ARC Centre of Excellence in Population Ageing Research

Working Paper 2021/06

A Group Regularisation Approach for Constructing Generalised Age-Period-Cohort Mortality Projection Models

Dilan SriDaran, Michael Sherris, Andrés M. Villegas and Jonathan Ziveyi

---

A Group Regularisation Approach for Constructing Generalised Age-Period-Cohort Mortality Projection Models

Dilan SriDaran[a], Michael Sherris[a], Andrés M. Villegas[a], Jonathan Ziveyi[a]

[a]*School of Risk and Actuarial Studies and ARC Centre of Excellence in Population Ageing Research (CEPAR), UNSW Sydney*

Abstract

Given the rapid reductions in human mortality observed over recent decades and the uncertainty associated with their future evolution, there have been a large number of mortality projection models proposed by actuaries and demographers in recent years. However, many of these suffer from being overly complex, thereby producing spurious forecasts, particularly over long horizons and for small, noisy datasets. In this paper, we exploit statistical learning tools, namely group regularisation and cross validation, to provide a robust framework to construct such discrete-time mortality models by automatically selecting the most appropriate functions to best describe and forecast particular datasets. Most importantly, this approach produces bespoke models using a trade-off between complexity (to draw as much insight as possible from limited datasets) and parsimony (to prevent overfitting to noise), with this trade-off designed to have specific regard to the forecasting horizon of interest. This is illustrated using both empirical data from the Human Mortality Database and simulated data, using code that has been made available within a user-friendly open-source `R` package **StMoMo**.

Keywords: Mortality projection, regularisation, cross validation, age-period-cohort model

1. Introduction

Humanity has made, and continues to make, significant progress in averting and delaying death, with life expectancy at birth rising by over five years globally since the turn of the millennium (World Health Organization 2018). Increasing longevity, whilst indicative of tremendous societal progress, does present economic and social risks to health care, social security, private pension systems, and life insurers, which has piqued the interest of actuaries and demographers. The main risk, however, is not increasing longevity per se, but rather the challenges tied to the inherent uncertainty of the future evolution of mortality rates. For example, if life expectancies are lower than anticipated, life insurers will incur large financial losses arising from earlier-than-expected insurance claims, while individuals may be unfairly overcharged for annuity contracts. Conversely, if future life expectancies are underestimated, which has occurred in recent decades (Blake, Cairns, and Dowd 2008), pension plans, annuity providers, and health care systems will be liable for larger-than-anticipated outward cash flows, and individuals may outlive their finite savings. Dynamic mortality modelling is therefore of critical importance, both to facilitate an understanding of the past and to provide a basis for future mortality projections.

Traditionally, amongst an actuarial readership, generalised age-period-cohort (GAPC) models have been a preferred approach (Villegas, Kaishev, and Millossovich 2018; Hunt and Blake 2020d). These models decompose the force of mortality for age x at time t, $\mu_{x,t}$, across the dimensions of age, x, period (or time), t, and cohort (or year-of-birth), $t-x$, via the following general specification:

$$\ln(\mu_{x,t}) = \alpha_x + \sum_{i=1}^{N} \beta^{(i)}(x)\kappa_t^{(i)} + \gamma_{t-x}. \tag{1}$$

GAPC models, which can be expressed as generalised (non-) linear models (Currie 2016), have proven particularly popular due to the ability to simulate future outcomes and develop probabilistic forecast intervals, as well as the

Email addresses: `dilan.sridaran@gmail.com` (Dilan SriDaran), `m.sherris@unsw.edu.au` (Michael Sherris), `a.villegas@unsw.edu.au` (Andrés M. Villegas), `j.ziveyi@unsw.edu.au` (Jonathan Ziveyi)

direct interpretability of parameters in terms of their demographic significance. For example, the general age shape component of mortality is captured by α_x, systematic longevity improvements across specific ages over time are reflected by $\beta_x^{(i)}$ and $\kappa_t^{(i)}$, $i = 1, \ldots, N$, and any effects embedded in an individual's year-of-birth are allowed for by γ_{t-x}. However, a major drawback to this family of models is the inherent subjectivity associated with the construction and selection of the most appropriate model within the family.

In theory, there are an infinite number of models that can be produced within the GAPC framework defined by Equation (1), corresponding to every possible combination of terms: α_x, N, $\beta_x^{(i)}$, and γ_{t-x}; where $\beta_x^{(i)}$ can be either non-parametric or parametric with various functional forms. In an ideal world, we would be able to consider every possible combination of terms to choose the optimal model for a given dataset and application. However, this is clearly not feasible due to time and computational constraints, and as a result, practitioners have been forced to rely on a small but growing selection of models, which are outlined in the Section 3. Citing these issues, Hunt and Blake (2014) propose an alternative "general procedure" to construct bespoke mortality models, which involves using "expert actuarial judgement" to manually select appropriate age-modulating terms to capture the historical mortality patterns identified within specific datasets. However, whilst data-driven and potentially powerful, the procedure lacks a formalised quantitative framework for implementation, thereby introducing room for subjectivity and inconsistency. Moreover, the overreliance on the historical goodness-of-fit to select the model specification does not necessarily translate into improved out-of-sample projection accuracy.

The main objective of this paper is to propose an alternative mortality model construction procedure – using regularisation and cross validation techniques – that mirrors the data-driven framework proposed by Hunt and Blake (2014), albeit through a different lens. Whilst Hunt and Blake (2014) start with a null model and iteratively add age-period terms, our framework begins with a full GAPC model containing a large suite of terms. This is then pared down to the final model using penalised fitting techniques, namely group regularisation (Yuan and Lin 2006; Breheny and Huang 2015), to automatically identify the number and type of terms to best describe mortality dynamics for specific datasets. In addition, to select the appropriate level of penalisation for a particular application, we introduce a block cross validation approach wherein the data is iteratively separated into that which is used for fitting and that which is used to assess predictions. By focusing on the predictive power of the models rather than on their historical goodness-of-fit, the cross validation approach provides a better assessment of their usefulness for mortality projection tasks.

To the best of our knowledge, this is the first time that group regularisation techniques are used to estimate GAPC mortality models. However, the use of penalised fitting procedures is not uncommon in the mortality literature. One classical strand of literature uses penalised fitting techniques to impose smoothness on the parameters of mortality models. For example, Delwarde, Denuit, and Eilers (2007) use penalised least-squares and penalised maximum likelihood to smooth the parameters in the Lee and Carter (1992) model and Poisson log-bilinear model (Brouhns, Denuit, and Vermunt 2002) for mortality projections, while Li et al. (2020) adopt similar techniques to smooth and fit other GAPC models. Another more modern stream of the literature closely related to our study uses regularisation fitting techniques borrowed from statistical learning to perform a data-driven selection of the variables present in a mortality model. Venter and Şahin (2018) fit age-period-cohort models with linear and cubic splines using the lasso (Tibshirani 1996) to avoid over-parameterisation and reduce predictive variance compared to maximum likelihood. To forecast mortality rates, Guibert, Lopez, and Piette (2019) and Li and Shi (2021) propose the use of a large Vector Autoregressive (VAR) model which they estimate using elastic-net penalisation (Zou and Hastie 2005) in order to impose sparsity on the model parameters. Hainaut and Denuit (2020) propose to represent the two dimensional mortality surface using a large wavelet-based decomposition and apply lasso penalisation to select the optimal wavelets. Similarly, Barigou, Loisel, and Salhi (2021) decompose the two dimensional mortality surface using a large polynomial basis expansion and, similar to us, use a combination of regularisation (elastic-net penalisation) and cross validation to obtain a parsimonious model better suited for mortality predictions.

One key feature of the work of Guibert, Lopez, and Piette (2019), Li and Shi (2021), Hainaut and Denuit (2020) and Barigou, Loisel, and Salhi (2021) is that the final structure of the mortality model is data-driven which translates into good forecasting performance relative to classical GAPC models. However, one main drawback of the data-driven approaches proposed to date is that the final models tend to be difficult to interpret, lacking the direct interpretability of the model parameters in terms of their demographic significance, which is arguably one of the features that has made GAPC models very popular among practitioners. In contrast, our research uses group regularisation to fit GAPC models directly and automatically select the most appropriate basis functions to capture mortality patterns, thereby automating the procedure introduced by Hunt and Blake (2014). By remaining within the familiar context of GAPC models, our approach retains the interpretability of the final model parameters in

terms of age, period and cohort effects, while still benefiting from the improved forecasting performance of data-driven approaches. Moreover, the final models produced by our data-driven model construction approach are easily integrable with existing tools for analysing GAPC models. In particular, our proposed model construction approach is readily available through open source code as an extension of the `R` package **StMoMo** (Villegas, Kaishev, and Millossovich 2018).

The growing array of models and construction procedures introduces another important element, namely model selection. This is especially pertinent when models provide conflicting forecasts using the same data, which is assumed to be comprised of a structural signal and an idiosyncratic noise. The objective of any predictive model is to capture all of the signal and none of the noise, to provide the best possible forecasts on independent data generated from the same distribution. Overly parsimonious models may not capture all of the signal, whilst over-parameterised models may capture some of the noise to produce volatile forecasts, particularly over longer horizons. Therefore, at its core, model selection relates to the ubiquitous trade-off between parsimony and goodness-of-fit, wherein the ability to explain in-sample data must be tempered to the extent that the fit is achieved using a highly complex model (Vandekerckhove, Matzke, and Wagenmakers 2015).

In practice, however, there is limited research on how to implement this trade-off in the context of mortality forecasting. Instead, irrespective of context, there remains a heavy reliance on goodness-of-fit measures in the existing literature to determine the "best" model for each data set (see Atance and Debón (2020) and Barigou, Loisel, and Salhi (2021) for a discussion). However, intuitively it may be expected that there is indeed no such thing as a single "best" mortality model, with there perhaps being optimal models for different applications. These applications include explanation and forecasting, where forecasting can be further defined by varying durations of forecasting horizons. Thus, whilst goodness-of-fit measures may be appropriate for selecting models for explanation, these chosen models may not generalise well to produce accurate forecasts. Accordingly, to select the optimal level of model complexity in our model construction approach, we introduce a cross validation framework which can be tailored to reflect different forecasting horizons. By focusing on the predictive power of the models, one can provide a better assessment of their usefulness for actuarial and financial applications such as pricing and reserving, where forecasting is arguably more pertinent than explanation. This will also allow us to provide an understanding of the key features of mortality models that are desired for different contexts. For example, in our extensive empirical analyse we are able to confirm the commonly-held heuristic that simpler more parsimonious models are preferred for noisier data and for longer forecasting horizons (Green and Armstrong 2015).

With this, our research extends the mortality modelling literature by producing a formalised framework that can be implemented by practitioners to: i) construct bespoke models to describe mortality dynamics in specific data sets; and ii) select appropriate models for different forecasting horizons.

The remainder of this article is organised as follows. In Section 2 we define the main notation used to refer to mortality data, while in Section 3 we formally introduce the GAPC family of models and provide a review of the main mortality models that it encompasses. In Section 4 we describe the regularisation and cross validation frameworks we propose for the construction of mortality projection models, which is subsequently illustrated using both empirical and simulated data. Specifically, in Section 5 we present an example of the application of the model construction procedure to data for USA male population. This is extended in Section 6 where an exhaustive empirical analysis is carried out on the predictive performance of our approach using data from 52 populations from the Human Mortality Database (2020). This is complemented in Section 7 with an examination of the performance of our approach using various simulated datasets. Finally, we conclude in Section 8 with a discussion of the key highlights of the paper and areas for further research.

2. Notation and data

Let calendar year t run from exact time t to $t+1$. The random variable, $D_{x,t}$, denotes the number of deaths aged x last birthday during calendar year t. The actual observed number of deaths is denoted by $d_{x,t}$, and the corresponding central and initial exposure at risk by $E^c_{x,t}$ and $E^0_{x,t}$, respectively. From these, one-year death probabilities can be estimated as $\hat{q}_{x,t} = d_{x,t}/E^0_{x,t}$. Similarly, central death rates can be estimated as $\hat{m}_{x,t} = d_{x,t}/E^c_{x,t}$. Under the strict assumption of a constant force of mortality between both integer ages and calendar years, the force of mortality, $\mu_{x,t}$, is equivalent to the central death rate, $m_{x,t}$.

The death and central exposure data are assumed to be arranged in matrices $\boldsymbol{D} = (d_{x,t})$ and $\boldsymbol{E^c} = (E^c_{x,t})$, which are each $n_a \times n_y$, so that we have n_a ages, n_y years, and $n_b = n_a + n_y - 1$ cohorts. Let $n = n_a n_y$ be the total

number of observations. From these, the matrix $\boldsymbol{\mu} = (d_{x,t}/E^c_{x,t})$ can be derived.

3. Generalised age-period-cohort models

Mortality modelling and forecasting techniques can be broadly classified as being expectations-based, explanatory, or extrapolative in nature, with an overarching literature summary presented in Booth and Tickle (2008). Amongst an actuarial readership, extrapolation has been favoured as the primary basis for mortality forecasting. Extrapolation inherently assumes that future mortality rates will represent a continuation of the past, and thus can be modelled via stochastic processes and univariate time series models (Box, Jenkins, and Reinsel 1994). This assumption is generally regarded as reasonable due to "historical regularities," although mortality jumps resulting from pandemics and medical discoveries do challenge this notion (Booth and Tickle 2008). However, despite these challenges, extrapolation remains popular within the context of mortality forecasting, with one obvious advantage being the ability to produce a prediction interval for the forecast distribution rather than a simple deterministic point estimate.

Within the class of extrapolative techniques, the most popular models decompose mortality rates into a combination of age, period (or time), and cohort (or year-of-birth) effects. This decomposition allows the evolution of mortality rates to incorporate the effects of ageing, the effects of medical, nutritional, and social improvements over time, and the lifelong effects carried from an individual's year-of-birth, arising from factors such as the implementation of nationwide healthcare and changing attitudes towards smoking (Li, Zhou, and Hardy 2015).

The literature on stochastic mortality models that exploit this age-period-cohort decomposition is vast. However, many authors have identified a number of similarities between the stochastic mortality models, with Hunt and Blake (2020d) defining them within a broad family known as Generalised Age-Period-Cohort (GAPC) models, which can be fitted using the `R` package **StMoMo** (Villegas, Kaishev, and Millossovich 2018). This draws upon the work of Currie (2016) that expresses mortality models within the standard terminology of generalised linear and non-linear models. By analogy, under the GAPC framework, a mortality model is comprised of four elements:

1. The *random component*: the numbers of deaths, $D_{x,t}$, is random and is drawn from either a Poisson or a Binomial distribution

$$D_{x,t} \sim \text{Poisson}(E^c_{x,t}\mu_{x,t}) \qquad \text{or} \qquad D_{x,t} \sim \text{Binomial}(E^0_{x,t}, q_{x,t}). \tag{2}$$

2. The *systematic component*: the predictor structure uses a linear or bilinear summation of a static age x effect α_x, N random period factors $\kappa^{(i)}_t$, $i = 1, \ldots, N$, and a random cohort factor γ_{t-x}, so that

$$\eta_{x,t} = \alpha_x + \sum_{i=1}^{N} \beta^{(i)}_x \kappa^{(i)}_t + \beta^{(0)}_x \gamma_{t-x}, \tag{3}$$

   where the age-modulating terms or factor loadings, $\beta^{(i)}_x$, can be either pre-specified functions of age, $\beta^{(i)}_x \equiv f^{(i)}(x)$, or non-parametric terms that need to be estimated.
3. The *link function*: the random and systematic components are linked using the canonical link function. For the Poisson distribution, the log link is used

$$\ln(\mu_{x,t}) = \eta_{x,t}, \tag{4}$$

   and for the Binomial distribution, the logit link is adopted

$$\text{logit}(q_{x,t}) = \eta_{x,t}. \tag{5}$$

4. The *parameter constraints*: in many situations the estimated parameters can be transformed using arbitrary constants without changing $\eta_{x,t}$. Thus, to ensure identifiability of the models, various constraints may need to be imposed.

Many of the models in the actuarial literature can be defined within the GAPC framework in Equation (3), where the differences are derived entirely from the inclusion or exclusion of static age, α_x, and cohort, γ_{t-x}, terms, the number of age-period factors, N, and the age-modulating functions being parametric, $f^{(i)}(x)$, or non-parametric, $\beta^{(i)}_x$. This is summarised in Table 1, which indicates the main features of ten common GAPC models. We review these models in what follows.

Table 1: Overview of different features of existing GAPC models.

