that the missingness is entirely under the data collector's control (Graham et al., 2006, p.324), and the fact that the missing data were intentionally not recorded (i.e., planned missingness) (Schafer and Graham, 2002).

For item (iii), 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 (iii) as Missing at Random (MAR). One may, however, argue that the missingness for item (iii) 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 MAR is not a perfect fit for item (iii) as the missingness may be correlated with age and/or exposure and is not completely out of control of the data collector. The available information does not permit us to prove or disprove this argument, and in this regard, the missing mechanism for item (iii) should be viewed as an assumption.


3.1 The 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:

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

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

\[
y_t = \alpha_t + \beta_{x,t} + \epsilon_t, \quad \epsilon_t \sim \text{MVN}(0, \sigma^2 T),
\]

where \( \sigma^2 \) takes three distinct values, reflecting the nature of the data for year \( t \):

\[
s^2_t = \begin{cases} 
  s^2 & t \in T_c \\
  s^2_{0.1\%} & t \in T_{0.1\%} \\
  s^2_{0.1\%} & t \in T_{0.1\%} 
\end{cases},
\]

where \( T_c \), \( T_{1\%} \), and \( T_{0.1\%} \) 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 national population, respectively. The conditional posterior distribution of \( s^2 \) 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 \( t \in T_c \), \( t \in T_{1\%} \), or \( t \in T_{0.1\%} \), whichever is appropriate.

The conditional posterior distributions of \( \alpha_s \), \( \beta_{x,t} \), \( c \) and \( d \) have to be modified accordingly. Specifically, \( s^2 \) in equations (6) to (9) and (14) to (17) is replaced by \( s^2_t \). We also need to replace \( s^2 \) by \( s^2_t \) in the algorithm for the sequential Kalman filter.

5 Estimation Results

In this section, we estimate the full model with the adaptations described in Sections 4.3 and 4.4 to the Chinese mortality data. The estimation is completed 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 the sake of space we show only those for males in the main text. The estimation results for females are provided in Annex E.

5.1 Convergence of Parameters

Using the initial values 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, 5,000 samples are recorded and used to form the joint posterior distribution. Further details concerning convergence of parameters are provided in Annex A.

5.2 Posterior Distributions of Parameters

The fan charts in Figure 3 display the posterior distributions of $\alpha_x$, $\beta_x$ and $\kappa_t$ 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%, ..., 90% credible intervals with progressively lighter shadings. The width of a fan chart indicates the level of parameter uncertainty entailed.

![Figure 3: The posterior distributions of $\alpha_x$, $\beta_x$, and $\kappa_t$ for $x = 0, \ldots, 99$ and $t = 1981, \ldots, 2014$.]

The pattern of the posterior means of $\alpha_x$'s 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 $\alpha_x$'s and $\beta_x$'s are smooth.

As expected, the posterior means of $\kappa_t$'s exhibit a downward trend. The variation in the uncertainty surrounding $\kappa_t$ 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 $\kappa_t$ is particularly wide.

Figure 4 depicts the posterior distributions of $s_{2.1}^2$, $s_{1.1}^2$, and $s_{0.1.1}^2$. The posterior means of $s_{2.1}^2$, $s_{1.1}^2$, and $s_{0.1.1}^2$ are highly 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 $s_{2.1}^2$ (which represents the extent of sampling uncertainty when the data are obtained from censuses) is the lowest, while the posterior mean of $s_{0.1.1}^2$ (which represents the extent of sampling uncertainty when the data are obtained from surveys of 0.1% of the population) is the highest.

Finally, Figure 5 shows the posterior distributions of the parameters in the state equation. It is noteworthy that $\mu$ is subject to a great deal of parameter uncertainty, with a posterior distribution that spans
6 Prediction Results

In this section, we present the prediction results that are derived 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 proposed adaptations.

6.1 Baseline Results

Predictions of future mortality are obtained from the posterior predictive distribution of \( y_{t_0+u} = y_{t_0} + \eta_u \), \( u = 1, 2 \). Assuming that historical and future log central death rates are independently distributed.
Figure 6: The 90% prediction intervals for the central death rates at ages 0, 10, 20, ..., and 90 for the 2015 to 2049 period, based on \( s^2 \) (solid lines), \( s^2_{\text{10\%}} \) (dashed lines) and \( s^2_{0.1\%} \) (dotted lines). The central prediction is presented by the line in the middle of the prediction intervals. The historical central death rates are shown in circles (for those obtained from censuses), squares (for those obtained from surveys of 1% of the population) and crosses (for those obtained from surveys of 0.1% of the population).

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

1. Calculate \( E[\mu|Y], E[\sigma^2|Y], E[\kappa_{t_0+n_y-1}|Y], E[\alpha|Y], E[\beta|Y] \) and \( E[s^2_{t_0+n_y-1+u}|Y] \) from the posterior distributions of the model parameters.

2. Simulate a realization of \( \kappa_{t_0+n_y-1+u} \) for each \( u = 1, 2, \ldots \) from its predictive distribution,

\[
N(E[\kappa_{t_0+n_y-1}|Y] + uE[\mu|Y], uE[\sigma^2|Y]).
\]
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 $\text{Var}(\varepsilon_{t})$ 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, the interval for the 1-year ahead projection is wide enough to encompass all but only one historical values.
7 Validation of the Proposed Modeling/Estimation Approach

7.1 Generating Pseudo Data Sets

We use pseudo data sets to validate the proposed modeling approach. Each pseudo data 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.

1. The exposure counts are taken as those in the actual data set (for Chinese males). This step implies that the pseudo data set is subject to exactly the same missing value problems described in items (i) and (ii) in Section 2.

This article is protected by copyright. All rights reserved
The central and 90% interval predictions of the log central death rates at ages 0, 10, 20, ..., and 90 in the 2015 to 2045 period, derived from the model without the adaptations proposed in Sections 4.3 and 4.4 (solid lines) and from the full model (based on $\kappa^2$) (dotted lines). The historical central death rates obtained from censuses are shown in circles.

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

$$E_{x,t} \exp(E[\alpha_x | Y] + E[\beta_x | Y]E[\kappa | Y]),$$

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

3. For each age-time cell in which the exposure count is available, calculate the simulated log central death rate as $\ln(\hat{D}_{x,t}) = \ln(\hat{D}_{x,t}/E_{x,t})$, where $\hat{D}_{x,t}$ denotes the simulated death count for age $x$.
and year \( t \). When \( D_{x,t} = 0 \), the corresponding log central death rate is recorded as "unreported", thereby resembling the missing data problem described in item (iii) in Section 2.

Reestimating our model to such pseudo data sets allows us to examine how our model may perform if the normality assumption does not hold and if heteroscedasticity exists (as the pseudo data generation procedure implies that the variances at different ages and in different calendar years are generally different).

Figure 11 summarizes the structure of one of the generated pseudo data sets. This structure is identical to that of the actual data set for Chinese males (the left panel of Figure 1), 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.

