Canadian Institute of Actuaries | Institut canadien des actuaires

Research paper

Constructing Two-Dimensional Mortality Improvement Scales for Canadian Pension Plans and Insurers: A Stochastic Modelling Approach

Prepared by:
Johnny Siu-Hang Li
Department of Statistics and Actuarial Science, University of Waterloo

Yanxin Liu
Department of Finance, University of Nebraska Lincoln, Nebraska, U.S.A

February 2019

Document 219015

Ce document est disponible en français

Constructing Two-Dimensional Mortality Improvement Scales for Canadian Pension Plans and Insurers: A Stochastic Modelling Approach[1]

Johnny Siu-Hang Li[2] and Yanxin Liu[3]

Abstract

Recently, the actuarial professions in Canada, the U.S. and the UK have adopted an innovative two-dimensional approach to projecting future mortality. In contrast to the conventional approach, the two-dimensional approach permits mortality improvement rates to vary not only with age but also with time. Despite being an important breakthrough, the newly proposed two-dimensional mortality improvement scales are subject to several significant limitations, most notably a heavy reliance on expert judgments and a lack of measures of uncertainty. In this paper, we aim to develop a method for producing two-dimensional mortality improvement scales with more solid statistical justifications. To this end, we propose a 'heat wave' model, in which short- and long-term mortality improvements are treated respectively as 'heat waves' that taper off over time and 'background improvements' that always exist. Using the model, one can derive two-dimensional mortality improvement scales with minimal expert judgment. Moreover, with likelihood-based inference methods, the uncertainty surrounding the best estimate of mortality improvement rates can be quantified.

1 Introduction

In the developed world, life expectancy has been rising steadily except during periods of war and worldwide pandemic outbreaks. To incorporate future mortality improvements into pricing and valuation, actuaries often rely on mortality improvement scales, which specify the expected rates of reduction in mortality by means of some simple mathematical formulas. A projection of future mortality can be obtained by applying an appropriate mortality improvement scale to a 'base mortality table', which reflects the current mortality level of the pool of pensioners or insured lives under consideration.

For quite some time, the Society of Actuaries (SOA) (1995) Scale AA has been widely used by pension plans in Canada and the U.S. for valuation purposes. Under Scale AA, death probabilities t years after the base year t_b are calculated using the following equation:

$$q_{x,t_b+t} = q_{x,t_b}(1 - AA_x)^t,$$

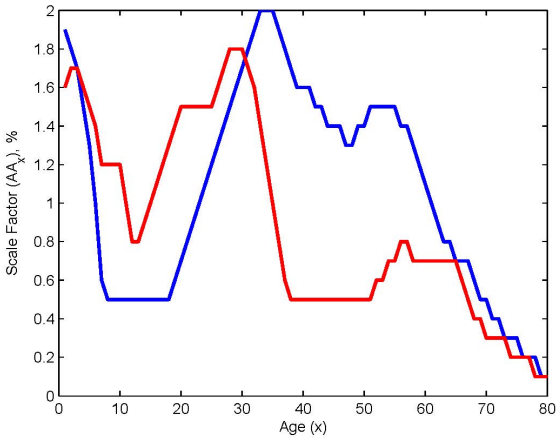
[1] Address correspondence to Johnny Siu-Hang Li, Department of Statistics and Actuarial Science, University of Waterloo, 200 University Ave. W., Waterloo, Ontario, Canada, N2L 3G1. Email: shli@uwaterloo.ca

[2] Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada.

[3] Department of Finance, University of Nebraska Lincoln, Nebraska, U.S.A.

where q_{x,t_b} is the mortality rate[4] for age x specified in the base table, and AA_x represents the expected mortality improvement rate at age x (see figure 1). One problem of Scale AA is that it understates the mortality improvement experienced in recent years. This problem can be seen in table 1, where we compare Scale AA with the annualized improvement rates that are realized over 1996–2005 (the decade after Scale AA was launched).

Figure 1: The values of AA_x in Scale AA, males and females.



Males are represented in blue.

Females are represented in red.

Table 1: Scale AA and the actual rates of reduction in Canadian mortality from 1996 to 2005 for various age groups. The actual rates of reduction are calculated using smoothed mortality rates that are derived by fitting a P-splines regression (Currie et al., 2004) to the raw data.

Age group	Scale AA	Actual Reduction (1996–2005)
Males		
-64	1.48%	2.71%
-69	1.36%	2.89%
-74	1.50%	2.84%
-79	1.28%	2.71%
Females		
-64	0.50%	1.77%
-69	0.50%	1.70%
-74	0.62%	1.69%
-79	0.74%	1.81%

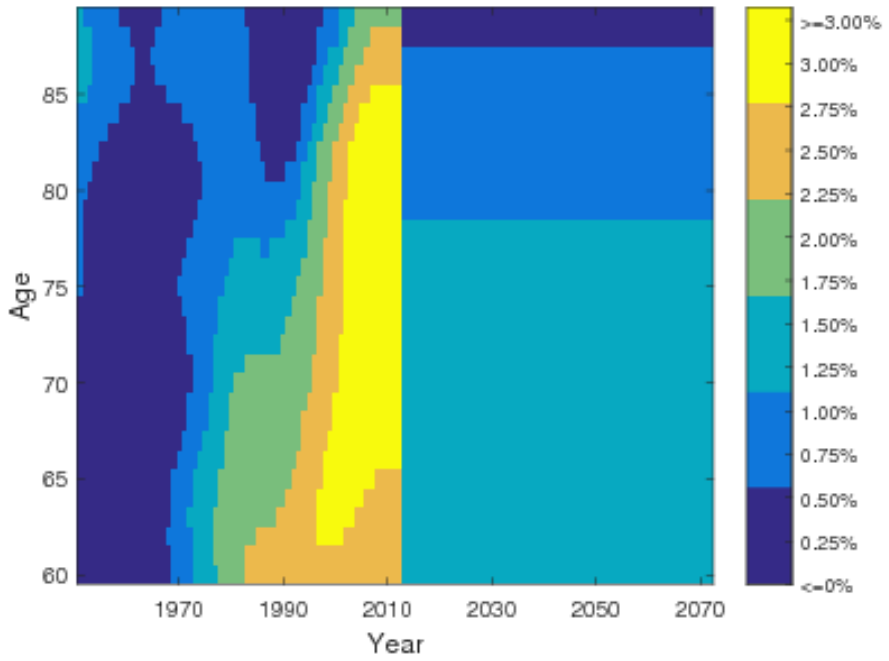
[4] More precisely, $q_{x,t}$ is the probability that an individual aged x exactly at time $t-1$ (i.e., the beginning of year t) dies during the time interval of $[t-1, t)$.

Another significant problem of Scale AA is that it does not yield a logical continuation of past mortality improvement rates. This problem can be visualized in figure 2, which shows the heat map of historical mortality improvement rates for Canadian males (the left portion) and the corresponding expected future mortality improvement rates implied by Scale AA (the right portion). The heat map can be interpreted as follows:

- Each row in the heat map represents mortality improvement rates for a specific age. The variation of colors along the vertical dimension thus represents the age effect of mortality improvement.
- Each column in the heat map represents mortality improvement rates for a specific calendar year. The variation of colors along the horizontal dimension therefore represents the period (time-related) effect of mortality improvement.
- Each diagonal (from lower-left to upper-right) in the heat map represents mortality improvement rates for a specific year of birth. Hence, the variation of colors among diagonals represents the cohort (year-of-birth-related) effect of mortality improvement.

Age, period and cohort effects are clearly observed in the historical mortality improvement rates. However, Scale AA only takes account of the age effect, because it simply assumes that future mortality improvement rates are constant over time. For the same reason, the transition of the heat map from past to future is not at all logical.

Figure 2: Heat map of historical mortality improvement rates for Canadian males (up to 2011) and mortality improvement rates implied by Scale AA for males (after 2011).



In recent years, the actuarial profession in North America has recognized the limitation of Scale AA, and has started to consider two-dimensional mortality improvement scales (of which the scale factors vary with both age and time instead of just age). Generally speaking, a two-dimensional mortality improvement scale is composed of the following three components.

