Quantifying longevity gaps using micro-level lifetime data

van Berkum, F.; Antonio, K.; Vellekoop, M.

DOI
10.1111/rssa.12631
Publication date
2021
Document Version
Final published version
Published in
Journal of the Royal Statistical Society. Series A: Statistics in Society

Received: 26 July 2018 | Accepted: 24 July 2020

DOI: 10.1111/rssa.12631

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

ORIGINAL ARTICLE

Quantifying longevity gaps using micro-level lifetime data

Frank van Berkum[1] | **Katrien Antonio**[1,2] | **Michel Vellekoop**[1]

[1]Faculty of Economics and Business, University of Amsterdam, Amsterdam, The Netherlands

[2]Faculty of Economics and Business, KU Leuven, Leuven, Belgium

Correspondence
Frank van Berkum, Faculty of Economics and Business, University of Amsterdam, The Netherlands.
Email: f.vanberkum@uva.nl

Abstract

Using flexible Poisson regressions, we analyse a huge micro-level lifetime dataset from a Dutch pension fund, including categorical, continuous and spatial risk factors collected on participants in the fund. The availability of granular lifetime data allows us to quantify the longevity gap between the national population and the fund on the one hand, and between participants within the fund on the other hand. We identify the most important risk factors using statistical criteria that measure the in- and out-of-sample performance of the regression models. We evaluate the financial performance of the models by introducing a novel type of backtest, which selects the risk factors that contribute most to an accurate prediction of future pension liabilities. For this portfolio, the most relevant risk factors (next to age and gender) are the salary, the time spent in disability and working at irregular hours. The resulting personalized mortality risk profiles show substantial differences between the remaining life expectancies for the most-favourable and least-favourable risk profiles. Our method to estimate these longevity gaps will help policymakers to assess wanted and unwanted consequences of longevity risk sharing in pension schemes.

KEYWORDS

generalized additive models, longevity gaps, portfolio-specific mortality

1 | INTRODUCTION

The study of factors influencing mortality has been the subject of many contributions in economics, medicine and actuarial science. For example, Brown and McDaid (2003) perform a qualitative literature review of 45 research papers that consider a variety of risk factors to explain differences in mortality after retirement. Relevant factors are (besides age and gender) education, income, tobacco and alcohol consumption, and marital status. Elo and Preston (1996) study adult US mortality in the period 1975–1985 with a logistic regression model for ${}_5q_x$, the probability of dying within a 5-year period for an x-year-old. For individuals with more than 16 years of education this probability is about 40% lower than for individuals with no or less than 8 years of education, even after correcting for factors such as income and race. Chetty et al. (2016) reveal a gap in the remaining period life expectancy at age 40 between the 1% richest and 1% poorest individuals in the US of about 14.6 years for males and 10.1 years for females, based on observations over the period 2001–2014.

Datasets that are collected by insurance companies and pension funds create an excellent opportunity to study such 'longevity gaps'. The future random lifetime of a participant plays a crucial role in the design of pension plans as well as the valuation of pension fund liabilities. The distribution of the future lifetime is summarized in life tables with entries q_x, the 1-year probability of dying (or: mortality rate) for an x-year-old. In most developed countries governmental bodies as well as professional associations of actuaries regularly publish age- and gender-specific population-wide mortality rates. For example, since the 1970s the Royal Dutch Actuarial Association publishes population-wide period life tables. Moreover, since 2007 the association also publishes mortality forecasts which are currently updated biennially (Antonio et al., 2017). Pension funds must value their liabilities with mortality rates suitable for the participants in the fund or portfolio, and can not simply use the published rates. Typically, participants in a fund experience mortality rates that are different from those applicable to the population due to differences in socioeconomic characteristics, life styles and professions. In view of the design of future-proof, sustainable pension systems, the micro-level lifetime data collected by pension fund managers are of utmost importance. In the era of big data and data analytics, careful analysis of the granular data collected on the lifetimes of individual participants, as well as their characteristics or risk factors, allows to quantify the longevity gap between the national population and the fund on the one hand, and among participants within the fund on the other hand. Pension funds are now able to incorporate the insights derived from such data when valuing their liabilities, for example, by distinguishing different risk profiles. In this paper we develop a statistical framework to construct such profiles and to assess their impact on death count predictions on the one hand and pension liability calculations on the other hand.

Historically, actuarial practice quantified portfolio-specific mortality using crude methods such as age shifting, which replaces the probability q_x applicable to the national population by q_{x+s} where s can be either positive or negative (Pitacco et al., 2009). Such approaches are easy to implement, but potentially highly inaccurate. Moreover, observed mortality in the portfolio is often expressed in terms of the value of accrued rights. Mortality is then characterized by the fraction of the total value of accrued rights that has been released in a year, see for example, Plat (2009). Using this alternative definition, the mortality rates (also referred to as *insured amount weighted mortality rates* when applied in insurance companies instead of pension funds) can be very volatile over the years due to the presence of individual mortality risk.[1] By using this alternative definition, more weight is given to members with high accrued rights when explaining historical portfolio mortality. Since the liabilities of a pension

[1]Individual mortality risk is the uncertainty associated with the binary outcome of survival *given* a fixed mortality rate.

fund are a function of accrued rights, this approach may lead to more accurate predictions than when the accrued rights are neglected. In this paper, we analyse the influence of accrued rights on the mortality rates of fund participants in a more explicit way by including these rights (or some related risk factor) directly in a regression model.

Gschlössl et al. (2011) and Richards et al. (2013) present first attempts to include risk factors in mortality models designed for lifetime data collected on individuals registered in an insurance portfolio or pension fund. Holford (1980) and Laird and Olivier (1981) (in statistical literature) and Brouhns et al. (2002) (in actuarial literature) demonstrate that the survival likelihood with piece-wise constant force of mortality (or hazard rate) is equivalent to a Poisson regression model for the number of deaths. Hence, Gschlössl et al. (2011) use such a Poisson regression model for the observed death counts in a German life insurance portfolio. Their dataset contains risk factors such as product type, policy duration and insured amount. First, they estimate a static baseline mortality rate in the portfolio, and then they explain remaining heterogeneity in the portfolio using categorical risk factors in a generalized linear model. Richards et al. (2013) explain the heterogeneity in mortality among participants of a German multi-employer pension scheme using a Makeham–Beard law for the force of mortality of an x-year-old. They let the parameters in this mortality law depend on risk factors such as scheme size and the health status of members. Whereas Gschlössl et al. (2011) and Richards et al. (2013) discretize the continuous risk factors available, Denuit and Legrand (2018) use generalized additive models (GAMs) (Hastie & Tibshirani, 1986; Wood, 2017) and directly include continuous covariates in the construction of mortality profiles.

Individual lifetime data observed in continuous time are often modelled using a Cox proportional hazard model (Cox-PH) in which a baseline hazard rate is shared by all individuals in the study (Cox, 1972). Covariates or risk factors are included in a multiplicative way and quantify differences between the survival functions of individuals. Czado and Rudolph (2002) use such Cox-PH models to analyse individual time-to-event data, possibly subject to right censoring, in the presence of time-fixed and time-varying risk factors. Our main goal is the estimation of mortality rates in the presence of multiple types of risk factors (including categorical, continuous and spatial), instead of the survival function in continuous time. Therefore, we focus on the modelling strategy in Denuit and Legrand (2018) and extend their framework to very large datasets, in the presence of multiple types of risk factors.

A different stream of literature investigates pension fund mortality with multiple population mortality models. Villegas and Haberman (2014) use multi-population extensions of the Lee and Carter (1992) model to predict mortality for five socioeconomic classes in England. These socioeconomic classes are based on characteristics such as income and education. Cairns et al. (2019) analyse mortality of Danish males for 10 socioeconomic groups constructed using an affluence index based on wealth and reported income. They use an extension of the gravity model from Dowd et al. (2011) to model mortality in the different affluence groups. However, both approaches are based on datasets that contain over 20 years of observations, and they consider separate time dynamics for each group. Such methods are no longer feasible if the dataset has only a few years of data available. Moreover, these papers consider the socioeconomic classes as exogenously given, while our contribution investigates the construction of such classes or risk profiles directly from micro-level lifetime data.

We work on a unique dataset from a large Dutch pension fund covering the period 2006–2011. In this dataset individuals in the fund are followed over time, and their risk factors such as age, gender, salary, disability status and postal code are recorded on a yearly basis. As a first contribution this paper shows how to construct mortality forecasts for individual risk profiles using this dataset. We first estimate a multiple population mortality model (Koninklijk Actuarieel Genootschap, 2014) which provides an appropriate baseline mortality level and at the same time allows to construct mortality forecasts for the Dutch population. Using the baseline, we explain remaining heterogeneity in the

portfolio using Poisson regression, as suggested by Gschlössl et al. (2011) and Denuit and Legrand (2018). We consider a wide variety of possible risk factors (including continuous and spatial variables) in a GAM framework to ensure that all information in these variables is adequately captured in our model.

As a second contribution, we provide a complete framework to determine which risk factors should be used to explain historical portfolio mortality data. We use the conditional Akaike information criterion (cAIC) for variable and model selection (Wood et al., 2016). However, a drawback of information criteria is that they focus solely on how well risk factors are able to explain in-sample variation. Therefore, we also use cross-validation checks and proper scoring rules to investigate the predictive power of the estimated regression models. Furthermore, we investigate the robustness of the estimated effects across cross-validated datasets, by comparing the recalibrated effects to the initial effect fitted on the complete dataset.