![Figure 11: A Lexis diagram summarizing the structure of one of the generated pseudo data sets.](image)

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

As implied by Step 2 in the procedure above, the data generating process underlying the pseudo data sets is the Lee-Carter model with parameters equal to the posterior parameter means derived from the actual data set. Hence, if our proposed modeling/estimation approach is appropriate, then when applied to the pseudo data sets, it should result in parameter estimates that are sufficiently close to the parameters in the underlying data generating process.

### 7.2 Validation Results

The solid lines in Figure 13 show the 90% credible intervals of \( \alpha_x, \beta_x \) and \( \kappa_t \) for \( x = 0, \ldots, 99 \) and \( t = 1981, \ldots, 2014 \), calculated using the posterior distributions derived from one of the pseudo data sets. They should be compared against the corresponding "true" parameters (i.e., those in the underlying data generating process), which are shown in dashed lines in the same figure.

For \( \alpha_x \) and \( \beta_x \), 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 of the "true" values of \( \alpha_x \) and \( \beta_x \), providing

---

This article is protected by copyright. All rights reserved
a strong support for using the proposed methods to estimate the model in the presence of missing data and finite sampling.

For $\kappa_t$, 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 the proposed estimation methods can estimate $\kappa_t$ reasonably accurately in the presence of missing data and finite sampling. Note also that they have comparable widths to those for the $\kappa_t$'s derived from the real Chinese mortality data set (see the right panel in Figure 3).

When applied to the rest of the generated pseudo data sets, the estimates of $\alpha_x$ and $\beta_x$ remain almost unchanged. The estimates of $\kappa_t$ vary observably when different data sets are used, but the 90% credible intervals consistently capture most of the "true" parameter values (see Figure 14).

We acknowledge that this analysis may not be comprehensive, as the Poisson pseudo data sets do not capture, for example, the possibility that the variance of a random death count is greater than the mean. In Annex B, we present two additional analyses that are based on pseudo data sets generated from (log-
Figure 14: The 90% credible intervals (solid lines) of \( \kappa_t \) for \( t = 1981 \ldots 2014 \), calculated using the posterior distributions derived from six of the generated pseudo data sets. The corresponding "true" parameter values are shown in dashed lines.

7.3 A Cautionary Note

As pointed out by Li and Chan (2005) and Zhou and Li (2013), mortality forecasts produced by the original Lee-Carter model are highly sensitive to the values of \( \kappa_t \) in the beginning and ending years of the data sample (i.e., \( \kappa_{t_0} \) 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, let us 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

\[
E[\kappa_{t_0+n_y-1}|Y] + u E[\mu|Y].
\]

The above 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) \( \mu \), which determines the gradient of the expected trajectory. Further-
more, as shown in expression (10), the mean of the conditional posterior distribution of \( \mu \) is

\[
\frac{K_{t_0 + n_y - 1} - K_{t_0}}{n_y - 1},
\]

which is determined exclusively by \( K_{t_0 + n_y - 1} \) and \( K_{t_0} \).

As discussed in the previous sub-section, the estimates of \( \kappa_t \) vary across the pseudo data sets. The colored jagged lines in Figure 15 represent the posterior means of \( \kappa_t \) for \( t = 1981, \ldots, 2014 \), obtained from three different pseudo data sets. The variation in the posterior means of \( \kappa_t \) 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. In the final year (2014) of the pseudo data sets, the exposure 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 pseudo data sets. This phenomenon can be observed in the colored dashed lines in Figure 15, which represent the expected trajectories of \( \kappa_{2015}, \kappa_{2016}, \ldots \) obtained from three of the generated pseudo samples. These expected trajectories have noticeably different levels and gradients. Some of them are quite different from the expected trajectory implied by the "true" parameters, which is represented by the black dashed line in the same figure.

![Figure 15: The posterior means of \( \kappa_{1981}, \ldots, \kappa_{2014} \) (colored solid lines), the expected trajectories of \( \kappa_{2015}, \kappa_{2016}, \ldots \) (colored dashed lines), and the line that passes through \( E[\kappa_{2014} | Y] \) and has a slope of \( E[\mu | Y] \) (colored dotted lines), obtained from three of the generated pseudo data sets (without any adjustment). The black lines represent the corresponding values implied by the "true" parameters.](image)

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

\[
E[\kappa_{2014} | Y] + uE[\mu | Y]
\]

to

\[
E[\kappa_{2010} | Y] + (u + 4)E[\mu | Y]
\]

(18)
and modify the conditional posterior mean of \( \mu \) from

\[
K_{2014} - K_{1981} = 33
\]

(19)

to

\[
K_{2010} - K_{1981} = 29
\]

while all of 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 \( K_{2015}, K_{2016}, \ldots \). The benefit of the proposed adjustments is demonstrated in Figure 16, from which we observe that with the proposed adjustments the resulting expected trajectories of \( K_{2015}, K_{2016}, \ldots \) exhibit less variation across the pseudo data sets and are all close to the expected trajectory implied by the "true" parameters.

Figure 16: The posterior means of \( K_{1981}, \ldots, K_{2014} \) (colored solid lines), the expected trajectories of \( K_{2015}, K_{2016}, \ldots \) (colored dashed lines), and the line that passes through \( E(\mu | Y) \) and has a slope of \( E(\mu | Y) \) (colored dotted lines) obtained from three of the generated pseudo data sets (with the proposed adjustments). The black lines represent the corresponding values 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 associated with the data set being accounted for. Significant effort has been made to ensure that the end result satisfies the three important criteria 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 is able to 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 into the model. This feature makes our proposed approach stands 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
A Bayesian Approach to Developing a Stochastic Mortality Model for China

Johnny S.-H. Li, Kenneth Q. Zhou, Xiaobai Zhu, Wai-Sum Chan, and Felix Wai-Hon Chan

Department of Statistics and Actuarial Science, University of Waterloo, Canada
Department of Economics, The University of Melbourne, Australia
Department of Finance, The Chinese University of Hong Kong, Hong Kong, P.R. China
Department of Professional Legal Education, The University of Hong Kong, Hong Kong, P.R. China

[Received July 2017, Final revision April 2019]

Abstract: Stochastic mortality models have a wide range of applications. They are particularly important for analyzing Chinese mortality, which is subject to rapid and uncertain changes. However, owing to data-related problems, stochastic modeling of Chinese mortality has not been given adequate attention. In this paper, 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, (1) acknowledge the fact that the data are obtained from different sources, and (2) mitigate the erratic behavior of the parameter estimates that arises from the data limitations. The proposed approach yields stochastic mortality forecasts that are in line with both the trend and variation of the historical observations. We further use simulated pseudo data sets with resembling limitations to validate the proposed approach. The validation result confirms our approach's success in dealing with the limitations of the Chinese mortality data.

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

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.

