factor methods

Instead of having a pointwise procedure, other types of adaptive approaches could be performed by using a global criterion. We could incorporate additional information into a global procedure by allowing the bandwidth to vary according to the reliability of the data, such as the variable kernel estimator proposed in Gavin *et al.* (1995, pp.190-193) for local polynomials. We present here the background of their approach which is extended in this article to likelihood techniques.

We can calculate a different bandwidth for each age at which the curve has to be estimated. The local bandwidth at each age is simply the global bandwidth multiplied by a local bandwidth factor to allow explicit dependence on this information. As we already obtained the local bandwidth factors, the process of using a global criterion decides the global value at which the bandwidth curve is located.

The aim is to allow the bandwidth to vary according to the reliability of the data, and to take into account the nature of the risk considered. The local bandwidth factors could depend on the exposure or the number of deaths per attained age, in case of annuities and death benefits, respectively.
For regions in which the exposure is large, a low value for the bandwidth results in an estimate that more closely reflects the crude rates. On the other hand, for regions in which the exposure is small, such as long duration, a higher value for the bandwidth allows the estimate of the true forces of mortality to progress more smoothly. This means that for long duration we are calculating local averages over a greater number of observations, which reduces the variance of the graduated rates but at the cost of a potentially higher bias.

The local bandwidth at each age is the global bandwidth multiplied by a local bandwidth factor, $h_i = h \times \delta_i^s$ for $i = 1, \ldots, n$. The variation in exposure or in deaths within a dataset can be enormous. To dampen the effect of this variation we choose

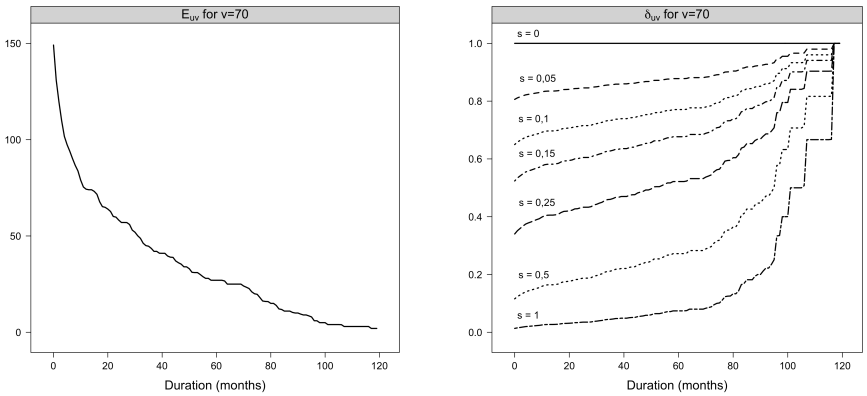
$$\delta_i^s \propto \widehat{\xi}_i^{-s}, \qquad \text{for } i = 1, \ldots, n \quad \text{and} \quad 0 \leq s \leq 1, \tag{15}$$

where s is a sensitivity parameter and

$$\widehat{\xi}_i = \begin{cases} E_i / \sum_{j=1}^n E_j & \text{for } i = 1, \ldots, n \quad \text{in case of annuities,} \\ D_i / \sum_{j=1}^n D_j & \text{for } i = 1, \ldots, n \quad \text{in case of death benefits.} \end{cases} \tag{16}$$

Choosing $s = 0$ reduces both models to the fixed parameter case, while $s = 1$ may result in very large smoothness variation depending on the particular dataset.
We choose the reciprocal of $\max\{\xi_i^{-s};\ i = 1, \ldots, n\}$ as the constant of proportionality in (15), so that $0 < \delta_i^s \leq 1$, for $i = 1, \ldots, n$. The observed exposure, or the observed deaths, decides the shape of the local bandwidth factor but the sensitivity parameter s determines the magnification of that shape, becoming more pronounced as s tends to 1.

Figure 2a shows the exposure for the age of occurrence 70 and Figure 2b displays the resulting smoothness tuning parameter for values of the sensitivity parameter of $0, 0.05, 0.1, 0.15, 0.25, 0.5$ and 1.

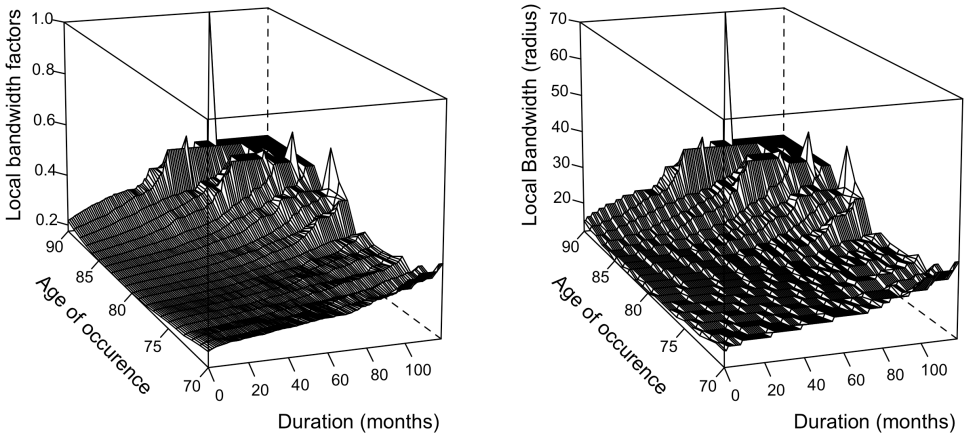


(a) Exposure $E_{u,70}$ (b) Local bandwidth factors $\delta_{u,70}$

Figure 2: $E_{u,70}$, and values of $\delta_{u,70}$, for various sensitivity parameters.

For $s = 0.15$, the minimum smoothness tuning parameter is about 0.5, at duration 0. This means that the bandwidth at the longest duration is about twice that at the shortest duration.

Figure 3 presents the value of $\delta_{u,v}$ for $s = 0.15$ and local bandwidth values (radius) derived. If there is a small exposure, then $\delta^s_{u,v}$ is large. It increases the smoothness tuning parameter and allows to apply more smoothing. The other way around if the amount of exposure is large.



(a) Local bandwidth factors $\delta_{u,v}$ (b) Local bandwidths (radius)

Figure 3: $\delta_{u,v}$ for $s = 0.15$ and the resulting local bandwidths.

Similarly to the local likelihood approach, we can apply the criteria presented in Section 3.4 to select the optimal constellation of smoothing parameters. As we already obtained the shape and the magnification of the local bandwidth factors, this process decides the global value at which the bandwidth curve is located.

5 Applications

The computations are carried out with the help of the software R, R Development Core Team (2013). The scripts are available on request.

5.1 The data

The data come from a portfolio of LTC-heavy-claimants of a French insurance company. We are concerned about the construction of the survival distribution.

