Stochastic modelling and projection of mortality improvements using a hybrid parametric/semi-parametric age–period–cohort model

Erengul Dodd, Jonathan J. Forster, Jakub Bijak & Peter W. F. Smith


To link to this article: https://doi.org/10.1080/03461238.2020.1815238
Stochastic modelling and projection of mortality improvements using a hybrid parametric/semi-parametric age–period–cohort model

Erengul Dodda, Jonathan J. Forstera, Jakub Bijaka, and Peter W. F. Smith

Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK; Mathematical Sciences, University of Southampton, Southampton, UK; ESRC Centre for Population Change, University of Southampton, Southampton, UK; Department of Statistics, University of Warwick, Coventry, UK; Department of Social Statistics and Demography, University of Southampton, Southampton, UK

ABSTRACT

We propose a comprehensive and coherent approach for mortality projection using a maximum-likelihood method which benefits from full use of the substantial data available on mortality rates, their improvement rates, and the associated variability. Under this approach, we fit a negative binomial distribution to overcome one of the several limitations of existing approaches such as insufficiently robust mortality projections as a result of employing a model (e.g. Poisson) which provides a poor fit to the data. We also impose smoothness in parameter series which vary over age, cohort, and time in an integrated way. Generalised Additive Models (GAMs), being a flexible class of semi-parametric statistical models, allow us to differentially smooth components, such as cohorts, more heavily in areas of sparse data for the component concerned. While GAMs can provide a reasonable fit for the ages where there is adequate data, estimation and extrapolation of mortality rates using a GAM at higher ages is problematic due to high variation in crude rates. At these ages, parametric models can give a more robust fit, enabling a borrowing of strength across age groups. Our projection methodology assumes a smooth transition between a GAM at lower ages and a fully parametric model at higher ages.

1. Introduction

Recent mortality improvements in most countries have led to higher life expectancies. Since this has significant social policy implications, for example, in areas such as pensions and health care, modelling and projecting mortality rates becomes imperative. For example, in the insurance industry, the risk of higher than expected annuity payments, or the so-called longevity risk needs to be quantified for solvency requirements. The mortality projections are not only important for pensions or health care but also in allocating resources and government planning, for example, for housing, education and labour market. This requires probabilistic models for mortality rates. In the past decade, a vast literature on probabilistic mortality models has been developed. However, very few of them are suitable for the entire age range. In this paper, instead of focusing only on older ages, we present a mortality projection methodology for the entire population.

CONTACT

Erengul Dodd, e.dodd@soton.ac.uk
Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, UK

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Probabilistic mortality projection models can be broadly summarised under four categories: generalised bilinear, generalised linear, semi-parametric and generalised additive models. The original Lee–Carter model (Lee & Carter 1992) is the pioneering method in this area. This model is an example of a bilinear model and has two factors, i.e. age and period, to model and forecast mortality rates. The other bilinear models include extensions of the Lee–Carter model, for example, three-factor models that encompass the cohort effect or Poisson error structure instead of an implied normality assumption (see, e.g. Brouhns et al. 2002, Renshaw & Haberman 2006). Whilst these models can provide a satisfactory fit, they also have some undesirable features. In particular, their parameter estimates can be sensitive to the range of years used for fitting, and they are challenging to estimate efficiently.

Alternatively, linear (rather than bilinear) models with age and period as factors were investigated by Renshaw & Haberman (2003). Different linear structures were developed and compared by Cairns et al. (2009, 2011a). Many studies show that coherent mortality forecasts can be obtained for different populations using these models (e.g. Cairns et al. 2011b, Li & Hardy 2011, Börger & Aleksic 2014). Currie et al. (2004) proposed modelling mortality as a smooth function in two dimensions (age and time) using P-spline methodology, although such an approach can be difficult to incorporate into a projection since then the projection ignores the majority of the observed historical mortality experience and is too sensitive to mortality in the base year, even after smoothing (Li et al. 2010). Besides, this approach does not allow for coherent projections for different populations (Börger & Aleksic 2014). Most of the mortality studies in the actuarial literature concentrate on the retirement ages (between early 60s and late 90s) due to the direct financial impact on pensions. Recently, Li & Liu (2019) and Richards (2020) considered mortality models for older ages. Approaches which fully account for uncertainty include Cairns et al. (2006, 2011b), Bennett et al. (2015) and Hilton et al. (2019).

