A comparative study of two-population models for the assessment of basis risk in longevity hedges

Andrés M. Villegas[1], **Steven Haberman**[2], **Vladimir K. Kaishev**[2], **and Pietro Millossovich**[2,3]

[1]**ARC Centre of Excellence in Population Ageing Research (CEPAR), UNSW Business School, University of New South Wales, Sydney, Australia**
[2]**Cass Business School, Faculty of Actuarial Science and Insurance, City, University of London, United Kingdom**
[3]**Department of Economics, Business, Mathematics and Statistics 'B. de Finetti', University of Trieste, Italy.**

ABSTRACT

Longevity swaps have been one of the major success stories of pension scheme de-risking in recent years. However, with some few exceptions, all of the transactions to date have been bespoke longevity swaps based upon the mortality experience of a portfolio of named lives. In order for this market to start to meet its true potential, solutions will ultimately be needed that provide protection for all types of members, are cost effective for large and smaller schemes, are tradable, and enable access to the wider capital markets. Index-based solutions have the potential to meet this need; however concerns remain with these solutions. In particular, the basis risk emerging from the potential mismatch between the underlying forces of mortality for the index reference portfolio and the pension fund/annuity book being hedged is the principal issue that has, to date, prevented many schemes progressing their consideration of index-based solutions. Two-population stochastic mortality models offer an alternative to overcome this obstacle as they allow market participants to compare and project the mortality experience for the reference and target populations and thus assess the amount of demographic basis risk involved in an index-based longevity hedge. In this paper, we systematically assess the suitability of several multi-population stochastic mortality models for assessing basis risks and provide guidelines on how to use these models in practical situations paying particular attention to the data requirements for the appropriate calibration and forecasting of such models.

1. INTRODUCTION

Recent years have seen a huge growth in longevity risk transfer, both in the insurer to reinsurer market, and from pension schemes to the insurance market. For example in 2014 £36.6bn of longevity risk was transferred from pension schemes to insurers and reinsurers via buy-ins, buy-outs and longevity swaps. Of this, £25.4bn related to longevity only transactions (longevity swaps), more than double the volume written in the preceding 3 years (Hymans Robertson LLP, 2015). An effective, growing market with sufficient capacity to meet demand would be to the benefit of all participants, whether to enable business to be done, or to manage risk.

To date most transactions have been "bespoke" deals, with the payouts linked directly to the actual experience or lifespans of the individuals being covered. But index-based solutions – where the payouts are linked to a longevity index or metric based on an external reference population – are possible. They have the potential to provide important benefits: lower costs, faster execution, potential for liquidity, and greater transparency.

In its simplest form an index based longevity swap involves a payment to the pension scheme or insurer that is based on the longevity experience of a reference index. An index-based swap provides a means to obtain (partial) protection from longevity risk both for pensioners but also deferred pensioners who are generally not covered by the "bespoke" transactions. In the case of life insurers they offer a potentially flexible way to manage exposure to longevity risk, or to facilitate a more capitally optimal balance between longevity and mortality risk. However, index-based swaps do not provide a perfect risk reduction due to the presence of basis risk, which arises from the differences in the mortality experiences of the reference population of the index and of the target population being hedged. As a result, the index based payments will not exactly match the actual annuity payments being made by the insurer or pension scheme.

There are three primary sources of basis risk driving the mismatch between the insurer or pension scheme liabilities and the longevity index hedge (LLMA, 2012):

- *Structuring risk* due to the payoff of the hedging instruments being different to that of the portfolio: for example the hedging instrument making annual payments whereas the portfolio pays annuities or pensions monthly, the hedge may be of shorter duration than the liabilities or it may contain some option-like features such as caps/floors or other non-linear payoff patterns.

- *Sampling risk* arising from the random outcomes in the mortality of the individual lives within the portfolio and the index population meaning the actual mortality experienced by the two populations will not be the same, other than by chance. The impact of sampling risk may be aggravated by concentration risk affecting the portfolio.

- *Demographic risk* owing to demographic and socio-economic differences in the composition of the actual portfolio being hedged and the index population referenced in the hedge, leading to different underlying mortality rates at the current moment – and in the future.

Well-established approaches for modelling the first two of these sources of basis risk exist. Structuring risk can be assessed by simulating the cashflows under the portfolio and the payoffs under the instrument, whilst sampling risk can be modelled by simulating the outcomes for the respective populations.

In contrast, there is no well-established approach for assessing demographic basis risk. Yet it is this risk which worries (re)insurers and pension schemes when they consider entering index-based longevity transactions (LLMA, 2012). The absence of an appropriate approach for quantifying such risk makes it very difficult to assess whether such a transaction looks good value for money, or what impact the transaction would have on the insurer's or pension scheme's overall risk profile and hence capital/funding requirements.

In the academic literature there have been a few contributions setting out possible approaches for quantifying longevity basis risk. Coughlan et al. (2011) propose a comprehensive framework for assessing the effectiveness of a longevity hedge, in which the first and key step entails a careful analysis of the historical experiences of the reference and target population to get an informed understanding of the mortality differences between the two populations. Li and Hardy (2011) investigate the use of a number of multipopulation extensions of the Lee-Carter model (Lee and Carter, 1992) for the assessment of basis risk and use the Augmented Common Factor model of Li and Lee (2005) to quantify the hedge effectiveness of an index-based q-forward longevity hedge. Li et al. (2015) propose a systematic approach for the construction of two-population mortality models that can be used for the quantification of the population basis risk in a standardised longevity hedge. In addition, recent years have seen a boom in the actuarial and demographic literature looking at the modelling of mortality in two (or more) related populations (e.g. Li and Lee (2005); Jarner and Kryger (2011); Plat (2009b); Cairns et al. (2011a); Dowd et al. (2011)). These two-population models, although not always proposed with the specific aim of assessing longevity basis risk, have the potential for allowing market participants to compare and project the mortality experience for the reference and target populations and thus assess the amount of demographic basis risk involved in an index-based longevity hedge. However, often the portfolio experience data will be sparse, posing a challenge for the accurate calibration and projection of the two-population model.

Our purpose in this paper is threefold. First, we provide a systematic and structured overview of existing multipopulation mortality modelling methodologies (c.f. Figure 1) scattered within the actuarial, demographic and statistical literature.

Our second goal is to summarize existing and formulate new criteria that a two-population mortality model should satisfy in order to be suitable for assessing basis risk.

Finally, our third goal is to systematically evaluate, contrast and select the model(s) that satisfy these criteria. We have done that by using prototype pension schemes with different size, history length and socio-economic composition. To the best of our knowledge, such a comprehensive analysis covering different characteristics of pension schemes and many alternative models has not been performed before. Our main finding is that two-populations mortality models are efficiently applied only if the scheme size exceeds 20,000-25,000 lives and its history length is at least 8-10 years. Given these conditions are satisfied, we found that the most appropriate models to be used for assessing basis risk are M7-M5 and CAE+Cohorts (see Table 3).

We believe that providing such an overview and comparison is an important contribution that will help researchers and industry practitioners interested in longevity risk modelling. Furthermore, we have shaped the framework under which basis risk assessment methodologies can reliably be used. Therefore, we have offered market participants involved in longevity transactions an invaluable analytical tool.

The paper is structured as follows. In Section 2 we introduce some notation. In Section 3 we provide an overview of the multipopulation mortality models that have been proposed in the literature. Then, to facilitate the comparison of models, we discuss in Section 4 a general modelling framework under which most two-population mortality models can be accommodated. In Section 5 we draw from the literature comparing single population mortality models to introduce a number of criteria that a good and practical two-population model for basis risk assessment should satisfy. We use these criteria in Section 6 to systematically evaluate the appropriateness of the possible two-population models for basis risk assessment. First, in Section 6.1, we evaluate the models against those criteria which relate to the theoretical properties of a model and can be evaluated without reference to a specific dataset. Then, in Section 6.2, we focus on those criteria which can only be evaluated after a model has been fitted to data. This systematic evaluation of the models will allow us to identify the main features of a good model for basis risk assessment and discuss the data requirements for the appropriate calibration and forecasting of such a model. Having identified some reasonable models for basis risk assessment, we examine in Section 7 the performance of these models in some simple illustrative hedge-effectiveness evaluation exercises, paying particular attention to the impact that different volumes of data may have on the assessment of basis risk. Finally, we conclude in Section 8 with a discussion of our main findings and future areas of research.

2. NOTATION

We denote by R the reference population backing the hedging instrument and by B the book population whose longevity risk is to be hedged. We assume that for the reference population the number of deaths at age x last birthday in calendar year t, D^R_{xt}, and the matching initial exposed to risk, E^R_{xt}, are available. The corresponding 1-year death rate for an individual in the reference population aged x last birthday and in calendar year t, denoted q^R_{xt}, can be estimated as $\hat{q}^R_{xt} = D^R_{xt} \big/ E^R_{xt}$. Similarly, the corresponding quantities for the book population are denoted D^B_{xt}, E^B_{xt} and $\hat{q}^B_{xt} = D^B_{xt} \big/ E^B_{xt}$. We assume that these data are available for a given set of ages and given numbers of years that can differ between the reference and the book populations. More precisely, we assume that D^R_{xt}, E^R_{xt} are available for consecutive ages $x = x_1, \ldots, x_l$ and consecutive calendar years $t = t_1, \ldots, t_{n_R}$, while in the book they are available for ages $x = x_1, \ldots, x_m$ and calendar years $t = u_1, \ldots, u_{n_B}$. Typically, data for the reference population will be available over a longer horizon than in the book, that is $n_R \geq n_B$. Also, the set of calendar years of data in the book may be a subset of the corresponding calendar years in the reference population i.e. we may find that $u_{n_B} \neq t_{n_R}$. Further, the ages available within the book may be a subset of those available in the reference population.

3. OVERVIEW OF AVAILABLE TWO-POPULATION MORTALITY MODELS

In order to be able to assess basis risk, we need a model that is able to capture the mortality trends in the reference population backing the hedging instrument and in the book population whose risk is to be hedged. That is to find a suitable two-population model for q^R_{xt} and q^B_{xt} which produces consistent stochastic forecasts of future mortality.

Many models have been proposed in the literature to represent the mortality evolution of two or more related populations. The majority of such models extend known single population models by specifying the correlation and interaction between the involved populations. Figure 1 contains a schematic representation of the multi-population models currently available in the published literature, broadly grouped according to three main categories, following the single population model they extend.

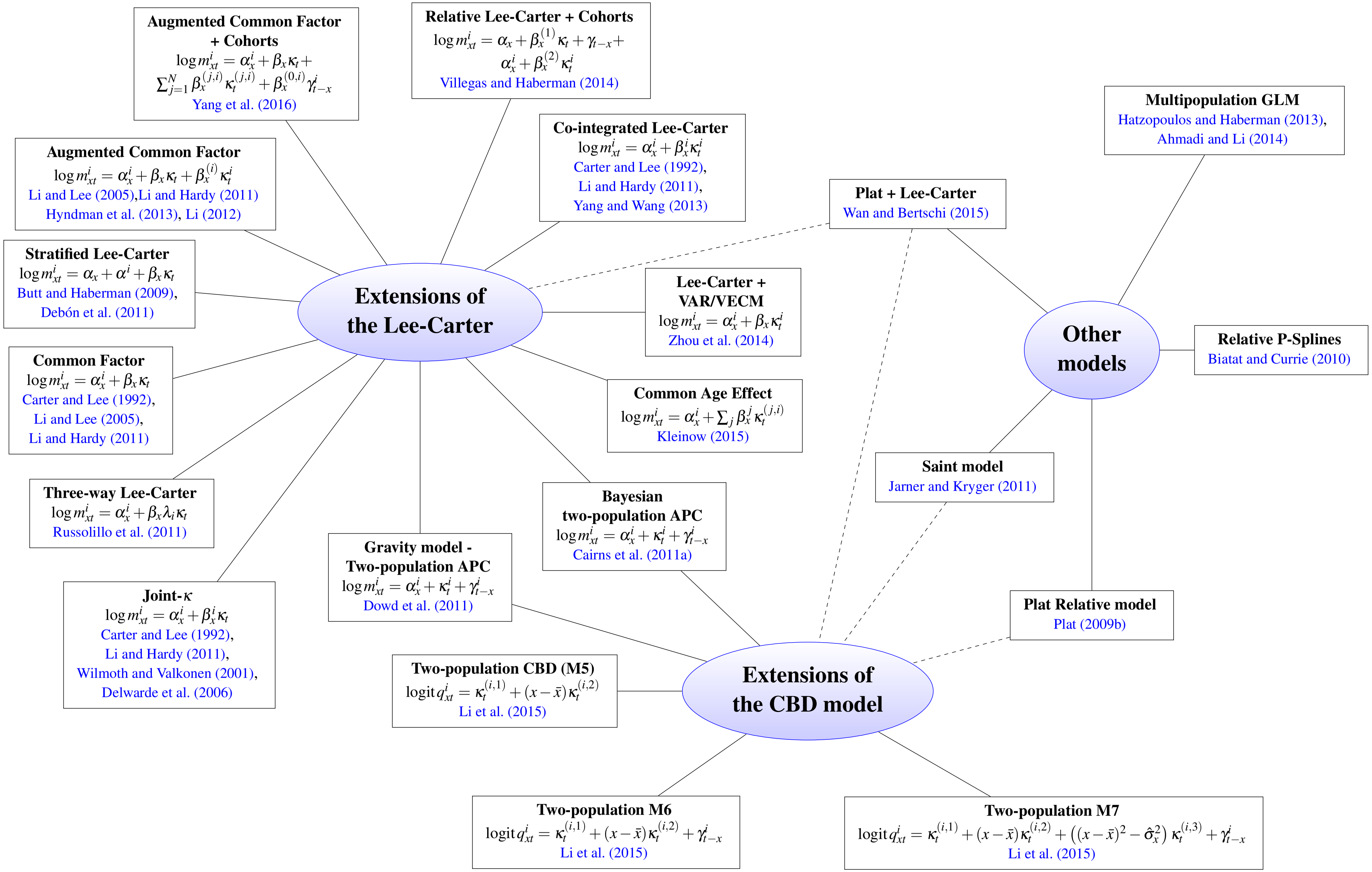

Extensions of the Lee-Carter
Extensions of the CBD model
Other models
Augmented Common Factor + Cohorts
$\log m^i_{xt} = \alpha^i_x + \beta_x\kappa_t + \sum_{j=1}^N \beta_x^{(j,i)}\kappa_t^{(j,i)} + \beta_x^{(0,i)}\gamma^i_{t-x}$
Yang et al. (2016)
Relative Lee-Carter + Cohorts
$\log m^i_{xt} = \alpha_x + \beta_x^{(1)}\kappa_t + \gamma_{t-x} + \alpha^i_x + \beta_x^{(2)}\kappa^i_t$
Villegas and Haberman (2014)
Co-integrated Lee-Carter
$\log m^i_{xt} = \alpha^i_x + \beta^i_x\kappa^i_t$
Carter and Lee (1992),
Li and Hardy (2011),
Yang and Wang (2013)
Multipopulation GLM
Hatzopoulos and Haberman (2013),
Ahmadi and Li (2014)
Augmented Common Factor
$\log m^i_{xt} = \alpha^i_x + \beta_x\kappa_t + \beta_x^{(i)}\kappa^i_t$
Li and Lee (2005),Li and Hardy (2011)
Hyndman et al. (2013), Li (2012)
Plat + Lee-Carter
Wan and Bertschi (2015)
Stratified Lee-Carter
$\log m^i_{xt} = \alpha_x + \alpha^i + \beta_x\kappa_t$
Butt and Haberman (2009),
Debón et al. (2011)
Lee-Carter + VAR/VECM
$\log m^i_{xt} = \alpha^i_x + \beta_x\kappa^i_t$
Zhou et al. (2014)
Relative P-Splines
Biatat and Currie (2010)
Common Factor
$\log m^i_{xt} = \alpha^i_x + \beta_x\kappa_t$
Carter and Lee (1992),
Li and Lee (2005),
Li and Hardy (2011)
Common Age Effect
$\log m^i_{xt} = \alpha^i_x + \sum_j \beta^j_x\kappa_t^{(j,i)}$
Kleinow (2015)
Saint model
Jarner and Kryger (2011)
Three-way Lee-Carter
$\log m^i_{xt} = \alpha^i_x + \beta_x\lambda_i\kappa_t$
Russolillo et al. (2011)
Bayesian two-population APC
$\log m^i_{xt} = \alpha^i_x + \kappa^i_t + \gamma^i_{t-x}$
Cairns et al. (2011a)
Gravity model - Two-population APC
$\log m^i_{xt} = \alpha^i_x + \kappa^i_t + \gamma^i_{t-x}$
Dowd et al. (2011)
Plat Relative model
Plat (2009b)
Joint-κ
$\log m^i_{xt} = \alpha^i_x + \beta^i_x\kappa_t$
Carter and Lee (1992),
Li and Hardy (2011),
Wilmoth and Valkonen (2001),
Delwarde et al. (2006)
Two-population CBD (M5)
$\operatorname{logit} q^i_{xt} = \kappa_t^{(i,1)} + (x-\bar{x})\kappa_t^{(i,2)}$
Li et al. (2015)
Two-population M6
$\operatorname{logit} q^i_{xt} = \kappa_t^{(i,1)} + (x-\bar{x})\kappa_t^{(i,2)} + \gamma^i_{t-x}$
Li et al. (2015)
Two-population M7
$\operatorname{logit} q^i_{xt} = \kappa_t^{(i,1)} + (x-\bar{x})\kappa_t^{(i,2)} + \left((x-\bar{x})^2 - \hat{\sigma}^2_x\right)\kappa_t^{(i,3)} + \gamma^i_{t-x}$
Li et al. (2015)

Figure 1. Overview of the multipopulation mortality modelling literature.