1. A short-term scale for the near future: the scale factors in the short-term scale are generally high, reflecting the rapid mortality improvement observed in recent decades.
2. A long-term scale for the distant future: compared to those in the short-term scale, the scale factors in the long-term (ultimate) scale are much lower, incorporating the view that rapid mortality improvement will not last forever.
3. A mid-term scale for the transitional phase: the scale factors in the mid-term scale are obtained by interpolating between those in the long- and short-term scales.

In Canada, the Office of the Chief Actuary (2014) developed a two-dimensional mortality improvement scale for the purpose of actuarial valuations of the Canada Pension Plan and Canada's national Old Age Security (OAS) Program. The Canadian Institute of Actuaries (CIA) has also made a huge effort on developing mortality improvement scales. In 2009, its Committee on Canadian Pensioners' Mortality Experience commissioned a research project, an outcome of which is a two-dimensional improvement scale that is developed using data from the Canada/Québec Pension Plan and an array of assumptions (Adam, 2012). Then, in 2014, the CIA (2014) launched the CPM-B scale, a two-dimensional mortality improvement scale created for actuarial valuations for a broad range of pension plans in Canada. In April 2017, the CIA (2017) released the MI-2017 scale, an update of the CPM-B scale, derived using more recent mortality data. Later in the year, the Actuarial Standards Board proposed that MI-2017 be promulgated for the purposes of the valuation of insurance contracts in Canada. For the reader's information, the MI-2017 scale is reviewed in Section 2.

In the U.S., the topic of two-dimensional improvement scales was first studied in 2012 by the SOA's Retirement Plans Experience Committee (RPEC), which developed Scale BB to replace the already obsolete Scale AA (SOA, 2012). In 2014, the RPEC created another two-dimensional mortality improvement scale known as MP-2014 (SOA, 2014), applicable to a broad range of retirement programs in the U.S. The MP-2014 scale was subsequently updated in 2015, 2016 and 2017. The updated scales are respectively known as MP-2015, MP-2016 and MP-2017 (SOA, 2015, 2016 and 2017). For the reader's information, the MP-2017 scale is reviewed in Section 2.

Although the move from one to two dimensions represents an important breakthrough, the existing two-dimensional mortality improvement scales are still subject to several significant limitations.

First, the existing two-dimensional scales provide only a best estimate of future mortality, but give no measure of uncertainty surrounding the best estimate. Without any measure of uncertainty, the scales do not aid users in setting Margins for Adverse Deviations (MfADs). For the same reason, the scales do not provide sufficient information for assessing the underlying longevity risk and developing risk management solutions (e.g., longevity swaps).

Second, the production of the existing two-dimensional scales cannot be regarded as an exact science. More specifically, the long-term scale factors in the existing two-dimensional scales were determined by expert judgments and/or by making reference to the long-term mortality assumptions used in other jurisdictions. Statistical justifications for the assumed long-term scale factors are yet to be sought. Likewise, the durations of the transitional phases in the existing two-dimensional scales were decided subjectively. As valuation results are typically sensitive to the assumed transitional phase duration, questions arise as to whether this important parameter can be estimated more rigorously.

In view of the aforementioned limitations, this study is set out to develop a stochastic mortality model that allows us to produce two-dimensional mortality improvement scales that (1) are in line with the spirit of the existing two-dimensional scales in the sense that short-, medium- and long-term scale factors are different, (2) involve less subjectivity, (3) come with measures of uncertainty, and (4) are easy to implement with spreadsheet programs and actuarial software. To this end, we propose the 'heat wave' model, which is composed of the following components:

i. Background improvements: similar to the concept of background radiation in physics, long-term mortality improvements are regarded as background improvements that always exist. They are modeled by usual parametric structures and time-series processes such as ARIMA.

ii. Heat waves: in line with the conjecture behind the existing two-dimensional scales, recent rapid mortality improvements are considered as 'heat waves', which are expected to diminish over time. They are modeled using methods similar to wavelets and Fourier series in time-series analysis. This component also decides the speed at which mortality improvement rates converge to their long-term values.

The distinction between background improvements and heat waves makes our proposed model stand out from typical stochastic mortality models in which the evolution of mortality is driven entirely by one or more time-series processes. Because the proposed model contains some bounded parameters, maximum likelihood estimates are not straightforward to obtain. To overcome this technical challenge, the barrier method for constrained optimization is utilized (Nocedal and Wright, 1999). Empirically, the estimated heat wave models yield scale factors that extend the patterns of historical mortality improvement rates in a logical manner.

To our knowledge, the heat wave model is the unique approach that satisfies all of the four criteria. Both the Lee–Carter model (Lee and Carter, 1992) and the Cairns–Blake–Dowd model (Cairns et al., 2006) imply scale factors that are invariant with time (i.e., one-dimensional). The cohort generalizations of these models (Cairns et al., 2009) do yield scale factors that vary with both age and time, but, as we demonstrate later in this article, the variation is not quite reasonable. The approach considered recently by the Continuous Mortality Investigation Bureau (2017a, 2017b) of the Institute and Faculty of Actuaries in the UK is somewhat more model-based and data-driven, but it still contains no measure of uncertainty. The method suggested by Cairns (2017) comes with measures of uncertainty, but it is not entirely statistically rigorous as it entails a subjective tweak of the drifts of the underlying time-series processes.

We have applied the proposed method to data for both Canada and the U.S., and for both genders. However, for the sake of space, we only report results that are generated using data for Canadian males. All of the required data (death counts and exposures) are obtained from the Human Mortality Database.

The rest of this paper is structured as follows. Section 2 briefly reviews the two-dimensional mortality improvement scales that are currently used in Canada and the U.S. Section 3 presents the heat wave model and discusses the demographic intuitions behind the model. Section 4 details the estimation procedure and presents the estimation results. Section 5 compares the heat wave model with several well-known existing stochastic mortality models. Section 6 explains how measures of uncertainty can be derived. Section 7 concludes the paper.

2 A Review of MI-2017 and MP-2017

The CIA's MI-2017 and the SOA's MP-2017 are constructed by the same principles, both of which contain a short-term scale for projecting mortality in the near future, a long-term scale for projecting mortality in the distant future, and a mid-term scale for projecting mortality in the transitional phase. We now describe how the three components in MI-2017 and MP-2017 are obtained.

2.1 MI-2017

According to the CIA (2017), the short-, mid- and long-term components of MI-2017 are constructed as follows:

- *Short-term*

  In MI-2017, the short-term rates are set equal to the mortality improvement rates implied by a two-dimensional Whittaker–Henderson graduation that is applied to data over the period of 1970 to 2015. The data are obtained from the Human Mortality Database (1970–2011) and the Canadian OAS Program plan (2012–2015).

  The short-term rates at and beyond age 105 are set to zero. For ages 96 to 104, the short-term rates are calculated by linearly interpolating between the improvement rates (obtained from the Whittaker–Henderson graduation) at age 95 and 105.

  The short-term rates are applicable up to year 2013, after which the mid-term scale kicks in. The two-year set-back (2013 instead of 2015) is because of an undesirable edge effect that may have possibly incurred in the graduation.

- *Long-term*

  In MI-2017, the long-term rates are obtained by trending past observations and considering a range of opinions from experts in the field. The long-term improvement rate for all ages up to and including 90 is set to 1.0% per annum. This value is trended linearly to 0.2% per annum at age 100, and further trended linearly to 0% at age 105. The long-term rate beyond age 105 is set to 0%.

- *Mid-term*

  In MI-2017, the transitional phases for different age groups are different. For ages 0 to 40, the mid-term scale applies to 2014 to 2023 (10 years). For ages 60 and above, the mid-term scale applies to 2014 to 2033 (20 years). For ages 41 to 59, the lengths of the transitional phases are obtained by linear interpolations.

  The mid-term rates are determined by cubic polynomials, one for each age. The four parameters in each cubic polynomial are calculated by using the corresponding improvement rates at the beginning and end points of the transitional phase, setting the slope of the polynomial to zero at the end of the transitional phrase, and setting the slope of the polynomial to the slope of the corresponding short-term rates between 2012 and 2013 (subject to a maximum absolute value of 0.003).

2.2 MP-2017

According to the SOA (2017), the short-, long- and mid-term components of MP-2017 are constructed as follows:

- *Short-term*

  In MP-2017, the short-term rates are obtained by applying a two-dimensional Whittaker–Henderson graduation to data over the period of 1951 to 2015. The short-term rates are applicable to projections up to 2013. The two-year set-back from 2015 is, again, to mitigate any potential for increased sensitivity around the edges of the graduated data.