In particular, Hilton et al. (2019) provide a Bayesian approach to producing mortality projections based on the use of generalised additive models (GAMs) for the majority of the age range, but with an inclusion of a parametric model at older ages where the data are sparse. Their approach allows for smooth functions of age and cohort effects, and provides estimates of mortality at young ages as well as extreme ages. Bayesian models incorporate multiple sources of uncertainty and expert opinion in a natural way. However, implementing a full Bayesian approach might be expensive, especially if computing marginal likelihoods requires high-dimensional integrals and posterior distributions are analytically intractable.

In this paper, we present the maximum likelihood approach to the methodology presented in Hilton et al. (2019). Additionally, we incorporate expert opinion in our projections. Moderating future mortality assumptions is especially important when the projection period is long. We model the mortality improvements instead of mortality rates allowing for overdispersion. This is because we believe mortality improvements can be modelled by a stable process which is required to be projected forward based on past experience. Projection of mortality improvement rates is advocated by Plat (2011), Haberman & Renshaw (2012, 2013) and more recently by Börger & Aleksic (2014). In the United Kingdom (UK), the Continuous Mortality Investigation (CMI) introduced age–period–cohort improvement (APCI) model as a new mortality projection method (CMI 2016). However, the CMI does not use the APCI model as a stochastic model to project future mortality rates. The CMI uses the APCI model to simply obtain the initial mortality improvements (separated by age-, period- and cohort-related improvements) for projections. Richards et al. (2019) implemented the APCI model as a fully stochastic model. They compared this model to the age–period, age–period–cohort and Lee–Carter models and found that the APCI model fits the data better than these other models considered in their paper.

In CMI (2016), the CMI recognises that the Office for National Statistics (ONS) dataset shows considerable overdispersion, relative to a Poisson error distribution. In the presence of overdispersion, modelling the observed number of deaths under a single parameter distribution such as a Poisson distribution (where the variance is restricted to be equal to the mean) will lead to underestimation of uncertainty. To allow for overdispersion, we use a more flexible negative binomial distribution in
modelling. We also impose smoothness in parameter series which vary over age, cohort, and time in an integrated way. GAMs, being a flexible class of semi-parametric statistical models, allow us to differentially smooth components, such as cohorts, more heavily in areas of sparse data for the component concerned.

While GAMs can provide a reasonable fit for the ages where there is adequate data, estimation and extrapolation of mortality rates using a GAM at higher ages is problematic due to high variation in crude rates. At these ages, parametric models can give a more robust fit, enabling a borrowing of strength across age groups. Our projection methodology is based on a smooth transition between a GAM at lower ages and a fully parametric model at higher ages. We model infant rates separately and propose a new method to model and predict them. Since spline-based methods are used widely in the literature (especially for the entire age range), as discussed above, we compare our results to the two-dimensional P-splines approach proposed by Currie et al. (2004) where relevant.

The rest of the paper is organised as follows: in Section 2, we introduce the data and our methodology for modelling the mortality improvements, and how a smooth transition from the smoothing spline to an old-age model is attained. In Section 3, we present our estimates for mortality rates and investigate the robustness of the proposed methodology. In Section 4, we explain how the mortality projections and the uncertainty around these projections are obtained. In Section 5, we incorporate the expert opinion and provide comparisons with the UK national population projections. Our conclusions are in Section 6.

2. The data and the model

We use UK population data between 1961 and 2013 obtained from the Human Mortality Database (2019). The data include the mid-year exposures and number of deaths for each year of age for males and females.

Here we propose a model that contains terms which specifically account for the variation of mortality differences over time and between different ages and cohorts. Let \( m_{xt} \) denote the central mortality rate at age \( x \) in year \( t \), then we consider as the initial model specification

\[
\log \frac{m_{xt}}{m_{x(t-1)}} = \alpha_x + \kappa_t^* + \gamma_{t-x}^*, \tag{1}
\]

where \( \alpha_x \) can be interpreted as a baseline annual mortality improvement at age \( x \), \( \kappa_t^* \) as the level of mortality improvement in year \( t \) and \( \gamma_{t-x}^* \) represents cohort differences in mortality improvement since cohorts are indexed by year of birth \((t-x)\).