The first ideas for modelling multiple populations go back to the seminal work of Carter and Lee (1992), who suggested three possible ways of extending their single population model (Lee and Carter, 1992) in order to forecast differentials in US mortality between men and women. The first and simplest approach suggested by Carter and Lee (1992) is to use independent Lee-Carter models for each population, and, if desired, to study in a later stage the dependence between the population-specific period effects. A second approach, the Joint-κ model, assumes that a single period component κ_t drives the mortality change for all the populations but assumes that the age-specific mortality pattern and the age-specific responses to changes in the level of mortality are population-specific. The third approach estimates the populations jointly using cointegration techniques.

Formally, the Joint-κ model assumes that the central death rate at time t for age x in population i, m^i_{xt}, is given by

$$\log m^i_{xt} = \alpha^i_x + \beta^i_x \kappa_t. \tag{1}$$

Several other models proposed in the literature can be thought of as restricted versions of the Joint-κ model in Equation (1). These include: the Three-way Lee-Carter model of Russolillo et al. (2011) which assumes that $\beta^i_x = \beta_x \lambda^i$; the Common Factor model introduced by Li and Lee (2005) where $\beta^i_x = \beta_x$; and the stratified Lee-Carter model proposed in Butt and Haberman (2009) where it is assumed that $\alpha^i_x = \alpha_x + \alpha^i$ and $\beta^i_x = \beta_x$.

The structure of the Joint-κ model and of its restricted versions imply that mortality improvements are perfectly correlated across populations. Moreover, the Common Factor and Stratified Lee-Carter models imply the same mortality improvements for all population at all times. However, since this is an unrealistic assumption for most datasets, Li and Lee (2005) have added a population specific factor to the Common Factor model, in the so-called Augmented Common Factor model:

$$\log m^i_{xt} = \alpha^i_x + \beta_x \kappa_t + \beta^i_x \kappa^i_t. \tag{2}$$

In Equation (2) the term $\beta^i_x \kappa^i_t$ captures the deviations of the rate of mortality change of population i from the long-term trend in mortality change implied by the common factor, $\beta_x \kappa_t$. In order to avoid divergence in the projected mortality, Li and Lee (2005) assume that the κ^i_t factors can be modelled using stationary processes such as a first order autoregressive process, AR(1). Under this modelling assumption the mortality rates of the different populations may wander apart in the short and medium terms, but tend to converge in the long-run. The Augmented Common Factor has spawned several variants and extensions. Hyndman et al. (2013) have introduced the product-ratio method which extends the Augmented Common Factor model by adopting a functional data approach and allowing more than one period index for the modelling of both the common factor and of the population-specific factors. Li (2012), who also considers multiple period indexes, uses a Poisson setting to estimate the parameters of the Augmented Common Factor model instead of the singular value decomposition approach originally employed by Li and Lee (2005). Recently, Yang et al. (2016) have extended the Poisson Augmented Common Factor to allow for possible cohort effects. Villegas and Haberman (2014) have considered a similar cohort variant of the Augmented Common Factor for the purpose of studying socio-economic differences in mortality.

As discussed in Li and Hardy (2011), to implement a two-population version of the co-integrated Lee-Carter model suggested by Carter and Lee (1992), one must first fit two independent single population Lee-Carter models to each of the populations,

$$\log m^i_{xt} = \alpha^i_x + \beta^i_x \kappa^i_t, \quad i = 1,2, \tag{3}$$

and then model jointly the period effects of the populations, κ_t^1 and κ_t^2, with a co-integrated bivariate process under the assumption of the existence of a common stochastic long-term trend linking the mortality of the two populations. In the same vein, Yang and Wang (2013) fit independent single population Lee-Carter models to multiple populations and then model simultaneously the period effects of the different populations using a Vector Error Correction Model. In order to impose further consistency in the forecast of the two-populations, Zhou et al. (2014) assume in (3) that both populations share the same age-sensitivity term, i.e. $\beta_x^i = \beta_x$. For modelling the period indexes κ_t^1 and κ_t^2, Zhou et al. (2014) consider three methods: a random walk with drift for κ_t^1 plus an AR(1) for $\kappa_t^2 - \kappa_t^1$ (abbreviated RWAR by the authors), a vector autoregressive model (VAR); and a Vector Error Correction Model (VECM). Similarly to Zhou et al. (2014), Kleinow (2015) has proposed a multiple population Common-Age-Effect model in which the age-sensitivity terms (age-effects) are common to all the populations.

Another alternative for modelling multi-population mortality is to extend the widely used single-population Cairns-Blake-Dowd (CBD) model of mortality (Cairns et al., 2006). This approach has recently been considered by Li et al. (2015) who introduce two-population versions of the CBD model and its variants. For instance, in a full two-population version of the M7 model (the CBD model with cohort and quadratic effects proposed in Cairns et al. (2009)), the one-year death rate for a person aged x at time t in population i, q_{xt}^i, is given by:

$$\operatorname{logit} q_{xt}^i = \kappa_t^{(i,1)} + (x-\bar{x})\kappa_t^{(i,2)} + \left((x-\bar{x})^2 - \hat{\sigma}_x^2\right)\kappa_t^{(i,3)} + \gamma_{t-x}^i, \quad i = 1,2, \tag{4}$$

where $\bar{x}$ is the average age in the data and $\hat{\sigma}_x^2$ is the average value of $(x-\bar{x})^2$. Li et al. (2015) also set out a systematic top-down procedure to evaluate if some of the stochastic factors in the two-population model can be shared by the two populations (e.g. by assuming in (4) that $\kappa_t^{(1,j)} = \kappa_t^{(2,j)}$ for some $j \in \{1,2,3\}$ or that $\gamma_{t-x}^1 = \gamma_{t-x}^2$). For model forecasting Li et al. (2015) consider the same three approaches used by Zhou et al. (2014).

In two closely linked studies looking at the mortality dynamics of a pair of related populations, Cairns et al. (2011a) and Dowd et al. (2011) have proposed the use of a two-population version of the Age-Period-Cohort (APC) model:

$$\log m_{xt}^i = \alpha_x^i + \kappa_t^i + \gamma_{t-x}^i, \quad i = 1,2. \tag{5}$$

In both studies, the spreads between the state variables underlying the mortality models of each population are modelled as mean-reverting processes (e.g. an AR(1)) allowing different short-run trends in the mortality rates, but parallel long-run improvements. Cairns et al. (2011a) employ a Bayesian framework permitting a single stage estimation of the unobservable state variables and the parameters of the stochastic process driving them. Dowd et al. (2011) use a planetary analogy in which the mortality of the two populations are attracted to each other by a dynamic gravitational force dependent on the relative size of the populations.

There are other studies examining the joint modelling of two populations which do not lie under the category of pure extensions of the Lee-Carter or CBD models. Several of these studies pursue a relative approach whereby a single-population model is first fitted to one of the populations and then a separate model is fitted to the ratio of the mortality rates in the two populations. For instance, Jarner and Kryger (2011) have proposed a methodology for modelling the mortality experience of a small population in conjunction with the mortality experience of a much larger reference population. They assume that the reference population follows a deterministic long-run trend which is shared with the small population, and then model short term deviations of the small population from that trend using a multivariate stationary time series. Similarly, Wan and Bertschi (2015) model the larger population using the multi-factor single population model proposed by Plat (2009a) and then model the spread between the larger population and the smaller population with a three factor Lee-Carter model. In a related study, Plat (2009b) introduces a model for forecasting portfolio specific mortality alongside the relevant national population. In this model, portfolio specific mortality forecasts are obtained by combining national mortality projections derived from a standard single-population CBD model, with forecasts of the ratio between portfolio mortality rates and national population mortality rates. It is worth noting that Jarner and Kryger (2011), Wan and Bertschi (2015) and Plat (2009b) adopt the same approach for modelling the factors driving the dynamics of the mortality ratios and use a vector autoregressive model of order 1, VAR(1), so to avoid any long-term divergence of the mortality in the two populations.

Some authors have considered the use in a multipopulation setting of other well-known single population modelling approaches. For instance, Biatat and Currie (2010) extend to two populations the P-spline methodology (Currie et al., 2004) that has been successfully applied in the single population case, while Hatzopoulos and Haberman (2013) and Ahmadi and Li (2014) use the framework of generalised linear models (GLM) to obtain coherent morality forecasts for multiple populations.

4. MODELLING THE REFERENCE AND THE BOOK POPULATION: A GENERAL FORMULATION

Along the same lines of the general formulation of single population models considered in Hunt and Blake (2015b) and Villegas et al. (2015), we have identified a general framework under which most two population models that have been introduced in the literature can be accommodated. However, in order to facilitate the comparison between models, the way such models are proposed here may slightly differ from their original formulation.

As in Jarner and Kryger (2011), we choose a relative approach where the reference population is modelled first, and then the book mortality dynamics are specified so as to incorporate features from the reference. This relative approach allows a data mismatch between the reference and the book and is well suited to the usual situation of the reference population being considerably larger than the book population. Moreover, since single population models for the reference population are readily available and extensively studied, it allows the focus of modelling to be on making an informed decision for the book part of the model whilst retaining a good fit to the reference population.

4.1. Reference population

Following Villegas et al. (2015), a general model for the reference population can be written as[1]

$$D^R_{xt} \sim \mathrm{Bin}(E^R_{xt}, q^R_{xt}),$$

$$\mathrm{logit}\, q^R_{xt} = \alpha^R_x + \sum_{j=1}^{N} \beta_x^{(j,R)} \kappa_t^{(j,R)} + \gamma^R_{t-x}. \tag{6}$$

In Equation (6) the term α^R_x determines the reference mortality level for age group x; the integer N indicates the number of age-period terms describing the mortality trend for the reference population; each time index $\kappa_t^{(j,R)}$ contributes to specifying the reference mortality trend with each coefficient $\beta_x^{(j,R)}$ dictating how mortality in the corresponding age group x reacts to a change in the time index $\kappa_t^{(j,R)}$; and the term γ^R_{t-x} accounts for the cohort effect in the reference population.

4.2. Book population

Given the reference population model, the mortality of the book population is then specified through

$$D^B_{xt} \sim \mathrm{Bin}(E^B_{xt}, q^B_{xt}),$$

$$\mathrm{logit}\, q^B_{xt} - \mathrm{logit}\, q^R_{xt} = \alpha^B_x + \sum_{j=1}^{M} \beta_x^{(j,B)} \kappa_t^{(j,B)} + \gamma^B_{t-x}. \tag{7}$$

Note that we are modelling the difference in the (logit of) mortality in the book and the reference populations. Therefore, in Equation (7) the term α^B_x determines the mortality level differences of the book population compared to the reference population for age group x with the mortality level in the book being $\alpha^R_x + \alpha^B_x$; the integer M (generally less than or equal to N) indicates the number of age-period terms describing the mortality trend differences between the book population and the reference population; each time index $\kappa_t^{(j,B)}$ contributes in shaping the difference in mortality trends with each coefficient $\beta_x^{(j,B)}$ dictating how mortality differences for age group x react to a change in the time index $\kappa_t^{(j,B)}$; and the term γ^B_{t-x} accounts for the differences in cohort effect in the two populations, with the cohort effect in the book being $\gamma^R_{t-x} + \gamma^B_{t-x}$.

Depending on how the model is specified, identification constraints may have to be added to (6) and (7) to ensure uniqueness of the parameter estimates. The estimation of the parameters of the model can be performed using maximum likelihood in two stages whereby the reference population part of the model is estimated in a first stage and then, conditional on the reference population parameters, the book population part of the model is estimated in a second stage.[2]

[1] Here, we have chosen to work with one-year death probabilities, q_{xt}. Therefore, it is most natural to use the logit function and model deaths using a Binomial distribution. However, if interested in central death rates, m_{xt}, or the force of mortality, μ_{xt}, then the general modelling framework can be easily reformulated using a log link function and a Poisson Distribution. In addition, based on our experience, no material differences are to be expected in the analysis if central death rates, m_{xt}, or the force of mortality, μ_{xt}, were considered instead.

[2] An alternative approach would be to estimate simultaneously the parameters in the reference and book populations. This would in principle not materially change the fitted parameters as it is expected that the

4.3. Time series dynamics

The modelling is completed by specifying the dynamics of the period indices and the cohort terms which are needed for forecasting and simulating future mortality. Although alternatives have been explored by some authors (see e.g. Zhou et al. (2014)) for the choice of the time series used in the dynamics, we stick to those commonly used in the literature.

Starting with the reference population, we assume that the period index is modelled as a multivariate random walk with drift (MRWD)

$$\kappa_t^R = \mathbf{d} + \kappa_{t-1}^R + \xi_t^R, \qquad \xi_t^R \sim N(0, \Sigma^R), \qquad \kappa_t^R = \left(\kappa_t^{(1,R)}, \ldots, \kappa_t^{(N,R)}\right)',$$

and that the cohort index is modelled as an integrated auto-regressive process ARIMA(1, 1, 0)

$$\Delta\gamma_c^R = \phi_0 + \phi_1 \Delta\gamma_{c-1}^R + \varepsilon_c^R, \qquad \varepsilon_c^R \sim N(0, \sigma_R^2),$$

where $\mathbf{d}$ is an N-dimensional vector of drift parameters; $\Delta\gamma_c^R$ denotes $\gamma_c^R - \gamma_{c-1}^R$ with $c = t - x$; ϕ_0 and ϕ_1 are the drift and autoregressive parameters associated with the cohort effect γ_c^R; and Σ^R is the $N \times N$ variance-covariance matrix of the multivariate white noise ξ_t^R.

As for the book population, we follow the assumption commonly made in the literature (Li and Lee, 2005; Plat, 2009b; Cairns et al., 2011a; Jarner and Kryger, 2011; Li and Hardy, 2011; Hyndman et al., 2013; Wan and Bertschi, 2015). More precisely we assume that in the long-run the two populations experience similar mortality improvements and therefore model the spread in the time indexes and cohort effects as stationary processes:

$$\kappa_t^B = \Phi_0 + \Phi_1 \kappa_{t-1}^B + \xi_t^B, \qquad \xi_t^B \sim N(0, \Sigma^B), \qquad \kappa_t^B = \left(\kappa_t^{(1,B)}, \ldots, \kappa_t^{(M,B)}\right)', \tag{8}$$

$$\gamma_c^B = \psi_0 + \psi_1 \gamma_{c-1}^B + \varepsilon_c^B, \qquad \varepsilon_c^B \sim N(0, \sigma_B^2),$$

where Φ_0 and Φ_1 are an M-dimensional vector and an $M \times M$ matrix of model parameters; Σ^B is the $M \times M$ variance-covariance matrix of the multivariate white noise ξ_t^B; and ψ_0 and ψ_1 are parameters associated to the cohort spread γ_c^B. Thus:

- The time indices κ_t^B are modelled as a vector auto-regressive process of order 1 (VAR(1)), for which we assume that the eigenvalues of the matrix Φ_1 are smaller than 1 in absolute value.

- The cohort difference γ_c^B follows an AR(1) process for which we assume that $\psi_1 < |1|$.

- We are assuming independence of the time series determining the reference population and those determining the difference between the reference and the book populations.[3]

Overall, the time series dynamics approach considered here corresponds to the RWAR approach discussed in Zhou et al. (2014) and Li et al. (2015).

book population has a small size relative to the reference population. Furthermore, for some models such as the two-population APC, CBD, M5 and M7 where the log-likelihood is separable, a two-stage estimation approach results in exactly the same parameter estimates as a joint estimation approach.

[3] Considering correlations between ξ_t^R and ξ_t^B or between ε_c^R and ε_c^B is in principle possible, as has been done in Cairns et al. (2011a) and in Li et al. (2015). However, we refrain from considering this due to the fact that the estimation of the appropriate covariance matrix may not be straightforward. This is the case

5. MODEL SELECTION CRITERIA

With over 20 two-population models currently proposed in the literature (see Figure 1), our main goal is to identify which model(s) are most likely to provide a satisfactory solution for assessing basis risk. In order to support this analysis, it is useful to test each model against certain criteria that a good and practical two-population model for basis risk assessment should satisfy. Building on the literature comparing single population models (e.g. Continuous Mortality Investigation (2007); Cairns et al. (2008, 2009, 2011b); Haberman and Renshaw (2011)), we consider the following criteria. The model should:

1. Produce a *non-perfect correlation* between mortality rates in the two populations.
2. Produce a *non-perfect correlation* between year-on-year changes in mortality at different ages.
3. Permit the *generation of sample paths* and the calculation of prediction intervals.
4. Have a structure that allows the incorporation of *parameter uncertainty* in simulations.
5. Permit the consideration of a *cohort effect* if necessary.
6. Be *compatible with the data* that are likely to be available when doing basis risk exercises.
7. Be *straightforward to implement* using standard statistical methods likely to be available to practitioners.
8. Be *transparent* enough so that the model assumptions, limitations and outputs are understood by the users and can be easily explained to non-experts.
9. Show a reasonable *goodness-of-fit* to historical data in both the reference population and the book population for a wide range of book populations.
10. Show a reasonable *goodness-of-fit* for metrics involving the two populations such as differences or ratios in mortality rates or life expectancies for a wide range of book populations.
11. Be relatively *parsimonious*.
12. Produce *plausible and reasonable central projections* of both single-population and two-population metrics.
13. Produce *plausible and reasonable forecast level of uncertainty* in projections of both single-population and two-population metrics, which are in line with historical levels of variability.

when the models involve multiple period effects; for example if the two-population M7 model defined in Equation (4) is used, then the covariance matrix is particularly large, containing up to 21 distinct elements associated with 6 period factors (see Li et al. (2015)). In addition, the estimation of the covariance matrix would be further complicated in the case when the time series for the reference and the book have different lengths, which is very frequent in practice.

