ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)

A Bayesian approach to developing a stochastic mortality model for China

Johnny Siu-Hang Li,

University of Waterloo, Canada, and University of Melbourne, Australia

Kenneth Q. Zhou and Xiaobai Zhu,

University of Waterloo, Canada

Wai-Sum Chan

Chinese University of Hong Kong, People's Republic of China

and Felix Wai-Hon Chan

University of Hong Kong, People's Republic of China

[Received July 2017. Final revision 2019]

Summary. Stochastic mortality models have a wide range of applications. They are particularly important for analysing Chinese mortality, which is subject to rapid and uncertain changes. However, owing to data-related problems, stochastic modelling of Chinese mortality has not been given adequate attention. We attempt to use a Bayesian approach to model the evolution of Chinese mortality over time, taking into account all of the problems associated with the data set. We build on the Gaussian state space formulation of the Lee–Carter model, introducing new features to handle the missing data points, to acknowledge the fact that the data are obtained from different sources and to mitigate the erratic behaviour of the parameter estimates that arises from the data limitations. The approach proposed yields stochastic mortality forecasts that are in line with both the trend and the variation of the historical observations. We further use simulated pseudodata sets with resembling limitations to validate the approach. The validation result confirms our approach's success in dealing with the limitations of the Chinese mortality data.

Keywords: Lee–Carter model; Multiple imputation; Sampling uncertainty; Sequential Kalman filter

1. Introduction

Since the reform and opening-up policy was implemented, China has made great strides in improving longevity. According to the World Bank, the life expectancy at birth of the unisex population of China has increased from 65.5 years in 1978 (when the reform and opening-up began) to 76.0 years in 2015, representing an average increase of 2.8 years per decade. The rising trend in life expectancy is welcomed by most people in the country where the pursuit of longevity is strongly embedded in the culture, but the uncertainty about how the trend may continue poses huge challenges to the government and many others.

Fuelled by the infamous one-child policy, improved longevity has accelerated the aging of the

Address for correspondence: Johnny Siu-Hang Li, Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada.
E-mail: shli@uwaterloo.ca

 0964–1998/20/183000

Chinese population, pushing China to the Lewis turning point (Lewis, 2013) at which labour demand outgrows labour supply. From a macroeconomic standpoint, the evolution of the trend in life expectancy affects the timing of this turning point, which in turn influences how much longer China can enjoy its 'demographic dividends' (output and other economic gains) resulting from a large proportion of the total population being in the working-age group (Cai, 2010; Peng, 2011). Further, as China is the most populous country in the world, its demographic transition has a strong influence on the supply of (low cost) labour and hence affordable consumer goods from a global perspective.

As outlined in the 1997 State Council Document 26, a significant part of the urban pension system in China is a defined benefit public plan, which comprises a pay as you go portion financed by employer contributions equal to 20% of wages plus a funded portion supported by employee contributions equal to 8% of wages. According to the Ministry of Human Resources and Social Security, this public plan had assets of ¥3993.7 billion at 2013 year end. Although this amount seems large, whether it is sufficient to cover the promised benefits depends very much on how mortality at pensionable ages changes. A faster-than-expected improvement in longevity may necessitate an increase in contribution rates in the future.

In the private sector, annuity products offering lifetime incomes have become increasingly popular in China, as evidenced by the remarkable increase in the total annuity benefit payout from ¥14.0 billion in 2010 to ¥21.5 billion in 2012 (China Insurance Regulatory Commission, 2011, 2013). The uncertainty surrounding the evolution of mortality affects systematically all life annuities that are sold in China and, in extreme circumstances, may result in a systemic failure of the Chinese insurance industry (Basel Committee on Banking Supervision, 2013). Recently, the Organisation for Economic Co-operation and Development (2014) has advocated using mortality-linked derivatives to manage the risk, but this solution requires both insurers and capital market investors to have a better understanding of the uncertainty surrounding the trend in life expectancy.

Policy makers and risk managers can be better prepared for these demographic headwinds with the aid of stochastic mortality models, which produce a best estimate forecast of future mortality and a range of possible deviations from the best estimate forecast. Although its practical relevance is clear, the development of stochastic mortality models for China has not been given much attention. Many of the existing studies of Chinese mortality, including those conducted by Banister and Hill (2004), Zhao (2012) and Zhao *et al.* (2013), focus primarily on (deterministically) trending past observations. Only a handful of attempts have been made to build stochastic mortality models for China, and, as explained later in this section, they are subject to some significant limitations.

We believe that the lack of studies of the stochastic modelling of Chinese mortality is due mainly to several data-related problems, which we now summarize. First, age- and gender-specific mortality data (death and exposure counts) for the population of China are not available for the pre-1981 period, leaving researchers with a rather short series of data to work with. Second, within the 1981–2014 period (the data series ends in 2014), a few years of data are either partially or completely missing, so that the already short data series is not continuous. Third, for a number of age–time cells (e.g. age 7 years, 2008, males) in the data set, exposure counts are provided but no death count is reported. Fourth, the source of data is inhomogeneous such that some data (those for 1981, 1989, 2000 and 2010) are obtained from nationwide censuses, whereas the rest are obtained from surveys of a fraction (1% or 0.1%) of the national population. As a consequence, the variation in the observed death rates is far from being constant.

Given the unique properties of the available data, developing stochastic mortality models for China is not only a practically relevant problem but also a methodological challenge. The goal

of this paper is to build a stochastic mortality model for China with all of these properties being accounted for. In particular, we believe that the model should meet the following three criteria.

(a) *Criterion 1: the model should exploit as much information as possible from the data.* Given that mortality projections are often made decades into the future and that the data series is quite short (34 years), none of the available data should be discarded. That said, the discontinuities in the data series (arising from the missing data) should be taken into account when estimating the model.

(b) *Criterion 2: the model should provide appropriate provisions for uncertainty.* Owing to the limited availability of data, parameter uncertainty tends to be significant and hence must be included in the resulting measures of forecast uncertainty. In particular, the additional parameter uncertainty arising from missing data should be incorporated during the course of estimation. Furthermore, although the survey data should be included, the fact that they are subject to additional sampling uncertainty must be acknowledged in the model structure.

(c) *Criterion 3: the model should be parsimonious and yield biologically reasonable projections.* Owing again to the limited availability of data, we prefer parsimonious models to sophisticated models. For instance, given only 34 years of data, at best we can only observe approximately 30% of a birth cohort (let alone the discontinuities in the observations arising from missing data), and hence it is difficult to justify using models with cohort effects such as the Renshaw–Haberman model (Renshaw and Haberman, 2006). Furthermore, as we aim to estimate the model to the full age range of 0–99 years, the model should be able to capture the fall and rise in mortality with age during childhood and adulthood years respectively. For this reason, we do not consider models such as the Cairns–Blake–Dowd model (Cairns *et al.*, 2006), which are applicable only to a restricted age range.

We now review the previous attempts to model the Chinese mortality data set, and we explain why they do not satisfy all of the criteria that we set out. For convenience of exposition, we divide the previous attempts into three broad categories.

The first category includes methods based on a subset of the available data in which there are no missing values. Methods that fall into this category include the pioneering work of Li *et al.* (2004), who estimated the Lee–Carter model (Lee and Carter, 1992) by using three years of complete mortality data that were obtained from nationwide censuses (those for 1989, 2000 and 2010). Despite being quite straightforward to implement, this approach does not satisfy criterion 1, because it completely disregards the survey data, which contain valuable information about the population (Lavely and Mason, 2006). In principle, the method that was proposed by Li *et al.* (2004) can be extended to incorporate the survey data (whereby some death counts are not reported) by using weighted least squares with a zero weight being assigned to the age–time cells with no reported death count (Wilmoth, 1993), but such an extension does not acknowledge the fact that the survey data are subject to greater sampling uncertainty. Furthermore, the work of Li *et al.* (2004) takes no account of parameter uncertainty and therefore does not meet criterion 2. This category of methods has been considered by other researchers including Jiang *et al.* (2013), who also used three years of census data, Huang and Browne (2017), who considered only the data from 1997 to 2011, and Wang and Huang (2011), who used only the data from 1994 to 2008. Their contributions also do not meet criterion 1.

The second category encompasses methods in which the missing values are filled with a *single* collection of proxies. This category is exemplified by the work of Ping *et al.* (2013), who fitted a multiple linear regression to the available data and then filled the missing values with the expected values implied by the fitted regression to obtain a *single* 'complete' data set. This

method fails to meet criterion 2, as it does not capture parameter uncertainty including the additional uncertainty that arises from the missing data. In part because of the understatement of uncertainty, Schafer and Graham (2002) argued that *ad hoc* methods such as that of Ping *et al.* (2013) may 'do more harm than good'. It is noteworthy that Ping *et al.* (2013) estimated their model with a two-stage approach, in which a singular value decomposition is applied to the 'complete' data set to obtain an estimate of the structural parameters, and then a random walk with drift is estimated to the time-varying indices for forecasting. Such a two-stage estimation approach for estimating stochastic mortality models has been criticized heavily by researchers including Leng and Peng (2016), who pointed out its inference pitfalls, and Czado *et al.* (2005), who asserted that it may lead to incoherence.

The third category consists of methods that borrow information from the historical mortality experience of other populations. One example is the contribution by Li (2014), who framed the method that was proposed by Li *et al.* (2004) in a Bayesian setting that allows the resulting mortality forecasts to be influenced by the information from another population. Specifically, Li (2014) adjusted the prior distribution of the volatility parameter in the random walk for the time-varying index in such a way that the resulting 'volatility-to-drift' ratio lines up with that of the reference population. Although this extension may yield more reasonable prediction intervals, it still retains some of the limitations of the original work of Li *et al.* (2004) (e.g. not satisfying criterion 1). Another example is the work of Li, Reuser, Kraus and Alho (2009), who developed a stochastic population forecast (which involves a stochastic mortality forecast) for China by using demographic information from European countries. The biggest problem with this category of methods is that it is difficult to justify why the demographic information of another population should (and can) be borrowed. This problem is particularly relevant to the modelling work for China, which has a distinct history of public health and economic development.

In this paper, we adopt a Bayesian approach to model stochastically the evolution of Chinese mortality. Bayesian multiple imputation is used to handle the discontinuities in the data set, so that all of the available data can be incorporated in the model, and therefore criterion 1 can be met. With Gibbs sampling, the approach proposed yields a joint posterior distribution of *all* the model parameters, which enables the user to gauge holistically the level of parameter uncertainty including the extra portion that arises from the missing data, so that criterion 2 can be satisfied. Although several Bayesian approaches to mortality forecasting are available (Czado *et al.*, 2005; Pedroza, 2006; Kogure *et al.*, 2009; Cairns *et al.*, 2011; Li, 2014; Girosi and King, 2008), we choose to follow the framework of Pedroza (2006) for two reasons:

(a) Pedroza's work is built on the Lee–Carter model, which appears to meet criterion 3 as it is relatively parsimonious compared with other models that are applicable to the full age range of 0–99 years;
(b) its Gaussian state space specification facilitates efficient estimation, as it enables us to obtain the conditional posterior distributions of the underlying latent factors (the time-varying indices) in the model with Kalman filtering and smoothing (Kim and Nelson, 1999).

Although we draw heavily from the work of Pedroza (2006), some significant adaptations are made to suit the situation that we are confronting.

First, we adapt the multiple-imputation algorithm that was proposed by Pedroza (2006). Pedroza (2006) and a few other researchers (e.g. Czado *et al.* (2005) and Cairns *et al.* (2011)) have pointed out that Bayesian approaches can be used to handle missing data in the context of stochastic mortality modelling, but none of them have applied these approaches to a real

mortality data set that is subject to problems that are similar to those we are facing. In a preliminary study, we find that the estimation algorithm of Pedroza (2006) simply does not converge when it is applied to the Chinese mortality data set. The non-convergence problem was also noted by Li (2014), who mentioned that

> 'if the missing data are treated as variables, the number of variables involved increases significantly, and it becomes very difficult for the simulation process to reach convergence'.

For this reason, he resorted to using only three years of data and (subjectively) adjusting the understated forecast uncertainty, instead of adopting a more rigorous approach whereby a joint posterior distribution of all the parameters is produced. When the adapted multiple-imputation algorithm is implemented, the typical version of the Kalman filter is no longer applicable. To overcome this technical challenge, we replace it with the sequential Kalman filter that was proposed by Koopman and Durbin (2000) and reformulate the model in a sequential representation accordingly.

Second, we introduce a Bayesian version of the cubic *B*-spline function that was considered by Renshaw and Haberman (2003) to smooth the pattern of the age response parameters (denoted by β_x in this and many other papers), which determine the expected rates of improvement in mortality at different ages. Although Bayesian formulations assume some sort of smoothness of age and time effects (Czado *et al.*, 2005), we find that the Bayesian estimates of β_xs for the Chinese population are very jagged across ages if additional smoothing is not imposed (see Section 6.3). This outcome is possibly because the data array that is used is small and has many discontinuities. A jagged pattern of β_xs is undesirable, because, for example, it is difficult to explain why mortality rates at adjacent ages evolve at highly different expected speeds. The adaptation that we propose ensures that the posterior mean of β_x follows a logical relationship with age, allowing the resulting model to meet criterion 3 better (biological reasonableness).

Third, we add a feature to address the fact that the Chinese mortality data are obtained from different sources. We find that, if this fact is not addressed, the resulting prediction intervals become erroneously wide and are therefore of limited usefulness (criterion 2 is not met) (see Section 6.3). The added feature permits the variance of the error term in the observation equation of the Lee–Carter model to be time inhomogeneous, taking one of the three possible values depending on the data for the year that is associated with the error term obtained from a nationwide census, a survey of 1% of the population or a survey of 0.1% of the population. The necessary modifications to the conditional posterior distributions and the sequential Kalman filter arising from this added feature are discussed and implemented.

When the full model with all the adaptations is applied to the Chinese mortality data set, the estimation algorithm converges fairly quickly and the resulting posterior distributions of the model parameters appear to be reasonable. Using the estimated full model, we generate forecasts of log(central death rates) over a horizon of 35 years. For every age, the central prediction forms a logical extension from the observed values, and the measure of forecast uncertainty is commensurate with the historical variability. We further use pseudodata sets to validate the ability of the proposed method for handling the data-related problems. The pseudodata sets are randomly generated from some assumed Lee–Carter parameters and are constructed in such a way that they have the mentioned limitations of the actual Chinese mortality data set. When applied to the pseudodata sets, the modelling and estimation approaches proposed produce parameter estimates that are sufficiently close to the parameters in the underlying data-generating process, confirming the approach's success in handling the data-related problems that we are confronting.

The remainder of this paper is organized as follows. Section 2 describes the Chinese mortality data set in more detail and explains how it is connected to the missing data mechanisms that were outlined in the work of Rubin (1976) and Mealli and Rubin (2015). Section 3 briefly reviews the work of Pedroza (2006). Section 4 presents our proposed approach to modelling Chinese mortality, with detailed descriptions of our technical innovations. Sections 5 and 6 discuss the estimation and prediction results, and demonstrate the importance of the adaptations that are made to the work of Pedroza (2006). Section 7 details the validation with pseudodata sets. Finally, concluding remarks and suggestions for future research are given in Section 8.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/1467985x/series-a-datasets`

2. The available data

The mortality data (age-specific death and mid-year population counts) that are used in this paper were obtained from the 1988–2015 China *Population and Employment Statistics Yearbooks*, issued by the National Bureau of Statistics of China. The data span an age range of 0–99 years and a time period of 1981–2014, covering 3400 age–time cells. However, as described below, the data for some of the age–time cells within this age range and time period are not available.

(a) No mortality data are available for 1982–1985, 1987, 1988 and 1990–1993.
(b) For 1989, 1994, 1997–1999, 2001–2004 and 2006–2009, the mortality data above age 89 years are unavailable; for 1996, the mortality data above age 85 years are unavailable.
(c) No death count is reported for some individual age–time cells (e.g. age 7 years, 2008, males).

Overall, there are 1197 and 1229 age–time cells in the data set for males and females respectively, with no death and/or population count, representing 35.21% and 36.15% of the total number of age–time cells respectively.

In addition to incompleteness, the available data are not always derived from the entire national population. Only the data for 1981, 1989, 2000 and 2010 are based on nationwide censuses. For 1986, 1995 and 2005 (midway between censuses), the data are based on surveys of 1% of the national population. For all other years, the data are based on surveys of 0.1% of the national population. The characteristics of the data set are summarized in Fig. 1.

For items (a) and (b), the missingness is due entirely to the fact that data were not collected, and is treated as missingness completely at random in this paper. This treatment is justified by the fact that the missingness is entirely under the data collector's control (Graham *et al.* (2006), page 324), and the fact that the missing data were intentionally not recorded (i.e. planned missingness) (Schafer and Graham, 2002).