Fueled by the infamous one-child policy, improved longevity has accelerated the aging of 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 dividend (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 No. 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 yuan at 2013 year-end. Although this amount seems large, whether it is sufficient...
Organisation for Economic Co-operation and Development

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 yuan in 2010 to 21.5 billion yuan in 2012 (China Insurance Regulatory Commission, 2011, 2013). The uncertainty surrounding the evolution of mortality affects all life annuities sold in China systematically and, in extreme circumstances, may result in a systemic failure of the Chinese insurance industry (Basel Committee on Banking Supervision, 2013). Recently, the BCCB (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.

Policymakers 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 post-1981 period, leaving researchers with a rather short data series 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 65 in 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.

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 rather 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.

Criterion 2: The model should provide appropriate provisions for uncertainty.

Due 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.

Criterion 3: The model should be parsimonious and yield biologically reasonable projections.

This article is protected by copyright. All rights reserved
Owing 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 to 99 years, the model should be able to capture the fall in mortality with age during childhood and early adulthood, respectively. For this reason, we do not consider models such as the Cairns-Blake-Dowd (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 explain why they do not satisfy all of the criteria 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 estimate the Lee-Carter model (Lee and Carter, 1992) using three years of complete mortality data that are 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 proposed by Li et al. (2004) can be extended to incorporate the survey data (whereby some death counts are not reported) using weighted least squares with a zero weight being assigned to the cells with no reported death counts (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 use three years of census data, Huang and Browne (2017) who consider 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 fit a multiple linear regression to the available data and then fill 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) argue 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) estimate 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 purposes. Such a two-stage estimation approach for estimating stochastic mortality models has been criticized heavily by researchers including Leng and Peng (2016) who point out its inference pitfalls and Czado et al. (2005) who assert 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 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) adjusts 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 et al. (2009), who develop a stochastic population forecast (which involves a stochastic mortality forecast) for China using data.

This article is protected by copyright. All rights reserved
graphic 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 modeling work for China, which has a distinct history of public health and economic development.

In this paper, we adopt a Bayesian approach to stochastically model 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 into the model, and therefore Criterion 1 can be met. With Gibbs sampling, the proposed approach yields a joint posterior distribution of all of the model parameters, which allows the user to holistically gauge 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: Pedroza’s work is built on the Lee-Carter model, which appears to meet Criterion 3 as it is relatively parsimonious compared to other models that are applicable to the full age range of 0 to 99. Its Gaussian state/space specification facilitates efficient estimation, as it allows 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), a number of significant adaptations are made to suit the situation we are confronting.

First, we adapt the multiple imputation algorithm proposed by Pedroza (2006). Pedroza (2006) and a few other researchers (e.g., Czado et al., 2005; Cairns et al., 2011) have pointed out that Bayesian approaches can be used to handle missing data in the context of stochastic mortality modeling, but none of them have applied these approaches to a real mortality data set that is subject to problems 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 mentions 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 resorts 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 of 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 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 considered by Renshaw and Haberman (2003) to smooth the pattern of the age response parameters (denoted by \( \beta_x \) in this and many other papers), which determine the expected rates of mortality improvement 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 \( \beta_x \) for the Chinese population are very jagged across ages if additional smoothing is not imposed. This outcome is possibly due to the fact that the data array used is small in size and has a lot of discontinuities. A jagged pattern of \( \beta_x \) is undesirable, because, for example, it is difficult to explain why mortality rates at adjacent ages evolve at highly different expected speeds. The adaptation we propose ensures that the posterior mean of \( \beta_x \) follows a logical relationship with age, allowing the resulting model to better meet Criterion 3 (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 not met). The added feature permits the variance

This article is protected by copyright. All rights reserved
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/1467986x/series-a-datasets

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 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 pseudo data sets to validate the ability of the proposed method for handling the data-related problems. The pseudo data sets are randomly generated from some assumed Lee-Carter parameters and are constructed in such a way that they possess the limitations of the actual Chinese mortality data set. When applied to the pseudo data sets, the proposed modeling estimation approach produces 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 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 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 modeling 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 made to the work of Pedroza (2006). Section 7 details the validation with pseudo data sets. Finally, concluding remarks and suggestions for future research are given in Section 8.

2. The Available Data

The mortality data (age-specific death and mid-year population counts) used in this paper are 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 0-89 and a time period of 1981-2014, covering 36,000 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.

- No death count is reported for a number of individual age-time cells (e.g., age 7, 2008, males).

Overall, there are 1,222,971,229 age-time cells in the data set for males (females), with no death and/or population count, representing 35.21% (36.15%) of the total number of age-time cells.

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 Figure 1.

For items (i) and (ii), the missingness is due entirely to the fact that data were not collected, and is treated as Missing Completely at Random (MCAR) in this paper. This treatment is justified by the fact

This article is protected by copyright. All rights reserved
that the missingness is entirely under the data collector's control (Graham et al., 2006, p. 324), and the fact that the missing data were intentionally not recorded (i.e., planned missingness) (Schafer and Graham, 2002).

For item (h), 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 (h) as Missing at Random (MAR). One may, however, argue that the missingness for item (h) 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 MAR is not a perfect fit for item (h) as the missingness may be correlated with age and/or exposure and is not completely out of control of the data collector. The available information does not permit us to prove or disprove this argument, and in this regard, the missing mechanism for item (h) should be viewed as an assumption.


3.1 The Lee-Carter Model in a Gaussian Formulation

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

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

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

This article is protected by copyright. All rights reserved.
independent and identically distributed

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

\[
\kappa_t = \mu + \kappa_{t-1} + \omega_t,
\]

(2)

where \( \mu \) is the drift term and \( \{ \omega_t \} \) is a sequence of \( \mathcal{N} \) normal random variables with a zero mean and a variance of \( \sigma^2 \). A non-stationary process is used instead of a stationary one for two main reasons. If a stationary (mean-reverting) process is used, the expected trajectory of future mortality would revert back to a certain (average) historical level. Such a trajectory seems counterintuitive. 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:

\[
\begin{align*}
\sum_{x=x_0}^{x_0+n_a-1} \beta_x &= 1 \\
\sum_{t=t_0}^{t_0+n_y-1} \kappa_t &= 0
\end{align*}
\]

(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 used in this paper, \( x_0 = 0 \), \( t_0 = 1981 \), \( n_a = 100 \) and \( n_y = 34 \).

