Annals of Actuarial Science, Vol. 6, part 2, pp. 284–306. © Institute and Faculty of Actuaries 2012
doi:10.1017/S174849951200005X
First published online 15 May 2012

A Bayesian smoothing spline method for mortality modelling

Arto Luoma*
University of Tampere, Finland

Anne Puustelli
University of Tampere, Finland

Lasse Koskinen
Financial Supervisory Authority of Finland and Helsinki School of Economics, Finland

Abstract

We propose a new method for two-dimensional mortality modelling. Our approach smoothes the data set in the dimensions of cohort and age using Bayesian smoothing splines. The method allows the data set to be imbalanced, since more recent cohorts have fewer observations. We suggest an initial model for observed death rates, and an improved model which deals with the numbers of deaths directly. Unobserved death rates are estimated by smoothing the data with a suitable prior distribution. To assess the fit and plausibility of our models we perform model checks by introducing appropriate test quantities. We show that our final model fulfils nearly all requirements set for a good mortality model.

Keywords

Cohort effect; Forecasting; Model checking; Parameter uncertainty; Stochastic mortality model

1 Introduction

Mortality forecasting is a problem of fundamental importance for the insurance and pensions industry. Due to the increasing focus on risk management and measurement for insurers and pension funds, stochastic mortality models have attracted considerable interest in recent years. A range of stochastic models for mortality have been proposed, for example the seminal models of Lee & Carter (1992), Renshaw & Haberman (2006) and Cairns *et al.* (2006b). Some models build on an assumption of smoothness in mortality rates between ages in any given year (e.g. Cairns *et al.*, 2006b), while others allow for roughness, (e.g. Lee & Carter 1992; Renshaw & Haberman, 2006).

In this paper we propose a new Bayesian method for two-dimensional mortality modelling. Our method is based on natural cubic smoothing splines, which are popular in statistical applications, since the smoothing problem can be solved using simple linear algebra. In this approach the distinct data values are taken as knots of the spline, and its smoothness is achieved by employing roughness penalty in a penalized likelihood function. In the Bayesian approach, the prior distribution takes the

*Correspondence to: Arto Luoma, School of Information Sciences, FIN-33014 University of Tampere, Finland. E-mail: arto.luoma@uta.fi

role of the roughness penalty term. A useful introduction to smoothing splines may be found, for example, in Green & Silverman (1994).

A more general penalized splines approach would employ a set of basis functions, such as B-splines. In the case that cubic B-splines are used, one may obtain the same solution as in the smoothing spline approach by using the same roughness penalty and by choosing the knots to be the distinct values of the data points. Compared to the general penalized splines approach our approach has the advantage that one does not need to optimize with respect to the number of knots and their locations. However, the drawback in our approach is that the matrices involved in computations become too large, unless one restricts the size of the estimation data set.

We use age-cohort data instead of age-period data, since we wish to preserve the sequential dependence of observations within each cohort. Therefore, we have to deal with imbalanced data, since more recent cohorts have fewer observations. We suggest an initial model for the observed death rates, and an improved model which deals with the numbers of deaths directly. We assume the number of deaths to follow a Poisson distribution, a common model for the number of deaths in a year in a particular cohort. Unobserved death rates are estimated by smoothing the data with one of our spline models. The proposed method is illustrated using Finnish mortality data for females, provided by the Human Mortality Database. We implement the Bayesian approach using the Markov chain Monte Carlo method (MCMC), or more specifically, the single-component Metropolis-Hastings algorithm.

The use of Bayesian methods is not new in this general context. Dellaportas *et al*. (2001) proposed a Bayesian mortality model in a parametric curve modelling context. Czado *et al*. (2005) and Pedroza (2006) provided Bayesian analyses for the Lee-Carter model using MCMC, with further work by Kogure & Kurachi (2010). More recently, Reichmuth & Sarferaz (2008) have applied MCMC to a version of the Renshaw & Haberman (2006) model. Schmid & Held (2007) present software which allows analysis of incidence count data with a Bayesian age-period-cohort model. Cairns *et al*. (2011) use the same model to compare results based on a two-population approach with single-population results. Currie *et al*. (2004) and Richards *et al*. (2006) assume smoothness in both age and cohort dimensions through the use of P-splines in a non-Bayesian set-up. Lang & Brezger (2004) introduce two-dimensional P-splines in a Bayesian set-up but in a different context.

Cairns *et al*. (2008) evaluated several types of stochastic mortality models using a checklist of criteria. These criteria are based on general characteristics and the ability of the model to explain historical patterns of mortality. None of the existing models met all of the criteria. However, Plat (2009) later proposed a model which apart from partly meeting the parsimony criteria meets all of the criteria. We also follow the same list in assessing the fit and plausibility of our model.

The plan of the paper is as follows. In the next section we describe the data and its use in estimation. In Section 3 we explain the smoothing problem and present the Bayesian formulation of the preliminary model, and in Section 4 we describe our final model. In Section 5 we introduce the estimation method and provide some convergence results. The model checks are described in Section 6, after which we conclude with a brief discussion.

2 Data

We use mortality data provided by the Human Mortality Database. This was created to provide detailed mortality and population data to those interested in the history of human longevity. In our

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

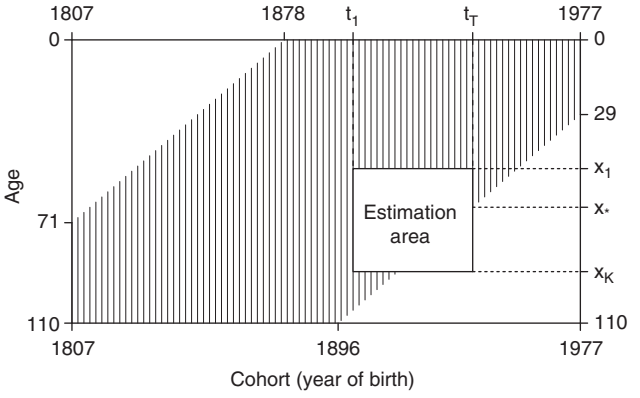

Figure 1. Age-cohort representation of the data set. The complete data set is indicated by the streaked area, and the imbalanced estimation set by the white rectangle.

work we use Finnish cohort mortality data for females. We use age-cohort data instead of age-period data, since we wish to take into account the dependence of consecutive observations within each cohort. In the complete data matrix the years of birth included are between 1807 and 1977; hence there are 171 different cohorts. The most recent data are from 2006. When the age group of persons 110 years and older is excluded, the dimensions of the data matrix become 110×171. These data are illustrated in Figure 1, in which the observed area is denoted by vertical lines and the unobserved by two white triangles in the upper left and lower right corners.

Our estimation method would produce huge matrices if all these data were used simultaneously. Therefore, we define estimation areas which are parts of the complete data set. A rectangular estimation area shown in Figure 1 indicates the cohorts and ages for which a smooth spline surface is fitted. The mortality rates are known for part of this area, and they are predicted for the unknown part. More specifically, an estimation area is defined by minimum age x_1, maximum age x_K, minimum cohort t_1 and maximum cohort t_T. The maximum age for which data are available in cohort t_T is denoted as x_*. Thus, the number of ages included is $K = x_K - x_1 + 1$ and the number of cohorts $T = t_T - t_1 + 1$.

Since the reader might be more familiar with age-period data, we have also plotted the data set in the dimensions of age and year in Figure 2. One should, however, remember that the figures in these two types of mortality tables are not computed in the same way. One figure in an age-period table is based on persons who have a certain (discrete) age during one calendar year and are born during two consecutive years, while each figure in an age-cohort table is based on data from two consecutive calendar years about persons born in a certain year (for details, see Wilmoth *et al.*, 2007).

3 Preliminary model

We start building our model in a simplified set-up. Let us denote the logarithms of observed death rates as $y_{xt} = \log(m_{xt})$ for ages $x = x_1, x_2, \ldots, x_K$ and cohorts (years of birth) $t = t_1, t_2, \ldots, t_T$. The observed death rates are defined as

$$m_{xt} = \frac{d_{xt}}{e_{xt}},$$

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

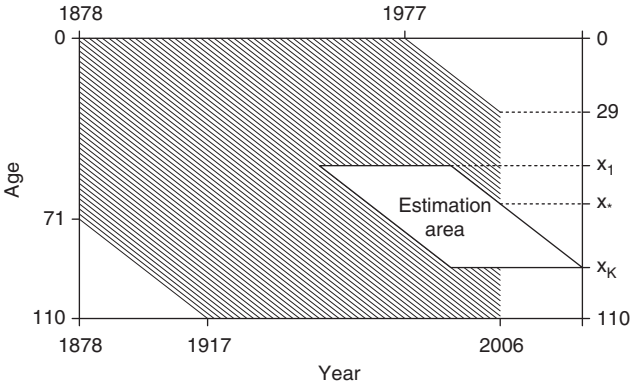

Figure 2. Age-period representation of the data set. The complete data set is indicated by the streaked area, and the imbalanced estimation set by the white parallelogram.

where d_{xt} is the number of deaths and e_{xt} the person years of exposure. In our preliminary set-up we model the observed death rates directly, while in our final set-up we model the theoretical, unobserved death rates μ_{xt}.