The period of observation contains 10 years. The data are composed of a mixture of pathologies. The pathologies are composed, among others, by dementia, neurological illness and terminal cancer. The data consist for 2/3 of women and 1/3 of men. We have no exogenous information about the LTC-claimants. We observe only the aggregated exposition and number of deaths over two dimensions. These are the age of occurrence v of the pathologies and the duration of the care u. The maximum duration of the pathologies is 119 months. Figures 4a, 4b, and 4c display the observed statistics of the dataset.

Moreover, we have at our disposal a benchmark obtained from an internal document Planchet (2012), Figure 4d. It gives an idea about the desirable shape that we aim to retain as it reasonable approximation of the *true* law. This robust reference has been constructed through backtesting and we can refer to this law in assessing the quality of the adjustment and the innovation obtained from the models.

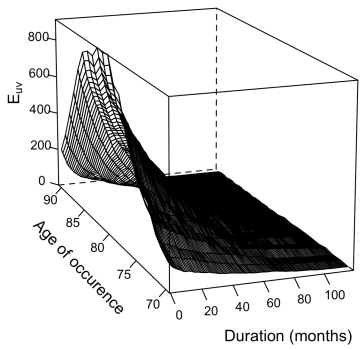
5.2 Smoothed surfaces and fits

Figure 4e presents the smoothed surface obtained with the local likelihood model with an Epanechnikov weight function, a polynomial of degree 2 and a bandwidth (radius) of 13 observations. The corresponding degrees of freedom υ are 29.25. The order of polynomial and the bandwidth have been chosen by minimizing the AIC criterion. The surface is relatively wiggly showing an inappropriate variance.

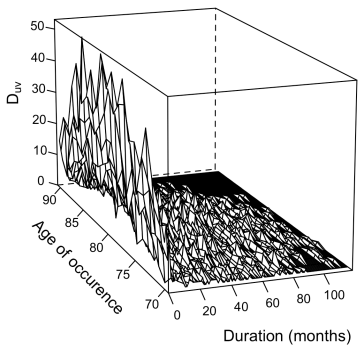
Figure 4f displays the smoothed surface obtained when fitting p-splines. The smoothing parameters $\lambda_u = 31.6$, $\lambda_v = 31.6$, have been chosen by minimizing the BIC criterion. It leads to $k_u = 24$, $k_v = 4$ for $\upsilon = 18.11$. The surface seems satisfactory, though the increase in the upper right corner (highest age of occurrence and longest duration) is not present as in the surface adjusted from Planchet (2012).

Figures 4g and 4h present the smoothed surface obtained with the adaptive local likelihood methods. For these applications, only the bandwidth is varying. The order of polynomial is still fixed at 2 and we use an Epanechnikov weight function. The fitted degrees of freedom υ are 10.76 and 16.16 respectively.

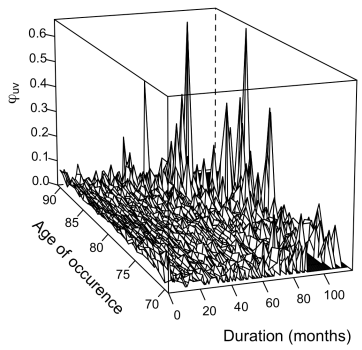
In general, only for the first months of the duration, the graduations are similar. After that, we obtain very different shapes according to the models.
The ICI rule and the local bandwidth factors seem the most satisfying methods in modeling the monotone phenomenon at the extreme ages, Figures 4g and 4h. The fitted degrees of freedom for the local bandwidth factors are larger than the ones obtained by the ICI rule indicating that the model is slightly more flexible and shows more features. The bandwidth values depend on the amount of exposure to represent effectively the remaining life expectancy in the regions where the amount of exposure is high. The corresponding bandwidths, in the left region, are relatively low, and they increase as the amount of exposure decreases. For regions in which the amount of exposure is low, a large value for the bandwidth results in an estimate that progress more smoothly.
As we already obtained the shape and the magnification of the local bandwidth factors, we used the AIC criterion to decide the global value at which the bandwidth curve is located. The sensitivity parameter s



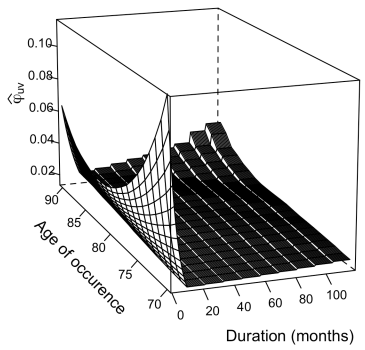
(a) Number of exposures to the risk, $E_{u,v}$



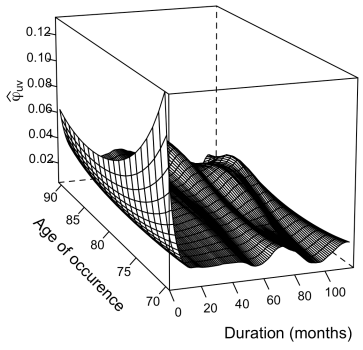
(b) Number of death, $D_{u,v}$



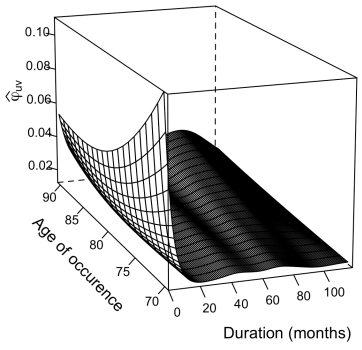
(c) Crude forces of mortality, $\varphi_{u,v}$



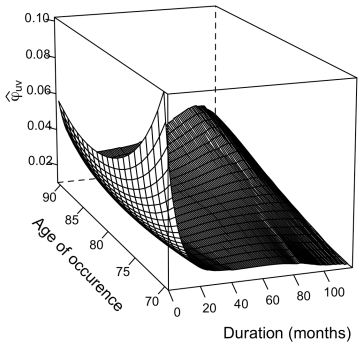
(d) $\widehat{\varphi}_{u,v}$, Planchet (2012)



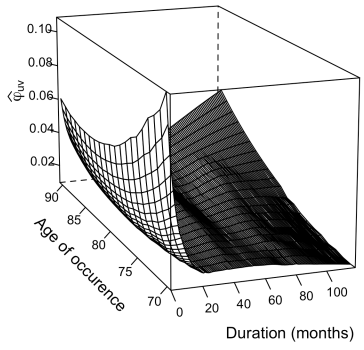
(e) $\widehat{\varphi}_{u,v}$, local likelihood



(f) $\widehat{\varphi}_{u,v}$, p-splines



(g) $\widehat{\varphi}_{u,v}$, ICI



(h) $\widehat{\varphi}_{u,v}$, local bandwidth factors

Figure 4: Observed statistics: $E_{u,v}$, $D_{u,v}$, $\varphi_{u,v}$ and smoothed forces of mortality $\widehat{\varphi}_{u,v}$ according to Planchet (2012), local likelihood, p-splines, ICI rule and local bandwidth factors methods

for the local bandwidth factors as well as the value c for the ICI rule have been chosen arbitrarily to be $0,15$ and 0.1 respectively. For higher value of s spurious features started to appear showing unacceptable variance, while for higher c, bias tends to show up.