Our final contribution is the design of a novel type of backtest to evaluate the performance of mortality models. Whereas we model observed death counts, risk managers in a pension fund are especially interested in an accurate prediction of the value of the liabilities. Therefore, we estimate the different models on all data excluding the most recent year. Using the estimated effects, we determine individual risk profiles and use these to construct prediction intervals for the liabilities needed at the end of this year. We compare these prediction intervals with the actual liabilities needed for surviving members. We also calculate the mean squared prediction error (MSPE) for the predicted liabilities. This enables us to select the risk factors that contribute most to an accurate prediction of the liabilities, which is of more importance for the risk management strategy of a fund than the predictive accuracy in terms of individual death counts.

The remainder of this paper is organized as follows. We discuss the dataset used in this paper in Section 2. In Section 3 we introduce the GAMs that we will use to explain observed portfolio mortality, and we discuss how we assess in-sample and out-of-sample performance. We present estimation and backtesting results in Section 4 and we conclude in Section 5.

2 | DATA

We use a large dataset from a Dutch pension fund which follows participants during the period 2006 to 2011. Each participant in the dataset is assigned a unique identifier which allows us to follow individual participants over time. At the end of each year it is recorded whether a participant is still alive and the observable risk factors of the participant are updated. In our analysis we include information from active participants[2], pensioners, and people who are fully or partially disabled. For some individuals, however, one or multiple risk factors are missing in one or several years. In this section we first discuss the observations and risk factors that are available in the dataset, and then we discuss the presence of, and causes for, missing data in the dataset.

2.1 | Mortality observations and risk factors

Table 1 lists the variables constructed from the dataset and Figure 1 shows the distribution of these variables in our dataset. We include the ages 20 to 90; we exclude lower ages because they are not relevant

[2]Inactive participants no longer work at a company linked to the pension fund, and these members therefore no longer pay premium and no longer accrue new pension rights, but retain their existing rights.

TABLE 1 A description of the mortality observations and risk factors. The percentage on the right shows for which fraction of the observations the information is available

	Mortality information	
D	1 if the participant died at the current age, 0 if the participant survived	(100.0%)
E	The fraction of the year lived by the participant at the current age	(100.0%)
	Risk factors	
Year	Year of the observation	(100.0%)
Age	Age of the participant	(100.0%)
Gender	Gender of the participant	(100.0%)
Sal	Logarithm of last observed FTE salary on annual basis, normalized per year, age and gender (if applicable: including an allowance for working at irregular hours)	(88.8%)
IA	The percentage of the last observed FTE salary earned through an allowance for working at irregular hours	(88.8%)
DisTime	The cumulative disability spell of the participant, adjusted for partial disability	(96.0%)
DisPerc	The time spent in disability as a percentage of total service years registered	(96.0%)
AFPP	The age at which the participant received his first pension payment	(11.6%)
(Long, Lat)	The longitude and latitude that correspond to the centre of the four-digit postal code of the participant	(100.0%)
Edu	Average education level at the postal code where the participant resides (obtained from Statistics Netherlands)	(95.4%)

for the liabilities of the pension fund, and higher ages are excluded because their exposure turned out to be negligible in our dataset (less than 0.1% of the total exposure). The dataset contains 11,321,861 individual observations on 2,162,619 unique individuals resulting in a total of 11,300,883 years lived, and during the observed period 41,611 deaths were recorded. The risk factor Age changes when participants celebrate their birthday, so we split the individual observations into observations before and after the birthday of the participants. This results in 22,625,142 observations with constant risk factors.

As discussed in the introduction, we expect that salary influences the level of mortality. There are several variables available in the dataset that can be used to capture this effect: salary earned during a year, the fraction of the year the participant worked (hereafter: parttime factor), and the amount of salary earned from working at irregular hours. We define full time equivalent (FTE) salary as

$$\text{FTE salary} = \frac{\text{parttime salary earned in a year}}{\text{parttime factor}} + \text{irregularity allowance}.$$

Salaries tend to increase with age and as a result of inflation. Therefore, salaries at different ages and in different years cannot be compared directly, and we believe the FTE salary should not directly be used as a risk factor. Instead, we construct the variable Sal which is a normalized version of the logarithm of the last observed FTE salary. For each participant, we subtract the mean from the logarithm of the FTE salary and divide the result by its standard deviation, where mean and variance of the log transformed FTE salary are determined per year, age and gender. For most pensioners salary information is not available, but when it is, this corresponds to the last salary earned as an active participant.

Participants with irregular working hours are more likely to have an irregular sleeping pattern, so participants who earn a larger fraction of their total salary through an irregularity allowance may have a higher mortality rate, see Costa (1996). We define the variable IA to include this effect, and it is computed as the irregularity allowance divided by FTE salary. We divide by FTE salary because we

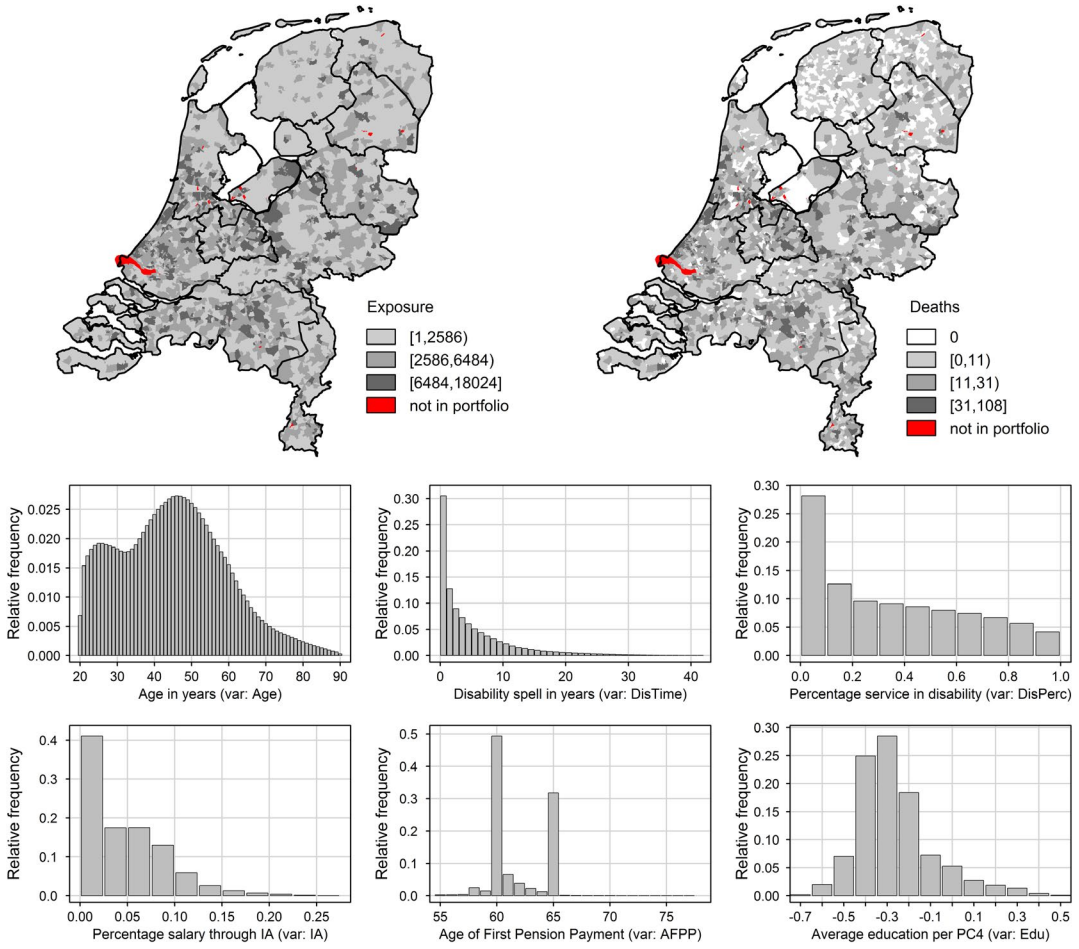

FIGURE 1 Empirical distribution of the variables in the dataset. Missing records are excluded, and for `DisTime` and `DisPer` we also do not show observations equal to zero. [Colour figure can be viewed at wileyonlinelibrary.com]

believe someone who works only 1 day a week at irregular hours should be treated differently from someone who works 5 days a week at irregular hours. This variable is positive for participants with an irregularity allowance (43.1% of the exposure), zero for participants who do not work at irregular hours (45.6%), and it is missing for other participants (11.3%).

The dataset also contains information on disability. In the Netherlands, people can be classified as being partially disabled. Someone who has been partially disabled at 40% for 3 years will have a cumulative disability spell of 1.2 years, and the years of service at the company will have increased by 1.8 years. The variable `DisTime` represents the cumulative disability spell for a participant, and `DisPerc` represents the fraction of working years spent in disability. The latter is defined as the cumulative disability spell divided by the sum of the number of service years and the cumulative disability spell.