3.1 The smoothing problem

Our goal is to smooth and predict logarithms of observed death rates. We fit a smooth two-dimensional curve $\theta(x,t)$, and denote its values at discrete points as $\theta_{x,t}$. In matrix form we may write

$$\mathbf{Y} = \boldsymbol{\Theta} + \mathbf{E},$$

where $\mathbf{Y}$ is a $K \times T$ matrix of observations, $\boldsymbol{\Theta}$ is a matrix of smoothed values, and $\mathbf{E}$ is a matrix of errors. We denote the columns of $\mathbf{Y}$, $\boldsymbol{\Theta}$ and $\mathbf{E}$ by $\boldsymbol{y}_j$, $\boldsymbol{\theta}_j$ and $\boldsymbol{\varepsilon}_j$, respectively. Concatenating the columns we obtain $\mathbf{y} = \mathrm{vec}(\mathbf{Y})$, $\theta = \mathrm{vec}(\boldsymbol{\Theta})$ and $\boldsymbol{\varepsilon} = \mathrm{vec}(\mathbf{E})$.

We further assume that the death rates within a cohort follow a multivariate normal distribution having an AR(1) correlation structure with autocorrelation coefficient ϕ. Thus,

$$\boldsymbol{\epsilon}_j \sim N(0, \sigma^2\mathbf{P}), \quad j = 1, 2, \ldots, T,$$

where $\mathbf{P}$ is a correlation matrix with elements $\rho_{rs} = \phi^{|r-s|}$. The observations in different cohorts are assumed to be independent.

In general, all observations are not available for all cohorts. For each j, we may partition $\mathbf{y}_j$ into observed $\mathbf{y}_{j1}$ and unobserved $\mathbf{y}_{j2}$, and $\boldsymbol{\theta}_j$ correspondingly into $\boldsymbol{\theta}_{j1}$ and $\boldsymbol{\theta}_{j2}$, and $\mathbf{P}$ into $\mathbf{P}_{\mathrm{j,r,s}}$, $r, s = 1,2$. The unobserved part of the data can be predicted using the result about the conditional distribution of the multivariate normal distribution:

$$\{\mathbf{y}_{j2} \mid \mathbf{y}_{j1}, \sigma^2, \phi\} \sim N(\boldsymbol{\theta}_{j,2.1}, \sigma^2\mathbf{P}_{j,22.1}),$$

where $\boldsymbol{\theta}_{j,2.1} = \boldsymbol{\theta}_{j2} + \mathbf{P}_{j,21}\mathbf{P}_{j,11}^{-1}(\mathbf{y}_{j1} - \boldsymbol{\theta}_{j1})$ and $\mathbf{P}_{j,22.1} = \mathbf{P}_{j,22} - \mathbf{P}_{j,21}\mathbf{P}_{j,11}^{-1}\mathbf{P}_{j,12}$.

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

When estimating $\boldsymbol{\theta}$ we wish to minimize the generalized sum of squares

$$SS_1 = \sum_{j=1}^{T} (\mathbf{y}_{j1} - \boldsymbol{\theta}_{j1})' \mathbf{P}_{j,11}^{-1} (\mathbf{y}_{j1} - \boldsymbol{\theta}_{j1}). \tag{1}$$

The vector of all observed mortality rates is $\mathbf{y}^{obs} = \mathbf{S}\mathbf{y}$, where $\mathbf{S}$ is a selection matrix selecting the known values from the complete data vector $\mathbf{y}$. The matrix $\mathbf{S}$ can be constructed from the identity matrix of size KT by including the ith row ($i = 1, 2, \ldots, KT$) if the ith element of $\mathbf{y}$ is known. Now we can write (1) as

$$SS_1 = (\mathbf{y}^{obs} - \mathbf{S}\boldsymbol{\theta})' (\mathbf{S}\mathbf{P}_* \mathbf{S}')^{-1} (\mathbf{y}^{obs} - \mathbf{S}\boldsymbol{\theta}), \tag{2}$$

where $\mathbf{P}_* = \mathbf{I}_T \otimes \mathbf{P}$.

In addition to maximizing fit, we wish to smooth $\boldsymbol{\Theta}$ in the dimensions of cohort and age. Specifically, we minimize the roughness functional

$$\int_{x_1}^{x_K} \left[\frac{\partial^2}{\partial x^2} \theta(x, t_j) \right]^2 dx \tag{3}$$

for each $j = 1,2,\ldots, T$ and

$$\int_{t_1}^{t_T} \left[\frac{\partial^2}{\partial t^2} \theta(x_k, t) \right]^2 dt \tag{4}$$

for each $k = 1,2,\ldots, K$.

If $\theta(x,t_j)$ is considered a smooth function of x obtaining fixed values at points $x_1, x_2, \ldots, x_K$, then using variational calculus it can be shown that the integral in (3) is minimized by choosing $\boldsymbol{\theta}(x,t_j)$ to be a cubic splines curve with knots at $x_1, x_2, \ldots, x_K$. Furthermore, this integral can be expressed as a squared form $\boldsymbol{\theta}_j' \mathbf{G}_K \boldsymbol{\theta}_j$, where $\mathbf{G}_K$ is a so-called roughness matrix with dimensions $K \times K$ (for proof, see Green & Silverman, 1994). Similarly, if $\theta(x_k,t)$ is a cubic splines curve with knots at $t_1, \ldots, t_T$, the integral in (4) equals $\boldsymbol{\theta}_{(k)}' \mathbf{G}_T \boldsymbol{\theta}_k$, where $\boldsymbol{\theta}_{(k)}$ denotes the kth row of $\boldsymbol{\Theta}$ and $\mathbf{G}_T$ is a $T \times T$ roughness matrix. Thus, we wish to minimize

$$SS_2 = \sum_{j=1}^{T} \boldsymbol{\theta}_j' \mathbf{G}_K \boldsymbol{\theta}_j = \boldsymbol{\theta}' (\mathbf{I}_T \otimes \mathbf{G}_K) \boldsymbol{\theta} \tag{5}$$

and

$$SS_3 = \sum_{k=1}^{K} \boldsymbol{\theta}_{(k)}' \mathbf{G}_T \boldsymbol{\theta}_{(k)} = \boldsymbol{\theta}' (\mathbf{G}_T \otimes \mathbf{I}_K) \boldsymbol{\theta}. \tag{6}$$

An $N \times N$ roughness matrix is defined as $\mathbf{G}_N = \boldsymbol{\nabla}_N \boldsymbol{\Delta}_N^{-1} \boldsymbol{\nabla}_N'$ where the non-zero elements of banded $N \times (N-2)$ and $(N-2) \times (N-2)$ matrices $\boldsymbol{\nabla}_N$ and $\boldsymbol{\Delta}_N$, respectively, are defined as follows:

$$\nabla_{i,i} = \frac{1}{x_{i+1} - x_i}, \quad \nabla_{i+1,i} = -\left(\frac{1}{x_{i+1} - x_i} + \frac{1}{x_{i+2} - x_{i+1}} \right), \quad \nabla_{i+2,i} = \frac{1}{x_{i+2} - x_{i+1}}$$

and

$$\Delta_{i,i+1} = \Delta_{i+1,i} = \frac{x_{i+2} - x_{i+1}}{6}, \quad \Delta_{i,i} = \frac{x_{i+2} - x_i}{3},$$

with data points $x_i, i = 1, \ldots, n$. In our case the data are given at equal intervals, implying that

$$\nabla_{i,i} = 1, \quad \nabla_{i+1,i} = -2, \quad \nabla_{i+2,i} = 1$$

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

and

$$\Delta_{i,i+1} = \Delta_{i+1,i} = \frac{1}{6}, \quad \Delta_{i,i} = \frac{2}{3}.$$

Combining the previous results, we obtain the bivariate smoothing splines solution for $\boldsymbol{\theta}$ by minimizing the expression $SS_1 + \lambda_1 SS_2 + \lambda_2 SS_3$, where SS_1, SS_2 and SS_3 are given in the equations (2), (5) and (6), respectively, and the parameters λ_1 and λ_2 control smoothing in the dimensions of age and cohort, respectively. Using matrix differentiation and the properties of Kronecker's product, it is easy to show that for fixed values of λ_1 and λ_2 the minimal solution is given by

$$\hat{\boldsymbol{\theta}} = \left[\mathbf{S}'(\mathbf{S}\mathbf{P}_*\mathbf{S}')^{-1}\mathbf{S} + \mathbf{A}\right]^{-1}\mathbf{S}'(\mathbf{S}\mathbf{P}_*\mathbf{S}')^{-1}\mathbf{y}^{obs}, \tag{7}$$

where

$$\mathbf{A} = \lambda_1(\mathbf{I}_T \otimes \mathbf{G}_K) + \lambda_2(\mathbf{G}_T \otimes \mathbf{I}_K). \tag{8}$$

In the special case that the data set is balanced ($\mathbf{S}$ is an identity matrix), the solution is simplified to $\hat{\boldsymbol{\theta}} = (\mathbf{I} + \mathbf{P}_*\mathbf{A})^{-1}\mathbf{y}$.