Model	α_x	N	$f^{(i)}(x)$	$\beta_x^{(i)}$	γ_{t-x}	Number of parameters
LC	✓	1		✓		$2n_a + n_y$
LC2	✓	2		✓		$3n_a + 2n_y$
APC	✓	1	✓		✓	$n_a + n_y + n_b$
RH	✓	1		✓	✓	$2n_a + n_y + n_b$
CBD		2	✓			$2n_y$
M7		3	✓		✓	$3n_y + n_b$
sPLAT	✓	2	✓		✓	$n_a + 2n_y + n_b$
cPLAT	✓	3	✓		✓	$n_a + 3n_y + n_b$
CBDx	✓	2	✓			$n_a + 2n_y$
M7x	✓	3	✓		✓	$n_a + 3n_y + n_b$

GAPC models rose to prominence via the revolutionary work of Lee and Carter (1992), which remains the most widely used projection models. The Lee-Carter (LC) model parsimoniously deconstructs the force of mortality into a bilinear summation of only age and period terms

$$\ln(\mu_{x,t}) = \alpha_x + \beta_x^{(1)}\kappa_t^{(1)}. \tag{6}$$

Crucially, the age parameters are explicitly assumed to be time-invariant, and thus forecasts are produced by simply extrapolating the period factor, $\kappa_t^{(1)}$, using autoregressive integrated moving average (ARIMA) models of the form

$$\Delta^d \kappa_t^{(1)} = \delta_0 + \sum_{k=1}^{p} \phi_k \Delta^d \kappa_{t-k}^{(1)} + \xi_t + \sum_{j=1}^{q} \delta_j \xi_{t-j}, \tag{7}$$

where Δ is the difference operator, ϕ_k are the autoregressive coefficients, δ_j are the moving average coefficients, and ξ_t is a Gaussian white noise process with variance σ_ξ^2. The random walk with drift is a common choice, given by

$$\kappa_t^{(1)} = \delta_0 + \kappa_{t-1}^{(1)} + \xi_t. \tag{8}$$

The LC model can also be extended to larger numbers, N, of age-period factors (Hunt and Blake 2020c), such as the $N = 2$ model by Renshaw and Haberman (2003), herein referred to as the Lee-Carter 2 (LC2) model

$$\ln(\mu_{x,t}) = \alpha_x + \beta_x^{(1)}\kappa_t^{(1)} + \beta_x^{(2)}\kappa_t^{(2)}. \tag{9}$$

Renshaw and Haberman (2006) further observe distinct ripple patterns when LC deviance residuals are plotted against cohorts, and thus propose the addition of a cohort term, γ_{t-x}, to produce the Renshaw-Haberman (RH) model

$$\ln(\mu_{x,t}) = \alpha_x + \beta_x^{(1)}\kappa_t^{(1)} + \beta_x^{(0)}\gamma_{t-x}. \tag{10}$$

However, this structure is prone to fitting instability (Hunt and Villegas 2015), with Haberman and Renshaw (2011) proposing setting $\beta_x^{(0)} = 1$ to remedy the issue. This simplified model can be regarded as a variant of the age-period-cohort (APC) model

$$\ln(\mu_{x,t}) = \alpha_x + \kappa_t^{(1)} + \gamma_{t-x}, \tag{11}$$

which was introduced to the actuarial readership by Currie (2006), despite having a longer history in other fields (Holford 1983). Importantly, in these models, the cohort term is also assumed to follow a univariate ARIMA model

$$\Delta^d \gamma_c = \delta_0 + \sum_{k=1}^{p} \phi_k \Delta^d \gamma_{c-k} + \xi_c + \sum_{j=1}^{q} \delta_j \xi_{c-j}. \tag{12}$$

Cairns, Blake, and Dowd (2006) introduce one of the most popular competitors to the LC model, known as the CBD model. This assumes that the logit of the one-year probability of death, $q_{x,t}$, is a linear function of age

$$\text{logit}(q_{x,t}) = \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)}. \tag{13}$$

This can also be viewed as a variant of the LC model, with $N = 2$ age-period terms and pre-specified age-modulating parameters $\beta_x^{(1)} \equiv f^{(1)}(x) = 1$ and $\beta_x^{(2)} \equiv f^{(2)}(x) = x - \bar{x}$; where $\bar{x}$ is the average of the age range. The lack of a static age term, α_x, does provide relative parsimony, however this tends to restrict its usage to modelling higher ages, which tend to be of greater interest to pension schemes. However, even across older ages, mortality rates are not strictly linear in age, and thus, Cairns et al. (2010) propose three further extensions that incorporate cohort effects and an age-curvature term

$$\begin{aligned}
\text{logit}(q_{x,t}) &= \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)} + \gamma_{t-x}, \\
\text{logit}(q_{x,t}) &= \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)} + ((x - \bar{x})^2 - \sigma_x^2)\kappa_t^{(3)} + \gamma_{t-x}, \\
\text{logit}(q_{x,t}) &= \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)} + (x - x_c)\gamma_{t-x},
\end{aligned} \tag{14}$$

where σ_x^2 is the variance of the age range used in the dataset, and x_c is a constant parameter to be estimated. These are commonly referred to as Models M6, M7, and M8 respectively, with M7 typically being favoured by practitioners.

A particularly promising model, herein referred to as the cPLAT, is proposed in Plat (2009). This model includes the static age term from the LC model that is suitable for full age ranges, the cohort effect of the RH model that adheres well to historical data, and the multiple age-period factors provided by the CBD model,

$$\ln(\mu_{x,t}) = \alpha_x + \kappa_t^{(1)} + (\bar{x} - x)\kappa_t^{(2)} + (\bar{x} - x)^+\kappa_t^{(3)} + \gamma_{t-x}. \tag{15}$$

In this model, Plat (2009) includes a put payoff, $f^{(3)}(x) = (\bar{x} - x)^+ = \max\{0, \bar{x} - x\}$, which means that $\kappa_t^{(3)}$ will have no effect on ages greater than $\bar{x}$. This is to reflect the notion that lower ages tend to behave differently from higher ages due to factors such as traffic accidents and other external causes of death. Therefore, when modelling only higher ages, Plat (2009) simplifies the structure (sPLAT) to

$$\ln(\mu_{x,t}) = \alpha_x + \kappa_t^{(1)} + (\bar{x} - x)\kappa_t^{(2)} + \gamma_{t-x}, \tag{16}$$

which is essentially the M6 model with a static age effect, α_x. Similarly, Hunt and Blake (2020c) and Dowd, Cairns, and Blake (2020) recently consider extending the CBD and M7 models by adding an age effect, α_x. Specifically, Hunt and Blake (2020c) consider the CBDx model

$$\ln(\mu_{x,t}) = \alpha_x + \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)}, \tag{17}$$

and Dowd, Cairns, and Blake (2020) the M7x model

$$\ln(\mu_{x,t}) = \alpha_x + \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)} + ((x - \bar{x})^2 - \sigma_x^2)\kappa_t^{(3)} + \gamma_{t-x}. \tag{18}$$

3.1. Forecasting mortality with GAPC models

In the general setting of GAPC mortality models, the time evolution of mortality is determined by the period effects $\kappa_t^{(i)}$, $i = 1, \ldots, N$, and the cohort effect γ_{t-x}. To forecast mortality rates, the standard approach (Villegas, Kaishev, and Millossovich 2018) is to assume that the period indices follow a multivariate random walk so that

$$\boldsymbol{\kappa}_t = \boldsymbol{\delta} + \boldsymbol{\kappa}_{t-1} + \boldsymbol{\xi}_t^{\kappa}, \qquad \boldsymbol{\kappa}_t = \begin{pmatrix} \kappa_t^{(1)} \\ \vdots \\ \kappa_t^{(N)} \end{pmatrix}, \qquad \boldsymbol{\xi}_t^{\kappa} \sim N(\mathbf{0}, \Sigma), \tag{19}$$

where $\boldsymbol{\delta} = \{\delta^{(1)}, \ldots, \delta^{(N)}\}'$ is an N-dimension column vector of drift parameters and Σ is the $N \times N$ variance-covariance matrix of the multivariate white noise $\boldsymbol{\xi}_t^{\kappa}$. For the cohort effect, the standard is to assume that they follow a univariate ARIMA model as described before in Equation (12).

Let t_{n_y} denote the last year where mortality data is available. To obtain mortality projections for forecast horizon h, one first obtains period index forecasts $\hat{\boldsymbol{\kappa}}_{t_{n_y}+h}$ and, if necessary, cohort index forecast $\hat{\gamma}_{t_y-x+h}$ using the respective time series model. The mortality projections at horizon h are then given by

$$\widehat{\ln(\mu_{x,t_{n_y}+h})} = \hat{\alpha}_x + \sum_{i=1}^{N} \hat{\beta}_x^{(i)} \hat{\kappa}_{t_{n_y}+h}^{(i)} + \hat{\beta}_x^{(0)} \hat{\gamma}_{t_{n_y}-x+h} \tag{20}$$

or

$$\widehat{\text{logit}(q_{x,t_{n_y}+h})} = \hat{\alpha}_x + \sum_{i=1}^{N} \hat{\beta}_x^{(i)} \hat{\kappa}_{t_{n_y}+h}^{(i)} + \hat{\beta}_x^{(0)} \hat{\gamma}_{t_{n_y}-x+h}. \tag{21}$$

4. Group regularisation for model construction

One limitation of the mortality modelling literature relates to the limited number of models that are considered when selecting the most appropriate model for a particular dataset. In theory, there are an infinite number of models that can be produced within the GAPC framework, corresponding to every possible combination of terms: α_x, N, $\beta_x^{(i)}$, $f^{(i)}(x)$, and γ_{t-x}; where $f^{(i)}(x)$ can take on various functional forms and parameterisations. Ideally, we would be able to consider every possible combination of terms, however this is currently not feasible due to computational constraints, so we rely on a small subset of models. Importantly, it is perhaps unreasonable to expect that all datasets can be entirely explained by one of the ten models summarised in Table 1. For example, a better fit for a particular dataset may be obtained by adjusting the "strike" price of the put option payoff in the cPLAT model to some $k \in \mathbb{Z}$, or through an alternative functional form entirely.

To overcome these issues, we exploit regularisation techniques to propose a framework to construct bespoke GAPC mortality models for specific datasets by selecting appropriate age-modulating functions, $f^{(i)}(x)$, from a large suite of basis functions. This resembles the data-driven framework proposed in Hunt and Blake (2014), however our approach is underpinned by regularisation and resampling approaches, which automatically determine the optimal degree of complexity for different applications by assessing how well a statistical model will generalise to unseen out-of-sample data (Hastie, Tibshirani, and Friedman 2001). With this, our approach makes more efficient use of data to draw statistical inferences and make predictions and places a stronger emphasis on out-of-sample performance for model selection.

4.1. Model framework

Whilst in the GAPC framework the number of deaths is assumed to be drawn from either a Poisson or Binomial distribution, we model the dependent variable, $\ln(D_{x,t}/E^c_{x,t})$, as a Gaussian generalised linear model. This decision is made for computational efficiency, noting that closed-form solutions for group regularisation estimates exist when using ordinary-least squares (Breheny and Huang 2015). Therefore, our general base model consists of a modified version of the GAPC framework described in Equations (2)-(5). Specifically, we consider the following three elements:

1. The *random component*: the logarithm of the force of mortality follows a Normal distribution so that

$$\ln(D_{x,t}/E^c_{x,t}) \sim \text{Normal}(\ln(\mu_{x,t}), \sigma^2). \tag{22}$$

2. The *link function*: the identity link (canonical) is used, such that

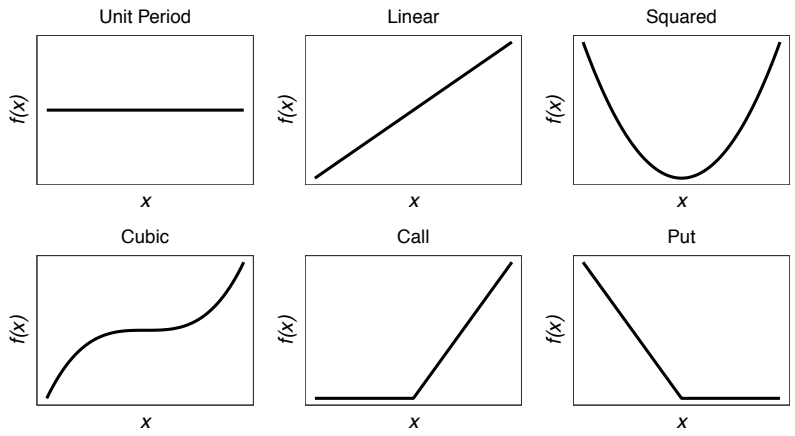
$$\eta_{x,t} = g(E[\ln(D_{x,t}/E^c_{x,t})]) = \ln(\mu_{x,t}). \tag{23}$$

3. The *systematic component*: the predictor captures age, period, and cohort effects, using

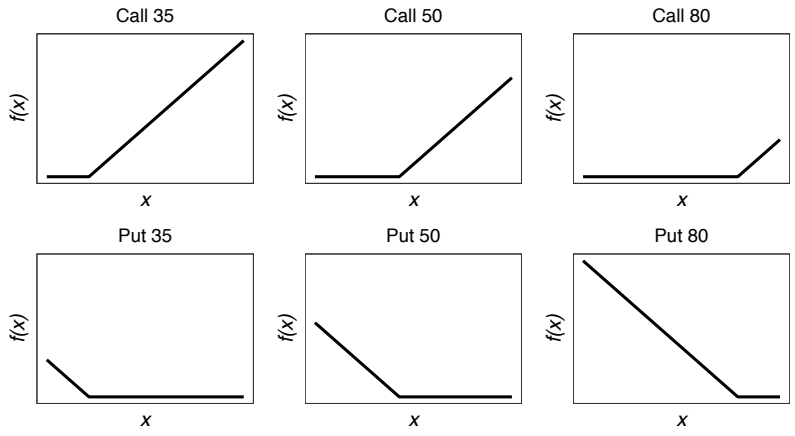
$$\eta_{x,t} = \alpha_x + \sum_{i=1}^{B} f^{(i)}(x)\kappa_t^{(i)} + \gamma_{t-x}, \tag{24}$$

   where the number of age-period terms, B, is very large. Here,

   - $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n_a})$ is a row vector of coefficients to capture patterns by age;
   - $\boldsymbol{\kappa^{(i)}} = (\kappa_1^{(i)}, \ldots, \kappa_{n_y}^{(i)})$ are row vectors of coefficients to capture systematic improvements over time;
   - $f^{(i)}(x)$ are pre-defined functions to modulate trends by age; and
   - $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{n_b})$ is a row vector of coefficients to capture fixed effects embedded in years-of-birth.



(a)



(b)

Figure 1: Examples of age-modulating basis functions, $f^{(i)}(x)$, included in initial group regularisation model specification (a) by type, and (b) by parameterisation (for call and put option payoffs).

Crucially, we include a large number, B, of age-period terms in the initial model specification, which allows the model to select from a wide range of potential trends.[1] This is done through the specification of the B, $f^{(i)}(x)$, factor loadings, which modulate how the random period factors, $\kappa_t^{(i)}$, impact different ages to determine how the age-period mortality curve changes through time. As a starting point, we consider five different continuous functional forms for $f^{(i)}(x)$, which are summarised in Table 2, and illustrated in Figure 1a. Within each functional form, we also consider various parameterisations illustrated in Figure 1b.

Table 2: Basis functions, $f^{(i)}(x)$, included in initial group regularisation model specification.

Name	Form
Unit Period	1
Linear	$x - \bar{x}$
Polynomial	$(x - \bar{x})^j$
Call Payoff	$(x - k)^+$
Put Payoff	$(k - x)^+$

[1]For example, in some of our empirical applications in Sections 5 and 6 we consider B to be as big as 67.

The unit period term determines the general level of mortality and is included to capture systematic improvements over time that affect all ages equally, arising from increasing social advances (Cairns, Blake, and Dowd 2006). The linear or slope term, however, could capture the recent trend towards an increasing concentration of deaths around the mean age of death, known as "rectangularisation" (Biffis 2005). Further, the inclusion of higher order polynomials is inspired by the curvature in $\text{logit}(q_{x,t})$ plots, mirroring various extensions to the CBD model (Cairns et al. 2010; Li and O'Hare 2017). Similarly, the call and put option payoffs are motivated by the cPLAT model (Plat 2009), however, we consider a greater range of strikes. We note that the framework and its associated implementation are flexible enough to allow the consideration of other functional forms beyond those in Table 2. For example, we could also consider the parametric forms in the "Age Function" toolkit of Hunt and Blake (2014), orthogonal piecewise linear splines as in Aro and Pennanen (2011), Wavelets as in Hainaut and Denuit (2020), or B-Splines (Currie, Durban, and Eilers 2004). We explore the use of this latter type of basis function in our empirical exercise in Sections 5 and 6.

4.2. Model identifiability

As with most GAPC models, parameter identifiability might be a concern, especially given the diversity of basis functions allowed within the initial model (see Hunt and Blake (2020c), Hunt and Blake (2020b) and Currie (2020) for a detailed discussion on this matter). For example, we can freely move the general level of mortality across the age, period and cohort parameters through the transformations

$$\left(\alpha_x, \kappa_t^{(i)}\right) \rightarrow \left(\alpha_x + b_i f^{(i)}(x), \kappa_t^{(1)} - b_i\right), \quad i = 1, \ldots, B, \tag{25}$$

and

$$(\alpha_x, \gamma_{t-x}) \rightarrow (\alpha_x + b, \gamma_{t-x} - b), \tag{26}$$

for b and b_i, $i = 1, \ldots, B$, arbitrary constants. Thus, we set $\kappa_1^{(i)} = 0$, $i = 1, \ldots, B$, so that α_x can be interpreted as the age shape of mortality in the first year of the data. Similarly, we set $\gamma_1 = 0$.

Besides the shifting of mortality levels, Hunt and Blake (2020b) show that if a GAPC model contains basis functions spanning polynomials up to order $M - 1$, then the model possesses invariant transformations which allow the arbitrary movement of polynomial trends of order M across the age, period and cohort parameters. However, we do not impose any explicit parameter constrains to deal with such invariant transformations, noting that we mitigate the possible impact of this identifiability issue on projections by modelling both the period and cohort index using time series models with an order of integration $d = 1$. As discussed in Currie (2020), such a choice of the order of integration implies that in most situations, the forecast mortality rate values would be invariant despite the possible non-identifiability of the model parameters.

