Annals of Operations Research
https://doi.org/10.1007/s10479-025-06779-2

ORIGINAL RESEARCH

A holistic approach to interpretable modelling and forecasting of human mortality by gender and country

Ana Debón[1] · Steven Haberman[2] · Gabriella Piscopo[3]

Received: 29 July 2024 / Accepted: 18 July 2025

Abstract
Studies from many countries find that gender differences in mortality rates and life expectancy vary by country. The multipopulation Lee-Carter family of models, a widely-used methodology, decompose mortality rates into age, time, and country components, offering valuable insights into mortality trends. We delve into the interpretability of the Lee-Carter multipopulation model, elucidating its ability to capture underlying mortality patterns and project future trajectories. Moreover, we extend our analysis by incorporating machine learning techniques to model the residuals of the Lee-Carter framework. The main contribution of the paper is to introduce these techniques in the context of the multiple population mortality models. Specifically, we employ Random Forest to refine joint mortality forecasts by country, effectively capturing complex nonlinear relationships in residuals and improving predictive performance. In this paper, we revisit these models using new statistical techniques and data sets from the Human Mortality Database. By leveraging advanced computational algorithms, we aim to enhance the accuracy of mortality rate predictions and account for residual patterns that may not be captured by the traditional Lee-Carter approach alone. Through empirical validation and comparative analyses, we demonstrate the efficacy of integrating machine learning into multiple population mortality forecasting, thereby contributing to the refinement and improvement of mortality modeling methodologies.

Keywords Multipopulation Lee-Carter models · Interpretability · Forecasting · Machine learning

1 Introduction

Since Lee and Carter proposed a stochastic approach to the modelling of dynamic life tables in 1992 and used this model to study projections of mortality rates in US (Lee & Carter, 1992), there have been several papers forecasting population mortality in other developed

Ana Debón and Steven Haberman contributed equally to this work.

Extended author information available on the last page of the article

Published online: 02 September 2025

countries such as Canada (Lee & Nault, 1993), Chile (Lee & Rofman, 1994), Japan (Wilmoth, 1996), Belgium (Brouhns et al., 2002), Austria (Carter & Prkawetz, 2001), England and Wales (Renshaw & Haberman, 2003), Australia (Booth & Tickle, 2003) and Spain (Guillen & Vidiella-i-Anguera, 2005; Debón et al., 2008). All these papers deal with separate models for each country, and fitting the Lee-Carter model to mortality data for each sex independently. Another perspective focuses on the mortality dynamics of two or more populations of similar size, see, for example, Li and Lee (2005), Delwarde et al. (2006), and Antonio et al. (2017).

Recent work that has considered a group of related countries or subnational mortality modelling includes Li and Lee (2005), Russolillo et al. (2011), Debón et al. (2011), Villegas and Haberman (2014), Danesi et al. (2015), Alexander et al. (2017), Bergeron-Boucher et al. (2018), Cairns et al. (2019), Wen et al. (2021) and Bégin et al. (2023). Among them, Li and Lee (2005), Russolillo et al. (2011) and Debón et al. (2011) have proposed models derived from the Lee-Carter model to obtain coherent mortality forecasts for a group of populations that could be applied to modelling mortality rates for the two-sexes mortality and for a group of countries. These models also facilitate the comparison of mortality from different countries due to the interpretability of their parameters. Models with explicitly interpreted parameters offer advantages in clarity, communication, hypothesis formulation, and practical applicability, making them valuable tools across various research and application fields.

In particular, the proposals of Russolillo et al. (2011) and Debón et al. (2011) have the advantage that their computational cost is very low, as they only need an ARIMA model for forecasting, the comparison between countries is reduced to a unique index, and they are robust models considering the outliers (Debón et al., 2011). The main advantages of multi-population mortality models are that pooling multiple populations can help exploit common information in datasets from different countries in order to identify a more stable trend, reduce statistical errors and increase coherence in forecasting. This summary is particularly relevant for actuaries and demographers, as it enables more efficient and interpretable cross-country mortality comparisons while maintaining robustness with respect to anomalies in the data. Additionally, the low computational cost makes these models practical for real-world applications, such as insurance pricing and reserving, pension forecasting, and public policy planning, where accurate and scalable mortality projections are essential. For these reasons, we think these types of model should be particularly considered for further development and extension.