Both the number of service years and the cumulative disability spell relate only to the active period at the pension fund; if a participant was active at another pension fund before joining the pension fund under consideration, that information is not available in this dataset. For the active participants and the fully or partially disabled participants, disability information is nearly always available but for retired members information on disability is often missing. In the dataset, 8.8% of the exposure has a positive disability spell, 87.2% has a disability spell equal to zero, and for 4.0% disability information is missing.

The official retirement age in the Netherlands during the observed period was 65 years, but many participants received pension payments before the age of 65. To stimulate the inflow of younger workers, many older workers close to the retirement age have had the opportunity to retire early (either fully or partially). We define the variable `AFPP` as the age at which the first pension payment was received, and we use this variable to investigate whether early retirement affects mortality. This variable is missing for participants who were not yet retired.

In the Netherlands an address is completely specified by a six-digit postal code and a number that specifies the house. Our dataset contains for nearly all participants a four-digit postal code (`PC`), which corresponds to a district in a city. Using the longitude (`Long`) and latitude (`Lat`) of the centre of the four-digit postal code we can estimate spatial heterogeneity in mortality rates.

Many studies have shown a link between education and mortality. Our dataset does not contain information on the education level of the participant, but we enrich our data with public data from the website of Statistics Netherlands on the level of education.[3] This dataset shows, per four-digit postal code, the fraction (i_c) of residents that has obtained education level c, where $c \in \{\text{Low, Medium, High}\}$. From this we construct the variable $\texttt{Edu} = -1 \cdot i_{\text{Low}} + 0 \cdot i_{\text{Medium}} + 1 \cdot i_{\text{High}}$, which is a weighted average of the education attained by the residents in a postal code. The variable ranges from -1 (if every resident is classified as having Low education) to $+1$ (if every resident is classified as having High education). The information on education is not provided in full for all postal codes, and this variable is therefore not available for all postal codes.

2.2 | Missing data

Our dataset contains missing data due to a number of different causes. With information collected on active, disabled and retired participants, the risk factors are often completely observed for active and disabled participants. However, for retired participants information regarding salary or disabled status is no longer recorded once pension payments start. For participants who retire during the observed period, we carry forward the salary and disability information from their last active period, since we believe this provides reliable information regarding their risk profile. This method of dealing with a missing observation is referred to as *last value carried forward* (LVCF) which is a single imputation method, see Bennett (2001). The occurrence of missing values is often dependent on `Age`, since for older participants variables are missing more often. Missing values for `Edu` are present since the education information is not available for all postal codes in the external dataset. In Section 3.1 we specifically discuss how we deal with missing variables in our model estimation procedure.

3 | A FRAMEWORK FOR STATISTICAL MODELLING OF PORTFOLIO MORTALITY

We analyse individual mortality data to investigate to what extent various risk factors can explain differences in observed mortality relative to some time-dependent population mortality rate, see also Denuit and Legrand (2018). By combining the estimated effects of multiple risk factors, we construct flexible participant-specific factors as opposed to portfolio-specific factors that depend on age and gender only. These participant-specific factors represent the relative difference in mortality between the participant and the general population.

[3]https://www.cbs.nl/-/media/imported/documents/2013/49/131203-opleiding-regelingen-verdachten-pc4-mw.xls

For notational convenience, we describe the framework for statistical modelling of portfolio mortality under the assumption that all participants celebrate their birthday on 1 January. In the online Supplementary Material we describe how the data definitions and model (assessment) formulas change if we take into account that birthdays occur throughout the year.

3.1 | Distributional assumptions and model estimation

We assume all participants celebrate their birthday on 1 January such that they have exact age x at the beginning of calendar year t with x and t integer. We have observations for participants $j = 1, \ldots, J_t$ where J_t equals the number of participants in year t. For each participant j, we define $\tau_{t,j,x}$ as the fraction of the year lived in year t at age x and define the corresponding indicator variable $\delta_{t,j,x}$ which equals 1 if the participant died in year t at age x and 0 otherwise.

3.1.1 | Model likelihood

We define the force of mortality for participant j at time t and age x by $\mu_{tjx} = \mu_{tjx}^{\text{pop}} \cdot \eta_{tjx}$. We assume that the baseline population force of mortality μ_{tjx}^{pop} is given for participant j aged x during calendar year t. The factor η_{tjx} represents the ratio between the population force of mortality and the force of mortality for participant j, and this factor must be estimated from the data.

In line with other literature (Cairns et al., 2009; Pitacco et al., 2009), we assume for integers t and x a constant force of mortality μ_{tjx} on the interval $[t, t + 1) \times [x, x + 1)$. The individual survival likelihood for participant j is then given by

$$p_{tjx} = \exp[-\mu_{tjx}] \tag{1a}$$

if he survives, and

$$\tau_{tjx} p_{tjx} \cdot \mu_{t+\tau_{tjx},\, x+\tau_{tjx}} = \exp[-\tau_{tjx}\mu_{tjx}]\mu_{tjx} \tag{1b}$$

if he dies. Substituting $\mu_{tjx} = \mu_{tjx}^{\text{pop}} \cdot \eta_{tjx}$ and aggregating over all participants, the likelihood for all individual survival observations is given by:

$$L(\eta_{tjx}) = \prod_{t=2006}^{2011} \prod_{j=1}^{J_t} \exp[-\tau_{tjx}\mu_{tjx}^{\text{pop}}\eta_{tjx}](\mu_{tjx}^{\text{pop}}\eta_{tjx})^{\delta_{tjx}}. \tag{2}$$

This likelihood is similar to the case where δ_{tjx} is a realization of a Poisson distributed random variable with parameter $\tau_{tjx}\mu_{tjx}^{\text{pop}}\eta_{tjx}$ where μ_{tjx}^{pop} is assumed known as described above, and η_{tjx} is to be estimated. We use this result to justify that we can use Poisson regression to estimate the unknown η_{tjx}, but we do not need to assume our observations to follow a Poisson distribution.

3.1.2 | Estimation of the model

In practice, the population force of mortality μ_{tjx}^{pop} is unknown. For population mortality estimation and forecasting we apply the model published by the Royal Dutch Actuarial Society which is documented

in Koninklijk Actuarieel Genootschap (2014) (hereafter referred to as the AG model). The AG model is a variant of the Li and Lee (2005) model, which we use to estimate $\widehat{\mu}^{\text{AG}}_{t,x,g}$ on population mortality data from 1970–2011, where the starting year (1970) of the calibration period is the same as in Koninklijk Actuarieel Genootschap (2014) and the end year (2011) is the endpoint of our dataset. Details of calibration of the AG model are available in the online Supplementary Material. When optimizing the likelihood in Equation (2), we use the fitted force of mortality from the AG model $\widehat{\mu}^{\text{AG}}_{t,x,g}$ for the unknown μ^{pop}_{tjx} with $g = \texttt{Gender}_j$. This way, our portfolio-specific mortality model uses an appropriate baseline mortality level.

The factor η_{tjx} is estimated using GAMs, introduced by Hastie and Tibshirani (1986) and popularized by Wood (2017). We include p categorical variables x^d_{ik} such as gender ($k = 1, \ldots, p$)[4], q smooth functions $f_l(\cdot)$ of one-dimensional continuous variables x^c_{il} such as age or salary ($l = 1, \ldots, q$), and a smooth function $g(\cdot, \cdot)$ of the two-dimensional variable $(x^{\text{long}}_i, x^{\text{lat}}_i)$ for the longitude and latitude coordinates of the centre of the postal code. If we use the subscript i to represent a combination (t, j, x), the model is specified as $D_i \sim \text{Poisson}(\lambda_i = \tau_i \mu^{\text{pop}}_i \eta_i)$ for which the additive predictor is given by:

$$\ln(\mathbb{E}(D_i)) - \ln(\tau_i \mu^{\text{pop}}_i) = \ln \eta_i = \beta_0 + \sum_{k=1}^{p} \beta_k x^d_{ik} + \sum_{l=1}^{q} f_l(x^c_{il}) + g(x^{\text{long}}_i, x^{\text{lat}}_i). \tag{3}$$

We define $\ln L(\boldsymbol{\beta})$ as the corresponding log likelihood, where $\boldsymbol{\beta}$ represents the vector of the unknown parameters for the categorical variables and for the smooth functions of the continuous variables. We use thin plate regression splines to estimate the smooth functions f and g. This means that a function $f(x)$ is represented as $\sum_{m=1}^{M} \beta_m b_m(x)$ and a function $g(x, y)$ as $\sum_{n=1}^{N} \beta_n \tilde{b}_n(x, y)$, for fixed M and N and known basis functions $b_m(x)$ and $\tilde{b}_n(x, y)$. Through this representation the model is reduced to a generalized linear model (GLM). To avoid overfitting a wiggliness penalty is added to the log likelihood, resulting in a penalized log likelihood. The wiggliness penalty is the product of the wiggliness of a function f or g and a corresponding smoothing parameter λ:

$$\begin{aligned} \ln L^{\text{Pen.}} = \ln L(\boldsymbol{\beta}) &+ \sum_{l=1}^{q} \lambda_l \int [f_l''(x)]^2 dx \\ &+ \lambda_g \int\int \left[\left(\frac{\partial^2 g}{(\partial x^{\text{long}})^2} \right)^2 + 2 \left(\frac{\partial^2 g}{\partial x^{\text{long}} \partial x^{\text{lat}}} \right)^2 + \left(\frac{\partial^2 g}{(\partial x^{\text{lat}})^2} \right)^2 \right] dx^{\text{long}} dx^{\text{lat}}. \end{aligned} \tag{4}$$

The unknown parameters $\boldsymbol{\beta}$ are estimated by optimizing the penalized log likelihood in (4). The generalized cross-validation (GCV) and the Akaike information criterion (AIC) are often used to select the smoothing parameters λ_l and λ_g. However, these methods are sensitive to misspecification of the correlation structure in the error terms which may result in over- or underfitting of the data, see for example, Krivobokova and Kauermann (2007) and Reiss and Ogden (2009). An alternative approach is to treat the smooth functions with a random effects specification, which means that the λ_l and λ_g can be estimated by marginal likelihood or restricted maximum likelihood (REML), see Wood (2011). We explored both methods and found that REML results in more robust parameter estimates. Therefore, we use REML to select the smoothing parameters. We use the `mgcv` package in `R` for all model estimations, see Wood (2017). The `bam` function in this package is specifically designed for estimation of GAMs on large dataset such as ours, see Wood et al. (2017) for details.