Model (1) is an age–period–cohort model for log-mortality differences (mortality logratios). Here we represent mortality improvements as logratios, rather than as relative differences, where model (1) would be expressed as

\[
\frac{m_{xt} - m_{x(t-1)}}{m_{x(t-1)}} = \alpha_x + \kappa_t^* + \gamma_{t-x}^*. \tag{2}
\]

For all but large mortality rates, differences between \( \log(m_{xt}/m_{x(t-1)}) \) and \( (m_{xt} - m_{x(t-1)})/m_{x(t-1)} \) are negligible. This model is similar in structure to models proposed by Renshaw & Haberman (2003), the non-spatial component of Bennett et al. (2015) and the APCl model of CMI (2016).

Note that in terms of mortality rates, model (1) can be expressed as

\[
\log m_{xt} = \mu_x + \alpha_x t + \kappa_t + \gamma_{t-x}, \tag{2}
\]

where there is a straightforward correspondence between the \( \kappa_t \) and \( \gamma_{t-x} \) parameters of models (1) and (2). More specifically, the cohort and period terms in (2) are the accumulated versions of their equivalents in (1) and \( \mu_x \) is the log-mortality rate at age \( x \) in year \( t = 0 \). Due to the linear relationship
between age, period and cohort components of the model, constraints are required in order to identify these effects. The identifiability constraints we use for model (1) are

$$\sum \kappa_t^* = 0, \quad (3)$$
$$\sum \gamma_{t-x}^* = 0, \quad (4)$$
$$\sum (t-x)\gamma_{t-x}^* = 0. \quad (5)$$

That is, the period effect is constrained to sum to zero. Similarly, the cohort effect is constrained so that the sum of effects is zero and displays zero growth over the whole range of cohorts.

In Section 3, we present parameter estimates for model (1), together with $\mu_x$. Note that this is equivalent to presenting the parameter estimates for model (2) on a mortality improvement scale (that is, with differenced cohort and period effects, $\kappa_t^*$ and $\gamma_{t-x}^*$, from model (1) rather than their summed equivalents). This model, being simply a generalised linear model, is easy to fit. Furthermore, its parameter estimates seem to be robust to the time window used to fit the models. Börger & Aleksic (2014) advocate the use of this model for projecting mortality, and we also find that it has the required properties of adequately and robustly fitting the observed data.

In the literature, it is very common to estimate the model parameters based on the Poisson log-likelihood. However, under the Poisson model, the variance is restricted to equal the mean, an assumption which is implausible for a large inhomogeneous population. A more flexible model would be a negative binomial model; in this paper, we assume

$$d_{xt} \sim \text{NegBinomial}(E_{xt}m_{xt}, a),$$

and therefore, the log-likelihood is

$$l(\theta, a) = \sum_{x,t} a \log \left( \frac{a}{E_{xt}m_{xt}(\theta) + a} \right) + \sum_{x,t} d_{xt} \log \left( \frac{E_{xt}m_{xt}(\theta)}{E_{xt}m_{xt}(\theta) + a} \right) - \log \Gamma(a + d_{xt}) - n \log \Gamma(a).$$

Here $d_{xt}$ is the observed death count, $E_{xt}$ is the central exposure to risk at age $x$ in year $t$, $a$ is the dispersion parameter such that the variance is $E_{xt}m_{xt}(\theta) + (E_{xt}m_{xt}(\theta))^2/a$, $\theta$ represents the model parameters $(\alpha_x, \kappa_t, \gamma_{t-x})$ and $n$ is the number of positive values of $E_{xt}$.

One disadvantage of model (1) is that the maximum-likelihood estimates of some of the model parameters do not vary smoothly (Figure 1). However, this can be easily overcome by adopting an estimation method which penalises roughness in the series of estimates for model (1) (e.g. penalised likelihood or Bayesian). One possible way of obtaining smoother estimates is to modify (2) to yield a generalised additive model of the form:

$$m_{xt} = \exp \left( s_{\mu}(x) + s_{\alpha}(x)t + \kappa_t + s_{\gamma}(t-x) \right). \quad (6)$$

In (6), $s_{\mu}$, $s_{\alpha}$ and $s_{\gamma}$ denote arbitrary smooth functions, which can be estimated by balancing goodness-of-fit to the observed data with smoothness of the corresponding function (Wood 2006). This can be fitted by using standard gam packages in R (e.g. using the gam function in mgcv package). The mgcv package allows us the choice of splines and the family of distributions. Here we use univariate penalised cubic regression splines and the negative binomial family. In (6), we use a non-smooth function of the time effect since we do not necessarily expect the mortality improvements to vary smoothly over time. Indeed, the data suggest that the year-specific contributions to mortality improvement are not correlated year by year and in the model fit without smoothness the time effect appears to be a white noise process (see Figure 1). However, at the mortality rate level, the period
effect would be a random walk. Thus, although we are not using smooth time effect on the mortality improvement level, there will be some (weak) smoothness on the mortality rate scale.