Artificial intelligence (AI) is transforming actuarial science, providing new opportunities to enhance mortality modelling, insurance pricing, and the estimation of reserves set aside to meet claims. Richman (2021) highlights how deep neural networks have improved predictive accuracy in key actuarial areas, including mortality forecasting, telematics data analysis, and loss reserving in general insurance. In this context, Alonso-García (2023) provides an overview of advances in mortality modeling over the past three decades, highlighting key methodological developments, challenges in estimation and forecasting, and the increasing role of machine learning and AI in refining multipopulation mortality models for actuarial and demographic applications.

In Levantesi and Nigri (2020) the authors introduce the development of new and sophisticated methods for mortality forecasting using a combination of Random Forest (RF) (Breiman, 2001) and two dimensional P-spline (Eilers & Marx, 1996) for both sexes in each country. Building on these advances, Bjerre (2022) explores the use of pure tree-based machine learning methods, specifically RF and Gradient Boosting, to model and forecast mortality, demonstrating their superior predictive performance over traditional stochastic models in most cases.

In this context, Alonso-García (2023) reviews the evolution of mortality models from the classic Lee-Carter model to modern AI-driven approaches, emphasizing the growing integration of machine learning techniques to improve the accuracy and applicability of these models in actuarial and demographic studies. The increasing role of deep learning is further exemplified by Nigri et al. (2019), who propose an integrated Lee-Carter model enhanced with Long Short-Term Memory (LSTM) networks to capture complex mortality dynamics better. Similarly, Garrido et al. (2024) have developed an LSTM-based coherent mortality forecasting model for developing countries, addressing mortality convergence trends across populations. Meanwhile, Euthum et al. (2024) leverage neural networks, including LSTM and Gated Recurrent Units (GRU), to model multipopulation mortality trends, incorporating socioeconomic factors to refine actuarial predictions. Richman and Wuthrich (2019) further validate the utility of recurrent neural networks (RNNs), demonstrating their ability to outperform traditional actuarial models in capturing mortality trends over time. Recently, De Mori et al. (2025) have highlighted the effectiveness of multi-task neural networks for multipopulation mortality forecasting, enhancing prediction accuracy by leveraging shared demographic trends across countries. The combination of these perspectives demonstrates how AI provides more powerful tools for actuarial data analysis and introduces new challenges and opportunities for the actuarial profession, particularly in ensuring model interpretability, robustness, and regulatory compliance.

Our paper seeks to extend the mortality analysis in Li and Lee (2005), Russolillo et al. (2011), and Debón et al. (2011) by modelling simultaneously the mortality from different countries for each sex, and boosting the models using RF. To do this, we analyse male and female mortality data corresponding to the period 1971-2020 in Italy, Spain and the United Kingdom (UK). The thirty years from 1971 to 2000 are used to fit the models and to forecast age-specific death probabilities for twenty years more from 2001 to 2020. By incorporating multiple countries into the analysis, we can effectively compare mortality trends across different demographic and socio-economic contexts, identifying common patterns and country-specific variations in a single index. Moreover, leveraging the power of machine learning, particularly RF, allows us to capture complex nonlinear relationships while maintaining the interpretability of the model, making it a valuable tool for both actuaries and demographers in understanding and predicting mortality dynamics. In this paper, we exploit, on one hand, the advantages of multipopulation approaches for mortality forecasting, and, on the other hand, the machine learning techniques applied to mortality. Thus, we propose as an original contribution the application of machine learning techniques to mortality projections not only by gender but also by country.

In summary, the plan of the paper is the following: Sect. 2 presents the methodologies to be used for trend estimation and for residual analysis. Sect. 3 is devoted to the results of applying the different methods to model the mortality data corresponding to the period 1971-2020 in Italy, Spain and The UK. The conclusions drawn from these results are presented in Sect. 4.

2 Methodology

Actuaries are often more interested in probabilities of death at age x in year t, q_{xt}, than other mortality measures (Currie, 2016) because most actuarial calculations directly involve those measures, although the results can be extended to the force of mortality μ_{xt} or central death mortality rates m_{xt}.