3.2 Bayesian formulation

Bayesian statistical inference is based on the posterior distribution, which is the conditional distribution of unknown parameters given the data. In order to compute the posterior distribution one needs to define the prior distribution, which is the unconditional distribution of parameters, and the likelihood function, which is the probability density of observations given the parameters. Bayes' theorem implies that the posterior distribution is proportional to the product of the prior distribution and the likelihood:

$$p(\boldsymbol{\eta} \mid \mathbf{y}) \propto p(\boldsymbol{\eta})p(\mathbf{y} \mid \boldsymbol{\eta}),$$

where $\mathbf{y}$ is the data vector and $\boldsymbol{\eta}$ the vector of all unknown parameters.

In our case, the likelihood is given by

$$p(\mathbf{y}^{obs} \mid \boldsymbol{\eta}) = (2\pi\sigma^2)^{-\frac{K_*}{2}} \mid \mathbf{S}\mathbf{P}_*\mathbf{S}' \mid^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}^{obs}-\mathbf{S}\boldsymbol{\theta})'(\mathbf{S}\mathbf{P}_*\mathbf{S}')^{-1}(\mathbf{y}^{obs}-\mathbf{S}\boldsymbol{\theta})}, \tag{9}$$

where K_* is the length of $\mathbf{y}^{obs}$.

In order to facilitate estimation we reparametrize the smoothing parameters as follows: $\lambda = \lambda_1$ and $\omega = \lambda_2/\lambda_1$, where λ_1 and λ_2 control the smoothing in the dimensions of age and cohort, respectively. Furthermore, we use the following hierarchical prior for $\boldsymbol{\eta}$:

$$p(\boldsymbol{\eta}) = p(\sigma^2)p(\lambda)p(\omega)p(\phi)p(\boldsymbol{\theta}|\sigma^2, \lambda, \omega, \phi),$$

where

$$p(\sigma^2) \propto \frac{1}{\sigma^2}$$
$$p(\lambda) \propto \lambda^{\alpha_1 - 1}e^{-\beta_1\lambda}$$
$$p(\omega) \propto \omega^{\alpha_2 - 1}e^{-\beta_2\omega}$$
$$p(\phi) \propto 1, \quad -1 < \phi < 1.$$

As hyperparameters we set $\alpha_1 = \beta_1 = 0.001$ and $\alpha_2 = \beta_2 = 10$. Thus, the prior of σ^2 is the standard uninformative improper prior used for positive parameters, and the priors of λ and ϕ are also fairly

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

uninformative. The prior of ω is instead more informative, having mean 1 and variance 0.1, since we found that the data do not contain enough information about ω, and with a looser prior we would face convergence problems in estimation. We made sensitivity analysis with respect to the prior of λ and found that increasing or decreasing the order of magnitude of α_1 and β_1 did not essentially affect the results.

The smoothing effect can now be obtained by choosing a conditional prior for $\boldsymbol{\theta}$ which is consistent with the smoothing splines solution. Such a prior contains information only on the curvature, or roughness, of the spline surface, not on its position or gradient. Thus, we assume that $\{\boldsymbol{\theta}|\sigma^2, \lambda, \omega, \phi\}$ is multivarite normal with density

$$p(\boldsymbol{\theta} \mid \sigma^2, \lambda, \omega, \phi) = (2\pi\sigma^2)^{-\frac{KT}{2}} \left|\lambda\left[(\mathbf{I}_T \otimes \mathbf{G}_{K,\gamma} + \omega(\mathbf{G}_T \otimes \mathbf{I}_K)\right]\right|^{\frac{1}{2}} e^{-\frac{\lambda}{2\sigma^2}\boldsymbol{\theta}'\left[(\mathbf{I}_T \otimes \mathbf{G}_{K,\gamma}) + \omega(\mathbf{G}_T \otimes \mathbf{I}_K)\right]\boldsymbol{\theta}}, \tag{11}$$

where $\mathbf{G}_{K,\gamma}$ is a positive definite matrix approximating $\mathbf{G}_K$. More specifically, we define $\mathbf{G}_{K,\gamma} = \mathbf{G}_K + \gamma\mathbf{X}\mathbf{X}'$, where $\gamma > 0$ can be chosen to be arbitrarily small, and $\mathbf{X} = (\mathbf{1}\ \mathbf{x})$ with $\mathbf{1} = (1,\ldots,1)'$ and $\mathbf{x} = (x_1,\ldots,x_K)'$. Initially, we use $\mathbf{G}_{K,\gamma}$ instead of $\mathbf{G}_K$, since otherwise $p(\theta \mid \sigma^2, \lambda, \omega, \phi)$ would be improper, which would lead to difficulties when deriving the conditional posteriors for λ and ω.

Multiplying the densities in (11) and (9) and picking the factors which include $\boldsymbol{\theta}$ we obtain the full conditional posterior for $\boldsymbol{\theta}$ up to a constant of proportionality:

$$p(\boldsymbol{\theta}|\mathbf{y}, \sigma^2, \lambda, \omega, \phi) \propto \mathrm{e}^{-\frac{1}{2\sigma^2}\left\{(\mathbf{y}^{obs} - \mathbf{S}\boldsymbol{\theta})'(\mathbf{S}\mathbf{P}_*\mathbf{S}')^{-1}(\mathbf{y}^{obs} - \mathbf{S}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}'\left[(\mathbf{I}_T \otimes \mathbf{G}_{K,\gamma}) + \omega(\mathbf{G}_T \otimes \mathbf{I}_K)\right]\boldsymbol{\theta}\right\}}. \tag{12}$$

Manipulating this expression and replacing $\mathbf{G}_{K,\gamma}$ with $\mathbf{G}_k$ we obtain

$$p(\theta \mid \mathbf{y}, \sigma^2, \lambda, \omega, \phi) \propto \mathrm{e}^{-\frac{1}{2\sigma^2}(\theta - \hat{\theta})'\mathbf{B}(\theta - \hat{\theta})},$$

where $\hat{\boldsymbol{\theta}}$ is given in (7) and $\mathbf{B} = \mathbf{A} + \mathbf{S}'(\mathbf{S}\mathbf{P}_*\mathbf{S}')^{-1}\mathbf{S}$. From this we see that the conditional posterior distribution of $\boldsymbol{\theta}$ is multivariate normal with mean $\hat{\boldsymbol{\theta}}$ and covariance matrix $\sigma^2\mathbf{B}^{-1}$ in the limiting case when $\mathbf{G}_{K,\gamma} \to \mathbf{G}_K$. This implies that the conditional posterior mode for $\boldsymbol{\theta}$ is equal to the smoothing splines solution provided in the previous section. Thus, using the multivariate prior described above, we can implement the roughness penalty of smoothing splines in the Bayesian framework.

In order to implement estimation using the Gibbs sampler, the full conditional posterior distributions of the parameters are needed. In the following, we will provide these for σ^2, λ, ω and ϕ in the limiting case when $\mathbf{G}_{K,\gamma} \to \mathbf{G}_K$.

The conditional posterior of σ^2 is

$$p(\sigma^2 \mid \mathbf{y}, \lambda, \omega, \phi) \propto (\sigma^2)^{-\left(\frac{K_* + KT}{2} + 1\right)} \mathrm{e}^{-\frac{1}{2\sigma^2}\left[(\mathbf{y}^{obs} - \mathbf{S}\boldsymbol{\theta})'(\mathbf{S}\mathbf{P}_*\mathbf{S}')^{-1}(\mathbf{y}^{obs} - \mathbf{S}\boldsymbol{\theta}) + \boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta}\right]},$$

which is the density of the scaled inverted χ^2 -distribution Inv$-\chi^2(v, b)$[1], where $\gamma = K_* + KT$ and $b = (SS_1 + \lambda SS_2 + \lambda\omega SS_3)/v$ with SS_1, SS_2 and SS_3 given in (2), (5) and (6).

[1] Notation $X \sim \text{Inv}-\chi^2(v, b)$ means that $vb/X \sim \chi^2_v$.

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

The conditional posterior of λ is

$$p(\lambda \mid \mathbf{y}, \boldsymbol{\theta}, \sigma^2, \omega, \phi) \propto \lambda^{\alpha_1 - 1 + \frac{KT}{2}} \mathrm{e}^{-\lambda\left[\beta_1 + \frac{1}{2\sigma^2}\boldsymbol{\theta}'(\mathbf{I}_T \otimes \mathbf{G}_K + \omega \mathbf{G}_T \otimes \mathbf{I}_K)\boldsymbol{\theta}\right]}, \tag{13}$$

which is the density of Gamma $(\alpha_1 + KT/2,\ \beta_1 + (SS_2 + \omega SS_3)/(2\sigma^2))$.