For item (c), the reason behind the unreported death counts is not documented in the yearbooks. Without knowing the exact reason for the missingness, we believe that it is prudent to treat item (c) as missingness at random. One may, however, argue that the missingness for item (c) is a consequence of finite sampling. When being drawn from only 0.1% of the national population, the exposure count in each age–time cell tends to be small, which may in turn lead to no realized death in certain age–time cells. If this argument is true, then missingness at random is not a perfect fit for item (c) as the missingness may be correlated with age and/or exposure and is not completely out of the control of the data collector. The available information does not

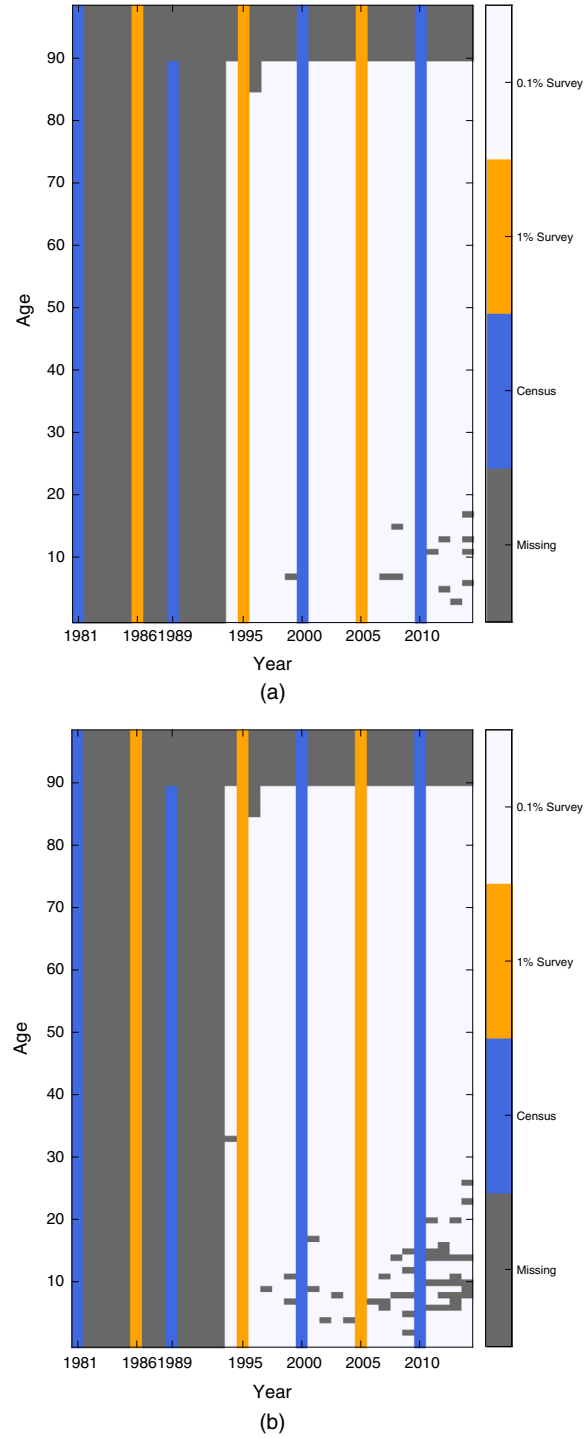


Fig. 1. Lexis diagrams summarizing the availability of mortality data for (a) Chinese males and (b) Chinese females: ▮, data obtained from censuses; ▮, data obtained from 1% surveys; ▯, data obtained from 0.1% surveys; ▮, missing data

permit us to prove or disprove this argument, and in this regard, the missingness mechanism for item (c) should be viewed as an assumption.

3. Review of Pedroza's (2006) approach

3.1. Lee–Carter model in a Gaussian formulation

The work of Pedroza (2006) is based on the Lee–Carter model (Lee and Carter, 1992), which can be expressed as follows:

$$\ln(m_{x,t}) = \alpha_x + \beta_x \kappa_t + \epsilon_{x,t}, \tag{1}$$

where $m_{x,t}$ is the central death rate for individuals aged x in year t, α_x represents the average level of mortality at age x, κ_t is a time-varying index governing the evolution of $\ln(m_{x,t})$ over time, β_x measures the sensitivity of $\ln(m_{x,t})$ to κ_t and $\epsilon_{x,t}$ is the error term, which has no correlation across both age and time.

Following the original work of Lee and Carter (1992), it is assumed that $\{\kappa_t\}$ follows a random walk with drift:

$$\kappa_t = \mu + \kappa_{t-1} + \omega_t, \tag{2}$$

where μ is the drift term and $\{\omega_t\}$ is a sequence of independent and identically distributed normal random variables with a zero mean and a variance of σ^2. A non-stationary process is used instead of a stationary process for two main reasons.

(a) If a stationary (mean-reverting) process is used, the expected trajectory of future mortality would revert to a certain (average) historical level. Such a trajectory seems counterintuitive.
(b) In demographic forecasting, it is conceivable that, when we look further into the future, we are more uncertain about what a demographic quantity may turn out to be. However, if a stationary process is used, the forecast uncertainty would not grow with the forecast horizon.

It is well known that the Lee–Carter model is subject to an identifiability problem. To stipulate parameter uniqueness, the following constraints are used:

$$\begin{aligned} \sum_{x=x_0}^{x_0+n_a-1} \beta_x &= 1, \\ \sum_{t=t_0}^{t_0+n_y-1} \kappa_t &= 0, \end{aligned} \tag{3}$$

where x_0 and t_0 are the youngest age and the earliest year covered by the data set respectively, and n_a and n_y are the numbers of ages and years covered by the data set respectively. For the data set that is used in this paper, $x_0 = 0$, $t_0 = 1981$, $n_a = 100$ and $n_y = 34$.

Pedroza (2006) chose to use a Gaussian formulation, meaning that the error term $\epsilon_{x,t}$ in equation (1) is assumed to follow a normal distribution with a zero mean and a variance of s^2. Although other distributional assumptions such as Poisson (Wilmoth, 1993; Brouhns *et al.*, 2002) or negative binomial (Li, Hardy and Tan, 2009) may be used instead, the Gaussian formulation is advantageous in the context of our research problem for the following reasons.

First, the Gaussian formulation permits us to use Gibbs sampling, with which samples of each model parameter can be directly drawn from its conditional posterior distribution, provided that an appropriate conjugate prior distribution is assumed. If a non-Gaussian formulation is used,

then more computationally intensive methods such as the Metropolis–Hastings algorithm are required.

Second, the Gaussian formulation implies that $\ln(m_{x,t})$ is normally distributed, thereby avoiding the possibility of imputing zero death rates (which will lead to a logarithm of 0 problem). If the Poisson formulation is used instead, the imputed death rate for a blank age–time cell may be 0, as a zero death count may be realized.

Third, the Gaussian formulation allows us to express the entire model (equations (1) and (2)) in a state space form, which can be adapted readily to suit the characteristics of the data set that we consider.

3.2. Gibbs sampling

The Gaussian Lee–Carter model contains $2n_a + n_y + 3$ parameters in total, including $\alpha_{x_0}, \ldots, \alpha_{x_0+n_a-1}$, $\beta_{x_0}, \ldots, \beta_{x_0+n_a-1}$, $\kappa_{t_0}, \ldots, \kappa_{t_0+n_y-1}$, μ, s^2 and σ^2. The primary objective of Gibbs sampling is to obtain the joint posterior distribution of all $2n_a + n_y + 3$ parameters.

We let $\mathbf{Y}$ be an $n_a \times n_y$ matrix containing the historical central death rates, and

$$\Theta = \{\theta_1, \theta_2, \ldots, \theta_{2n_a+n_y+3}\}$$

be the set of all of the parameters in the model. We further use $\Theta_{-\theta_j}$, $j = 1, \ldots, 2n_a + n_y + 3$, to represent Θ excluding its jth entry. In each iteration of Gibbs sampling, a sample of θ_j is drawn from its (full) conditional posterior distribution $\pi(\theta_j|\mathbf{Y}, \Theta_{-\theta_j})$. The process is repeated for each $j = 1, \ldots, 2n_a + n_y + 3$, yielding a realization of the joint posterior distribution of Θ (i.e. $\pi(\Theta|\mathbf{Y})$). After a large number of iterations, an empirical joint posterior distribution of Θ is obtained.

The conditional posterior distribution of θ_j depends on the assumed conditional prior distribution $\pi(\theta_j|\Theta_{-\theta_j})$. For α_xs, β_xs, μ, s^2 and σ^2, improper prior distributions are assumed, i.e. $\pi(\theta_j|\Theta_{-\theta_j}) \propto 1$ for α_xs, β_xs and μ, and $\pi(\theta_j|\Theta_{-\theta_j}) \propto \theta_j^{-1}$ for s^2 and σ^2. The derivation of the conditional posterior distributions of these parameters can be found in Pedroza (2006). For κ_ts, the conditional posterior distributions are obtained with a Kalman filter (Harvey, 1991), which is detailed in the next subsection.

At the end of each Gibbs sampling iteration, parameters α_x, β_x and κ_t are rescaled so that the two constraints that are specified in equation (3) are satisfied. In principle, it is possible to include the constraints in the prior distributions for κ_t and β_x. However, in an application that involves a large volume of missing data, this alternative method (which is explained in the on-line annex F) would result in numerical instability.

3.3. Kalman filter

Given a sample of α_xs, β_xs, μ, s^2 and σ^2, a sample of κ_ts can be generated by using a Kalman filter. To implement a Kalman filter, we first rewrite the model (equations (1) and (2)) in a state space form as follows: observation equation,

$$\mathbf{y}_t = \boldsymbol{\alpha} + \boldsymbol{\beta}\kappa_t + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathrm{MVN}(0, s^2\mathbf{I}); \tag{4}$$

state equation,

$$\kappa_t = \mu + \kappa_{t-1} + \omega_t, \qquad \omega_t \sim N(0, \sigma^2), \tag{5}$$

where $\mathbf{y}_t = (y_{x_0,t}, \ldots, y_{x_0+n_a-1,t})' = (\ln(m_{x_0,t}), \ldots, \ln(m_{x_0+n_a-1,t}))'$, $\boldsymbol{\alpha} = (\alpha_{x_0}, \ldots, \alpha_{x_0+n_a-1})'$, $\boldsymbol{\beta} = (\beta_{x_0}, \ldots, \beta_{x_0+n_a-1})'$, $\boldsymbol{\epsilon}_t = (\epsilon_{x_0,t}, \ldots, \epsilon_{x_0+n_a-1,t})'$ and $\mathbf{I}$ is an $n_a \times n_a$ identity matrix. The Kalman filter can be separated into two parts: the filtering process and the smoothing process.

3.3.1. The filtering process

The filtering process predicts and updates κ_t on the basis of $\mathbf{Y}_t = (\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t_0})$, which contains all information up to and including time t. We let $k_t := E(\kappa_t|\mathbf{Y}_t)$ and $p_t := \mathrm{var}(\kappa_t|\mathbf{Y}_t)$. The filtering process derives k_t and p_t by using the following recursion, which runs from $t = t_0$ to $t = t_0 + n_y - 1$:

$$\begin{aligned} k_t &= \mu + k_{t-1} + \mathbf{g}_t \boldsymbol{\eta}_t, \\ p_t &= (1 - \mathbf{g}_t \boldsymbol{\beta})(p_{t-1} + \sigma^2), \\ \boldsymbol{\eta}_t &= \mathbf{y}_t - \boldsymbol{\alpha} - \boldsymbol{\beta}(\mu + k_{t-1}), \\ \mathbf{F}_t &= \boldsymbol{\beta}(p_{t-1} + \sigma^2)\boldsymbol{\beta}' + s^2\mathbf{I}, \\ \mathbf{g}_t &= (p_{t-1} + \sigma^2)\boldsymbol{\beta}'\mathbf{F}_t^{-1} \end{aligned}$$

where $\boldsymbol{\eta}_t$ is an $n_a \times 1$ column vector representing the prediction error of $\mathbf{y}_t$ based on k_{t-1}, $\mathbf{F}_t$ is the $n_a \times n_a$ covariance matrix of the prediction error and $\mathbf{g}_t$ is a $1 \times n_a$ row vector denoting the Kalman gain.

To begin the filtering process, the values of k_{t_0-1} and p_{t_0-1} are needed. Pedroza (2006) set k_{t_0-1} to 5 and p_{t_0-1} to 10. Having completed the filtering process, we have a realization of k_t and p_t for each $t = t_0, \ldots, t_0 + n_y - 1$, with which the smoothing process can be executed.

3.3.2. The smoothing process

The smoothing process further smooths κ_t with information beyond time t. It is implemented with the backward smoothing algorithm (Carter and Kohn, 1994), which runs (backwards) from $t = t_0 + n_y - 1$ to $t = t_0$.

Let $h_t := E(\kappa_t|\mathbf{Y}_t, \kappa_{t+1})$ and $v_t := \mathrm{var}(\kappa_t|\mathbf{Y}_t, \kappa_{t+1})$, for $t = t_0, \ldots, t_0 + n_y - 2$, be the smoothed conditional expectation and variance of κ_t respectively. All relevant information beyond time t is incorporated in h_t and v_t. Note that we do not need to include κ_{t+2} and onwards in h_t and v_t, because $\{\kappa_t\}$ follows a random walk. The backward smoothing algorithm is implemented as follows.

Step 1: for $t = t_0 + n_y - 1$, draw a sample of κ_t from a normal distribution with a mean of $k_t = E(\kappa_t|\mathbf{Y}_t)$ and a variance of $p_t = \mathrm{var}(\kappa_t|\mathbf{Y}_t)$.
Step 2: for $t = t_0 + n_y - 2, \ldots, t_0$,

(a) calculate h_t and v_t by using the value of κ_{t+1} that is obtained from the previous step and the equations

$$\begin{aligned} h_t &= k_t + p_t(p_t + \sigma^2)^{-1}(\kappa_{t+1} - \mu - k_t), \\ v_t &= p_t - p_t^2(p_t + \sigma^2)^{-1}; \end{aligned}$$

(b) draw a sample of κ_t from $N(h_t, v_t)$.

Having completed a run of the backward smoothing algorithm, a sample of $\kappa_{t_0}, \ldots, \kappa_{t_0+n_y-1}$ is obtained. It is then combined with the sample of α_xs, β_xs, μ, s^2 and σ^2 to form a realization of the joint posterior distribution of the $2n_a + n_y + 3$ parameters.

3.4. Multiple imputation

Pedroza (2006) further suggested using the method of multiple imputation when fitting the model to a mortality data set that is subject to missing data problems. When multiple imputation is used, the Gibbs sampling procedure is modified as follows.

Step 1: at the beginning of each iteration of Gibbs sampling, impute all the missing data points. Specifically, if the central death rate at age x^* and year t^* is missing, impute the value of $\ln(m_{x^*,t^*})$ from

$$\ln(m_{x^*,t^*}|\Theta) \sim N(\alpha_{x^*} + \beta_{x^*}\kappa_{t^*}, s^2),$$

where α_{x^*}, β_{x^*}, κ_{t^*} and s^2 are taken as the values drawn in the previous iteration.
Step 2: the imputed values of all of the missing data points are combined with the observed data points to form a 'complete' data set, which is then used to obtain a realization of the model parameters by using the algorithms that were outlined in the previous subsections.
Step 3: repeat steps 1 and 2 to obtain an empirical posterior distribution of the model parameters.

4. The proposed stochastic model for Chinese mortality

4.1. Abridged multiple imputation

Pedroza (2006) did not implement the multiple-imputation method by using a real data set with missing data points. When we apply the method directly to the Chinese mortality data set, we find that the posterior distribution of the model parameters does not converge.

The non-convergence problem may be attributed to the fact that the proportion of missing data is too large (see Section 2). What we encounter is reminiscent of the demonstrations by Schafer (1997), which show that, when some data are missing, the joint posterior distribution is not necessarily proper if non-informative prior distributions are assumed, so convergence is not guaranteed, and that the non-convergence problem is exacerbated if

(a) the number of data points is small compared with the number of parameters and/or
(b) the proportion of missing data is high.

Note that both conditions (a) and (b) apply to the situation that we are confronting.

To circumvent this problem, we choose to impute only the individual missing (unreported) values (item (c) in Section 2) but not the missing values that appear in blocks (items (a) and (b) in Section 2). The rationale is that, as discussed in Section 2, the missing values that appear in blocks fit the definition of missingness completely at random. Not imputing data that are missing at random does not affect the validity of statistical inferences and predictions (Graham (2012), page 48; Van Buuren (2012), page 8). We emphasize that (as shown in subsequent sections) we still obtain a joint posterior distribution of all the model parameters even though not all of the missing values are imputed.

When this abridged multiple imputation is used, the conditional posterior distributions of α_x, β_x and σ^2 must be modified to reflect that they are conditioned on a partially imputed incomplete data set. The modified conditional posterior distributions are as follows.

(a) The conditional posterior distribution of α_x is $N(a_\alpha, b_\alpha)$, where

$$a_\alpha = \left(\sum_{t=t_0}^{t_0+n_y-1} \frac{\mathbb{1}_{x,t}}{s^2} \right)^{-1} \sum_{t=t_0}^{t_0+n_y-1} \frac{\ln(m_{x,t}) - \beta_x\kappa_t}{s^2} \mathbb{1}_{x,t}, \tag{6}$$

$$b_\alpha = \left(\sum_{t=t_0}^{t_0+n_y-1} \frac{\mathbb{1}_{x,t}}{s^2} \right)^{-1}, \tag{7}$$

and $\mathbb{1}_{x,t}$ is an indicator function that equals 1 if the central death rate for age x and year t is available, and 0 otherwise.