14. Produce parameter estimates and model forecasts that are *robust* relative to the period of data and range of ages employed.

Most of the above criteria coincide with the criteria that a good single population model should satisfy; we thus refer the reader to Continuous Mortality Investigation (2007, Section 8) and Cairns et al. (2008, Section 3) for a detailed discussion of their relevance. By contrast, criteria 1, 10, 12 and 13, referring to correlations between the mortality rates in the two populations and to the performance of the models in relation to two-population metrics, are new. The latter criteria are of prime importance to the application of two-population models in the assessment of basis risk in standardised longevity hedges. On the one hand, if a model assumes a perfect correlation between mortality rates in the two populations then it will imply that the reference population provides a perfect match for the book population, trivially leading to no (or very little) demographic basis risk. On the other hand, since demographic basis risk emerges from the mismatch in the mortality of the reference and the book population, it is critical that the two-population model shows a good fit to metrics involving the two populations, and that forecast levels of uncertainty and central trajectories for these metrics are plausible and consistent with historical differences between the populations.

We note however, that a two-population model which might not be suitable for basis risk assessment, may be an appropriate model for other applications in which some of the above criteria would be superfluous. For example, consider the case of valuing the liabilities of a pension book with sparse data, where we may consider a two-population model to borrow information from a larger reference population with the objective of improving the accuracy in the projections of the pension schemes' mortality. In this situation, having a non-perfect correlations between the mortality of the two populations would be unnecessary and the performance of the model relative to two-population metrics would be of lesser importance.

6. IDENTIFYING AN APPROPRIATE TWO-POPULATION MODEL

Given the wealth of models available and the large number of criteria, we have followed a two-stage filtering process to identify the model structures likely to be suitable for basis risk assessment. In a first stage, we focus on criteria 1 to 8 which refer to theoretical properties of a model and can be evaluated without reference to a specific dataset. Then, in a second stage, we focus on criteria 9 to 14 which can only be evaluated after a model has been fitted to data. More specifically, in the second stage of filtering we evaluate the goodness of fit, the reasonableness of the output, the forecasting performance and the robustness of those models which pass the first stage of filtering.

6.1. Stage 1 of filtering: Criteria requiring no data to assess

We first evaluate all the candidate models against those criteria that can be assessed independently of data or the actual fitting of the models. This process permits the identification of a number of models which could be rejected, either because their theoretical properties are not suitable for basis risk assessment or because they are unlikely to be accessible to the wider industry.

6.1.1. Non-perfect correlation between mortality rates in the two populations

A perfect correlation between the mortality rates q_{xt}^{B} and q_{xt}^{R} implies that the two populations move in tandem, with changes in the mortality of the book population matched by changes in the mortality of the reference population.[4] This will result in the model spuriously suggesting that there is no (or very little) demographic basis risk. This is the case for those Lee-Carter based models with a single common period effect for both populations, leading us to view the Stratified Lee-Carter, the Common Factor Model, the Three-way Lee-Carter, and the Joint-κ model as inadequate models for assessing demographic basis risk.

6.1.2. Non-perfect correlation between year-on-year changes in mortality at different ages

This criterion refers to the correlation between $q^R_{x,t+1} - q^R_{x,t}$ and $q^R_{y,t+1} - q^R_{y,t}$ (or between $q^B_{x,t+1} - q^B_{x,t}$ and $q^B_{y,t+1} - q^B_{y,t}$) for $x \neq y$. As noted by Cairns et al. (2008), a model that assumes a perfect correlation between changes in mortality at different ages would incorrectly suggest that holding a derivative instrument linked to a single age would provide just as good a hedge as holding several instruments linked to a range of different ages. Disregarding this issue can result in a misassessment of the structuring basis risk underlying a longevity hedge.

Lee-Carter type models with a single period effect and no cohort effect, such as the Cointegrated Lee-Carter and the Lee-Carter+VAR/VECM, have a trivial age correlation structure. In addition, the two-population APC model in Equation (5) implies that there is perfect correlation at all ages except at the youngest ages, where there is potentially additional randomness arising from the arrival of new cohorts with an unknown cohort effect (see Cairns et al. (2009)). In contrast, two-population extensions of the CBD model allow for imperfect correlations between annual changes in mortality at different ages due to the presence of multiple period factors.

We do not discard however any model due to its age correlation structure for two reasons. In many instances it may only be required to perform an indicative assessment of the demographic basis risk associated with an index-based hedge, without necessarily considering in detail the precise structuring of the hedge. Further, in order to assess model risk, it may be useful to consider an alternative model to the one used in structuring the hedge.

6.1.3. Generation of sample paths

Mortality sample paths are required for the assessment of the uncertainty in the cash-flows of a mortality-linked security as well as for the pricing and structuring of a longevity hedge. A distinguishing feature of the P-Spline model of Biatat and Currie (2010) and of the multipopulation GLM of Ahmadi and Li (2014) is that they assume that mortality follows a deterministic time trend, meaning that these models cannot generate sample paths. Hence, we do not consider these two models any further.

6.1.4. Parameter uncertainty

Given that in most cases the amount of data for the book populations is limited, the parameters of the models may be subject to significant estimation error. It is thus important to be able to consider the impact that parameter risk can have on forecasts levels of uncertainty and on hedge effectiveness. With the exception of the Bayesian two-population

[4] Note that the correlation between the mortality rates q^B_{xt} and q^R_{xt} may not be perfect, although it will be close to one, even when correlation is perfect on the logit scale used by the models introduced in Section 4. Also note that having a perfect correlation between the populations does not necessarily imply that he two populations experience exactly the same mortality improvements. For instance, the Joint-κ and the Three-way Lee-Carter models allow for improvement rate differentials, but imply a perfect correlation between the populations.

model of Cairns et al. (2011a) which naturally accounts for parameter uncertainty, none of the studies we have reviewed considers parameter uncertainty. Nevertheless, for most of the models it is possible to incorporate parameter uncertainty using bootstrapping techniques such as the ones proposed in Brouhns et al. (2005), Koissi et al. (2006) and Renshaw and Haberman (2008). We should mention however that, unlike the Bayesian framework, bootstrapping is related to the effect of sampling variation in the data. Therefore by means of bootstrapping it is not possible to assess the parameter uncertainty arising from the time series processes, but rather only that due to sampling variation.

6.1.5. Cohort effect

For some countries, including England and Wales, it is important that models allow for the now well-accepted cohort effect, separating out general improvements over time to those specific to a given birth cohort. Although not all the models include a cohort effect, they can in principle be extended to include such an effect.

6.1.6. Compatibility with available data

The data requirements of some of the models are incompatible with the likely available data. For instance, it is unlikely that the book population will provide the same length of history as the reference population, hindering the application of models which cannot easily deal with such a scenario. In particular, this requirement leads to the rejection of two further Lee-Carter based models, namely the Lee-Carter VAR/VECM and the Co-integrated Lee-Carter.

6.1.7. Ease of implementation and transparency

Ease of implementation and transparency are essential for a model to be of general use by practitioners. Accordingly, these two criteria lead to the rejection of several other models. In particular, the Multipopulation GLM of Hatzopoulos and Haberman (2013) is considered to be impractical for basis risk assessment as it is a complex model which is computationally involved to implement and may be difficult to communicate to non-experts. In addition, we disregard the Plat+Lee-Carter model of Wan and Bertschi (2015) (apart from other reasons discussed later) because it combines a parametric structure for the reference with a non-parametric structure for the book, and we believe that for the sake of interpretability of the parameters both parts of the model should be within the same class of models. Finally, although the Bayesian two-population APC model of Cairns et al. (2011a) is particularly amenable to the short history and modest exposures sizes of most book datasets, the implementation and transparency issues related to the underlying Bayesian approach have led us to rule out this model. However, some of the features of the approach of Cairns et al. (2011a) will still be investigated subsequently in this paper through a maximum-likelihood implementation of the two-population APC model.

6.2. Stage 2 of filtering: Criteria requiring data to assess

After carrying out the initial data-independent assessment, the following 10 models can be identified as candidates which are worth testing against the data dependent criteria: the Augmented Common Factor model and its cohort extension, the Relative Lee-Carter model with cohorts, the Common Age Effect Model, the two-population APC (Gravity model), the two-population M5, the two-population M6, the two-population M7, the Saint model, and the Plat relative model.

The second stage of filtering entails the evaluation of the historical goodness-of-fit, the (subjective) evaluation of the reasonableness of the forecast level of uncertainty produced

Table 1. Description of the book datasets used for model testing. Q1 represents the least deprived quintile of England and Q5 the most deprived quintile.

Dataset	Description	Percentage of exposure by IMD quintile				
		Q1	Q2	Q3	Q4	Q5
Typical Lives	This is the typical IMD split we would expect to see in a book population weighted by lives (head-count)	23%	22%	21%	20%	14%
Typical Amounts	This uses the same split as the typical (lives) but weighted by individual pension amounts to approximate the effect of a typical portfolio's liability distribution amongst the IMDs	30%	25%	20%	15%	10%
Extreme Wealthy	This reflects the split by IMD (on an amounts weighted basis) that we would expect to see in a very affluent book population	45%	30%	20%	5%	0%
Extreme Deprived	This reflects the split by IMD (on a lives weighted basis) that we would expect to see in a book skewed towards lower socio-economic groups	10%	15%	15%	25%	35%

by the models, and the evaluation of the forecasting performance and robustness of the models.

6.2.1. Data

The evaluation of the criteria in this stage requires data for model fitting. We have used as the reference population data the England and Wales male mortality experience as obtained from the Human Mortality Database (2013). For the purposes of our analysis we have focused on a subset of these data covering calendar years 1961-2010 and those older ages most relevant to longevity hedging, namely ages 60-89.

For the book population we use synthetic datasets generated based on England mortality data by quintiles of the Index of Multiple Deprivation 2007 (IMD 2007)[5] and the socio-economic composition observed within individual occupational pension schemes of the Club Vita dataset.[6] Specifically, the synthetic datasets used throughout this paper have been generated by randomly sampling from the national IMD data to obtain a dataset of exposure size, history length, and IMD profile desired. The technical details of this data sampling process are described in Appendix A. The use of synthetic data as opposed to actual pension scheme data facilitates a more thorough assessment of the models. Concretely, synthetic datasets permit us to control some key characteristics of the book population data while changing others. For instance, it allows us to vary the history length and exposure size of the book data whilst keeping the socio-economic and age composition constant. Moreover, synthetic datasets let us rely on the longer history of the national IMD mortality data to perform backtesting exercises such us those described in Section 6.2.7.

[5] A detailed analysis of the mortality data used in this paper can be seen in Villegas and Haberman (2014) or in Lu et al. (2014). For further information on the Index of Multiple Deprivation see Noble et al. (2007).

[6] Club Vita is an organisation which provides longevity analytics to pension schemes. The schemes in the Club Vita dataset span a wide range of sizes including some of the largest DB schemes in the UK and (as at September 2014) consists of nearly 6 million member records.

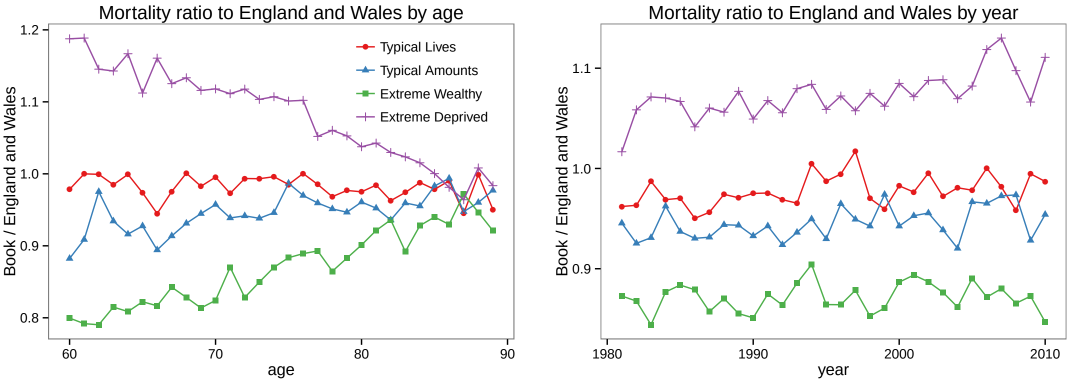


Figure 2. Ratio of the mortality in each of the four synthetic book datasets to the mortality in England and Wales. The left graph shows this ratio by age while the one on the right presents the time evolution of this ratio.

For the assessment of the goodness-of-fit of the models, we consider four different synthetic datasets to reflect the variety of socio-economic mixes observed in real pension schemes and annuity books. In each case, the socio-economic splits are motivated by the profiles seen within the Club Vita dataset. Table 1 describes the socio-economic profiles of these datasets. In all cases we use sample books with historical exposures of 100,000 male lives per year, which we believe is reasonable proxy for the largest exposure any pension scheme or insurer is likely to have. We also assume that book data are available for the period 1981-2010 and ages 60 to 89. Finally, we use the age distribution of the English population to split by age the total exposure of each of the sample schemes.

Figure 2 depicts the ratio of the mortality in each of the four datasets to the mortality in England and Wales. We note that the ordering of the ratios in the four datasets is consistent with their socio-economic mixes: the “Extreme Wealthy” dataset has below average mortality (ratio < 1), the “Extreme Deprived” dataset has above average mortality (ratio > 1), and the “Typical Lives” and “Typical Amounts” datasets exhibit a mortality ratio close to 1 due to the similarity of their socio-economic mix with that of England and Wales. It is also worth noticing that none of the datasets shows any very marked increasing or decreasing time trend in the mortality ratios, albeit there is a slight upward trend in the “Extreme Deprived” dataset. This is consistent with the slower mortality improvements for the two most deprived quintiles of England reported by Villegas and Haberman (2014).

6.2.2. Model fitting

To facilitate the fitting of the 10 models that passed our first-stage filtering, we have followed the general modelling framework described in Section 4 whereby each model can be viewed as a model for the reference population combined with a model for the book population (or perhaps more accurately, a model for the mortality ratio between reference and book). As such, the fitting and the assessment of the goodness-of-fit of a model can be carried out in two stages: fitting and assessing the goodness-of-fit of the reference model, followed by the fitting and the assessment of the goodness-of-fit of the book part of the model.[7] We note that conclusions regarding the goodness-of-fit of the model to

[7] All the model fitting performed in this paper has been carried out using the **R** package StMoMo (Villegas et al., 2015) which facilitates the implementation of stochastic mortality models using the unifying framework of generalised (non)linear models.

Table 2. Mathematical description of the models selected for the reference population.

Model	**Formula**
LC + Cohorts	$\operatorname{logit} q_{xt}^R = \alpha_x^R + \beta_x^R \kappa_t^R + \gamma_{t-x}^R$
APC	$\operatorname{logit} q_{xt}^R = \alpha_x^R + \kappa_t^R + \gamma_{t-x}^R$
M7	$\operatorname{logit} q_{xt}^R = \kappa_t^{(1,R)} + (x-\bar{x})\kappa_t^{(2,R)} + \left((x-\bar{x})^2 - \hat{\sigma}_x^2\right)\kappa_t^{(3,R)} + \gamma_{t-x}^R$

the reference may lead us to slightly modifying the original formulation of certain of the two-population models before assessing the goodness-of-fit of the book part of the model. The specific modifications for each particular two-population model are described later in this section.

6.2.3. Selection of reference population

In order to identify an appropriate model for the England and Wales reference population, we have carried out an extensive evaluation of the goodness-of-fit of a number of candidate single population models. However, for the sake of brevity, we present here only the conclusion of this evaluation, but details can be followed in Haberman et al. (2014, Section 6.2.2.3).

Consistently with the existing literature which compares single population mortality models for the England and Wales population (see e.g. Cairns et al. (2009) and Haberman and Renshaw (2011)), we have found that the three models presented in Table 2 are appropriate for modelling the mortality in the reference population. In Table 2, the model labelled LC+Cohorts is one of the Renshaw and Haberman (2006) cohort extensions of the Lee-Carter model while the APC model is a special case of the LC+Cohorts where it is assumed that $\beta_x^R = 1$. Model M7 is an extension of the original CBD model and was proposed in Cairns et al. (2009). A common characteristic of these three models is that they all include a cohort term to capture the well-known effect of year-of-birth on England and Wales mortality (Willets, 2004).

6.2.4. Goodness-of-fit for book population

In line with the models selected for the reference population, we have adapted several of the candidate two-population models before carrying out further goodness-of-fit assessments. Specifically, we have made the following adaptations:

- The Common Age Effect model, as proposed in Kleinow (2015), does not include a cohort effect. Therefore, given that there is strong evidence of a cohort effect in England and Wales, in our testing we extend this model to include such an effect. The reference population model is then a LC+Cohorts model.

- Similarly, for the Augmented Common Factor model we should consider a cohort effect, but doing so would turn the model into the Relative Lee-Carter model with cohorts. Consequently, the Augmented Common Factor model is not considered further in the analysis.

- In the two-population M5 and the two-population M6 models we replace the corresponding M5 and M6 models for the reference population with an M7 model.

- For the Relative Plat model we assume an M7 model for the reference population as opposed to the M5 model originally assumed by Plat (2009b). In addition, while Plat (2009b) models directly the mortality ratio between the reference and the book population, q^B_{xt}/q^R_{xt}, we model the difference in logits of mortality rates, $\operatorname{logit} q^B_{xt} - \operatorname{logit} q^R_{xt}$.