Figures 5 and 6 present the smooth fits obtained from the different models for various ages of occurrence and durations.

The approaches produce relatively similar graduations for regions having a large exposition, Figure 5b. In contrast, the illustrations suggest that a significant model risk affects the results where the exposition is limited. The low exposition should favor the use of parametric models relying on some strong expert arguments similarly to the methodology used to obtain the surface in Planchet (2012). We confront this benchmark to the fits obtain by more rigorous methods. The fit obtained from global local likelihood, and not as strongly the p-splines, present an unacceptably high variance. It shows the inapplicability to model such datasets with global methods or to select the smoothing parameters by relying explicitly on a criterion. The local bandwidth factors method has the capability to model the forces of mortality in the first months of duration relatively well, Figure 6a, and the sharp increase at the highest extremes of the age of occurrence and duration, Figures 5c and 6c. The ICI rule and p-splines fail to model these features. Hence, for regions having a low exposition, the benefits of the adaptive local likelihood technique with local bandwidth factors become apparent. The approach has the advantage to be more rigorous compared to parametric methods relying on expert opinions.

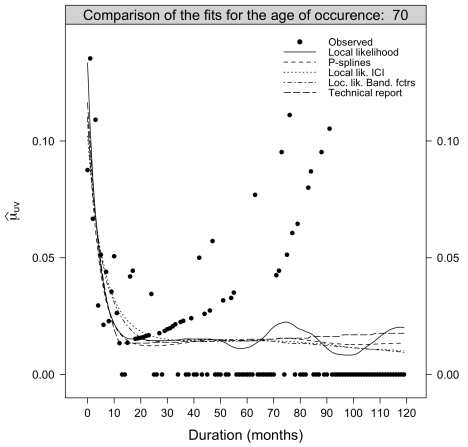
5.3 Analysis of the residuals

Figure 7 presents the residuals of the 5 models for the age of occurrence 70 as well as the ones obtained from the adjusted surface from Planchet (2012).

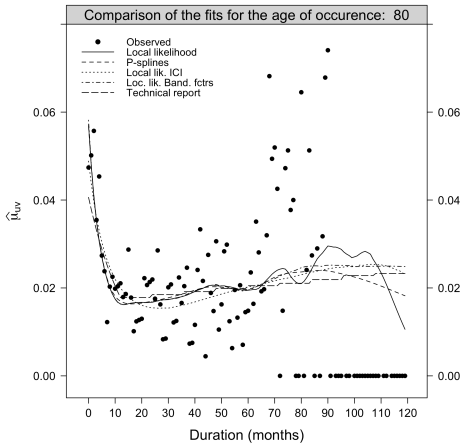
The pattern of the residuals displayed for each model is roughly similar. We superimposed a *loess* smooth curve on the response and Pearson residuals. These smooths help search for clusters of residuals that may indicate lack of fit. By reducing the noise, our attention may be more readily drawn to features that have been missed or not properly modeled by the smooth. Here the process is not to judge a fit adequate if a smooth curve on its residuals plot is flat. A flat curve means simply that no systematic, reproducible lack of fit has been detected. The fit may well be too noisy, and stays to close to an interpolation since trends in small parts of the data are interpreted as more widespread trends. Then for small datasets, the fit is very nearly interpolating the data which results in unacceptably high variance. Strong patterns appear in the response residuals in Figure 7. It indicates a lack of fit in this region. However, this is not surprising as most of the deaths at the longest durations are zero for the age of occurrence 70.

The Pearson residuals are mainly in the interval $[-2, 2]$, which indicates that the models adequately capture the variability of the dataset.

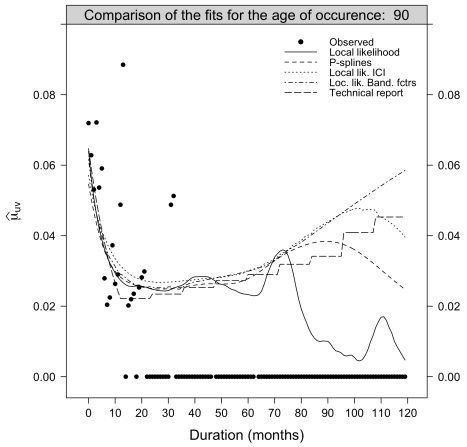
The deviance residuals present, for the longest durations, several successive residuals having the same sign. It illustrates that the forces of mortality are over-smoothed locally. As the sign is negative, from 80 to 119 months, we strongly overestimate the forces of mortality. However, we would have excepted such a pattern as we observe zero deaths at the highest extreme of the duration of the care.



(a) Age of occurrence $v = 70$

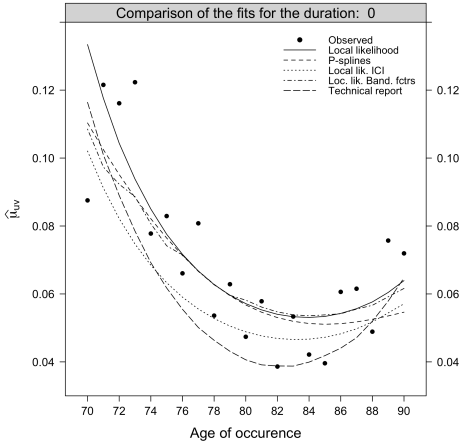


(b) Age of occurrence $v = 80$

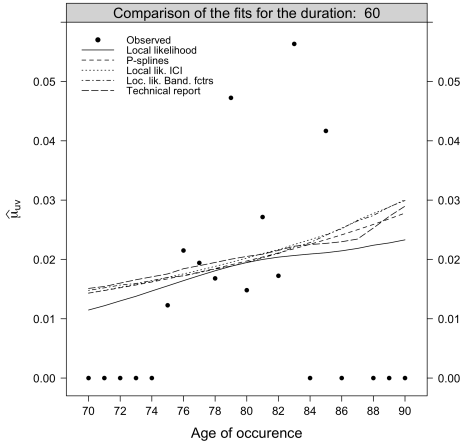


(c) Age of occurrence $v = 90$

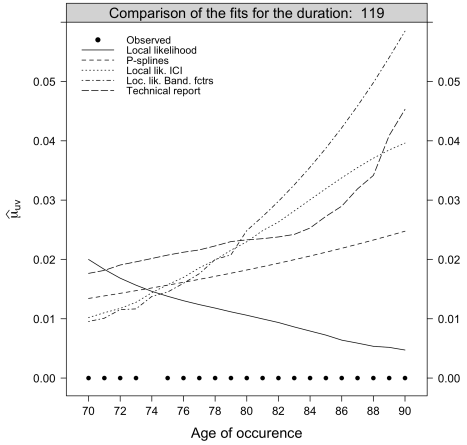
Figure 5: Observed forces of mortality and smooth fits for various ages of occurrence.