Pedroza (2006) chooses to use a Gaussian formulation, meaning that the error term \( \varepsilon_{x,t} \) 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, et al., 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 log of zero problem). If the Poisson formulation is used instead, the imputed death rate for a blank age-time cell may be very small 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 we consider. That is

\[3.2 \text{ Gibbs Sampling}\]

The Gaussian Lee-Carter model contains \( 2n_a + n_y + 3 \) parameters in total, including

\[\alpha_{x_0}, \ldots, \alpha_{x_0+n_a-1}, \beta_{x_0}, \ldots, \beta_{x_0+n_y-1}, k_{t_0}, \ldots, k_{t_0+n_y-1}, \mu, s^2 \text{ and } \sigma^2.\]

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 \( n_0 \times n_y \) 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 \( \Theta_{-j} = \{ \theta_1, \ldots, \theta_{2n_a+n_y+3} \} \) to represent \( \Theta \) excluding its \( j \)th entry. In each iteration of Gibbs sampling, a sample of \( \theta_j \) is drawn from its conditional
The conditional posterior distribution of $\theta_j$ depends on the assumed conditional prior distribution $\pi(\theta_j|\Theta_{-\theta_j})$. For $\alpha_x, \beta_x, \mu, s^2$ and $\sigma^2$, improper prior distributions are assumed, that is, $\pi(\theta_j|\Theta_{-\theta_j}) \propto 1$ for $\alpha_x, \beta_x, \mu$, and $\pi(\theta_j|\Theta_{-\theta_j}) \propto \theta_j^{-1}$ for $s^2$ and $\sigma^2$. The derivation of the conditional posterior distributions of these parameters can be found in the paper by Pedroza (2006). For $\kappa$, 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 $\alpha_x, \beta_x$ and $\kappa$ are rescaled so that the two constraints specified in equation (3) are satisfied. In principle, it is possible to include the constraints into the prior distributions for $\kappa$ and $\beta_x$. However, in an application that involves a large volume of missing data, this alternative method (explained in Annex F) would result in numerical instability.

3.3 Kalman Filter

Given a sample of $\alpha_x, \beta_x, \mu, s^2$ and $\sigma^2$, a sample of $\kappa$ can be generated 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:

$$ y_t = \alpha + \beta \kappa_t + \epsilon_t, \quad \epsilon_t \sim \text{MVN}(0, \sigma^2 I), $$

(4)

$$ \kappa_t = \mu + \kappa_{t-1} + \omega_t, \quad \omega_t \sim \text{N}(0, \omega^2), $$

(5)

where $y_t = (y_{x_0,t}, \ldots, y_{x_0+n_a-1,t})'$, $x = (\alpha_{x_0}, \ldots, \alpha_{x_0+n_a-1})'$, $\beta = (\beta_{x_0}, \ldots, \beta_{x_0+n_a-1})'$, $\epsilon_t = (\epsilon_{x_0,t}, \ldots, \epsilon_{x_0+n_a-1,t})'$ and $I$ 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 $\kappa_t$ on the basis of $Y_t = (y_t, y_{t-1}, \ldots, y_0)$, which contains all information up to and including time $t$. We let $k_t := \mathbb{E}(\kappa_t|Y_t)$ and $p_t := \text{var}(\kappa_t|Y_t)$. The filtering process derives $k_t$ and $p_t$ using the following recursion, which runs from $t = t_0$ to $t = t_0 + n_y - 1$:

$$ k_t = \mu + k_{t-1} + \gamma_t \eta_t, $$

$$ p_t = (1 - g_t \beta)(p_{t-1} + \sigma^2), $$

$$ \eta_t = y_t - \alpha - \beta (\mu + k_{t-1}), $$

$$ F_t = \beta(p_{t-1} + \sigma^2)\beta' + s^2 I, $$

$$ g_t = (p_{t-1} + \sigma^2)\beta F_t^{-1}, $$

where $\eta_t$ is an $n_a \times 1$ column vector representing the prediction error of $y_t$ based on $k_{t-1}$, $F_t$ is the $n_a \times n_a$ covariance matrix of the prediction error and $g_t$ 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.

Throughout the appendix, a conditional posterior distribution refers to a full conditional posterior distribution.
3.3.2 The Smoothing Process

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

1. For \( t = t_0 + n_y - 1 \), draw a sample of \( \kappa_t \) from a normal distribution with a mean of \( k_t = \text{E}(\kappa_t | Y_t) \) and a variance of \( p_t = \text{Var}(\kappa_t | Y_t) \).

2. For \( t = t_0 + n_y - 2, \ldots, t_0 \) that is
   
   (a) calculate \( h_t \) and \( v_t \) using the value of \( \kappa_{t+1} \) obtained from the previous step and the following equations:
   
   \[
   h_t = k_t + p_t (p_t + \sigma^2)^{-1} (\kappa_{t+1} - \mu - k_t) \\
   v_t = p_t (p_t + \sigma^2)^{-1}
   \]

   (b) draw a sample of \( \kappa_t \) 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 \( \alpha_x, \beta_x, \mu, s^2 \) and \( \sigma^2 \) to form a realization of the joint posterior distribution of the \( 2n_a + n_y + 3 \) parameters.

3.4 Multiple Imputation

Pedroza (2006) further suggests 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.

1. At the beginning of each iteration of Gibbs sampling, impute all \( m \) the missing data points. Specifically, if the central death rate at age \( x^* \) and year \( t^* \) is missing, impute the value of \( \text{ln}(m_{x^*, t^*}) \) from
   \[
   \text{ln}(m_{x^*, t^*} | \theta) \sim N(\alpha_{x^*, t^*} + \beta_{x^*, t^*} \kappa_{t^*}, s^2)
   \]

   where \( \alpha_{x^*, t^*} \), \( \beta_{x^*, t^*} \), \( \kappa_{t^*} \) and \( s^2 \) are taken as the values drawn in the previous iteration.

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 using the algorithms outlined in the previous subsections.

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 The Abridged Multiple Imputation

Pedroza (2006) does not implement the multiple imputation method 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.

This article is protected by copyright. All rights reserved
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 the non-convergence problem is exacerbated if (a) the number of data points is small compared to 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 we are confronting.

To circumvent this problem, we choose to impute only the individual missing (unreported) values (item (e) in Section 2) but not the missing values that appear in blocks (items (f) and (h) in Section 2). The rationale is that, as discussed in Section 2, the missing values that appear in blocks fit the definition of MCAR. Not imputing data that are MCAR does not affect the validity of statistical inferences and predictions (Graham (2012), p. 158; Van Buuren (2012), p. 35). We emphasize that (as shown in subsequent sections) we still obtain a joint posterior distribution of all $\mathbf{\theta}$ 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 $\alpha_x$, $\beta_x$ and $\sigma^2$ have to 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 $\alpha_x$ is $N(\alpha, b_\alpha)$, where

$$a_\alpha = \left( \sum_{t=t_0}^{t_0+n_y-1} \frac{1}{s^2} \right)^{-1} \left( \sum_{t=t_0}^{t_0+n_y-1} \ln \left( \frac{n_{x,t}}{s^2} \right) \right)$$

$$b_\alpha = \left( \sum_{t=t_0}^{t_0+n_y-1} \frac{1}{s^2} \right)^{-1}$$

and $1_{x,t}$ 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 $\beta_x$ is $N(\alpha, b_\beta)$, where

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

and

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

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

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

and a rate parameter of

$$b_s = \frac{\sum_{x=x_0}^{x_0+n_a-1} \sum_{t=t_0}^{t_0+n_y-1} \left( \ln \left( \frac{n_{x,t}}{s^2} \right) \right)}{2}.$$
The conditional posterior distributions of $\mu$ and $\sigma^2$ require no modification. They are provided below for completeness.