4.3. Generalised linear model representation

Leveraging the result from Currie (2016), we can express the systematic component as a generalised linear model. Let $\boldsymbol{\eta} = \text{vec}(\ln(\boldsymbol{\mu}))$; here, the vec operator stacks the columns of a matrix in column order on top of each other so $\boldsymbol{\eta}$ is column vector of dimensions $n \times 1$ with $n = n_a n_y$. We can then write Equation (24) in matrix form

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} = \sum_{j=0}^{B+1} \boldsymbol{X}_j \boldsymbol{\beta}_j, \tag{27}$$

where the coefficients vector, $\boldsymbol{\beta}$, can be decomposed into $B + 2$ non-overlapping groups

$$\boldsymbol{\beta} = (\boldsymbol{\beta_0}', \ \boldsymbol{\beta_1}', \ \boldsymbol{\beta_2}', \ \ldots \ , \ \boldsymbol{\beta_B}', \ \boldsymbol{\beta_{B+1}}')' = (\boldsymbol{\alpha}, \ \boldsymbol{\kappa^{(1)}}, \ \boldsymbol{\kappa^{(2)}}, \ \ldots \ , \ \boldsymbol{\kappa^{(B)}}, \ \boldsymbol{\gamma})'. \tag{28}$$

The first group refers to the vector of static age coefficients, $\boldsymbol{\alpha}$. The next B groups denote each set period of period coefficients, $\boldsymbol{\kappa^{(i)}}$, $i = 1, \ldots, B$, while the final group denotes the vector of cohort coefficients, $\boldsymbol{\gamma}$. The design matrix, $\boldsymbol{X}$, can also be divided

$$\boldsymbol{X} = [\boldsymbol{X_0} : \boldsymbol{X_1} : \boldsymbol{X_2} : \ldots : \boldsymbol{X_B} : \boldsymbol{X_{B+1}}], \tag{29}$$

where $\boldsymbol{X}_j$ refers to the portion of the design matrix related to the predictors in Group j. For example, the first portion, $\boldsymbol{X_0}$ $(n \times n_a)$, defines the design matrix corresponding to $\boldsymbol{\alpha}$. Similarly, $\boldsymbol{X_i}$ $(n \times n_y)$ defines each $\boldsymbol{\kappa^{(i)}}$ for $i = 1, 2, ..., B$, and $\boldsymbol{X_{B+1}}$ $(n \times n_b)$ defines $\boldsymbol{\gamma}$. The Kronecker product is a compact way of writing these matrices (Currie 2016):

> "*If* $\boldsymbol{A}$*,* $m_1 \times n_1$*, and* $\boldsymbol{B}$*,* $m_2 \times n_2$*, are matrices, then their Kronecker product, written* $\boldsymbol{A} \otimes \boldsymbol{B}$*, is formed by replacing the element* $a_{i,j}$ *of* $\boldsymbol{A}$ *with the block* $a_{i,j}\boldsymbol{B}$ *for every element of* $\boldsymbol{A}$*. The Kronecker product is therefore* $m_1 m_2 \times n_1 n_2$*.*"

Using this definition, we can write the static age effects as

$$\boldsymbol{X_0} = \boldsymbol{I_{n_y}} \otimes \boldsymbol{1_{n_a}}, \tag{30}$$

and each period effect can be written as

$$\boldsymbol{X_i} = \boldsymbol{1_{n_y}} \otimes \boldsymbol{x_i}, \tag{31}$$

where $\boldsymbol{1_J}$ is a vector of 1's of length J, $\boldsymbol{I_J}$ is the identity matrix of dimension J, and $\boldsymbol{x_i} = (f^{(i)}(x_1), \ldots, f^{(i)}(x_{n_a}))'$. There is no compact notation for $\boldsymbol{X_{B+1}}$, but it can be built noting that row i contains 0s except for a single 1 which occurs in column c if the data point belongs to cohort c (Currie 2016).

4.4. Group regularisation fitting

Typically, to estimate parameters in GAPC-style models, $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\alpha}},\ \hat{\boldsymbol{\kappa}}^{(1)},\ \hat{\boldsymbol{\kappa}}^{(2)},\ \ldots\ ,\ \hat{\boldsymbol{\kappa}}^{(B)},\ \hat{\boldsymbol{\gamma}})'$, we maximise the log-likelihood or minimise the residual sum of squares. These will estimate all the coefficients in the model, noting that all values will likely be nonzero. Whilst suitable for existing pre-defined GAPC models, these approaches are not appropriate for our framework. If every estimated coefficient from our model were nonzero, we would have a highly over-parameterised and complex model with $n_a + B \times n_y + n_b$ parameters. This would be problematic due to interpretability and predictive accuracy, commonly referred to as the "$p > n$" problem (James et al. 2014).

Instead, we estimate parameters using group regularisation (Yuan and Lin 2006; Breheny and Huang 2015), which minimises the objective function

$$Q(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{\eta}) = L(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{\eta}) + \sum_{j=1}^{B+1} p_\lambda(\|\boldsymbol{\beta_j}\|), \tag{32}$$

where $L(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{\eta})$ denotes the residual sum of squares, and $p_\lambda(\cdot)$ is a penalty function applied to the Euclidean (ℓ_2) norm of each group to prevent the model from becoming too complex.

The key result is that the penalty function carries out group selection, meaning that within a group, coefficients will either all be zero or all be nonzero:

$$\textbf{If Group } j \textbf{ is selected, then } \hat{\boldsymbol{\beta}}_{\boldsymbol{jk}} \neq \boldsymbol{0} \textbf{ for all } \boldsymbol{k}\textbf{, and otherwise } \hat{\boldsymbol{\beta}}_{\boldsymbol{jk}} = \boldsymbol{0} \textbf{ for all } \boldsymbol{k}.$$

For example, if Group 1 is not selected, $\hat{\kappa}_t^{(1)}$ will be zero for all t, and the age-period pair $f^{(1)}(x)\kappa_t^{(1)}$ will be eliminated from the model. Therefore, this estimation approach will select basis functions from the initial large model to best capture the period and cohort patterns in each dataset.

There are alternative ways of defining the penalty function, $p_\lambda(\cdot)$, including the group lasso penalty (Yuan and Lin 2006), the smoothly clipped absolute deviation penalty (Fan and Li 2001), and the minimax concave penalty (Zhang 2010). We have decided to use the minimax concave penalty (MCP),

$$p_{\lambda,\gamma}(\theta) = \begin{cases} \lambda\theta - \frac{\theta^2}{2\nu}, & \text{if } \theta \leq \nu\lambda, \\ \frac{1}{2}\nu\lambda^2, & \text{if } \theta > \nu\lambda, \end{cases} \tag{33}$$

which is a function of the tuning parameter, ν, and the penalty term, λ. The MCP not only selects which groups of basis function to include, but also shrinks estimated coefficients to prevent the model from over-fitting to noise, and thus each unique (ν, λ) defines a unique fitted model. However, for ease of computation, we fix the tuning parameter to be $\nu = 3$, which aligns with common practice (Breheny and Huang 2015). Thus, λ solely controls the trade-off between goodness-of-fit and parsimony; the larger the value of λ, the heavier the penalisation, resulting in a more parsimonious model (fewer included basis functions and smaller estimated coefficients).

It is worth noting that the objective function in Equation (32) depends on the original scale of the design matrix $\boldsymbol{X}$, which in turn depends on the original scale of the basis functions $f^{(i)}(\cdot)$, $i = 1, \ldots, B$. To abstract from the original scale of the basis functions, the group-specific design matrices $\boldsymbol{X_j}$, $j = 1, \ldots, B+1$, are orthonormalised prior to fitting and then transformed back to the original scale after fitting the model (Breheny and Huang 2015).

4.5. Block cross validation for λ selection

The value of λ in the group regularisation defines a regularisation path containing GAPC mortality models with varying levels of complexity among which we need to choose the model considered best for a given application. To determine the optimal degree of penalisation, we propose a block cross-validation-style approach (Racine 2000; Bergmeir and Benítez 2012; Bergmeir, Costantini, and Benítez 2014) that aims to assess how well a model is likely to perform over a given forecasting horizon. This approach, which is in the spirit of the resampling methods discussed in Atance and Debón (2020), involves iteratively dividing the in-sample data into training data to fit the models and test data to assess forecasting abilities.

First, a portion of the data (roughly one-third) at the end is held back for validation, denoted as $\mathcal{V}$ and shown by the green points in Figure 2. This corresponds to time periods $t = n_T + 1, n_T + 2, \ldots, n_T + n_E$. The remaining data ($t = 1, 2, \ldots, n_T$) is used for training and testing, where the test sets can take different widths to reflect differing forecasting horizons, as illustrated in Figure 2. For example, to select a model to forecast a 1-year horizon, test sets are defined as 1-year blocks (top row of Figure 2). By contrast, to select a model to forecast a 3-year horizon, test sets are defined as 3-year blocks (bottom row of Figure 2). Therefore, the training sets (blue) for forecasting horizon h are defined as

$$\mathcal{T}^k(h) = \{(x,t)\}_{t \notin [k+1, k+h]}, \tag{34}$$

and the test sets (red) as

$$\mathcal{F}^k(h) = \{(x,t)\}_{t=k+h}, \tag{35}$$

for all $k = 1, 2, ..., n_T - h - 2, n_T - h - 1$.

Inspired by Bergmeir, Hyndman, and Koo (2018), we use data both to the left and right of each test set to train the model, which contrasts heavily to the more common rolling window out-of-sample evaluation illustrated in Figure 3 (Tashman 2000; Dowd et al. 2009). This allows us to conduct significantly more test sets for a given dataset than if we were restricted to fitting models using only data to the left, since this would place a limit on how far to the left test sets can occur. Moreover, we also minimise the portion of data that is left completely unused (grey), thereby extracting as much information as possible and producing an error estimate that is less variable than that of the traditional out-of-sample procedure (Bergmeir, Costantini, and Benítez 2014).

Given that every training set will contain at least one observation from every age (row) and cohort (major diagonal) of the dataset, the full vector of coefficients will be produced for the age, $\hat{\boldsymbol{\alpha}}$, and cohort, $\hat{\boldsymbol{\gamma}}$, terms. In other words, we will be able to estimate a coefficient corresponding to every age group and to every cohort. However, there will not be any observations from the non-training years, and thus all $\hat{\boldsymbol{\kappa}}^{(i)}$ will not be estimated fully. Figure 4

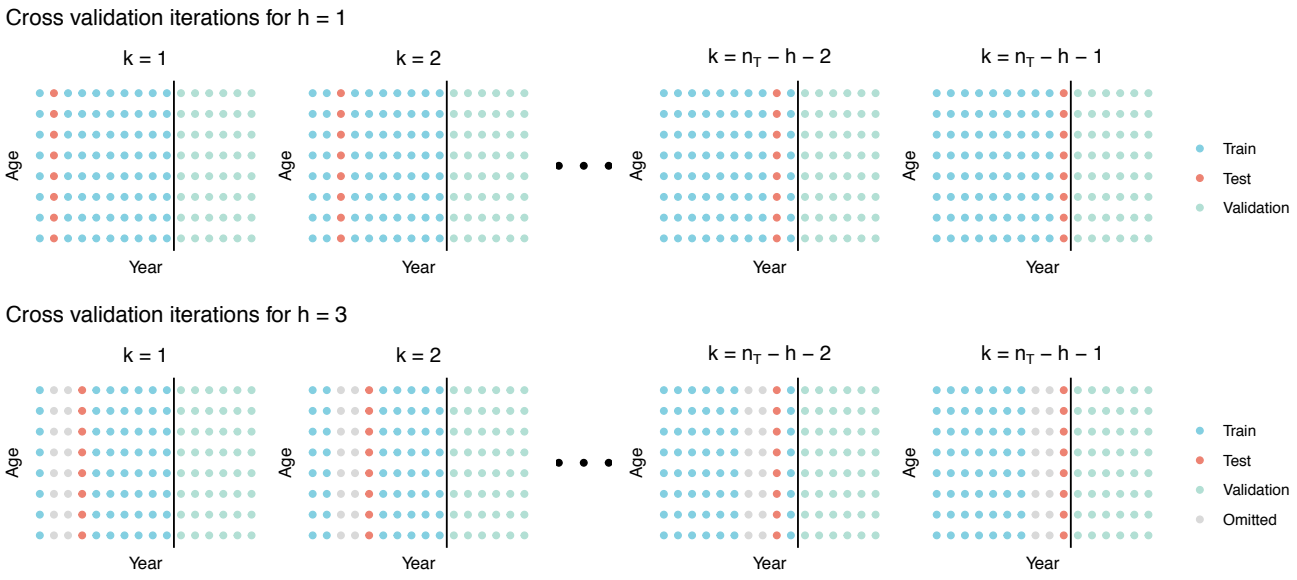


Figure 2: Iterations of cross validation for horizon $h = 1$ (top row) and $h = 3$ (bottom row).

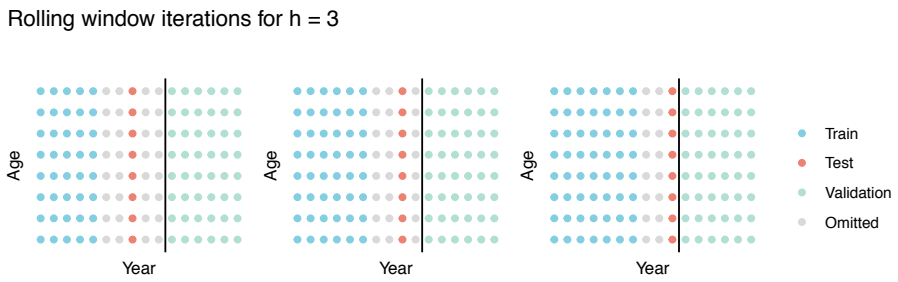


Figure 3: Iterations of a rolling window approach for horizon $h = 3$.

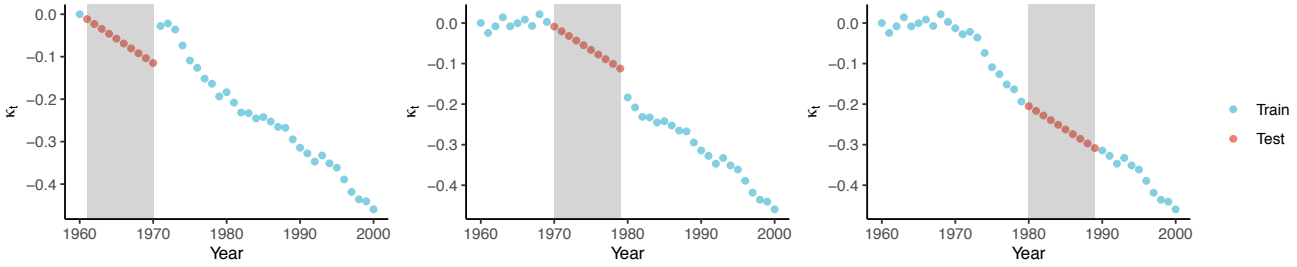


Figure 4: Illustration of time-series imputation approach. The blue points correspond to $\hat{\kappa}_t$ estimates that can be estimated directly from the training set, and the grey area to the non-training years that cannot be estimated with the red points indicating the values imputed using a forward fill procedure.

illustrates three examples, where the blue circles correspond to $\hat{\kappa}_t^{(i)}$ values that can be estimated directly from the training data, and the grey shaded regions to non-training years. Therefore, in order to assess the ability of a model to predict the test sets, we need to impute the missing values. This is achieved by fitting a random walk with drift using all the period indices that are observable (blue circles), both to the left and right of the missing region

$$\kappa_t^{(i)} = \kappa_{t-1}^{(i)} + \delta^{(i)} + \epsilon_t^{(i)}, \tag{36}$$

where $\epsilon_t^{(i)}$ is Gaussian white noise process and $\delta^{(i)}$ is the drift. We can then predict the red circles using a forward fill procedure, basing predictions on the value from the previous period

$$\hat{\kappa}_t^{(i)} = \hat{\kappa}_{t-1}^{(i)} + \hat{\delta}^{(i)}. \tag{37}$$

With these estimates, we can then compute the fitted/predicted values

$$\widehat{\ln(\mu_{x,t})} = \hat{\alpha}_x + \sum_{i=1}^{B} f^{(i)}(x)\hat{\kappa}_t^{(i)} + \hat{\gamma}_{t-x}, \tag{38}$$

and the mean squared error (MSE) for the test set

$$\mathrm{CV}_k(h) = \frac{1}{n_a} \sum_{(x,t)\in\mathcal{F}^k(h)} (\widehat{\ln(\mu_{x,t})} - \ln(\mu_{x,t}))^2. \tag{39}$$

This procedure is iterated for all test sets, $\mathcal{F}^k(h)$, $k = 1, 2, ..., n_T - h - 2, n_T - h - 1$, to determine the cross validation error at horizon h

$$\mathrm{CV}(h) = \frac{1}{n_T - h - 1} \sum_{k=1}^{n_T-h-1} \mathrm{CV}_k(h). \tag{40}$$

The model with the lowest cross validation error is then selected as the best model for forecasting horizon h. We can test a grid of penalties, $\mathbf{\Lambda} = \{\lambda_\ell\}_{\ell=1}^{L}$, and select, for every horizon $h = 1, \ldots, H$, the optimal value which minimises this error:

$$\hat{\lambda}(h) = \underset{\lambda\in\mathbf{\Lambda}}{\operatorname{argmin}}\, \mathrm{CV}(h;\lambda). \tag{41}$$

4.6. Rolling window model evaluation

After selection, we must then evaluate our chosen model using completely independent and unseen data to ensure that it performs competitively relative to its peers. We implement out-of-sample evaluations with a rolling window approach, akin to Dowd et al. (2009), using the validation data, $\mathcal{V}$.