We consider a set of probabilities of death in the form of dynamic life tables for different countries denoted by i. We wish to produce smoother estimates, $\hat{q}_{xti}$, of the true but unknown mortality probabilities q_{xti} from the set of crude probabilities of death, $\dot{q}_{xti}$, for each age x and year t in each region i. The crude probability at age x is typically based on the corresponding number of deaths recorded, d_{xti}, relative to those initially exposed to risk, E_{xti}.

2.1 Lee-Carter models

The classical Lee-Carter model was applied to the annual age-specific central mortality rates, m_{xt}, in Lee and Carter (1992). In that paper, the logit transformation of the annual age-specific probability of death, q_{xt}, $\text{logit}(q_{xt}) = \log\left(\frac{q_{xt}}{1-q_{xt}}\right)$, is modeled as follows,

$$\text{logit}(q_{xt}) = a_x + b_x k_t + \epsilon_{xt}. \tag{1}$$

Cossette et al. (2007) have used the complementary log-log (cloglog) transformation but this choice may be somewhat arbitrary as Haberman and Renshaw (2008) have pointed out. There are different link functions: the justification here for choosing the logit link is the fact that as we work with probabilities the application of this link is guaranteed to provide estimates between 0 and 1.

In the above expression Eq. (1), a_x and b_x are age-dependent parameters and k_t is a mortality index specific for each year. The errors, ϵ_{xt}, reflect age-specific historical influences that are not captured by the model.

This model presents a problem of identifiability (Lee & Carter, 1992); therefore, some constraints must be imposed on the parameters to get a single solution. Although Lee and Carter (1992) propose the normalization $\sum_x b_x = 1$ and $\sum_t k_t = 0$, we propose other constraints, namely that $b_{x_0} = 0$ where x_0 is the first age and $k_{t_0} = 0$ where t_0 is the first time point because they are easier to implement in R.

The standard Lee and Carter (1992) model is often estimated using Singular Value Decomposition (SVD), which provides a computationally efficient decomposition but lacks a probabilistic foundation. Generalized non-linear models (GNM) offer more flexibility in incorporating distributional assumptions and additional constraints on parameters (Currie, 2016). Empirical studies, such as Debón et al. (2008, 2010a), confirm that GNM generally provides superior goodness-of-fit and robustness with respect to irregularities in mortality data. Given these advantages, GNM is preferred over SVD for mortality modelling within the Lee-Carter framework.

The analysis was undertaken using purpose-written code in R Core Team (2024). The fitting procedure of a range of Lee-Carter models using gnm library (Turner & Firth, 2023) can be found in Debón et al. (2010a) and Currie (2016). In addition, Currie (2016) compared Poisson models for the force of mortality and Binomial models for the rate of mortality over six countries, and, in most of them, the Binomial models outperform the Poisson ones in terms of goodness of fit. However, using the Binomial family, the results show overdispersion as the Deviance statistic is greater than the degrees of freedom in the model; therefore, the quasiBinomial family is recommended. The following models have been fitted following these same procedures and recommendations.

2.1.1 Additive model

By analogy with the Lee-Carter models and in order to obtain a parsimonious multipopulation mortality model, Debón et al. (2011) propose adding a factor index that specifically modifies mortality for each member of the group. The proposed model is,

$$\text{logit}(q_{xti}) = a_x + b_x k_t + I_i + \epsilon_{xti}. \tag{2}$$

Again, we have an identifiability problem in the model that has been effectively addressed within the model above through the imposition of constraints, specifically setting k_{t_0}=0 and b_{x_0}=1, while stipulating I_1=0. Therefore, taking the population 1 of the group as a reference and making $I_1 = 0$, the index I_i means the additive change necessary for transforming the $\text{logit}(q_{xt1}) = a_x + b_x k_t$ in population 1 to that of a population *i*. In addition, this model assumes that differences in the mortality of specific populations are age and time-independent.