The conditional posterior distribution of $\mu$ is $\mathcal{N}(a_\mu, b_\mu)$, where

$$ a_\mu = \frac{\kappa_{t_0+n_y-1} - \kappa_{t_0}}{n_y - 1} \quad \text{and} \quad b_\mu = \frac{\sigma^2}{n_y - 1}. $$

(10)

The conditional posterior distribution of $\sigma^2$ is an inverse gamma distribution with shape parameter of 

$$ a_\sigma = n_y - 1 $$

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.

1. Obtain a crude estimate of each missing death exposure count by linearly interpolating between adjacent (non-missing) values. A completed data set is then obtained.
2. Get crude estimates of the model parameters using Poisson maximum likelihood (Wilmoth, 1993; Brouhns et al., 2002).
3. Using the crude estimates from the previous step as initial values, apply the Gibbs sampling algorithm to the completed data set.
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.

The procedure above allows the Gibbs sampling algorithm from which the final estimated model is derived to start at a state that is close enough to the stationary state.

4.2 The Sequential Kalman Filter

In the abridged multiple imputation, we do not impute item (8) defined in Section 2. As a result, we do not always have a complete vector of observed imputed death rates $\hat{y}_t$, which is required in the third step of the filtering process described in Section 3.3.1.

To overcome this technical challenge, we replace the original Kalman filter with the sequential Kalman filter 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 $y_t$ in each step, incorporating the observed/imputed log central death rates into $k_t$ and $p_t$ one at a time.

Poisson maximum likelihood is regarded as a standard (frequentist-type) method for estimating the Lee-Carter model when there is 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 essentially the same if another method of obtaining crude estimates is used.
To implement the sequential Kalman filter, we first rewrite the state space form of the Lee-Carter model in a sequential representation:

\[
\ln(m_{x,t}) = \alpha_x + \beta_x \kappa_{x,t} + \epsilon_{x,t},
\]

\[
\kappa_{x,t} = \begin{cases} 
\kappa_{x-1,t} & \text{if } x_0 < x \leq x_0 + n_{a} - 1, \\
\mu + \kappa_{x_0+n_{a}-1,t-1} + \omega_t & \text{if } x = x_0 
\end{cases}
\]

where \( \epsilon_{x,t} \sim \mathcal{N}(0, s^2) \) and \( \omega_t \sim \mathcal{N}(0, \sigma^2) \) 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 \( \mathcal{Y}_{x,t} = \{y_{x,t}, y_{x-1,t}, \ldots, y_{x_0,t}, \mathcal{Y}_{t-1}\} \), 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 \( k_{x,t} := \hat{\mathbb{E}}(\kappa_t | \mathcal{Y}_{x,t}) \) and \( p_{x,t} := \sqrt{\text{Var}(\kappa_t | \mathcal{Y}_{x,t})} \) as follows.

For \( x = x_0 \), we have

\[
\begin{align*}
&k_{x_0,t} = \mu + k_{x_0+n_{a}-1,t-1} + g_{x_0,t} \eta_{x_0,t} \\
p_{x_0,t} = \sqrt{(1 - g_{x_0,t}^2) p_{x_0+n_{a}-1,t-1} + \sigma^2} \\
&\eta_{x_0,t} = y_{x_0,t} - \alpha_{x_0} - \beta_{x_0} (\mu + k_{x_0+n_{a}-1,t-1}) \\
g_{x_0,t} = \frac{(p_{x_0+n_{a}-1,t-1} + \sigma^2) \beta_{x_0}}{(p_{x_0+n_{a}-1,t-1} + \sigma^2) \beta_{x_0}^2 + s^2}
\end{align*}
\]

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

\[
\begin{align*}
k_{x_0,t_0} &= \frac{y_{x_0,t_0} - \alpha_{x_0}}{\beta_{x_0}} \\
p_{x_0,t_0} &= \frac{s^2}{\beta_{x_0}^2}
\end{align*}
\]

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 \( x_0 < x \leq x_0 + n_{a} - 1 \), we have

\[
\begin{align*}
k_{x,t} &= k_{x-1,t} + g_{x,t} \eta_{x,t} \\
p_{x,t} &= (1 - g_{x,t}^2) p_{x-1,t} \\
&\eta_{x,t} = y_{x,t} - \alpha_x - \beta_x k_{x-1,t} \\
g_{x,t} &= \frac{p_{x-1,t} \beta_x}{p_{x-1,t} \beta_x^2 + s^2}
\end{align*}
\]

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 to 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), p. 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 get the values of \( k_{t_0}, \ldots, k_{t_0+n_y-1} \) as

\[
k_{x_0+n_{a}-1,t} = \mathbb{E}(\kappa_t | \mathcal{Y}_{x_0+n_{a}-1,t}) = \mathbb{E}(\kappa_t | \mathcal{Y}_t) = k_t
\]

according to the definitions of \( k_{x,t}, k_t, \mathcal{Y}_t \) 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

\[
\begin{align*}
  k_{x,t} &= \begin{cases} 
    \mu + k_{x_0 + n_{x-1},t-1}, & x = x_0 \lor \\
    k_{x-1,t}, & x = x_0 + 1, \ldots, x_0 + n_\alpha - 1
  \end{cases} \\
  p_{x,t} &= \begin{cases} 
    p_{x_0 + n_{x-1},t-1} + \sigma^2, & x = x_0 \lor \\
    p_{x-1,t}, & x = x_0 + 1, \ldots, x_0 + n_\alpha - 1
  \end{cases}
\end{align*}
\]

respectively, so that no information concerning \(y_{x,t}\) is incorporated into \(k_{x,t}\) and \(p_{x,t}\).

### 4.3 Smoothing the Age-Response Parameters

Our preliminary work shows that without any special treatment, the Bayesian estimates of \(\beta_x\) derived from the Chinese mortality data set exhibit a very jagged pattern across ages. With such estimates of \(\beta_x\), the projected future mortality rates do not possess a smooth, logical age pattern. The same problem is also encountered in the work of Huang and Browne (2017) who apply the Lee-Carter model to a portion of the Chinese mortality data set. The abnormally jagged pattern of \(\beta_x\) 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 \(\beta_x\) across ages is needed.

There are a few methods for smoothing the pattern of \(\beta_x\) 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 \(\beta_x\) is defined as follows:

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

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 \(\left(\ln(x + 1) - \ln(x_j + 1)\right)^3\) 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

\[
\beta = A c
\]

where \(c = (c_0, \ldots, c_{r+3})^\top\), and \(A\) is an \(n_{x_0 + n_{x-1}} \times (r + 4)\) matrix of which the \(i\)th row is given by \((1, \ln(x_0 + i), \ln(x_0 + i)^2, \ln(x_0 + i)^3, \ln(x_0 + i) - \ln(x_1 + 1))^\top, \ldots, (\ln(x_0 + i) - \ln(x_r + 1))^\top\).