[4]We use the superscript d to emphasize that we use dummy coding of the categorical variables.

Identification problems exist for risk factors that are available for all participants. The smooth function for such a risk factor can be shifted by a constant c and the intercept can be shifted by $-c$ without affecting the model fit. Therefore, risk factors that are available for all participants are modelled with smooth functions centred around zero.

For the categorical variables we quantify the uncertainty in the estimates by constructing confidence intervals based on assumed large sample normality. For the smooth components of the GAM we construct component-wise Bayesian confidence intervals, including the uncertainty in the intercept (Marra & Wood, 2012).

3.1.3 | Treatment of missing variables in GAM estimation

Our estimation of a Poisson GAM is complicated by the presence of missing observations as discussed in Section 2. If a variable has missing records, we use the indicator method (Bennett, 2001) and multiply the smooth effect for that variable with a dummy variable that indicates whether the variable is available or not. For example, a model with `Age` and `Sal` is specified as

$$\ln\left(\mathbb{E}(D_i)\right) = \ln\left(\tilde{E}_i\right) + \beta_0 + f_{\text{Age}}(\texttt{Age}_i) + \text{I}[\texttt{Sal}_i \text{ is available}] \cdot f_{\text{Sal}}(\texttt{Sal}_i),$$

with $\text{I}[\texttt{Sal}_i \text{ is available}] = 1$ if salary is known for observation i and 0 otherwise. Smooth effects for variables with missing values are not applicable to all observations, and therefore these smooth effects cannot be shifted by a constant c with a counteracting shift of $-c$ in the intercept. Identification problems are therefore not present for smooth effects that are multiplied by a numerical value, and these are therefore not centred around zero.

Alternative imputation methods turned out to be infeasible for our dataset. If one can determine a good approximation for the joint distribution of all covariates with missing variables, likelihoods can be calculated by integrating over this joint distribution. In a recent paper, Ungolo et al. (2019) use this method for an actuarial application similar to ours, for a dataset which contains missing data on two risk factors with a small number of outcomes. This approach is not feasible for our dataset due to the presence of multiple continuous risk factors with missing observations.

Other imputation strategies such as predictive mean matching (van Buuren & Groothuis-Oudshoorn, 2011)) or random forests (Stekhoven & Bühlmann, 2012) impute missing values with a predictive model which incorporates random variation. These methods turned out to be infeasible for our dataset using current standard implementations. Alternative multiple imputation strategies such as Bayesian linear regression (van Buuren, 2018) may result in implausible imputed datasets: the structure of the data is not guaranteed to be preserved (e.g. imputed values can become negative where only non-negative values are allowed).

3.2 | Mortality predictions

Combining the estimated effects for the risk factors included in the regression model, we obtain participant-specific factors η_{tjx} which represent the relative difference in mortality between the general population and the participant. These participant-specific factors may change over time due to the presence of covariates or risk factors that change over time. Age is the most obvious example of a risk factor that changes over time, but other risk factors such as postal code or salary may have also changed during the observed period.

When predicting the future mortality rate of a participant, we let the age change over time, but keep all other risk factors fixed. As such, we estimate the participant-specific factors η_{tjx} by combining the estimated effects for the predicted risk factors for future t. Then, participant-specific forecasts of μ_{tjx} are obtained by multiplying a population-wide mortality forecasts μ_{tjx}^{pop} with the predicted participant-specific factors η_{tjx} . An assumption underlying this approach is that the general population and the participants within the portfolio follow the same mortality trend (i.e. there is no basis risk as defined in Haberman et al. , 2015). Our primary aim is to improve best estimate predictions/valuation, and basis risk is therefore not the focus in this article. The limited number of years of data available makes it less suitable for long-term predictions.

3.3 | Model assessment

First, we introduce several statistical measures to determine the added value of the inclusion of a risk factor when explaining observed mortality. However, for a pension fund it is more relevant to accurately predict the value of its future liabilities than to predict the future numbers of deaths. Therefore, we also introduce a novel backtest that targets an accurate prediction of these liabilities. This has, to the best of our knowledge, not been used before when testing models for portfolio-specific mortality.

3.3.1 | In-sample model fit

We investigate how well-different models are able to explain the observations. When models are estimated on the complete dataset, we compute the log likelihood ($\ln L$) on individual observations, that is, the natural logarithm of (2).

The penalized log likelihood in (4) uses penalty parameters λ_l and λ_g to control the wiggliness of the functions f_l and g. If these parameters are zero, the choices for these functions are unrestricted, with M and N degrees of freedom respectively. But as the penalty parameters increase towards infinity, the optimal functions converge to straight lines, which have only two degrees of freedom. This illustrates that when we use an information criterion which penalizes extra parameters, we should not include the number of parameters M and N but what is known as the *effective degrees of freedom*, which takes the wiggliness penalty into account. Our estimates for these are found using the approach of Wood et al. (2016). With these values we can then calculate the cAIC) which is defined as

$$\text{cAIC} = -2\ln \quad L + 2k, \tag{5}$$

with n the number of observations included in the likelihood function (i.e. $n = 22{,}632{,}277$) and k the effective degrees of freedom. In a similar way we define the Bayesian information criterion (BIC) as

$$\text{BIC} = -2\ln \quad L + k\ln n. \tag{6}$$

3.3.2 | Cross-validation statistics and robustness analysis

Czado et al. (2009) discuss proper scoring rules to evaluate out-of-sample performance of models for count data. We focus on the log score which can be interpreted as an out-of-sample log likelihood statistic. Denote by F_{-t} all observations in the dataset excluding observations from year t. For each $t \in \mathcal{T} = \{2006, \ldots, 2011\}$ we estimate the model using F_{-t} which yields an estimate $\widehat{\eta}_{j,x}^{-t}$ for participant j in year t at age x. Pension funds and insurance companies typically update their assumptions once a

year when new data become available. Therefore, we have chosen to split the dataset into a training dataset and a test dataset based on the calendar year, since this corresponds to their annual update process.

Similar to (1a) and (1b), we define L_{tj} as the likelihood of observed death or survival for participant $j = 1, \ldots, L_t$ in calendar year t given the predictive distribution that follows from F_{-t}. If τ_{tjx} represents the fraction of year t that participant j was alive at age x (assuming he was alive at the start of that year), L_{tj} is given by:

$$\ell_{tj} = \exp\left[-\tau_{tjx}\mu_{tjx}^{\text{pop}}\widehat{\eta}_{j,x}^{-t}\right]\left(\tau_{tjx}\mu_{tjx}^{\text{pop}}\widehat{\eta}_{j,x}^{-t}\right)^{\delta_{tjx}}, \tag{7}$$

and the log score for year t is then defined by:

$$\ln \mathrm{S}_t = -\frac{1}{L_t}\sum_{i=1}^{L_t}\ln \ell_{tj}, \tag{8}$$

with $\sum_{t=2006}^{2011} L_t = 11,325,511$. Finally, we aggregate this over the different years into the time-averaged value $\ln \mathrm{S} = \frac{1}{6}\sum_{t=2006}^{2011} \ln \mathrm{S}_t$.

We also investigate the robustness of estimated effects. For risk factor l, define $\widehat{f}_l^{-t}(x_l)$ as the effect estimated using F_{-t}. We compare the estimated effects $\widehat{f}_l^{-t}(x_l)$ with the 80% confidence interval for $\widehat{f}_l(x_l)$ estimated on the complete dataset. If the estimated effect is robust (i.e. consistent through time), the estimates $\widehat{f}_l^{-t}(x_l)$ are close to (and show a similar pattern as) $\widehat{f}_l(x_l)$.