(b) The conditional posterior distribution of β_x is $N(a_\beta, b_\beta)$, where

$$a_\beta = \left(\sum_{t=t_0}^{t_0+n_y-1} \frac{\kappa_t^2 \mathbb{1}_{x,t}}{s^2} \right)^{-1} \sum_{t=t_0}^{t_0+n_y-1} \frac{\ln(m_{x,t}) - \alpha_x}{s^2} \kappa_t \mathbb{1}_{x,t} \tag{8}$$

and

$$b_\beta = \left(\sum_{t=t_0}^{t_0+n_y-1} \frac{\kappa_t^2 \mathbb{1}_{x,t}}{s^2} \right)^{-1}. \tag{9}$$

(c) The conditional posterior distribution of s^2 is an inverse gamma distribution with a shape parameter of

$$a_s = \frac{\sum_{x=x_0}^{x_0+n_a-1} \sum_{t=t_0}^{t_0+n_y-1} \mathbb{1}_{x,t}}{2}$$

and a rate parameter of

$$b_s = \frac{\sum_{x=x_0}^{x_0+n_a-1} \sum_{t=t_0}^{t_0+n_y-1} \{\ln(m_{x,t}) - \alpha_x - \beta_x \kappa_t\}^2 \mathbb{1}_{x,t}}{2}.$$

The conditional posterior distributions of μ and σ^2 require no modification. They are provided below for completeness.

(a) The conditional posterior distribution of μ is $N(a_\mu, b_\mu)$, where

$$\begin{aligned} a_\mu &= \frac{\kappa_{t_0+n_y-1} - \kappa_{t_0}}{n_y - 1}, \\ b_\mu &= \frac{\sigma^2}{n_y - 1}. \end{aligned} \tag{10}$$

(b) The conditional posterior distribution of σ^2 is an inverse gamma distribution with a shape parameter of

$$a_\sigma = \frac{n_y - 1}{2}$$

and a rate parameter of

$$b_\sigma = \frac{\sum_{t=t_0+1}^{t_0+n_y-1} (\kappa_t - \kappa_{t-1} - \mu)^2}{2}.$$

In the presence of missing data, initial values are critically important to Gibbs sampling (Carpenter and Kenward, 2013). If inappropriate initial values are used, the rate of convergence can be very slow. To improve convergence, we choose the initial values with the following procedure.

Step 1: obtain a crude estimate of each missing death or exposure count by linearly interpolating between adjacent (non-missing) values. A 'complete' data set is then obtained.

Step 2: obtain crude estimates of the model parameters by using Poisson maximum likelihood (Wilmoth, 1993; Brouhns *et al.*, 2002). Poisson maximum likelihood is regarded as a standard

(frequentist-type) method for estimating the Lee–Carter model when there are no missing data. We have experimented with other methods such as singular value decomposition, which was used in the original work of Lee and Carter (1992). The final estimation results are esssentially the same if another method of obtaining crude estimates is used.

Step 3: using the crude estimates from the previous step as initial values, apply the Gibbs sampling algorithm to the 'complete' data set.

Step 4: using the posterior means of the parameters from the previous step as initial values, apply the Gibbs sampling algorithm with multiple imputation to the actual data set (in which some central death rates are missing). This step yields the final estimated model.

This procedure enables the Gibbs sampling algorithm from which the final estimated model is derived to start at a state that is sufficiently close to the stationary state.

4.2. Sequential Kalman filter

In the abridged multiple imputation, we do not impute item (b) defined in Section 2. As a result, we do not always have a complete vector of observed or imputed death rates $\mathbf{y}_t$, which is required in the third step of the filtering process that was described in Section 3.3.1.

To overcome this technical challenge, we replace the original Kalman filter with the sequential Kalman filter that was proposed by Koopman and Durbin (2000). Instead of running the filtering process in a vector form, the sequential Kalman filter uses only one element in $\mathbf{y}_t$ in each step, incorporating the observed or imputed log(central death rates) into k_t and p_t one at a time.

To implement the sequential Kalman filter, we first rewrite the state space form of the Lee–Carter model in a sequential representation:

$$\ln(m_{x,t}) = \alpha_x + \beta_x \kappa_{x,t} + \epsilon_{x,t}, \tag{11}$$

$$\kappa_{x,t} = \begin{cases} \kappa_{x-1,t} & \text{if } x_0 < x \leqslant x_0 + n_a - 1, \\ \mu + \kappa_{x_0+n_a-1,t-1} + \omega_t & \text{if } x = x_0, \end{cases} \tag{12}$$

where $\epsilon_{x,t} \sim^{\text{IID}} N(0, s^2)$ and $\omega_t \sim^{\text{IID}} N(0, \sigma^2)$ as before. In the sequential representation, each log(central death rate) is driven by a hidden state $\kappa_{x,t}$ that depends on both age and time. However, as equation (12) implies, in a given year t the values of $\kappa_{x,t}$ are the same for all ages.

We let

$$\mathcal{Y}_{x,t} = \{y_{x,t}, y_{x-1,t}, \ldots, y_{x_0,t}, \mathbf{Y}_{t-1}\},$$

which contains all information from years t_0 to $t-1$ and partial information (from age x_0 to x) in year t. The sequential filtering process updates $k_{x,t} := E(\kappa_t | \mathcal{Y}_{x,t})$ and $p_{x,t} := \text{var}(\kappa_t | \mathcal{Y}_{x,t})$ as follows.

For $x = x_0$, we have

$$\begin{aligned} k_{x_0,t} &= \mu + k_{x_0+n_a-1,t-1} + g_{x_0,t}\eta_{x_0,t}, \\ p_{x_0,t} &= (1 - g_{x_0,t}\beta_{x_0})(p_{x_0+n_a-1,t-1} + \sigma^2), \\ \eta_{x_0,t} &= y_{x_0,t} - \alpha_{x_0} - \beta_{x_0}(\mu + k_{x_0+n_a-1,t-1}), \\ g_{x_0,t} &= \frac{(p_{x_0+n_a-1,t-1} + \sigma^2)\beta_{x_0}}{(p_{x_0+n_a-1,\,t-1} + \sigma^2)\beta_{x_0}^2 + s^2} \end{aligned}$$

when $t = t_0 + 1, \ldots, t_0 + n_y - 1$ and, as we use diffuse initial values for the sequential Kalman filter, we have

$$k_{x_0,t_0} = \frac{y_{x_0,t_0} - \alpha_{x_0}}{\beta_{x_0}},$$
$$p_{x_0,t_0} = \frac{s^2}{\beta_{x_0}^2}.$$

when $t = t_0$, where $\eta_{x,t}$ represents the prediction error of $y_{x,t}$, and $g_{x,t}$ denotes the Kalman gain. Further, for $x_0 < x \leqslant x_0 + n_a - 1$, we have

$$k_{x,t} = k_{x-1,t} + g_{x,t}\eta_{x,t},$$
$$p_{x,t} = (1 - g_{x,t}\beta_x)\, p_{x-1,t},$$
$$\eta_{x,t} = y_{x,t} - \alpha_x - \beta_x k_{x-1,t},$$
$$g_{x,t} = \frac{p_{x-1,t}\beta_x}{p_{x-1,t}\beta_x^2 + s^2}.$$

The recursion runs from $t = t_0$ to $t = t_0 + n_y - 1$ and, for each t, from $x = x_0$ to $x = x_0 + n_a - 1$. In contrast with the original filtering process, the sequential filtering process does not involve inversions of large dimension matrices and is therefore more efficient (Koopman and Durbin (2000), page 287).

Having completed the entire recursion, we obtain the values of $k_{x,t}$ for $x = x_0, \ldots, x_0 + n_a - 1$ and $t = t_0, \ldots, t_0 + n_y - 1$, from which we obtain the values of $k_{t_0}, \ldots, k_{t_0+n_y-1}$ as

$$k_{x_0+n_a-1,t} = E(\kappa_t | \mathcal{Y}_{x_0+n_a-1,t}) = E(\kappa_t | \mathbf{Y}_t) = k_t$$

according to the definitions of $k_{x,t}$, k_t, $\mathbf{Y}_t$ and $\mathcal{Y}_{x,t}$. The values of $k_{t_0}, \ldots, k_{t_0+n_y-1}$ are then fed into the backward smoothing algorithm (Section 3.3.2) in which a sample of $\kappa_{t_0}, \ldots, \kappa_{t_0+n_y-1}$ is drawn.

The sequential filtering process can be modified readily to suit the abridged multiple imputation. If a data point (say $y_{x,t}$) is missing and not imputed, then we change the equations for $k_{x,t}$ and $p_{x,t}$ in the recursion to

$$k_{x,t} = \begin{cases} \mu + k_{x_0+n_a-1,t-1}, & x = x_0, \\ k_{x-1,t}, & x = x_0 + 1, \ldots, x_0 + n_a - 1, \end{cases}$$

and

$$p_{x,t} = \begin{cases} p_{x_0+n_a-1,t-1} + \sigma^2, & x = x_0, \\ p_{x-1,t}, & x = x_0 + 1, \ldots, x_0 + n_a - 1 \end{cases}$$

respectively, so that no information concerning $y_{x,t}$ is incorporated in $k_{x,t}$ and $p_{x,t}$.

4.3. Smoothing the age–response parameters

Our preliminary work shows that, without any special treatment, the Bayesian estimates of β_x derived from the Chinese mortality data set exhibit a very jagged pattern across ages. With such estimates of β_x, the projected future mortality rates do not have a smooth, logical age pattern. The same problem was also encountered in the work of Huang and Browne (2017), page 42, who applied the Lee–Carter model to a portion of the Chinese mortality data set. The abnormally jagged pattern of β_x is in part because the length of the data set is too short and in part because

some exposure counts (those in the years for which the data are obtained from surveys) are too small. To ensure that the resulting mortality forecasts are demographically reasonable, an adaptation that smooths the pattern of β_x across ages is needed.

There are a few methods for smoothing the pattern of β_x across ages during the course of estimation, including the penalized log-likelihood approach (Delwarde *et al.*, 2007) and the cubic *B*-splines method (Renshaw and Haberman, 2003). We choose to extend the cubic *B*-splines method to a Bayesian set-up.

The cubic *B*-splines function for smoothing β_x is defined as follows:

$$\beta_x = c_0 + c_1 \ln(x+1) + c_2 \ln(x+1)^2 + c_3 \ln(x+1)^3 + \sum_{j=1}^{r} c_{3+j} (\ln(x+1) - \ln(x_j+1))_+^3, \quad (13)$$

where c_j for $j = 0, 1, \ldots, r+3$ are the coefficients, r denotes the number of knots, x_j for $j = 1, 2, \ldots, r$ are the chosen knot ages, and $(\ln(x+1) - \ln(x_j+1))_+^3$ equals $\{\ln(x+1) - \ln(x_j+1)\}^3$ if $x > x_j$, and 0 otherwise. To extend the method to a Bayesian set-up, we first rewrite equation (13) in a matrix form as

$$\boldsymbol{\beta} = \mathbf{A}\mathbf{c},$$

where $\mathbf{c} = (c_0, \ldots, c_{r+3})'$, and $\mathbf{A}$ is an $n_a \times (r+4)$ matrix of which the ith row is given by

$$(1, \ln(x_0+i), \ln(x_0+i)^2, \ln(x_0+i)^3, (\ln(x_0+i) - \ln(x_1+1))_+^3, \ldots, (\ln(x_0+i) - \ln(x_r+1))_+^3).$$

Our goal is to derive the conditional posterior distribution of $\mathbf{c}$, which will in turn produce posterior distributions of β_xs whose expectations are smooth across ages. Assuming an improper conditional prior distribution, the conditional posterior distribution of $\mathbf{c}$ is MVN($\mathbf{a_c}, \mathbf{B_c}$), where

$$\mathbf{a_c} = \left(\sum_{t=t_0}^{t_0+n_y-1} \frac{\kappa_t^2}{s^2} \mathbf{A}' \mathbf{I}_t \mathbf{A} \right)^{-1} \left(\sum_{t=t_0}^{t_0+n_y-1} \frac{\kappa_t}{s^2} \mathbf{A}' \mathbf{I}_t (\mathbf{y}_t - \boldsymbol{\alpha}) \right), \quad (14)$$

$$\mathbf{B_c} = \left(\sum_{t=t_0}^{t_0+n_y-1} \frac{\kappa_t^2}{s^2} \mathbf{A}' \mathbf{I}_t \mathbf{A} \right)^{-1}, \quad (15)$$

and $\mathbf{I}_t$ is an $n_a \times n_a$ diagonal matrix, of which the ith diagonal element equals 0 if $m_{x_0+i-1,t}$ is missing and not imputed, and 1 otherwise.

The jaggedness is less significant for α_x, but for consistency we also apply a cubic *B*-splines function with the same number of knots, r, and the same knot ages $(x_1, \ldots x_r)$ to α_x. In a matrix form, the cubic *B*-splines function for smoothing α_x is given by

$$\boldsymbol{\alpha} = \mathbf{A}\mathbf{d},$$

where $\mathbf{d} = (d_0, \ldots, d_{r+3})'$ represents the vector of coefficients. Assuming an improper conditional prior distribution, the conditional posterior distribution of $\mathbf{d}$ is MVN($\mathbf{a_d}, \mathbf{B_d}$), where

$$\mathbf{a_d} = \left(\sum_{t=t_0}^{t_0+n_y-1} \frac{1}{s^2} \mathbf{A}' \mathbf{I}_t \mathbf{A} \right)^{-1} \left(\sum_{t=t_0}^{t_0+n_y-1} \frac{1}{s^2} \mathbf{A}' \mathbf{I}_t (\mathbf{y}_t - \boldsymbol{\beta}\kappa_t) \right) \quad (16)$$

and

$$\mathbf{B_d} = \left(\sum_{t=t_0}^{t_0+n_y-1} \frac{1}{s^2} \mathbf{A}' \mathbf{I}_t \mathbf{A} \right)^{-1}. \quad (17)$$

The knot ages $(x_1, \ldots x_r)$ are evenly spaced between ages 0 and 70 years. We used evenly spaced knots between ages 0 and 70 years and no knot between ages 70 and 99 years for two reasons. First, too much flexibility beyond age 70 years may result in a counterintuitive outcome that projected mortality rates at older ages are not monotonically increasing with age. Second, placing all knots between age 0 and 70 years enables us to capture the accident hump better. The number of knots, r, is determined by using the deviance information criterion (Spiegelhalter *et al.*, 2002), which is designed to compare different Bayesian models. The Bayesian deviance is defined by

$$D(\Theta) = -2\ln\{\Pr(\mathbf{Y}|\Theta)\} + 2\ln\{f(\mathbf{Y})\},$$

where $f(\mathbf{Y})$ is a fully specified standardizing term which depends on $\mathbf{Y}$ only and, following Spiegelhalter *et al.* (2002), the deviance information criterion is defined as

$$\mathrm{DIC} = E\{D(\Theta)|\mathbf{Y}\} + E\{D(\Theta)|\mathbf{Y}\} - D\{E(\Theta|\mathbf{Y})\},$$

where $E\{D(\Theta)|\mathbf{Y}\}$ measures the goodness of fit and $E\{D(\Theta)|\mathbf{Y}\} - D\{E(\Theta|\mathbf{Y})\}$ penalizes the use of more parameters. Note that this definition of DIC coincides with the seventh definition of DIC in the work of Celeux *et al.* (2006). We can compute $E\{D(\Theta)|\mathbf{Y}\}$ and $D\{E(\Theta|\mathbf{Y})\}$ readily by using the empirical joint posterior distribution of Θ. There is no need to determine $f(\mathbf{Y})$, because it does not depend on the model and hence is cancelled out in model comparisons. We choose the number of knots, r, that yields the smallest DIC.

4.4. Incorporating sampling uncertainty that depends on the source of data

As discussed in Section 2, a good proportion of the available data is based on surveys of 1% or 0.1% of the national population instead of nationwide censuses. Fig. 2 shows the log(central death rates) from age 0 to 99 years for 2010 (for which the data are based on a nationwide census), 2005 (for which the data are based on a survey of 1% of the national population) and 2014 (for which the data are based on a survey of 0.1% of the national population). It is clear that the observed log(central death rates) in the three years chosen are subject to different amounts of volatility.

Therefore, for the data set in question, it is inappropriate to assume that $\epsilon_{x,t}$ has a variance of s^2 which does not depend on whether the data for year t are obtained from a census or survey. If the same value of s^2 is used for every year, we shall overestimate the variance of $\epsilon_{x,t}$ in years for which the data are obtained from censuses, which will in turn lead to overly wide prediction intervals when predicting the underlying future death rates for the entire national population.