- *Long-term*

  In MP-2017, the long-term rate is set to 1.0% per annum for all ages up to and including 85. The long-term rate is trended linearly to 0.85% per annum at age 95, and further trended to 0% at age 115. The choice of the 1.0% long-term rate has remained unchanged since the MP-20xx scale was first released in 2014. According to the SOA (2014), the choice is based on the long-term averages of the U.S. population's historical mortality improvement rates, the US Social Security Administration's intermediate-cost assumption, and expert opinions.

- *Mid-term*

  In MP-2017, the mid-term rates are derived from a 'double cubic interpolation' methodology that blends the short- and long-term rates. Each mid-term rate is computed as the simple arithmetic average of the values developed from two separate cubic interpolations. The first interpolation, which spans 10 years, is performed across a fixed age path. The second interpolation, which spans 20 years, is performed along a fixed year-of-birth path. The reader is referred to the SOA (2014) for details concerning how the cubic polynomials are calibrated.

3 The Heat Wave Model

3.1 Definition

Let $m_{x,t}$ be the underlying central rate of death at age x and in calendar year t. Suppose that the data set under consideration spans an age range of $[x_0, x_1]$ and a sample period of $[t_0, t_1]$. The heat wave model is defined as follows:

$$\ln(m_{x,t}) = a_x + b_x k_t + c_x \sum_{j=t_0}^{t} f(x, t; \vec{\theta}), \tag{1}$$

where a_x, b_x and c_x are age-specific parameters, k_t is a time-varying parameter, and f is function of age and time with a parameter vector $\vec{\theta}$.

In equation (1), $a_x + b_x k_t$ is the basic Lee–Carter structure, which, in our modelling approach, captures the background mortality improvement. As in the original Lee–Carter model, the evolution of k_t is captured by a random walk with drift:

$$k_t = d + k_{t-1} + \varepsilon_t, \tag{2}$$

where d is the drift term and $\{\varepsilon_t\}$ is a sequence of independent and identically distributed normal random variables with a zero mean and a constant variance. We also use the original Lee–Carter identifiability constraints, $\sum_{x=x_0}^{x_1} b_x = 1$ and $\sum_{t=t_0}^{t_1} k_t = 0$, to ensure parameter uniqueness.

On the other hand, $f(x, t; \vec{\theta})$ captures the heat wave. We set this function to the probability density function of a normal distribution:

$$f(x, t; \vec{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{((t-t_0)-(\mu+(x-x_0)h))^2}{2\sigma^2}\right), \tag{3}$$

where $\vec{\theta} = (\mu, \sigma, h)'$ is the vector of heat wave parameters that determine the location, size and pattern of the heat wave. Although other parametric functions may be used, we choose to use this particular function because it leads to improvement rates that are straightforward to interpret.

For notational convenience, we let

$$g(x, t; \vec{\theta}) = \sum_{j=t_0}^{t} f(x, t; \vec{\theta}),$$

which can be understood as the cumulative effect of the heat wave up to and including time $t > t_0$. In terms of $g(x, t; \vec{\theta})$, the heat wave model can be rewritten more compactly as

$$\ln(m_{x,t}) = a_x + b_x k_t + c_x g(x, t; \vec{\theta}). \tag{4}$$

3.2 Implied Mortality Improvement Rates

The interpretation of the heat wave model becomes clearer when we consider the expected mortality improvement implied by the model. Switching off the random term in equation (2), the change in the log central death rate for age x from year $t-1$ to t is given by

$$\ln(m_{x,t}) - \ln(m_{x,t-1}) = b_x d + c_x f(x,t;\vec{\theta}) \quad (5)$$

$$= b_x d + c_x \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{((t-t_0)-(\mu+(x-x_0)h))^2}{2\sigma^2}\right), \quad (6)$$

for $t = t_1+1, t_1+2, \ldots$.

It is clear that the first term ($b_x d$) in the above captures the long-term (background) improvement, which, by definition, depends on age (x) but not time (t). We conjecture that the long-term improvement rates for all ages are of the same sign. For this reason, we require $b_x > 0$ for all $x = x_0, \ldots, x_1$.[5] With this constraint, we anticipate that k_t is downward sloping, which in turn means that d is negative.

The other term corresponds to the heat wave, which reduces asymptotically to zero as t tends to infinity. Parameter c_x allows the impact of the heat wave to be age dependent. In the extreme case when $c_x = 0$, the mortality improvement at age x is completely unaffected by the heat wave. As $f(x,t;\vec{\theta})$ is non-negative, we require $c_x < 0$ for all $x = x_0, \ldots, x_1$ so that the heat wave absorbs excess mortality improvement (instead of deterioration). All of the three heat wave parameters (μ, σ, h) are highly interpretable.

First, σ controls the speed at which the heat wave tapers off. As the normal density becomes fairly close to zero (0.054) at two standard deviations above mean, we may regard $2 \times \sigma$ as the approximate time for the improvement rates to converge from their peak values to their long-term values (i.e., the convergence period). Of course, σ must be strictly positive. In addition, as the convergence period can neither be too short (say less than eight years) nor too long (say more than 60 years), we further require $4 < \sigma < 30$.

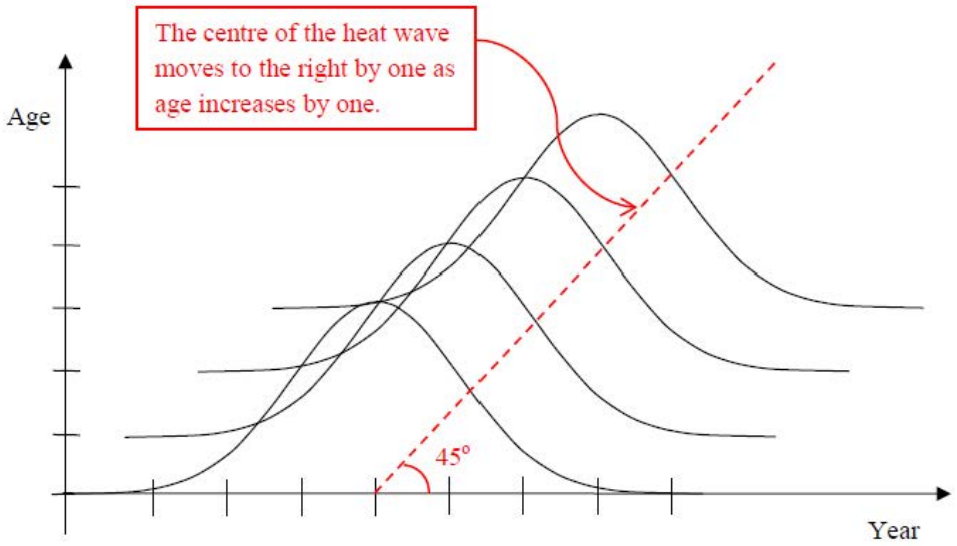
Second, μ determines the location of the heat wave. For age x_0 (the lowest age in the age range under consideration), the peak of mortality occurs in year $t_0 + \mu$. Given the patterns in typical mortality heat maps, this peak should be observed within the sample period of the data set, and for this reason we require $1 < \mu < t_1 - t_0$. More generally, for $x = x_0, \ldots, x_1$, the peak of the heat wave occurs in year $t_0 + \mu + (x - x_0)h$.

Finally, h reflects the mix between period and cohort effects in the heat wave. The meaning of h is illustrated in figure 3. In one extreme when $h = 1$, the centre of a heat wave increases by one year as age increases by one year. Consequently, the heat waves align perfectly diagonally, suggesting that the excess mortality improvement over the background rates is due entirely to cohort effects. In the other extreme when $h = 0$, the centre of a heat wave does not change with age. Consequently, the heat waves align perfectly vertically, indicating that the excess mortality improvement over the background rates is an outcome of period effects only.