- For the Saint model, instead of the frailty-type model considered originally by Jarner and Kryger (2011) which we believe is too complex to be accessible to practitioners and does not permit the generation of sample paths, we use an M7 model for the reference population.

For comparison purposes, in some of our additional goodness-of-fit and reasonableness testing we will consider the Common Factor Model with added cohorts. This model, which was previously deemed inappropriate as it unrealistically implies zero basis risk, is useful for illustrating some of the undesirable characteristics in a model for basis risk assessment.

Table 3 summarises the models whose goodness-of-fit will be investigated further. The parameter constraints associated with these model structures are described in Appendix B. The Common Factor model with cohorts (CF+Cohorts), the Common Age Effect model with cohorts (CAE+Cohorts), and the relative Lee-Carter model with cohorts (RelLC+Cohorts) belong to the Lee-Carter family of models. The CF+Cohorts only allows for level differences between the reference and the book population, whilst the CAE+Cohorts and the RelLC+Cohorts also allow for improvement differences. Nevertheless, the latter two models differ in the specification of the age-modulating factor β^B_x accompanying the book-specific time index κ^B_t: in the RelLC+Cohorts β^B_x is estimated directly from the observed logit difference of mortality between the book and reference data while in the CAE+Cohorts β^B_x is borrowed from the reference population model, i.e., $\beta^B_x \equiv \beta^R_x$.

The Gravity model corresponds to a Binomial-logit version of the two-population APC introduced in Equation (5).

Models M7-M5, M7-M6, M7-M7, M7-Saint, and M7-Plat (which are the implemented versions of the two-population CBD, the two-population M6, the two-population M7, the Saint model, and the Relative Plat model, respectively) all belong to the CBD family of models. These models differ in the type of differences between the book and the reference population that are allowed for in the parametric age functions: M7-M5 and M7-M6 allow only for level and slope differences with M6 also allowing for cohort differences; M7-Saint, M7-M7 allow for level, slope and curvature differences with M7 also allowing for cohort differences; and M7-PLAT is a constrained version of M7-M5 assuming that at age 100 there is no difference between the reference and the book.

A good two-population model should show a reasonable fit to the historical mortality rates in both the reference population and the book population. In addition, the model should show a good fit to metrics involving the two populations such as differences or ratios of mortality rates. This last criterion is very relevant as demographic basis risk emerges from the mismatch in the mortality of the reference and the book population.

When assessing the quality of the fit of the models with respect to the book population and with respect to two-population metrics, we have found that the traditional graphical diagnostic of model residuals is not very informative. In principle, this can be attributed to the fact that cohort and age patterns in the book population residuals may be confounded with the sampling noise in the book data. Alternatively, the examination of plots of fitted vs.

Table 3. Mathematical description of the two-population models considered for goodness-of-fit assessment. CF+Cohorts = Common Factor model with cohorts; CAE+Cohorts = Common Age Effect model with cohorts; RelLC+Cohorts = Relative Lee-Carter model with cohorts; M7-X = Two-population model where the reference population follows an M7 model and the book-reference difference is specified through a model of type X. See Table 2 for the corresponding reference population models.

Original Model	**Model Name**	**Reference Population (See Table 2)**	**Book-Reference Difference Formula** $\mathbf{logit\,q^B_{xt} - logit\,q^R_{xt}}$
Common Factor	CF+Cohorts	LC+Cohorts	α_x^B
Common Age Effect	CAE+Cohorts	LC+Cohorts	$\alpha_x^B + \beta_x^R \kappa_t^B$
Relative Lee-Carter with cohorts	RelLC+Cohorts	LC+Cohorts	$\alpha_x^B + \beta_x^B \kappa_t^B$
Gravity	Gravity (APC)	APC	$\alpha_x^B + \kappa_t^B + \gamma_{t-x}^B$
Two-population M5	M7-M5	M7	$\kappa_t^{(1,B)} + (x-\bar{x})\kappa_t^{(2,B)}$
Two-population M6	M7-M6	M7	$\kappa_t^{(1,B)} + (x-\bar{x})\kappa_t^{(2,B)} + \gamma_{t-x}^B$
Two-population M7	M7-M7	M7	$\kappa_t^{(1,B)} + (x-\bar{x})\kappa_t^{(2,B)} + \left((x-\bar{x})^2 - \hat{\sigma}_x^2\right)\kappa_t^{(3,B)} + \gamma_{t-x}^B$
Saint model	M7-Saint	M7	$\kappa_t^{(1,B)} + (x-\bar{x})\kappa_t^{(2,B)} + \left((x-\bar{x})^2 - \hat{\sigma}_x^2\right)\kappa_t^{(3,B)}$
Plat relative model	M7-Plat	M7	$\frac{100-x}{100-\bar{x}}\kappa_t^{(1,B)}$

observed period survival probabilities in the book and the corresponding plots for ratios of period survival probabilities in the book and the reference can give useful insight into the goodness-of-fit of the models. As an illustration, Figure 3 depicts, for a selection of models, the fitted and observed 30 years period survival probabilities at age 60 for the "Extreme Wealthy" and the "Extreme Deprived" sample schemes as well as the corresponding fitted and observed ratios of period survival probabilities between both sample schemes and the England and Wales reference. Figure 3 is representative of the detailed analyses we have carried out and which have helped with our assessment of the performance of the different models under consideration.

The left panel in Figure 3 shows that, with the exception of the M7-Plat model which shows a slight underestimation in the later years when fitted to the "Extreme Deprived" scheme, all the other models show a similar and reasonable fit to the period survival probabilities in the book. By contrast, when considering ratios of survival probabilities the models show very different performances. In particular, from the right panel of Figure 3

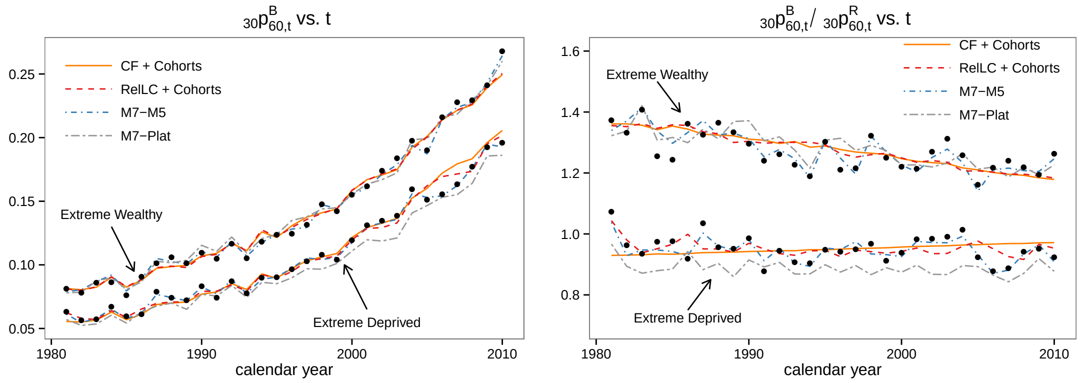


Figure 3. Fitted vs. observed 30 year period survival probabilities at age 60 for the "Extreme Wealthy" and the "Extreme Deprived" sample schemes. Left panel presents result for the book populations while the right panel presents results the corresponding results for ratios with respect to the England and Wales reference population. Dots in the graphs represent observed quantities.

we note:

- For the "Extreme Deprived" dataset the M7-Plat model shows a stark bias in the fitted ratios consistent with the underestimation seen in the period survival probabilities in the book population. In an attempt to improve the fit of the M7-Plat model, instead of assuming that crossing of mortality between the reference and book population occurs at the prefixed age 100, we have treated the age of crossing as an additional parameter that needs to be estimated from the data. This has however not eliminated the bias issues suggesting that the M7-Plat model might be too restrictive for some datasets. Therefore we do not consider the M7-Plat model further as a candidate for basis risk assessment.

- The CF+Cohorts and the RelLC+Cohorts models produce very smooth ratios of survival probabilities which seem to understate the observed volatility in the ratios. Whilst the poor performance of the CF+Cohorts model was expected due to the perfect correlation between populations assumed by this model, the poor performance of RelLC+Cohorts was not.

- Further investigation of the parameters of the RelLC+Cohorts indicates that the over-smoothed fitted ratios can be linked to the presence of a book-specific non-parametric β_x^B which needs to be estimated from the book data. The estimation of this term requires large amounts of data, and, hence, with the relatively small population sizes of the book populations, the estimated β_x^B values tend to be erratic and lack precision. In particular, there exists the possibility that β_x^B fluctuates around 0 (see Figure 4) which results in mortality differentials between the book and the reference cancelling out when aggregated measures of mortality such as survival probabilities and life expectancies are calculated. Given that this over fitting of the β_x^B may result in an inappropriate perfect correlation between the reference and the book populations, we consider that the RelLC+Cohorts is inadequate for basis risk assessment. This conclusion extends to other models with non-parametric β_x^B

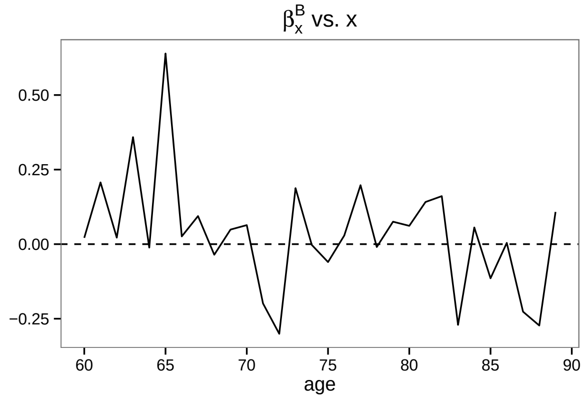


Figure 4. Fitted age modulating parameter β_x^B for the RelLC+Cohorts fitted to the "Extreme Wealthy" scheme.

Table 4. Effective number of parameters and AIC for the book part of different two-population models fitted to the four test books.

Model	Number of reference parameters	Number of book parameters	Typical Lives	Typical Amounts	Extreme Wealthy	Extreme Deprived
CF+Cohorts	185	30	7008 (1)	7001 (1)	6921 (1)	7146 (4)
CAE+Cohorts	185	59	7036 (2)	7026 (2)	6950 (2)	7130 (3)
Gravity	156	116	7090 (5)	7077 (5)	7010 (5)	7182 (6)
M7-M5	226	60	7043 (3)	7049 (3)	6971 (3)	7102 (1)
M7-M6	226	117	7106 (6)	7099 (6)	7033 (6)	7166 (5)
M7-M7	226	146	7123 (7)	7128 (7)	7052 (7)	7188 (7)
M7-Saint	226	90	7069 (4)	7074 (4)	6991 (4)	7117 (2)

parameters such as the Augmented Common Factor model and the Plat+Lee-Carter model.

The graphic testing of the goodness-of-fit of the models leaves us with six potential candidate models for basis risk assessment. These models are: CAE+Cohorts, Gravity, M7-M5, M7-M6, M7-M7, and M7-Saint. The balance between goodness-of-fit and parsimony of these models is investigated in Table 4 where we show the AIC values[8] for the book part of each model when applied to the four sample schemes, together with the corresponding ranking across models (in brackets). From Table 4 we note the following:

- The CF+Cohorts, which is the simplest model among all the models fitted, tops the AIC ranking for three out of four datasets. However, as noted before, this model is not suitable for basis risk assessment since it assumes that the reference and book populations are perfectly correlated. One may nevertheless consider this model for other applications where the correlation between the populations is not important.

[8] The AIC value is computed as $AIC = 2\nu_B - 2\mathscr{L}_B$ where $\mathscr{L}_B$ is the Binomial log-likelihood of the book part of the model under the assumption that the reference population is treated as a known offset and ν_B is the number of book-specific parameters of the models.

- Among all other models, the CAE+Cohort and M7-M5 show the best compromise between goodness-of-fit and parsimony, consistently ranking in the top three places and with very similar performance.

- M7-Saint and M7-M7, which have a quadratic age term in the book model, are always outperformed by the M7-M5 model. This suggests that when considering models from the CBD-Family it is necessary to allow for differences in level of mortality and a gradient by age, but that an additional parameter for the curvature by age is not necessary, i.e., it is sufficient to inherit the curvature from the reference population. Thus, we eliminate the M7-M7 and M7-Saint models from our list of candidate models.

- The Gravity model, M7-M6 and M7-M7, which have a book-specific cohort effect, have the worst trade-off between goodness-of-fit and parsimony. This suggests that we should generally reject models with a book cohort effect on grounds of parsimony. However, for the moment we shall retain the Gravity model (two-population APC) which, among models with book-specific cohort effect, shows the best compromise between goodness-of-fit and parsimony. This will enable us to investigate how forecasts levels of uncertainty and hedge effectiveness may be impacted by allowing for a book-specific cohort effect.

6.2.5. Plausibility of forecast central trends and levels of uncertainty

So far, we have shortlisted the CAE+Cohorts, Gravity and M7-M5 based on their theoretical properties, practicality and goodness-of-fit performance. However, the outcome of a basis risk assessment exercise will be strongly driven by the expected level of uncertainty around the central forecast of the demographic and financial quantities underlying the index-based hedge. It is then crucial to check that these models produce reasonable forecast for both single and two-population metrics. This entails judging whether or not the forecast central trajectories and patterns of uncertainty look plausible and are in line with historical variability.

Following Cairns et al. (2011b), we assess this property by examining fan charts of the forecasts produced by the models. Fan charts allow us to examine any distinctive visual feature of the forecasts of the models, as well as the differences between models. Each fan chart presents 95% prediction intervals and depicts the forecast output from the stochastic mortality models by also presenting 80% and 50% prediction intervals.

In producing the model simulations underlying the fan charts, we have considered the following two sources of uncertainty (risk): i) *process risk* (PR) arising from the possible future trajectories of the time series of the period and cohort indices and ii) *parameter uncertainty* (PU) arising from the estimation of the parameters of the model. Process risk is taken into account by simulating trajectories of the period and cohort indices,[9] while parameter uncertainty is allowed for by using a Binomial adaptation of the residual bootstrapping approach proposed by Koissi et al. (2006).[10] We note that due to

[9] To model process risk we use a multivariate adaptation of Algorithm 2 in Haberman and Renshaw (2009) without provision for parameter error. We note that Algorithm 2 in Haberman and Renshaw (2009) is itself an adaptation of the prediction interval approach of Cairns et al. (2006).

[10] We note that in adapting the bootstrap we follow Renshaw and Haberman (2008) and solve for the observed numbered of deaths instead of the fitted number of deaths as done by Koissi et al. (2006). The details of the residual bootstrapping approach under a Binomial framework are described in Debón et al. (2010, Section 3).

the considerable exposure of the England and Wales population, we deliberately ignore parameter uncertainty in the reference population.

Rather than analysing forecasts of mortality rates, we concentrate on the forecast of life expectancies and survival rates. According to Coughlan et al. (2011), these two aggregate quantities are more appropriate than individual mortality rates for gaining insight into the basis risk associated with longevity hedges. On the one hand, life expectancies and survival rates are more closely related to the hedge effectiveness objective than mortality rates, as, for instance, in a pensioner population life expectancy corresponds to the number of years over which a pension needs to be paid while survival rates correspond to the number of pensioners who are still alive to receive pension. On the other hand, these aggregate metrics smooth out a lot of the noise associated with individual mortality rates.

Figure 5 presents fan charts of 30 year curtailed period life expectancies at age 60,

$$\overset{\uparrow}{e}{}^{i}_{60,\overline{30}|}(t) = \sum_{h=1}^{30} \prod_{j=0}^{h-1} (1 - q^{i}_{60+j,t}), \qquad i = R, B,$$

along with fan charts for the value of a cohort survivor index,

$$S^{i}(65,t) = \prod_{j=0}^{t-1} (1 - q^{i}_{65+j,2011+j}), \qquad i = R, B,$$

for the reference population ($i = R$) and for the "Extreme Wealthy" test book ($i = B$). Figure 5 also shows matching fan charts of the difference between the period life expectancies in the book and the reference population, $\overset{\uparrow}{e}{}^{B}_{60,\overline{30}|}(t) - \overset{\uparrow}{e}{}^{R}_{60,\overline{30}|}(t)$, and of the ratio of the book and reference population survivor indexes, $S^{B}(65,t)/S^{R}(65,t)$. The survivor index, $S^{i}(65,t), i = R, B$, measures the proportion from a group of males aged 65 at the start of 2011 who are still alive at the start of year $2011 + t$. We note that $S^{R}(65,t)$ and $S^{B}(65,t)$ do not involve any forecasts of the cohort effects as the relevant cohort effects, γ^{R}_{1946} and γ^{B}_{1946} in the case of the Gravity model, are known at the start of 2011.

To assist in the assessment of the levels of uncertainty produced by the models, Table 5 presents the forecast variance of period life expectancies in 2020 at age 60, $\overset{\uparrow}{e}{}^{R}_{60,\overline{30}|}(2020)$ and $\overset{\uparrow}{e}{}^{B}_{60,\overline{30}|}(2020)$, while Table 6 presents the forecast variance of the 25 year cohort life expectancy for someone aged 65 in 2011 in the reference and book populations,

$$\overset{\nearrow}{e}{}^{i}_{65,\overline{25}|}(2011) = \sum_{t=1}^{25} S^{i}(65,t) = \sum_{t=1}^{25} \prod_{j=0}^{t-1} (1 - q^{i}_{65+j,2011+j}), \qquad i = R, B.$$

From Figure 5 and Tables 5, 6 we can see that:

- For all the models the central forecast and their levels of uncertainty for the life expectancies and the survivor indexes in the reference and the book are reasonable and consistent. We note however that there are noticeable differences between the models with the CAE+cohorts and CF+cohorts projecting longer life expectancies and higher survival probabilities with slightly smaller uncertainty (narrower fan widths and smaller variances) than the other two models. Notably, M7-M5 produces wider fans for the reference population than the other three models. This reflects the existence of more random period effects in M7-M5 than in the CAE+cohorts, the CF+cohorts and the APC (Gravity) model.