(a) Duration $u = 0$



(b) Duration $u = 60$



(c) Duration $u = 119$

Figure 6: Observed forces of mortality and smooth fits for various durations.

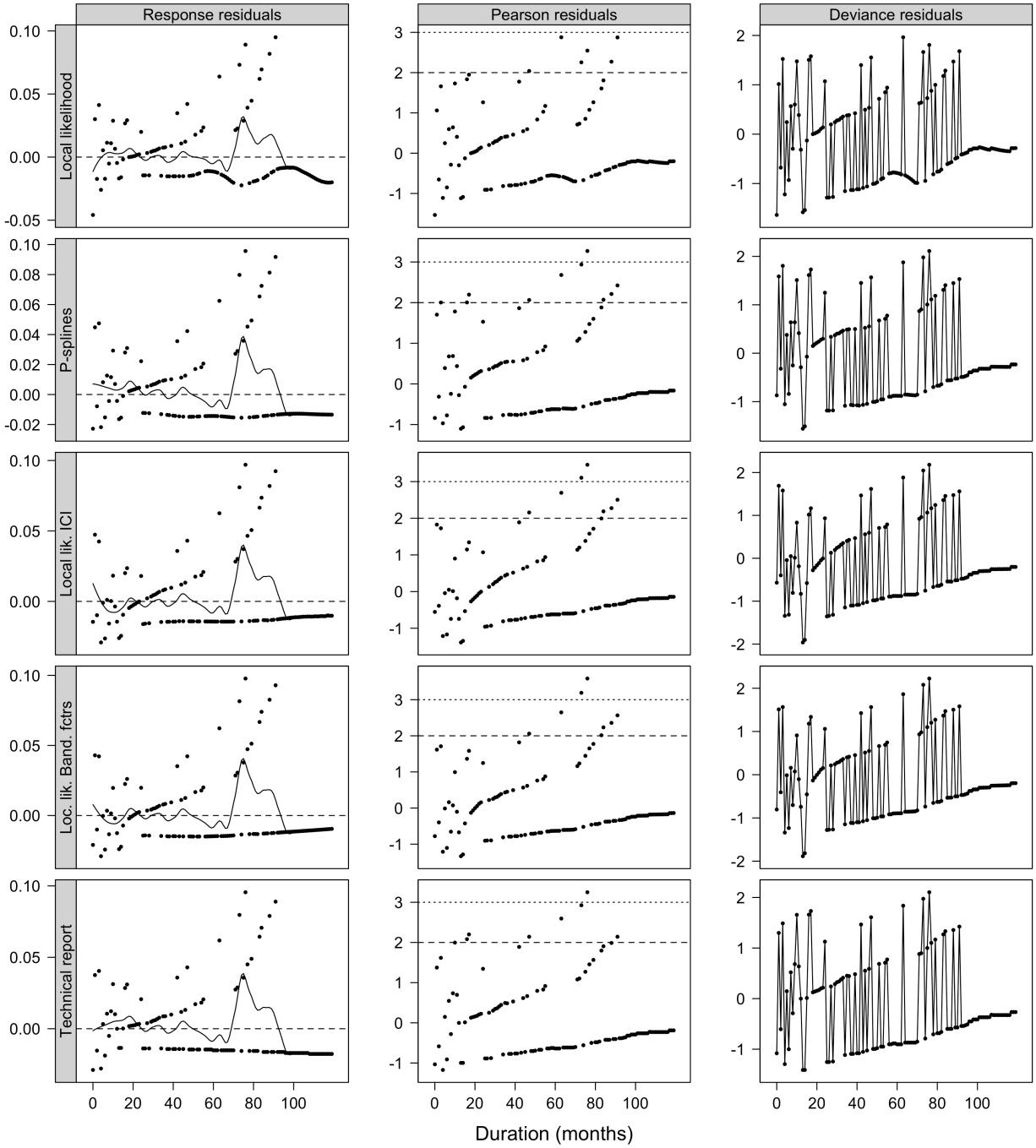


Figure 7: Response, Pearson and deviance residuals for the age of occurrence 70

6 Comparisons

6.1 Tests to compare graduations

We continue the comparisons by applying the tests proposed by Forfar *et al.* (1988, p.56-58) and Debón *et al.* (2006, p.231). We have also obtained the values of the mean absolute percentage error $MAPE$ and R^2 used in Felipe *et al.* (2002). We compare the crude mortality rates to the graduated series to see whether the approaches lead to similar graduation. Table 3 presents the results.

		Local lik.	p-splines	Adapt. lik. ICI	Adapt. band. factors	Planchet (2012)
Fitted DF v		29.25	18.11	10.76	16.16	NA
Deviance		2259.91	2291.89	2440.81	2311.04	2409.99
Standardised sesiduals	> 2	108	117	127	115	129
	> 3	25	31	38	32	42
Signs test	+(−)	910(1610)	907(1613)	891(1629)	900(1620)	908(1612)
	p-value	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
Runs test	Nb of runs	1028	1016	971	1019	1013
	Value	-6.25	-6.71	-8.19	-6.47	-6.64
	p-value	$4.05e-10$	$1.98e-11$	$2.57e-16$	$9.69e-11$	$3.10e-11$
χ^2		2433.56	2487.09	2616.08	2458.71	2644.09
R^2		0.2406	0.2340	0.2433	0.2476	0.2234
$MAPE$ (%)		46.11	46.62	48.18	46.99	47.38

Table 3: Comparisons between the smoothing approaches.

The approaches display different results. The global local likelihood approach, having the highest degrees of freedom, has the capacity to show many features in the data. Therefore, the values of the various tests are the *best*. It has the lowest deviance, lowest number of standardized residuals exceeding the thresholds 2 and 3, highest number of runs, best mix of the residuals between positive and negative signs, highest value for the run test. In addition, the approach results in the minimum χ^2 and $MAPE$. Conversely, the adaptive local likelihood using the ICI rule yields the smallest degrees of freedom. As a consequence, it leads to the highest deviance and $MAPE$, lowest mix of the residuals between positive and negative signs, lowest value for the run test. We observe that the results obtained with the benchmark Planchet (2012) are relatively bad. It has the highest number of standardized residuals exceeding the thresholds 2 and 3, highest χ^2 and MAPE, and lowest R2.

The results for the p-splines and the adaptive approach using the local bandwidth factors are similar, even though we have seen that the adaptive method has a better ability to model the mortality patterns (high mortality for the first month of duration of the care and increase at the extreme highest of the duration and age of occurrence), while the p-splines model has higher degrees of freedom. Hence the adaptive method using the local bandwidth factors approach would be privileged.