3.3.3 | Predicted life expectancies

We further investigate the impact of including risk factors in the mortality model on remaining life expectancies of participants in the fund. We compute remaining life expectancies for different ages and for males and females separately. We consider a risk profile as a combination of all risk factors except age and gender, represented by profile r instead of i.[5] We define the remaining cohort life expectancy for risk profile r with age x at the beginning of calendar year t and for gender g as

$$LE_g(r,t,x) \approx \frac{1}{2} + \sum_{y=x+1}^{\infty} S_g(r,t,x,y) \tag{9}$$

with cumulative survival probabilities

$$S_g(r,t,x,y) = \exp\left[-\sum_{s=0}^{y-x+1}\widehat{\mu}_{t+s,x(r)+s}^{\text{AG},g(r)}\cdot\widehat{\eta}_r\right]. \tag{10}$$

3.3.4 | Financial backtest

For this backtest, we assume that the management of the pension fund at the beginning of the year 2011 wants to predict the value of the liabilities at the end of the year. For simplicity we ignore any

[5]The difference between risk profile i and risk profile k is that the risk factors `Age` and `Gender` are fixed.

cash flows during the year 2011, but this will not materially affect the results since the liabilities are mostly determined by cash flows further in the future. We use data from the years 2006–2010 as the training sample, and data from 2011 as the test sample. On 1 January 2011, we use all information available up to that point to approximate for each participant separately the value of the liabilities on 31 December 2011. Using the risk profile of individuals and the estimated effects from the training sample, we predict the probability that a participant survives the year 2011. The observations in the test sample tell us which participants died in the year 2011 and who survived until 31 December 2011. Using the test set we know the liabilities at the end of the year, and we can compare how well-different model specifications are able to predict the liabilities required for the surviving participants in the test sample. This helps to determine the added value of a risk factor when predicting the value of liabilities. Below we outline the highlights or our proposed financial backtest, more details are available in the online Supplementary Material.

For participant j we denote the yearly benefit by b_j which is paid in year t if $x(j,\ t) \geq x_r$, with $x(j,\ t)$ the age of participant j at the beginning of year t and $x_r = 65$ the retirement age in the Netherlands in 2011. We define an annuity a_j that represents the present value of a cash flow of 1 if participant j is still alive at that time t and $x(j,\ t) \geq x_r$, such that the present value of the liabilities for participant j is given by $b_j a_j$.

Define $\widehat{\eta}_{j,x}^{-2011}$ as the fitted factor for participant j aged x during the year 2011, which is estimated using the training sample. We model the uncertainty of participant j surviving the year 2011 using a Bernoulli distributed random variable $Y_{2011,j}$, with $P(Y_{2011,j} = 1) = p_{2011,j}$ which are independent for different j's. Under the assumption of a constant force of mortality μ_{tx} on the interval $[t,\ t+1) \times [x,\ x+1)$, the 1-year survival probability for participant j is given by $p_{2011,j} = \exp\left[-\widehat{\mu}_{2011,x}^{\mathrm{AG},g} \cdot \widehat{\eta}_{j,x}^{-2011}\right]$. The stochastic value of the liabilities Γ on 31 December 2011 is then given by

$$\Gamma = \sum_{j=1}^{J_{2011}} \left(Y_{2011,j} \cdot b_j a_j + (1 - Y_{2011,j}) \cdot 0\right), \tag{11}$$

with J_{2011} the number of members alive at the beginning of the year 2011. The fund faces a liability with present value $b_j a_j$ at the end of 2011 for those participants who survive, whereas this liability is released for participants who die during 2011. J_{2011} is larger than two million, and based on a simulation study we have verified that Γ behaves like a normally distributed random variable. However, the skewness of Γ is close to but not equal to zero. We use a parametric approximation to construct prediction intervals for Γ, and we allow for the non-zero skewness by using the skew-normal distribution for this approximation, see for example Vernic (2006).

We define the indicator variable $\mathrm{I}_{2011,j}$ that is 1 if participant j was still alive at 31 December 2011 and zero otherwise. The actual liabilities per 31 December 2011 are then calculated as:

$$\tilde{\Gamma} = \sum_{j=1}^{J_{2011}} \mathrm{I}_{2011,j} \cdot b_j a_j. \tag{12}$$

If a model is able to accurately predict the evolution in the liabilities over a 1-year horizon, the actual liabilities $\tilde{\Gamma}$ will often lie within the prediction interval for Γ.

Within a predictive distribution for the liabilities, underestimated values and overestimated values may cancel out. Therefore, we also calculate the MSPE as:

$$\begin{aligned}\text{MSPE} &= \sum_{j=1}^{J_{2011}} (\mathrm{I}_{2011,j} \cdot b_j a_j - p_{2011,j} \cdot b_j a_j)^2 \Big/ \sum_{j=1}^{J_{2011}} (b_j a_j)^2 \\ &= \sum_{j=1}^{J_{2011}} \underbrace{(b_j a_j)^2}_{\text{'weights'}} \underbrace{(\mathrm{I}_{2011,j} - p_{2011,j})^2}_{\text{'errors'}} \Big/ \underbrace{\sum_{j=1}^{J_{2011}} (b_j a_j)^2}_{\text{normalizingconstant}} .\end{aligned} \tag{13}$$

Through this definition of the MSPE, the prediction error in a participant's survival is weighted by the value of the liabilities needed for that participant. This way we ensure that participants who contribute more to the liabilities of a pension fund are given more weight in this statistic.

4 | RESULTS

We estimate portfolio factors using the procedure described in Section 3.1, and we define two reference specifications:

1. Portfolio mortality equals population mortality ($\eta_i = 1$ for all i);
2. The relative difference with population mortality is the same for all participants, that is, the regression model in Equation (3) only includes the constant β_0 ($\eta_i = \exp(\beta_0)$ for all i).

These two reference models allow us to quantify the relative importance of including different risk factors when explaining historical portfolio mortality. Besides these reference models, we consider all single variable specifications and a selection of multiple variable specifications.

The variables that we include as explanatory variables are `Gender`, `Age`, `DisTime`, `DisPerc`, `Sal`, `IA`, `AFPP`, `Edu` and (`Long`, `Lat`). The variable `Year` is not included, because the dataset covers only a few years. We consider only main effects, but the analysis can easily be extended to include interactions between risk factors such as gender-specific salary effects. Interactions between continuous variables are also possible, but this greatly increases the number of possible specifications and increases the computational cost.

4.1 | Estimation results

Table 2 shows the model fit (the log likelihood, effective degrees of freedom, conditional AIC and BIC) and cross-validation statistics (the time-averaged log score) for a selection of single and multiple variable specifications.

4.1.1 | Variable selection, model fit and estimated factors

The two reference specifications are shown in the top 2 rows of Table 2, and the model fit improves considerably if we allow for differences between mortality in the population and the portfolio. The cAIC and BIC almost always improve if we add a single variable on top of the constant, but the BIC does not improve if we add `Gender` or `PC`. It is no surprise that `Gender` has little additional explanatory power, since gender is already included in the baseline mortality level $\mu_{t,x}^{\text{pop},g}$. The smooth effect

TABLE 2 Estimation results for a selection of models. For ln$L$ larger is better, for the other statistics smaller is better. The horizontal lines separate models with the same number of risk factors. Note that all regression models (row 2 and below) include a constant

	Model	ln L	EDF	cAIC	BIC	ln S ($\times 10^{-3}$)
1.	No model	−222,719	0.0	445,437	445,437	19.66964
2.	Constant	−222,036	1.0	444,073	444,088	19.61136
3.	Gender	−222,029	2.0	444,062	444,092	19.61081
4.	Age	−221,940	2.5	443,885	443,922	19.60350
5.	DisTime	−220,430	6.6	440,873	440,973	19.47026
6.	DisPerc	−220,442	8.8	440,902	441,034	19.47168
7.	Sal	−221,069	5.2	442,149	442,226	19.52643
8.	IA	−221,733	4.3	443,475	443,539	19.58459
9.	AFPP	−221,916	6.0	443,844	443,934	19.60153
10.	Edu	−221,962	6.7	443,938	444,039	19.60533
11.	PC	−221,946	35.9	443,964	444,500	19.60729
12.	DisPerc-DisTime	−220,414	13.9	440,855	441,062	19.46984
13.	DisPerc-Gender	−220,417	9.9	440,854	441,001	19.46952
14.	DisPerc-Age	−220,383	13.4	440,793	440,994	19.46750
15.	DisPerc-Sal	−219,861	11.9	439,746	439,923	19.42094
16.	DisPerc-IA	−220,393	12.1	440,811	440,991	19.46754
17.	DisPerc-AFPP	−220,357	10.6	440,736	440,894	19.46439
18.	DisPerc-Edu	−220,385	14.1	440,797	441,007	19.46701
19.	DisPerc-PC	−220,378	32.9	440,822	441,314	19.46869
20.	DisPerc-Sal-Gender	−219,843	12.9	439,712	439,904	19.41943
21.	DisPerc-Sal-Age	−219,807	16.5	439,647	439,894	19.41727
22.	DisPerc-Sal-IA	−219,793	14.5	439,614	439,830	19.41511
23.	DisPerc-Sal-AFPP	−219,784	14.0	439,596	439,805	19.41439
24.	DisPerc-Sal-Edu	−219,823	16.9	439,679	439,931	19.41790
25.	DisPerc-Sal-PC	−219,808	33.0	439,682	440,174	19.41854
26.	DisPerc-Sal-IA-AFPP	−219,717	16.6	439,467	439,714	19.40868
27.	DisPerc-Sal-IA-Edu	−219,755	19.5	439,550	439,840	19.41217
28.	DisPerc-Sal-IA-PC	−219,742	34.6	439,553	440,070	19.41280
29.	DisPerc-Sal-IA-AFPP-Edu-PC	−219,641	36.8	439,356	439,905	19.40397

for `PC` requires a large number of effective degrees of freedom, and the trade-off between improvement in fit and complexity does not result in improved BIC statistics. The results for cAIC and BIC are similar, but as is to be expected, there are some cases in which the information criteria suggest different model specifications.