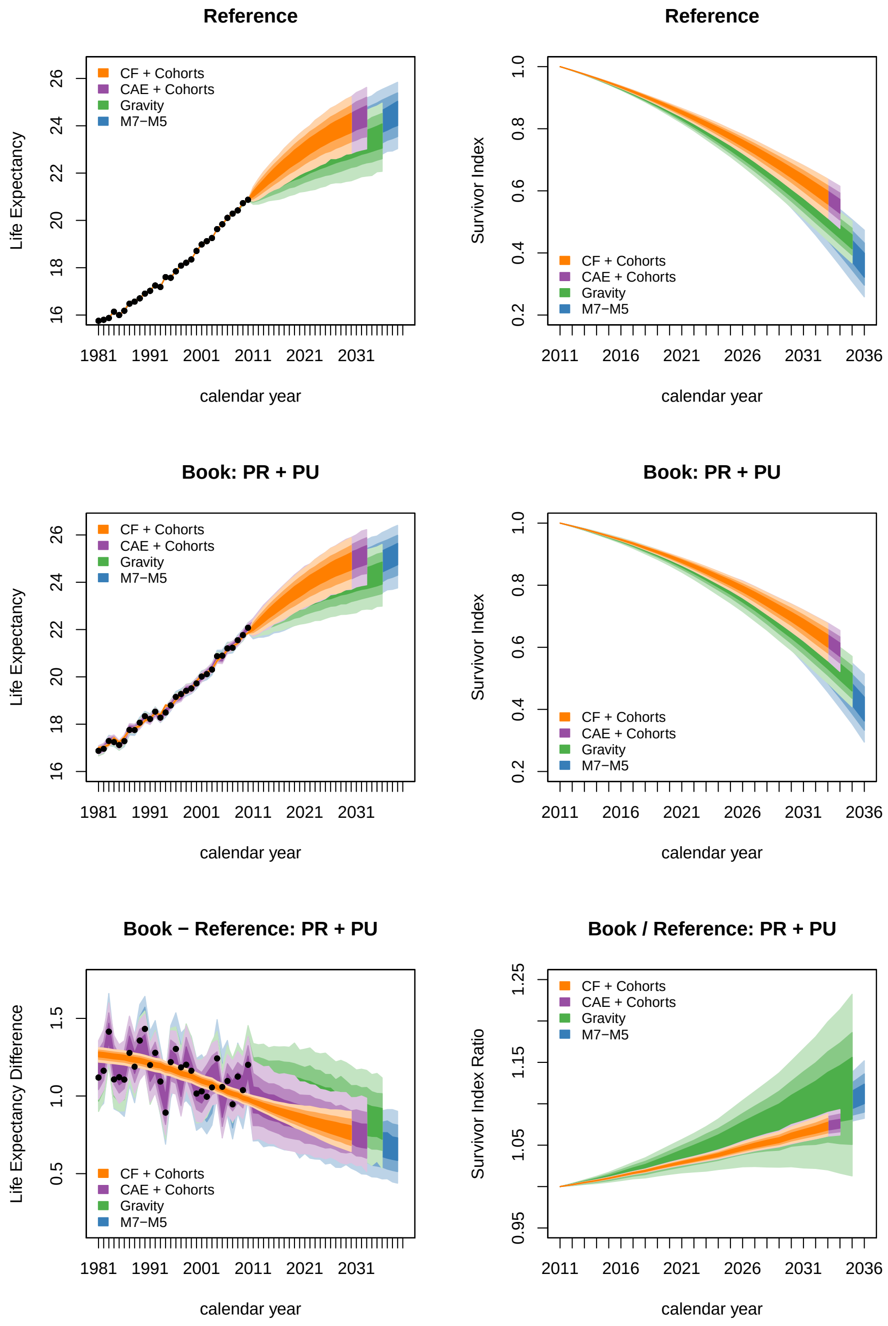


Figure 5. Fan charts of 30 year period curtailed life expectancy at age 60, $\overset{\uparrow}{e}^{i}_{60,\overline{30}|}(t), i = R, B$, and of the cohort survivor index, $S^i(65,t), i = R, B$, for the England and Wales reference population and the "Extreme Wealthy" book using different mortality models.

- The levels of uncertainty in the difference in life expectancy and in the ratio of survivor indexes vary considerably across models. In particular, the unreasonably tight fan widths produced by the CF+cohorts for differences in period life expectancy confirm the issues with models assuming a perfect correlation between the reference and the book populations.

Table 5. Forecast mean and variance of 30 year period curtailed life expectancy at age 60 in 2020 for the England and Wales male reference population and for the "Extreme Wealthy" test book using different mortality models. The columns labelled "Difference" present the mean and variance of the difference in life expectancy between the reference and the book populations.

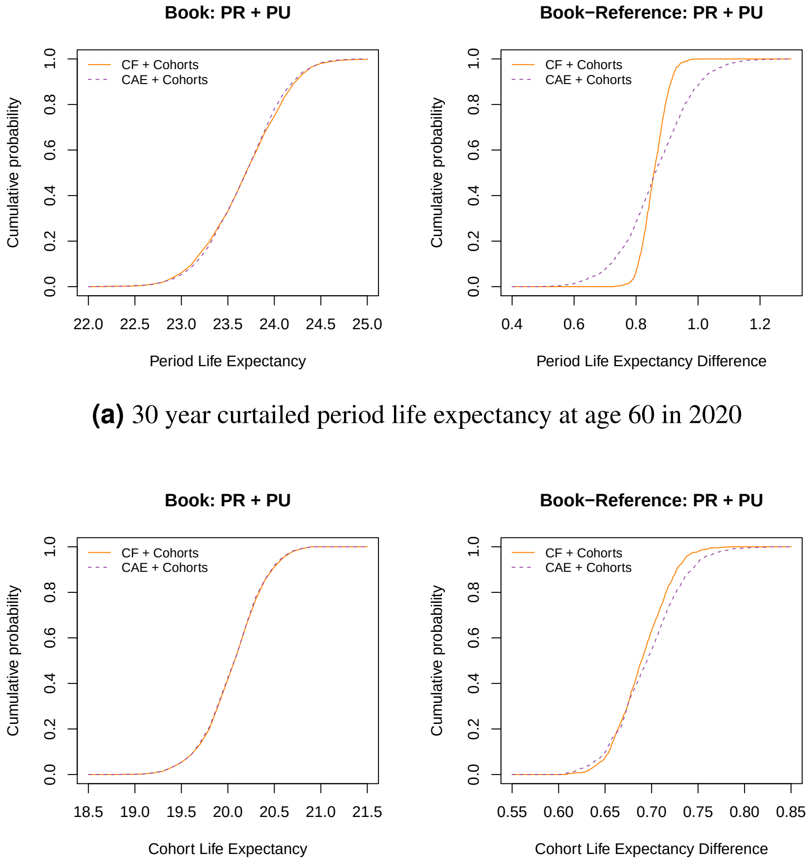
	Mean of period life expectancy at age 60 in 2020			Variance of period life expectancy at age 60 in 2020		
Model	**Reference**	**Book**	**Difference**	**Reference**	**Book**	**Difference**
CF+Cohorts	22.82	23.68	0.86	0.2040	0.1718	0.0017
CAE+Cohorts	22.82	23.68	0.86	0.2040	0.1878	0.0135
Gravity	22.12	23.13	1.01	0.2251	0.2065	0.0206
M7-M5	22.33	23.19	0.86	0.2403	0.2230	0.0169

Table 6. Forecast mean and variance of 25 year cohort curtailed life expectancy for the cohort aged 65 in 2011 for the England and Wales male reference population and for the "Extreme Wealthy" test book using different mortality models. The columns labelled "Difference" present the mean and variance of the difference in life expectancy between the reference and the book populations.

	Mean of cohort life expectancy for the cohort aged 65 in 2011			Variance of cohort life expectancy for the cohort aged 65 in 2011		
Model	**Reference**	**Book**	**Difference**	**Reference**	**Book**	**Difference**
CF+Cohorts	20.36	21.05	0.69	0.1229	0.1062	0.0008
CAE+Cohorts	20.36	21.06	0.70	0.1229	0.1066	0.0014
Gravity	19.47	20.34	0.87	0.1252	0.2027	0.0905
M7-M5	19.54	20.27	0.73	0.1830	0.1677	0.0015

- The levels of uncertainty for the ratio of survivors in the book and reference population produced by the Gravity (APC) model, which is the only model that allows for a book specific cohort effect, are completely unreasonable. This suggests that, unless there is strong reason to believe in the existence of a different cohort effect in the book to the reference population, the parameter uncertainty in fitting a book-specific cohort term will greatly outweigh any benefits in terms of goodness-of-fit to historical experience.

- The close alignment between the fans of models CF+cohorts and CAE+cohorts deserves further investigation. For these two models, which share the same reference population model, we plot in Figure 6 the simulated empirical cumulative distribution considering both process risk and parameter risk for the 30 year period curtailed life expectancy at age 60 in 2020, $\overset{\uparrow}{e}^{B}_{60,\overline{30}|}(2020)$, and for the 25 year cohort life expectancy for someone aged 65 in 2011 in the book population, $\overset{\nearrow}{e}^{B}_{65,\overline{25}|}(2011)$, together with the corresponding simulated empirical cumulative distribution of the difference in period and cohort life expectancies between the book and the reference populations. While for the book population the empirical distributions are practically indistinguishable, there are notable dissimilarities in the empirical distributions for the difference in both period and cohort life expectancies, suggesting that the uncertainty in the book



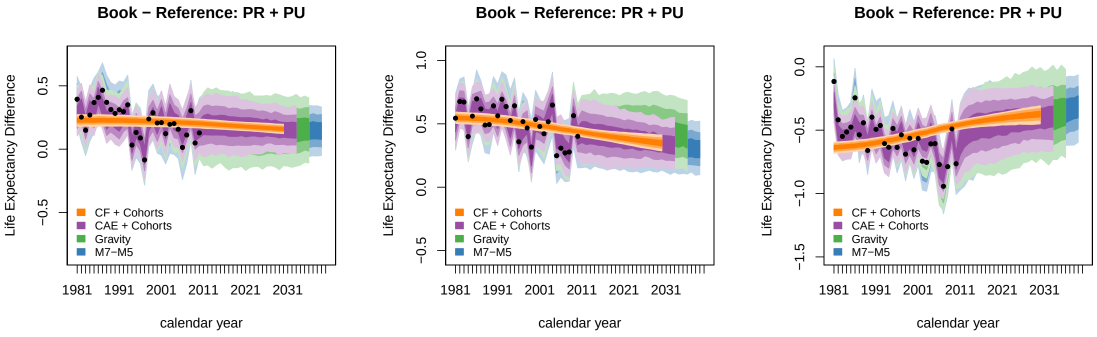
(a) 30 year curtailed period life expectancy at age 60 in 2020

(b) 25 year curtailed cohort life expectancy at age 65 in 2011

Figure 6. Cumulative distribution function of the curtailed period and cohort life expectancy in the "Extreme Wealthy" test book and cumulative distribution function of the corresponding difference in period and cohort life expectancies between the book and reference populations. The cumulative distribution functions account for both process and parameter risk.

survivor index is dominated by the uncertainty in the reference part of the model. Furthermore, the discrepancies in the mean and variances of the life expectancy for the book population forecast by both models are immaterial. These results allow us to conclude that although the CF+Cohorts model is unsuitable for basis risk assessment due to its implicit perfect correlation between the book and reference populations, this model might be a reasonable alternative in applications where only single population metrics are of interest such as when valuing pension liabilities or pricing annuities.

Finally, in order to investigate the generalisability of our conclusion to the other test datasets, we show in Figure 7 fan charts of the difference in period life expectancies with respect to the England and Wales reference for the "Typical Lives", "Typical Amounts" and "Extreme Deprived" books. While the forecasts for the "Typical Lives" and "Typical Amounts" books look plausible, the forecasts for the 'Extreme Deprived" book look completely unreasonable, with the models failing to project the increase in life expectancy differences observed over the 1981-2010 period. Recalling Table 1, 60% of the "Extreme



(a) Typical Lives **(b)** Typical Amounts **(c)** Extreme Deprived

Figure 7. Fan charts of 30 year period curtailed life expectancy differences at age 60 between the different test books and the reference population.

Deprived" population belong to the two most deprived quintiles of England which have seen a significant increase in relative mortality differentials with the respect to England and Wales (see Villegas and Haberman (2014)). This suggests that the non-divergence assumption embedded in the (vector) autoregressive process of order 1 used for forecasting the period index in the book κ_t^B (recall Equation (8)) may be inappropriate for the "Extreme Deprived" book.

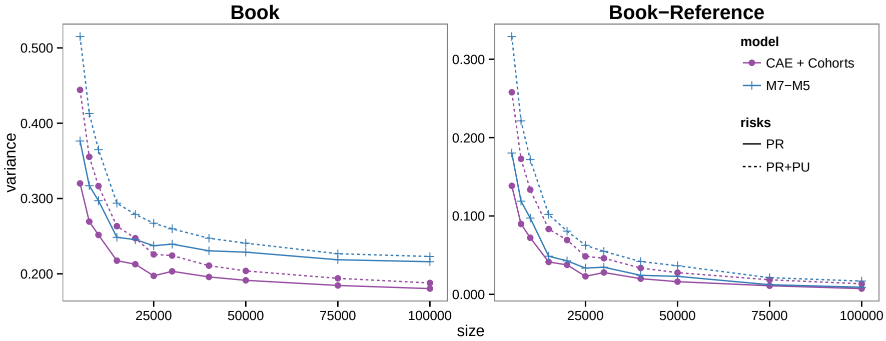
Overall, although all the models produce plausible trends and forecast levels of uncertainty for single population metrics, for two-population metrics only models CAE+Cohorts and M7-M5 produce plausible results. In addition, there are big enough differences between the models for us to acknowledge model risk as an important issue.

6.2.6. Forecast levels of uncertainty by book size

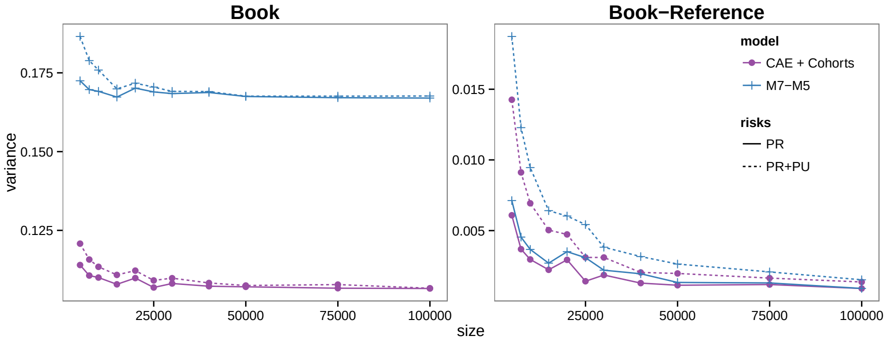
The analysis of the plausibility of the forecast levels of uncertainty performed so far has been based on a fairly large book population with 100,000 exposed lives per year between ages 60 and 89. However, for smaller exposures of the book population the sampling noise in the data is bigger, leading to more uncertainty in the estimates of the parameters of the models. This additional variability arising from a smaller population size can potentially have a material impact on the plausibility of the forecast levels of uncertainty. To explore this phenomenon, we investigate how the contribution of the different sources of uncertainty to the total level of risk varies by population size. Figure 8 shows, for models CAE+Cohorts and M7-M5 and the "Extreme Wealthy" test book, the variation by book size of the variance of $\overset{\uparrow}{e}{}^{B}_{60,\overline{30}|}(2020)$ and of $\overset{\uparrow}{e}{}^{B}_{60,\overline{30}|}(2020) - \overset{\uparrow}{e}{}^{R}_{60,\overline{30}|}(2020)$ (top left and top right plots), and the variance of $\overset{\nearrow}{e}{}^{B}_{65,\overline{25}|}(2011)$ and of $\overset{\nearrow}{e}{}^{B}_{65,\overline{25}|}(2011) - \overset{\nearrow}{e}{}^{R}_{65,\overline{25}|}(2011)$ (bottom left and bottom right plots). From this figure we can see how:

- The differences in the levels of uncertainty produced by the models are evident, with the M7-M5 producing higher variance than the CAE+Cohorts. These differences are particularly notable for cohort life expectancies in the book population.

- The magnitude of the variance of both period and cohort life expectancies starts to stabilise around a book size of 25,000 lives. This is particularly noticeable when considering only process risk.



(a) 30 year curtailed period life expectancy at age 60 in 2020



(b) 25 year curtailed cohort life expectancy at age 65 in 2011

Figure 8. Variance by population size of the curtailed period and cohort life expectancy in the "Extreme Wealthy" test book and of the corresponding difference in period and cohort life expectancies between the book and reference populations using different models and considering different sources of risk.

- For book sizes smaller than 15,000 lives, process risk is unrealistic producing artificially high variances.

These observations suggest that to avoid a distorted assessment of the levels of uncertainty, models CAE+Cohorts and M7-M5 should only be used when the book exposure is higher than 20,000-25,000 lives. Furthermore, as we show in Section 7, insisting on using the models with modest exposure numbers may result in a misstated assessment of demographic basis risk.