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

\[
\hat{c}_c = \left(\sum_{t=t_0}^{t_0+n_{x-1}} k_t^2 / s^2 A^\top_t A\right)^{-1} \sum_{t=t_0}^{t_0+n_{x-1}} k_t / s^2 A^\top_t (y_t - \alpha), \quad (14)
\]

and

\[
B_c = \left(\sum_{t=t_0}^{t_0+n_{x-1}} k_t^2 / s^2 A^\top_t A\right)^{-1}, \quad (15)
\]
and \( I_t \) is an \( n \times n \) diagonal matrix, of which the \( i \)th diagonal element equals \( 0 \) if \( m_{x(i)+1,i} \) is missing and not imputed, and \( 1 \) otherwise.

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

\[
\alpha = Ad,
\]

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

\[
a_d = \left( \sum_{t=t_0}^{t_0+n_2-1} \frac{1}{s^2} A'T_t A \right)^{-1} \left( \sum_{t=t_0}^{t_0+n_2-1} \frac{1}{s^2} A'T_t (y_t - \beta \kappa_t) \right)
\]

and

\[
B_d = \left( \sum_{t=t_0}^{t_0+n_2-1} \frac{1}{s^2} A'T_t A \right)^{-1}.
\]

The knot ages \( (x_1, \ldots, x_r) \) are evenly spaced between \( 0 \) and \( 70 \). The number of knots \( r \) is determined using the deviance information criterion (Spiegelhalter et al., 2002), which is designed to compare different Bayesian models. The Bayesian deviance is defined by

\[
D(\Theta) = -2 \ln(\text{Pr}(Y|\Theta)) + 2 \ln f(Y),
\]

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

\[
\text{DIC} = \text{E}[D(\Theta)|Y] + \text{E}[D(\Theta)|Y] - D[\text{E}(\Theta|Y)],
\]

where \( \text{E}[D(\Theta)|Y] \) measures the goodness of fit and \( \text{E}[D(\Theta)|Y] - D[\text{E}(\Theta|Y)] \) penalizes the use of more parameters. We can compute \( \text{E}[D(\Theta)|Y] \) and \( \text{E}[D(\Theta)|Y] - D[\text{E}(\Theta|Y)] \) readily using the empirical joint posterior distribution of \( \Theta \). Note that there is no need to determine \( f(Y) \), 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 are based on surveys of 1% or 0.1% of the national population instead of nationwide censuses. Figure 2 shows the log central death rates from age 0 to 99 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 are subject to different amounts of volatility.

We used evenly spaced knots between age 0 to 70 and no knot between age 70 to 99 for two reasons. First, too much flexibility beyond age 70 may result in a counterevolutionary outcome that projected mortality rates at older ages are not monotonically increasing with age. Second, placing all knots between age 0 and 70 allows us to better capture the accident hazard.

We used the original definition of DIC (Spiegelhalter et al., 2002), which coincides with the seventh DIC definition in the work of Celeux et al. (2006).
This article is protected by copyright. All rights reserved
genders have similar properties, and therefore for the sake of space we show only those for males in the main text. The estimation results for females are provided in Appendix E.

5.1 Convergence of Parameters

Using the initial values 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, 50,000 samples are recorded and used to form the joint posterior distribution. Further details concerning convergence of parameters are provided in Appendix A.

5.2 Posterior Distributions of Parameters

The fan charts in Figure 3 display the posterior distributions of $\alpha_x$, $\beta_x$ and $\kappa_t$ 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%, ... 90% credible intervals with progressively lighter shadings. The width of a fan chart indicates the level of parameter uncertainty entailed.

![Figure 3](image)

Figure 3: The posterior distributions of $\alpha_x$, $\beta_x$ and $\kappa_t$ for $x = 0, \ldots, 99$ and $t = 1981, \ldots, 2014$.

The pattern of the posterior means of $\alpha_x, \beta_x$ 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 $\alpha_x$ and $\beta_x$ are smooth.

As expected, the posterior means of $\kappa_t$ exhibit a downward trend. The variation in the uncertainty surrounding $\kappa_t$ 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 $\kappa_t$ is particularly wide.

Figure 4 depicts the posterior distributions of $s_0^2$, $s_1^2$ and $s_{0.1}^2$. The posterior means of $s_0^2$, $s_1^2$ and $s_{0.1}^2$ are highly 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 $s_0^2$ (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, Figure 5 shows the posterior distributions of the parameters in the state equation. It is noteworthy that $\mu$ is subject to a great deal of 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 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 proposed adaptations.

6.1 Baseline Results

Predictions of future mortality are obtained from the posterior predictive distribution of $y_{t_0+n_0-1:k}$, $k = 1, 2, \ldots$. Assuming that historical and future log central death rates are independently distributed.
given $\Theta$, the posterior predictive distribution of $Y_{t_0+n_y-1+u}$ is given by

$$f(Y_{t_0+n_y-1+u}|Y) = \int f(Y_{t_0+n_y-1+u}|\Theta) \pi(\Theta|Y) d\Theta,$$

where $f(Y_{t_0+n_y-1+u}|\Theta)$ is the density function of $\text{MVN}(\alpha + \beta \kappa_{t_0+n_y-1+u}, \sigma_t^2 + \kappa_{t_0+n_y-1+u}$). Alternately, if one is 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 one 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 $\varepsilon_{t_0+n_y-1+1}$ is zero no matter which one of the three options is chosen.

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

1. Obtain a realization of $\Theta$ from its empirical posterior distribution.
2. Given the values of $\mu$, $\sigma^2$ and $\kappa_{t_0+n_y-1}$ in the realization of $\Theta$ from the previous step, simulate a realization of $\kappa_{t_0+n_y-1+u}$ from its conditional predictive distribution $\text{MVN}(\kappa_{t_0+n_y-1+u} + u\mu, u\sigma^2)$.
3. Given the realizations of $\kappa_{t_0+n_y-1+u}$ and $\Theta$ from the previous two steps, simulate a realization of $Y_{t_0+n_y-1+u}$ from its conditional predictive distribution $\text{MVN}(\alpha + \beta \kappa_{t_0+n_y-1+u}, \sigma_t^2 + \kappa_{t_0+n_y-1+u}$).

Repeat Steps 1 to 3 to obtain a large number of realizations of $Y_{t_0+n_y-1+u}$, which collectively form an empirical posterior predictive distribution of $Y_{t_0+n_y-1+u}$.