As shown in Figure 5, we fit the models using all in-sample data (blue) and forecast on the unseen validation set (red), which we assume contains n_E years. We then roll forward one period $n_E - h$ times until all data is exhausted. Similar to the cross validation procedure, forecasting is enabled by extrapolating the period and cohort parameters over the required time horizon, with a difference being that this is now achieved using conventional time series forecasting as discussed in Section 3.1. The out-of-sample mean squared errors at horizon h for each validation set can then be computed, using

$$\text{MSE}_j(h) = \frac{1}{n_a} \sum_{(x,t)\in\mathcal{V}^j(h)} (\widehat{\ln(\mu_{x,t})} - \ln(\mu_{x,t}))^2, \tag{42}$$

where $\mathcal{V}^j(h) = \{(x,t)\}_{t=n_T+j+h-1}$ for $j = 1,\ldots,n_E - h + 1$. The overall out-of-sample mean squared error at horizon h is then given by:

$$\text{MSE}(h) = \frac{1}{n_E - h + 1} \sum_{j=1}^{n_E-h+1} \text{MSE}_j(h). \tag{43}$$

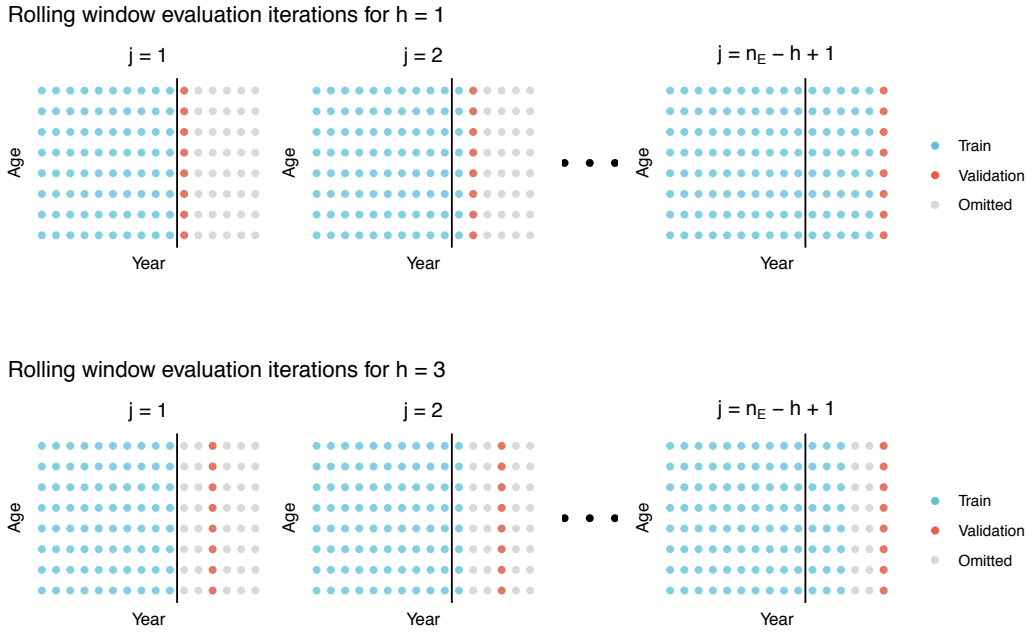


Figure 5: Iterations of rolling window out-of-sample evaluation for horizon $h = 1$ (top row) and $h = 3$ (bottom row).

5. Illustration using data for United States males

In this section, we illustrate the proposed group regularisation construction framework using United States males mortality data obtained from the Human Mortality Database (HMD). We focus on modelling adult age ranges, 20 to 89, using years 1960 to 2000 as our in-sample data, and years 2001 to 2017 as our validation set.

5.1. Initial basis functions

To implement the group regularisation framework, we consider a base model with $B = 37$ basis functions taken from the functional forms in Table 2 as follows:

$$f^{(1)}(x) = 1; \tag{44}$$
$$f^{(2)}(x) = x - \bar{x}; \tag{45}$$
$$f^{(2+j)}(x) = (x - \bar{x})^{j+1}, \quad j = 1, \ldots, 9; \tag{46}$$
$$f^{(11+r)}(x) = (x - 20 - 5r)^{+}, \quad r = 1, \ldots, 13; \tag{47}$$
$$f^{(24+p)}(x) = (20 + 5p - x)^{+}, \quad p = 1, \ldots, 13. \tag{48}$$

5.2. Regularisation path

To search for the optimal level of regularisation, we consider a grid of 25 values of λ equally spaced on the log scale and in the range $\left[e^{-9}, e^{-3.5}\right]$. This range is wide enough to produce a regularisation path which starts with a very over-parametrised model and ends in a very parsimonious model. The resulting regularisation path is shown in Figure 6. This plot shows the relative size of each Group j of coefficients $\|\boldsymbol{X}_j\boldsymbol{\beta}_j\|/\|\boldsymbol{X}\boldsymbol{\beta}\|$ as a function of $\ln(\lambda)$. In this particular exercise, we have decided not to penalise the $\boldsymbol{\alpha}$ parameters so that the model always contains a static age term α_x independently of the amount of regularisation. This is because we are considering the age range 20-89, where the static age term is always believed to be important so that the accident hump at younger ages is appropriately captured (Hunt and Blake 2020d; Plat 2009).

Small values of λ correspond to low levels of regularisation and more complex models. As such, to the left of Figure 6 when $\lambda = 0.0001234098$, we start with the model:

$$\begin{aligned}\ln(\mu_{x,t}) = \alpha_x + \kappa_t^{(1)} + (\bar{x} - x)\kappa_t^{(2)} + (x - 70)^{+}\kappa_t^{(21)} + (x - 85)^{+}\kappa_t^{(24)} + \\ (25 - x)^{+}\kappa_t^{(25)} + (30 - x)^{+}\kappa_t^{(26)} + (45 - x)^{+}\kappa_t^{(29)} + \gamma_c,\end{aligned} \tag{49}$$

which, in addition to the default static age term and the cohort effect, includes 7 of the 37 possible age-period terms. Noticeably, this very complex model includes terms that capture the idiosyncrasies of mortality trends and noise at very young and very old ages via the $\kappa_t^{(21)}$, $\kappa_t^{(24)}$, $\kappa_t^{(25)}$, and $\kappa_t^{(26)}$ terms. As we increase the value of λ, the

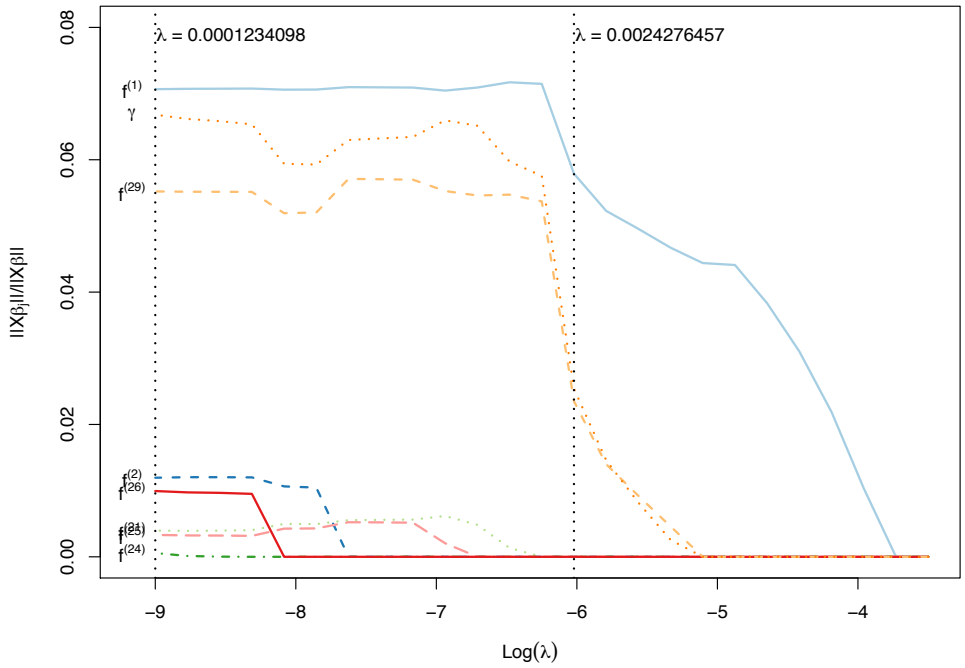


Figure 6: Regularisation path of mortality models for United States males fitted to ages 20 to 89 for years 1960 to 2000.

additional regularisation induces the elimination of some age-period terms, resulting in more parsimonious models. For example, when we reach a value of $\lambda = 0.0024276457$, we obtain the model

$$\ln(\mu_{x,t}) = \alpha_x + \kappa_t^{(1)} + (45 - x)^+ \kappa_t^{(29)} + \gamma_c, \tag{50}$$

which, as opposed to the seven age-period terms of the initial model, only includes two age-period terms capturing the general level of mortality and the differential improvement of individuals younger than age 45. We note that this model is very similar to the sPLAT model in Equation (15), albeit without the linear term and with the strike of the put function at 45 rather than $\bar{x} = 54.5$.

5.3. Selected values of λ

Given the regularisation path of models, the next step is to select a model suitable for each forecasting horizon using the block cross validation approach described in Subsection 4.5. The selected $\hat{\lambda}(h)$ for each horizon h is the value that minimises the cross validation MSE. Figure 7 illustrates this for horizons $h = 1, 5$, and 10, where the selected values correspond to $\hat{\lambda}(1) = 0.0001234098$, $\hat{\lambda}(5) = 0.001220697$, and $\hat{\lambda}(10) = 0.0024276457$, respectively. Note, for λ values to the right of each respective $\hat{\lambda}$, we have under-fitting, whilst to the left, we have over-fitting. Table 3 presents the selected $\hat{\lambda}(h)$ for all horizons of interest along with a summary of the corresponding model structure.

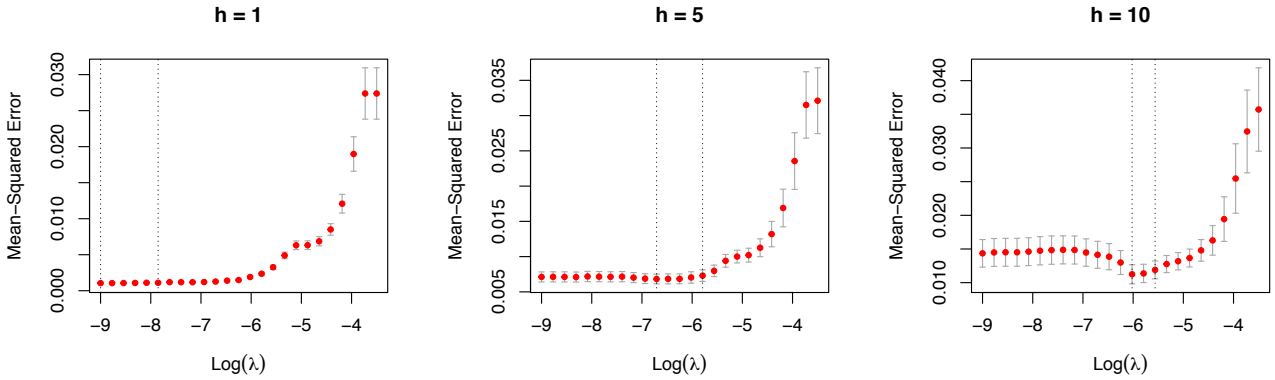


Figure 7: Cross validation MSE for forecasting horizons $h = 1$, $h = 5$ and $h = 10$ for United States males data fitted to ages 20 to 89 for years 1960 to 2000.

Table 3: Summary of optimal model structures for group regurlarisation for ages 20-89 in the period 1960-2000 in USA males.

Horizon (h)	1	2	3	4	5	6	7	8	9	10
$\hat{\lambda}(h)$	0.0001	0.0001	0.0002	0.0012	0.0012	0.0015	0.0024	0.0024	0.0024	0.0024
Period terms	7	7	6	3	3	3	2	2	2	2
Cohort	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

In the case of 1-year and 2-year forecasting horizons, the selected model has seven age-period terms and corresponds to the specification in Equation (49). This contrasts with the 7 to 10-year forecasting horizons for which the selected model corresponds to the simpler structure in Equation (50), which has only two age-period terms. In general, as indicated by Table 3, the longer the forecasting horizon, the higher the value of $\hat{\lambda}(h)$, the fewer the number of period terms, and the more parsimonious the selected model specification.

5.4. Out-of-sample evaluation

Having selected appropriate models for each horizon, we now evaluate the out-of-sample performance of these models and their performance relative to existing GAPC models. We evaluate the forecasting performance with the rolling window approach described in Section 4.6 using the validation data from 2001 to 2017. Thus, for horizon

$h = 1$, there are 17 years for evaluation, which reduces to 8 years for horizon $h = 10$ namely, 2010, 2011, ..., 2017, whose forecasts are derived using as last training point the years 2000, 2001, ..., 2007, respectively.

We consider as benchmarks eight of the ten GAPC models presented in Table 1. We have discarded the CBD and M7 models as they are both designed to fit higher ages (typically ages 50 and beyond), and thus are not entirely appropriate for the 20 to 89 age range considered in our experiments. For consistency, we have implemented all the classical GAPC models under a log-Poisson setting within the `R` package **StMoMo**. Appendix A includes the constraints used to ensure the identifiability of the models.

In the evaluation exercise, we consider two versions of the group regularisation (GR) framework obtained by varying the number and form of the initial basis functions. The first version, herein referred to as GR1, corresponds to the implementation described previously in Subsection 5.1 which uses the $B = 37$ basis functions in Equations (44)-(48).

The second version, referred to as GR2, explores the use of a larger pool of initial basis functions in order to enrich the types of trends that the model can capture. To do so, in addition to the 37 basis functions in Equations (44)-(48), we consider additional cubic B-Splines basis functions (Currie, Durban, and Eilers 2004), which are piecewise polynomials commonly used in the graduation and forecasting of mortality (Macdonald, Richards, and Currie 2018). Figure 8 depicts the basis functions we add. These are comprised of four groups of B-splines, each of which contain 2, 4, 8, and 16 functions respectively. Thus, the total number of basis in the initial model under the GR2 implementation is $B = 37 + 2 + 4 + 8 + 16 = 67$.

For the forecasting of the period functions in GAPC and GR models, we follow the standard approach and use a multivariate random walk with drift (Cairns et al. 2010; Haberman and Renshaw 2011; Villegas, Kaishev, and Millossovich 2018). In addition, for models with a cohort effect, we forecast γ_{t-x} using an ARIMA model as in Equation (12) with level of differentiation $d = 1$, autoregressive order $p = 1$ and moving average order $q = 0$. As discussed before in Subsection 4, choosing $d = 1$ tempers the possible impact of model identifiability on projections (Currie 2020). Moreover, setting $p = 1$ is in line with the level auto regression found by others when modelling cohort effects (Cairns et al. 2008, 2011; Hunt and Blake 2020a).

Table 4 presents the out-of-sample mean squared error (MSE) by forecasting horizon for each of the eight GAPC models and the two implementations of the group regularisation framework. Among the standard GAPC models,

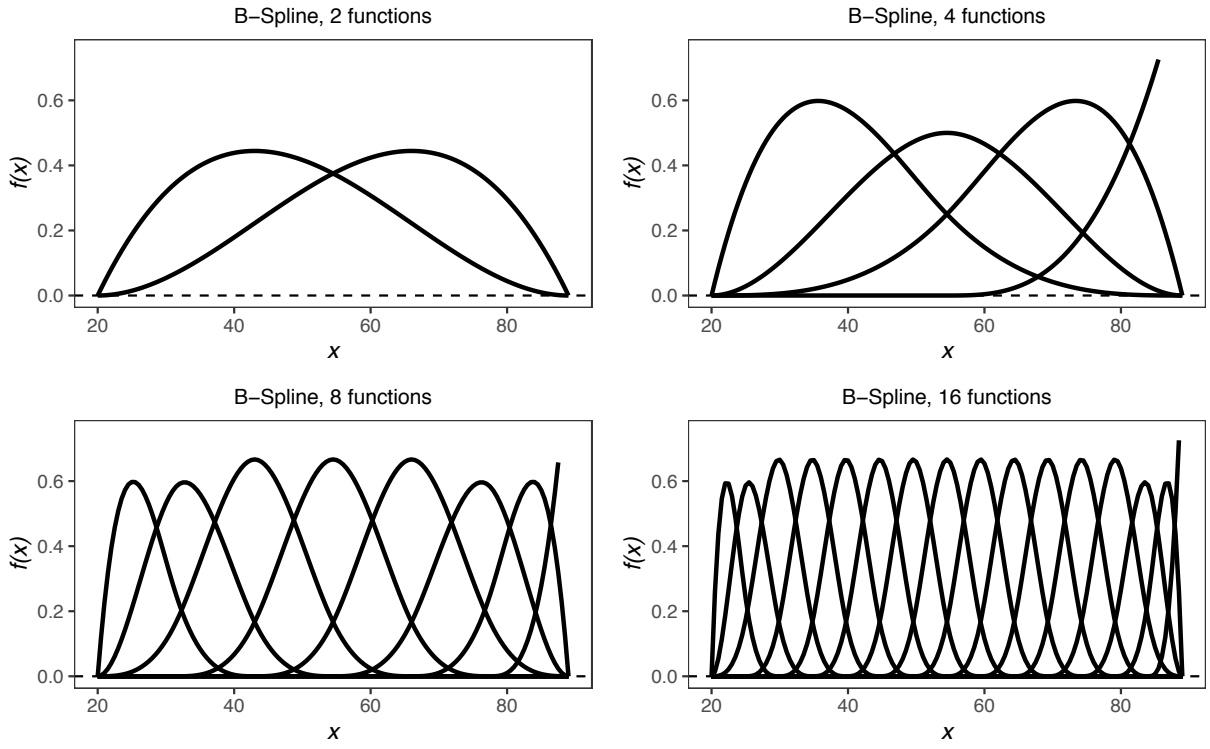


Figure 8: Cubic B-splines basis functions added to the initial model for the GR2 version of the group regularisation implementation.