6.2.7. Forecasting performance and robustness

A good mortality model should not only produce forecasts that appear reasonable ex-ante, but should also provide good ex-post forecast, that is forecasts that do not deviate significantly from realised outcomes. In addition, these forecasts should be robust relative to the choice of period for the data employed in producing the forecasts. To assess the forecasting accuracy of the models, we first carry out a backtesting exercise in the spirit of Booth et al. (2006) and Jarner and Kryger (2011, Section 4). This exercise entails the fitting and forecasting of the models using data for the period 1981 to 2010 for different history lengths, book sizes, and IMD compositions in the book population; and the evaluation of different metrics of forecasting performance.

Specifically, the different models were fitted to history lengths ranging from 5 years to

20 years,[11] book sizes ranging from 5,000 lives to 100,000 exposed lives between ages 60 to 89 and the four test IMD compositions described before in Table 1. The forecasting performance of the models is evaluated by comparing the actual 30 year curtailed period life expectancies at age 60 in the book population, $\overset{\uparrow}{e}{}^{B}_{60,\overline{30}|}(t)$, and the actual differences in 30 year curtailed period life expectancies at age 60 between the book and the reference population, $\overset{\uparrow}{e}{}^{B}_{60,\overline{30}|}(t) - \overset{\uparrow}{e}{}^{R}_{60,\overline{30}|}(t)$, with their corresponding predicted counterparts over the rest of the period until 2010. Forecast bias (actual-fitted) is summarised by averaging across years, book sizes and forecasting horizon. The matching absolute errors are also averaged to provide a measure of forecast accuracy.

The forecast bias (mean errors) and the forecast accuracy (mean absolute error) for both period life expectancy in the book and differences in period life expectancy between the book and the reference, plotted against history length are shown in Figure 9. We note the following:

- Models CAE+cohorts and CF+Cohorts stand out as the best models for forecasting period life expectancies in the book with the smallest bias and with the smallest mean absolute error. The close alignment between the mean errors and mean absolute errors of these two models reflects the fact that they share the same reference population model.

- History length has a material impact on the out-of-sample performance of the models. With the exception of model CF+Cohorts which does not require the forecasting of any book specific time index, the forecasting performance of the models for history lengths shorter than 8 years is poor. The noticeably poorer performance of model M7-M5 for the shorter history lengths is explained by the fact that this model has two period indices for the book, implying a more complex and data demanding time series process for the forecasting.

- For differences in life expectancies and when we have more than 8 years of history, the models perform very similarly both in terms of bias and accuracy.

- The bias in forecasting differences between the "Extreme Deprived" population and the England and Wales reference population is considerably higher than the bias for the other three test book compositions. This higher bias gives further evidence for concluding that the non-divergence assumption may be inadequate for the "Extreme Deprived" book.

In order to check the robustness of the models we will examine the stability of forecasts towards the inclusion of additional years at the right end of the data window, using a contracting horizon backtest as proposed by Dowd et al. (2010). Figure 10 shows plots of forecasts of the 30 year period curtailed life expectancy at age 60 in 2010, $\overset{\uparrow}{e}{}^{B}_{60,\overline{30}|}(2010)$, for the four test book populations with 100,000 exposed lives, made in 1985, 1986, ..., 2009. Equivalent plots for the difference between the book and reference population,

[11] For instance when considering a history length of 5 years the models were fitted using data for the book population covering the periods 1981-1985, 1982-1986, 1983-1987,..., 2003-2007, 2004-2008, 2005-2009. In all cases, the reference population data was assumed to start in 1961 and end in the same year as the book population data.

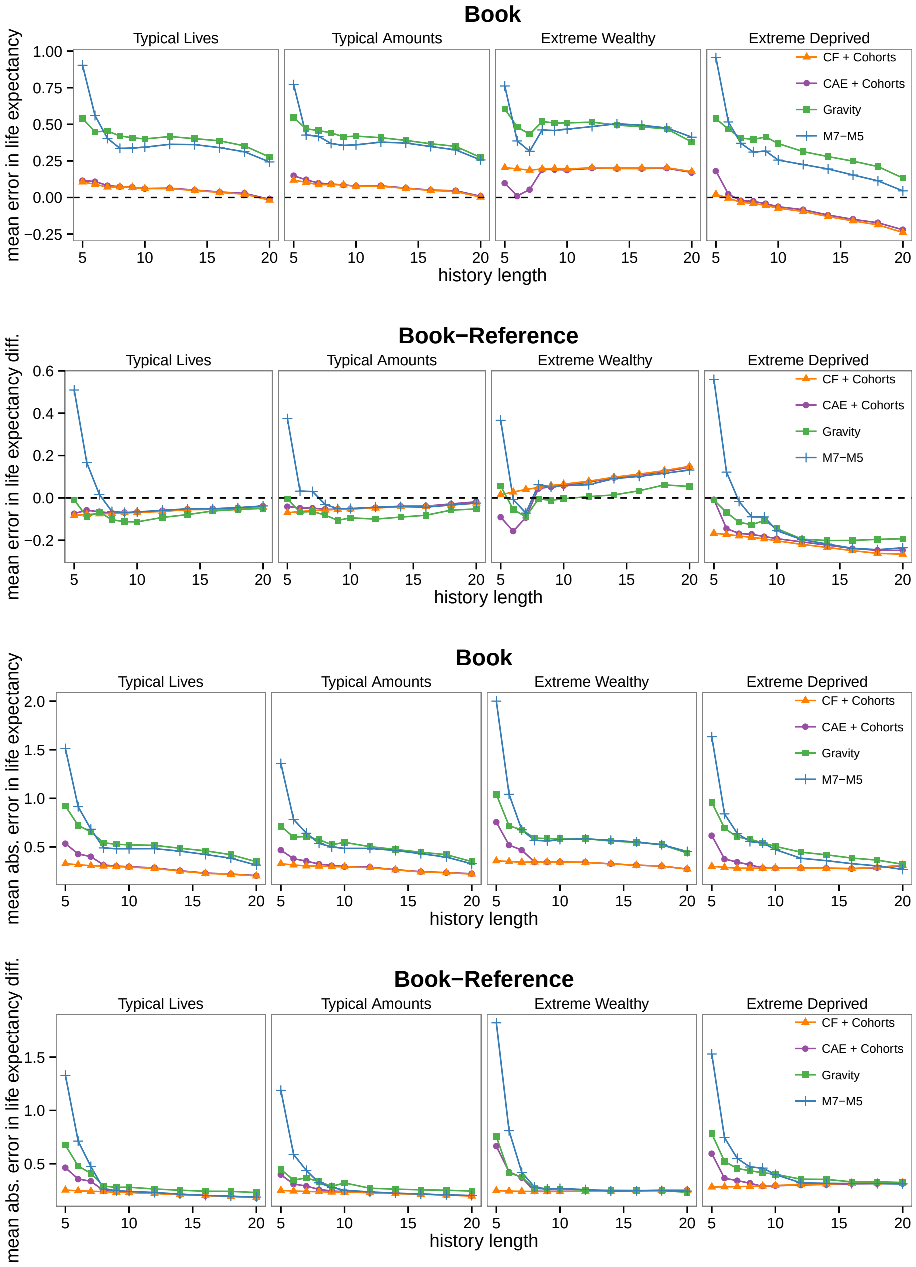


Figure 9. Mean error (actual - fitted) and mean absolute error in the forecast of 30 year period curtailed life expectancy at age 60 in the book and in differences in 30 year period curtailed life expectancy at age 60 between the book and the reference. The results are averaged across years, book sizes and forecast horizons ranging from 1 year to 15 years ahead.

$\overset{\uparrow}{e}^{B}_{60,\overline{30}|}(2010) - \overset{\uparrow}{e}^{R}_{60,\overline{30}|}(2010)$, are also included. For all book populations and models, we see that the forecasts for the book population are well-behaved, in the sense that they converge in a stable manner towards the realised outcome. For forecasts of differences in life expectancy we see a similar stable behaviour, albeit the discrepancies in the forecasts for the "Extreme Deprived" population are noticeable and consistent with our previous findings regarding the unsuitability of the non-divergence assumption for this book (recall

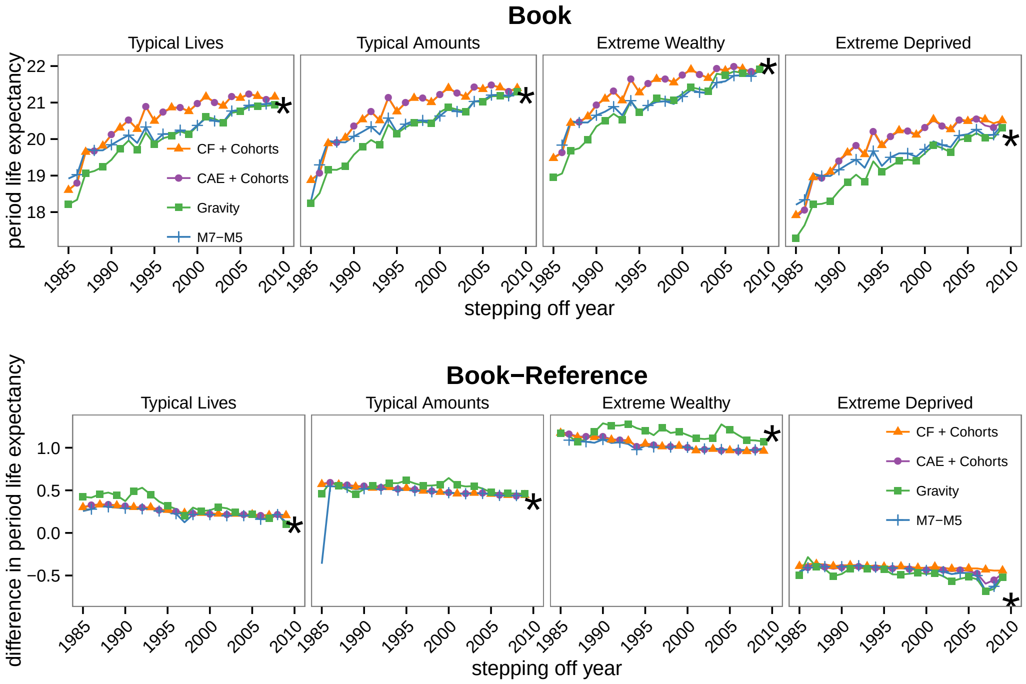


Figure 10. Forecast of the 30 year period curtailed life expectancy at age 60 in 2010 for the four different book populations using different fitting periods. The stepping-off year is the final year used in fitting the models. The realised life expectancy for 2010 is represented by a star.

Figure 7). Similar analysis for other book sizes show that the models are robust, provided that the length of the fitting period is longer than 10 years (i.e. for stepping off years after 1990).

7. QUANTIFYING BASIS RISK

In this section we examine the performance of the models when used for assessing basis risk. We also discuss the impact that different volumes of data may have on the parameter uncertainty and on the assessment of basis risk. In presenting the basis risk analysis we follow the five-step hedge effectiveness framework proposed in Coughlan et al. (2011).

7.1. Steps 1 and 2: Hedging objectives and hedging instruments

Most hedging exercises either consider *value hedges* or *cash flow hedge* aiming, respectively, to mitigate the variability of the *cash flows* or the variability of the *value* of these cash flows. We consider thus two separate simple examples reflecting the objectives of a *value hedge* and of a *cash flow hedge*:

Value hedge example. Noting that period life expectancy corresponds to an annuity value using a zero percent interest rate and no mortality improvements, for the value hedge case we assume that the quantity at risk to be hedged is the 30 year curtailed period life expectancy at age 60 in 2020, $\overset{\uparrow}{e}{}^{B}_{60,\overline{30}|}(2020)$, i.e. over a horizon of 10 years. The hedging instrument to reduce the liability risk is the 30 year curtailed period life expectancy at age 60 in 2020 for the England and Wales population, $\overset{\uparrow}{e}{}^{R}_{60,\overline{30}|}(2020)$. This exercise is similar in spirit to the hedge effectiveness analysis performed in Cairns (2013) and Cairns et al. (2014).

Cash flow hedge example. To reflect a cash flow hedge situation we consider that the quantity at risk to be hedged is the 25 year curtailed cohort life expectancy at age 65 in 2011, $\overset{\nearrow}{e}{}^{B}_{65,\overline{25}|}(2011)$, which can be interpreted as the sum of the cash flows payable for 25 years to a pensioner aged 65 in 2011 and who belongs to a pension plan that pays £1 at the end of each year. The hedging instrument to reduce the liability risk is the 25 year curtailed cohort life expectancy at age 65 in 2011 for the England and Wales population. $\overset{\nearrow}{e}{}^{R}_{65,\overline{25}|}(2011)$. This exercise is similar in spirit to the hedge effectiveness analysis performed in Li and Hardy (2011).

Although very simple, these two examples should be informative of the performance of the models for hedge effectiveness assessment while avoiding the idiosyncrasies of specific pension benefit structures or more realistic hedging instruments.

7.2. Step 3: Method for hedge effectiveness assessment

Following Li and Hardy (2011), Cairns (2013) and Cairns et al. (2014), we use the variance as our measure of risk. Alternatively, a tail based risk measure such as Value-at-Risk, as in Li and Hardy (2011) and Coughlan et al. (2011), or expected shortfall could be considered. However, because our focus is on the comparison of competing models and for the sake of simplicity, we have implemented the variance as a measure of hedge effectiveness. Therefore, if L denotes the random unhedged liability and H represent the value of the index-linked hedging instrument, we assume that the hedger wishes to minimise the variance of $L-hH$, where h is the number of units (hedge ratio) held of the hedging instrument. We define thus the relative risk reduction (hedge effectiveness) as $R^2(h) = 1 - \operatorname{var}(L-hH)\big/\operatorname{var}(L)$. It can be proved (see, for example, Cairns et al. (2014)) that the optimal hedge ratio is $h^* = \operatorname{cov}(L,H)\big/\operatorname{var}(H)$ with optimal relative risk reduction $R^2(h^*) = 1 - \operatorname{var}(L-h^*H)\big/\operatorname{var}(L) = \rho^2$, where ρ is the correlation coefficient between L and H.

For our examples, it will thus suffice to analyse the correlation between L and H, i.e., the correlation between $\overset{\uparrow}{e}{}^{B}_{60,\overline{30}|}(2020)$ and $\overset{\uparrow}{e}{}^{R}_{60,\overline{30}|}(2020)$ for the value-hedging example and the correlation between $\overset{\nearrow}{e}{}^{B}_{65,\overline{25}|}(2011)$ and $\overset{\nearrow}{e}{}^{R}_{65,\overline{25}|}(2011)$ for the cash flow-hedging example.

7.3. Step 4: Calculation of hedge effectiveness

For each stochastic two-population model under consideration we compute the correlations between L and H based on 1,000 simulated mortality scenarios. Our experience suggests that a higher number of simulations does not lead to significant differences in the results. In the analysis that follows, we contemplate three cases concerning the sources of risks considered in the simulations: i) only process risk (PR); ii) process risk and parameter uncertainty (PR+PU); and iii) process risk, parameter uncertainty and sampling risk (PR+PU+SR). Process risk and parameter risk are considered using the techniques described before in Section 6.2.5, while sampling risk is considered by randomly sampling the number of deaths from a Binomial distribution once parameter uncertainty and process risk have been taken into account.

Specifically, for the value-hedging example, we assume that the future exposures $E^{B}_{x,2010+t}$, $t = 1,..,10$, are equal to the average age-specific book exposure over the data period used in fitting the mortality model and then simulate the number of deaths using the conditional Binomial assumption:

$$D^B_{x,2010+t}|q^B_{x,2010+t} \sim \mathrm{Bin}(E^B_{x,2010+t}, q^B_{x,2010+t}), \quad t = 1, \ldots, 10.$$

For the cash-flow hedge example, we take sampling risk into account by treating the cohort of pensioners as a random survivorship group. Thus, if l_x denotes the number of pensioners who survive to age x and given a simulated mortality scenario $\{q^B_{65,2011}, q^B_{66,2012}, \ldots, q^B_{89,2035}\}$, we model sampling risk with the following Binomial death process:

$$l_{65+t} \sim \mathrm{Bin}(l_{65+t-1}, 1 - q^B_{65+t-1,2010+t}), \quad t = 1, \ldots, 25,$$

with l_{65} equal to 5% of the total exposure in the book between ages 60 to 89 (e.g. if the total book exposure is 100,000 lives we take $l_{65} = 5,000$). We have chosen 5% as from the total English male population aged 60 to 89 broadly 5% is aged 65.

7.4. Step 5: Interpretation of results

We concentrate on hedge effectiveness results for models CAE+Cohorts and M7-M5 which have been identified as the best performing models after the systematic model assessment we have carried out in Section 6. However, in spite of the implausible projections produced by models CF+Cohorts and APC (gravity) model, we shall also compute hedge effectiveness metrics for these two models to illustrate the issues that may arise if we insist on using these models for basis risk assessment.