Figure 6 presents the predictions of central death rates at ages 0, 10, 20, ..., and 90 over a horizon of 35 years. Three 90% prediction intervals, which are respectively derived by setting $s^2; t = 2015, 2049$, to $s_{0.1}^2$ (solid lines), $s_{0.01}^2$ (dashed lines) and $s_{0.01}^2$ (dotted lines), are displayed in each panel. The line in the middle of the prediction intervals represents the central prediction. Also depicted in Figure 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 Figure 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 $s^2_{0.1}$) seems to be sufficient for capturing the variation in the circles (historical values obtained from censuses), and the other two prediction intervals (based on $s^2_{0.1}$ 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 presented in the previous subsection include three sources of uncertainty, namely, the stochastic uncertainty in the random walk for $\{\kappa_t\}$, the parameter uncertainty implied by the joint posterior distribution of the parameters and the sampling uncertainty 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.

This article is protected by copyright. All rights reserved.
Figure 6. The 90% prediction intervals for the central death rates at ages 0, 10, 20, ..., and 90 for the 2015-2049 period, based on $s^2$ (solid lines), $s^2_{1.15}$(dashed lines) and $s^2_{0.15}$ (dotted lines). The central prediction is presented by the line in the middle of the prediction intervals. The historical central death rates are shown in circles (for those obtained from censuses) squares (for those obtained from surveys of 1% of the population) and crosses (for those obtained from surveys of 0.1% of the population).

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

1. Calculate $E(Y|X)$, $E(X^2|Y)$, $E(s^2_{0.15}+n_{y}-1|Y)$, $E(s^2|Y)$, $E(e^2|Y)$ and $E(s^2_{0.15}+n_{y}-1+u|Y)$ from the posterior distributions of the model parameters.

2. Simulate a realization of $\kappa_{0.15}+n_{y}-1+u$ for each $u = 1, 2, \ldots$ from its predictive distribution, $N\left[E(\kappa_{0.15}+n_{y}-1|Y) + u\left[E(\mu|Y) + uE(\sigma^2|Y)\right]\right]$. 
Given the result from Step 2, simulate a realization of $\mu_{0+n_u-1+u}$ for each $u = 1, 2, \ldots$ from its predictive distribution,

$$
\text{MVN}(E(\alpha|Y), E(\beta|Y), E(\sigma^2|Y)).
$$

4. Repeat Steps 2 and 3 to obtain an empirical predictive distribution of $\mu_{0+n_u-1+u}$ 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.

Figure 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 entailed in the prediction of future Chinese mortality.

Finally, we remark that the parameter uncertainty associated with $\mu$ plays a particularly important role in long horizon forecasts. This phenomenon can be understood by considering the fact that the posterior predictive variance of $\mu_{0+n_u-1+u}$, given by

$$
\text{Var}(\mu_{0+n_u-1+u}|Y) = \text{Var}(E(\mu_{0+n_u-1+u}|\theta)|Y) + E(\text{Var}(\mu_{0+n_u-1+u}|\theta)|Y)
$$

is related to $u^2$ times the posterior variance of $\mu$.

### 6.3 The Importance of the Proposed Adaptations

The results 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.

Figure 8 compares the posterior distributions of $\beta$ when the two adaptations are used and not used. Without the B-splines function for smoothing $\beta$, the resulting pattern of the posterior means of $\beta$ 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 $\beta$ at ages 9 and 10 are $-0.0003$ and $0.0128$, respectively, which mean the log central death rates at these two ages are expected to evolve in opposite directions, an outcome that makes no demographic sense. The problem can also be observed in Figure 9, in which we compare the projected age patterns of future mortality generated from the full model and the model without the proposed adaptations.

Figure 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 to those derived from the full model, the central predictions generated from the restricted model are not so in line with the corresponding historical values. The problem is most apparent in the panel for age 50, where we observe the central projection and the line joining the circles (historical values obtained from censuses) have rather different slopes. A likely cause of the problem is a misestimation of $\beta$, which arises as the B-splines function is removed.

The prediction intervals presented in Figures 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, the interval for the 1-year-ahead projection is wide enough to encompass all but only one historical values.
7 Validation of the Proposed Modeling/Estimation Approach

7.1 Generating Pseudo Data Sets

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

The exposure counts are taken as those in the actual data set (for Chinese males). This step implies that the pseudo data set is subject to exactly the same missing value problems described in items (1) and (2) in Section 2.
Figure 10: The central and 90% interval predictions of the log central death rates at ages 0, 10, 20, ..., and 90 in the 2015 to 2049 period, derived from the model without the adaptations proposed in Sections 4.3 and 4.4 (solid lines) and from the full model (based on \( a_x^2 \)) (dotted lines). The historical central death rates obtained from censuses are shown in circles.

Step 1/2: The death count at age \( x \) and in year \( t \) is obtained by simulating a realization of a Poisson distribution with a mean of

\[
E_{x,t} \exp \left[ E(\alpha_x | Y) + E(\beta_x | Y) + E(\kappa_t | Y) \right],
\]

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

Step 3/3: 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 \( x \)

This article is protected by copyright. All rights reserved.
and year $t$. When $D_{x,t} = 0$, the corresponding log central death rate is recorded as unreported, thereby resembling the missing data problem described in item (iii) in Section 2.

Restoring our model to such pseudo data sets allows us to examine how our model may perform if the normality assumption does not hold and if heteroscedasticity exists (as the pseudo data generation procedure implies that the variances at different ages and in different calendar years are generally different).

Figure 11 summarizes the structure of one of the generated pseudo data sets. This structure is identical to that of the actual data set for Chinese males (the left panel of Figure 1), 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.

**Figure 11**: A Lexis diagram summarizing the structure of one of the generated pseudo data sets.

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

As implied in Step 2 in the procedure above, the data generating process underlying the pseudo data 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-estimation approach is appropriate, then when applied to the pseudo data sets, it should result in parameter estimates that are sufficiently close to the parameters in the underlying data generating process.

7.2 Validation Results

The solid lines in Figure 13 show the 90% credible intervals of $\alpha_x$, $\beta_x$ and $\kappa_t$ for $x = 0, \ldots, 99$ and $t = 1981, \ldots, 2014$, calculated using the posterior distributions derived from one of the pseudo data sets. They should be compared against the corresponding true parameters (i.e., those in the underlying data generating process), which are shown in dashed lines in the same figure.

For $\alpha_x$ and $\beta_x$, 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 $\alpha_x$ and $\beta_x$, providing

This article is protected by copyright. All rights reserved
strong support for using the proposed methods to estimate the model in the presence of missing data and finite sampling.

For $\kappa_t$, 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 the proposed estimation methods can estimate $\kappa_t$ reasonably accurately in the presence of missing data and finite sampling. Note also that they have comparable widths to those for the $\tau_{x,t}$ derived from the real Chinese mortality data set (see the right panel in Figure 3).

When applied to the rest of the generated pseudo data sets, the estimates of $\alpha_x$ and $\beta_x$ remain almost unchanged. The estimates of $\kappa_t$ vary observably when different data sets are used, but the 90% credible intervals consistently capture most of the true parameter values (Figure 14).