Consequently, this is the interpretation of the parameters:

1. a_x coefficients describe the shape of the age profile in population 1 for the period t_0.
2. The values k_t represent the trend of mortality in population 1 during the period.
3. The evolution of b_x gives an idea of how fast the ratios decrease in response to changes in k_t. Noting that, for many developed countries, the trend in k_t has been downward over time.

$$\frac{d\text{logit}(q_{xti})}{dt} = b_x \frac{dk_t}{dt},$$

4. The succession of values I_i allows for comparing mortality patterns between population 1 and other populations i. Positive values mean higher mortality probabilities than population 1, and negative values mean the opposite.

2.1.2 Multiplicative model

In contrast to the previous model, the Russolillo et al. (2011) proposal, called here the multiplicative model, includes the population as a multiplicative index and assumes that differences in mortality in specific populations are dependent on age and time. This model can easily be compared with Eq. (2) as both share the same number of parameters. Its expression is,

$$\mathrm{logit}(q_{xti}) = a_x + b_x k_t I_i + \epsilon_{xti}. \tag{3}$$

Therefore, the index I_i stands for the specific change shown in each population by the increments, $b_x k_t$, taking place concerning the general behaviour of the logit of mortality, a_x. The problem of identifiability in Eq. (3) is solved by setting k_{t_0}=0, b_{x_0}=1 and I_1=1. Then, the following is the interpretation of the parameters,

1. a_x coefficients describe the average shape of the age profile for the period t_0.
2. The evolution of b_x gives an idea of how fast the age decreases in response to changes in k_t for population 1,
3. The evolution of $b_x I_i$ gives an idea of how fast the age-specific death probabilities decrease in response to changes in k_t for population *i*. I_i values smaller than 1 mean higher mortality probabilities than population 1, and I_i values higher than 1 mean the opposite.
4. The values k_t represent the trend of mortality during the period.

2.2 Li and Lee models

Li and Lee (2005) propose a more complex variant of the original Lee-Carter model for estimating mortality in countries that form part of a group instead of considering them individually. This model is expressed by,

$$log(m_{xti}) = a_{xi} + b_x k_t + b_{xi} k_{ti} + \epsilon_{xti}. \tag{4}$$

To avoid long-term divergence in mortality forecasting for the group, all populations must have the same b_x and drift term for k_t. At the same time, a_x can be estimated separately for each population and denoted as a_{xi}. Debón et al. (2011) adapt the reduced version of this model called the common factor model by Li and Lee (2005) to the $\mathrm{logit}(q_{xti})$, that is,

$$\mathrm{logit}(q_{xti}) = a_{xi} + b_x k_t + \epsilon_{xti}, \tag{5}$$

the problem of identifiability is solved with the following restrictions: $k_{t_0} = 0$ and $b_0 = 1$. The interpretation of the parameters is as follows,

1. a_{xi} coefficients describe the average shape of the age profile in the region *i*.
2. The evolution of b_x gives an idea of how fast the ratios decrease in response to changes in k_t for the country.
3. The values k_t represent the general trend of mortality during the period.

The Li-Lee based models and, in general, multipopulation models depend on the assumption that mortality rates are similar between countries. This hypothesis is not always supported (Grigoriev et al., 2010) and has to be verified case by case.

2.2.1 Li-Lee additive and multiplicative models

We propose to modify model (5) including a population index I_i in an additive way,

$$\operatorname{logit}(q_{xti}) = a_{xi} + b_x k_t + I_i + \epsilon_{xti}, \tag{6}$$

or in a multiplicative way,

$$\operatorname{logit}(q_{xti}) = a_{xi} + b_x k_t I_i + \epsilon_{xti}, \tag{7}$$

with the corresponding restrictions $I_1 = 0$ and $I_1 = 1$. In this way, the differences between the countries are collected through the index I_i. Eq. (4) is equivalent to Eq. (6) where $b_{xi}k_{ti}$ is synthetized into I_i . The model fitting is easily obtained with the gnm library, providing as an initial point the SVD solution of the model in Eq. 5. These models considerably improve the comparison of countries using I_i with similar interpretations as in the above corresponding models - additive as in Eq. (2) and multiplicative as in Eq. (3).