For the highest ages $x$, for which observed mortality experience is sparse, we recommend that the baseline mortality $\mu_x$ and the age-specific mortality differences $\alpha_x$ are estimated by using parametric models, for example, log-linear model or logistic model, with parameters estimated from the mortality data for the older ages. The resulting log-linear model, with $\mu_x = \mu + \mu_xx$ and $\alpha_x = \alpha + \alpha_xx$, has the form

$$m_{xt} = \exp(\mu + \mu_xx + (\alpha + \alpha_xx)t) \exp(\kappa_t + s_y(t - x)), \quad \text{for } x \geq x_0,$$

and the logistic model has the form

$$m_{xt} = \frac{\beta \exp(\mu + \mu_xx + (\alpha + \alpha_xx)t) \exp(\kappa_t + s_y(t - x))}{1 + \exp(\mu + \mu_xx + (\alpha + \alpha_xx)t)}, \quad \text{for } x \geq x_0,$$  

where $x_0$ is an optimal age to make the transition from smooth to linear model, and $\kappa_t$ and $s_y(t - x)$ are the estimates obtained from fitting (6) to the main body of data ($0 < x < x_0$). In (7) and (8), the sum of $\kappa_t$ and $s_y(t - x)$ can be considered as a non-standard ‘offset’. Note that in (8) these terms only appear in the numerator. This is to ensure the same interpretation of these parameters in both (6) and (8) and since they were log-linear parameters in (6), they should be log-linear parameters (and not logistic parameters) in both parts of the model. We can fit (8) using a general optimisation function (e.g. optim or nlm) in R. Here, we use the nlm function and assume the number of deaths follows a negative binomial distribution as before.
The log-linear model has, therefore, the following estimates of the baseline mortality
\[
\mu_x = \begin{cases} 
    s_\mu(x) & x < x_0 \\
    \mu + \mu_x x & x \geq x_0
\end{cases}
\]
and mortality improvement
\[
\alpha_x = \begin{cases} 
    s_\alpha(x) & x < x_0 \\
    \alpha + \alpha_x x & x \geq x_0
\end{cases}
\]
for both males and females. For the logistic model these estimates are, respectively,
\[
\mu_x = \begin{cases} 
    s_\mu(x) & x < x_0 \\
    \log \left( \frac{\beta \exp (\mu + \mu_x x)}{1 + \exp (\mu + \mu_x x)} \right) & x \geq x_0
\end{cases}
\]
and
\[
\alpha_x = \begin{cases} 
    s_\alpha(x) & x < x_0 \\
    \log \left( \frac{\beta \exp (\mu + \mu_x x + \alpha + \alpha_x x)}{1 + \exp (\mu + \mu_x x + \alpha + \alpha_x x)} \right) - \log \left( \frac{\beta \exp (\mu + \mu_x x)}{1 + \exp (\mu + \mu_x x)} \right) & x \geq x_0
\end{cases}
\]
also for both males and females.

We treat infant (age 0) mortality separately. Here, we exclude the period effect \( \kappa \) and fit the model
\[
m_{0t} = \exp (\mu_0 + \alpha_0 t) \exp \left( s_\gamma (t) \right), \tag{9}
\]
where \( s_\gamma (t) \) is the estimate of the cohort effect for \( x = 0 \) obtained from fitting (6) to the main body of data \((0 < x < x_0)\). In fitting this negative binomial generalised linear model, we use \( \text{glm.nb} \) function in the \textsc{mass} package in R. We have investigated the dependence of infant mortality rates on both the time and the cohort effects, estimated from the rest of the data. It transpired that infant mortality has a unique pattern of period variation, and therefore, the time effect was a very weak predictor for the infant rates. On the other hand, the infant mortality exhibits a strong dependence on the cohort effect estimated from the rest of the data (see Figure 9).

Note that in this paper we only present the results using the logistic model for older ages. Therefore, our proposed model combines three components, (9), (6) and (8), corresponding to infants \((x = 0)\), a majority of ages \((0 < x < x_0)\) and the oldest ages \((x \geq x_0)\), respectively.