Large improvements in model fit come from adding `DisTime`, `DisPerc` or `Sal`. With `Salary` and `DisPerc` included in the model (e.g. row 15 in Table 2), the information criteria and log score statistic improve considerably compared to the first two rows. If we keep adding variables, the

statistics improve further, but the improvements are much smaller than those from adding Salary and DisPerc.

The model with the risk factors DisPerc, Sal, IA, AFPP, Edu and PC performs well for the model assessment criteria introduced in Section 3.3 (see row 29 in Table 2 for those results). Figure 2 shows the estimated effects for this model. The grey area in the graphs represents the 80% confidence interval based on the calibration using the complete dataset. The coloured lines represent the estimated effect if a single year is left out of the dataset. Models with Salary and DisTime perform similarly to models with Salary and DisPerc, but we prefer to use DisPerc over DisTime since DisPerc is more uniformly spread, see Figure 1. Moreover, exploratory marginal model fits revealed a widening confidence interval for larger values of DisTime, while for all values of DisPerc the confidence interval width did not change.

In Figure 2 we observe that the effects of Sal and IA are in line with intuition: higher salary leads to lower mortality, and more hours worked at irregular times leads to higher mortality, and the estimated effects for these risk factors are strong. The maximal relative difference in the force of mortality for participants with high and low salary is about $\exp[0.4 - (-0.4)] \approx 2.2$, and the effect of IA is only slightly lower.

The estimated effect of AFPP in Figure 2 indicates that early retirement increases the level of mortality compared to retiring at the regular retirement age. Participants may have different motivations for retiring early. Wealthy participants may have chosen to retire early because they no longer *need* to work, whereas participants with bad health may have retired early because they *can* no longer work. For the first group we expect lower mortality and for the second group we expect higher mortality. In a marginal model (not shown), AFPP attempts to capture both combined effects, while in a model that also includes the risk factors Sal and DisPerc, AFPP only captures the remaining effect. The most common values are $\texttt{AFPP} = 60$ and $\texttt{AFPP} = 65$, and retiring at the age of 60 (early retirement) increases the force of mortality by about 5% compared to retiring at 65 (the official retirement age). Note, however, that retired participants experience lower mortality than participants who are not yet retired, ceteris paribus.

In preliminary analysis (not shown) we investigated the use of $\texttt{Dis} \in \{\texttt{DisPerc}, \texttt{DisTime}\}$ in the regression models. Participants with a disability spell equal to zero ($\texttt{Dis} = 0$) are included in the reference group, for participants with missing disability information (Dis is missing) we include a dummy variable, and for participants with a non-zero disability spell ($\texttt{Dis} > 0$) a smooth effect is included.[6] Consequently, in a model that only includes the variable Dis, the factor η_i is thus represented as:

$$\ln\eta_i = \beta_0 + \mathrm{I}[\texttt{Dis is missing}] \cdot \beta_1 + \mathrm{I}[\texttt{Dis} > 0] \cdot f_{\mathrm{Dis}}(\texttt{Dis}_i).$$

In model specifications that include other variables, the variable Dis is included in the factor η_i accordingly.

The estimated effect for DisPerc in the final model shown in Figure 2 is not monotone. In 2005 the Dutch government introduced new legislation regarding income protection provided by the state. Under this new legislation people can be classified as being partially disabled, and for this group a suitable replacement job is searched. Three main classes are distinguished: less than 35% disabled, between 35% and 80% disabled, and above 80% disabled, and the state pension depends on the disability class. DisPerc is a variable aggregated over time and can thus not be directly linked to the disability

[6] If a participant has a cumulative disability spell equal to zero, then both DisTime and DisPerc are zero. The same principle holds for participants with missing disability information or with a positive disability spell.

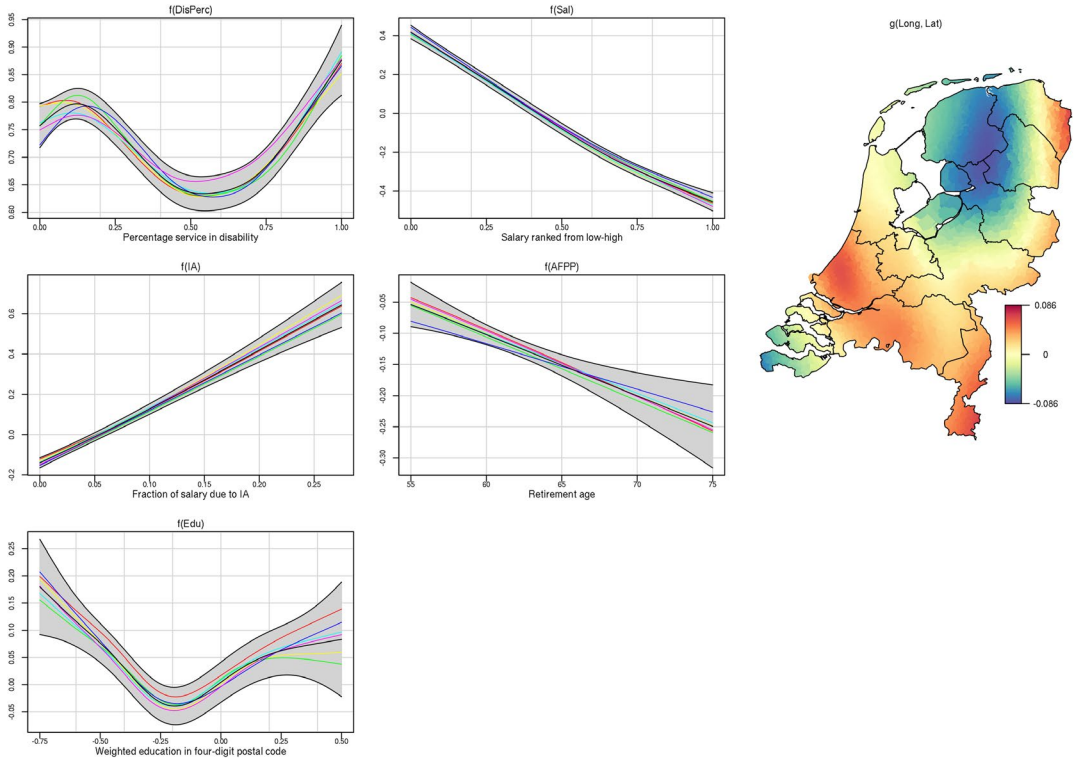

FIGURE 2 Estimated effects for the model with `DisPerc`, `Sal`, `IA`, `AFPP`, `Edu` and `PC` included (row 29 in Table 2). The first two columns show the estimated smooth effects for `DisPerc`, `Sal`, `IA`, `AFPP` and `Edu`; the grey area bounded by the black lines represents the 80% confidence interval for the effects estimated on the complete dataset, and the coloured lines represent the estimated effects if a single observation year is excluded. The third column shows the estimated spatial effect. [Colour figure can be viewed at wileyonlinelibrary.com]

classes distinguished in the legislation. The different treatment of these three disability classes probably causes the curvature in the estimated effect for `DisPerc`.

The third column in Figure 2 shows the estimated spatial effect based on longitude and latitude information. The map clearly highlights regions with higher mortality and regions with lower mortality.

The results shown in Table 2 are obtained by optimizing the likelihood as specified in Equation (2) using Poisson regression. We challenged the assumption of equidispersion underlying the Poisson regression by investigating the presence of overdispersion in our data. However, we did not find any evidence for this, see the online Supplementary Material.

4.1.2 | Cross-validation

The last column in Table 2 shows the cross-validation statistics (smaller ln S is better). The scoring rule does not include a penalty for adding parameters. Therefore, if a variable captures an effect that is consistent through time, including that variable will lead to a more accurate predictive distribution and thus to improved results for the scoring rule, while including variables that represent uninformative noise or which are not consistent over time may deteriorate those results. Results for the scoring rule improve if we add variables, so the variables capture effects that are consistent through time.

TABLE 3 Predicted remaining cohort life expectancies for different risk profiles. The model used for estimating factors includes the variables DisPerc, Sal, IA, AFPP, Edu and PC_Group, and remaining life expectancies are computed in the year 2012 for males (*M*) and females (*F*) at the age of 25 and 65. To limit the dimensions of the risk profiles, we have assumed AFPP and Edu missing in all risk profiles, and we have chosen a postal code where the estimated effect is nil

IA	DisPerc	Sal	$LE_M(25)$	$LE_F(25)$	$LE_M(65)$	$LE_F(65)$
No	No	0.90	68.4	70.5	26.7	29.2
		0.50	65.9	68.1	24.1	26.8
		0.10	62.6	65.1	20.9	23.9
	5%	0.90	62.1	64.6	20.4	23.4
		0.50	59.4	62.1	18.0	21.1
		0.10	55.8	58.7	15.2	18.3
10%	No	0.90	67.4	69.5	25.7	28.2
		0.50	64.9	67.2	23.1	25.9
		0.10	61.6	64.1	19.9	23.0
	5%	0.90	61.0	63.6	19.4	22.5
		0.50	58.3	61.0	17.1	20.2
		0.10	54.6	57.5	14.3	17.5
Baseline (general population)			62.5	65.1	19.6	22.7

Furthermore, from Figure 2 we observe that the coloured lines show a similar pattern as the grey areas, which means that the estimated effects for the risk factors are stable over time.