Table 4: Out-of-Sample Mean Square Error for USA males by forecasting horizon with model ranking in brackets. Values are presented as 1000×MSE.

h	CBDx	LC	APC	sPLAT	RH	LC2	cPLAT	M7x	GR1	GR2
1	8.1 (10)	6.69 (9)	2.93 (7)	1.83 (6)	1.53 (4)	5.85 (8)	1.39 (3)	1.61 (5)	1.32 (2)	1.09 (1)
2	10.46 (10)	8.42 (8)	4.3 (7)	3.29 (4)	3.11 (3)	8.44 (9)	3.05 (2)	3.66 (6)	3.46 (5)	2.91 (1)
3	13.16 (10)	10.34 (8)	5.83 (5)	4.84 (1)	5.65 (4)	10.93 (9)	4.92 (2)	6.09 (7)	6.02 (6)	5.05 (3)
4	16.01 (10)	12.38 (8)	7.3 (5)	6.26 (1)	9.27 (7)	13.38 (9)	6.48 (2)	8.39 (6)	7.29 (4)	7.01 (3)
5	18.85 (10)	14.35 (8)	8.6 (3)	7.42 (1)	13.37 (7)	15.99 (9)	7.75 (2)	10.54 (6)	9.35 (5)	8.67 (4)
6	21.38 (10)	16.25 (8)	9.8 (4)	8.46 (1)	15.95 (7)	18.5 (9)	8.97 (2)	12.74 (6)	11.32 (5)	9.06 (3)
7	23.15 (10)	18.2 (7)	10.54 (5)	8.78 (1)	20.43 (8)	20.92 (9)	9.42 (2)	14.2 (6)	10.33 (4)	9.82 (3)
8	24.17 (9)	19.78 (7)	10.4 (5)	7.97 (1)	24.48 (10)	23.13 (8)	8.61 (3)	14.4 (6)	9.68 (4)	8.13 (2)
9	26.15 (9)	21.25 (7)	10.25 (5)	7.34 (2)	28.9 (10)	25.19 (8)	7.88 (3)	14.99 (6)	9.06 (4)	6.5 (1)
10	28.41 (9)	22.65 (7)	11.14 (4)	7.08 (1)	33.52 (10)	27.23 (8)	7.65 (2)	16.1 (6)	8.96 (3)	11.48 (5)

the sPLAT and cPLAT model have the best forecasting performance at every horizon, with the sPLAT model topping the ranking at 7 of the 10 forecasting horizons.

The group regularisation (GR) models perform competitively, outperforming the CBDx, LC, LC2, and M7x models at every horizon and the APC and RH at most horizons. Moreover, the GR approaches show a very consistent performance across all horizons, with MSE values which are very close to the MSE values of the sPLAT and cPLAT models. This is not surprising in the case of GR1, as for the USA male dataset for horizons $h = 7, 8, 9, 10$, the GR1 approach identified a model which closely resembles the cPLAT model (recall Equation (50)).

It is worth noting that for horizons 1 and 2, the GR2 approach outperforms all nine other competing models by a significant margin. By including the 67 additional basis functions depicted in Figure 8, the GR2 model can explore more complex models that are able to better capture the cross-sectional patterns in the mortality data. Hence, the improved performance of the GR2 over the rest of the models highlights the benefits of allowing for a larger variety of age-specific trends and complexity of model when conducting short-term mortality forecasts.

6. Application to 52 datasets from the Human Mortality Database

To further explore the proposed group regularisation model construction framework, in this section we apply this approach to 52 different empirical datasets from the Human Mortality Database (HMD), corresponding to 26 countries and two genders. At the time of writing, the HMD had detailed mortality data for 41 countries or areas. The years of data vary by country, with only 28 countries or areas having data for the standardised period 1960 to 2017. Of the 28 remaining countries, we discard Iceland and Luxembourg as they have multiple data cells with zero deaths, leaving us with 26 countries. Similar to the previous USA male data example, we use the years 1960 to 2000 as our in-sample data and reserve the years 2001 to 2017 for out-of-sample evaluation.

6.1. Summary of model structures

The results of applying the group regularisation approach reveal that the optimal model varies not only by country, but also within country, depending on the forecasting horizon. To illustrate this, Table 5 reports the average characteristics of the model structures produced by the GR1 approach for different horizons.

Table 5: Summary of optimal model structures according to GR1 for ages 20-89 in the period 1960-2000 in 52 datasets.

Horizon (h)	1	2	3	4	5	6	7	8	9	10
Avg. $\hat{\lambda}(h)$	0.0018	0.0025	0.0029	0.0034	0.0047	0.0051	0.0052	0.0055	0.0057	0.0059
Avg. period terms	4.77	4	3.63	2.88	2.67	2.69	2.44	2.44	2.29	2.27
% with cohort	98%	94%	90%	83%	75%	73%	69%	69%	60%	60%

Notably, across all 52 datasets, $\hat{\lambda}(h)$ tends to increase with the horizon of interest. This means that simpler more parsimonious models are constructed for longer forecasting periods. For example, while at horizon 1 the models have on average 4.77 period terms, at horizon 10, the average number of period terms reduces to 2.27. It is also

worth noting that the longer the horizon, the less likely the optimal structure tends to include a cohort effect. While at short horizons the optimal structure will almost certainly include a cohort effect, at horizon 10, only 60% of the 52 datasets will have a cohort effect.

These results conform with the hypothesis of parsimony being preferred for longer horizons. More specifically, for shorter forecasting horizons, the predictive success of a model is determined by its ability to accurately fit the cross section of the mortality curve. By contrast, as the forecasting horizon increases, it becomes more important to filter out the year-on-year noise in the data, and having a closer fit to the initial mortality curves becomes less relevant.

As an illustration of the diversity of models produced by the GR framework, Table 6 shows for the 52 datasets the optimal model produced by the GR1 approach for forecasting horizon 10. In some cases, the optimal models are similar in structure to existing GAPC models. For example, for USA females, Japanese males, UK (GBR-NP) females, Belarusian males, Austrian males, and Bulgarian males, the structure of the optimal model is the same as the sPLAT model. Similarly, for Swedish males, Czech females, and Irish females, the optimal model structure coincides with the CBDx model. In most cases, though, the optimal models vary quite remarkably from existing models, but with the following commonalities in the type of basis functions used for the different countries:

- For 79% of the populations, the model contains a unit period term, $\kappa_t^{(1)}$, capturing changes in the general level of mortality that affect all ages.
- For 42% of the populations, the model contains a linear term, $(x-\bar{x})\kappa_t^{(2)}$ which captures changes in the slope of the mortality curve.

- Only 6% of the models include a quadratic term, $(x-\bar{x})^2\kappa_t^{(3)}$, which captures the curvature of the mortality curve.
- 37% of models include at least one of the terms, $(x-25)^+\kappa_t^{(23)}$, $(x-30)^+\kappa_t^{(24)}$ or $(x-35)^+\kappa_t^{(25)}$, that are specifically targeting declining mortality rates at ages below 35, known as the "flattening of the accident hump."

Existing GAPC models are unable to capture the right combination of the above trends, stressing the merits of our data-driven approach and the pitfalls of using pre-defined structures for all data sets.

6.2. Out-of-sample evaluation

Whilst the previous section implies that a wide range of bespoke models may be more appropriate to explain historical data, it is important to assess whether the GR-constructed models generalise well out-of-sample and if they perform competitively relative to existing GAPC models. For the 52 datasets, we replicate the out-of-sample evaluation described in Section 5.4.

Figure 9 presents heatmaps reporting the ranking of the eight GAPC models and the two GR approaches for forecasting horizons $h = 1, 5$, and 10. In this figure, the GAPC models are ordered in terms of their complexity as indicated by their number of parameters. Similarly, the countries are ordered according to their population size. To facilitate the interpretation of these model rankings, Figure 10 presents the mean ranking of each of the 12 models as a function of the 10 forecasting horizons of interest.

Over the 1-year horizon, the GR2 approach produces the best result with a median rank of 2, and is in the top three of models in 41 of the 52 datasets. Interestingly, the best performance of the GR2 approach is among the countries with the biggest populations. The GR1 approach also has a very strong performance at short horizons, surpassing most GAPC models, but is almost always outperformed by the GR2 approach. This reaffirms the benefits of allowing for a larger variety of age-specific trends when focusing on short-term forecasting so that the cross-sectional variability of the mortality curve is better captured. The cPLAT and M7x models are the two best-performing GAPC models, whilst the more parsimonious CBDx, APC, and LC models perform relatively poorly. This reaffirms our hypothesis that complex models are preferred for shorter horizons.

As we increase the forecasting horizon, the absolute values of the forecasting errors increase for all models, as would be expected. However, the relative rankings of the models also change, which is of far more interest. The strong performance of our GR approaches remains consistent, with the GR1 in particular having the best median rank at horizons 5 and 10, reflecting the ability of our regularisation approach to filter out the year-on-year noise in

Table 6: Optimal model structures according to GR1 for 10-year forecasting horizon using data for ages 20-89 in the period 1960-2000.

Data	Form
USA (M)	$\alpha_x + \kappa_t^{(1)} + (45-x)^+\kappa_t^{(29)} + \gamma_{t-x}$
USA (F)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + \gamma_{t-x}$
JPN (M)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + \gamma_{t-x}$
JPN (F)	$\alpha_x + \kappa_t^{(1)}$
ITA (M)	$\alpha_x + \kappa_t^{(1)} + (40-x)^+\kappa_t^{(28)} + \gamma_{t-x}$
ITA (F)	$\alpha_x + \kappa_t^{(1)}$
GBR-NP (M)	$\alpha_x + \kappa_t^{(1)} + (45-x)^+\kappa_t^{(29)} + \gamma_{t-x}$
GBR-NP (F)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + \gamma_{t-x}$
FRATNP (M)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (x-\bar{x})^2\kappa_t^{(3)} + (25-x)^+\kappa_t^{(25)} + \gamma_{t-x}$
FRATNP (F)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (x-80)^+\kappa_t^{(23)} + (25-x)^+\kappa_t^{(25)} + \gamma_{t-x}$
ESP (M)	$\alpha_x + (x-25)^+\kappa_t^{(12)} + (x-75)^+\kappa_t^{(22)} + (25-x)^+\kappa_t^{(25)} + (30-x)^+\kappa_t^{(26)} + (45-x)^+\kappa_t^{(29)} + \gamma_{t-x}$
ESP (F)	$\alpha_x + \kappa_t^{(1)}$
POL (M)	$\alpha_x + \kappa_t^{(1)} + (35-x)^+\kappa_t^{(27)} + \gamma_{t-x}$
POL (F)	$\alpha_x + (x-25)^+\kappa_t^{(12)} + (55-x)^+\kappa_t^{(31)} + \gamma_{t-x}$
AUS (M)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (x-\bar{x})^2\kappa_t^{(3)} + (25-x)^+\kappa_t^{(25)} + (35-x)^+\kappa_t^{(27)} + \gamma_{t-x}$
AUS (F)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (x-\bar{x})^4\kappa_t^{(5)} + \gamma_{t-x}$
NLD (M)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (x-70)^+\kappa_t^{(21)} + (25-x)^+\kappa_t^{(25)} + \gamma_{t-x}$
NLD (F)	$\alpha_x + \kappa_t^{(1)}$
BEL (M)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (40-x)^+\kappa_t^{(28)} + \gamma_{t-x}$
BEL (F)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (25-x)^+\kappa_t^{(25)} + \gamma_{t-x}$
CZE (M)	$\alpha_x + \kappa_t^{(1)} + (40-x)^+\kappa_t^{(28)}$
CZE (F)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (25-x)^+\kappa_t^{(25)}$
PRT (M)	$\alpha_x + (x-25)^+\kappa_t^{(12)} + (30-x)^+\kappa_t^{(26)} + (70-x)^+\kappa_t^{(34)} + \gamma_{t-x}$
PRT (F)	$\alpha_x + \kappa_t^{(1)}$
HUN (M)	$\alpha_x + (x-25)^+\kappa_t^{(12)} + (30-x)^+\kappa_t^{(26)} + (85-x)^+\kappa_t^{(37)} + \gamma_{t-x}$
HUN (F)	$\alpha_x + (x-\bar{x})^2\kappa_t^{(3)}$
BLR (M)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + \gamma_{t-x}$
BLR (F)	$\alpha_x + (x-25)^+\kappa_t^{(12)} + (40-x)^+\kappa_t^{(28)}$
SWE (M)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)}$
SWE (F)	$\alpha_x + \kappa_t^{(1)}$
AUT (M)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + \gamma_{t-x}$
AUT (F)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})^3\kappa_t^{(4)} + (25-x)^+\kappa_t^{(25)} + \gamma_{t-x}$
CHE (M)	$\alpha_x + \kappa_t^{(1)} + (x-75)^+\kappa_t^{(22)} + (25-x)^+\kappa_t^{(25)} + (40-x)^+\kappa_t^{(28)} + \gamma_{t-x}$
CHE (F)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)}$
BGR (M)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + \gamma_{t-x}$
BGR (F)	$\alpha_x + (80-x)^+\kappa_t^{(36)}$
DNK (M)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (25-x)^+\kappa_t^{(25)} + \gamma_{t-x}$
DNK (F)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (25-x)^+\kappa_t^{(25)} + \gamma_{t-x}$
SVK (M)	$\alpha_x + \kappa_t^{(1)} + (35-x)^+\kappa_t^{(27)} + \gamma_{t-x}$
SVK (F)	$\alpha_x + \kappa_t^{(1)} + (45-x)^+\kappa_t^{(29)} + \gamma_{t-x}$
FIN (M)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (x-\bar{x})^4\kappa_t^{(5)} + \gamma_{t-x}$
FIN (F)	$\alpha_x + \kappa_t^{(1)}$
NOR (M)	$\alpha_x + \kappa_t^{(1)} + (x-70)^+\kappa_t^{(21)} + (35-x)^+\kappa_t^{(27)} + \gamma_{t-x}$
NOR (F)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (30-x)^+\kappa_t^{(26)}$
IRL (M)	$\alpha_x + \kappa_t^{(1)} + (35-x)^+\kappa_t^{(27)} + \gamma_{t-x}$
IRL (F)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)}$
LTU (M)	$\alpha_x + \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (25-x)^+\kappa_t^{(25)} + \gamma_{t-x}$
LTU (F)	$\alpha_x + (40-x)^+\kappa_t^{(28)}$
LVA (M)	$\alpha_x + \kappa_t^{(1)} + (x-60)^+\kappa_t^{(19)} + (40-x)^+\kappa_t^{(28)}$
LVA (F)	$\alpha_x + (70-x)^+\kappa_t^{(34)} + (80-x)^+\kappa_t^{(36)}$
EST (M)	α_x
EST (F)	α_x

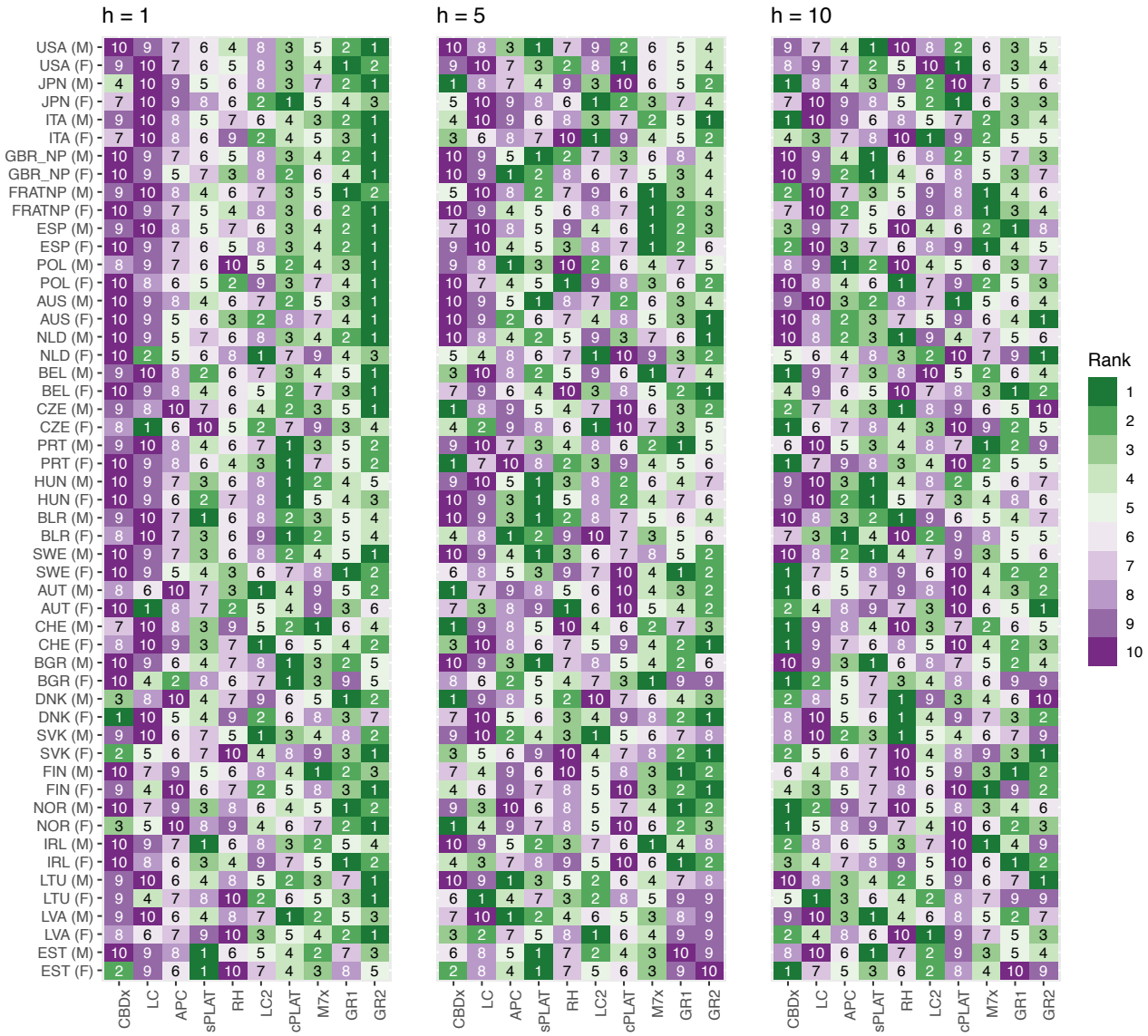


Figure 9: Ranking of models according to out-of-sample MSE at $h = 1$, $h = 5$, and $h = 10$, for different datasets.

the mortality trends to produce a parsimonious model which concentrates on the secular longer-term trends.[2] In contrast, the other models show a marked shift, with more parsimonious models, such as the APC and the CBDx, now performing significantly better.