We acknowledge that this analysis may not be comprehensive, as the Poisson pseudo data sets do not capture, for example, the possibility that the variance of a random death count is greater than the mean. In Appendix B, we present two additional analyses that are based on pseudo data sets generated from (log-}
7.3 A Cautionary Note

As pointed out by Li and Chan (2005) and Zhou and Li (2013), mortality forecasts produced by the original Lee-Carter model are highly sensitive to the values of $\kappa_t$ in the beginning and ending years of the data sample (i.e., $\kappa_{t_0}$ 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, let us 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

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

The above 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) $\mu$, which determines the gradient of the expected trajectory. Further discussion of this issue is deferred to a forthcoming article.
more, as shown in expression (10), the mean of the conditional posterior distribution of \( \mu \) is

\[
\frac{K_{t0+n_y-1} - K_{t0}}{n_y - 1},
\]

which is determined exclusively by \( K_{t0+n_y-1} \) and \( K_{t0} \).

As discussed in the previous subsection, the estimates of \( \kappa_t \) vary across the pseudo data sets. The colored jagged lines in Figure 15 represent the posterior means of \( \kappa_t \) for \( t = 1981, \ldots, 2014 \), obtained from three different pseudo data sets. The variation in the posterior means of \( \kappa_t \) 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. In the final year (2014) of the pseudo data 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 pseudo data sets. This phenomenon can be observed in the colored dashed lines in Figure 15, which represent the expected trajectories of \( \kappa_{2015}, \kappa_{2016}, \ldots \) obtained from three of the generated pseudo samples. These expected trajectories have noticeably different levels and gradients. Some of them are quite different from the expected trajectory implied by the true parameters, which is represented by the black dashed line in the same figure.

Figure 15: The posterior means of \( \kappa_{1981}, \ldots, \kappa_{2014} \) (colored solid lines), the expected trajectories of \( \kappa_{2015}, \kappa_{2016}, \ldots \) (colored dashed lines), and the line that passes through \( E[\kappa_{2014}|Y] \) and has a slope of \( E[\mu|Y] \) (colored dotted lines) obtained from three of the generated pseudo data sets (without any adjustment). The black lines represent the corresponding values implied by the true parameters.

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

\[
E[\kappa_{2014}|Y] + uE[\mu|Y]
\]

to

\[
E[\kappa_{2010}|Y] + (u + 4)E[\mu|Y]
\]
and modify the conditional posterior mean of $\mu$ from

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

to

$$\frac{K_{2010} - K_{1981}}{29},$$

(19)

while all of 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 $K_{2015}$, $K_{2016}$, \ldots. The benefit of the proposed adjustments is demonstrated in Figure 16, from which we observe that with the proposed adjustments the resulting expected trajectories of $K_{2015}$, $K_{2016}$, \ldots exhibit less variation across the pseudo data sets and are all close to the expected trajectory implied by the true parameters.

\textbf{Figure 16: The posterior means of $K_{1981}$, ..., $K_{2014}$ (colored solid lines), the estimated trajectories of $K_{2015}$, $K_{2016}$, ..., (colored dashed lines) and the line that passes through $E[K_{2014}|Y]$ and has a slope of $E[\mu|Y]$ (colored dotted lines), obtained from three of the generated pseudo data sets (with the proposed adjustments). The black lines represent the corresponding values 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 associated with the data set being accounted for. Significant effort has been made to ensure that the end result satisfies the three important criteria 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 tailored the data set an estimation procedure, synthesizing Gibbs sampling, an abridged multiple imputation algorithm and a sequential Kalman filter. The estimation procedure is able to 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 into 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

This article is protected by copyright. All rights reserved.
parameters (including $\kappa_t$ 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 $\text{Var}(\varepsilon_{x,t})$, allowing it to vary according to how the data for year $t$ is 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 to other stochastic mortality models that are applicable to the full age range of 0 to 99. 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 behavior that may arise from a jagged pattern of the parameter estimates.

Unlike two-stage estimation approaches whereby Box-Jenkins method (Box et al., 2015) may be used to identify the optimal time-series process for the estimates of $\kappa_t$ 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 $\text{Var}(\varepsilon_{x,t})$ is fixed along the age dimension. However, allowing $\text{Var}(\varepsilon_{x,t})$ to age-specific would significantly increase the number of parameters that have to be estimated.

We do not consider cohort effects due to the limited availability of data. For the data set in question, we can only observe at most about one-third of a birth cohort, and the limited observation is not even continuous because of the missing data. Any cohort-specific characteristics identified from such a data set is likely to be spurious. When more years of data become available, it is warranted to extend our modeling 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, 2000; 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 somewhat similar to the Lee-Carter model 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. While 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 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 $s_{t'}^2$.

We have validated our modeling/estimation approach using pseudo data sets generated from Poisson and and student t distributions. The estimation results derived from the pseudo data sets are not much different from those obtained using the original data set, suggesting 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: (1) approximate the model with a linear Gaussian or mixture Gaussian alternative, (2) use the sequential Kalman filter to draw samples of the state variables from the alternative model, and (3) use the Metropolis-Hastings algorithm to accept or reject the drawn samples. If the departure from the normality assumption is significant, then the method of particle Markov Chain Monte Carlo 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 modeling by Fung et al. (2017). One drawback of this alternative method is that it is very time-consuming to implement.

Another limitation of our modeling approach is that it does not consider the possibility of structural changes. Although there exist 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 U-shaped age patterns of mortality decline, which has attracted a lot of 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 the recent paper by Cairns et al. (2016).

References


This article is protected by copyright. All rights reserved


This article is protected by copyright. All rights reserved


<table>
<thead>
<tr>
<th>Year (t)</th>
<th>Value</th>
<th>ln m_{0,t}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2040</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year (t)</th>
<th>Value</th>
<th>ln m_{10,t}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2040</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year (t)</th>
<th>Value</th>
<th>ln m_{20,t}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2040</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year (t)</th>
<th>Value</th>
<th>ln m_{30,t}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2040</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year (t)</th>
<th>Value</th>
<th>ln m_{40,t}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2040</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year (t)</th>
<th>Value</th>
<th>ln m_{50,t}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2040</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year (t)</th>
<th>Value</th>
<th>ln m_{60,t}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2040</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year (t)</th>
<th>Value</th>
<th>ln m_{70,t}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2040</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year (t)</th>
<th>Value</th>
<th>ln m_{80,t}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2040</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year (t)</th>
<th>Value</th>
<th>ln m_{90,t}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2040</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Author/s:
Li, JS-H; Zhou, KQ; Zhu, X; Chan, W-S; Chan, FW-H

Title:
A Bayesian approach to developing a stochastic mortality model for China

Date:
2019-10-01

Citation:

Persistent Link:
http://hdl.handle.net/11343/285820