4.2 | Predicted cohort life expectancies

The model with the risk factors DisPerc, Sal, IA, AFPP, Edu and PC performs well both in-sample and out-of-sample, and the estimated effects for this model are robust. Using the estimated effects from this model (see Figure 2), we compute remaining cohort life expectancies (LE) as in Equation (9) for a set of risk profiles which are shown in Table 3. This allows us to quantify the impact of different variables on LE, and to compare our results with existing literature. We do not distinguish between different levels of PC, AFPP and Edu because doing so would complicate a clear presentation of the results. Therefore, for the variables AFPP and Edu we assume these are missing, and for PC we assume a postal code that has a limited estimated effect, when calculating the remaining life expectancies. The most-favourable group in this table is defined by (IA = No, DisPerc = "No", Sal = 0.90) and the least-favourable group by (IA = 10%, DisPerc = 5%, Sal = 0.10).

The results are striking. For the remaining life expectancy at age 25, the difference between low and high salary risk profiles is between 5 and 6 years, and the difference is of similar size for being/having been disabled or not. Chetty et al. (2016) found a difference of 14.6 years between the 1% richest and 1% poorest on a dataset that is not restricted to people in a particular pension fund. Instead, their data covers all individuals with positive household earnings in the USA. The USA is a more heterogeneous group than the participants within the Dutch pension fund, and we use different quantiles for salary, so the different scale in the longevity gaps is not surprising. Furthermore, even the effect of postal code on life expectancy is not negligible since it amounts to approximately 1 year, and this result is in line with figure 5 from Chetty et al. (2016).

Cairns et al. (2019) compute the partial period life expectancy from age 55 to 90 for different affluence groups. These affluence groups are defined using a combination of reported wealth and income. In their work the difference in partial period life expectancy between the lowest and highest affluence groups is about 6.5 years, which is also in line with our results. Furthermore, RIVM (2014) reports differences up to 6 years in period life expectancy at birth between low and highly educated people. This range is similar to what we observe for low and high salaries.

Our approach allows us not only to study main effects, but also favourable and unfavourable combinations. The difference in remaining life expectancy for $x = 25$ between the most-favourable group in Table 3 and the least-favourable group is 13.8 and 13.0 years for males and females respectively. These differences persist over time: at age 65 the differences are 12.4 and 11.7 years for males and females respectively. This means that if participants were to retire at age 65, males in the most-favourable group benefit 26.7 years from their pension which is almost twice as long as males in the least-favourable group, who benefit 14.3 years on average. We have not taken into account the effect from all risk factors that can be included in the model, so for more specific risk profiles the differences may be even larger.

In these calculations we have assumed that the variables `Sal`, `DisPerc` and `IA` remain constant throughout the entire lifetime of a participant. Although this is a strong assumption, it is the best assumption we can make based on the available data.

4.3 | Financial backtest

Figure 3 shows the results of the financial backtest. For each model we computed the mean, variance and skewness of the value of the liabilities (predicted at 1 January 2011). The mean is represented by the large dot, the horizontal lines represent the 90% prediction intervals under the assumption of skewed normality for the value of the liabilities. The vertical dashed line indicates the actual value observed at the end of the year (i.e. the target for our predictions at the beginning of the year). The MSPE as defined in Equation (13) is shown on the right-hand side of the figure.

If we use participant factors equal to one (i.e. mortality for each participant is the same as population mortality), we underestimate the liabilities; the actual liabilities are far outside the prediction interval. Predictions improve if we take participant mortality into account, so this clearly shows the usefulness of adjustments to population mortality rates. However, the liabilities are still underestimated. It is surprising that for models with only a single variable `Age`, `Gender`, `DisTime`, `DisPerc`, `AFPP`, `PC` or `IA` included, the predicted liabilities are very similar to the predicted liabilities when only an intercept, and thus a fixed correction to population mortality, is included. Obviously, the liabilities predicted at the level of individual participants differ when moving from a model with only an intercept to a model incorporating one of these risk factors.

However, if we include `Sal` as an explanatory variable the prediction of the liabilities improves significantly. Since accrued benefits are correlated with salary, this was to be expected. This is also the reason why practitioners in insurance companies tend to work with mortality rates weighted by insured amounts, as explained in Plat (2009).

The model estimation results in Table 2 suggest to use a model which includes `Sal` and either `DisTime` or `DisPerc`, and possibly `IA`, `Edu` and `PC` as additional explanatory variables. However, in Figure 3 we see that including information on disability when `Sal` is already taken into account decreases the predicted value of liabilities and the backtesting results worsen (compared to the model where only `Sal` is included). This is further investigated in Figure 4 where we show the predicted value of liabilities for specifications with `DisPerc` and/or `Sal` included, but we distinguish between participants with known and unknown disability and/or salary information:

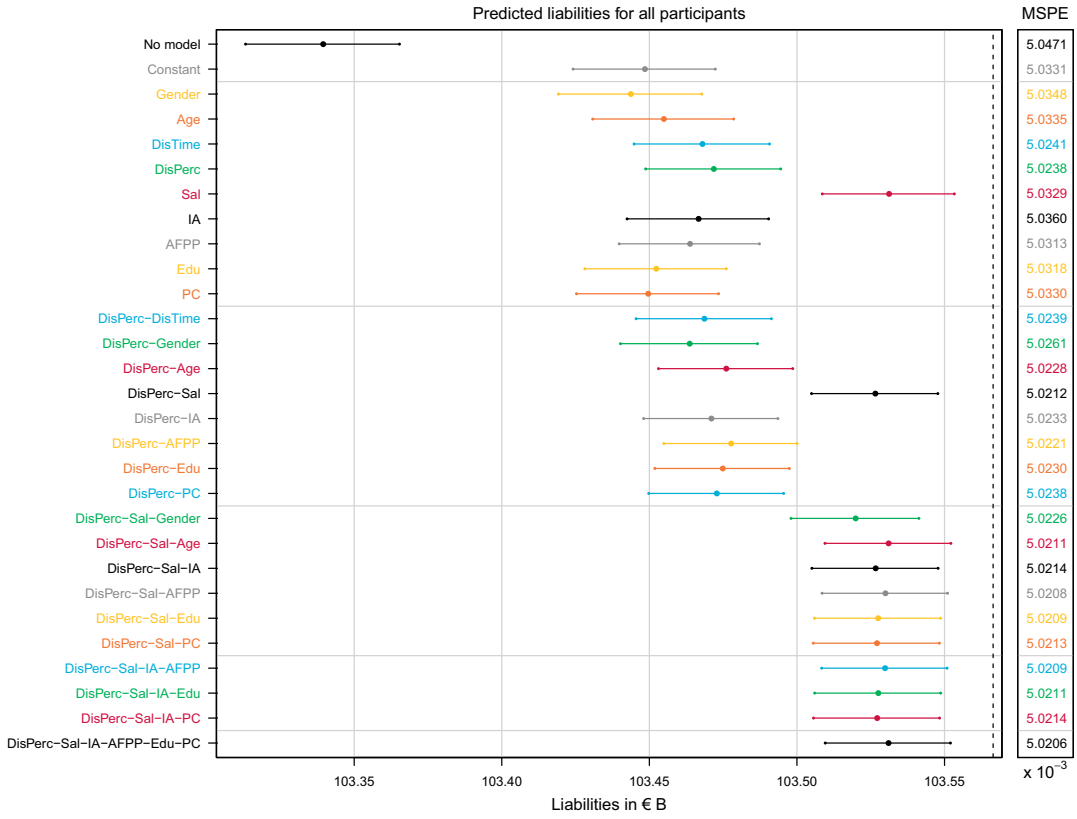

FIGURE 3 Results for the financial backtest. For each model we show the expected value and 90% confidence interval for the predicted liabilities needed on 31 December 2011 (predicted at January 1st 2011). The vertical dotted line represents the liabilities for the participants who were still actually alive on 31 December 2011. The mean squared prediction error is shown for each model on the right-hand side of the figure. [Colour figure can be viewed at wileyonlinelibrary.com]

1. If only `DisPerc` is included, liabilities are underestimated for participants without a positive disability spell, because the correlation between mortality and accrued rights is not taken into account. The actual liabilities fall within the prediction interval for participants with a positive disability spell;
2. If only `Sal` is included, predicted liabilities increase for participants with known salary information compared to the predicted liabilities for the other single factor regression models. For participants without positive disability spell the liabilities are underestimated, and for participants with positive disability spell the liabilities are overestimated. The actual liabilities fall outside the prediction interval for all groups;
3. If both `DisPerc` and `Sal` are included, the predicted liabilities are closer to the actual liabilities for nearly all groups compared to the liabilities predicted using the single factor regression models. For people with known salary the prediction is very close to the actual liabilities.