Interestingly, the popular cPLAT and LC models rank within the bottom four models almost exclusively for the longer horizons. Notably, one of the only datasets where the cPLAT model performs reasonably well over a 10-year horizon is for the United States populations, which is actually the data that Plat (2009) used to construct this model. This lends support to the notion that if we examine data and build a model based on the patterns observed within it, as Plat (2009) did for the United States, we can expect to yield strong forecasting results. However, we cannot just assume that a model that explains United States mortality well will fit well to all other countries. Thus, we must repeat this procedure for each dataset, which is effectively what our data-driven approach aims to do.

[2]At horizon 5, however, the GR approaches have a poor performance for both genders in Estonia, Latvia, and Lithuania, and for Slovakian males and Bulgarian females. For the Baltic states in particular, this poor performance can partially be explained by the inability of our approach to capture the change in trend that occurred after Soviet independence in 1991, when mortality passed from a stable or increasing trend to a declining trend more in line with the trends in Western European countries (Karanikolos et al. 2012).

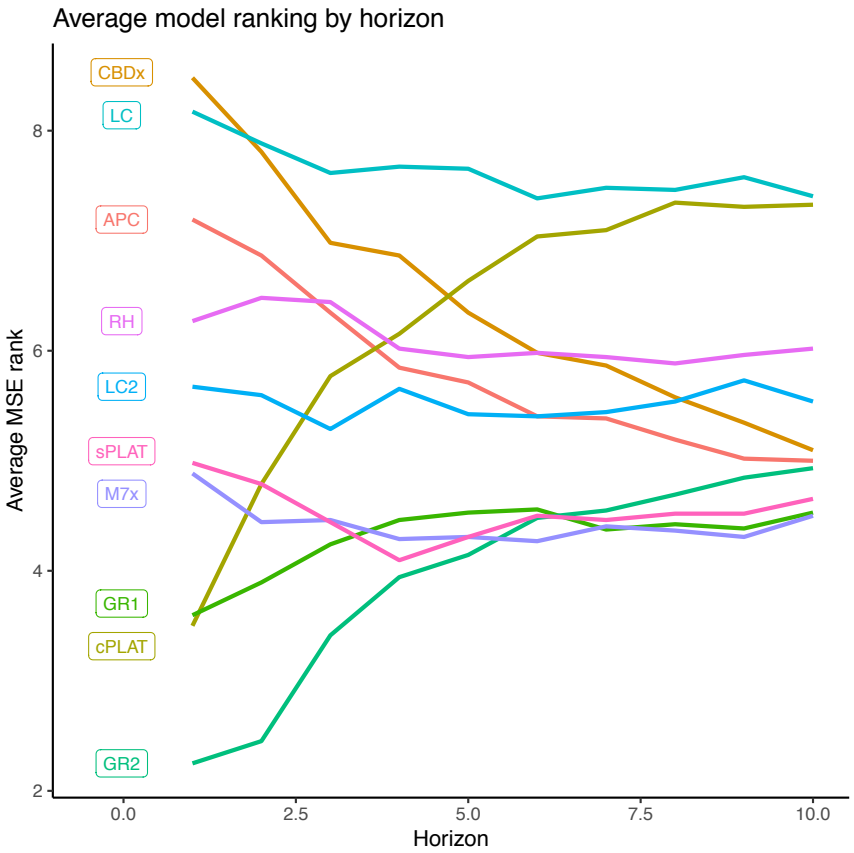


Figure 10: Average of out-of-sample MSE rank across all 52 datasets for each model.

7. Simulation study

To validate the findings from the preceding section in a more robust manner, we consider the use of simulated datasets. These are created by first simulating mortality rates from a specific GAPC model, and subsequently sampling the number of deaths from a Poisson distribution to reflect random variation, with the detailed simulation process described in Appendix B. This framework provides a controlled setting to isolate and understand specific effects, namely (a) the complexity of the underlying process from which mortality rates are generated, (b) the population size from which deaths are sampled (or level of noise), and (c) the length of the forecasting horizon. For each of the following subsections, we vary systematically one of these three elements and use as a reference example for comparison the case where rates are generated from an APC model, using a population size of 5 million to predict rates at a forecasting horizon of 10 years. For each combination of inputs, we run 100 simulations.[3] For this section, the RH, LC2, and GR2 models are not considered due to being computationally taxing.

7.1. Effect of generator model

Figure 11 illustrates forecasting errors using data generated by a number of GAPC models, including models defined in the literature (APC, LC, and cPLAT), and others that have been randomly constructed using a suite of basis functions. These models have been created to reflect varying numbers of age-period terms, the inclusion/exclusion of a cohort term, and a range of basis functions with different forms and parameterisations. This includes a basis function not considered in the GR approach in the bottom right panel, namely a unit period term applying to ages 50 and below only.

[3]While 100 simulations seems to be a relatively small number of simulations, we keep this number small due to the intensive computational requirements of the exercise. Moreover, our focus is on model rankings rather than absolute values of performance quantities. As such, the conclusions are likely to remain unchanged if we increase the number of simulations.

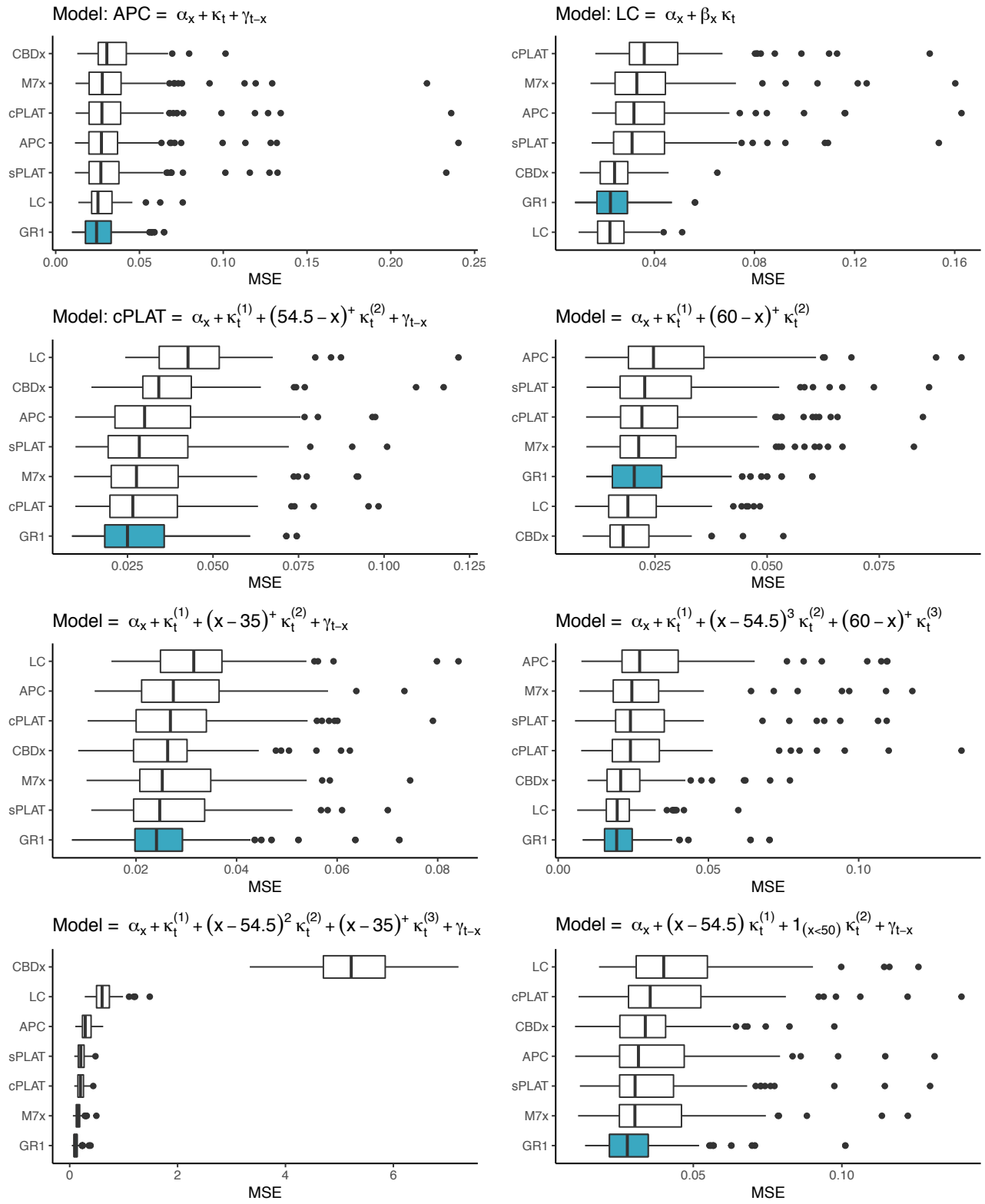


Figure 11: Boxplot of mean squared errors from 100 simulated datasets, involving (a) rates generated from different GAPC models, (b) using a population size of 5 million, and (c) a forecasting horizon of 10 years.

As expected, in the first three panels, the model that generated each dataset performs very well with regards to both median errors and the range of errors across the 100 simulations. However, the performance of models on data generated from another model is of far greater interest. Notably, no single GAPC model performs strongly across all scenarios, with the complex models, such as the cPLAT and M7x, tending to perform stronger on the more complex datasets, and simpler more parsimonious models, such as the LC and APC, performing better on the simpler datasets. More strikingly, this study highlights the significance of the cohort term, where models without a cohort term perform extremely poorly where cohort effects are prevalent in the underlying data. This is most pronounced in the bottom left panel of Figure 11, where the LC and CBDx models generate unwieldy errors.

By contrast, the GR1 model performs strongly and consistently, with the lowest median error in six of the eight scenarios, and ranking within the top three models every time. This is hardly surprising, given the adaptability of its framework to select more period and/or cohort terms where more complex patterns exist in the underlying data.

7.2. Effect of population size

When population size decreases (or the level of noise increases), it is obvious that forecasting errors will increase in absolute terms, as illustrated in Figure 12. However, the changes in the relative performance of the models is of greater interest. For very small populations, complex models have a tendency to overfit the Poisson noise in the data, resulting in a poor forecasting performance, with higher median forecasting errors and increased tendencies to produce significantly large errors or anomalies. In contrast, the more parsimonious LC and CBDx models perform stably for these smaller populations as their reduced number of parameters allows them to better filter the Poisson noise.

The GR1 model, however, remains competitive under all scenarios, registering the lowest median error on the 500,000 and 5 million populations, and recording extremely similar median errors to the best models for the larger 50 million and 500 million populations. Notably, in all scenarios, the GR1 approach tends to select a model structure that is similar to or exactly matches that of the APC model that generated the data, which explains the strikingly similar results of these two models for larger populations. Interestingly, however, the GR1 model significantly outperforms the APC model for the smaller populations, which owes to the ability of the regularisation approach to

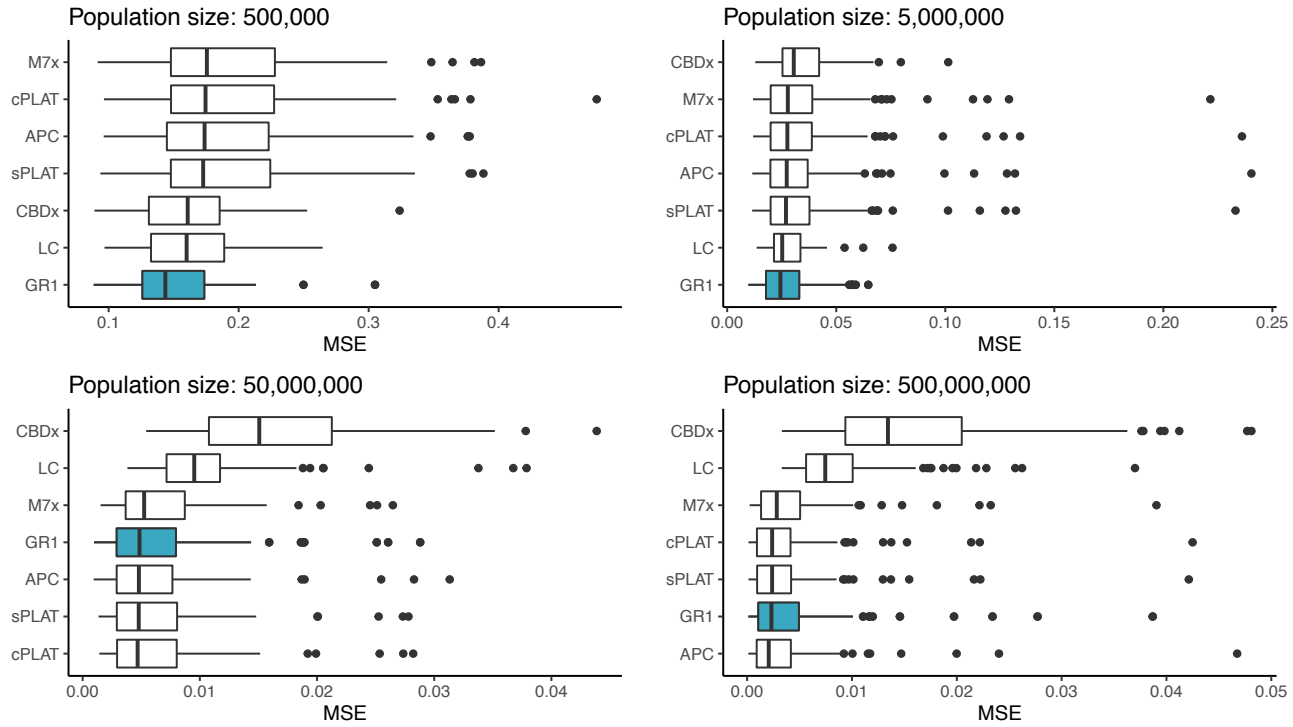


Figure 12: Boxplot of mean squared errors from 100 simulated datasets, involving (a) rates generated from an APC model, (b) using varying population sizes, and (c) a forecasting horizon of 10 years.

not only select basis functions, but to also shrink parameter estimates to avoid over-fitting to noise in the training data.

7.3. Effect of forecasting horizon

Figure 13 illustrates forecasting errors for varying forecasting horizons, where for consistency, the number of years of training data was amended in each scenario such that the GR1 approach had 30 test sets per horizon. For the short-term 1-year horizon, all models perform strongly in absolute terms, with similar median errors observed for all models. However, even for this horizon, it is observed that there is a greater tendency for the complex models such as the cPLAT, M7x, and sPLAT to produce large, unwieldy errors, owing to the potential to overfit to noise given the large number of parameters. This poor and volatile performance of the complex less parsimonious models is even more pronounced as the length of the forecasting horizon increases, where there is a striking difference between the two simplest GAPC models – CBDx and LC – and the others. This mirrors the behaviour observed in Section 6.

For all horizons, the GR1 approach performs strongly. This again owes to the ability to simplify the underlying structure of the model as the forecasting horizon increases.