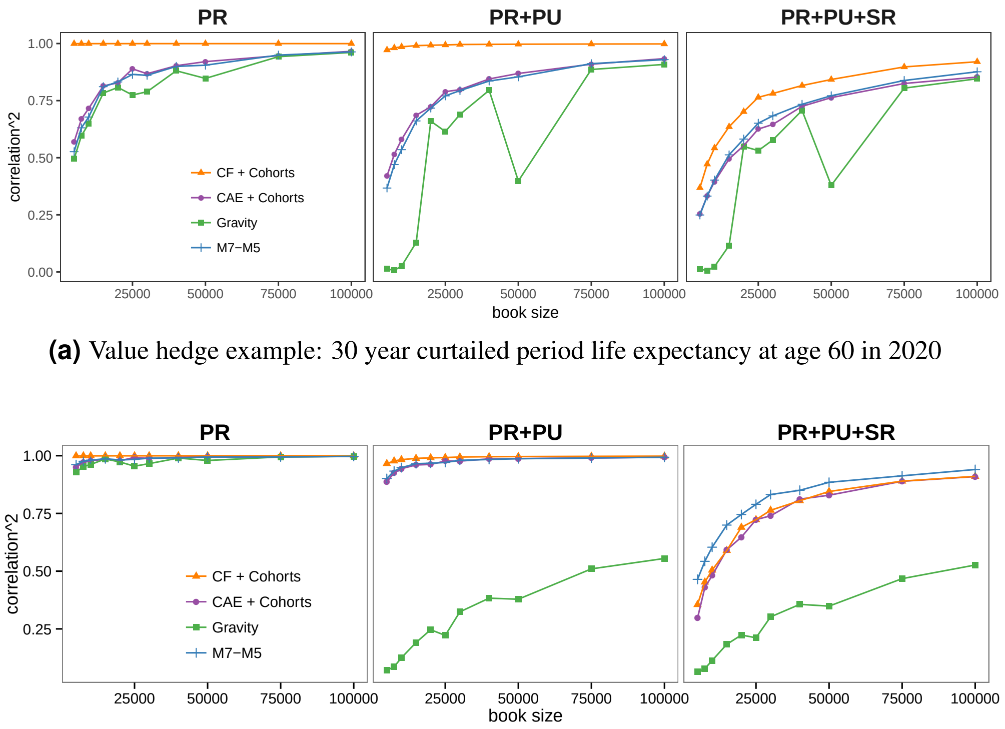
7.4.1. Hedge effectiveness by book population size

As discussed in section 6.2.5, population size has a material impact in the parameter uncertainty of the models. In addition, it is expected that the higher sampling risk associated with smaller populations will reduce the effectiveness of a standardised longevity-hedge. To investigate this phenomenon, we present in Figure 11 hedge effectiveness results for the "Extreme Wealthy" test book considering population sizes ranging from 5,000 to 100,000 exposed lives between ages 60 to 89, and considering different sources of risk. In all cases, data for the period 1981-2010 (i.e. a history length of 30 years) is used for fitting the models. From Figure 11 we note the following:

- As expected, the inherent perfect correlation of the CF+Cohorts results in an unrealistic zero or close to zero basis risk when sampling risk is ignored.

- The previously raised issues in relation to the parameter uncertainty in the estimation of book-specific cohort parameters become evident, with the APC (Gravity) model showing implausibly low hedge-effectiveness once parameter risk is taken into account. This is especially noticeable for the cash flow-hedge example which involves cohort-type quantities. In this case, even for populations as big as 100,000 lives, the hedge effectiveness produced by the model are below 60% while other models produce hedge-effectiveness of more than 85%.

- For book sizes smaller than 15,000 lives, process risk is unrealistically high (recall Figure 8) distorting the assessment of basis risk and producing artificially low hedge effectiveness for the value-hedge example.



(a) Value hedge example: 30 year curtailed period life expectancy at age 60 in 2020

(b) Cash flow hedge example: 25 year curtailed cohort life expectancy at age 65 in 2011

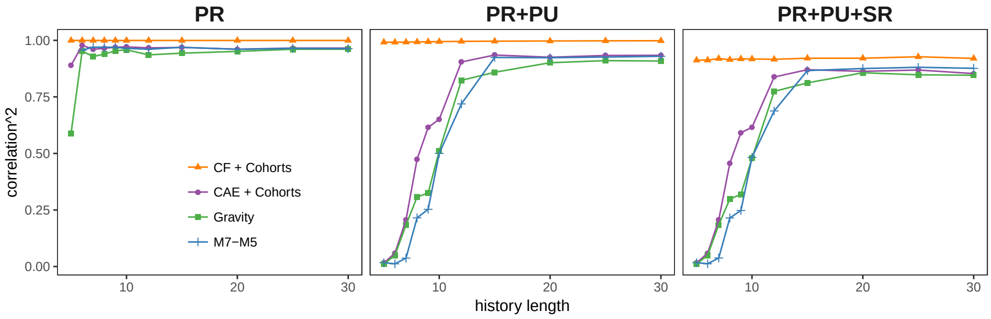
Figure 11. Squared correlation, ρ^2, between the liability L and the hedging instrument H, as a function of book population size. All values correspond to the "Extreme Wealthy" socio-economic composition where data for the period 1981-2010 have been used to fit the models.

- The impact of sampling risk on correlations is material, with hedge effectiveness values for both the value-hedge and the cash flow-hedge falling rapidly for populations below 10,000 lives. We also note that for the cash flow-hedge example, sampling risk is the main determinant of basis risk. In fact, the CF+Cohorts model which implies zero basis risk before accounting for sampling risk, results in virtually the same risk reductions as models CAE+Cohort and M7-M5 once sampling risk is accounted for.

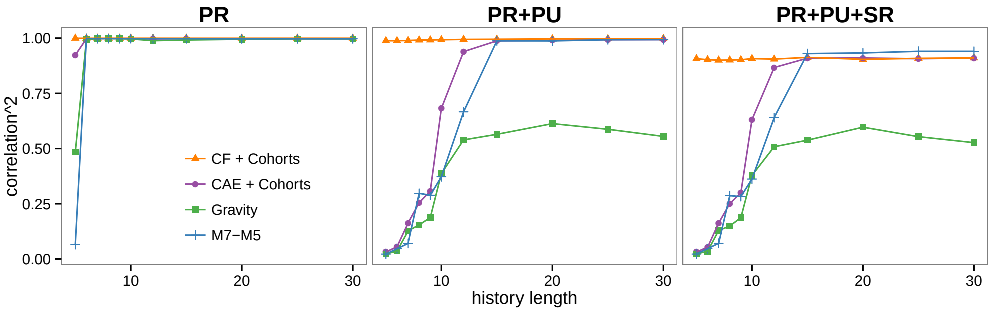
- Although models M7-M5 and CAE+Cohorts can give rather different mortality forecasts, these differences seem to attenuate in applications, with the two models producing very similar hedge effectiveness values once all risks have been taken into account.

7.4.2. Hedge effectiveness by history length

We now investigate how the number of years of available data in the book population impacts the evaluation of hedge effectiveness. Figure 12 presents hedge effectiveness results for the "Extreme Wealthy" test book considering history lengths ranging from 5 years to 30 years, and considering different sources of risk. In all cases a book population size of 100,000 lives is used for fitting the models. In this figure we can see how history length has a significant impact on hedge effectiveness assessment. For history lengths shorter than 10-12 years and once parameter uncertainty has been considered, risk reductions



(a) Value hedge example: 30 year curtailed period life expectancy at age 60 in 2020



(b) Cash flow hedge example: 25 year curtailed cohort life expectancy at age 65 in 2011

Figure 12. Squared correlation, ρ^2, between the liability L and the hedging instrument H, as a function of history length of the book population data used to fit the models. All values correspond to the "Extreme Wealthy" socio-economic composition with a book size of 100,000 annual exposed lives between ages 60 to 89.

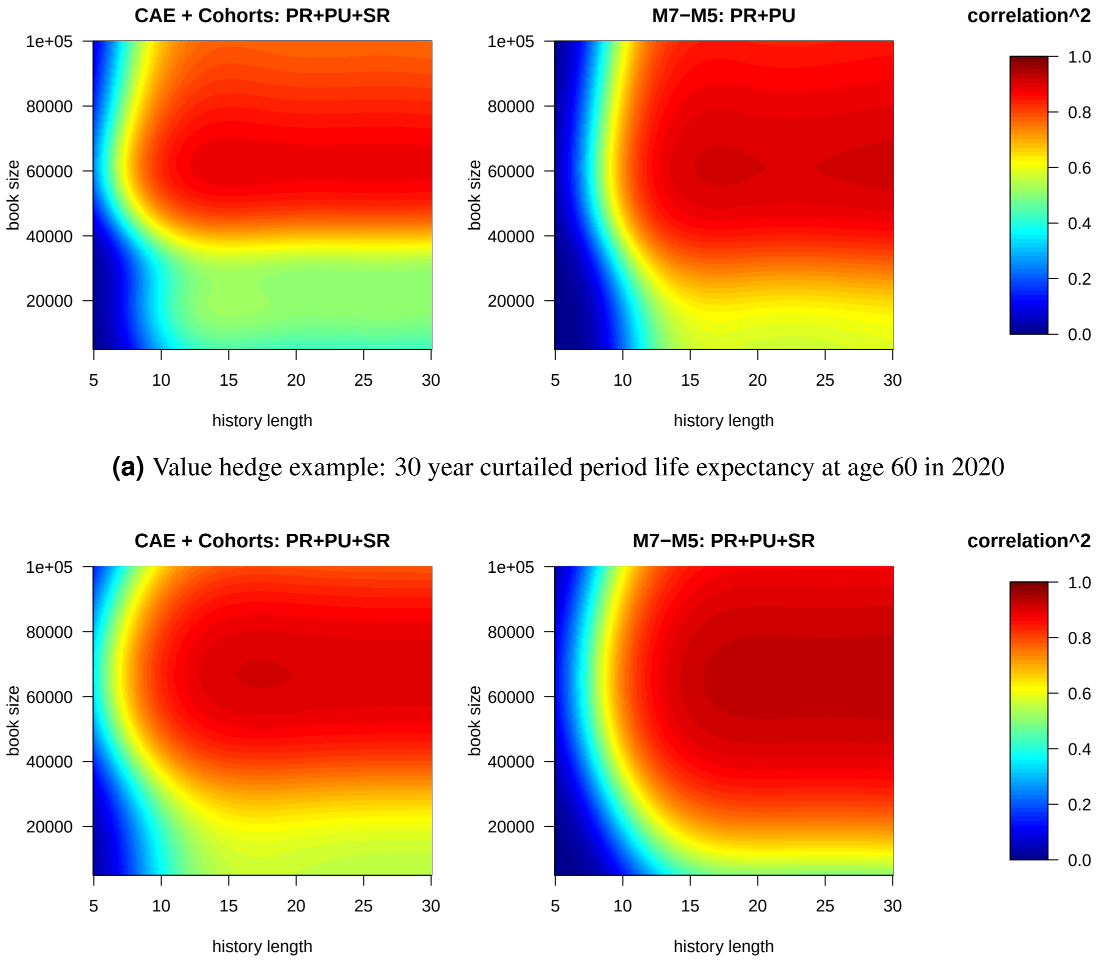
fall rapidly for model CAE+Cohorts, M7-M5 and APC (Gravity). This reinforces the previously discussed issues of fitting time series models when historical data are scarce.

7.4.3. Interaction between book size and history length

We end this section by investigating the interaction between book size and history length in the assessment of hedge effectiveness. To do so, we have fitted the CAE+Cohort and M7-M5 model to the "Extreme Wealthy" test book considering all possible combinations between book sizes 5 000, 7 500, 10 000, 15 000, 20 000, 25 000, 30 000, 40 000, 50 000, 75 000, and 100 000 and history lengths 5, 6, 7, 8, 9, 10, 12, 15, 20, 25 and 30 years. Figure 13 presents heatmaps depicting for both the value-hedge and cash-flow hedge examples the resulting hedge effectiveness values once all sources of risks have been considered. To ease the identification of patterns, correlations have been smoothed along the book size and history length dimensions.[12] From Figure 13 we note the following:

- The interaction between history length and book size is minimal, with hedge effectiveness falling rapidly for history lengths shorter than 10-12 years and book sizes smaller than 15,000-25,000 exposed lives above age 60.
- While for model M7-M5 correlations start to fall significantly for history length below 12 years, for the the CAE+cohorts correlations only start to show a material

[12] Smoothing has been performed using a generalised additive model of the form $\text{logit}\,\rho = s(size) + s(length)$, where, s denotes a penalised spline. For smoothing we have used **R** package **mgcv** (Wood, 2015).



(a) Value hedge example: 30 year curtailed period life expectancy at age 60 in 2020

(b) Cash flow hedge example: 25 year curtailed cohort life expectancy at age 65 in 2011

Figure 13. Smoothed squared correlation, ρ^2, between the liability L and the hedging instrument H, as a function of the book size and history length of the book population. All values correspond to the “Extreme Wealthy” socio-economic composition.

decline for history length below 10 years. This suggest that when historical data are limited, models with fewer book specific period indexes should be preferred over models with multiple period book specific period terms. However, in all cases the fitting of two-populations models should only be pursued when book data exceeds 8-10 years of history.

- For a book size over 25,000 lives and history length above 12 years, the hedge effectiveness reductions for both the value-hedge and the cash-flow hedge examples are above 70%, suggesting that index-based hedges can be a meaningful alternative for hedging longevity risk.

8. DISCUSSION AND CONCLUSIONS

The main conclusions of our systematic assessment of the alternative two-population mortality models for basis risk assessment can be summarised as follows. First, as can be expected, none of the models satisfy all the desirable practical criteria of a practical for assessing basis risk laid down in Section 5. However, M7-M5 and CAE+Cohorts stand out as the models which provide the most suitable balance between flexibility, simplicity, parsimony, goodness-of-fit to data, and forecasting performance. Both models produce reasonable best estimate projections with plausible levels of uncertainty, but with sufficient differences to suggest that model risk should be recognised as an important issue.

As our analysis suggests, the paucity of book data implies that it is difficult to estimate the age modulating terms $\beta_x^{(j,B)}$ without resulting in non-robust and erratic parameter estimates. Thus, any parameter which moderates the sensitivity of the book to time trends at different ages should be inherited from the reference population (i.e. $\beta_x^{(j,B)} \equiv \beta_x^{(j,R)}$). Furthermore, unless there is a strong reason to believe in the existence of a different cohort effect in the book than in the reference population, the parameter uncertainty in fitting a (non-parametric) book-specific cohort term will greatly outweigh any benefits in terms of goodness-of-fit to historical experience.

The fitting of two-population models should in principle only be pursued when two requirements are met. First the book annual exposure should be over 20,000-25,000 lives, since for smaller exposures the impact of parameter uncertainty may result in a biased estimate of basis risk. Secondly, there should be at least 8-10 years of reliable book data, since for shorter history lengths the quality of the forecasts is likely to be poor.

The above conclusions are underpinned by the analysis based on England and Wales population data and the profile of sample schemes drawn from the Club Vita database. We would expect many of the key conclusions to hold for other populations, although specific results (such as AIC rankings) are necessarily dependent on the choice of data.

We end this paper by making a number of general comments arising from our investigations.

Our previous sections have suggested that M7-M5 or the CAE+cohorts are appropriate models when undertaking the modelling of the mortality of the reference and the book populations in a basis risk assessment exercise. However, this need not preclude the consideration of additional models. Indeed, the modeller may wish to look at alternative models as part of sensitivity testing; or in order to gain a better understanding of model risk; or to err on the side of adding more features into the model than historic back-testing alone might suggest these features may be needed as part of a personal belief regarding the complexity of mortality. Further, as time goes on, new models will enter the actuarial literature and our work can help integrate those models into a basis risk assessment. Therefore, we next provide some general guidelines for the construction of two-population models for basis risk assessment.

When building a two-population model for assessing longevity basis risk, it is usual to find that the reference population is considerably larger and has a longer back history of data than the book population. It is therefore natural to start by selecting an appropriate model for the reference population. Once the reference population model is chosen a reasonable approach would be to select the book part of the model from within the same model family of the reference part. This will ensure a correspondence between the model parameters in the book and the reference populations which facilitates interpretation of the parameters of the models and makes the subsequent analysis more comprehensive and consistent in both populations. Our research on different models has also identified the following facts: it is in general enough to include at most two book-specific time-dependent terms; any parameter which moderates the sensitivity of the book to these time trends at different ages should be inherited from the reference book (i.e. $\beta_x^{(j,B)} \equiv \beta_x^{(j,R)}$); finally, it is generally appropriate not to include a book specific cohort effect. In mathematical terms, if the preferred reference population model is given by

$$\text{logit}\, q_{xt}^{R} = \alpha_x^R + \sum_{j=1}^{N} \beta_x^{(j,R)} \kappa_t^{(j,R)} + \gamma_{t-x}^{R},$$

then a good starting point for the book model would in general be of the form:

$$\text{logit}\, q_{xt}^{B} - \text{logit}\, q_{xt}^{R} = \alpha_x^B + \sum_{j=1}^{M} \beta_x^{(j,R)} \kappa_t^{(j,B)}.$$

We would usually expect M to be at most two as it is unlikely that the book population can support more than two time series i.e. $M \leq \min(2,N)$.[13] By way of example, if we choose to model the reference population using the single population model described in Börger et al. (2013):

$$\text{logit}\, q_{xt}^{R} = \alpha_x^R + \kappa_t^{(1,R)} + (x-\bar{x})\kappa_t^{(2,R)} + (x_{\text{young}} - x)^{+}\kappa_t^{(3,R)} + (x - x_{\text{old}})^{+}\kappa_t^{(4,R)} + \gamma_{t-x}^{R},$$

where x_{young} and x_{old} are predefined constant, then a suitable starting point for the book model would be

$$\text{logit}\, q_{xt}^{B} - \text{logit}\, q_{xt}^{R} = \alpha_x^B + \kappa_t^{(1,B)} + (x-\bar{x})\kappa_t^{(2,B)}.$$

Our systematic analysis of the two-population mortality literature has focused exclusively on the use of these models for the assessment of basis risk in longevity hedges. Furthermore, we have implicitly assumed that the target book population is a subset or is closely related to the reference population on which the index is based. Hence, our conclusions may not necessarily extend directly to other applications of two-population mortality models and the evaluation of the suitability of a model will largely depend on the task at hand (e.g whether it is a basis risk assessment exercises or not) and on the nature of the relationship between the two populations being modelled. As we have repeatedly discussed in this paper, simpler models that are not suitable for basis risk assessment (e.g, because of their implied perfect correlation between the populations) may be suitable for other applications such us when valuing pension liabilities or pricing annuities. In addition, the use of two-population mortality models for assessing the basis risk in longevity hedges where the mortality in one country is hedged with the mortality of another country[14] would require a deep understanding of the differences between the two countries' mortality. Such differences may not be captured by the structure of the two-population models we have proposed and the relative approach we have pursued may have to be substituted by a simultaneous modelling of the two countries' mortality, for instance along the lines of the work of Li et al. (2015) or the GLM modelling approach of Hatzopoulos and Haberman (2013).