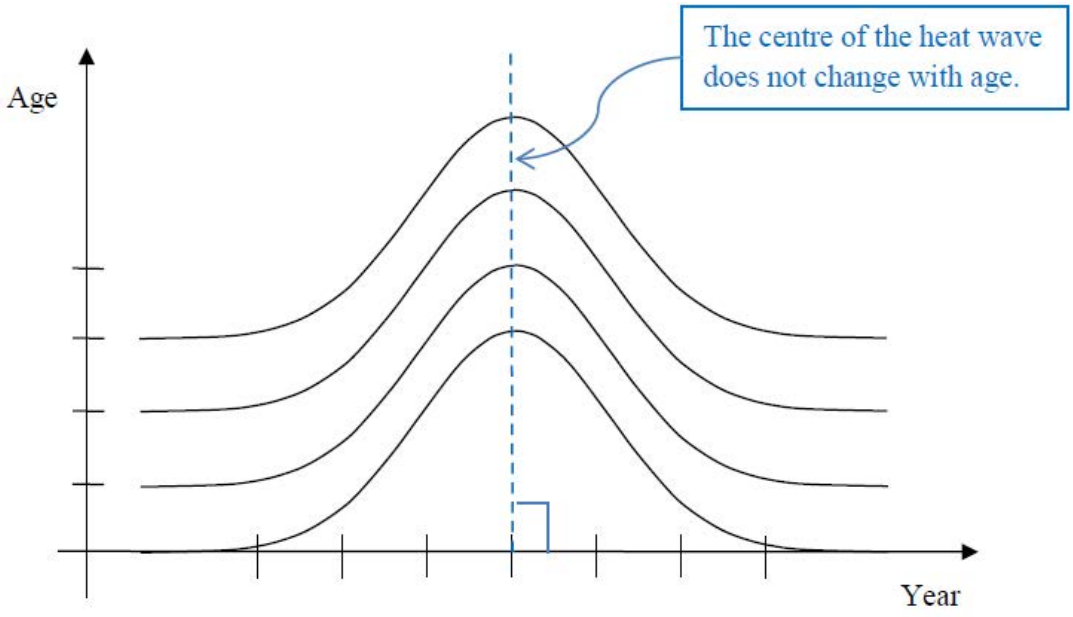
---

[5] We cannot require $b_x < 0$ for all $x = x_0, \ldots, x_1$, because of the identifiability constraint $\sum_{x=x_0}^{x_1} b_x = 1$ used.

Figure 3: Illustrative patterns of the heat waves when $h = 1$ (upper panel) and $h = 0$ (lower panel).



$h = 1$



$h = 0$

In practice, instead of absolute changes in log death rates, scale factors are expressed in terms of relative (percentage) changes in death rates; that is,

$$R(x,t) = 1 - \frac{m_{x,t}}{m_{x,t-1}},$$

$t = t_1 + 1, t_1 + 2, \ldots$, where $R(x,t)$ stands for the mortality improvement scale factor for age x and calendar year t. It follows from equation (6) that

$$R(x,t) = 1 - \exp\left(b_x d + c_x \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{((t-t_0)-(\mu+(x-x_0)h))^2}{2\sigma^2}\right)\right), \quad (7)$$

$t = t_1 + 1, t_1 + 2, \ldots$, under the heat wave model when the random innovations in equation (2) are switched off. This function can be implemented straightforwardly in spreadsheet programs and actuarial software.

4 Estimation

4.1 Estimation Method

We now explain how the heat wave model may be estimated.

First, we derive a likelihood function for estimating the parameters in equation (1). The likelihood function is based on a Poisson death count assumption. Let $D_{x,t}$ be the observed number of death counts at age x and year t, and $E_{x,t}$ be the corresponding number of exposures.[6] The likelihood function l is obtained by assuming that $D_{x,t}$ is a realization of a Poisson distribution with a mean of $E_{x,t}m_{x,t}$, where $m_{x,t}$ follows the specification in equation (1).

Second, we use the barrier method to incorporate the inequality constraints:

$$b_x > 0 \text{ and } c_x < 0, \qquad x = x_0, \ldots, x_1.$$

In more detail, we subtract barrier functions for parameters b_x and c_x from the original likelihood function l, forming an objective function $l^{(B)}$ which is maximized to obtain model parameter estimates. Each barrier function is created in such a way that it approaches positive infinity as its associated parameter approaches its boundary, thereby preventing the resulting parameter estimate from exceeding the boundary.

Third, initial values for the optimization (maximization) process are chosen. We choose the initial values of μ, h, and σ by considering the patterns of historical mortality improvements. The initial values of other parameters (c_x, a_x, b_x, and k_t) are obtained by running a (partial) maximum likelihood estimation that is conditioned on the chosen initial values of μ, h, and σ. Good initial values can stabilize and expedite the optimization process.

Fourth, we maximize the objective function $l^{(B)}$ using an interative Newton's method, in which parameters are updated one at a time. At the end of each iteration (i.e., when all of the parameters in equation (1) are updated), the estimates of b_x and k_t are rescaled so that they

[6] The observed value of $m_{x,t}$ is the ratio of $D_{x,t}$ to $E_{x,t}$.

sum to 1 and 0, respectively.[7] The iterations stop when the change in $l^{(B)}$ is smaller than a pre-specified tolerance level, say 10^{-6}.

Finally, given the estimates of $k_{t_0}, \ldots, k_{t_1}$, the drift term d in equation (2) can be estimated readily. Using the method of conditional least squares, the best estimate of d is simply $(k_{t_1} - k_{t_0})/(t_1 - t_0 - 1)$.

4.2 Estimation Results

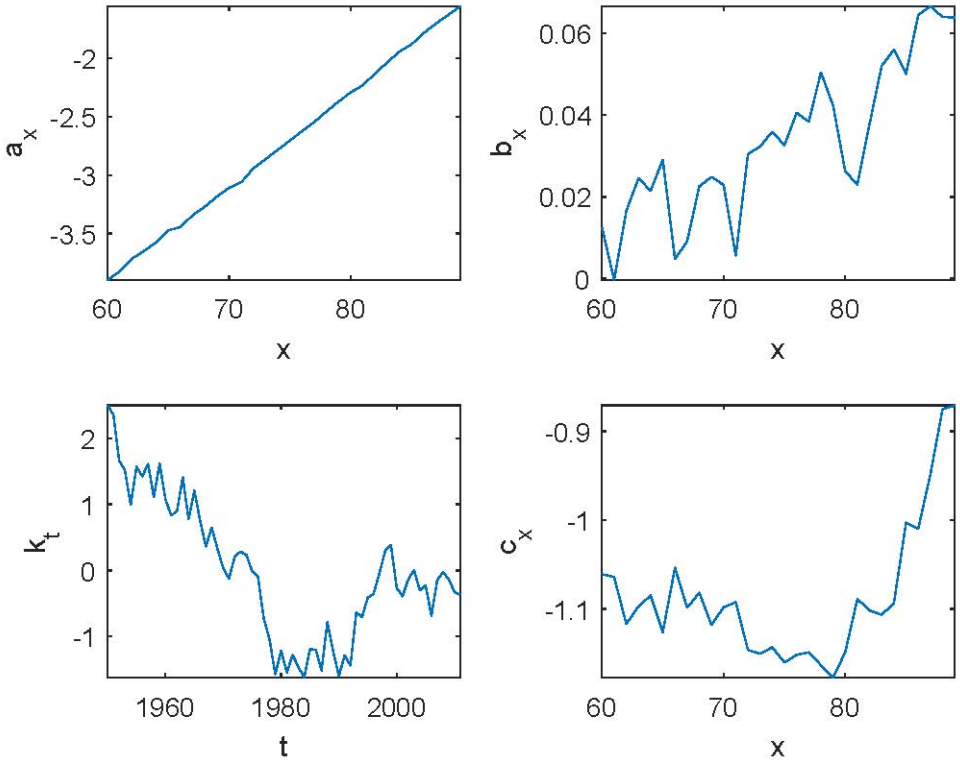
4.2.1 Parameter Estimates

We fit the heat wave model to data from Canadian male population over a sample period of $[t_0, t_1] = [1950,2011]$ and a sample age range of $[x_0, x_1] = [60,89]$. The estimates of μ, σ, and h are 44.5238, 14.4002 and 0.7549, respectively. The estimate of h indicates that the heat wave (i.e., the rapid mortality improvement observed in the past two decades) is an outcome of both cohort and period effects, with cohort effects being more influential (75% vs. 25%). The estimate of σ suggests that the convergence period (i.e., the period over which mortality improvement rates converge from their current values to their long-term values) is approximately $2 \times 14.4002 \approx 29$ years.