The conditional posterior of ω is

$$p(\omega \mid \mathbf{y}, \boldsymbol{\theta}, \sigma^2, \lambda, \phi) \propto \omega^{\alpha_2 + T - 2}\left[\prod_{k=1}^{K-2}\prod_{j=1}^{T-2}\left(1 + \omega\frac{\mu_j}{\nu_k}\right)\right]^{\frac{1}{2}} \mathrm{e}^{-\omega\left[\beta_2 + \frac{\lambda}{2\sigma^2}\boldsymbol{\theta}'(\mathbf{G}_T \otimes \mathbf{I}_K)\boldsymbol{\theta}\right]}, \tag{14}$$

where $\mu_j, j = 1, \ldots, T-2$ and $\nu_K, k = 1, \ldots, K-2$, are the nonzero eigenvalues of $\mathbf{G}_T$ and $\mathbf{G}_K$, respectively. This is not a standard distribution, but since it is log-concave, it is possible to generate values from it using adaptive rejection sampling, introduced by Gilks & Wild (1992).

Finally, the conditional posterior of ϕ, given by

$$p(\phi|\mathbf{y}, \boldsymbol{\theta}, \sigma^2, \lambda, \omega) \propto (1-\phi^2)^{-\frac{1}{2}(K_* - T)} \mathrm{e}^{-\frac{1}{2\sigma^2}(\mathbf{y}^{obs} - \mathbf{S}\boldsymbol{\theta})'(\mathbf{S}\mathbf{P}_*\mathbf{S}')^{-1}(\mathbf{y}^{obs} - \mathbf{S}\boldsymbol{\theta})},$$

is not of standard form, and it is therefore difficult to generate random variates from it directly. Instead, we may employ a Metropolis step within the Gibbs sampler.

Now, the estimation algorithm is implemented so that the parameters λ, ω, σ^2 and $\boldsymbol{\theta}$ are updated one by one using Gibbs steps, and ϕ is updated using a Metropolis step. Further details will be given in Section 5.

4 The final model

In our final set-up we are able to control for unsystematic mortality risk in addition to systematic risk. Unsystematic risk means that even if the true mortality rate were known, the numbers of deaths would remain unpredictable. When the population becomes larger, the unsystematic mortality risk becomes smaller due to diversification.

4.1 Formulation and estimation

In the final model the inference is rendered more accurate by modelling the observed numbers of deaths directly. Specifically, we assume that

$$d_{xt} \sim \text{Poisson}(\mu_{xt} e_{xt}),$$

where d_{xt} is the number of deaths at age x and cohort t, μ_{xt} is the theoretical death rate (also called intensity of mortality or force of mortality) and e_{xt} is the person years of exposure. This is an approximation, since neither the death rate nor the exposure is constant during any given year. Our purpose is to model $\theta_{xt} = \log(\mu_{xt})$ with a smooth spline surface. Compared to the preliminary model we have replaced m_{xt} with μ_{xt} and removed the error term and its autocorrelation structure.

Similarly to the preliminary model, we obtain the smoothing effect by using a suitable conditional prior distribution for $\boldsymbol{\theta}$. Specifically, we obtain $p(\boldsymbol{\theta}|\lambda,\omega)$ by replacing σ^2 with 1 in equation (11). For λ and ω we use the same prior distributions as earlier, given by (10), and their conditional posteriors are obtained from (13) and (14) when σ^2 is set at 1. However, here we use hyperparameters $\alpha_1 = \beta_1 = 10^{-6}$, since removing σ^2 changes the scale of λ several orders of magnitude.

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

Now the full conditional posterior distribution of $\boldsymbol{\theta}$ may be written as

$$p(\boldsymbol{\theta}|\mathbf{d}^{obs},\lambda,\omega) \propto \exp\left\{\sum_{t=t_1}^{t_T}\sum_{x=x_1}^{x_{K_t}}[d_{xt}\theta_{xt}-e_{xt}\exp(\theta_{xt})]-\frac{1}{2}\boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta}\right\}, \tag{15}$$

where $\mathbf{d}^{obs}$ is a vector of observed death numbers, and K_t the number of ages for which data are available in cohort t. The double sum in this expression comes from the likelihood function and the squared form from the prior distribution.

This model can be estimated similarly to the preliminary model, using Gibbs sampling. However, since the conditional distribution in (15) is non-standard, it is difficult to sample from it directly. Here we may use a Metropolis-Hastings step within the Gibbs sampler. As a proposal distribution we may use a multivariate normal approximation to (15), given by

$$J(\boldsymbol{\theta}|\lambda,\omega) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{y}^{obs}-\mathbf{S}\boldsymbol{\theta}\right)\left(\mathbf{S}\boldsymbol{\Sigma}\mathbf{S}'\right)^{-1}\left(\mathbf{y}^{obs}-\mathbf{S}\boldsymbol{\theta}\right)-\frac{1}{2}\boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta}\right\},$$

where $\mathbf{y}^{obs}$ is a vector of observed log death rates, $\boldsymbol{\Sigma}$ is a diagonal matrix with approximate variances of log death rates, denoted as v_{xt}, $x = x_1,\ldots,x_K$, $t = t_1,\ldots,t_T$, as its diagonal elements, and $\mathbf{S}$ is a selection matrix defined in Section 3.1. We obtain v_{xt} by applying the delta method to the relevant transformation of the underlying Poisson variable. More specifically, we use $v_{xt} = 1/(e_{xt}\exp(\tilde{\theta}_{xt}))$, where $\tilde{\theta}_{xt} = (\sum_{i=-1}^{1}\sum_{j=-1}^{1} y_{x+i,t+j})/9$ is an initial approximation to the log death rate.

Thus, the proposal $\boldsymbol{\theta}^*$ is distributed as

$$\boldsymbol{\theta}^* \sim \mathrm{MVN}(\mathbf{C}^{-1}\mathbf{S}'(\mathbf{S}\boldsymbol{\Sigma}\mathbf{S}')^{-1}\mathbf{y}^{obs},\mathbf{C}^{-1}),$$

where $\mathbf{C} = \mathbf{A} + \mathbf{S}'(\mathbf{S}\boldsymbol{\Sigma}\mathbf{S}')^{-1}\mathbf{S}$, and is accepted with probability

$$\min\left(1,\frac{p(\boldsymbol{\theta}^*\,|\,\mathbf{d}^{obs},\lambda,\omega)/J(\boldsymbol{\theta}^*\,|\,\lambda,\omega)}{p(\boldsymbol{\theta}\,|\,\mathbf{d}^{obs},\lambda,\omega)/J(\boldsymbol{\theta}\,|\,\lambda,\omega)}\right).$$

The whole algorithm is once more a special case of the single-component Metropolis-Hastings. Further details on this algorithm will be provided in the next section.

5 Estimation

5.1 Estimation procedure

Our estimation procedure is a single-component (or cyclic) Metropolis-Hastings algorithm. This is one of the Markov Chain Monte Carlo (MCMC) methods, which are useful in drawing samples from posterior distributions. Generally, MCMC methods are based on drawing values from approximate distributions and then correcting these draws to better approximate the target distribution, and hence they are used when direct sampling from a target distribution is difficult. A useful reference for different versions of MCMC is Gilks *et al.* (1996).

The Metropolis-Hastings algorithm was introduced by Hastings (1970) as a generalization of the Metropolis algorithm (Metropolis *et al.*, 1953). Also the Gibbs sampler proposed by Geman & Geman (1984) is its special case. The Gibbs sampler assumes the full conditional distributions of the

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

target distribution to be such that one is able to generate random numbers or vectors from them. The Metropolis and Metropolis-Hastings algorithms are more flexible than the Gibbs sampler; with them one only needs to know the joint density function of the target distribution with density $p(\theta)$ up to a constant of proportionality.

With the Metropolis algorithm the target distribution is generated as follows: first a starting distribution $p_0(\theta)$ is assigned, and from it a starting-point θ^0 is drawn such that $p(\theta^0) > 0$. For iterations $t = 1,2,\ldots$, a proposal θ^* is generated from a jumping distribution $J(\theta^* \mid \theta^{t-1})$, which is symmetric in the sense that $J(\theta_a \mid \theta_b) = J(\theta_b \mid \theta_a)$ for all θ_a and θ_b. Finally, iteration t is completed by calculating the ratio

$$r = \frac{p(\theta^*)}{p(\theta^{t-1})} \tag{16}$$

and by setting the new value at

$$\theta^t = \begin{cases} \theta^* & \text{with probability min(r, 1)} \\ \theta^{t-1} & \text{otherwise.} \end{cases}$$

It can be shown that, under mild conditions, the algorithm produces an ergodic Markov Chain whose stationary distribution is the target distribution.

Metropolis-Hastings algorithm generalizes the Metropolis algorithm by removing the assumption of symmetric jumping distribution. The ratio r in (16) is replaced by

$$r = \frac{p(\theta^*) / J(\theta^* | \theta^{t-1})}{p(\theta^{t-1}) / J(\theta^{t-1} | \theta^*)}$$

to correct for the asymmetry in the jumping rule.