In all our mortality projections and simulations we have employed the usual assumption that the spread between the mortality in the reference and the book will conform to the

[13] Note that the M7-M5 model and the CAE+cohorts can be derived from this form by applying the previous rules if we start by modelling the reference population using an M7 model or a LC+Cohorts model, respectively.

[14] An example of this is the Kortis bond where UK mortality is hedged using US mortality, see Hunt and Blake (2015a).

non-divergence hypothesis in the long run, i.e., that the ratio of q_{xt}^{B} to q_{xt}^{R} will tend to a limiting distribution as $t \to \infty$. We have captured this non-divergence constraint via the use of a (vector) autoregressive process for the time series indices (κ_t^B) in the book part of the model, implying that in the long-run the spread between the logit of mortality for the book and the reference population will revert from the current level to the historical mean. Although the investigation of the appropriateness of this assumption is out of the scope of this paper, the unsatisfactory results we have obtained when modelling the "Extreme Deprived" book population suggest that such an assumption may not be appropriate in all cases. In addition, the non-divergence assumption implies that the variance of the difference in (logit) mortality between the book and the reference population is bounded, potentially understating demographic basis risk and hence overstating the hedge effectiveness. We thus encourage further research looking at alternative choices of times series model and at the implications that such choices may have on hedge-effectiveness.

Our investigations indicate that the accurate calibration and projection of a two-population model requires that the annual exposure in the book population is over 20,000-25,000 lives and that there are at least 10-12 years of reliable book data. However, in practice a large proportion of pension scheme books and life company portfolios will not meet these data requirements leaving open the question of how to assess longevity basis risk and hedge effectiveness for such populations. If book size is the main issue, then a Bayesian approach such as those considered in Cairns et al. (2011a) and in Antonio et al. (2015) may offer an alternative. But, if the problem is the lack of sufficiently long historical data, indirect approaches where the book is modelled indirectly by reference to a bigger population with a more reliable and longer mortality experience could be the way through. Such approaches have recently been considered in the "mixing" approach proposed by Ahcan et al. (2014) and in the "characterisation" approach introduced in Haberman et al. (2014), and we believe that this line of research deserves further consideration.

Finally, although we have only considered very stylised longevity hedges, our hedge-effectiveness results show that index-based hedges have the potential to provide an effective and flexible solution to mitigate longevity risk. We hope that our research has shed light on the assessment of basis risk and contributes to moving forward the market of standardised longevity transactions.

ACKNOWLEDGEMENTS

This work forms part of a larger research project completed by a joint team from Cass Business School and Hymans Robertson LLP in response to a commission from the Longevity Basis Risk Working Group of the Institute & Faculty of Actuaries and the Life & Longevity Markets Association – see Haberman et al. (2014). We thank Steven Baxter, Andrew Gaches, Sveinn Gunnlaugsson and Mario Sison from Hymans Robertson LLP for their significant contribution to this research. We also thank the Longevity Basis Risk Working Group and Andrew Hunt for their feedback which has helped improved this work. Earlier versions of the paper have been presented at EM Lyon, University of Venice, University of New South Wales, Georgia State University, University of Hamburg, PARTY 2015, IME 2015 in Liverpool, Longevity 11 in Lyon and the Conference on Actuarial Risks in the Solvency II Era in Bologna.

REFERENCES

Ahcan, A., Medved, D., Olivieri, A., and Pitacco, E. (2014). Forecasting mortality for small populations by mixing mortality data. *Insurance: Mathematics and Economics*, 54:12–27.

Ahmadi, S. S. and Li, J. S.-H. (2014). Coherent mortality forecasting with generalized linear models: A modified time-transformation approach. *Insurance: Mathematics and Economics*, 59:194–221.

Antonio, K., Bardoutsos, A., and Ouburg, W. (2015). Bayesian Poisson log-bilinear models for mortality projections with multiple populations. *European Actuarial Journal*, 5(2):245–281.

Biatat, V. and Currie, I. D. (2010). Joint models for classification and comparison of mortality in different countries. In *Proceedings of 25rd International Workshop on Statistical Modelling, Glasgow*, pages 89–94.

Booth, H., Hyndman, R. J., Tickle, L., and de Jong, P. (2006). Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. *Demography*, 15:289–310.

Börger, M., Fleischer, D., and Kuksin, N. (2013). Modeling the mortality trend under modern solvency regimes. *ASTIN Bulletin*, 44(1):1–38.

Brouhns, N., Denuit, M., and Van Keilegom, I. (2005). Bootstrapping the Poisson log-bilinear model for mortality forecasting. *Scandinavian Actuarial Journal*, (3):212–224.

Butt, Z. and Haberman, S. (2009). Ilc: A collection of R functions for fitting a class of Lee-Carter mortality models using iterative fitting algorithms. *Actuarial Research Paper, Cass Business School.*

Cairns, A. J. G. (2013). Robust hedging of longevity risk. *Journal of Risk and Insurance*, 80:621–648.

Cairns, A. J. G., Blake, D., and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *Journal of Risk and Insurance*, 73(4):687–718.

Cairns, A. J. G., Blake, D., and Dowd, K. (2008). Modelling and management of mortality risk: a review. *Scandinavian Actuarial Journal*, (2):79–113.

Cairns, A. J. G., Blake, D., Dowd, K., and Coughlan, G. D. (2011a). Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin*, 41:29–59.

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., and Khalaf-Allah, M. (2011b). Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, 48(3):355–367.

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

Cairns, A. J. G., Dowd, K., Blake, D., and Coughlan, G. D. (2014). Longevity hedge effectiveness: a decomposition. *Quantitative Finance*, 14(2):217–235.

Carter, L. R. and Lee, R. D. (1992). Modeling and forecasting US sex differentials in mortality. *International Journal of Forecasting*, 8(3):393–411.

Continuous Mortality Investigation (2007). Working Paper 25 – Stochastic projection methodologies: Lee–Carter model features, example results and implications.

Coughlan, G. D., Khalaf-Allah, M., Ye, Y., Kumar, S., Cairns, A. J. G., Blake, D., and Dowd, K. (2011). Longevity hedging 101: A framework for longevity basis risk analysis and hedge effectiveness. *North American Actuarial Journal*, 15(2):150–176.

Currie, I. D., Durban, M., and Eilers, P. H. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, 4(4):279–298.

Debón, A., Martínez-Ruiz, F., and Montes, F. (2010). A geostatistical approach for dynamic life tables: The effect of mortality on remaining lifetime and annuities. *Insurance: Mathematics and Economics*, 47(3):327–336.

Debón, A., Montes, F., and Martínez-Ruiz, F. (2011). Statistical methods to compare mortality for a group with non-divergent populations: an application to Spanish regions. *European Actuarial Journal*, 1(2):291–308.

Delwarde, A., Denuit, M., Guillén, M., and Vidiella-i Anguera, A. (2006). Application of the Poisson log-bilinear projection model to the G5 mortality experience. *Belgian Actuarial Bulletin*, 6(1):54–68.

Dowd, K., Cairns, A. J. G., Blake, D., Coughlan, G. D., Epstein, D., and Khalaf-Allah, M. (2010). Backtesting stochastic mortality models: An ex-post evaluation of multi-period-ahead density forecasts. *North American Actuarial Journal*, 14(3):281–298.

Dowd, K., Cairns, A. J. G., Blake, D., Coughlan, G. D., and Khalaf-allah, M. (2011). A gravity model of mortality rates for two related populations. *North American Actuarial Journal*, 15(2):334–356.

Haberman, S., Kaishev, V. K., Millossovich, P., Villegas, A. M., Baxter, S., Gaches, A., Gunnlaugsson, S., and Sison, M. (2014). Longevity Basis Risk: A methodology for assessing basis risk. *Institute and Faculty of Actuaries Sessional Research Paper*. Available from: `http://www.actuaries.org.uk/documents/longevity-basis-risk-methodology-assessing-basis-risk`.

Haberman, S. and Renshaw, A. (2009). On age-period-cohort parametric mortality rate projections. *Insurance: Mathematics and Economics*, 45(2):255–270.

Haberman, S. and Renshaw, A. (2011). A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics*, 48(1):35–55.

Hatzopoulos, P. and Haberman, S. (2013). Common mortality modeling and coherent forecasts. An empirical analysis of worldwide mortality data. *Insurance: Mathematics and Economics*, 52(2):320–337.

Human Mortality Database (2013). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available from: `www.mortality.org`.

Hunt, A. and Blake, D. (2015a). Modelling longevity bonds: Analysing the Swiss Re Kortis bond. *Insurance: Mathematics and Economics*, 63:12–29.

Hunt, A. and Blake, D. (2015b). On the Structure and Classification of Mortality Models Mortality Models. *Pension Institute Working Paper*. Available from: `http://www.pensions-institute.org/workingpapers/wp1506.pdf`.

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

Hymans Robertson LLP (2015). Buy-outs, buy-ins and longevity hedging, Q4 2014. Available from: `http://www.hymans.co.uk/media/591924/150317-managing-pension-scheme-risk-q4-2014.pdf`.

Hyndman, R. J., Booth, H., and Yasmeen, F. (2013). Coherent mortality forecasting: the product-ratio method with functional time series models. *Demography*, 50(1):261–283.

Jarner, S. F. and Kryger, E. M. (2011). Modelling adult mortality in small populations: The Saint Model. *ASTIN Bulletin*, 41(2):377–418.

Kleinow, T. (2015). A Common Age Effect Model for the Mortality of Multiple Populations. *Insurance: Mathematics and Economics*, 63:147–152.

Koissi, M.-C., Shapiro, A., and Hognas, G. (2006). Evaluating and extending the Lee-Carter model for mortality forecasting: Bootstrap confidence interval. *Insurance: Mathematics and Economics*, 38(1):1–20.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87(419):659–671.

Li, J. (2012). A Poisson common factor model for projecting mortality and life expectancy jointly for females and males. *Population Studies*, 67(1):111–126.

Li, J. S.-H. and Hardy, M. R. (2011). Measuring basis risk in longevity hedges. *North American Actuarial Journal*, 15(2):177–200.

Li, J. S.-H., Zhou, R., and Hardy, M. (2015). A step-by-step guide to building two-population stochastic mortality models. *Insurance: Mathematics and Economics*, 63:121–134.

Li, N. and Lee, R. D. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, 42(3):575–594.

LLMA (2012). Basis risk in longevity hedging: parallels with the past. *Institutional Investor Journals*, 2012(1):39–45.

Lu, J. L. C., Wong, W., and Bajekal, M. (2014). Mortality improvement by socio-economic circumstances in England (1982 to 2006). *British Actuarial Journal*, 19(1):1–35.

Noble, M., Mclennan, D., Wilkinson, K., Whitworth, A., Exley, S., Barnes, H., and Dibben, C. (2007). *The English indices of deprivation 2007*. Department of Communities and Local Government, London.

Plat, R. (2009a). On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45(3):393–404.

Plat, R. (2009b). Stochastic portfolio specific mortality and the quantification of mortality basis risk. *Insurance: Mathematics and Economics*, 45(1):123–132.

Renshaw, A. and Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, 38(3):556–570.

Renshaw, A. and Haberman, S. (2008). On simulation-based approaches to risk measurement in mortality with specific reference to Poisson Lee-Carter modelling. *Insurance: Mathematics and Economics*, 42(2):797–816.

Russolillo, M., Giordano, G., and Haberman, S. (2011). Extending the Lee-Carter model: a three-way decomposition. *Scandinavian Actuarial Journal*, (2):96–117.

Villegas, A. M. and Haberman, S. (2014). On the modeling and forecasting of socioeconomic mortality differentials: an application to deprivation and mortality in England. *North American Actuarial Journal*, 18(1):168–193.

Villegas, A. M., Kaishev, V., and Millossovich, P. (2015). StMoMo : An R Package for Stochastic Mortality Modelling. Available from: `http://cran.r-project.org/package=StMoMo`.

Wan, C. and Bertschi, L. (2015). Swiss coherent mortality model as a basis for developing longevity de-risking solutions for Swiss pension funds: A practical approach. *Insurance: Mathematics and Economics*, 63:66–75.

Willets, R. (2004). The cohort effect: Insights and explanations. *British Actuarial Journal*, 10(4):833–877.

Wilmoth, J. and Valkonen, T. (2001). A parametric representation of mortality differentials over age and time. In *Fith seminar of EAPS Working Group on Differential in Health, Morbidity and Mortality in Europe*.

Wood, S. (2015). Package mgcv. Available from: `http://cran.r-project.org/web/packages/mgcv/index.html`.

Yang, B., Li, J., and Balasooriya, U. (2016). Cohort extensions of the Poisson common factor model for modelling both genders jointly. *Scandinavian Actuarial Journal*, (2):93–112.

Yang, S. S. and Wang, C.-W. (2013). Pricing and securitization of multi-country longevity risk with mortality dependence. *Insurance: Mathematics and Economics*, 52(2):157–169.

Zhou, R., Wang, Y., Kaufhold, K., Li, J. S.-H., and Tan, K. S. (2014). Modeling period effects in multi-population mortality models: applications to Solvency II. *North American Actuarial Journal*, 18(1):150–167.

A. GENERATION OF SYNTHETIC DATA

In this appendix we present a possible procedure for generating, based on a reference dataset, synthetic mortality datasets which have a given exposure size with a given distribution of this exposure across population subgroups.

Assume that we have a reference dataset containing observed number of deaths D_{xtg} in year t for people age x in subgroup g with matching central exposures E_{xtg} and matching death rates $\mu_{xtg} = D_{xtg} \big/ E_{xtg}$. Let $C_t^{'}$ be the target total exposure for year t in the synthetic dataset and $(w_{tg_1}^{'}, \ldots, w_{tg_m}^{'})$ be a vector of weights adding to one which represents the desired splitting in year t of this exposure among the subgroups.

The synthetic central exposures $E_{xtg}^{'}$ in year t for people age x in subgroup g are obtained as

$$E_{xtg}^{'} = C_t^{'} \frac{\sum_g E_{xtg}}{\sum_x \sum_g E_{xtg}} w_{tg}^{'} = C_t^{'} \frac{E_{xt}}{E_t} w_{tg}^{'},$$

where $E_{xt} = \sum_g E_{xtg}$ are the total exposed to risk at age x in year t across all groups and $E_t = \sum_x \sum_g E_{xtg}$ are the total exposed to risk in year t across all groups and ages. Hence the exposure for the reference dataset is being used to obtain the split by age for a particular year and group. The corresponding synthetic number of deaths $D_{xtg}^{'}$ is generated by drawing a random sample from a Poisson distribution with mean $E_{xtg}^{'} \mu_{xtg}$. It should be mentioned that the use of raw death rates may inflate the variability in the simulated numbers of deaths. However, in the present application, based on the large UK population, the extent of this additional variability is limited. Different applications based on smaller populations may require the preliminary smoothing of death rates.

B. MODEL FITTING CONSTRAINTS

Some of the models require parameter constraints to ensure identifiability of the parameters. Table 7 presents the parameter constraints imposed to the reference population models and Table 8 shows the parameter constraints imposed to the book part of the two-population models. It is well known that cohort extensions of the Lee-Carter model have robustness and stability issues with models being very sensitive to changes in the data or the fitting algorithm (see e.g. Hunt and Villegas (2015)). Therefore, when implementing the

LC+Cohorts model we follow the approach suggested in Hunt and Villegas (2015) which helps resolve many of the stability issues.

Table 7. Parameter constraints for the reference population models.

Model	**Constraints**
LC+Cohorts	$\sum_x \beta_x^R = 1, \sum_t \kappa_t^R = 0, \sum_{t-x} \gamma_{t-x}^R = 0, \sum_{t-x}(t-x)\gamma_{t-x}^R = 0$
APC	$\sum_t \kappa_t^R = 0, \sum_{t-x} \gamma_{t-x}^R = 0, \sum_{t-x}(t-x)\gamma_{t-x}^R = 0$
M7	$\sum_{t-x} \gamma_{t-x}^R = 0, \sum_{t-x}(t-x)\gamma_{t-x}^R = 0, \sum_{t-x}(t-x)^2\gamma_{t-x}^R = 0$

Table 8. Parameter constraints for the book part of the models.

Model	**Constraints**
CF+Cohorts	-
CAE+Cohorts	$\sum_t \kappa_t^B = 0$
RelLC+Cohorts	$\sum_x \beta_x^B = 1, \sum_t \kappa_t^B = 0$
Gravity (APC)	$\sum_t \kappa_t^B = 0, \sum_{t-x} \gamma_{t-x}^B = 0, \sum_{t-x}(t-x)\gamma_{t-x}^B = 0$
M7-M5	-
M7-M6	$\sum_{t-x} \gamma_{t-x}^B = 0, \sum_{t-x}(t-x)\gamma_{t-x}^B = 0$
M7-M7	$\sum_{t-x} \gamma_{t-x}^B = 0, \sum_{t-x}(t-x)\gamma_{t-x}^B = 0, \sum_{t-x}(t-x)^2\gamma_{t-x}^B = 0$
M7-Saint	-
M7-PLAT	-