The estimates of the non-heat-wave parameters are presented in figure 4. Similar to the original Lee–Carter model, the estimate of a_x increases fairly linearly with age. The estimates of c_x over the sample age range suggest that mortality improvement at younger ages tends to be more responsive to the heat wave. Based on the estimated values of $k_{t_0}, \ldots, k_{t_1}$, the estimate of the drift d in equation (2) is -0.0471.

[7] This step is taken as we use the original Lee–Carter identifiability constraints ($\sum_{x=x_0}^{x_1} b_x = 1$ and $\sum_{t=t_0}^{t_1} k_t = 0$) to stipulate parameter uniqueness.

Figure 4: The estimates of a_x, b_x, c_x and k_t for $x = 60, \ldots, 89$ and $t = 1950, \ldots, 2011$.



4.2.2 Goodness-of-fit

We now analyze the goodness-of-fit of the heat wave model. As the heat wave model is developed from the original Lee–Carter model, in this analysis we use the Lee–Carter as the benchmark model.

The maximized log-likelihood values for the heat wave model and the original Lee–Carter model are -10243.4 and -11386.0, respectively. The difference in the log-likelihood values suggests that the heat wave model provides a better fit to the historical data.

Noting that the heat wave model contains more parameters than the original Lee–Carter model, we also compare the two models in terms of the Bayesian Information Criterion (BIC; Schwarz, 1978), which is defined as

$$\mathrm{BIC} = \hat{l} - 0.5 \times n_p \times n_d, \tag{8}$$

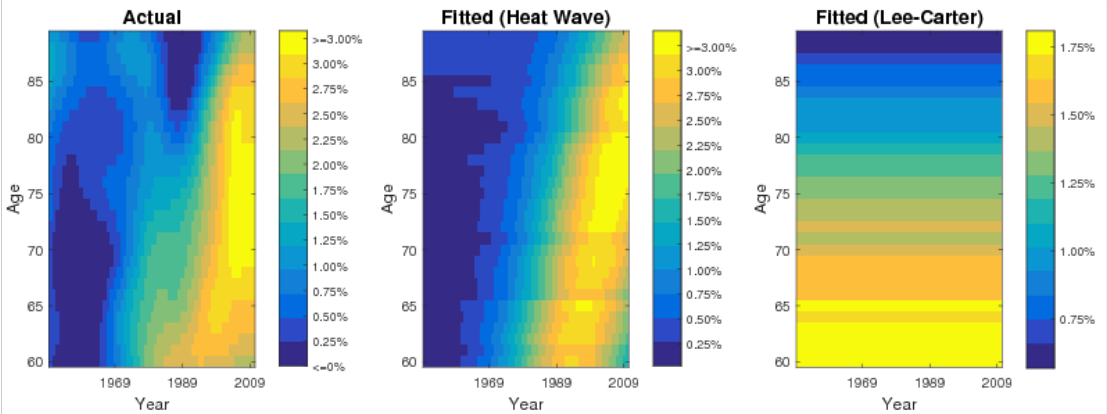
where $\hat{l}$ denotes the maximized log-likelihood, n_p represents the number of model parameters and $n_d = (x_1 - x_0 + 1) \times (t_1 - t_0 + 1)$ is the number of data points used. The BIC includes a penalty for the use of model parameters: the larger the number of model parameters is, the larger the penalty is. Given how the BIC is defined, a model with a higher BIC value is preferred. The values of the BIC for the heat wave model and the original Lee–Carter model are -10819.3 and -11837.8, respectively. These BIC values suggest that the heat wave model outperforms the original Lee–Carter model even when the use of additional model parameters is taken into consideration.

Figure 5 compares the actual mortality improvement rates over the sample period with the fitted mortality improvement rates produced by the heat wave and the original Lee–Carter models. For the heat wave model, the fitted mortality improvement rates are calculated using equation (7) and the estimated model parameters. For the original Lee–Carter model, the fitted mortality improvement rates are calculated using the following formula and the estimated model parameters:

$$R(x,t) = 1 - \exp(b_x d). \tag{9}$$

It can be observed that compared to the original Lee–Carter model, the heat wave model can much more accurately capture the non-trivial pattern of the historical mortality improvement rates.

Figure 5: The actual mortality improvement rates (left panel) and the fitted mortality improvement rates produced by the heat wave model (middle panel) and the original Lee–Carter model (right panel) over the sample period of 1950 to 2011.



4.2.3 Mortality Projections

We now turn to mortality projections. The top two panels of figure 6 display the heat maps of the expected future mortality improvement rates implied by the heat wave model and the original Lee–Carter model, respectively. [8] To facilitate analyses, the historical mortality improvement rates are also included in the heat maps.

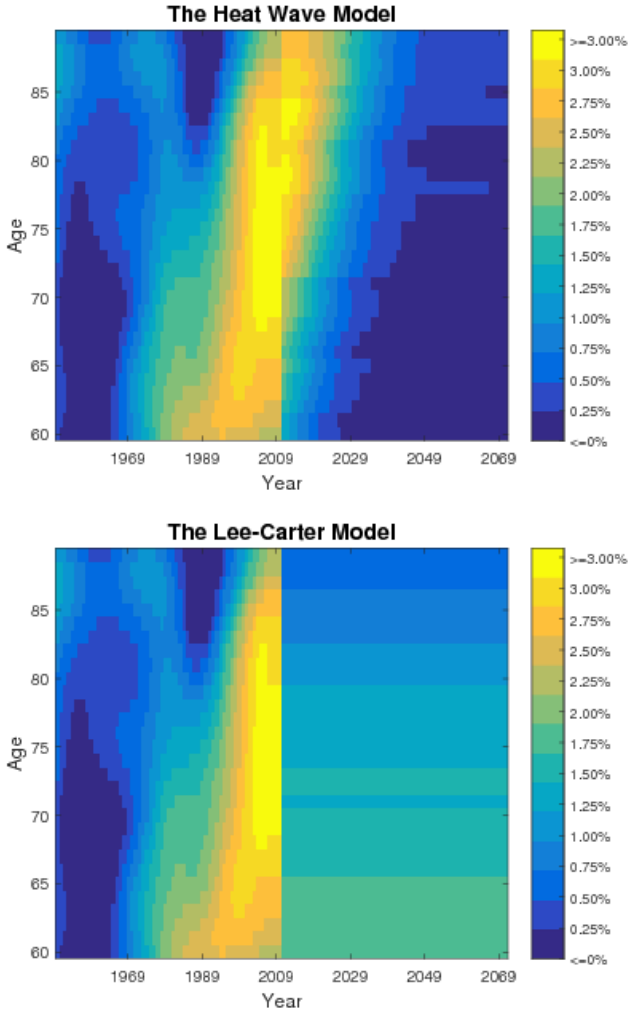
The pattern of the expected future mortality improvement rates generated from the heat wave model appears to be a logical extension to that of the historical mortality improvement rates. Features including cohort effects (variation of colors across the diagonal dimension) are preserved. In contrast, as equation (9) implies, the original Lee–Carter model produces a projected heat map that shows no variation along both the horizontal and vertical dimensions.

[8] For ease of exposition, the historical mortality improvement rates shown in the diagrams are pre-smoothed. It is important to note that the heat wave model and the Lee-Carter model are calibrated to raw data rather than pre-smoothed data.

This overly simple pattern does not seem to be a reasonable extension to the pattern of the heat map of historical mortality improvement rates.

In the lowest panel of figure 6 we show the heat map of the expected future mortality improvement rates specified in MI-2017. Compared to the heat wave model, MI-2017 produces less conservative mortality improvement rates in the short term (the warm colors observed in the heat map of historical mortality improvement rates are not found in the heat map produced by MI-2017), but the opposite is true in the long term. We remark that a smoother transition from historical to projected MI-2017 improvement rates is observed in the heat maps provided in the CIA (2017) report. The difference may be attributed to the fact that the smoothed improvement rates in those heat maps are produced by a different method (Whittaker-Henderson graduation).

Figure 6: The heat maps of the expected mortality improvement rates (2012 and onwards) implied by the heat wave model (upper panel), the original Lee–Carter model (middle panel), and MI-2017 (lower panel). The actual mortality improvement rates over the sample period are also shown.