This example shows that focusing on a single backtest at an aggregate level may lead to suboptimal decisions regarding the selection of risk factors. For one risk profile we may underestimate the liabilities while for another risk profile the liabilities may be overestimated. Therefore, we also calculated the MSPE and include the results on the right-hand side of Figure 3. We calculated the

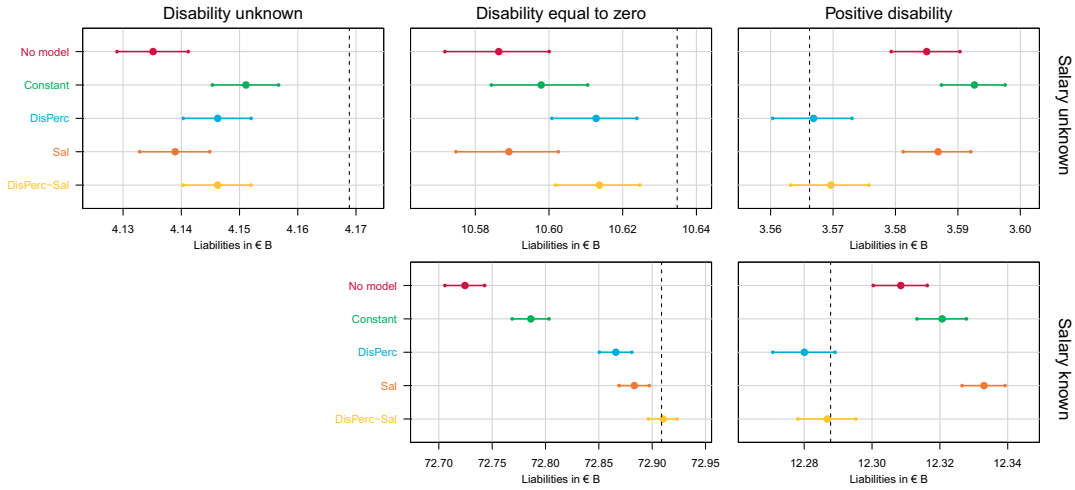

FIGURE 4 The financial backtest split between participants with and without disability and/or salary information. There are no participants with known salary information but unknown disability information. Notes: see Figure 3. [Colour figure can be viewed at wileyonlinelibrary.com]

MSPE using liabilities predicted at individual level, so within the MSPE under- and overestimation cannot cancel each other out. The MSPE for the model with only `Sal` included is worse than the one obtained when only `DisPerc` is included. Furthermore, if we include both `DisPerc` and `Sal` the MSPE improves substantially. When both `Sal` and `DisPerc` have been included, the MSPE improves only marginally when other risk factors are added. Based on Figures 3 and 4 we conclude that the MSPE yields useful information regarding which models provide accurate liability predictions on an individual level. However, the prediction intervals provide additional insights for subgroups of the portfolio and reveal for which groups the liabilities are under- or overestimated. Therefore, we recommend to use both tools to obtain a complete view on how different models perform relative to each other.

5 | CONCLUSION

This paper shows how to estimate mortality rates for individual risk profiles using a large dataset of individual records registered over time. The scale of our empirical analysis strengthens our conclusion that salary, disability and working at irregular hours all have a particularly strong impact on the level of mortality. We quantified significant longevity gaps between the most-favourable and least-favourable risk profiles in the fund: a difference between remaining life expectancies of about 13 years at age 25 and 12 years at age 65.

To identify relevant risk factors we did not only rely on statistical measures (evaluated in- and out-of-sample), but we also designed a financial backtest that targets the (monetary) accuracy of the predictions. For our data, information on salary and time spent in disability of participants turns out to be particularly relevant for accurate prediction of the pension liabilities.

This paper illustrates the importance of intensive and careful data collection and storage. The sophistication of modern statistical methods allows us to improve mortality estimates by including explanatory variables beyond the traditional age-gender setting. As such, we are able to quantify the

longevity gap between the national population and a subgroup on the one hand, and among participants within that group on the other hand. Pension funds can incorporate these insights when valuing their liabilities. Moreover, having access to similar data collected at the level of the national population, the strategy outlined in this paper will support policymakers involved in the design of pension reforms, for example by quantifying the longevity gaps between different categories of professions.

ACKNOWLEDGEMENTS

The authors gratefully thank the Dutch pension fund for making the dataset on the micro-level lifetime data available. The authors also gratefully thank two anonymous referees and the editor for valuable comments on an earlier version. Support by Netspar, the Network for Studies on Pensions, Ageing and Retirement, is gratefully acknowledged.

REFERENCES

Antonio, K, Devriendt, S, de Boer, W., de Vries, R., De Waegenaere, A., Kan, H.-K. et al. (2017) Producing the Dutch and Belgian mortality projections: A stochastic multi-population standard. *European Actuarial Journal*, 7(2), 297–336.

Bennett, D.A. (2001) How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25(5), 464–469.

Brouhns, N., Denuit, M. & Vermunt, J. K. (2002) A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, 31(3), 373–393.

Brown, R.L. & McDaid, J. (2003) Factors affecting retirement mortality. *North American Actuarial Journal*, 7(2), 24–43.

Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A. et al. (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., Kallestrup-Lamb, M., Rosenskjold, C., Blake, D. & Dowd, K. (2019) Modelling socioeconomic differences in the mortality of Danish males using a new affluence index. *ASTIN Bulletin*, 49, 555–590. doi: 10.1017/asb.2019.14.

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N. et al. (2016) The association between income and life expectancy in the United States, 2001-2014. *Journal of the American Medical Association*, 315(16), 1750–1766.

Costa, G. (1996) The impact of shift and night work on health. *Applied Ergonomics*, 27(1), 9–16.

Cox, D. R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 32(2), 187–220.

Czado, C. & Rudolph, F. (2002) Application of survival analysis methods to long-term care insurance. *Insurance: Mathematics and Economics*, 31, 395–413.

Czado, C., Gneiting, T. & Held, L. (2009) Predictive model assessment for count data. *Biometrics*, 65, 1254–1261.

Denuit, M. & Legrand, C. (2018) Risk classication in life and health insurance: Extension to continuous covariates. *European Actuarial Journal*, 8, 245–255.

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

Elo, I.T. & Preston, S.H. (1996) Educational differentials in mortality: United States, 1979–85. *Social Science & Medicine*, 42(1), 47–57.

Gschlössl, S., Schoenmaekers, P. & Denuit, M. (2011) Risk classification in life insurance: Methodology and case study. *European Actuarial Journal*, 1, 23–41.

Haberman, S., Kaishev, V., Millossovich, P., Villegas, A., Baxter, S., Gaches, A. et al. (2015) A methodology for assessing basis risk—Abstract of the London discussion. *British Actuarial Journal*, 20(3), 461–490.

Hastie, T. & Tibshirani, R. (1986) Generalized additive models. *Statistical Science*, 1, 297–310.

Holford, T.R. (1980) The analysis of rates and of survivorship using log-linear models. *Biometrics*, 36(2), 299–305.

Koninklijk Actuarieel Genootschap. (2014) Projection table AG 2014. Available at: http://www.ag-ai.nl/view.php?action=view&Pagina_Id=625.

Krivobokova, T. & Kauermann, G. (2007) A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, 102(480), 1328–1337.

Laird, N. & Olivier, D. (1981) Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76(374), 231–240.

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

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

Marra, G. & Wood, S. (2012) Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1), 53–74.

Pitacco, E., Denuit, M., Haberman, S. & Olivieri, A. (2009) *Modelling longevity dynamics for pensions and annuity business*. New York: Oxford University Press.

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

Reiss, P.T. & Ogden, R.T. (2009) Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society, Series B*, 71(2), 505–523.

Richards, S.J., Kaufhold, K. & Rosenbusch, S. (2013) Creating portfolio-specific mortality tables: A case study. *European Actuarial Journal*, 3, 295–319.

RIVM. (2014) Socio-economic differences in remaining life expectancy. Available at: http://www.eengezondernederland.nl/Heden_en_verleden/Levensverwachting (Accessed 30th October 2017).

Stekhoven D.J. & Bühlmann, P. (2012) MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118.

Ungolo, F., Christiansen, M.C., Kleinow, T. & MacDonald, A.S. (2019) Survival analysis of pension scheme mortality when data are missing. *Scandinavian Actuarial Journal*, 6, 523–547.

van Buuren, S. (2018) Flexible imputation of missing data. London: Chapman and Hall/CRC.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011) mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67.

Vernic, R. (2006) Multivariate skew-normal distributions with applications in insurance. *Insurance: Mathematics and Economics*, 38, 413–426.

Villegas, A. M. & 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.

Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B*, 73(1), 3–36.

Wood, S. N. (2017) *Generalized additive models: An introduction with R*, 2 edition. London: Chapman and Hall/CRC.

Wood, S.N., Pya, N. & Säfken, B. (2016) Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516), 1548–1563.

Wood, S.N., Li, Z., Shaddick, G. & Augustin, N.H. (2017) Generalized additive models for gigadata: Modeling the U.K. black smoke network daily data. *Journal of the American Statistical Association*, 112(519), 1199–1210.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.
Data S1

How to cite this article: van Berkum F, Antonio K, Vellekoop M. Quantifying longevity gaps using micro-level lifetime data. *J R Stat Soc Series A*. 2021;184:548–570. https://doi.org/10.1111/rssa.12631