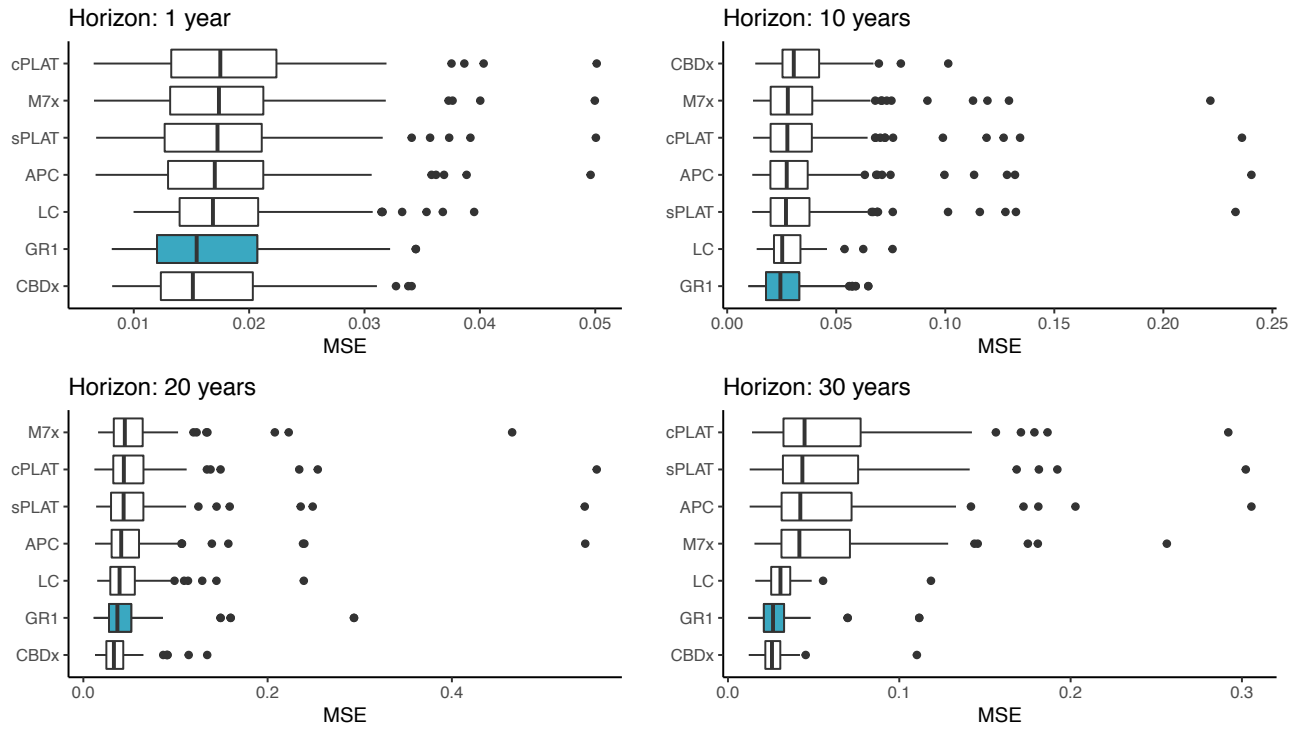


Figure 13: Boxplot of mean squared errors from 100 simulated datasets, involving (a) rates generated from an APC model, (b) using a population size of 5 million, and (c) varying forecasting horizons.

8. Conclusions

In this paper we have introduced and applied a data-driven framework to construct bespoke GAPC mortality models. Our approach considers a large suite of parametric functions, which users can modify based on their views of "demographic significance", to best capture patterns in specific datasets. The procedure is fitted using group regularisation and cross validation techniques to determine the optimal trade-off between complexity and parsimony for each dataset and application. These techniques select which basis functions to include in the final model and determine how heavily to shrink the estimated coefficients to prevent over-fitting.

Our exhaustive empirical analysis revealed that no pre-defined model can perform well for every dataset and application, and thus our regularised data-driven models are far superior to existing models on average. Moreover,

our empirical analysis revealed an interesting divide between simple more parsimonious models and complex less parsimonious models, with parsimonious models outperforming complex models as the forecasting horizon increases and the population size decreases. This link between parsimony and forecasting performance was particularly noticeable at shorter forecasting horizons where allowing for more complex age-specific trends proved to significantly improve the forecasting performance of GAPC models.

However, it is important to stress that the key contribution of this paper is neither the models nor the rankings generated, but rather the framework that we have introduced. We must stress that whilst we have obtained very strong results across all datasets, this paper does not purport to provide all the answers or the perfect models. Instead, we simply aim to provide an alternative approach to thinking about mortality modelling through the lens of statistical learning techniques, which has been made available in an open-source `R` package **StMoMo**, and can be extended and built upon with additional research.

We fitted our models using an ordinary least squares setting for computational efficiency. However, GAPC models are traditionally fitted under a Poisson or Binomial framework, which is particularly advantageous at higher ages (Brouhns, Denuit, and Vermunt 2002). As an intermediary step, we may consider using weighted least squares. However, we aim to develop a Poisson-fitting framework, taking advantage of emerging computational tools for the fitting of regularised generalised linear models, such as the **smurf** package (Devriendt et al. 2020). Furthermore, future research may also consider the use of more complex time series models for the period indices, such as ARIMA(p, d, q) models. Accordingly, the construction framework could be extended to not only select the optimal basis functions, but also the optimal time series specification of the period terms.

Finally, we have only used the MSE of the log mortality rates for model selection and evaluation. However, this can be adapted to other criteria, such as errors in survival probabilities, life expectancies, or implied annuity prices. Further, in many financial applications, forecast intervals are of greater interest than point estimates. Therefore, future research may consider interval-based measures, such as interval scorecards, which count the number of times the realised mortality rate falls within a confidence range (Gneiting and Raftery 2007).

Acknowledgments

This research is funded by Society of Actuaries Center of Actuarial Excellence 2016 Research Grant on Longevity Risk and the Australian Research Council Centre of Excellence in Population Ageing Research (CEPAR) project number CE110001029.

References

Aro, Helena, and Teemu Pennanen. 2011. "A user-friendly approach to stochastic mortality modelling." *European Actuarial Journal* 1: 151–67.

Atance, David, and Ana Debón. 2020. "A comparison of forecasting mortality models using resampling methods." *Mathematics*, 1–21. https://doi.org/10.3390/math8091550.

Barigou, Karim, Stéphane Loisel, and Yahia Salhi. 2021. "Parsimonious predictive mortality modeling by regularization and cross-validation with and without Covid-type effect." *Risks* 9 (5).

Bergmeir, Christoph, and José M. Benítez. 2012. "On the use of cross-validation for time series predictor evaluation." *Information Sciences* 191: 192–213. https://doi.org/10.1016/j.ins.2011.12.028.

Bergmeir, Christoph, Mauro Costantini, and José M. Benítez. 2014. "On the usefulness of cross-validation for directional forecast evaluation." *Computational Statistics and Data Analysis* 76: 132–43. https://doi.org/10.1016/j.csda.2014.02.001.

Bergmeir, Christoph, Rob Hyndman, and Bonsoo Koo. 2018. "A note on the validity of cross-validation for evaluating autoregressive time series prediction." *Computational Statistics and Data Analysis* 120: 70–83. https://doi.org/10.1016/j.csda.2017.11.003.

Biffis, Enrico. 2005. "Affine processes for dynamic mortality and actuarial valuations." *Insurance: Mathematics and Economics* 37 (3): 443–68. https://doi.org/10.1016/j.insmatheco.2005.05.003.

Blake, David, Andrew Cairns, and Kevin Dowd. 2008. "The birth of the life market." *Asia-Pacific Journal of Risk and Insurance* 3 (1): 6–36. https://doi.org/10.1080/152165499307332.

Booth, Heather, and Leonie Tickle. 2008. "Mortality modelling and forecasting: A review of methods." https://doi.org/10.1017/S1748499500000440.

Box, George, Gwilym Jenkins, and Gregory Reinsel. 1994. "Time Series Analysis - Forecasting and Control." *Prentice Hall New Jersey 1994*. https://doi.org/10.1016/j.ijforecast.2004.02.001.

Breheny, Patrick, and Jian Huang. 2015. "Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors." *Statistics and Computing* 25 (2): 173–87. https://doi.org/10.1007/s11222-013-9424-2.

Brouhns, Natacha, Michel Denuit, and Jeroen Vermunt. 2002. "A Poisson log-linear regression approach to the construction of projected life tables." *Insurance: Mathematics and Economics* 31: 373–93.

Cairns, Andrew, David Blake, and Kevin Dowd. 2006. "A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration." *The Journal of Risk and Insurance* 73 (4): 687–718.

Cairns, Andrew, David Blake, Kevin Dowd, Guy Coughlan, David Epstein, and Marwa Khalaf-Allah. 2011. "Mortality density forecasts: An analysis of six stochastic mortality models." *Insurance: Mathematics and Economics* 48 (3): 355–67. https://doi.org/10.1016/j.insmatheco.2010.12.005.

Cairns, Andrew, David Blake, Kevin Dowd, Guy Coughlan, David Epstein, Alen Ong, and Igor Balevich. 2010. "A quantitative comparison of stochastic mortality models using data from England and Wales and the United States." *North American Actuarial Journal* 13 (1).

Cairns, Andrew J. G., David Blake, Kevin Dowd, Guy D. Coughlan, David Epstein, and Marwa Khalaf-Allah. 2008. "Mortality density forecasts: An analysis of six stochastic mortality models." *Pensions Institute Discussion Paper PI-0801*. http://papers.ssrn.com/sol3/papers.cfm?abstract%7B/_%7Did=1340353.

Currie, Iain. 2006. "Smoothing and forecasting mortality rates with p-splines." *Https://Www.macs.-hw.ac.uk/Iain*.

———. 2016. "On fitting generalized linear and non-linear models of mortality." *Scandinavian Actuarial Journal*, 356–83. https://doi.org/10.1080/03461238.2014.928230.

———. 2020. "Constraints, the identifiability problem and the forecasting of mortality." *Annals of Actuarial Science* 14 (2): 537–66. https://doi.org/10.1017/S1748499520000020.

Currie, Iain, Maria Durban, and Paul Eilers. 2004. "Smoothing and forecasting mortality rates." *Statistical Modelling* 4: 279–98.

Delwarde, Antoine, Michel Denuit, and Paul Eilers. 2007. "Smoothing the Lee-Carter and Poisson log-bilinear models for mortality forecasting: A penalized log-likelihood approach." *Statistical Modelling* 7 (1): 29–48. https://doi.org/10.1177/1471082X0600700103.

Devriendt, Sander, Katrien Antonio, Tom Reynkens, and Roel Verbelen. 2020. "Sparse Regression with Multi-Type Regularized Feature Modeling." *Insurance: Mathematics and Economics*.

Dowd, Kevin, Andrew Cairns, David Blake, Guy Coughlan, David Epstein, and Marwa Khalaf-Allah. 2009. "Backtesting stochastic mortality models: An ex post evaluation of multiperiod-ahead density forecasts." *North American Actuarial Journal* 14 (3): 281–98.

Dowd, Kevin, Andrew J. G. Cairns, and David Blake. 2020. "CBDX: a workhorse mortality model from the Cairns–Blake–Dowd family." *Annals of Actuarial Science*, 1–16. https://doi.org/10.1017/s1748499520000159.

Fan, Jianqing, and Runze Li. 2001. "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American Statistical Association* 96 (456). https://doi.org/10.1198/016214501753382273.

Gneiting, Tilmann, and Adrian Raftery. 2007. "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association* 102: 359–78.

Green, Kesten, and J. Scott Armstrong. 2015. "Simple versus complex forecasting: The evidence." *Journal of Business Research* 68 (8): 1678–85. https://doi.org/10.1016/j.jbusres.2015.03.026.

Guibert, Quentin, Olivier Lopez, and Pierrick Piette. 2019. "Forecasting mortality rate improvements with a high-

dimensional VAR." *Insurance: Mathematics and Economics* 88: 255–72. https://doi.org/10.1016/j.insmatheco.2019.07.004.

Haberman, Steven, and Arthur Renshaw. 2011. "A comparative study of parametric mortality projection models." *Insurance: Mathematics and Economics* 48 (1): 35–55. https://doi.org/10.1016/j.insmatheco.2010.09.003.

Hainaut, Donatien, and Michel Denuit. 2020. "Wavelet-based feature extraction for mortality projection." *ASTIN Bulletin*, no. 2009: 1–33. https://doi.org/10.1017/asb.2020.18.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. https://doi.org/10.1007/b94608.

Holford, T R. 1983. "The estimation of age, period and cohort effects for vital rates." *Biometrics*. https://doi.org/10.2307/2531004.

Human Mortality Database. 2020. "University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany)." www.mortality.org.

Hunt, Andrew, and David Blake. 2014. "A General procedure for constructing mortality models." *North American Actuarial Journal* 18 (1): 116–38. https://doi.org/10.1080/10920277.2013.852963.

———. 2020a. "A Bayesian approach to modeling and projecting cohort effects." *North American Actuarial Journal* 0 (0): 1–20. https://doi.org/10.1080/10920277.2019.1649157.

———. 2020b. "Identifiability in age/period/cohort mortality models." *Annals of Actuarial Science* 14 (2): 550—–536. https://doi.org/10.1017/S1748499520000123.

———. 2020c. "Identifiability in age/period mortality models." *Annals of Actuarial Science*, 1–39. https://doi.org/10.1017/s1748499520000111.

———. 2020d. "On the structure and classification of mortality models." *North American Actuarial Journal* 0 (0): 1–20. https://doi.org/10.1080/10920277.2019.1649156.

Hunt, Andrew, and Andrés M. Villegas. 2015. "Robustness and convergence in the Lee-Carter model with cohorts." *Insurance: Mathematics and Economics* 64: 186–202.

Hyndman, Rob, Heather Booth, Leonie Tickle, and John Maindonald. 2017. "Package 'demography'." http://cran.r-project.org/package=demogra.

James, Gareth, Daniela Witen, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning with Applications in R*. https://doi.org/10.1016/j.peva.2007.06.006.

Karanikolos, Marina, David A. Leon, Peter C. Smith, and Martin McKee. 2012. "Minding the gap: Changes in life expectancy in the Baltic States compared with Finland." *Journal of Epidemiology and Community Health* 66 (11): 1043–9. https://doi.org/10.1136/jech-2011-200879.

Lee, Ronald, and Lawrence Carter. 1992. "Modeling and forecasting U.S. Mortality." *Journal of the American Statistical Association* 87 (419): 673–74.

Li, Han, and Colin O'Hare. 2017. "Semi-parametric extensions of the Cairns–Blake–Dowd model: A one-dimensional kernel smoothing approach." *Insurance: Mathematics and Economics* 77: 166–76.

Li, Hong, and Yanlin Shi. 2021. "Mortality forecasting with an age-coherent sparse VAR model." *Risks* 9 (35). https://doi.org/10.1017/asb.2020.39.

Li, Johnny S. H., Rui Zhou, Yanxin Liu, George Graziani, R Dale Hall, Jennifer Haid, Andrew Peterson, and Laurence Pinzur. 2020. "Drivers of mortality dynamics: Identifying age/period/cohort components of historical U.S. mortality improvements." *North American Actuarial Journal* 24 (2): 228–50. https://doi.org/10.1080/10920277.2020.1716808.

Li, Johnny Siu Hang, Rui Zhou, and Mary Hardy. 2015. "A step-by-step guide to building two-population stochastic mortality models." *Insurance: Mathematics and Economics* 63: 121–34. https://doi.org/10.1016/j.insmatheco.2015.03.021.

Macdonald, Angus S, Stephen J Richards, and Iain D Currie. 2018. *Modelling Mortality with Actuarial Applications*. Cambridge University Press.

Plat, Richard. 2009. "On stochastic mortality modeling." *Insurance: Mathematics and Economics* 45 (3): 393–404. https://doi.org/10.1016/j.insmatheco.2009.08.006.

Racine, Jeff. 2000. "Consistent cross-validatory model-selection for dependent data: hv-block cross-validation." *Journal of Econometrics* 99 (1): 39–61. https://doi.org/10.1016/S0304-4076(00)00030-0.

Renshaw, Arthur, and Steven Haberman. 2003. "On the forecasting of mortality reduction factors." *Insurance: Mathematics and Economics* 32: 379–401. https://doi.org/10.1016/S0167-6687(03)00118-5.

———. 2006. "A cohort-based extension to the Lee-Carter model for mortality reduction factors." *Insurance: Mathematics and Economics* 38 (3): 556–70. https://doi.org/10.1016/j.insmatheco.2005.12.001.

Tashman, Leonard J. 2000. "Out-of-sample tests of forecasting accuracy: An analysis and review." *International Journal of Forecasting* 16 (4): 437–50. https://doi.org/10.1016/S0169-2070(00)00065-0.

Tibshirani, Robert. 1996. "Regression selection and shrinkage via the lasso." *Journal of the Royal Statistical Society* 58 (1): 267–88. https://doi.org/10.2307/2346178.

Vandekerckhove, Joachim, Dora Matzke, and Eric-Jan Wagenmakers. 2015. "Model comparison and the principle of parsimony." *The Oxford Handbook of Computational and Mathematical Psychology*. https://doi.org/10.1093/oxfordhb/9780199957996.013.14.

Venter, Gary, and Şule Şahin. 2018. "Parsimonious parameterization of age-period-cohort models by Bayesian shrinkage." *ASTIN Bulletin* 48 (1): 89–110. https://doi.org/10.1017/asb.2017.21.