### 3. Estimation of mortality rates

Dodd et al. (2018) suggest that for England and Wales mortality data, an optimal age \((x_0)\) at which to make the transition from smooth to logistic model is 93 for males and 91 for females, based on 2010–2012 mortality data. We assume that these thresholds are fixed over time. This is a strong assumption and the transition age \(x_0\) at each year might be included in the model as an unknown parameter. However, the added complexity required for different threshold ages may not be justified since our preliminary investigation shows a negligible effect on mortality projections.

Figure 1 presents the maximum-likelihood estimates of the parameters of model (1) under the Poisson distribution and negative binomial distribution for males aged between 1 and 92 years, using data for the period 1961–2013.

We compare the fit of model (1) to the observed data with the fit of the two-dimensional P-spline methodology proposed by Currie et al. (2004), which we will simply call the P-spline method from now on. With regard to an assessment of model fit, Figure 2 presents the square of Pearson residuals from the P-spline method, and from model (1) assuming the Poisson distribution.
Figure 2. Comparison of residuals, data for males, 1961–2013: the P-spline approach (left panel) and model (1) assuming the Poisson distribution (right panel).

Figure 3. Comparison of residuals, data for males, 1961–2013: the P-spline approach allowing for overdispersion (left panel) and model (1) assuming the negative binomial distribution (right panel).
It can be observed from Figure 2 that model (1) fits the data at least as well as the P-spline method. Indeed, by conventional goodness-of-fit measures (residual deviance), model (1) fits significantly better than the P-spline model, even allowing for its increased complexity in terms of the number of degrees of freedom required for parameter estimation. Model (1) seems to do a better job of estimating mortality in the age range 15–20 (at the start of the ‘accident hump’ related to external causes of death during early adulthood, especially for men). Both models have difficulty fitting the 1919 cohort (see Cairns et al. 2016), but arguably this cohort is of limited significance for population projection. Both models, however, fail to fit when assessed by conventional goodness-of-fit measures such as Pearson’s chi-squared statistic. Evidence for this is the large number of Pearson residuals with a squared value greater than 9. On the other hand, estimates which allow for overdispersion, either model (1), fitted by maximising a negative binomial likelihood, or a P-spline fitted by quasi-likelihood produce residuals within a much more acceptable range (see Figure 3). Therefore, we use the negative binomial model to estimate the model parameters in our analyses.

One advantage of the P-spline approach is that it provides estimates of mortality rates that vary smoothly over age and time, as illustrated in Figure 4, which accounts for overdispersion through maximum Poisson quasi-likelihood estimation. Significant cohort effects, and cohorts with large annual mortality improvements can also be seen in this figure. Under this P-spline method, the projections will only depend on the most recent years, and they would be largely insensitive to historical data.

For model (1), the maximum-likelihood estimates of some of the model parameters, illustrated in Figure 1, do not vary smoothly, and as a consequence, the estimated mortality rates, presented in Figure 5, are also more irregular than would be desirable.

Figure 4. Heat map of the fitted mortality improvements for the P-spline model allowing for overdispersion through a penalised quasi-likelihood method, data for males, 1961–2013.
To obtain smoother estimates, we use model (6). Figure 6 displays the estimates for the resulting smooth model (6), superimposed over the corresponding estimates for model (2) on a mortality improvement scale. Note that it is the differenced cohort and period effects ($\kappa^*_t$ and $\gamma^*_t-x$ from model (1)) that are plotted rather than their summed equivalents. Not surprisingly, the estimates for model (6) are much more regular and have the desired smoothness, and the fitted mortality rates, displayed in Figure 7, are also smoother. There is an increase in residual deviance, but this is compensated by a corresponding decrease in the effective complexity of the model. Note also that the vertical strips correspond to the year effect, which we do not smooth.

Figure 8 presents the combined estimates of the parameters of models (9), (6) and (8). Under the logistic model (8) mortality rates flatten off, converging to a limiting rate $\beta$ as $x$ tends to infinity. We estimate $\beta$ as 1.48 for males and 0.99 for females – the values which we set as constant over time.

The observed and fitted infant mortality using model (9) are displayed in Figure 9. If the cohort effect is not considered in the model, the estimates of infant mortality rates would be a straight line (on the log scale). By borrowing information on cohort effect from the rest of the data by age, we can identify the cohort improvements in infant mortality rates, e.g. around 1990.