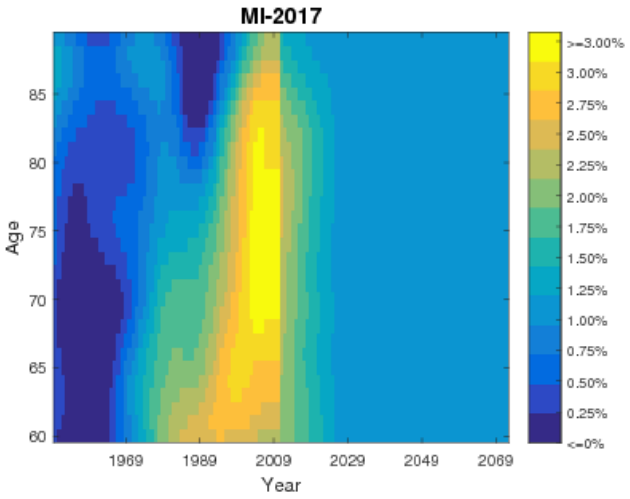
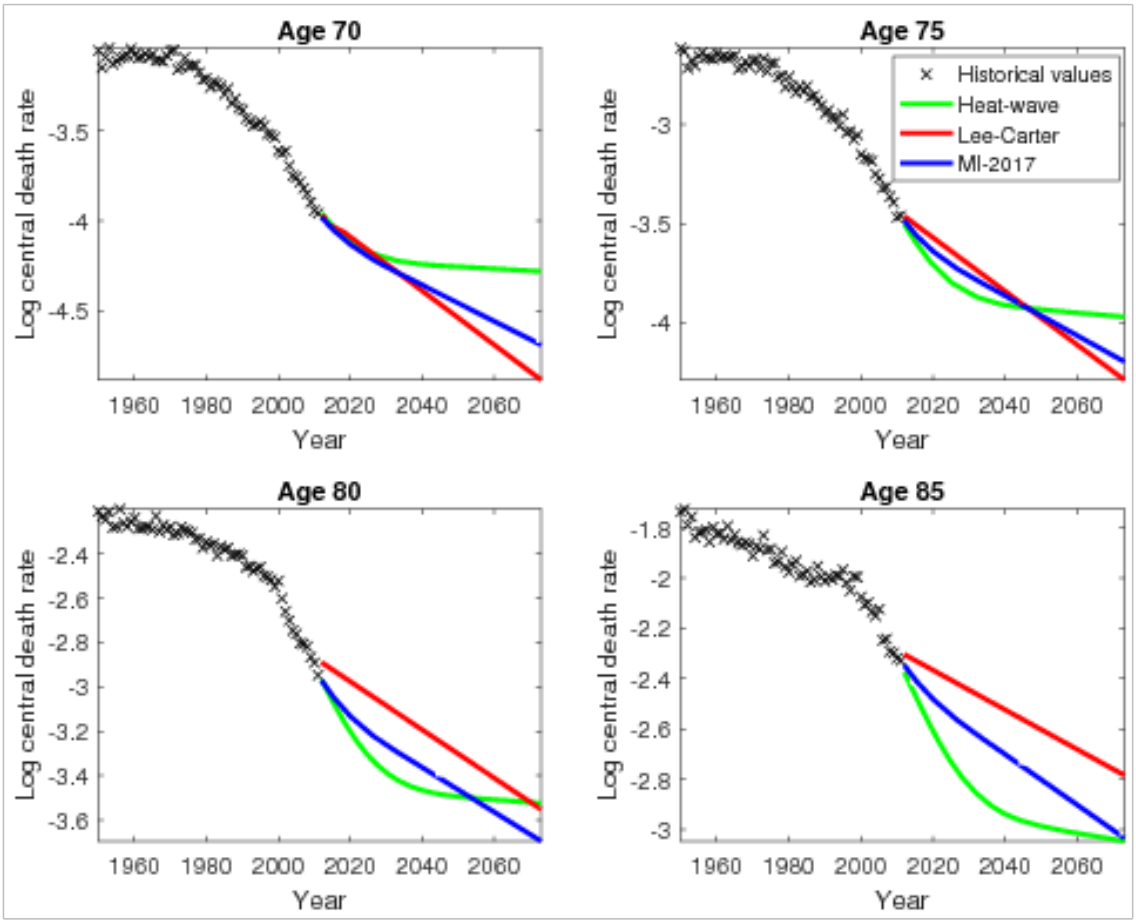


Figure 7 compares the age-specific central rates of death projected by the heat wave model, the original Lee–Carter model and MI-2017. The original Lee–Carter model yields a purely linear projection, with a gradient that is not quite in line with the pace of reduction observed in the past two decades. Both the heat wave model and MI-2017 produce non-linear projections, but with different degrees of conservatism. The heat wave model implies more aggressive mortality improvement at the higher end of the age range, but the opposite is true in at the lower end. Also, in the (very) long run MI-2017 tends to yield lower projected mortality rates, as its ultimate scale factors are generally higher than the long-term improvement rates implied by the heat wave model.

Figure 7: Age-specific central rates of death (in log scale), 2012 and onwards, projected by the heat wave model, the original Lee–Carter model and MI-2017.



5 Comparison with Existing Stochastic Mortality Models

In this section, we compare the heat wave model with several other existing stochastic mortality models that are widely used in the literature. These models include:

- The original Cairns–Blake–Dowd model (aka Model M5; Cairns et al., 2006)

$$\ln\left(\frac{q_{x,t}}{1-q_{x,t}}\right) = k_t^{(1)} + k_t^{(2)}(x-\bar{x})$$

- The Cairns–Blake–Dowd model with a quadratic age effect (aka Model M6; Cairns et al., 2009)

$$\ln\left(\frac{q_{x,t}}{1-q_{x,t}}\right) = k_t^{(1)} + k_t^{(2)}(x-\bar{x}) + \gamma_c$$

- The Cairns–Blake–Dowd model with a quadratic age effect and a cohort effect (aka Model M7; Cairns et al., 2009)

$$\ln\left(\frac{q_{x,t}}{1-q_{x,t}}\right) = k_t^{(1)} + k_t^{(2)}(x-\bar{x}) + k_t^{(3)}((x-\bar{x})^2 - \hat{\sigma}_x^2) + \gamma_c$$

- The Plat model (Plat, 2009)

$$\ln(m_{x,t}) = a_x + k_t^{(1)} + k_t^{(2)}(\bar{x}-x) + k_t^{(3)}(\bar{x}-x)_+ + \gamma_c$$

- The simplified Plat model (Plat, 2009)

$$\ln(m_{x,t}) = a_x + k_t^{(1)} + k_t^{(2)}(\bar{x}-x) + \gamma_c$$

In the above, $q_{x,t} \approx 1 - \exp(-m_{x,t})$ represents the conditional probability of death at age x and in year t, a_x is an age-specific parameter, $k_t^{(i)}$, $i = 1,2,3$, is a stochastic factor that depends on time (t), γ_c is a stochastic factor that depends on year of birth $(c = t - x)$, $\bar{x}$ represents the mean age over the sample age range, $\hat{\sigma}_x$ is the mean of $(x-\bar{x})^2$ over the sample age range, and $(\bar{x}-x)_+$ represents the minimum of $(\bar{x}-x)$ and zero.

The models are fitted to the data for Canadian male population using the method of Poisson maximum likelihood. The identifiability constraints used are the same as those used in the original work of Cairns et al. (2009) and Plat (2009). We use a random walk with drift to model the evolution of k_t over time, and an ARMA(1,1) process to model the evolution of γ_c over year of birth.

Let us first compare the goodness-of-fit produced by the models. Table 2 reports the value of the BIC (defined in equation (8)) produced by each of the estimated models. Compared to the five alternative models, the heat wave model yields the highest (least negative) BIC value. The result suggests that the heat wave model outperforms not only the original Lee–Carter model but also five other commonly used stochastic mortality models in terms of goodness-of-fit, with the number of parameters taken into consideration.