To take this special characteristic of the data set into account, we further extend the model by permitting the variance of $\epsilon_{x,t}$ to be time inhomogeneous. In our extension, the observation equation (equation (4)) of the model is modified to

$$\mathbf{y}_t = \boldsymbol{\alpha} + \boldsymbol{\beta}\kappa_t + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathrm{MVN}(0, s_t^2\mathbf{I}),$$

where s_t^2 takes three distinct values, reflecting the nature of the data for year t:

$$s_t^2 = \begin{cases} s_{\mathrm{c}}^2 & t \in \mathcal{T}_{\mathrm{c}}, \\ s_{1\%}^2 & t \in \mathcal{T}_{1\%}, \\ s_{0.1\%}^2 & t \in \mathcal{T}_{0.1\%}, \end{cases}$$

where $\mathcal{T}_{\mathrm{c}}$, $\mathcal{T}_{1\%}$ and $\mathcal{T}_{0.1\%}$ represent the collections of years for which the data are obtained from nationwide censuses, surveys of 1% of the national population and surveys of 0.1% of the

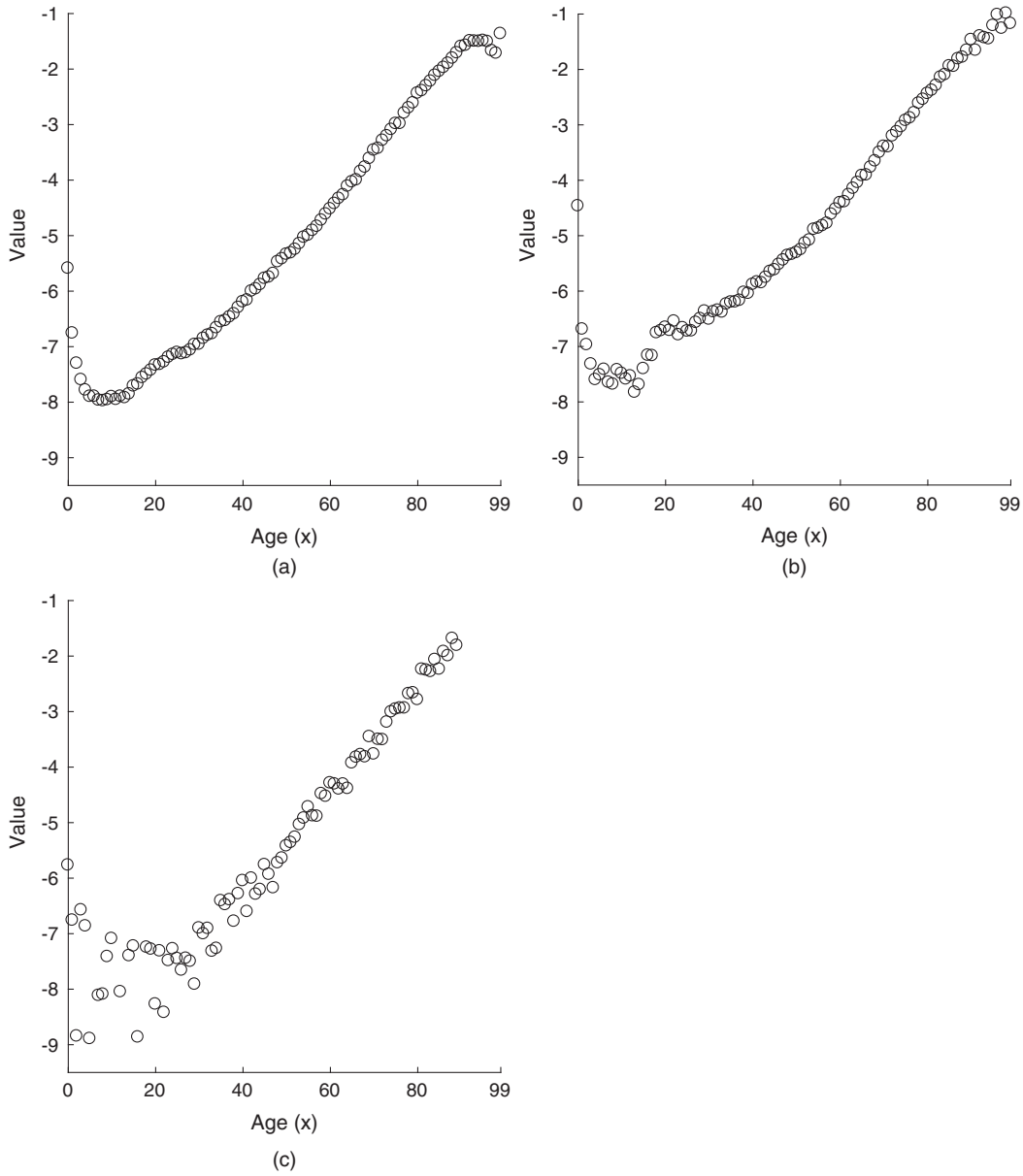


Fig. 2. Log(central death rates) from age 0 to 99 years for (a) 2010 (for which the data are based on a nationwide census), (b) 2005 (for which the data are based on a survey of 1% of the national population) and (c) 2014 (for which the data are based on a survey of 0.1% of the national population)

national population respectively. The conditional posterior distribution of s_t^2 is the same as that of s^2, except that the second summation in the numerators of a_s and b_s is taken over either $t \in \mathcal{T}_{\mathrm{c}}$, $t \in \mathcal{T}_{1\%}$ or $t \in \mathcal{T}_{0.1\%}$, whichever is appropriate.

The conditional posterior distributions of α_x, β_x, **c** and **d** must be modified accordingly. Specifically, s^2 in equations (6)–(9) and (14)–(17) is replaced by s_t^2. We also need to replace s^2 by s_t^2 in the algorithm for the sequential Kalman filter.

5. Estimation results

In this section, we estimate the full model with the adaptations that were described in Sections 4.3 and 4.4 to the Chinese mortality data. The estimation is completed by using Gibbs sampling with the abridged multiple imputation (Section 4.1) and the sequential Kalman filter (Section 4.2). The estimation results for both genders have similar properties, and therefore for brevity we show only those for males in the main text. The estimation results for females are provided in the on-line annex E.

5.1. Convergence of parameters

Using the initial values that were determined by the procedure described in Section 4.1, the Gibbs sampling algorithm converges fairly quickly (within 25 iterations). As usual in Gibbs sampling, we treat the first 500 drawn values as 'burn-in' and discard them. To mitigate auto-correlation, only one sample is recorded in every 100 values drawn after the burn-in period. In total, 5000 samples are recorded and used to form the joint posterior distribution. Further details concerning convergence of parameters are provided in the on-line annex A.

5.2. Posterior distributions of parameters

The fan charts in Fig. 3 display the posterior distributions of α_x, β_x and κ_t for $x = 0, \ldots, 99$ and $t = 1981, \ldots, 2014$. Each fan chart shows the equal-tail 10% credible interval with the heaviest shading, surrounded by the $20\%, 30\%, \ldots, 90\%$ credible intervals with progressively lighter shadings. The width of a fan chart indicates the level of parameter uncertainty that is entailed.

The pattern of the posterior means of α_xs is generally in line with a typical age pattern of mortality. With the aid of the cubic B-splines functions, the patterns of the posterior means of both α_xs and β_xs are smooth.

As expected, the posterior means of κ_ts exhibit a downward trend. The variation in the uncertainty surrounding κ_t over time is highly in line with the structure of the data set. In 1982–1985, 1987, 1988 and 1990–1993 for which no mortality data are available, the width of the fan chart of κ_t is particularly wide.

Fig. 4 depicts the posterior distributions of s_c^2, $s_{1\%}^2$ and $s_{0.1\%}^2$. The posterior means of s_c^2, $s_{1\%}^2$ and $s_{0.1\%}^2$ are considerably different, confirming the need for allowing the variance of $\epsilon_{x,t}$ to change according to how the data are obtained. As expected, the posterior mean of s_c^2 (which represents the extent of sampling uncertainty when the data are obtained from censuses) is the lowest, whereas the posterior mean of $s_{0.1\%}^2$ (which represents the extent of sampling uncertainty when the data are obtained from surveys of 0.1% of the population) is the highest.

Finally, Fig. 5 shows the posterior distributions of the parameters in the state equation. It is noteworthy that μ is subject to much parameter uncertainty, with a posterior distribution that spans both positive values (which imply mortality improvement) and negative values (which imply mortality deterioration).

6. Prediction results

In this section, we present the prediction results that are derived by using the data for Chinese males. We first present the baseline results that are obtained from the full model, followed by a demonstration of how the prediction results may change if we ignore parameter uncertainty or omit the adaptations proposed.

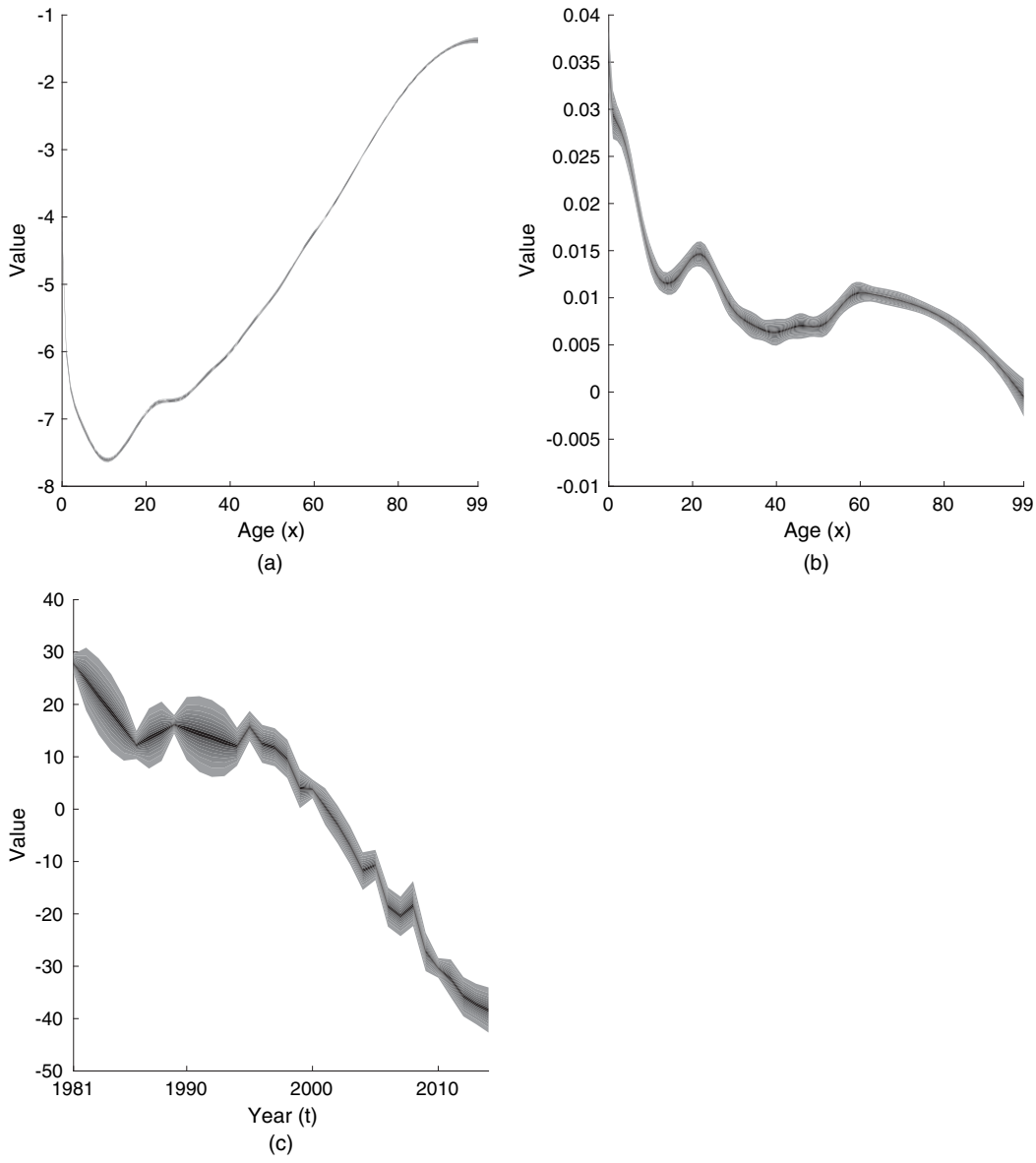


Fig. 3. Posterior distributions of (a) α_x, (b) β_x and (c) κ_t for $x = 0, \ldots, 99$ and $t = 1981, \ldots, 2014$

6.1. Baseline results

Predictions of future mortality are obtained from the posterior predictive distribution of $\mathbf{y}_{t_0+n_y-1+u}$, $u = 1, 2, \ldots$. Assuming that historical and future log(central death rates) are independently distributed given Θ, the posterior predictive distribution of $\mathbf{y}_{t_0+n_y-1+u}$ is given by

$$f(\mathbf{y}_{t_0+n_y-1+u}|\mathbf{Y}) = \int f(\mathbf{y}_{t_0+n_y-1+u}|\Theta)\pi(\Theta|\mathbf{Y})\mathrm{d}\Theta, \qquad u = 1, 2, \ldots,$$

where $f(\mathbf{y}_{t_0+n_y-1+u}|\Theta)$ is the density function of $\mathrm{MVN}(\boldsymbol{\alpha} + \boldsymbol{\beta}\kappa_{t_0+n_y-1+u}, s^2_{t_0+n_y-1+u}\mathbf{I})$: the distribution of $\mathbf{y}_{t_0+n_y-1+u}$ given Θ.

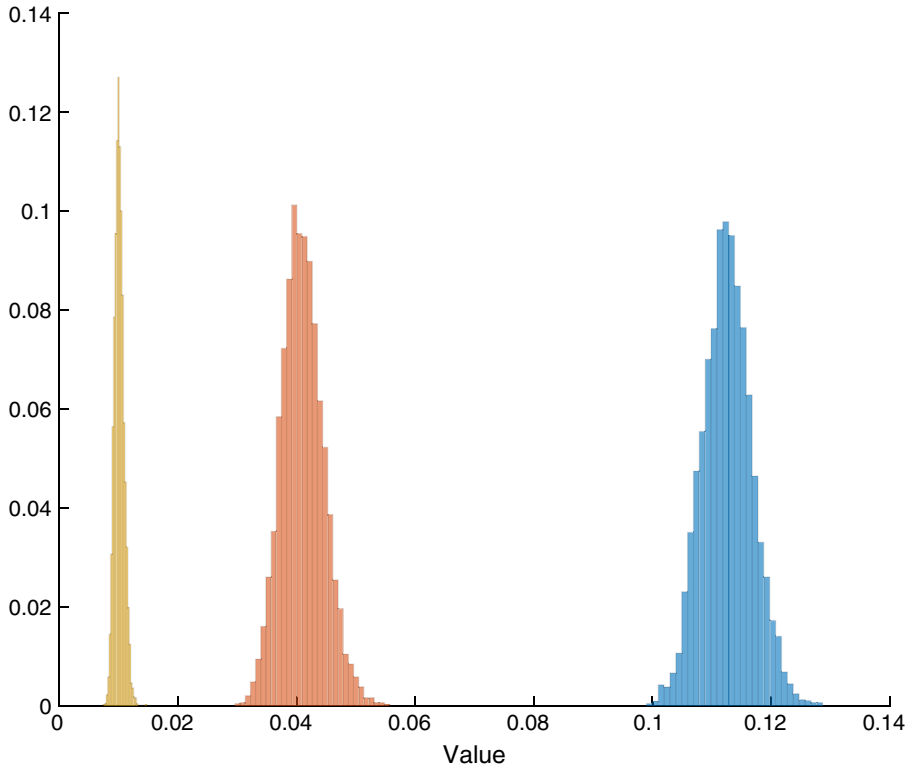


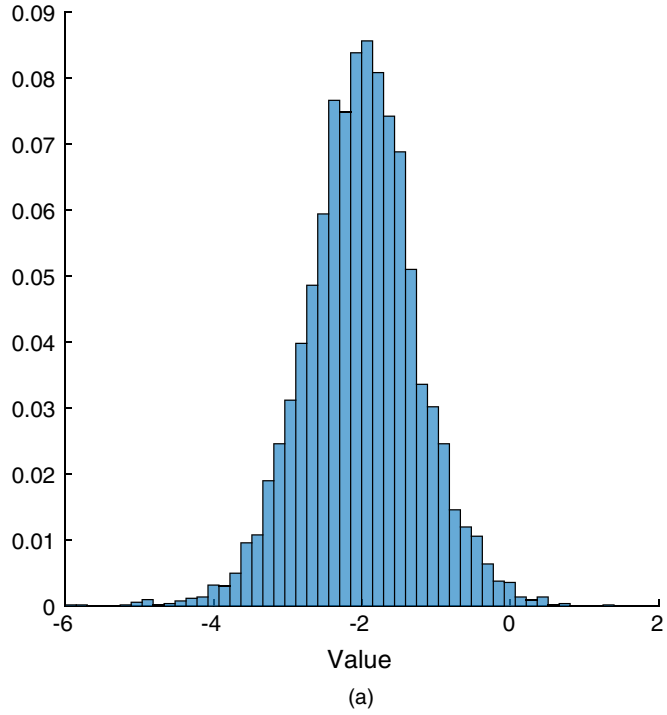
Fig. 4. Posterior distributions of s^2_{c} (), $s^2_{1\%}$ () and $s^2_{0.1\%}$ ()

The posterior predictive distribution depends on the choice of $s^2_{t_0+n_y-1+u}$. When predicting mortality for the entire population, one should set $s^2_{t_0+n_y-1+u}$ to s^2_{c}. Alternatively, if we are interested in knowing how the observed central death rates in year t_0+n_y-1+u may look if they are sampled from 1% (or 0.1%) of the population, then we may set $s^2_{t_0+n_y-1+u}$ to $s^2_{1\%}$ (or $s^2_{0.1\%}$). The choice of $s^2_{t_0+n_y-1+u}$ affects only the prediction intervals but not the central prediction (the posterior predictive mean), as the expected value of ϵ_{x,t_0+n_y-1+u} is 0 no matter which one of the three options is chosen.