The tests and quantities carried out in Table 3 show the strengths and weaknesses of each model to adjust the observed mortality. It is up to potential users of the table to decide the weights they place on the different criteria. However, regarding the wide ranging set of model selection criteria, we can conclude about the relevancy of the models in adjusting the observed mortality. Because of the appropriate data-driven choice of the adaptive smoothing parameters, the adaptive local likelihood models adapt neatly to the complexity of the mortality surface and are the closest to the benchmark, Planchet (2012). The adaptive local likelihood method using the bandwidth factors models well the high mortality during the first months of duration and the increase at the extreme high duration and age of occurrence compared to the other methods. Having 13 degrees of freedom less than the local likelihood model, the adaptive bandwidth factors model is then less flexible however the tests presented in Table 3 show relatively good results.

6.2 Comparing single indices summarizing the lifetime probability distribution

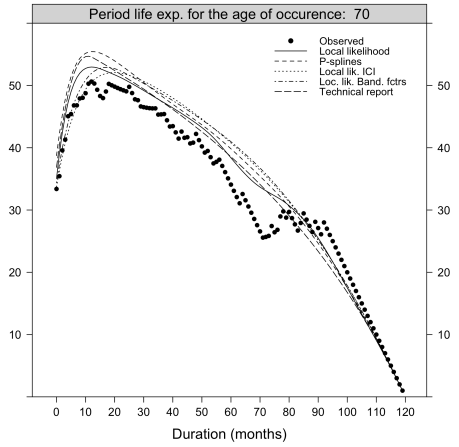
We end these comparisons by presenting some figures summarizing the lifetime probability distribution. Figures 8 and 9 display the life expectancy obtained from the different models for various ages of occurrence and durations.

At age of occurrence 70, with the exception of the adaptive local likelihood using the ICI rule and local bandwidth factors, the models are over-estimating the period life expectancy for the first months of duration (until 10 months), Figure 8a. This is particularly visible for the p-splines and the adjusted surface obtained in Planchet (2012). The over-estimation is general at age of occurrence 80, Figure 8b. The shapes of the life expectancy differ much at age of occurrence 90, Figure 8c, where the global local likelihood tends to estimate a more rectangular shape.

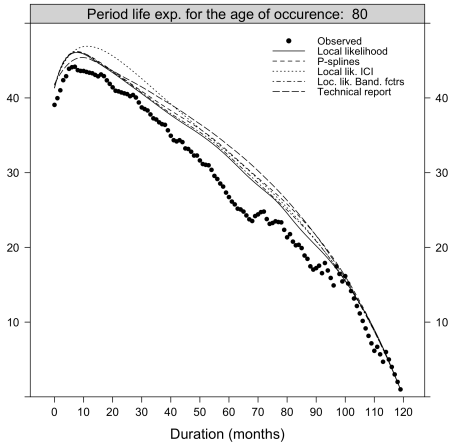
The shape and trend of the life expectancies are similar when we observe a large amount of exposure (first months of duration of the care), Figure 9a. The high correlation of the pathologies with the age of occurrence can explained the concave shape observed for the life expectancies during the first months of the care. The lowest ages of occurrence are marked by a relatively high mortality mainly due to the death of the individuals suffering from terminal cancer, while the highest ages concern principally the dementia. At the 60th month of the care, the life expectancy is decreasing rapidly, Figure 9b. However, while the ICI rule and local bandwidth factors produce similar patterns, the shapes and trends given by the other models diverge markedly, the local likelihood predicting a rise of the life expectancy for the highest age of occurrence. This pattern is also present, although less markedly, in Figure 9c.

Figure 10 shows the median month at death, Figure 10a, standard deviation of the random life time, Figure 10b and entropy, Figure 10c, as a function of the age of occurrence of the claim.
In Figure, 10a displaying the median month at death as a function of the age of occurrence of the claim, we observe a concave shape similar to Figure 9a. This phenomenon shows, once more, the correlation between the age of occurrence and the pathologies. The adaptive local likelihood using the ICI rule, having the lowest degrees of freedom, mostly under-estimates the median month at death compared to the others models.
After a steady increase, the standard deviation of the random lifetime is slowing down at age of occurrence 82, and decreases until 90 years old, Figure 10b. It is explained by the fact that we observe most of the deaths at the lowest age of occurrence and duration, while the number of deaths is zero, and thus stable, at the highest age and duration.
Figure 10c shows the entropy. The values decline as the deaths become more concentrated. We observe that the deaths predicted by the adaptive local likelihood models (ICI rule and bandwidth factors) are the most stretched. Conversely, the adjusted number of deaths obtained by Planchet (2012) are more concentrated.

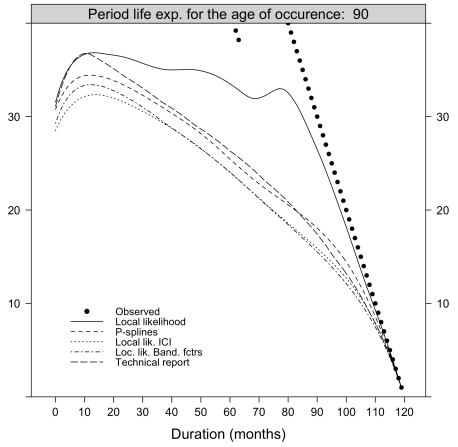
Table 4 summarizes the indices. For the period life expectancy, e_{70}, e_{80}, and e_{90}, the observations suggest an increase with the age, which, based on our knowledge, is unrealistic. We are more likely to look for a concave shape, predicted by the models as displayed in Figure 9a. On average, the models agree on the same period life expectancy, around 38 months, and under-estimate the observed one.
The median month at death, $\text{Med}(T_0)$, estimated by the models varies in average slightly from 25 to 27 months. However, for a particular age of occurrence, such as $\text{Med}(T_0(70))$, the difference between the models (p-splines and ICI rule) can grow until 6 months.
All the models sensibly estimated the same average standard deviation of the random life time, σ_0, which corresponds to the observed standard deviation, around 0.22.
Finally, all the models agree on the estimated average entropy $H(T_0)$, between 0.035 to 0.037. The entropy