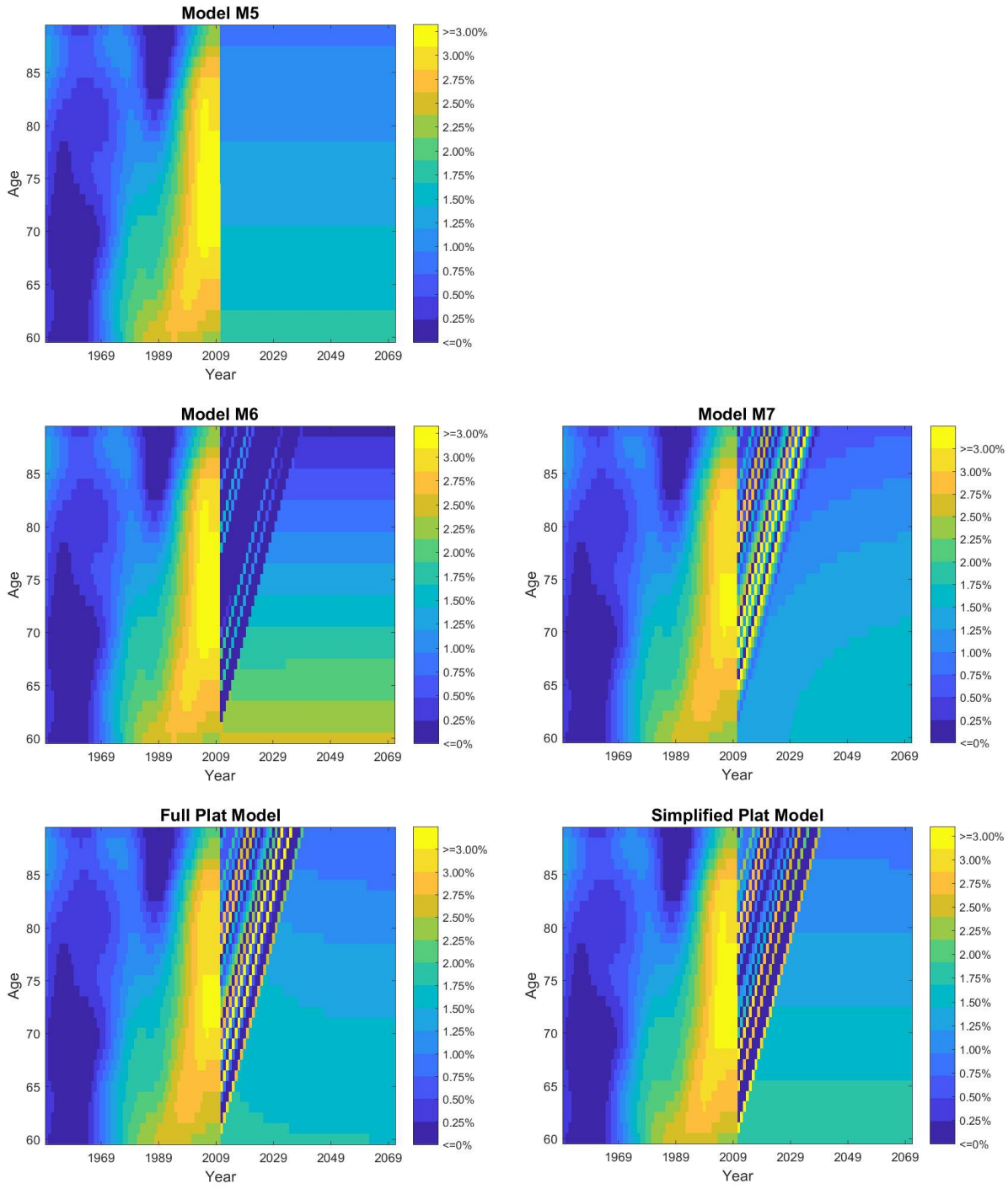
Table 2: The values of the BIC produced by the heat wave model and five other stochastic mortality models fitted to the data from Canadian male population.

Model	BIC
The Cairns–Blake–Dowd model	−11899
The Cairns–Blake–Dowd model with a cohort effect	−11031
The Cairns–Blake–Dowd model with quadratic age and cohort effects	−11157
The Plat model	−11137
The simplified Plat model	−10989
The heat wave model	−10819

We then examine the expected future mortality improvement rates implied by the five alternative models (figure 8). Model M5 suffers from the same problem as the original Lee–Carter model and Scale AA. It yields expected future mortality improvement rates that depend neither on time nor year of birth. This problem arises from the theoretical fact that under Model M5, the (expected) difference between $\ln(q_{x,t}/(1-q_{x,t}))$ and $\ln(q_{x,t-1}/(1-q_{x,t-1}))$ does not depend on t.

The other four models, all of which incorporate cohort effects, result in projected heat maps that exhibit some diagonal patterns. However, these diagonal patterns do not appear to be natural extensions of the patterns observed in the past. More importantly, these models do not fit into the framework of two-dimensional mortality improvement scales, in which short-term scale factors converge smoothly and gradually to the long-term scale factors.

Figure 8: The heat maps of the expected mortality improvement rates (2012 and onwards) implied by Model M5, Model M6, Model M7, the Plat model and the simplified Plat model. The actual mortality improvement rates over the sample period are also shown.



6 Measures of Uncertainty

One important aspect about the heat wave model is that it is able to produce measures of uncertainty surrounding the expected future mortality improvement rates. This section outlines the derivation of such measures of uncertainty.

As equation (5) indicates, under the heat wave model the expected change in the log central death rate at age x between years $t-1$ and t is

$$v(\vec{\theta}^*) = b_x d + c_x f(x,t;\vec{\theta}),$$

which is a function of a vector of six model parameters,

$$\vec{\theta}^* = (b_x, c_x, \mu, \sigma, h, d)'.$$

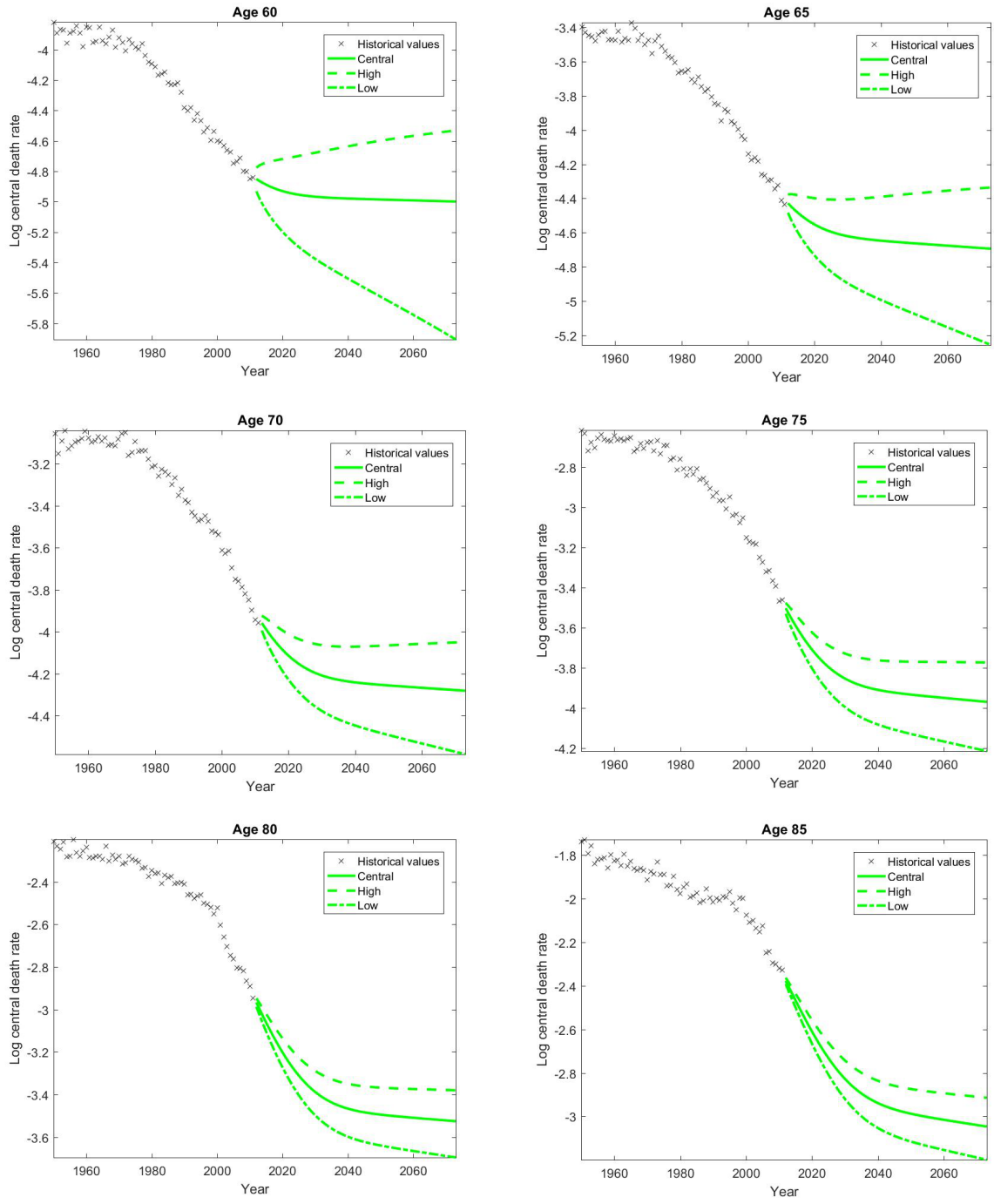
Of course, the true values of these six parameters are never known. Because we can only evaluate v using the estimated values of the six model parameters, the calculated mortality improvement rates are subject to the uncertainty surrounding the estimates of the six model parameters.