Villegas, Andrés, Vladimir Kaishev, and Pietro Millossovich. 2018. "StMoMo: An R package for stochastic mortality modeling." *Journal of Statistical Software* 84 (3). https://doi.org/10.18637/jss.v084.i03.

World Health Organization. 2018. "Global Health Observatory (GHO) Data." *Http://Www.who.int/-Gho/Mortality*.

Yuan, Ming, and Yi Lin. 2006. "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society. Series B: Statistical Methodology*. https://doi.org/10.1111/j.1467-9868.2005.00532.x.

Zhang, Cun Hui. 2010. "Nearly unbiased variable selection under minimax concave penalty." *Annals of Statistics* 38 (2): 894–942. https://doi.org/10.1214/09-AOS729.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society* 67 (2): 301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Appendix A. Identifiability constraints for GAPC models

Table A.7 shows the constraints we use to identify the eight GAPC models tested in the paper. For LC, LC2, and CBDx we follow the approach discussed in Hunt and Blake (2020c), while for APC, sPLAT, cPLAT, and M7x we follow Hunt and Blake (2020b). For the RH model, we impose the constraints suggested by Hunt and Villegas (2015) in order to improve the stability and robustness of the model.

Table A.7: Identifiability constraints for GAPC models.

Model	**Constraints**
LC	$\sum_t \kappa_t^{(1)} = 0$, $\sum_x \beta_x^{(1)} = 1$
LC2	$\sum_t \kappa_t^{(1)} = 0$, $\sum_t \kappa_t^{(2)} = 0$, $\sum_t \kappa_t^{(1)}\kappa_t^{(2)} = 0$, $\sum_x \lvert\beta_x^{(1)}\rvert = 1$, $\sum_x \lvert\beta_x^{(2)}\rvert = 1$, $\sum_x \beta_x^{(1)}\beta_x^{(2)} = 0$
APC	$\sum_t \kappa_t^{(1)} = 0$, $\sum_c \gamma_c = 0$, $\sum_c c\gamma_c = 0$
RH	$\sum_t \kappa_t^{(1)} = 0$, $\sum_x \beta_x^{(1)} = 1$, $\sum_c \gamma_c = 0$, $\sum_c c\gamma_c = 0$
sPLAT	$\sum_t \kappa_t^{(1)} = 0$, $\sum_c \gamma_c = 0$, $\sum_c c\gamma_c = 0$
cPLAT	$\sum_t \kappa_t^{(1)} = 0$, $\sum_c \gamma_c = 0$, $\sum_c c\gamma_c = 0$, $\sum_c c^2\gamma_c = 0$
CBDx	$\sum_t \kappa_t^{(1)} = 0$, $\sum_t \kappa_t^{(2)} = 0$
M7x	$\sum_t \kappa_t^{(1)} = 0$, $\sum_t \kappa_t^{(2)} = 0$, $\sum_t \kappa_t^{(3)} = 0$, $\sum_c \gamma_c = 0$, $\sum_c c\gamma_c = 0$, $\sum_c c^2\gamma_c = 0$, $\sum_c c^3\gamma_c = 0$

Appendix B. Simulating datasets

The simulation of the datasets in Section 7 involves first simulating mortality rates from a specific GAPC model and then randomly generating the number of deaths from a Poisson distribution to reflect sampling variation, or noise. For illustration, here we outline a seven-step procedure to simulate data from the APC model:

$$\eta_{x,t} = \ln(\mu_{x,t}) = \alpha_x + \kappa_t + \gamma_c; \tag{B.1}$$

however, this procedure can be extended and applied to any GAPC model.

Step 1: Generate reasonable estimates for $\boldsymbol{\alpha}$, $\boldsymbol{\kappa}$, and $\boldsymbol{\gamma}$.

We first need to fit the generator model, from which mortality rates can be simulated. This is achieved by fitting an APC model to USA female data for the years 1960 to 2017, ages 20 to 89, and cohorts 1871 to 1997. Using the Poisson maximum likelihood approach embedded in the **R** package **StMoMo** by Villegas, Kaishev, and Millossovich (2018), we can estimate

$$\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_2 0, \ldots, \hat{\alpha}_{89}), \qquad \hat{\boldsymbol{\kappa}} = (\hat{\kappa}_{1960}, \ldots, \hat{\kappa}_{2017}), \qquad \hat{\boldsymbol{\gamma}} = (\hat{\gamma}_{1871}, \ldots, \hat{\gamma}_{1997}). \tag{B.2}$$

Step 2: Fit time series models to $\boldsymbol{\kappa}$ and $\boldsymbol{\gamma}$.

The simulation of mortality rates requires the projection of the period index, $\boldsymbol{\kappa}$, and cohort index, $\boldsymbol{\gamma}$. We assume that each follows a univariate random walk with drift

$$\kappa_t = \kappa_{t-1} + \delta_0 + \epsilon_t \qquad \text{and} \qquad \gamma_c = \gamma_{c-1} + \delta_1 + \varepsilon_c, \tag{B.3}$$

where ϵ_t and ε_c are Gaussian white noise processes with variance σ_ϵ^2 and σ_ε^2 respectively. With these models, we can simply and easily estimate $\hat{\delta}_0$ and $\hat{\delta}_1$.

Step 3: Simulate $\boldsymbol{\kappa}$ and $\boldsymbol{\gamma}$ over a 100-year horizon.

From the fitted time series models, we can then simulate values of the period index, $\dot{\kappa}_{2017+s}$, and the cohort index, $\dot{\gamma}_{1997+s}$, over the next 100 years ($s = 1, \ldots, 100$)

$$\dot{\kappa}_{2017+s} = \hat{\kappa}_{2017} + s\hat{\delta}_0 + \sum_{t=1}^{s} \epsilon_t \qquad \text{and} \qquad \dot{\gamma}_{1997+s} = \hat{\gamma}_{1997} + s\hat{\delta}_1 + \sum_{c=1}^{s} \varepsilon_c. \tag{B.4}$$

Step 4: Simulate mortality rates, $\dot{\mu}_{x,t}$, for 100-year horizon.

With this, we can calculate the associated value of the predictor over the 100-year horizon

$$\dot{\eta}_{x,2016+s} = \hat{\alpha}_x + \sum_{i=1}^{N} f^{(i)}(x)\dot{\kappa}^{(i)}_{2016+s} + \dot{\gamma}_{2016+s-x}. \tag{B.5}$$

These can then be converted into the age-specific forces of mortality (rates), $\dot{\mu}_{x,2016+s}$, by inverting the canonical log-link function

$$\dot{\mu}_{x,2016+s} = e^{\dot{\eta}_{x,2016+s}}. \tag{B.6}$$

Step 5: Generate exposure matrix, $\boldsymbol{E^c}$.

For simplicity, we assume that the exposure is time-invariant, $E^c_{x,t} \equiv E^c_x$. Therefore, $\boldsymbol{E^c}$ has 100 columns, each of which are identical. We assume that the age structure mirrors the observed exposures for the USA females in 2017, given by $E^c_{x,2017}$. Therefore

$$E^c_x = p \times \frac{E^c_{x,2017}}{\sum_x E^c_{x,2017}}, \tag{B.7}$$

where p denotes the total population size being simulated.

Step 6: Generate death count matrix, $\boldsymbol{D}$.

We then produce the matrix of death observations using the simulated central mortality rates by randomly drawing from a Poisson distribution

$$D_{x,2016+s} \sim \text{Poisson}(E^c_x \dot{\mu}_{x,2016+s}). \tag{B.8}$$

Step 7: Repeat procedure 100 times for the given combination of generator model and population.

Appendix C. Group regularisation in StMoMo

We have implemented the model construction framework introduced in this paper as an extension of the `R` package **StMoMo** (Villegas, Kaishev, and Millossovich 2018). The implementation of the group regularised fitting is underpinned by the use of the `R` package **grepreg** (Breheny and Huang 2015) which provides fast and stable group descent algorithms for general group regularisation purposes. In this Appendix, we illustrate the replication of some of the results associated with the USA males example presented in Section 5.

Currently, the group regularisation functions are available in the `GroupLasso` branch of the development version of **StMoMo**, available in Github at https://github.com/amvillegas/StMoMo/tree/GroupLasso. This branch will be integrated to the CRAN version of the package after peer review of this paper. The `GroupLasso` development version can be installed with the following commands:

```
install.packages("devtools")
devtools::install_github("amvillegas/StMoMo", ref = "GroupLasso")
```

The code below defines a base model with the $B = 37$ basis functions in Equations (44)-(48):

```
library(StMoMo)
strikes <- seq(25, 85, 5)
baseModel <- StMoMo(
  link = "log-Gaussian",
  staticAgeFun = TRUE,
  periodAgeFun = c("1", genPoly(1:10), genCall(strikes), genPut(strikes)),
  cohortAgeFun = "1"
)
```

The `genPoly`, `genCall` and `genPut` functions are utility functions to generate, respectively, basis functions, $f^{(i)}(x)$, of polynomial form (Equation (46)), call payoffs (Equation (47)), and put payoffs (Equation (48)). There is also the `genSpline` utility function to generate B-Spline basis such as those depicted in Figure 8.

To fit the models to USA male data, we need to download these data from the Human Mortality Database. This can be done using the package **demography** (Hyndman et al. 2017) with the code:

```
library(demography)
USAdata <- hmd.mx(country = "USA", username = username, password = password)
dataStMoMo <- StMoMoData(USAdata, series = "male")
dataStMoMo$Dxt <- round(dataStMoMo$Dxt)
```

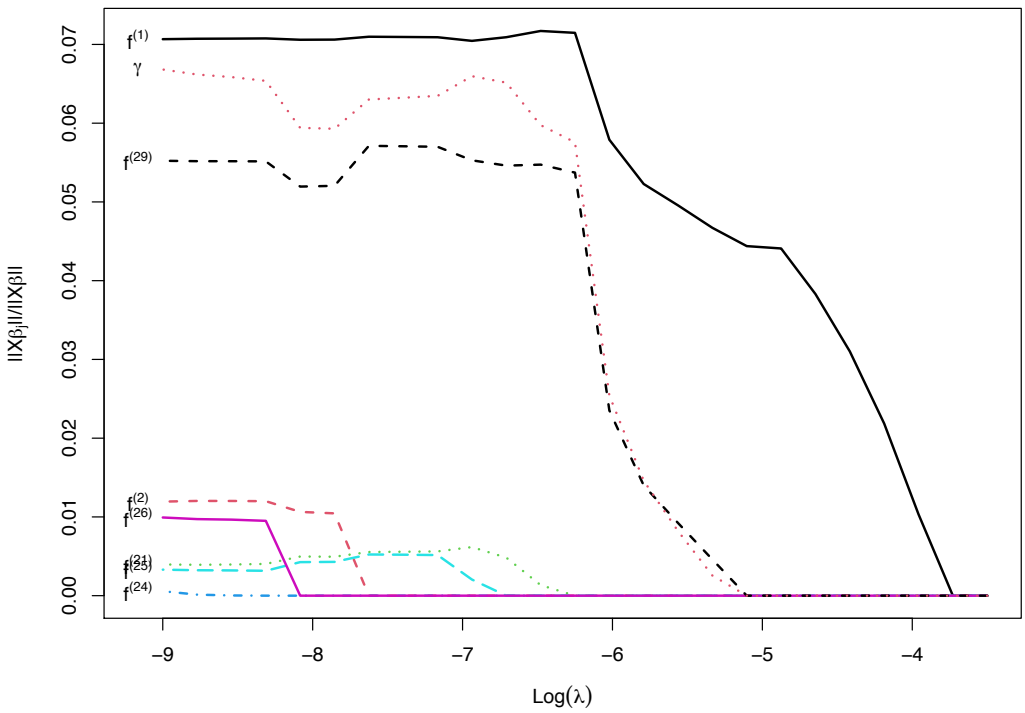
The `username` and `password` above are for the Human Mortality Database and should be replaced appropriately.

The main function for fitting a regularisation path is `grpfit`. We can use this function to fit the regularisation path induced by a grid of 25 values of λ in the range $\left[e^{-9}, e^{-3.5}\right]$ applied to data for the period 1960 to 2000 and the age range 20 to 89:

```
lambda <- exp(seq(-3.5, -9, length.out = 25))
GRfit <- grpfit(
  baseModel,
  lambda = lambda,
  data = dataStMoMo,
  ages.fit = 20:89,
  years.fit = 1960:2000
)
```

We can then plot the regularisation path as follows:

```
plot(GRfit, lwd = 2)
```

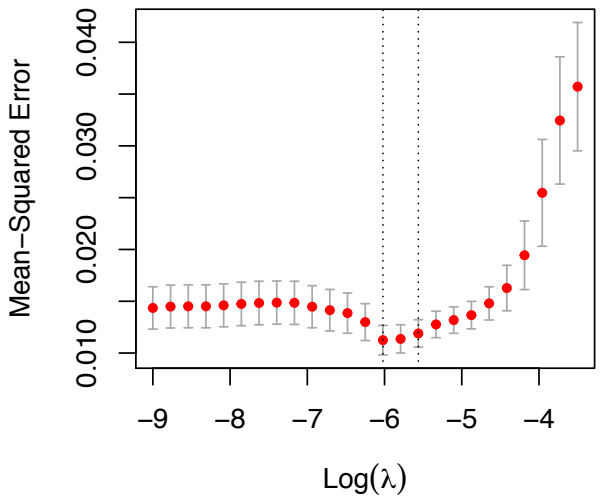


The cross validation framework described in Section 4.5 is implemented in function `cv.grpStMoMo`. For example, to select an appropriate level of regularisation for forecasting horizon $h = 10$, we can use function `cv.grpStMoMo` as follows:

```
GRCV_10 <- cv.grpStMoMo(
  baseModel,
  h = 10,
  lambda = lambda,
  data = dataStMoMo,
  ages.train = 20:89,
  years.train = 1960:2000
)
```

Since cross validation entails repeated refitting of the models, the above code can take a couple of minutes to run. We can then plot the cross validation error as a function of λ with the code:

```
plot(GRCV_10)
```

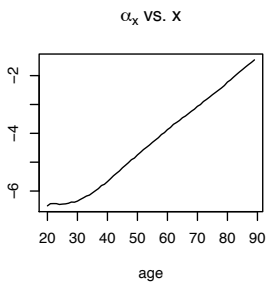


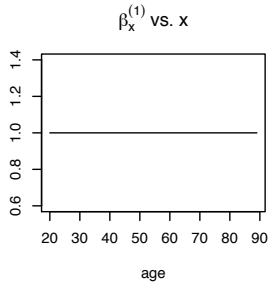
The value of λ with the minimum cross validation error is reported in `GRCV_10$lambda.min` and the corresponding index within the grid is reported in `GRCV_10$min`. This index can be used in combination with function `extractStMoMo` to extract the model that minimises the cross validation error at horizon $h = 10$:

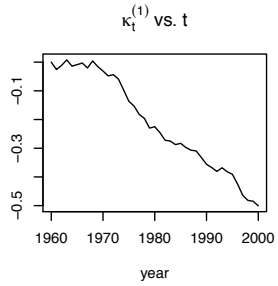
```
GRfit_10 <- extractStMoMo(GR1fit, k = GRCV_10$min)
```

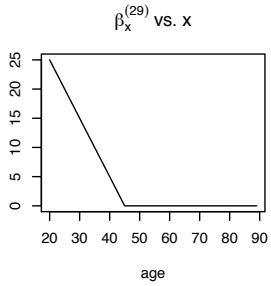
The output of `extractStMoMo` is a standard **StMoMo** object which can be used with the standard functions of **StMoMo** such as `plot`, `fitted`, `forecast` and `simulate`. For example, we can plot the selected model for horizon $h = 10$ as follows:

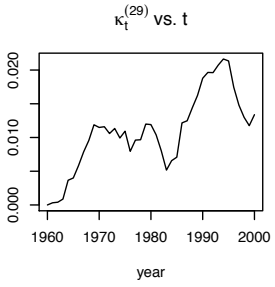
```
plot(GRfit_10, nCol = 4)
```

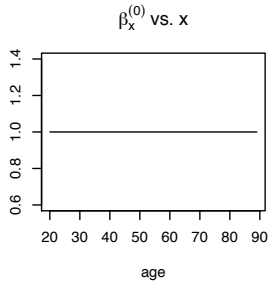

α_x vs. x
−2
−4
−6
20 30 40 50 60 70 80 90
age

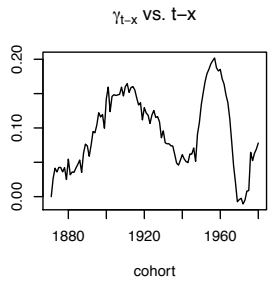

$\beta_x^{(1)}$ vs. x
1.4
1.2
1.0
0.8
0.6
20 30 40 50 60 70 80 90
age


$\kappa_t^{(1)}$ vs. t
−0.1
−0.3
−0.5
1960 1970 1980 1990 2000
year


$\beta_x^{(29)}$ vs. x
25
20
15
10
5
0
20 30 40 50 60 70 80 90
age


$\kappa_t^{(29)}$ vs. t
0.020
0.010
0.000
1960 1970 1980 1990 2000
year


$\beta_x^{(0)}$ vs. x
1.4
1.2
1.0
0.8
0.6
20 30 40 50 60 70 80 90
age


γ_{t-x} vs. t−x
0.20
0.10
0.00
1880 1920 1960
cohort