In the single-component Metropolis-Hastings algorithm the simulated random vector is divided into components or subvectors which are updated one by one. If the jumping distribution for a component is its full conditional posterior distribution, the proposals are accepted with probability one. In the case that all the components are simulated in this way, the algorithm is called a Gibbs sampler. As stated above, in the case of our preliminary model we can simulate all parameters except ϕ directly, and may therefore use a Gibbs sampler with one Metropolis step. As the jumping distribution of ϕ we use the normal distribution $N(\phi^{t-1}, 0.05^2)$. For the final model we use a Gibbs sampler with one Metropolis-Hastings step for $\boldsymbol{\theta}$. The proposal distribution and its acceptance probability were already given in Section 4.

5.2 Empirical results

All the computations in this article were performed and figures produced using the R computing environment (R Development Core Team, 2010). The functions and data needed to replicate the results can be found at http://mtl.uta.fi/codes/mortality. A minor drawback is that we cannot use all available data in estimation but must restrict ourselves to a relevant subset. This is due to the huge matrices involved in computations if many ages and cohorts are included in the data set. For example, if we used our complete data set, whose dimensions are $T = 110$ and $K = 171$, we would have to deal with Kronecker product matrices of dimension 18810×18810. This would require

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

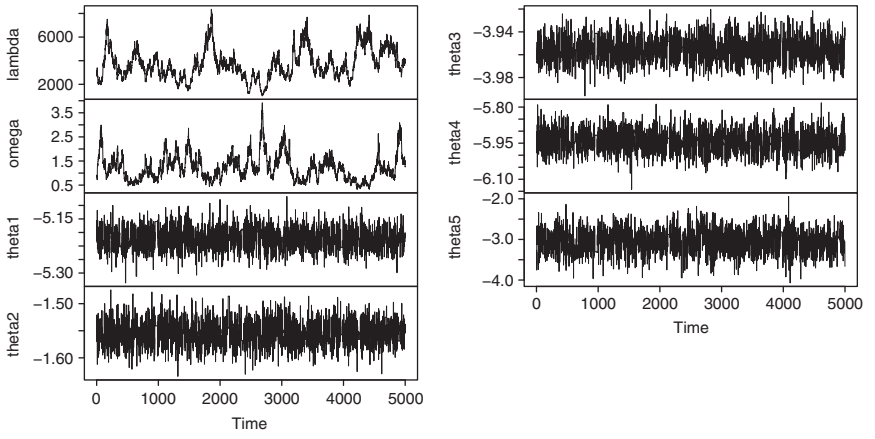

Figure 3. Posterior simulations of the final model.

5 GB of memory for storing one matrix and much more for computations. Although we can alleviate the storage problem and also speed up the computations using sparse matrix methods, we still cannot use the complete data set. In our implementation we use the R package SparseM.

To assess the convergence of the simulated Markov chain to its stationary distribution we used 5 representative values of $\boldsymbol{\theta}$, denoted as $\theta_1,\ldots,\theta_5$, from each corner and the middle of the data matrix, in addition to the upper level parameters. The value θ_5 is in the lower right corner of the matrix and corresponds to an unobserved data item.

For each data set and both models we assessed the convergence of iterative simulation using three simulated sequences with 5000 iterations. In the case of the final model we discarded 1500 first iterations of each chain as a burn-in period, while in the case of the preliminary model the convergence was more rapid and we discarded only 200 iterations.

Figure 3 shows one simulated chain corresponding to the final model and the data set with ages 50–90 and cohorts 1901–1941. The series of λ and ω do not mix well, that is, they are fairly autocorrelated. To obtain accurate results for the estimates, more iterations would be needed. However, the chains converge to their stationary distribution fairly quickly, as indicated by the values of the potential scale reduction factor, which is a convergence diagnostic introduced by Gelman & Rubin (1992). The diagnostic values for the final model are less than 1.1 indicating approximate convergence. Summaries of the estimation results for both preliminary and final model as well as the diagnostics are provided in Appendix 1.

6 Model checking

Cairns *et al.* (2008) provide a checklist of criteria against which a stochastic mortality model can be assessed. We will follow this list as we assess the fit and plausibility of our two models. The list is as follows:

1. Mortality rates should be positive.
2. The model should be consistent with historical data.

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

3. Long-term dynamics under the model should be biologically reasonable.
4. Parameter estimates should be robust relative to the period of data and range of ages employed.
5. Model forecasts should be robust relative to the period of data and range of ages employed.
6. Forecast levels of uncertainty and central trajectories should be plausible and consistent with historical trends and variability in mortality data.
7. The model should be straightforward to implement using analytical methods or fast numerical algorithms.
8. The model should be relatively parsimonious.
9. It should be possible to use the model to generate sample paths and calculate prediction intervals.
10. The structure of the model should make it possible to incorporate parameter uncertainty in simulations.
11. At least for some countries, the model should incorporate a stochastic cohort effect.
12. The model should have a non-trivial correlation structure.

Both of our models fulfil the first item in the list, since we model log death rates. To assess the consistency of the models with historical data we will introduce three Bayesian test quantities in Section 6.1.

A model is defined by Cairns *et al.* (2006a) to be biologically reasonable if the mortality rates are increasing with age and if there is no long-run mean reversion around a deterministic trend. Our spline approach implies that the log death rate increases linearly beyond the estimable region. The preliminary model allows for short-term mean reversion (or autocorrelation) for the observed death rate, while there is no mean reversion at all in the final model.

The fourth and fifth points in the list, that is, the robustness of parameter estimates and model forecasts, will be studied in Sections 6.2 and 6.3. The figures of posterior predictions in Section 6.3 help assess the plausibility and uncertainty of forecasts and their consistency with historical trends and variability.

Implementing the models is fairly straightforward but involves several algorithms. Basically, we use the Gibbs sampler, and supplement it with rejection sampling and Metropolis and Metropolis-Hastings steps, which are needed to update certain parameters or parameter blocks. A further complication is that we have to use sparse matrix methods to increase the maximum size of the data set.

In the Bayesian approach one typically uses posterior predictive simulation, in which parameter uncertainty is taken into account, to generate sample paths and calculate prediction intervals. This will be explained in detail in Section 6.3.

The hierarchical structure of the spline models makes them parsimonious: on the upper level the preliminary model has 4 parameters, the final model only 2. Both models also incorporate a stochastic cohort effect. The preliminary model incorporates an AR(1) structure for observed mortality, while the final model has no correlation structure for deviations from the spline surface.

One should note, however, that the spline model in itself implies a covariance structure. In the one-dimensional case the Bayesian smoothing spline model can be interpreted as a sum of a linear trend and integrated Brownian motion (Wahba, 1978). The prior distribution does not contain

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

information on the intercept or slope of the trend but implies the covariance structure of the integrated Brownian motion. Similarily, in our two-dimensional case, the spline surface can be interpreted as a sum of a plane and deviations from this plane. The conditional prior of $\boldsymbol{\theta}$, given the smoothing parameters, does not include information on the plane but implies a specific spatial covariance structure for the deviations.

6.1 Tests for the consistency of the model

In the Bayesian framework, posterior predictive simulations of replicated data sets may be used to check the model fit (see Gelman *et al.*, 2004). Once several replicated data sets y^{rep} have been produced, they may be compared with the original data set y. If they look similar to y, the model fits.

The discrepancy between data and model can be measured using arbitrarily defined test quantities. A test quantity $T(y,\theta)$ is a scalar summary of parameters and data which is used to compare data with predictive simulations. If the test quantity depends only on data and not on parameters, then it is said to be a test statistic. If we already have N posterior simulations $\theta_i, i = 1,\ldots,N$, we can generate one replication y_i^{rep} using each θ_i, and compute the test quantities $T(y,\theta_i)$ and $T(y_i^{rep},\theta_i)$. The Bayesian p-value is defined to be the posterior probability that the test quantity computed from a replication, $T(y^{rep},\theta)$, will exceed that computed from the original data, $T(y,\theta)$. This test may be illustrated by a scatter plot of $(T(y,\theta_i), T(y_i^{rep},\theta_i))$, $i = 1,\ldots,N$, where the same scale is used for both coordinates. Further details on this approach can be found in Chapter 6 of Gelman *et al.* (2004) or Chapter 11 of Gilks *et al.* (1996).

In the case of our preliminary model, a replication of data is generated as follows: First, $\boldsymbol{\theta}$, σ^2 and ϕ are generated from their joint posterior distribution. Then, using these parameter values, a replicated data vector $\mathbf{y}^{rep}$ is generated from the multivariate normal distribution $N(\theta, \mathbf{I} \otimes \sigma^2\mathbf{P})$. Finally, the elements of $\mathbf{y}^{rep}$ which correspond to the observed values in $\mathbf{y}^{obs}$ are selected. In the case of the final model, $\boldsymbol{\theta}$ is first generated and then the numbers of deaths d_{xt} and exposures e_{xt} are generated recursively by starting from the smallest age included in the estimation data set. The numbers for the smallest age are not generated but they are taken to be the same as in the estimation set. Finally, the replicated death rates are computed as $y_{xt} = \log(d_{xt}/e_{xt})$, and the values corresponding to the observed values in $\mathbf{y}^{obs}$ are selected. Further details about this procedure are provided in Appendix 2.