Finally, we investigate the robustness of the proposed methodology by exploring the sensitivity of the estimated mortality rates in a later year (2011) to changes in the data used to estimate the model. Two different approaches are taken. In the first, we compare the estimates of 2011 mortality rates and 2011–2012 mortality improvements for model (6) fitted for ages $1 \leq x < x_0$ by using data from 1961 to 2013, with the equivalent estimates fitted for 1971–2013 and 1981–2013 (see Figure 10). Then, we compare the estimates of 2011 mortality rates and 2011–2012 mortality improvements obtained by using data from 1961 to 2011 with the equivalent estimates fitted to the 1961–2012 and 1961–2013 series (see Figure 11). In Figures 10–12, we also provide a comparison with the P-spline approach.
Both in Figures 10 and 11, there is a big difference between our proposed model and the P-spline method in terms of the fit for 2011. Under the P-spline method, there is a quite dramatic mortality improvement at around age 20, which we are not picking up in our model. This is because under the P-spline model the big mortality improvement at late teens in current years are projected forward as an age-specific improvement, without taking the historical data into account (see Figure 4). Since our model takes the whole period into account instead of only the current years, we do not see such a dramatic improvement in mortality around age 20. This is the largest difference between the future mortality projections produced by the two models.

In the proposed projection methodology, existing cohort components are included in projections, but the period effects (and any as yet unobserved cohorts) are projected as zero, which is consistent with the model. We present the mortality improvements for 2011–2012 in Figures 10 and 11, excluding the estimated period effect for 2012 for comparison with the P-spline approach. Our actual proposal would be to project forward from the final year of observed data (2013) in which case the base mortality rates and mortality improvements are presented in Figure 12.

Note that the scale of Figure 12 is different than the scale of Figure 10, and therefore, the difference between different historical fitting periods is more obvious. The 1961–2013 or 1971–2013 datasets broadly show similar patterns. However, if we ignore the data from 1961 to 1980, we move to a regime where there are much bigger mortality improvements at age around 20. As a result, not surprisingly, when we lose almost 40% of the data we see some sensitivity in mortality improvements.
4. Projection of mortality rates

Based on the parameter estimates of a model such as (6), providing point projections over any future time horizon is straightforward. Such a projection only requires extrapolation of the time effects $\kappa_t$ for future years $t$, and the cohort effects $\gamma_{t-x}$ for future birth cohorts. The identifiability constraint we imposed on the period effect in (1) implies that the accumulated period effect in (2) is constrained to zero in the final observed year, and therefore, the estimated $\kappa_t$ series approximates a random walk with zero-mean increments (see Figure 6), then it is reasonable to forecast the period effect to be zero for future $t$. Uncertainty about these forecasts is incorporated by assuming normally distributed increments with variance, $\sigma^2_{\kappa}$, estimated from the $\kappa_t$ series. A similar argument suggests that it is reasonable to also set cohort effects for future cohorts to zero. We do not include uncertainty about future cohorts, as these cohorts are likely to have a negligible effect on population mortality over the forecast horizon.

The confidence intervals can be calculated using the standard deviation obtained from the covariance matrix of the estimated model parameters and the variance of the observed period effect. More precisely for main model (6), we have

\[
\text{Var}(\log \hat{m}_{xt}) = \text{Var}(s_{\mu}(x) + s_{\theta}(x)t + s_{\gamma}(t-x)) + \sigma^2_{\kappa},
\]

(10)

where the first term on the right-hand side is the variance of a linear function of GAM parameters which can be computed using standard output from GAM fitting in R. For the old-age model, applying the delta method (see, e.g. Schervish 1995), we have

\[
\text{Var}(\log \hat{m}_{xt}) = \Delta^T \Sigma \Delta + \sigma^2_{\kappa},
\]

(11)
where $V$ is the covariance matrix of the model parameters $(\beta, \mu, \mu_X, \alpha, \alpha_X)$,

$$
\Delta_{xt} = 
\begin{pmatrix}
1 \\
\delta_{xt} \\
x \delta_{xt} \\
t \delta_{xt} \\
x t \delta_{xt}
\end{pmatrix}
$$

and

$$
\delta_{xt} = 1 - \frac{\exp(\mu + \mu_X x + (\alpha + \alpha_X)t)}{1 + \exp(\mu + \mu_X x + (\alpha + \alpha_X)t)}.
$$

We present our 2025 and 2055 projections for males and females in Figure 13.