(a) Age of occurrence $v = 70$

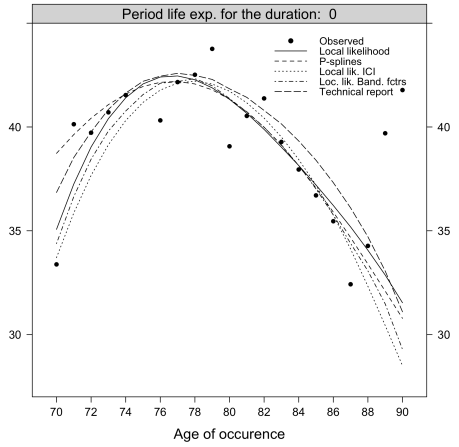


(b) Age of occurrence $v = 80$

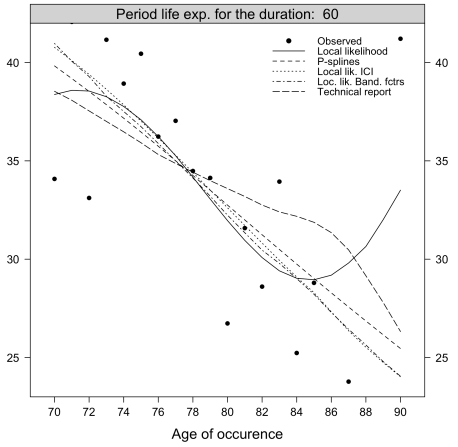


(c) Age of occurrence $v = 90$

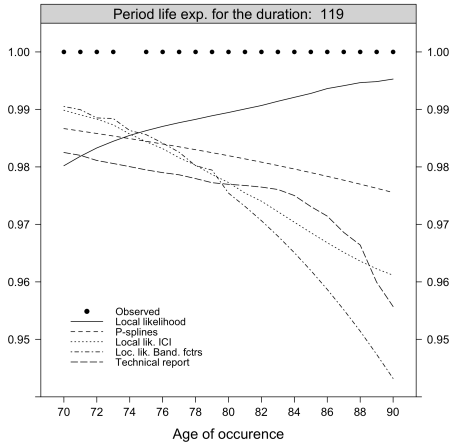
Figure 8: Observed and predicted period life expectancy for various ages of occurrence.



(a) Duration $u = 0$

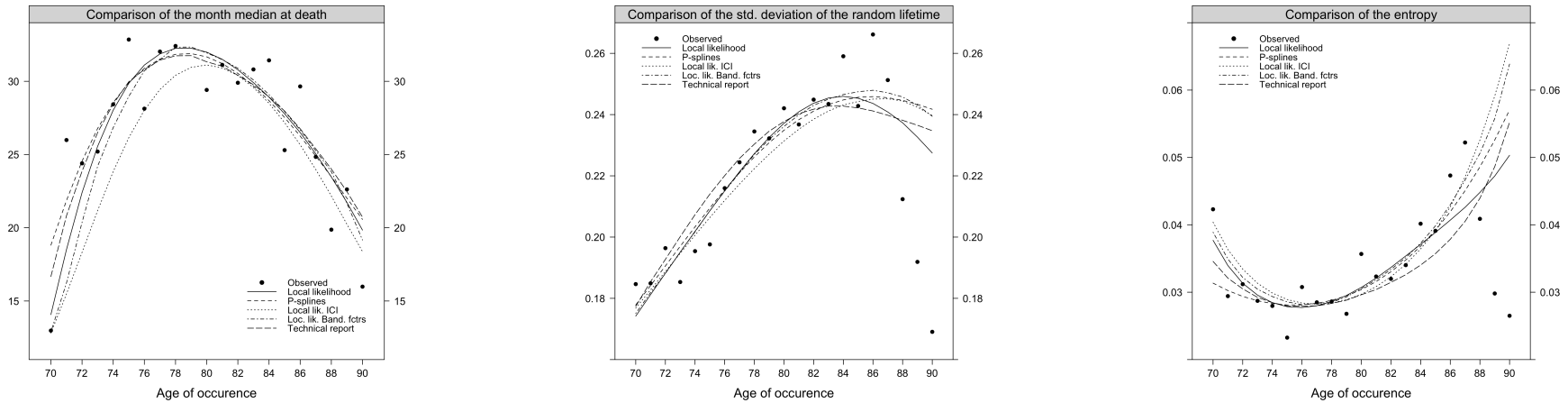


(b) Duration $u = 60$



(c) Duration $u = 119$

Figure 9: Observed and predicted period life expectancy for various duration.



(a) Median month at death (b) Std. dev. of life time (c) Entropy

Figure 10: Median month at death, standard deviation of the random life time and entropy with the age of occurrence of the claim.

	Observed	Local lik.	p-splines	Adapt. lik. ICI	Adapt. band. factors	Planchet (2012)
e_{70}	33.38	35.07	38.73	33.71	34.38	36.84
e_{80}	39.06	41.33	41.34	41.72	41.37	41.84
e_{90}	41.78	31.53	30.78	28.50	29.29	31.09
Average	39.44	38.59	38.80	37.88	38.12	39.23
Med$(T_0(70))$	12.98	14.07	18.81	12.81	12.89	16.67
Med$(T_0(80))$	29.41	31.99	31.67	31.12	31.94	31.33
Med$(T_0(90))$	15.98	19.84	20.57	18.38	19.16	20.73
Med(T_0)	27.07	26.83	27.29	24.99	26.40	27.18
$\sigma_0(70)$	0.1846	0.1742	0.1778	0.1768	0.1751	0.1774
$\sigma_0(80)$	0.2421	0.2372	0.2349	0.2314	0.2361	0.2377
$\sigma_0(90)$	0.1691	0.2274	0.2417	0.2397	0.2394	0.2347
σ_0	0.2196	0.2231	0.2250	0.2228	0.2251	0.2250
$H(T_0(70))$	0.0423	0.0377	0.0313	0.0404	0.0389	0.0346
$H(T_0(80))$	0.0357	0.0307	0.0304	0.0297	0.0305	0.0296
$H(T_0(90))$	0.0265	0.0503	0.0570	0.0669	0.0638	0.0552
$H(T_0)$	0.0337	0.0350	0.0353	0.0374	0.0370	0.0341

Table 4: Single figure indices to summarize the lifetime probability distributions.

estimated from the adjusted surface obtained in Planchet (2012) suggests that the estimated deaths are less stretched than the models predictions.

6.3 Consequences of uncertainty associated to the survival law

The pricing and reserving as well as the management of LTC portfolios are very sensitive to the choice of the mortality table adopted. For example, experts in LTC insurance know that a simple shift of half a year in the calculation of an age, i.e. corresponding to a modification of the roundness convention on the calculation of ages, impacts the premium of 2 %. When provisioning the insurer faces adverse deviations due to two distinct factors:

i. The random fluctuations of the observed mortality rates around the relevant expected values, i.e. the adjusted mortality rates, consequences of the finite size of the population exposed to the risk. The risk of random fluctuations (often called process risk) is diversifiable (one should better said pooling). Its financial impact decreases, in relative terms, as the portfolio size increases.

ii. The inaccuracy of the underlying survival law is called the table risk. It is the risk of unanticipated aggregate mortality, arising from the uncertainty in modeling LTC-claimants survival law. The table risk can be thought as the risk of systematic deviations referring not only to a parameter risk but, as well, to any other sources leading to a misinterpretation of the life table resulting for example from an evolution of medical techniques or a change in rules of acceptance. The risk of systematic deviations cannot be hedged by increasing the portfolio size. Actually, in relative terms, its severity does not reduce as the portfolio size increases, since deviations concern all the insureds in the same direction.