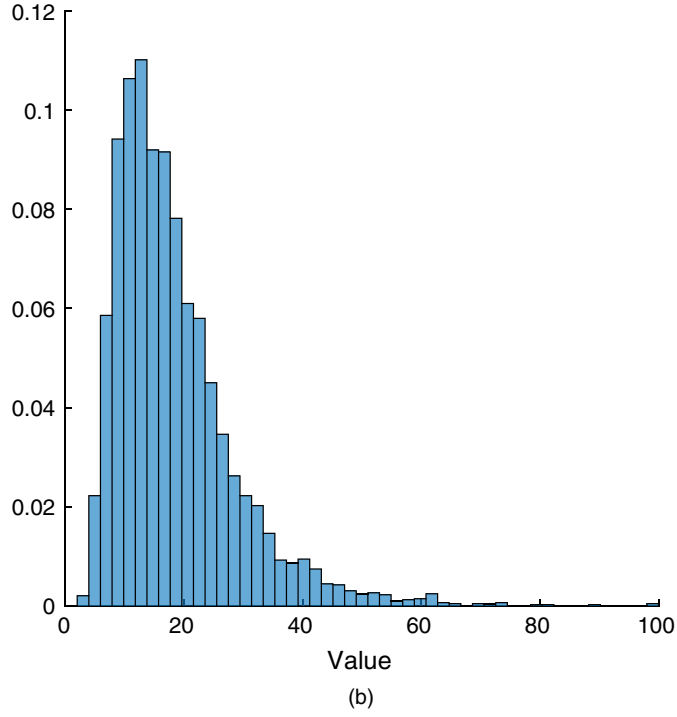
We obtain the posterior predictive distribution numerically with the following procedure.

Step 1: obtain a realization of Θ from its empirical posterior distribution.
Step 2: given the values of μ, σ^2 and $\kappa_{t_0+n_y-1}$ in the realization of Θ from the previous step, simulate a realization of $\kappa_{t_0+n_y-1+u}$ from its conditional predictive distribution $N(\kappa_{t_0+n_y-1}+u\mu, u\sigma^2)$.
Step 3: given the realizations of $\kappa_{t_0+n_y-1+u}$ and Θ from the previous two steps, simulate a realization of $\mathbf{y}_{t_0+n_y-1+u}$ from its conditional predictive distribution $\mathrm{MVN}(\boldsymbol{\alpha}+\boldsymbol{\beta}\kappa_{t_0+n_y-1+u}, s^2_{t_0+n_y-1+u}\mathbf{I})$.
Step 4: repeat steps 1–3 to obtain a large number of realizations of $\mathbf{y}_{t_0+n_y-1+u}$, which collectively form an empirical posterior predictive distribution of $\mathbf{y}_{t_0+n_y-1+u}$.

Fig. 6 presents the predictions of central death rates at ages $0, 10, 20, \ldots,$ 90 years over a horizon of 35 years. Three 90% prediction intervals, which are respectively derived by setting $\{s^2_t; t=2015,\ldots,2049\}$, to s^2_{c} (the full curves), $s^2_{1\%}$ (the broken curves) and $s^2_{0.1\%}$ (the dotted



(a)



(b)

Fig. 5. Posterior distributions of (a) μ and (b) σ^2

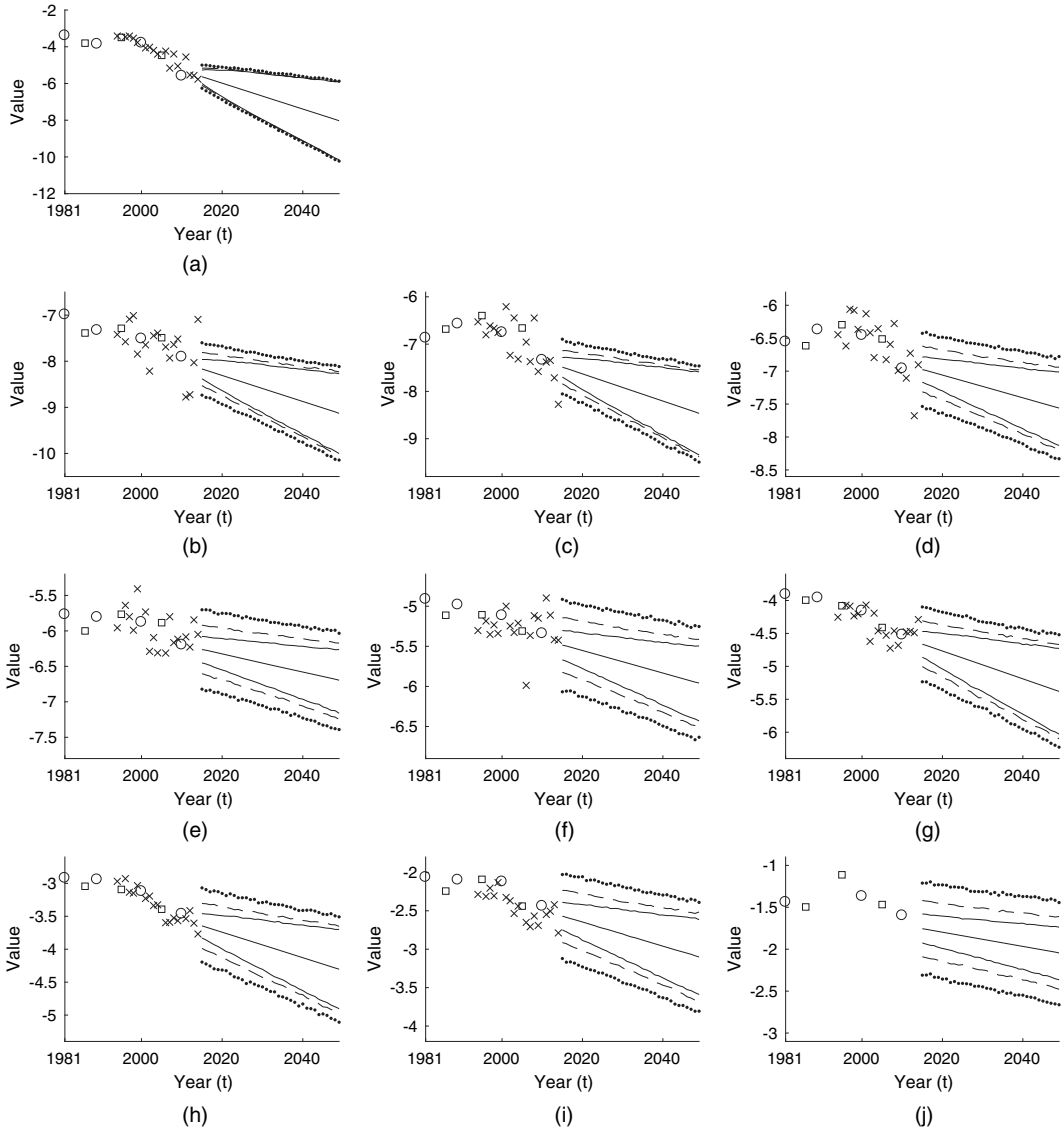


Fig. 6. 90% prediction intervals for the central death rates at ages (a) 0, (b) 10, (c) 20, (d) 30, (e) 40, (f) 50, (g) 60, (h) 70, (i) 80 and (j) 90 years for the 2015–2049 period, based on s_c^2 (———), $s_{1\%}^2$ (— —) and $s_{0.1\%}^2$ (· · · · · ·) (the central prediction is presented by the line in the middle of the prediction intervals): ○, historical central death rates for those obtained from censuses; □, historical central death rates for those obtained from surveys of 1% of the population; ×, historical death rates for those obtained from surveys of 0.1% of the population

curves), are displayed in each panel. The line in the middle of the prediction intervals represents the central prediction. Also depicted in Fig. 6 are the observed historical log(central death rates): those obtained from nationwide censuses are shown as circles, those obtained from surveys of 1% of the national population are shown as squares and those obtained from surveys of 0.1% of the national population are shown as crosses.

Two key observations can be made from Fig. 6. First, the central prediction appears to be a

logical progression of the historical values. The gradient of the central prediction is particularly close to that of the line joining the circles (historical values obtained from censuses).

Second, the widths of the prediction intervals are commensurate with the variability of the observed historical values. In each panel, the narrowest prediction interval (based on s^2_{c}) seems to be sufficient for capturing the variation in the circles (historical values obtained from censuses), and the other two prediction intervals (based on $s^2_{1\%}$ and $s^2_{0.1\%}$) appear to be adequate in capturing the additional variation in the squares and crosses (historical values obtained from surveys).

6.2. The importance of parameter uncertainty

The prediction intervals that were presented in the previous subsection include three sources of uncertainty, namely the stochastic uncertainty in the random walk for $\{\kappa_t\}$, the parameter uncertainty that is implied by the joint posterior distribution of the parameters and the sampling uncertainty that is captured by the error term in the observation equation. In this subsection, we demonstrate the importance of incorporating parameter uncertainty when forecasting Chinese mortality by comparing the previously presented prediction intervals with the corresponding prediction intervals that do not incorporate any parameter uncertainty.

We use the following procedure to generate prediction intervals that do not include parameter uncertainty.

Step 1: calculate $E(\mu|\mathbf{Y})$, $E(\sigma^2|\mathbf{Y})$, $E(\kappa_{t_0+n_y-1}|\mathbf{Y})$, $E(\boldsymbol{\alpha}|\mathbf{Y})$, $E(\boldsymbol{\beta}|\mathbf{Y})$ and $E(s^2_{t_0+n_y-1+u}|\mathbf{Y})$ from the posterior distributions of the model parameters.

Step 2: simulate a realization of $\kappa_{t_0+n_y-1+u}$ for each $u = 1, 2, \ldots$ from its predictive distribution,

$$N\{E(\kappa_{t_0+n_y-1}|\mathbf{Y}) + u\,E(\mu|\mathbf{Y}), u\,E(\sigma^2|\mathbf{Y})\}.$$

Step 3: given the result from step 2, simulate a realization of $\mathbf{y}_{t_0+n_y-1+u}$ for each $u = 1, 2, \ldots$ from its predictive distribution,

$$\mathrm{MVN}\{E(\boldsymbol{\alpha}|\mathbf{Y}) + E(\boldsymbol{\beta}|\mathbf{Y})\kappa_{t_0+n_y-1+u}, E(s^2_{t_0+n_y-1+u}|\mathbf{Y})\mathbf{I}\}.$$

Step 4: repeat steps 2 and 3 to obtain an empirical predictive distribution of $\mathbf{y}_{t_0+n_y-1+u}$ for each $u = 1, 2, \ldots$. The lower 5% and upper 95% percentiles of the empirical predictive distributions form 90% prediction intervals that do not include any parameter uncertainty.

Fig. 7 compares the prediction intervals of log(central death rates) with and without parameter uncertainty. When parameter uncertainty is removed, the prediction intervals at all ages are significantly narrower. The reduction in width is more significant for longer horizon predictions. These findings suggest that parameter uncertainty is important when considering the Chinese mortality data set, and that ignoring it may lead to an understatement of the true level of uncertainty that is entailed in the prediction of future Chinese mortality.

Finally, we remark that the parameter uncertainty that is associated with μ plays a particularly important role in long horizon forecasts. This can be understood by considering the fact that the posterior predictive variance of $\kappa_{t_0+n_y-1+u}$, given by

$$\begin{aligned}\mathrm{var}(\kappa_{t_0+n_y-1+u}|\mathbf{Y}) &= \mathrm{var}\{E(\kappa_{t_0+n_y-1+u}|\Theta)|\mathbf{Y}\} + E\{\mathrm{var}(\kappa_{t_0+n_y-1+u}|\Theta)|\mathbf{Y}\}\\ &= \mathrm{var}(\kappa_{t_0+n_y-1} + u\mu|\mathbf{Y}) + E(u\sigma^2|\mathbf{Y})\\ &= \mathrm{var}(\kappa_{t_0+n_y-1}|\mathbf{Y}) + u^2\mathrm{var}(\mu|\mathbf{Y}) + 2u\,\mathrm{cov}(\kappa_{t_0+n_y-1}, \mu|\mathbf{Y}) + u\,E(\sigma^2|\mathbf{Y}),\end{aligned}$$

is related to u^2 times the posterior variance of μ.

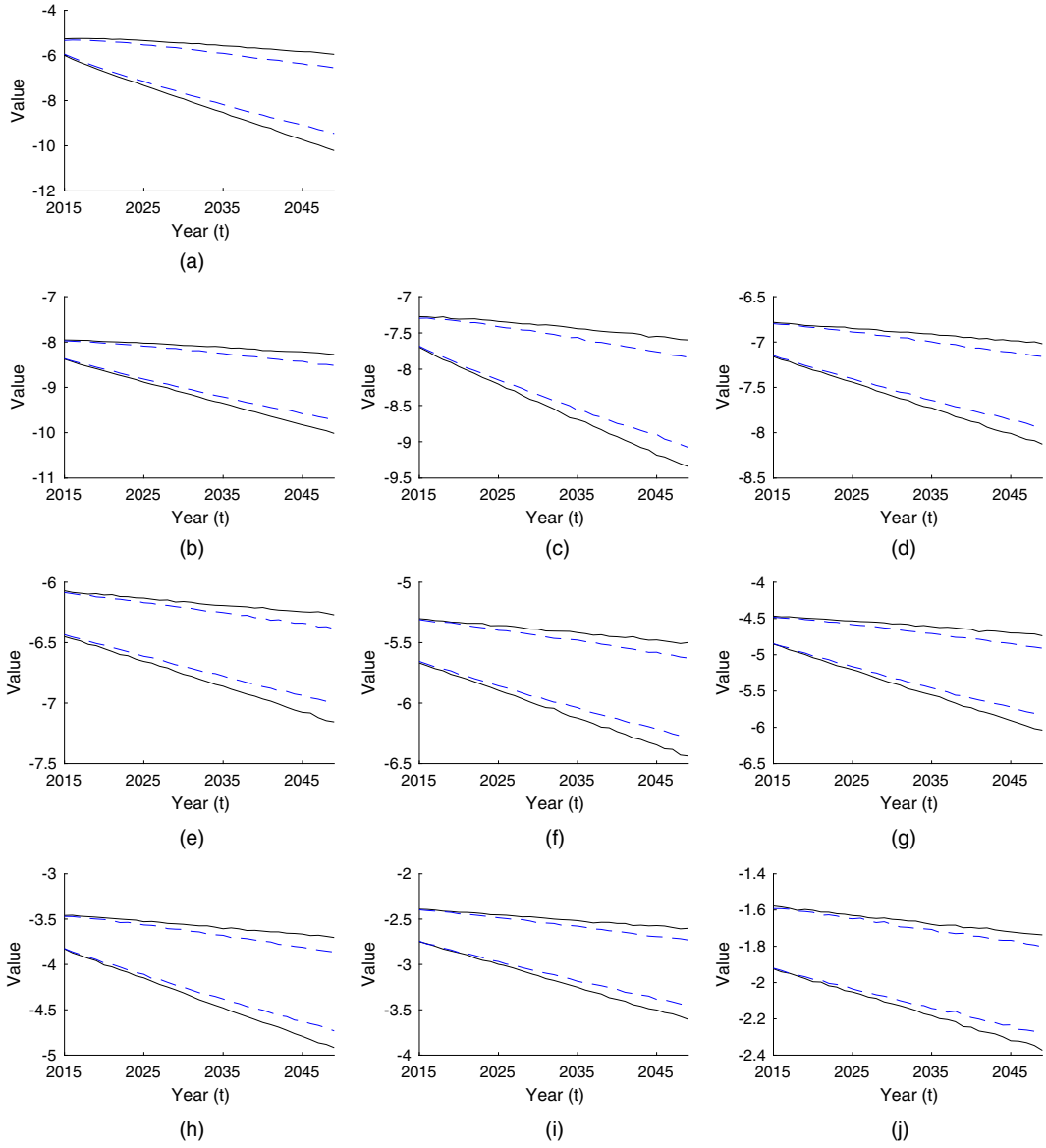


Fig. 7. 90% prediction intervals of log(central death rates) at ages (a) 0, (b) 10, (c) 20, (d) 30, (e) 40, (f) 50, (g) 60, (h) 70, (i) 80 and (j) 90 years over a horizon of 35 years, with parameter uncertainty (———) and without parameter uncertainty (— —): all of the prediction intervals shown are derived by using $s_t^2 = s_c^2$ for $t = 2015, \ldots, 2049$

6.3. Importance of the proposed adaptations

The results that were presented in Sections 5 and 6.1 are based on the full model, which includes the adaptations for smoothing the age-specific parameters and incorporating sampling uncertainty that depends on the source of data. In this subsection, we demonstrate how the estimation and prediction results may change if these two adaptations are switched off.