In Figures 14 and 15, we compare the projections for 2025 and 2055 to the respective values from the 2014-based national population projections of the ONS, which use the past and projected data from the period and cohort life tables, in a range of variants: principal, high, and low (ONS 2015a, 2015b). The projection methodology of the ONS is based on a P-spline model, and the technical details can be found in ONS (2016).

We understand that the discontinuity of the ONS rates after age around 110 is due to merging the estimated $q_x$ for different constituent countries of the UK, where $q_x$ denotes the probability of dying by age $x + 1$ given that an individual attains age $x$. When calculating $q_{xt}$ we use the following approximation:

$$
q_{xt} \approx 1 - \exp(-m_{xt}).
$$
Figure 9. Estimates of infant mortality rates using model (9) (solid lines), compared with observed rates (+), data for 1961–2013. (Left panel) males, (right panel) females.

Figure 10. Estimated 2011 mortality rates (left panel), and 2011–2012 mortality improvements (right panel) for males, using model (6) and different historical fitting periods.

There is some discrepancy between ONS projections and our projections, especially around the accident hump. As mentioned before this is because of the current high improvement rates that are projected forward under the P-spline method. For both males and females, from age around 60 onwards the projections are consistent with each other.
Figure 11. Estimated 2011 mortality rates (left panel), and 2011–2012 mortality improvements (right panel) for males, using model (6) and different recent fitting periods.

Figure 12. Estimated 2013 mortality rates (left panel), and 2013–2014 mortality improvements (right panel) for males, using model (6) and different historical fitting periods.

Figure 13. Mortality projections for 2025 (left panel) and 2055 (right panel) for males and females.
Figure 14. Comparison of 2025 mortality projections for males (left panel) and females (right panel) with ONS projections.

Figure 15. Comparison of 2055 mortality projections for males with ONS projections.

5. Incorporating expert opinion

The ONS projections are moderated by experts, whereas our proposed model so far is completely data driven. We believe that the use of expert knowledge is a useful tool for moderating predictions provided by the model. A slightly modified version of the proposed approach can allow us to make full use of all available sources of information, including expert opinion.

The ONS projections assume a convergence in annual mortality improvement to a constant value over a fixed time horizon (currently 25 years) across all ages, for every cohort born after 1960. For older cohorts, the convergence is imposed on the series of cohort mortality improvements. Our modification remains in the spirit of the approach proposed earlier, with the age-specific mortality improvements ($\alpha_x$) converging to a common, expert-specified value and the cohort effects converging...
to zero over a 25-year time horizon. We incorporate the convergence using the same weight function:

$$w_h = \begin{cases} 
1 - 3 \left( \frac{h}{25} \right)^2 + 2 \left( \frac{h}{25} \right)^3 & \text{for } 0 \leq h \leq 25, \\
0 & \text{for } h \geq 25,
\end{cases}$$

(12)

where $h$ represents the projection period (i.e. $h = 0$ corresponds to the last year of observed data). Therefore, for example, the age-specific mortality improvements incorporating the experts in period $h$, $\hat{\alpha}_{x,h}^e$, becomes

$$\hat{\alpha}_{x,h}^e = \begin{cases} 
\alpha^e (1 - w_h) + \hat{\alpha}_x w_h & \text{for } 0 \leq h \leq 25, \\
\alpha^e & \text{for } h \geq 25.
\end{cases}$$

(13)

Here $\hat{\alpha}_x$ is the estimated year-on-year improvement by age effect using (6), (8) and (9). In line with the ONS mortality assumptions, we assume the value of the target expert-based mortality improvement is 1.2%, independent of age and sex, i.e. $\alpha^e = -1.2\%$.

Where expert opinion is incorporated in the forecasts, it is also incorporated in the uncertainty with the expert uncertainty and parameter uncertainty being weighted correspondingly. The additional variance for the expert opinion, i.e. $(\sigma_e \sum_{h=1}^{H} (1 - w_h))^2$ where $H$ is the number of projection years and $\sigma_e$ is the standard deviation of the expert-based mortality improvement, can simply be added to the right-hand side of Equation (10). In this paper, we assume $\sigma_e = 0.6\%$.