Focusing on the table risk, we can measure the sensibility of the reserve to the survival law by computing relative difference between the remaining period life expectancy for a given age and when applying a

reduction of 1 % to the conditional probabilities of death obtained with the proposed models. Table 5 presents the results for the age of occurrence of a pathology 70, 80 and 90.

	Local lik.	p-splines	Adapt. lik. ICI	Adapt. band. factors	Planchet (2012)
$v = 70$	1.1242	1.0339	1.1530	1.1402	1.0755
$v = 80$	0.8620	0.8689	0.8685	0.8641	0.8579
$v = 90$	1.0598	1.0002	1.0307	1.0182	1.0162
Average	0.9385	0.9246	0.9454	0.9351	0.9228

Table 5: Relative difference (in %) between the period life expectancy and when applying a reduction of 1 % to the conditional probabilities of death.

In a general manner, we observe that for regions with a large exposition, such as age of occurrence 80, the consequence of a deviation of the conditional probabilities of death is limited. In contrast, for regions having a low exposition, the impact is larger than the deviation initially applied. Between the models, the influence remains close as it is mainly determined byt he shape of the survival curve.

In a recent article, Planchet and Tomas (2013) analyze the consequences of an error of appreciation on the LTC-claimants survival probabilities in terms of level of reserves with the same data presented here. They introduce the risk of systematic deviations arising from the uncertainty on the conditional probability of death directly with a semi-parametric approach.
They measure the impact of uncertainty on the reserve by computing the relative difference between the 95 % quantile of the simulated remaining lifetime obtained from simulations to its expectation. They found that the impact of uncertainty is relatively linear on the remaining life expectancy for a given age of occurrence. With a level of volatility of 9 % the resulting uncertainty is approximatively 12 % when the pathology occurred at age 80. When computing the quantile at 99.5 %, the difference with the expected remaining lifetime is around 19.5 %. It means that the capital required for covering the uncertainty is approximatively 19.5 % of the best estimate. In applying a reduction of 20 % on the conditional probabilities of death with the same logic as the disability / morbidity shock described in the QIS5 specifications, CEIOPS (2010), the remaining life expectancy of an LTC-claimant when the pathology occurred at age 80 increases from 41.4 to 49.5 months, meaning a gain of 19.6 %. The authors conclude that setting the volatility of the disturbance at 9 % appears to be relatively coherent with the calibration of the standard formula.
In addition, it appears that the results are insensitive to the underlying structure of the survival law as the computation concerns the core of the distribution, i.e. the general form of the survival law. In consequence, the underlying structure has no impact.

Using the general approximation proposed by Guibert *et al.* (2010), Planchet and Tomas (2013) compute the ratio between the SCR and the best estimate of the reserve as a function of the portfolio size for the underwriting risk. They found that for an age of occurrence 80, the minimal SCR is around 38 % of the best estimate of the reserve for a portfolio of LTC-claimants of infinite size, i.e. when ignoring the risk of random fluctuations. For a size of 100 LTC-claimants, the minimal SCR is 144 % of the best estimate of the reserve. In addition, they observe that for a a portfolio of small size, the SCR is rapidly decreasing with the age of occurrence.
Unlike computing the quantiles of the distribution of the reserve, the results obtained with the general approximation proposed by Guibert *et al.* (2010) are very sensitive to the choice of the underlying survival

law. It highlights the impact of the structure of the survival law on the underwriting risk, in particular the importance of the tail of distribution.

7 Conclusions

In this study, we are interested in modeling the mortality of Long-Term Care (LTC) claimants belonging only to one state of severeness (heavy claimants). Practitioners often use empirical methods that rely heavily on experts opinion. We have proposed methods not depending on experts advice and allowing to extract more pertinently the information from the data. We illustrated how adaptive local likelihood methods can be used to graduate mortality surfaces. Tests and single indices summarizing the lifetime distributions have been used to compare the graduated forces of mortality obtained from adaptive local likelihood to global non-parametric methods such as local likelihood and p-splines models. In addition, we have briefly addressed the consequence of table risk arising from the uncertainty in modeling LTC-claimants survival law.

Using locally adaptive parameters instead of a global smoothing one may be advantageous for several reasons. The estimator can adapt to the structure of the regression function and to the reliability of the data, smoothing more when the volume of observations is low and less when it is high.

The intersection of confidence intervals (ICI) rule has been introduced as a locally adaptive pointwise method. The critical value controls the bias-variance tradeoff. Because a larger class of estimators is available, it may in turn affect the variability. Hence, the set of window sizes contains relatively large bandwidths. The choice of the set of window sizes is done subjectively, based on the the mechanism generating the data and on the performance of the smoother used in the fitting. Another drawback in applying such methods is that they require more computer time than a global procedure. Specifically, the computational effort is multiplied by the number of observations.

A technique closely related to the ICI rule is the Lepski method. This procedure uses the standard deviation of the difference $\widehat{\psi}_{\lambda_1}(x_i) - \widehat{\psi}_{\lambda}(x_i)$ for some $\lambda \leq \lambda_1$ until a significance difference is found. Chichignoud (2010, Section 1.5) provides an extensive discussion of the technique in his recent Ph.D Thesis. The discussion and the implementation of the Lepski method for graduating experience data originating from life insurance is a topic of ongoing research.

The bandwidth correction factors method allows the estimated forces of mortality to include explicitly the extra information provided by the changing amounts of exposure. The observed exposure decided the shape of the smoothness parameter. The magnification of that shape has been determined by a sensitivity parameter which we chose subjectively for practical reasons. The global bandwidth parameter is used to control the absolute level of the bandwidth curve. We used a global criterion instead of pointwise methods. It appears that the procedure has the ability to model relatively well the mortality pattern where the other models fail to model these features.

In global procedures as well as for locally adaptive procedures, there is no deterministic method to obtain the constellation of smoothing parameters with the classical selectors. Residual analysis and goodness of fit diagnostics are just as important for locally adaptive procedures as they are for global procedures. The purpose for which the mortality table is required must be kept clearly in mind, and the final choice of graduation is always a matter of judgment.

Rather then having a parametric model relying heavily on experts advice, having an adaptive model can be a benefit. However, the relative merit of the procedures would depend on the purpose for which the mortality table has been computed. If we are essentially exploring the data, then additional information derived might not justify the effort. However, the potential uses of adaptive approaches suggest that they have much to offer as part of the actuarial toolkit.

Acknowledgment

The authors wish to thank Professor A.Charpentier (UQaM) 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.

References

Camarda, C. G. (2008). *Smoothing methods for the analysis of mortality development*. Ph.D. thesis, Universidad Carlos III de Madrid.

CEIOPS (2010). QIS5 technical specifications. Technical report, European Commission - Internal Market and Services DG.

Chichignoud, M. (2010). *Performances statistiques d'estimateurs non-linéaires*. Ph.D. thesis, Université de Provence - U.F.R Mathématiques.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**, 596–610.