Fig. 8 compares the posterior distributions of β_xs when the two adaptations are used and not

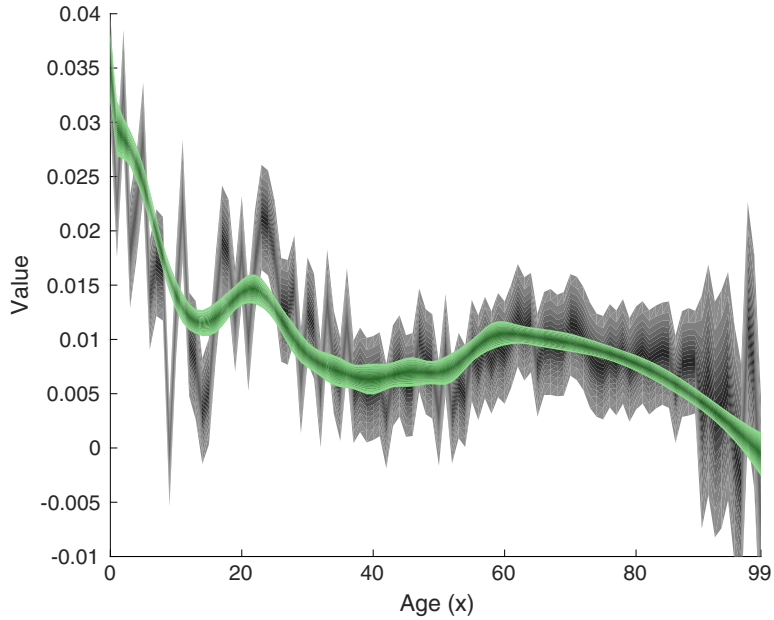


Fig. 8. Posterior distributions of β_xs when the adaptations proposed in Sections 4.3 and 4.4 are used (▬) and not used (▬)

used. Without the *B*-splines function for smoothing β_x, the resulting pattern of the posterior means of β_xs becomes highly erratic. Such a pattern leads to counterintuitive projection results. For example, in the absence of the *B*-splines function, the posterior means of β_x at ages 9 and 10 years are -0.0003 and 0.0128 years respectively, which means that the log(central death rates) at these two ages are expected to evolve in opposite directions, which is an outcome that makes no demographic sense. The problem can also be observed in Fig. 9, in which we compare the projected age patterns of future mortality generated from the full model and the model without the adaptations proposed.

Fig. 10 displays the central and 90% interval predictions of log(central death rates), derived from the restricted model that does not incorporate the two proposed adaptations. Compared with those derived from the full model, the central predictions that are generated from the restricted model are not so in line with the corresponding historical values. The problem is most apparent in Fig. 10(f) for age 50 years, where we observe that the central projection and the line joining the circles (historical values obtained from censuses) have quite different slopes. A likely cause of the problem is a misestimation of β_xs, which arises as the *B*-splines function is removed.

The prediction intervals that are presented in Fig. 10 appear to be too wide if we view them as prediction intervals for the log(central death rates) of the national population. As discussed earlier, when predicting future mortality for the entire population, the uncertainty due to sampling from a fraction of the population should be excluded, and for this reason a flexible specification of $\text{var}(\epsilon_{x,t})$ is proposed. When the restricted model is used, the prediction intervals reflect the piece of uncertainty that should not be taken into consideration and are therefore too wide. The problem is particularly obvious for shorter-term projections. For example, at age 50 years, the interval for the 1-year-ahead projection is sufficiently wide to encompass all except only one historical values.

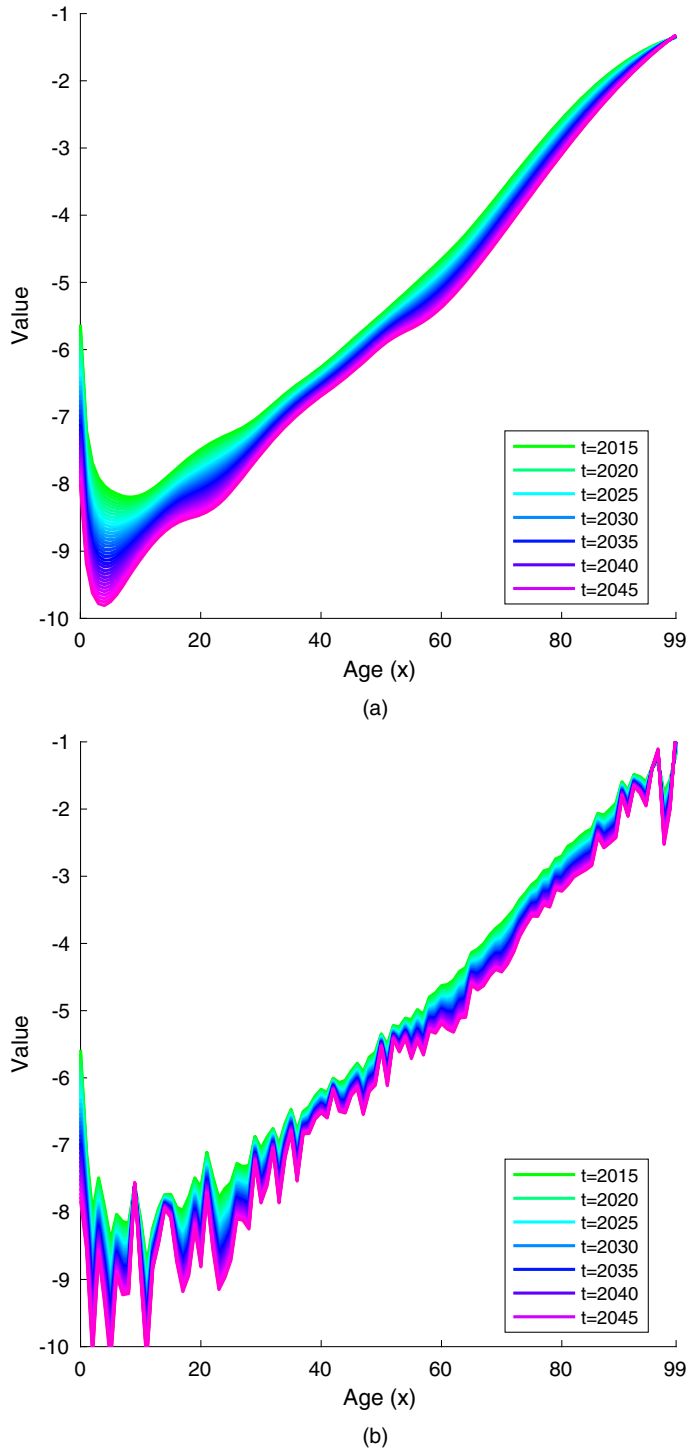


Fig. 9. Central prediction of log(central death rates) in 2015, 2020, . . . , 2049, derived from (a) the full model and (b) the model without the adaptations proposed in Sections 4.3 and 4.4

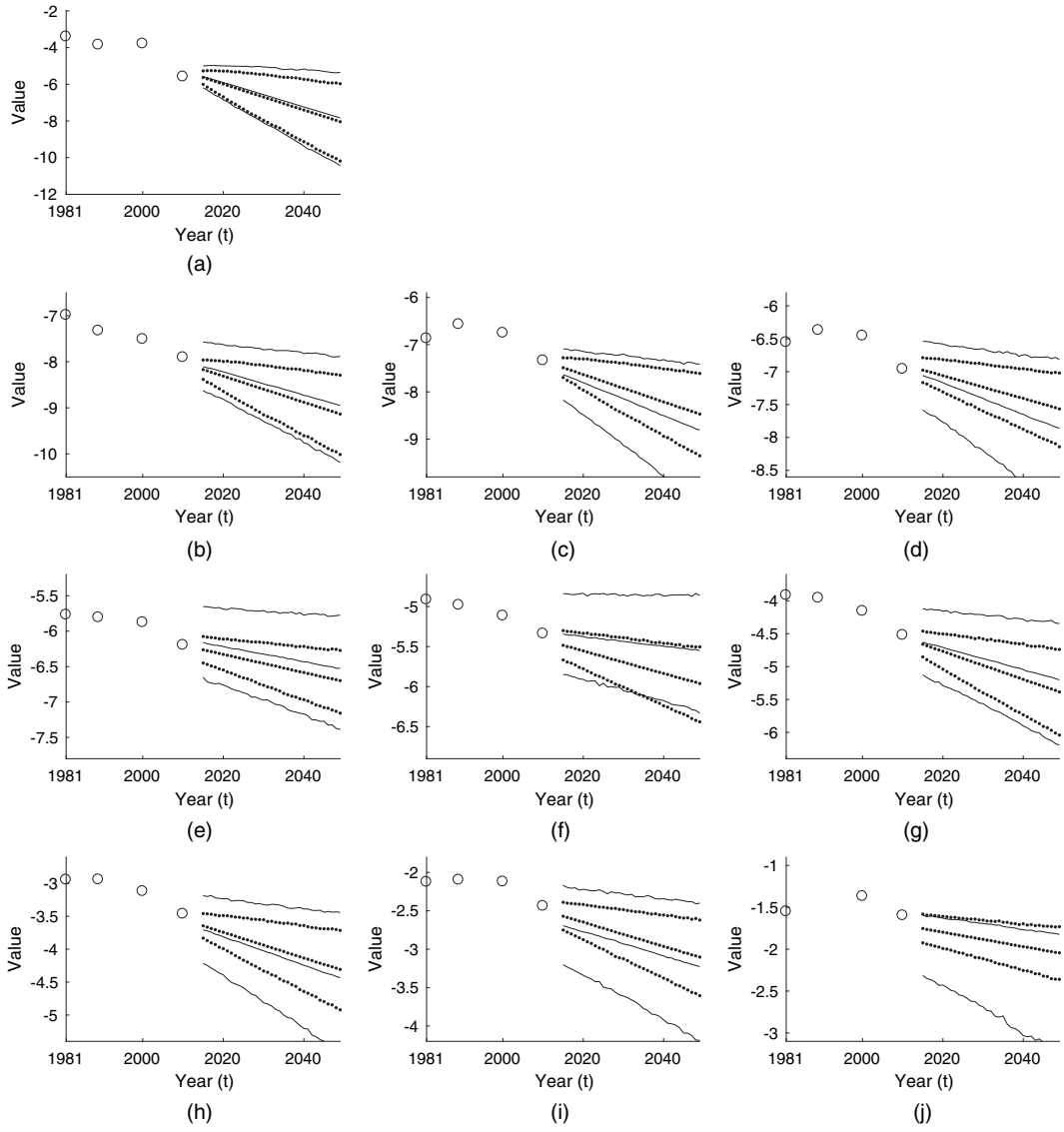


Fig. 10. Central and 90% interval predictions of the log(central death rates) at ages (a) 0, (b) 10, (c) 20, (d) 30, (e) 40, (f) 50, (g) 60, (h) 70, (i) 80 and (j) 90 years in the 2015–2049 period, derived from the model without the adaptations proposed in Sections 4.3 and 4.4 (———) and from the full model (based on s_c^2) (· · · · · ·): ○, historical central death rates obtained from censuses

7. Validation of the proposed modelling and estimation approaches

7.1. Generating pseudodata sets

We use pseudodata sets to validate the modelling approach proposed. Each pseudodata set has the same dimension as the actual data set (i.e. $n_a = 100$ and $n_y = 34$) and is generated with the following procedure.

Step 1: the exposure counts are taken as those in the actual data set (for Chinese males). This

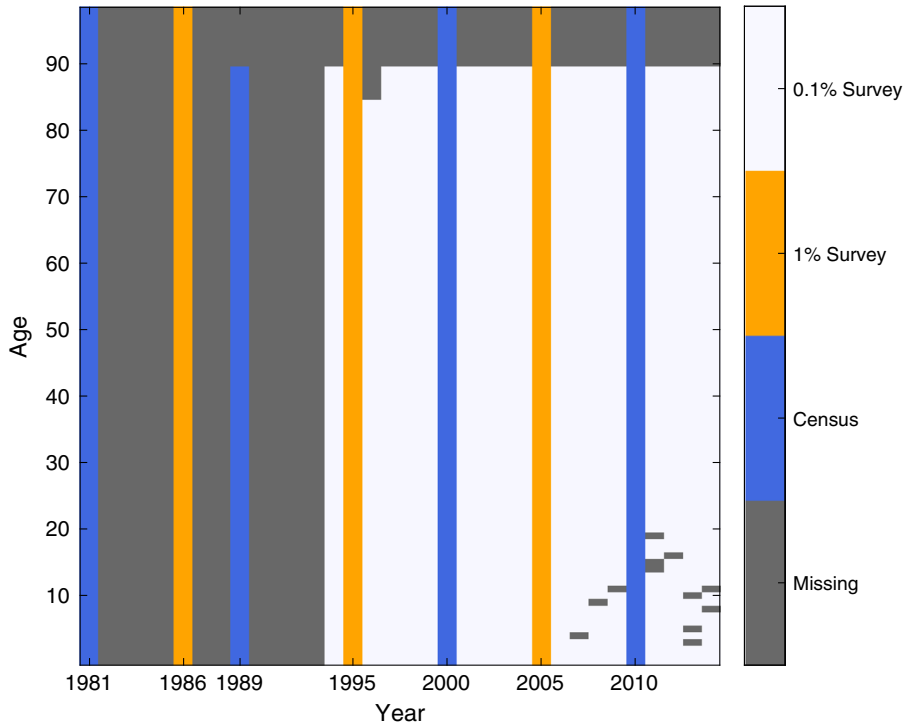


Fig. 11. Lexis diagram summarizing the structure of one of the generated pseudodata sets

step implies that the pseudodata set is subject to exactly the same missing value problems as described in items (a) and (b) in Section 2.

Step 2: the death count at age x and in year t is obtained by simulating a realization of a Poisson distribution with a mean of

$$E_{x,t} \exp\{E(\alpha_x|\mathbf{Y}) + E(\beta_x|\mathbf{Y})E(\kappa_t|\mathbf{Y})\},$$

where $E_{x,t}$ is the corresponding exposure count (obtained from the previous step), and $E(\alpha_x|\mathbf{Y})$, $E(\beta_x|\mathbf{Y})$ and $E(\kappa_t|\mathbf{Y})$ are the posterior means of α_x, β_x and κ_t calculated in Section 5.2 respectively.

Step 3: for each age–time cell in which the exposure count is available, calculate the simulated log(central death rate) as $\ln(\tilde{m}_{x,t}) = \ln(\tilde{D}_{x,t}/E_{x,t})$, where $\tilde{D}_{x,t}$ denotes the simulated death count for age x and year t. When $\tilde{D}_{x,t} = 0$, the corresponding log(central death rate) is recorded as 'unreported', thereby resembling the missing data problem that was described in item (c) in Section 2.

Re-estimating our model with such pseudodata sets enables us to examine how our model may perform if the normality assumption does not hold and if heteroscedasticity exists (as the pseudodata generation procedure implies that the variances at different ages and in different calendar years are generally different).

Fig. 11 summarizes the structure of one of the pseudodata sets generated. This structure is identical to that of the actual data set for Chinese males (Fig. 1(a)), except that the locations and exact number of the individual missing values are different owing to the Poisson variations in step 2 of the procedure above.

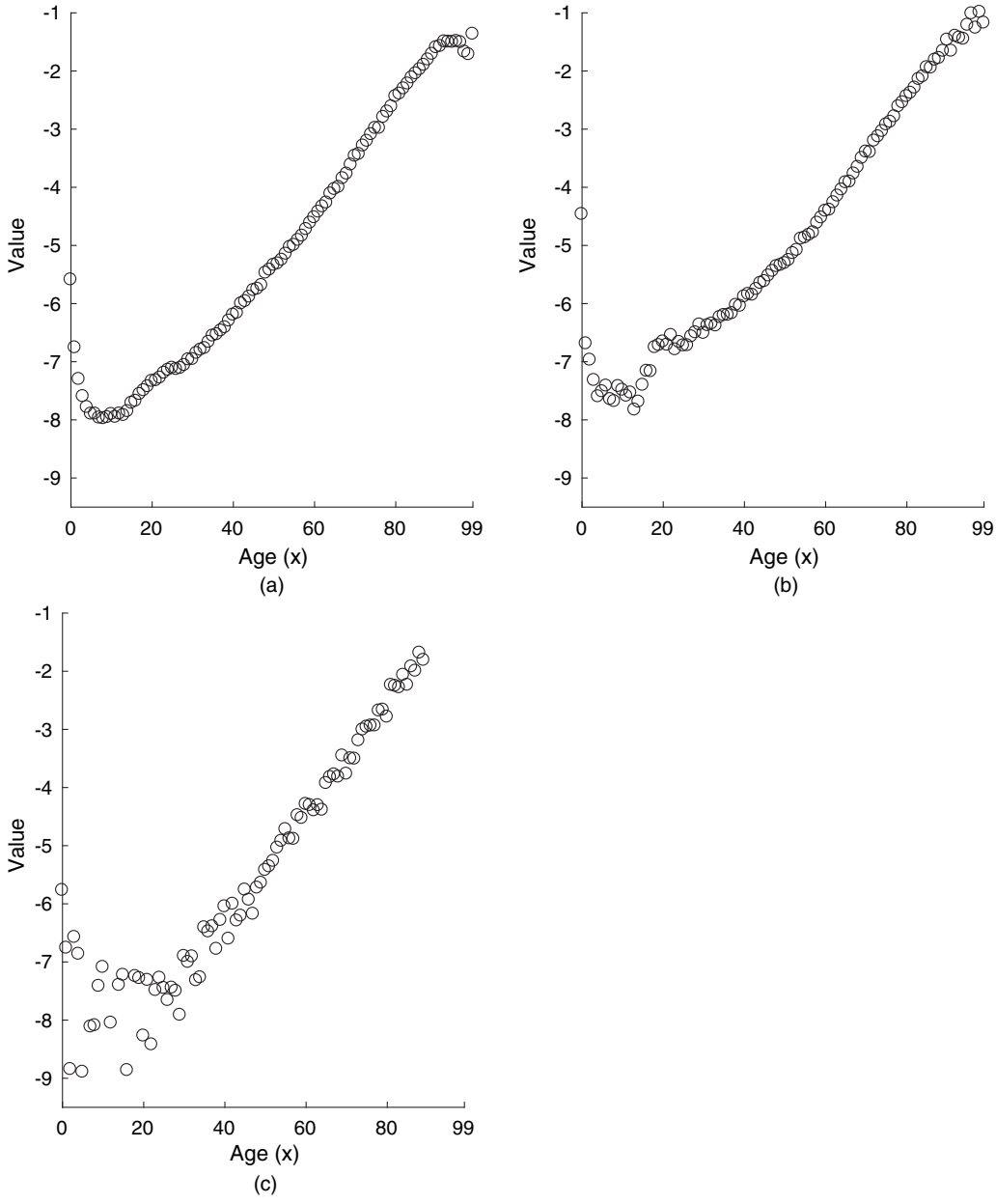


Fig. 12. Log(central death rates) for (a) 2010, (b) 2005 and (c) 2014 in one of the generated pseudodata sets

Fig. 12 shows the log(central death rates) for 2010, 2005 and 2014 in one of the generated pseudodata sets. By comparing them with the corresponding values in the actual data set (see Fig. 2), we conclude that the pseudodata set closely reproduces the three different levels of sampling uncertainty that are found in the actual data set.