Up to this point, males and females were modelled separately. However, when projecting the mortality rates, one final adjustment is made to fix any divergence between male and female rates at very old ages (see Figure 13). This is because we do not believe that the male and female mortality rates will start to diverge at very high ages following steady convergence up to this point. This also applies where the divergence occurs after the rate functions cross. Therefore, starting from the age where the difference between male and female mortality rates starts to increase (if that occurs), we keep the difference between the male and female mortality rates at a constant value and obtain weighted mortality rates using

$$m_{xt}^* = m_{xt}^m w_{xh}^l + (m_{xt}^f + d_t) (1 - w_{xh}^l)$$

and

$$m_{xt}^f = (m_{xt}^m - d_t) w_{xh}^l + m_{xt}^f (1 - w_{xh}^l)$$

Figure 16. Mortality projections for 2025 (left panel) and 2055 (right panel) for males and females.
where

\[ w_{xh}^f = \frac{m_{xh}^m}{m_{xh}^m + m_{xh}^f}. \]

Here \( m_{xh}^m \) and \( m_{xh}^f \) denote the male and female mortality rates before the weighting is applied, respectively, and the ‘\( * \)’ in the superscript refers to the adjusted mortality rates. We define \( d_t \) as the smallest positive difference between male and female mortality rates in year \( t \) over the age range \( x \) (considering only the ages above 50 in our application). Finally, \( l_{xh} \) represents the expected number of survivors to exact age \( x \) in year \( h \) from a birth population of size \( l_0 = 100,000 \) and the weights are used as a proxy for survivorship probabilities at age \( x \) and any future year for males and females.

The comparison of 2025 and 2055 projections after these adjustments can be seen in Figure 16.

**Figure 17.** Comparison of 2025 mortality projections for males (left panel) and females (right panel) with ONS projections.

**Figure 18.** Comparison of 2055 mortality projections for males with ONS projections.
Figure 19. Comparison of out-of-sample mortality rates (+) for males with the mortality projections under the proposed model and ONS projections.
Figure 20. Comparison of out of sample mortality rates (+) for females with the mortality projections under the proposed model and ONS projections.
In Figures 17 and 18, we compare our 2025 and 2055 mortality projections, this time incorporating the expert opinion and other adjustments we mentioned, with the 2014-based ONS projections. These adjustments, especially incorporating the expert opinion, visibly decreases the difference between two sets of projections. Also, restricting continuous year-on-year improvements by age enables us to avoid implausible long-term projections at ages where we estimate high mortality improvements from the model.

Finally, in order to investigate the forecast performance of the proposed model, we present the 2014–2017 forecasts against the out of sample outcomes obtained from ONS (2018). Note that we used mortality data between 1961 and 2013 to estimate the model parameters. In Figures 19 and 20, we present the mortality projections for males and females under the proposed model, the realised mortality rates for a 4-year forecast horizon and the 2014-based mortality projections of the ONS for the same horizon. From these figures, it can be seen that for the majority of the ages, mortality projections under the proposed model adhere acceptably well to the realised mortality rates both for males and females.

6. Conclusion

In this paper, we have developed a method for estimating and projecting mortality rates, which takes advantage of the ease with which a wide range of smooth and parametric models can routinely be fitted. We model mortality improvement using generalised additive models for ages where we have reliable data and a parametric model at older ages where the data are sparse. Our methodology is based on a smooth transition between a GAM at lower ages and a fully parametric model at higher ages. To obtain the estimates, we use the maximum-likelihood method. The approach described in this paper provides a computationally straightforward way of estimating and projecting mortality rates across the whole age range, including older ages where data are sparse or non-existent. Furthermore, our approach allows uncertainty and expert opinion to be coherently incorporated.

Under a fully Bayesian estimation method, all different sources of uncertainty – in data series, model parameters, including the choice of the model cut-off $x_0$, as well as expert judgement – would be treated jointly in a coherent, fully probabilistic manner. However, this would come at a considerable expense in terms of computing effort. In contrast, our method is very simple and is therefore computationally very cheap and easy to implement, as it requires no Markov chain Monte Carlo sampling and is based on pre-existing R functions. For that reason, the approach proposed in this paper offers an appealing alternative for implementing a sophisticated and robust analytical method in actuarial practice.

Acknowledgments

All the views presented in this paper are of the authors only.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The majority of this work was supported through a research contract to review the methodology for projecting mortality between the UK Office for National Statistics (ONS) and the University of Southampton. This included helpful discussions with ONS colleagues and members of the Mortality Assumptions Expert Users Group, convened by ONS. Additional support was provided by the ESRC Centre for Population Change – Phase II (ES/K007394/1).
References