We introduce three test quantities to check the model fit. The first measures the autocorrelation of the observed log death rate and the second and third its mean square error:

$$AC(y,\boldsymbol{\theta}) = \frac{\sum_{t=t_1}^{t_T} \sum_{x=x_1}^{x_K-1} (y_{x+1,t}-\theta_{x+1,t})(y_{xt}-\theta_{xt})}{\sum_{t=t_1}^{t_T} K_t},$$

where K_t is the number of observations in cohort t, and

$$MSE_1(y,\boldsymbol{\theta}) = \frac{\sum_{t=t_1}^{t_T} \sum_{x=x_1}^{x_{K_t}} (y_{xt}-\theta_{xt})^2}{\sum_{t=t_1}^{t_T} K_t}, \quad MSE_2(y,\boldsymbol{\theta}) = \frac{\sum_{t=t_1}^{t_T} \left(y_{x_{K_t}t}-\theta_{x_{K_t}t}\right)^2}{T}.$$

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

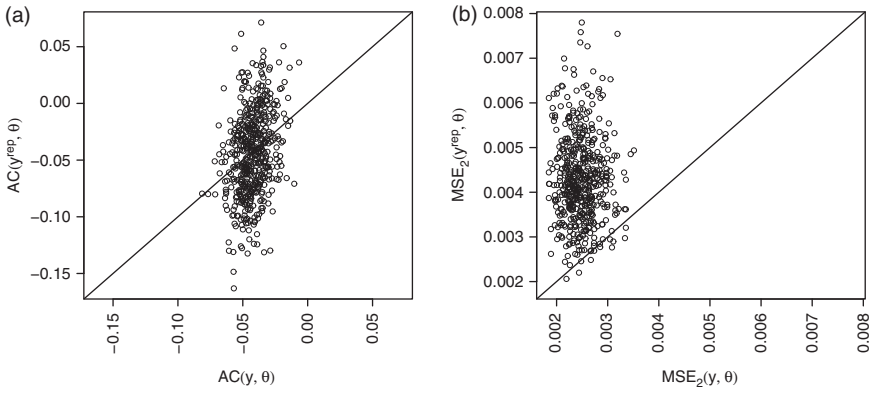

Figure 4. Goodness-of-fit testing for the preliminary model. (a) Autocorrelation test. (b) MSE test.

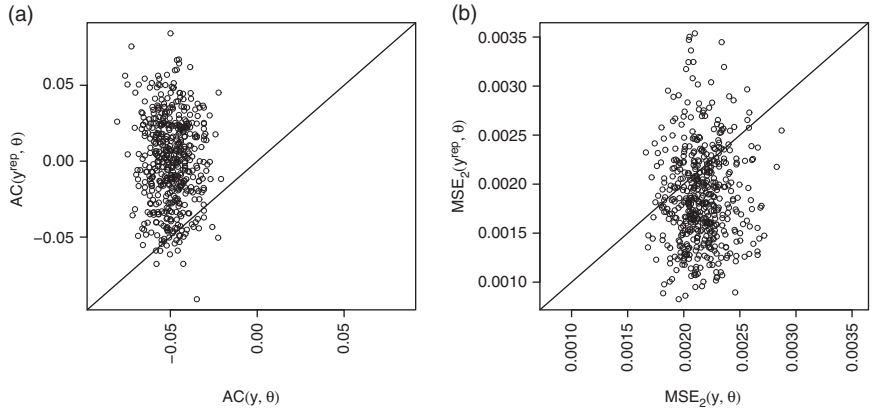

Figure 5. Goodness-of-fit testing for the final model. (a) Autocorrelation test. (b) MSE test.

Figure 4 and 5 show the results when using the data set with ages 50–90 and cohorts 1901–1941. Each figure is based on 500 simulations. If the original data and replicated data were consistent, about half the points in the scatter plot would fall above the 45° line and half below. Figure 4(a) indicates that the preliminary model adequately explains the autocorrelation observed in the original data set, while Figure 5(a) suggests that there might be slight negative autocorrelation in the residuals not explained by the model. However, since the Bayesian p-value, which is the proportion of points above the line, is approximately 0.95, there is no sufficient evidence to reject the assumption of independent Poisson observations.

The test statistic MSE_1 measures the overall fit of the models, and both models pass it (figures not shown). The test statistic MSE_2 measures the fit at the largest ages of the cohorts. From Figure 5(b) we see that the final model passes this test. However, Figure 4(b) suggests that under the preliminary model the MSE_2 simulations based on the original data are smaller than those based on replicated data sets ($p_B = 0.98$). The reason here is that the homoscedasticity assumption of logarithmic mortality data is not valid. The validity of the homoscedasticity and independence assumptions could be further assessed by plotting the standardized residuals (not shown here).

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

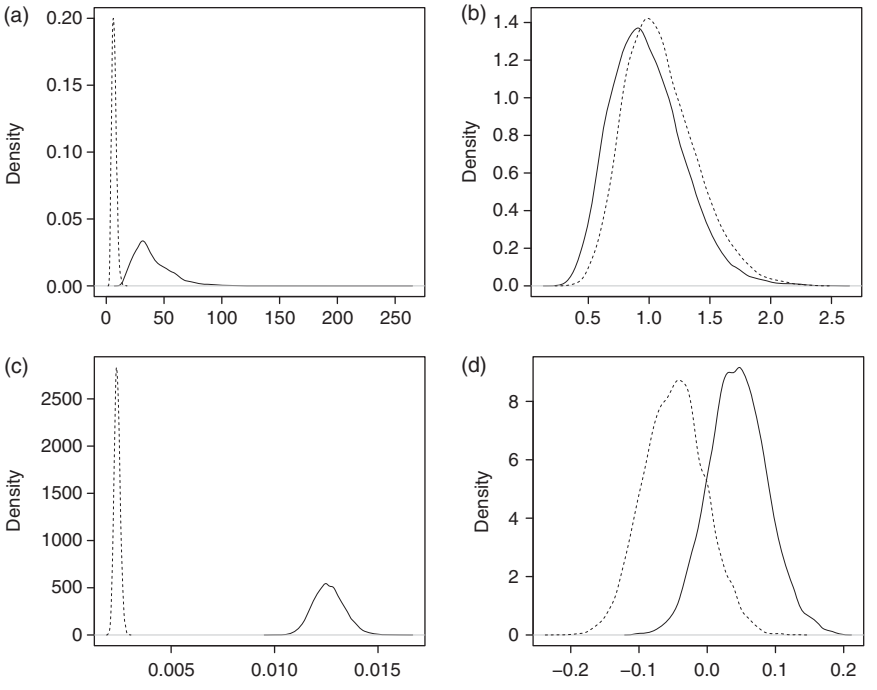

Figure 6. Distributions of (a) λ, (b) ω, (c) σ^2 and (d) ϕ for the preliminary model. The solid line corresponds to the younger (ages 40–70) and the dashed line the older (ages 60–90) age group.

6.2 Robustness of the parameter estimates

The robustness of the parameters may be studied by comparing the posterior distributions when two different but equally sized data sets are used. Here we used two data sets with ages 40–70 and 60–90, and cohorts 1917–1947 and 1886–1916, respectively. We refer to these as the younger and older age groups, respectively. Figure 6 (c) indicates that the variance parameter σ^2 of the preliminary model is clearly higher for the younger age group. This results from the fact that the variance of observed log mortality becomes smaller when the age grows. This also causes a robustness problem for λ, since its posterior distribution is dependent on that of σ^2. Also ϕ seems to be somewhat unrobust.

Figure 7 (a) indicates that under the final model the posterior of λ is more concentrated on small values for the older age group. This is compensated by smaller values of ω for the younger group, which indicates that the smoothing effect in the cohort dimension is similar in both groups. However, the difference between the age groups is not as clear as in the case of the preliminary model. Besides, the range of the distribution is fairly large in both cases.

6.3 Forecasting

Our procedure for forecasting mortality is as follows. We first select a rectangular estimation area which includes in its lower right corner the ages and cohorts for which the death rates are to be predicted. Thus we have in our estimation set earlier observations from the same age as the predicted age and from the same cohort as the predicted cohort. An example of an estimation area is shown in Figure 1.

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

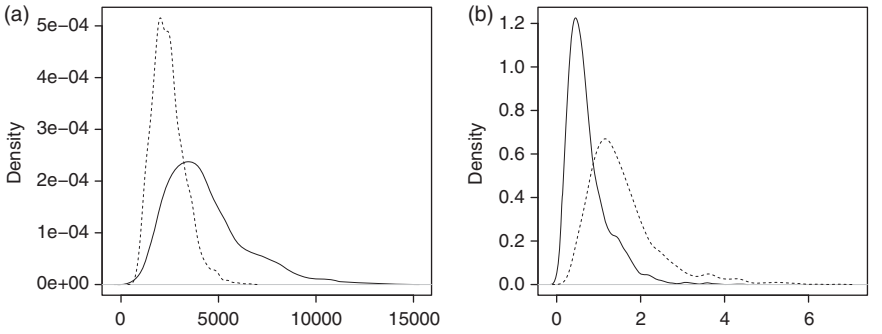

Figure 7. Distributions of (a) λ and (b) ω for the final model. The solid line corresponds to the younger (ages 40–70) and the dashed line the older (ages 60–90) age group.