As implied by step 2 in the procedure above, the data-generating process underlying the pseudodata sets is the Lee–Carter model with parameters equal to the posterior parameter means

derived from the actual data set. Hence, if our proposed modelling and estimation approaches are appropriate, then, when applied to the pseudodata sets, they should result in parameter estimates that are sufficiently close to the parameters in the underlying data-generating process.

7.2. Validation results

The full curves in Fig. 13 show the 90% credible intervals of α_x, β_x and κ_t for $x = 0, \ldots, 99$

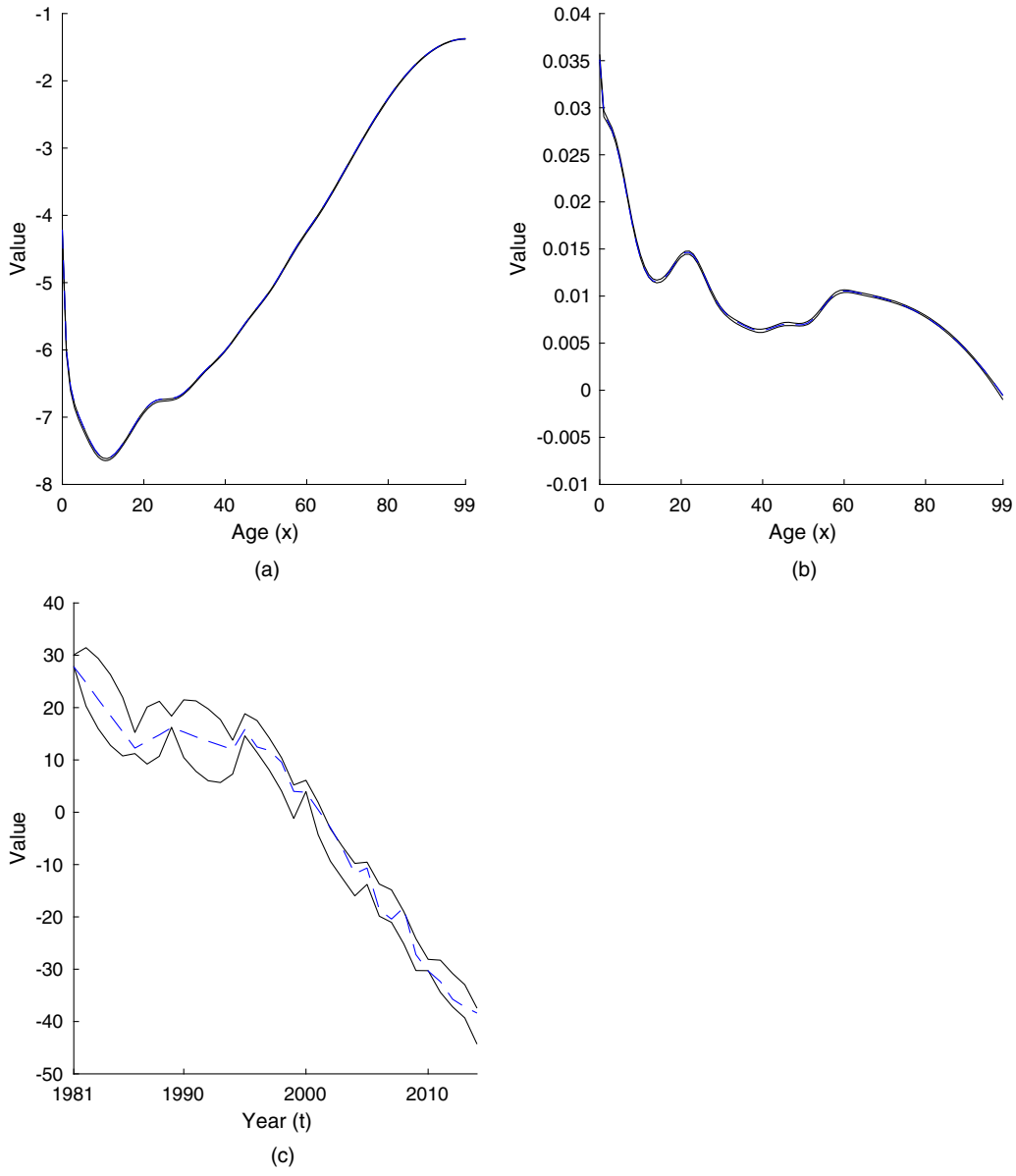


Fig. 13. 90% credible intervals (———) of (a) α_x, (b) β_x and (c) κ_t for $x = 0, \ldots, 99$ and $t = 1981, \ldots, 2014$, calculated by using the posterior distributions derived from one of the pseudodata sets: — —, corresponding 'true' parameter values

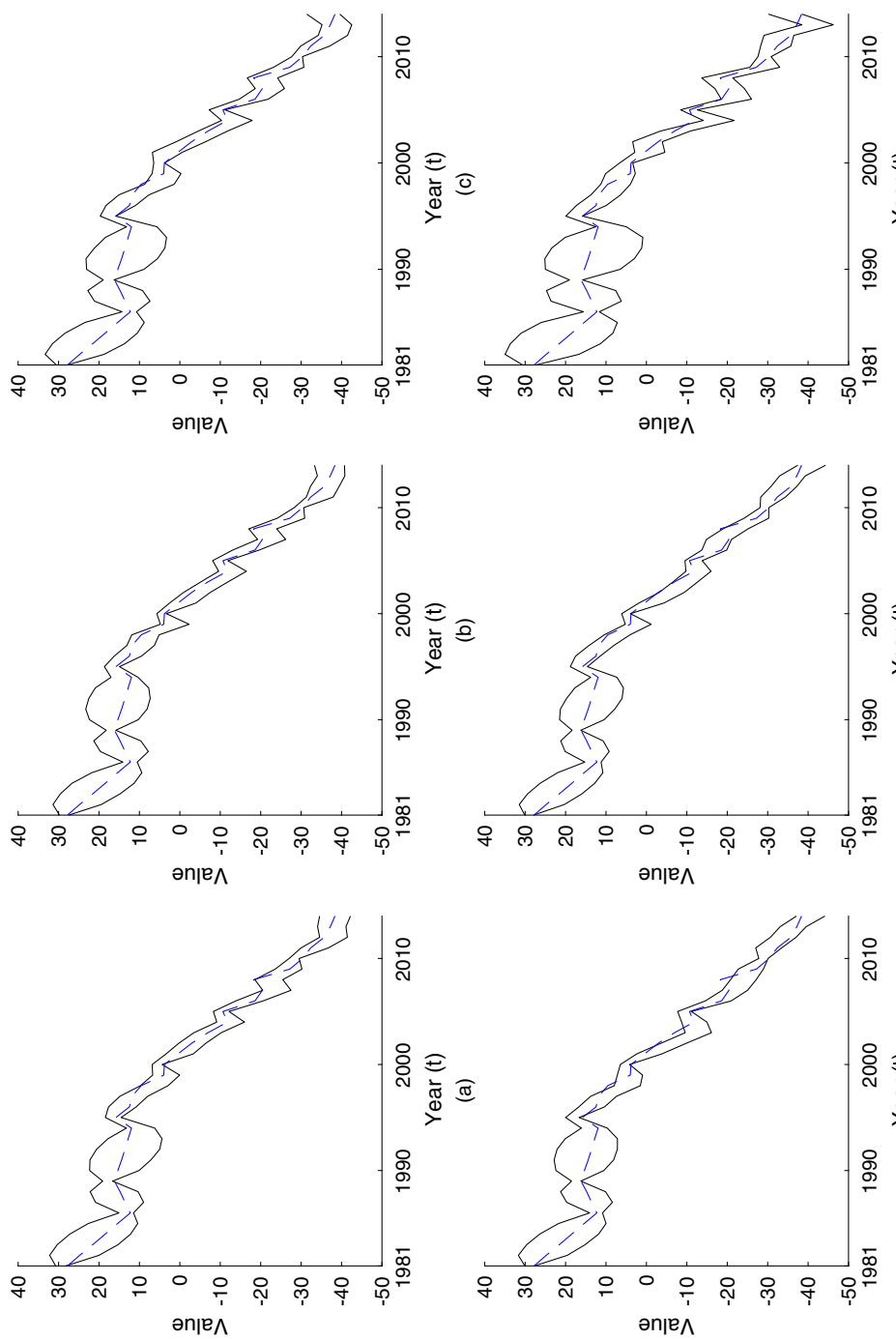


Fig. 14. 90% credible intervals (——) of κ_t for $t = 1981, \ldots, 2014$, calculated by using the posterior distributions derived from six of the generated pseudodata sets: — —, 'true' parameter values

years and $t = 1981, \ldots, 2014$, calculated by using the posterior distributions that were derived from one of the pseudodata sets. They should be compared against the corresponding 'true' parameters (i.e. those in the underlying data-generating process), which are shown in broken curves in Fig. 13.

For α_x and β_x, the 90% credible intervals are rather narrow, suggesting that the extent of parameter uncertainty arising from finite sampling and missing values is small for these age-specific parameters. Despite being narrow, the 90% credible intervals capture all the 'true' values of α_x and β_x, providing strong support for using the proposed methods to estimate the model in the presence of missing data and finite sampling.

For κ_t, the 90% credible intervals are wider, indicating that finite sampling and missing values create more uncertainty surrounding this parameter. The 90% credible intervals enclose most (33 out of 34) of the 'true' parameter values, suggesting that the proposed estimation methods can estimate κ_t reasonably accurately in the presence of missing data and finite sampling. Note also that they have comparable widths with those for the κ_ts derived from the real Chinese mortality data set (see Fig. 3(b)).

When applied to the rest of the generated pseudodata sets, the estimates of α_x and β_x remain almost unchanged. The estimates of κ_t vary observably when different data sets are used, but the 90% credible intervals consistently capture most of the 'true' parameter values (Fig. 14).

We acknowledge that this analysis may not be comprehensive, as the Poisson psuedodata sets do not capture, for example, the possibility that the variance of a random death count is greater than the mean. In the on-line annex B, we present two additional analyses that are based on psuedodata sets generated from (log-)normal distributions and Student t-distributions instead. For all the tests performed, the parameter estimates that were obtained from the pseudodata sets are highly similar to those derived from the actual data set. The normality assumption that we made therefore appears to be reasonable.

7.3. A cautionary note

As pointed out by Li and Chan (2005) and Zhou and Li (2013), mortality forecasts that are produced by the original Lee–Carter model are highly sensitive to the values of κ_t in the beginning and ending years of the data sample (i.e. κ_{t_0} and $\kappa_{t_0+n_y-1}$). This problem still exists even if the model is estimated with Bayesian methods instead of singular value decomposition or maximum likelihood. To demonstrate, consider the u-step-ahead forecast of log(central death rates), all of which are critically dependent on the posterior predictive mean of $\kappa_{t_0+n_y-1+u}$, given by

$$E(\kappa_{t_0+n_y-1}|\mathbf{Y}) + u\,E(\mu|\mathbf{Y}).$$

This expression depends entirely on (the posterior mean of) $\kappa_{t_0+n_y-1}$, which governs the initialization of the forecast, and (the posterior mean of) μ, which determines the gradient of the expected trajectory. Furthermore, as shown in expression (10), the mean of the conditional posterior distribution of μ is

$$\frac{\kappa_{t_0+n_y-1} - \kappa_{t_0}}{n_y - 1},$$

which is determined exclusively by $\kappa_{t_0+n_y-1}$ and κ_{t_0}.

As discussed in Section 7.2, the estimates of κ_t vary across the pseudodata sets. The coloured jagged curves in Fig. 15 represent the posterior means of κ_t for $t = 1981, \ldots, 2014$, obtained from three different pseudodata sets. The variation in the posterior means of κ_t is very subtle for years when the exposure size equals the population size (1981, 1989, 2000 and 2010) but is more pronounced for years when the exposure size is only a fraction of the population size.

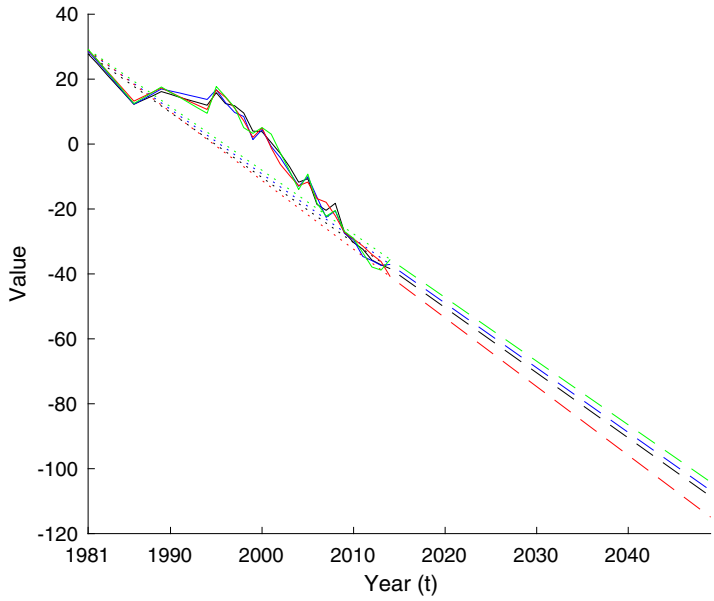


Fig. 15. Posterior means of $\kappa_{1981},\ldots,\kappa_{2014}$ (———, pseudodata set 1; ———, pseudodata set 2; ———, pseudodata set 3), expected trajectories of κ_{2015}, $\kappa_{2016},\ldots$ (— —, — —, — —) and the line that passes through $E(\kappa_{2014}|\mathbf{Y})$ and has a slope of $E(\mu|\mathbf{Y})$ (·····, ·····, ·····) obtained from three of the generated pseudodata sets (without any adjustment): ———, — —, · · · · · ·, corresponding values implied by the 'true' parameters

In the final year (2014) of the pseudodata sets, the exposures sizes are 0.1% of the respective population sizes, so the expected trajectories (posterior predictive means) of $\kappa_{2015}, \kappa_{2016}, \ldots$ may vary considerably across different pseudodata sets. This phenomenon can be observed in the coloured broken lines in Fig. 15, which represent the expected trajectories of $\kappa_{2015}, \kappa_{2016}, \ldots$ obtained from three of the generated pseudosamples. These expected trajectories have noticeably different levels and gradients. Some of them are quite different from the expected trajectory that is implied by the 'true' parameters, which is represented by the black broken line in Fig. 15.