Let $\hat{\vec{\theta}}^*$ be the estimate of the vector of six model parameters. Using the multivariate delta method, the variance of $v(\vec{\theta}^*)$ can be approximated as a function of the information matrix of $\vec{\theta}^*$ and the partial derivatives of $v(\vec{\theta}^*)$ with respect to $\vec{\theta}^*$. A high/low estimate of a mortality improvement rate (expressed in terms of the change in log central death rates) can be calculated as the best estimate of the mortality improvement rate plus/minus a multiple (say three) times the square root of the corresponding variance; that is,

$$v(\hat{\vec{\theta}}^*) \pm 3\sqrt{\mathrm{Var}(v(\hat{\vec{\theta}}^*))}.$$

Using the method described above, we obtain high/low estimates of future mortality improvement rates for Canadian males under the heat wave model. In figure 9 we show the projected paths of central death rates at various ages that are derived using the high/low estimates of the mortality improvement rates. The result presented in figure 9 gives an idea as to how high/low death rates may turn out to be in the future, and may also aid in setting margins for adverse deviation.

Figure 9: Age-specific central rates of death (in log scale), 2012 and onwards, projected using the central, high, and low estimates of mortality improvement rates implied by the heat wave model.



7 Conclusion

In this paper, we introduce the heat wave model for modelling and projecting mortality. This new model is built on a unique view that overall mortality improvement is composed of 'background improvements' and 'heat waves'. The former is captured by a simple Lee–Carter structure, whereas the latter is modeled by a parametric function that bears some similarity to a Fourier series.

The construction of the heat wave model fits very well into the framework of two-dimensional mortality improvement scales, a mortality projection framework that has been promulgated recently by several actuarial organizations including the CIA for use in actuarial valuation. The heat wave model produces scale factors that extend logically from the historical mortality improvement rates, converging smoothly from higher short-term values to lower long-term (ultimate) values. As demonstrated in Section 5, none of the existing stochastic mortality models (including the Lee–Carter model, the Cairns–Blake–Dowd model and its variants, and two versions of the Plat model) considered in this paper can produce scale factors with such desirable properties.

The heat wave model complements the current methods for deriving two-dimensional mortality improvement scales in two significant aspects. First, the heat wave model is entirely data-driven, requiring much fewer subjective judgments. Estimated from historical data, parameter σ indicates the length of the convergence period, and parameters b_x and d inform the long-term (ultimate) mortality improvement rate at each age. Second, rather than just a single best estimate, the heat wave model produces also a measure of uncertainty surrounding the best estimate. As illustrated in Section 6, with the heat wave model one can derive high/low scale factor estimates, which may be used for setting MfADs in practice.

From a statistical viewpoint, the heat wave model also represents a significant improvement over many of the existing stochastic mortality models. Compared to six alternative models that are widely used in the literature, the heat wave model provides a significantly better fit to the data set under consideration, even when the number of model parameters is taken into account. Admittedly, the heat wave model is more challenging to estimate due to its inequality parameter constraints, but we have overcome the estimation challenge using the barrier method.

Two limitations of the heat model are noted. First, compared to the Lee-Carter model, the estimation procedure for the heat wave model is admittedly more involved. Initial values have to be carefully chosen to expedite convergence, and barrier functions have to be specified to ensure the inequality contraints on some of the parameters. Second, given how the model is constructed, it is difficult to test its forecasting performance, particularly over a long forecast horizon.

In future research, the heat wave model may be improved in a few directions. First, it would be interesting to investigate if functions other than the normal density function may better capture the heat wave and produce a more remarkable goodness-of-fit. Second, because some data sets may contain more than one heat wave, it is warranted to extend the current version of the heat wave model to incorporate multiple heat waves and to develop a procedure to determine the number of heat waves that should be incorporated. Finally, as the barrier method is not the only

optimization method that can handle inequality constraints, it would be useful to explore alternative methods for maximizing the log-likelihood functions of the heat wave model and its future variants.

Acknowledgments

The authors gratefully acknowledge financial support from the 2016–17 CIA Academic Research Grant.

References

Adam, L. (2012). The Canadian Pensioners Mortality Table, mortality trends as at December 31, 2007. Available at www.cia-ica.ca/docs/default-source/2013/213012e.pdf.

Cairns, A.J.G. (2017). A Flexible and Robust Approach to Modelling Single Population Mortality. Paper presented at Longevity 13, Taipei, September 2017.

Cairns, A.J.G., Blake, D., and Dowd, K. (2006). A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration. *Journal of Risk and Insurance* 73: 687–718.

Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A., and Balevich, I. (2009). A Quantitative Comparison of Stochastic Mortality Models Using Data from England and Wales and the United States. *North American Actuarial Journal* 13: 1–35.

Canadian Institute of Actuaries (2014). Final Report on Canadian Pensioners' Mortality. Available at www.cia-ica.ca/docs/default-source/2014/214013e.pdf.

Canadian Institute of Actuaries (2017). Draft Report – Task Force Report on Mortality Improvement. Available at www.cia-ica.ca/docs/default-source/2017/217044e.pdf.

Continuous Mortality Investigation Bureau (2017a). CMI Mortality Projections Model: CMI_2016. Mortality Projections Committee Working Paper 97.

Continuous Mortality Investigation Bureau (2017b). CMI Mortality Projections Model: Methods. Mortality Projections Committee Working Paper 98.

Currie I.D., Durban, M., and Eilers, P.H.C. (2004). Smoothing and Forecasting Mortality Rates. *Statistical Modelling* 4: 279–298.

Lee, R.D., and Carter, L.R. (1992). Modelling and Forecasting US Mortality. *Journal of the American Statistical Association* 87(419): 659–671.

Nocedal, J., and Wright, S. (1999). *Numerical Optimization*. New York, NY: Springer.

Office of the Chief Actuary (2014). Mortality Projections for Social Security Programs in Canada. Actuarial Study No. 12. Available at www.osfi-bsif.gc.ca/eng/oca-bac/as-ea/Pages/mpsspc.aspx.

Plat, R. (2009). On Stochastic Mortality Modelling. *Insurance: Mathematics and Economics* 45: 393–404.

Schwarz, G.E. (1978). Estimating the Dimension of a Model. *Annals of Statistics* 6(2): 461–464.

Society of Actuaries (1995). 1994 Group Annuity Mortality Table and 1994 Group Annuity Reserving Table. *Transactions of the Society of Actuaries*: XLVII, 865–915.

Society of Actuaries (2012). Mortality Improvement Scale BB Report. Available at www.soa.org/files/research/exp-study/research-mortality-improve-bb-report.pdf.

Society of Actuaries (2014). Mortality Improvement Scale MP-2014 Report. Available at www.soa.org/Files/Research/Exp-Study/research-2014-mp-report.pdf.

Society of Actuaries (2015). Mortality Improvement Scale MP-2015. Available at www.soa.org/Files/Research/Exp-Study/research-2015-mp-report.pdf.

Society of Actuaries (2016). Mortality Improvement Scale MP-2015. Available at www.soa.org/Files/Research/Exp-Study/mortality-improvement-scale-mp-2016.pdf

Society of Actuaries (2017). Mortality Improvement Scale MP-2017. Available at www.soa.org/Files/Research/Exp-Study/mortality-improvement-scale-mp-2017.pdf.