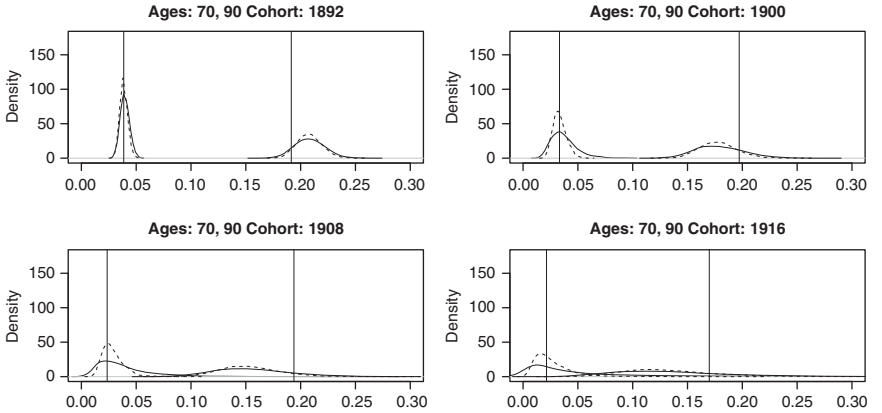

Figure 8. Posterior predictive distributions of the death rates at ages 70 and 90, based on the preliminary model. The solid curves correspond to the larger data set (cohorts 1876–1916, and ages 30–70 when the death rate at age 70 is predicted, and ages 50–90 when the death rate at age 90 is predicted) and the dashed curves the smaller (cohorts 1886–1916, and ages 40–70 when the death rate at age 70 is predicted, and ages 60–90 when the death rate at age 90 is predicted). The vertical lines indicate the realized death rates.

In the Bayesian approach, forecasting is based on the posterior predictive distribution. In the case of our preliminary model, a simulation from this distribution is drawn as follows: First, $\boldsymbol{\theta}$, σ^2 and ϕ are generated from their joint posterior distribution. Then the unobserved data vectors $\mathbf{y}_{j2}$, $j = 1,2,\ldots,T$, (which are to be predicted) are generated from their conditional multivariate normal distributions, given the observed data vectors $\mathbf{y}_{j1}$ and the parameters $\boldsymbol{\theta}$, σ^2 and ϕ. These distributions were provided in Section 3. In the case of our final model, $\boldsymbol{\theta}$ is first generated. Then the numbers of deaths d_{xt} and the exposures e_{xt} are generated recursively starting from the most recent observed values within each cohort. In this way we obtain simulation paths for each cohort and a predictive distribution for each missing value in the mortality table. Further details are provided in Appendix 2.

In studying the accuracy and robustness of forecasts, we use estimation areas similar to those used earlier. However, we choose them so that we can compare the predictive distribution of the death rate with its realized value. The estimation is done as if the triangular area in the right lower corner

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

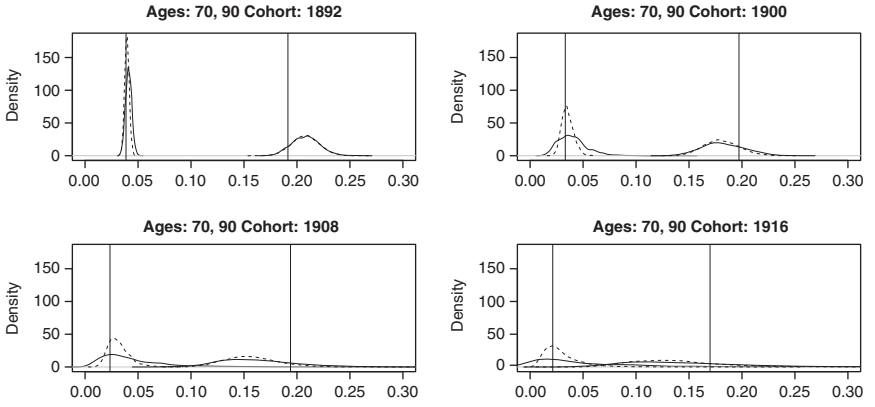

Figure 9. Posterior predictive distributions of the death rates at ages 70 and 90, based on the final model. The solid curves correspond to the larger data set (cohorts 1876–1916; ages 30–70 when the death rate at age 70 is predicted, and ages 50–90 when the death rate at age 90 is predicted) and the dashed curves the smaller (cohorts 1886–1916; ages 40–70 when the death rate at age 70 is predicted, and ages 60–90 when the death rate at age 90 is predicted). The vertical lines indicate the realized death rates.

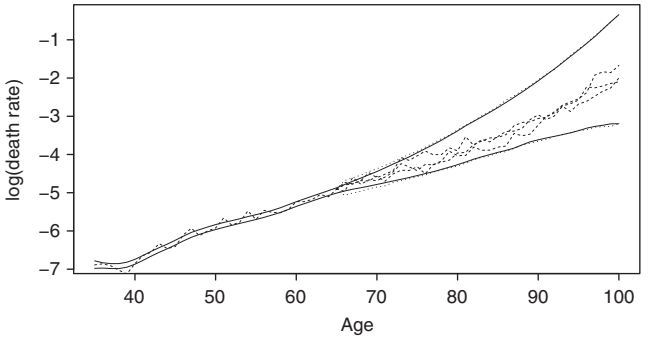

Figure 10. Posterior predictions with the preliminary model for ages 66−100 and cohort 1941. The solid lines represents the 95% posterior limits for θ, the dotted lines the 95% posterior predictive limits for the observed log death rate, and the dashed lines the observed log death rate for ages 35−65 and three predictive paths for ages 66−100.

of the estimation area, indicated in Figure 1, were not known. The posterior predictive distributions shown in Figure 8 are based on the preliminary model, while those in Figure 9 are based on the final model. The four cases in both figures correspond to forecasts 1, 9, 17 and 25 years ahead, for cohorts 1892, 1900, 1908 and 1916, respectively, when the death rate at ages 70 and 90 are forecast. The distributions indicated by solid lines are based on larger estimation sets than those indicated by dashed lines.

It may be seen that increasing uncertainty is reflected by the growing width of the distributions. Furthermore, the size of the estimation set does not considerably affect the distributions when the death rate at age 90 is predicted, while when it is predicted at age 70, the smaller data sets produce more accurate distributions. The obvious reason is that in the latter case the larger estimation set

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

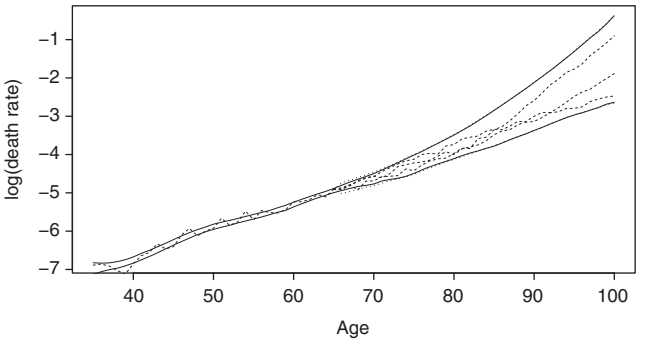

Figure 11. Posterior predictions with the final model for ages 66−100 and cohort 1941. The solid lines represent the 95% posterior limits for θ, the dotted lines the 95% posterior predictive limits for the observed log death rate, and the dashed lines the observed log death rate for ages 35−65 and three predictive paths for ages 66−100.

contains observations from the age interval 30–40 in which the growth of mortality is less regular than at larger ages, inducing more variability in the estimated model. In all cases, the realized values lie within the 90% prediction intervals.

Figure 10 and 11 show posterior predictive simulations for the log death rate when the preliminary and the final model is used, respectively. In each case, the estimation region includes cohorts 1901–1941 and ages 35–100. Three paths of posterior simulations are shown for cohort 1941, for which the data are available until age 65. As may be seen, the variability of the predictions resembles that of the observed path. Furthermore, the 95% posterior limits for the log death rate (θ_{xt}) and the 95% posterior predictive limits for the observed log death rate (y_{xt}) are shown. These two types of limits differ substantially only in the beginning of the forecast horizon. The prediction belt is narrower for the final model, which reflects better model fit.

7 Conclusions

In this article we have introduced a new method to model mortality data in both age and cohort dimensions with Bayesian smoothing splines. The smoothing effect is obtained by means of a suitable prior distribution. The advantage in this approach compared to other splines approaches is that we do not need to optimize with respect to the number of knots and their locations. In order to take into account the serial dependence of observations within cohorts, we use cohort data sets, which are imbalanced in the sense that they contain fewer observations for more recent cohorts. We consider two versions of modelling: first, we model the observed death rates, and second, the numbers of deaths directly.