For the real Chinese mortality data set, the data in the last year (2014) were obtained from a survey of 0.1% of the population. Thus, as Fig. 15 suggests, the resulting mortality forecasts are influenced by the large amount of uncertainty that is associated with κ_{2014}. One possible way to mitigate this problem is to adjust the posterior predictive mean of κ_{2014+u}, $u = 1, 2, \ldots$, from

$$E(\kappa_{2014}|\mathbf{Y}) + u\, E(\mu|\mathbf{Y})$$

to

$$E(\kappa_{2010}|\mathbf{Y}) + (u+4)\, E(\mu|\mathbf{Y}) \tag{18}$$

and to modify the conditional posterior mean of μ from

$$\frac{\kappa_{2014} - \kappa_{1981}}{33}$$

to

$$\frac{\kappa_{2010} - \kappa_{1981}}{29}, \tag{19}$$

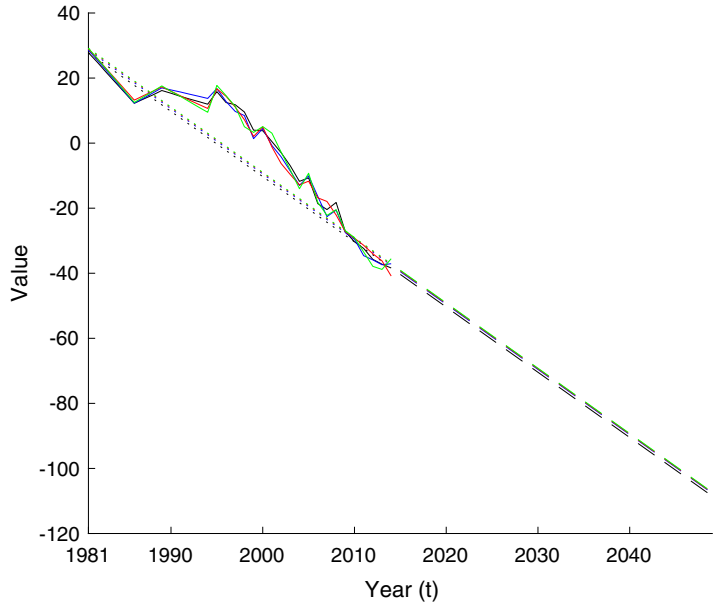


Fig. 16. Posterior means of $\kappa_{1981}, \ldots, \kappa_{2014}$ (———, pseudodata set 1; ———, pseudodata set 2; ·····, pseudodata set 3), expected trajectories of κ_{2015}, $\kappa_{2016}, \ldots$ (– –, – –, – –) and the line that passes through $E(\kappa_{2014}|\mathbf{Y})$ and has a slope of $E(\mu|\mathbf{Y})$ (·····, ·····, ·····) obtained from three of the generated pseudodata sets (with the proposed adjustments): ———, — —, · · · · · ·, corresponding values implied by the 'true' parameters

while all the other components in the model are kept unchanged. As the data for 1981 and 2010 are obtained from nationwide censuses, expressions (18) and (19) are less influenced by sampling uncertainty, and so is the expected trajectory of $\kappa_{2015}, \kappa_{2016}, \ldots$. The benefit of the adjustments proposed is demonstrated in Fig. 16, from which we observe that with the proposed adjustments the resulting expected trajectories of $\kappa_{2015}, \kappa_{2016}, \ldots$ exhibit less variation across the pseudodata sets and are all close to the expected trajectory that is implied by the 'true' parameters.

8. Conclusion

We have developed in this paper a Bayesian approach to model the evolution of Chinese mortality over time, with all of the problems that are associated with the data set being accounted for. Significant effort has been made to ensure that the end result satisfies the three important criteria that were set out in Section 1.

The first criterion is that the model should exploit as much information as possible from the data set. We have tailor made for the data set an estimation procedure, synthesizing Gibbs sampling, an abridged multiple-imputation algorithm and a sequential Kalman filter. The estimation procedure can handle the various types of discontinuities in the data set, so that all of the available age-specific mortality data (from 1981 to 2014) can be incorporated in the model. This feature makes our proposed approach stand out from those in the literature, which have disregarded a significant portion of the available data.

The second criterion is that the model should provide appropriate provisions for uncertainty. To our knowledge, our work represents the first attempt to obtain a joint posterior distribution

of *all* of the model parameters (including κ_t for all $t = 1981, \ldots, 2014$) for the Chinese mortality data set. The joint posterior distribution offers us a comprehensive measure of parameter uncertainty, including the additional uncertainty arising from the missing data. In addition, we have introduced a flexible specification of $\text{var}(\epsilon_{x,t})$, allowing it to vary according to how the data for year t are obtained. This flexible specification yields more reasonable prediction intervals for the population's underlying log(central death rates).

The third criterion is that the model should be parsimonious and yield biologically reasonable projections. We have built our proposed model from (a Gaussian state space version of) the Lee–Carter model, which is relatively parsimonious compared with other stochastic mortality models that are applicable to the full age range of 0–99 years. Also, we have introduced a Bayesian version of the cubic *B*-spline function of Renshaw and Haberman (2003) to smooth the pattern of the estimates (posterior means) of the age response parameters. This feature prevents the mortality forecasts from exhibiting any anti-intuitive behaviour that may arise from a jagged pattern of the parameter estimates.

Unlike two-stage estimation approaches whereby the Box–Jenkins method (Box *et al.*, 2015) may be used to identify the optimal time series process for the estimates of κ_t from the first stage, Bayesian (single-stage) estimation approaches require us to fix the time series process for $\{\kappa_t\}$ ahead of the estimation. We have specified a random walk with drift for $\{\kappa_t\}$, in part because this simple process has been shown to work well for various populations (Tuljapurkar *et al.*, 2000) and in part because the short data series does not seem to support more sophisticated processes. We suggest revisiting the process for $\{\kappa_t\}$ when more years of data become available. We also acknowledge that it may not be ideal to assume that $\text{var}(\epsilon_{x,t})$ is fixed along the age dimension. However, allowing $\text{var}(\epsilon_{x,t})$ to be age specific would significantly increase the number of parameters that must be estimated.

We do not consider cohort effects due to the limited availability of data. For the data set in question, we can observe at most only about a third of a birth cohort, and the limited observation is not even continuous because of the missing data. Any cohort-specific characteristic that is identified from such a data set is likely to be spurious. When more years of data become available, it is warranted to extend our modelling approach to incorporate cohort effects. To accomplish such an extension, there is a need to reformulate a stochastic mortality model that bears cohort effects (e.g. Renshaw and Haberman (2006) and Cairns *et al.* (2009)) into a Gaussian state space form. The recent work of Fung *et al.* (2017) may be used as a starting point.

The recent model of Hilton *et al.* (2019) merits some discussion. Ignoring its cohort term, this model is quite similar to the Lee–Carter model that we consider in terms of structure and number of parameters. Formulated as a generalized additive model, the model of Hilton *et al.* (2019) is advantageous in being less challenging to estimate and incorporate parameter smoothing. Although Hilton *et al.* (2019) developed the estimation procedure in a Bayesian setting, they did not consider the possibility of having missing data. To utilize their model for our application, an external imputation model for imputing missing death counts would seem necessary. Furthermore, it is not clear how the model of Hilton *et al.* (2019) can take the source of data (nationwide censuses, surveys of 1% of the population or surveys of 0.1% of the population) into account. Without this feature, it would be difficult to interpret the prediction intervals that are produced by the model when it is estimated from data from different sources. In contrast, when using our model in which s_t^2 is devised to incorporate sampling uncertainty that depends on the source of data, the user can easily produce, for example, prediction intervals for the underlying mortality rates (i.e. mortality rates that are estimated from the entire population) by setting s_t^2 to s_c^2.

We have validated our modelling and estimation approaches by using pseudodata sets generated from Poisson and Student t-distributions. The estimation results that are derived from the pseudodata sets are not much different from that obtained by using the original data set, suggesting that the normality assumption seems to be adequate in our application. Admittedly, the normality assumption might be inadequate in other applications. If the departure from the normality assumption is only moderate, then the following estimation procedure may be executed:

(a) approximate the model with a linear Gaussian or mixture Gaussian alternative,
(b) use the sequential Kalman filter to draw samples of the state variables from the alternative model and
(c) use the Metropolis–Hastings algorithm to accept or reject the samples drawn.

If the departure from the normality assumption is significant, then the particle Markov chain Monte Carlo method may be used instead. This alternative estimation method is discussed in the work of Andrieu *et al.* (2010) and has been applied in the context of mortality modelling by Fung *et al.* (2017). One drawback of this alternative method is that it is very time consuming to implement.

Another limitation of our modelling approach is that it does not consider the possibility of structural changes. Although there are statistical tests for structural changes in historical mortality (Coelho and Nunes, 2011; Li *et al.*, 2011; Li and Li, 2017; Van Berkum *et al.*, 2016), none of them can be applied to data sets with discontinuities. Further, we have not considered the possible 'rotation' of age patterns of mortality decline, which has attracted considerable attention since it was first studied by Li *et al.* (2013). The investigation of these issues in the context of Chinese mortality is left for future research. Finally, we note that the issue of data quality is beyond the scope of this paper. Future research warrants an analysis of the quality of the Chinese mortality data set, drawing on Cairns *et al.* (2016).

Acknowledgements

The authors thank the journal Editors and three referees for their very helpful comments. This work is supported by research grants from the Natural Sciences and Engineering Research Council of Canada (discovery grant RGPIN-356050-2013) and the Society of Actuaries Center of Actuarial Excellence Program.

References

Andrieu, C., Doucet, A. and Holenstein, R. (2010) Particle Markov chain Monte Carlo methods (with discussion). *J. R. Statist. Soc.* B, **72**, 269–342.

Banister, J. and Hill, K. (2004) Mortality in China 1964-2000. *Popln Stud.*, **58**, 55–75.

Basel Committee on Banking Supervision (2013) Longevity risk transfer market: market structure, growth drivers and impediments, and potential risks. *Technical Report*. Basel Committee on Banking Supervision, Basel.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C. and Ljung, G. M. (2015) *Time Series Analysis: Forecasting and Control*, 5th edn. Hoboken: Wiley.

Brouhns, N., Denuit, M. and Vermunt, J. K. (2002) A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insur. Math. Econ.*, **31**, 373–393.

Cai, F. (2010) Demographic transition, demographic dividend, and Lewis turning point in China. *China Econ. J.*, **3**, no. 2, 107–119.

Cairns, A. J., Blake, D. and Dowd, K. (2006) A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *J. Risk Insur.*, **73**, 687–718.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A. and Balevich, I. (2009) A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *Nth Am. Act. J.*, **13**, 1–35.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D. and Khalaf-Allah, M. (2011) Bayesian stochastic mortality modelling for two populations. *Astin Bull.*, **41**, 29–59.
Cairns, A. J. G., Blake, D., Dowd, K. and Kessler, A. R. (2016) Phantoms never die: living with unreliable population data. *J. R. Statist. Soc.* A, **179**, 975–1005.
Carpenter, J. R. and Kenward, M. G. (2013) *Multiple Imputation and Its Application*. Chichester: Wiley.
Carter, C. K. and Kohn, R. (1994) On Gibbs sampling for state space models. *Biometrika*, **81**, 541–553.
Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006) Deviance information criteria for missing data models. *Baysn Anal.*, **1**, 651–673.
China Insurance Regulatory Commission (2011) *Yearbook of China's Insurance*. China Insurance Yearbook Publishing House.
China Insurance Regulatory Commission (2013) *Yearbook of China's Insurance*. China Insurance Yearbook Publishing House.
Coelho, E. and Nunes, L. C. (2011) Forecasting mortality in the event of a structural change. *J. R. Statist. Soc.* A, **174**, 713–736.
Czado, C., Delwarde, A. and Denuit, M. (2005) Bayesian Poisson log-bilinear mortality projections. *Insur. Math. Econ.*, **36**, 260–284.
Delwarde, A., Denuit, M. and Eilers, P. (2007) Smoothing the Lee-Carter and Poisson log-bilinear models for mortality forecasting: a penalized log-likelihood approach. *Statist. Modllng*, **7**, 29–48.
Fung, M. C., Peters, G. W. and Shevchenko, P. V. (2017) A unified approach to mortality modelling using state-space framework: characterisation, identification, estimation and forecasting. *Ann. Act. Sci.*, **11**, 343–389.
Girosi, F. and King, G. (2008) *Demographic Forecasting*. Princeton: Princeton University Press.
Graham, J. W. (2012) *Missing Data: Analysis and Design*. New York: Springer Science and Business Media.
Graham, J. W., Taylor, B. J., Olchowski, A. E. and Cumsille, P. E. (2006) Planned missing data designs in psychological research. *Psychol. Meth.*, **11**, 323–343.
Harvey, A. C. (1991) *Forecasting, Structural Time Series Models and the Kalman Filter*, 1st edn. Cambridge: Cambridge University Press.
Hilton, J., Dodd, E., Forster, J. J. and Smith, P. W. F. (2019) Projecting UK mortality by using Bayesian generalized additive models. *Appl. Statist.*, **68**, 29–49.
Huang, F. and Browne, B. (2017) Mortality forecasting using a modified continuous mortality investigation mortality projections model for China I: methodology and country-level results. *Ann. Act. Sci.*, **11**, 20–45.
Jiang, Q., Song, W. and Sánchez-Barricarte, J. J. (2013) Forecasting China's mortality. *Popln Rev.*, **52**, 87–98.
Kim, C.-J. and Nelson, C. R. (1999) *State-space Models with Regime Switching*, 1st edn. Cambridge: MIT Press.
Kogure, A., Kitsukawa, K. and Kurachi, Y. (2009) A Bayesian comparison of models for changing mortalities toward evaluating longevity risk in Japan. *Asia-Pacif. J. Risk Insur.*, **3**, no. 2, 1–21.
Koopman, S. J. and Durbin, J. (2000) Fast filtering and smoothing for multivariate state space models. *J. Time Ser. Anal.*, **21**, 281–296.
Lavely, W. and Mason, W. M. (2006) An evaluation of the one percent clustered sample of the 1990 census of China. *Demog. Res.*, **15**, 329–346.
Lee, R. D. and Carter, L. R. (1992) Modeling and forecasting U.S. mortality. *J. Am. Statist. Ass.*, **87**, 659–671.
Leng, X. and Peng, L. (2016) Inference pitfalls in Lee-Carter model for forecasting mortality. *Insur. Math. Econ.*, **70**, 58–65.
Lewis, W. A. (2013) *Theory of Economic Growth*, vol. 7. New York: Routledge.
Li, J. (2014) An application of MCMC simulation in mortality projection for populations with limited data. *Demog. Res.*, **30**, 1–48.
Li, S.-H. and Chan, W.-S. (2005) Outlier analysis and mortality forecasting: the United Kingdom and Scandinavian countries. *Scand. Act. J.*, no. 3, 187–211.
Li, J. S.-H., Chan, W.-S. and Cheung, S.-H. (2011) Structural changes in the Lee-Carter mortality indexes: detection and implications. *Nth Am. Act. J.*, **15**, 13–31.
Li, J. S.-H., Hardy, M. R. and Tan, K. S. (2009) Uncertainty in mortality forecasting: an extension to the classical Lee-Carter approach. *Astin Bull.*, **39**, 137–164.
Li, N., Lee, R. and Gerland, P. (2013) Extending the Lee-Carter method to model the rotation of age patterns of mortality decline for long-term projections. *Demography*, **50**, 2037–2051.
Li, N., Lee, R. and Tuljapurkar, S. (2004) Using the Lee-Carter method to forecast mortality for populations with limited data. *Int. Statist. Rev.*, **72**, 19–36.
Li, H. and Li, J. S.-H. (2017) Optimizing the Lee-Carter approach in the presence of structural changes in time and age patterns of mortality improvements. *Demography*, **54**, 1073–1095.
Li, Q., Reuser, M., Kraus, C. and Alho, J. (2009) Ageing of a giant: a stochastic population forecast for China, 2006-2060. *J. Popln Res.*, **26**, 21–50.
Mealli, F. and Rubin, D. B. (2015) Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, **102**, 995–1000.
Organisation for Economic Co-operation and Development (2014) Mortality assumptions and longevity risk: implications for pension funds and annuity providers. *Technical Report*. Organisation for Economic Co-operation and Development Publishing.
Pedroza, C. (2006) A Bayesian forecasting model: predicting U.S. male mortality. *Biostatistics*, **7**, 530–550.

Peng, X. (2011) China's demographic history and future challenges. *Science*, **333**, 581–587.
Ping, A., Tan, K. S. and Li, J. S.-H. (2013) Forecasting mortality in the presence of missing data: an application to Chinese population. *9th Int. Longevity Risk and Capital Markets Solutions Conf., Beijing*.
Renshaw, A. E. and Haberman, S. (2003) On the forecasting of mortality reduction factors. *Insur. Math. Econ.*, **32**, 379–401.
Renshaw, A. E. and Haberman, S. (2006) A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insur. Math. Econ.*, **38**, 556–570.
Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*. New York: Chapman and Hall.
Schafer, J. L. and Graham, J. W. (2002) Missing data: our view of the state of the art. *Psychol. Meth.*, **7**, 147–177.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.
Tuljapurkar, S., Li, N. and Boe, C. (2000) A universal pattern of mortality decline in the G7 countries. *Nature*, **405**, 789–792.
Van Berkum, F., Antonio, K. and Vellekoop, M. (2016) The impact of multiple structural changes on mortality predictions. *Scand. Act. J.*, no. 7, 581–603.
Van Buuren, S. (2012) *Flexible Imputation of Missing Data*. Boca Raton: CRC Press.
Wang, X.-J. and Huang, S.-L. (2011) Comparison and selection of stochastic mortality models in China. *Popln Econ.*, **184**, 82–86.
Wilmoth, J. R. (1993) Computational methods for fitting and extrapolating the Lee-Carter model of mortality change. *Technical Report*. Department of Demography, University of California at Berkeley, Berkeley.
Zhao, B. B. (2012) A modified Lee-Carter model for analysing short-base-period data. *Popln Stud.*, **66**, 39–52.
Zhao, B. B., Liang, X., Zhao, W. and Hou, D. (2013) Modeling of group-specific mortality in China using a modified Lee-Carter model. *Scand. Act. J.*, no. 5, 383–402.
Zhou, R. and Li, J. S.-H. (2013) A cautionary note on pricing longevity index swaps. *Scand. Act. J.*, no. 1, 1–23.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'A Bayesian approach to developing a stochastic mortality model for China'.