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. Härdle and M. G. Schimek, eds.

Courbage, C. and Roudaut, N. (2011). Long-term care insurance: The French example. *European Geriatric Medecine*, **2**(1), 22–25.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, **34**(2), 187–220.

Currie, I. D. and Durbán, M. (2002). Flexible smoothing with p-splines: a unified approach. *Statistical Modelling*, **2**(333-349).

Currie, I. D., Durbán, M., and Eilers, P. H. C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, **4**, 279–298.

Currie, I. D., Durbán, M., and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimentional smoothing. *Journal of the Royal Statistical Society*, **68**(Part 2), 259–280.

Czado, C. and Rudolph, F. (2002). Application of survival analysis methods to long-term care insurance. *Insurance: Mathematics & Economics*, **31**(3), 395–413.

de Boor, C. (2001). *A practical guide to splines*. New York: Springer Verlag, (revised ed.) edition.

Debón, A., Montes, F., and Sala, R. (2006). A comparison of nonparametric methods in the graduation of mortality: Application to data from the Valencia region (Spain). *International statistical Review*, **74**(2), 215–233.

Deléglise, M.-P., Hess, C., and Nouet, S. (2009). Tarification, provisionnement et pilotage d'un portefeuille dépendance. *Bulletin Français d'Actuariat*, **9**(17), 70–108.

Delwarde, A., Kachkhdze, D., Olie, L., and Denuit, M. (2004). Modèles linéaires et additifs généralisés, maximum de vraisemblance local et méthodes relationelles en assurance sur la vie. *Bulletin Français d'Actuariat*, **6**(12), 77–102.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, **11**(2), 89–102.

Eilers, P. H. C. and Marx, B. D. (2002). Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics*, **11**(4), 758–783.

Eilers, P. H. C., Currie, I. D., and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, **50**, 61–76.

Fan, J., Heckman, N. E., and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi likelihood functions. *Journal of the American Statistical Association*, **90**(429), 141–150.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M., and Engel, J. (1997). Local polynomial regression: optimal kernels and asymptotic minimax efficieny. *Annals of the Institute of Statistical Mathematics*, **49**(1), 79–99.

Felipe, A., Guillén, M., and Pérez-Marín, A. (2002). Recent mortality trends in the Spanish population. *British Actuarial Journal*, **8**(4), 757–786.

Forfar, D., McCutcheon, J., and Wilkie, A. (1988). On graduation by mathematical formula. *Journal of the Institute of Actuaries*, **115**(part I(459)), 643–652.

Gauzère, F., Commenges, D., Barberger-Gateau, P., Letenneur, L., and Dartigues, J.-F. (1999). Maladie et dépendance: description des évolutions par des modèles multi-états. *Population*, **54**(2), 205–222.

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*, **47**, 173–209.

Goldenshulger, A. and Nemirovski, A. (1997). On spatially adaptive estimation of nonparametric regression. *Mathematical methods of statistics*, **6**(2), 135–170.

Gschlössl, S., Schoenmaekers, P., and Denuit, M. (2011). Risk classification in life insurance: methodology and case study. *European Actuarial Journal*, **1**(1), 23–41.

Guibert, Q., Planchet, F., and Juillard, M. (2010). Un cadre de référence pour un modèle interne partiel en assurance de personnes : application à un contrat de rentes viagères. *Bulletin Français d'Actuariat*, **10**(20), 5–34.

Haberman, S. and Renshaw, A. E. (1996). Generalized linear models and actuarial science. *Journal of the Royal Statistical Society*, **45**(4), 407–436.

Helms, F., Czado, C., and Gschlössl, S. (2005). Calculation of ltc premiums based on direct estimates of transition probabilities. *ASTIN Bulletin*, **35**(2), 455–469.

Hickman, J. C. and Miller, R. B. (1977). Notes on bayesian graduation. *Transactions of the Society of Actuaries*, **29**, 7–49.

Katkovnik, V. (1999). A new method for varying adaptive bandwidth selection. *IEEE Transactions on signal processing*, **47**(9), 2567–2571.

Katkovnik, V., Foi, A., Egiazarian, K. O., and Astola, J. T. (2005). Anisotropic local likelihood approximations: Theory, algorithm, applications. In E. R. Dougherty, J. T. Astola, and K. O. Egiazarian, editors, *Proceedings of the SPIE - Image processing algorithms and systems IV*, volume 5672, pages 181–192.

Kessler, D. (2008). The long-term care insurance market. *The Geneva Papers on Risk and Insurance - Issues and Practice*, **33**(1), 33–40.

Kirkby, J. G. and Currie, I. D. (2010). Smooth models of mortality with period shocks. *Statistical Modelling*, **10**(2), 177–196.

Lang, S. and Umlauf, N. (2010). Applications of multilevel structured additive regression models to insurance aata. In Y. Lechevallier and G. Saporta, editors, *Proceedings of COMPSTAT'2010*, pages 155–164. Physica-Verlag HD.

Lang, S., Kragler, P., Haybach, G., and Fahrmeir, L. (2002). Bayesian space-time analysis of health insurance data. In Springer, editor, *In Schwaiger, M., O.Opitz, eds.: Exploratory Data Analysis in Empirical Research*.

Levantesi, S. and Menzietti, M. (2012). Managing longevity and disability risks in life annuities with long term care. *Insurance: Mathematics & Economics*, **50**(3), 391–401.

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

Marx, B. D. and Eilers, P. H. C. (1998). Direct generalized additive smoothing with penalized likelihood. *Computational Statistics & Data Analysis*, **28**, 193–209.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, volume 37 of *Monographs on Statistics and Applied Probability*. Boca Raton: Chapman & Hall / CRC Press, second edition.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, **135**, 370–384.

Planchet, F. (2012). Analyse de la survie des dépendants. Confidentiel Version 1.9, Institut de Science Financière et d'Assurances - Université Claude Bernard Lyon 1, 50 Avenue Tony Garnier - 69366 Lyon Cedex 07 - France.

Planchet, F. and Tomas, J. (2013). Uncertainty on survival probabilities and solvency capital requirement : application to long-term care insurance. *Cahiers de Recherche de l'ISFA*, **2013**(4), 1–11. Working paper.

Planchet, F. and Winter, P. (2007). L'utilisation des splines bidimensionnels pour l'estimation de lois de maintien en arrêt de travail. *Bulletin Français d'Actuariat*, **13**(7), 83–106.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Renshaw, A. E. (1991). Actuarial graduation practice and generalized linear and non-linear models. *Journal of Institute of Actuaries*, **118**, 295–312.

Richards, S. J., Kirkby, J. G., and Currie, I. D. (2006). The importance of year of birth in two dimensional mortality data. *British Actuarial Journal*, **12**(1), 5.

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). Essays on boundaries effects and practical considerations for univariate graduation of mortality by local likelihood models. *Insurance and Risk Management*, **80**(2), 203–261.