To assess the fit and plausibility of our models we follow the checklist provided by Cairns *et al.* (2008). The Bayesian framework allows us to easily assess parameter and prediction uncertainty using the posterior and posterior predictive distributions, respectively. In order to assess the consistency of the models with historical data we introduce test quantities. We find that our models are biologically reasonable, have non-trivial correlation structures, fit the historical data well, capture the stochastic cohort effect, and are parsimonious and relatively simple. Our final model

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

has the further advantages that it has less robustness problems with respect to parameters, and avoids the heteroscedasticity of standardized residuals. A further remedy for the unrobustness of the smoothing parameters might be generalizing the model to allow for dependence between these parameters and age.

A minor drawback is that we cannot use all available data in estimation but must restrict ourselves to a relevant subset. This is due to the huge matrices involved in computations if many ages and cohorts are included in the data set. However, this problem can be alleviated using sparse matrix computations. Besides, for practical applications using "local" data sets should be sufficient.

In conclusion, we may say that our final model meets well the mortality model selection criteria proposed by Cairns *et al.* (2008) except that it has a somewhat local character. This locality is partly due to limitations on the size of the estimation set and partly due to slight robustness problems related to the smoothing parameters and forecasting uncertainty.

Acknowledgments

The authors are grateful to referees for their insightful comments and suggestions, which substantially helped in improving the manuscript. The second author of the article would like to thank the Finnish Academy of Science and Letters, Väisälä Fund, for the scholarship during which she could complete this project.

References

Cairns, A.J.G., Blake, D. & Dowd, K. (2006a). Pricing death: Frameworks for the valuation and securitization of mortality risk. *ASTIN Bulletin*, **36**, 79–120.

Cairns, A.J.G., Blake, D. & Dowd, K. (2006b). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance*, **73**, 687–718.

Cairns, A.J.G., Blake, D. & Dowd, K. (2008). Modelling and management of mortality risk: a review. *Scandinavian Actuarial Journal*, **2**, 79–113.

Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D. & Khalaf-Allah, M. (2011). Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin*, **41**, 29–59.

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

Czado, C., Delwarde, A. & Denuit, M. (2005). Bayesian Poisson log-linear mortality projections. *Insurance: Mathematics and Economics*, **36**, 260–284.

Dellaportas, P., Smith, A.F.M. & Stavropoulos, P. (2001). Bayesian analysis of mortality data. *Journal of the Royal Statistical Society. Series A*, **164**, 275–291.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Gelman, A. & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–511.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Gilks, W. R. & Wild, P. (1992). *Applied Statistics*, 41, 337–348.

Gilks, W.R., Richardson, S. & Spiegelhalter, D., eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Green, P. & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). www.mortality.org or www.humanmortality.de [Accessed April, 2009].

Kogure, A. & Kurachi, Y. (2010). A Bayesian approach to pricing longevity risk based on risk-neutral predictive distributions. *Insurance: Mathematics and Economics*, **46**, 162–172.

Lee, R.D. & Carter, L.R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, **87**, 659–675.

Lang, S. & Brezger, A. (2004). Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Pedroza, C. (2006). A Bayesian forecasting model: predicting U.S. male mortality. *Biostatistics*, **7**, 530–550.

Plat, R. (2009). On stochastic mortality modeling. *Insurance: Mathematics and Economics*, **45**, 393–404.

Reichmuth, W. & Sarferaz, S. (2008). Bayesian demographic modelling and forecasting: An application to US mortality. SFB 649 Discussion paper 2008–052.

Renshaw, A.E. & Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, **38**, 556–570.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Richards, S.J., Kirkby, J.G. & Currie, I.D. (2006). The importance of year of birth in two-dimensional mortality data. *British Actuarial Journal*, **12**, 5–38.

Schmid, V.J. & Held, L. (2007). Bayesian age-period-cohort modeling and prediction – BAMP. *Journal of Statistical Software*, **21**, 8. http://www.jstatsoft.org [Accessed April 2012].

Wahba, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society B*, **40**, 364–372.

Wilmoth, J.R., Andreev, K., Jdanov, D. & Glei, D.A. (2007). Methods Protocol for the Human Mortality Database. http://www.mortality.org [Accessed April 2009].

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

Appendix 1

The posterior simulations were performed using the R computing environment. The following outputs were obtained using the summary function of the add-on package MCMCpack:

Table 1. Estimation results of the preliminary mortality model.

```
Number of chains = 3
Sample size per chain = 4800

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

            Mean        SD  Naive SE Time-series SE
lambda 16.940473 5.0795384 4.233e-02      4.001e-01
omega   1.010386 0.2781601 2.318e-03      1.765e-02
sigma2  0.004243 0.0001720 1.433e-06      3.552e-06
phi    -0.047612 0.0295280 2.461e-04      6.206e-04
theta1 -5.163545 0.0335889 2.799e-04      4.310e-04
theta2 -1.552719 0.0331673 2.764e-04      2.790e-04
theta3 -3.958345 0.0121454 1.012e-04      1.156e-04
theta4 -5.903402 0.0333465 2.779e-04      2.878e-04
theta5 -3.099490 0.2857720 2.381e-03      3.118e-03

2. Quantiles for each variable:

             2.5%       25%       50%       75%     97.5%
lambda  8.643654 13.198089 16.351764 20.291032 28.209809
omega   0.551153  0.814618  0.979278  1.175660  1.634720
sigma2  0.003923  0.004125  0.004237  0.004355  0.004597
phi    -0.104964 -0.067280 -0.047333 -0.027680  0.011365
theta1 -5.229142 -5.186360 -5.163524 -5.141287 -5.097599
theta2 -1.617886 -1.574978 -1.552572 -1.530468 -1.488354
theta3 -3.982101 -3.966441 -3.958372 -3.950303 -3.934482
theta4 -5.969645 -5.925643 -5.903172 -5.880918 -5.837735
theta5 -3.659055 -3.288343 -3.100254 -2.910501 -2.526937



Potential scale reduction factors:

       Point est. 97.5% quantile
lambda       1.02            1.08
omega        1.03            1.08
sigma2       1.00            1.01
phi          1.00            1.01
theta1       1.00            1.00
theta2       1.00            1.00
theta3       1.00            1.00
theta4       1.00            1.00
theta5       1.00            1.00

Multivariate psrf

1.02
```


https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

Table 2. Estimation results of the final mortality model.

```
Number of chains = 3
Sample size per chain = 3500

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

           Mean        SD  Naive SE Time-series SE
lambda 4069.801 1.325e+03 1.293e+01      9.781e+01
omega     1.008 4.595e-01 4.485e-03      3.461e-02
theta1   -5.173 3.929e-02 3.834e-04      1.029e-03
theta2   -1.553 2.365e-02 2.308e-04      5.993e-04
theta3   -3.955 1.054e-02 1.029e-04      2.501e-04
theta4   -5.888 4.601e-02 4.490e-04      1.155e-03
theta5   -3.081 2.822e-01 2.754e-03      7.924e-03

2. Quantiles for each variable:

            2.5%       25%      50%      75%    97.5%
lambda 1909.1259 3187.9277 3839.586 4794.966 7185.706
omega     0.3991    0.6733    0.923    1.245    2.160
theta1   -5.2483   -5.1999   -5.173   -5.148   -5.096
theta2   -1.5975   -1.5692   -1.553   -1.537   -1.506
theta3   -3.9751   -3.9614   -3.955   -3.947   -3.933
theta4   -5.9788   -5.9185   -5.886   -5.858   -5.798
theta5   -3.6424   -3.2617   -3.076   -2.902   -2.520


Potential scale reduction factors:

       Point est. 97.5% quantile
lambda       1.09            1.27
omega        1.08            1.25
theta1       1.01            1.02
theta2       1.00            1.00
theta3       1.00            1.00
theta4       1.00            1.01
theta5       1.01            1.04

Multivariate psrf

1.09
```

Appendix 2

In the case of the final model, the numbers of deaths d_{xt} and the exposures e_{xt} should be forecast for the ages and cohorts for which they are unknown. Furthermore, these values should be generated when replications of the original estimation data set are produced.

In the case of forecasting, we use an iterative procedure to generate d_{xt} and e_{xt}, starting from the most recent observation of death rate within each cohort. In the case of data replication, we start from the smallest age available in the data set. In each case, the initial cohort size is estimated on the basis of the relationship

$$q_{xt} = 1 - \exp(-\mu_{xt}),$$

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press

where q_{xt} is the probability that a person in cohort t dies at age x. The same equality applies for the maximum likelihood estimates of q_{xt} and μ_{xt}, given by $\hat{q}_{xt} = d_{xt} / n_{xt}$ and $m_{xt} = d_{xt}/e_{xt}$, where n_{xt} is the number of persons reaching age x in cohort t. Thus, we obtain the formula

$$\frac{d_{xt}}{n_{xt}} = 1 - \exp\left(-\frac{d_{xt}}{e_{xt}}\right), \tag{17}$$

from which we may solve n_{xt} when d_{xt} and e_{xt} are known.

Further, the number of persons alive is updated recursively as $n_{x+1,t} = n_{xt} - d_{xt}$, and the number of deaths is generated from the binomial distribution:

$$d_{x+1,t} \sim \text{Bin}\left(n_{x+1,t}, q_{x+1,t}\right).$$

Then $e_{x+1,t}$ is solved using (17).

https://doi.org/10.1017/S174849951200005X Published online by Cambridge University Press