2.3 Residual analysis: Random Forest

The residuals of the above models can be modelled with a non-parametric technique called a Random Forest (RF), an ensemble of classification and regression trees (CART). The CART construction for regression, which applies in our case, using the ANOVA method, involves the following steps (Therneau & Atkinson, 2023). The technique identifies the single variable that optimally divides the dataset into two groups, aiming for maximal homogeneity within each group and maximal dissimilarity between them. The dataset is then partitioned accordingly, and this process iterates recursively on each variable until it reaches a minimum size or further improvement becomes unattainable.

The predictive supervised learning method RF improves the prediction capacity by combining independent CART using the bagging technique Breiman (1996)[1] and by randomly selecting, at each node, a subset of m_{try} variables ($1 \leq m_{\text{try}} \leq p$, with p the total number of predictors). Only this subset is searched when choosing the best split, so m_{try} governs the trade-off between tree strength and inter-tree correlation: smaller values encourage diversity (higher bias, lower correlation), whereas larger values strengthen individual trees (lower bias, higher correlation). Common defaults are $m_{\text{try}} = \sqrt{p}$ for classification and $m_{\text{try}} = p/3$ for regression Breiman (2001). The RF then resamples the training observations with replacement, and the unsampled observations are called out-of-bag (OOB) observations, which are used to measure the prediction error of the method. A more detailed description of CART and RF can be found in Hastie et al. (2009).

The RF algorithm captures complex, non-linear relationships between a dependent variable and its predictors by growing an ensemble of bootstrapped decision trees, each split on a random subset of covariates; it then produces a prediction by averaging the conditional responses across all trees, thereby reducing variance and enhancing out-of-sample accuracy. In our analysis, the dependent variable is the mortality-model residuals, while Year, Age, and Country serve as explanatory covariates.

The RF training process was performed by employing cross-validation with $k = 5$ to avoid overfitting, using the train function of the caret R-package (Kuhn & Max, 2008). We adopted $k = 5$because it is the default in caret and a widely recommended bias-variance

[1] Bagging is the acronym of "bootstrap aggregating". Bagging predictors involves a key step of creating multiple versions of a predictor. The different versions are created by making bootstrap samples from the original learning set to train the predictor, and after those versions are combined into a single, aggregated predictor.

compromise (Hastie et al., 2009). Therefore, the "mtry" hyperparameter of the RF models was optimised. For our three-dimensional data (indexed by x, t, and i), the 5-fold cross-validation process works as follows. The entire dataset is divided into five equally sized subsets. In each of the five iterations, one subset is held out as a validation set while the remaining four are used to train the RF model. This systematic rotation ensures that all parts of the data serve as unseen validation data exactly once. As a result, the model's predictive performance is robustly evaluated to optimise the "mtry" hyperparameter.

Moreover, RF analysis empowers us to assess the importance of variables within the regression model (Breiman, 2001). Liaw and Wiener (2002) introduced two metrics, Inc-MSE% and IncNodePurity, within their randomForest R-package:

- %IncMSE (Mean increase MSE): obtained by calculating how much the prediction error (MSE) of the OOB observations increases when the data of one variable are permuted, leaving the rest unchanged.
- IncNodePurity (Increase Node Purity): is the total decrease in node impurities due to splitting the variable, averaged across all trees. For regression, node impurity is measured by the mean squared error (MSE).

The %IncMSE measure is better as it is more robust than IncNodePurity. IncNodePurity is only used if the extra computing time is unacceptable.

However, the importance of the variables only establishes a ranking of variables. Partial dependent plots (PDPs) offer a simple solution to quickly understand the relationship between outcome and predictors of interest (Greenwell, 2017). PDPs are graphical representations of the predicted outcome of a model vs features to show whether the relationship is linear or not, positive or negative. PDPs are especially useful for visualizing the relationships discovered by complex machine learning algorithms such as a RF.

In this paper logit residuals with expression,

$$r_{xti} = \text{logit}(\dot{q}_{xti}) - \text{logit}(\hat{q}_{xti}) \tag{8}$$

have been modeled.

We select the logit scale because it provides a convenient and additive framework: differences (residuals) on the logit scale can be directly added to the predicted logits to adjust the estimates. Moreover, since the RF model is non-parametric, it does not require the residuals to follow a symmetric or Normal distribution, allowing us to capture deviations effectively without imposing strict distributional assumptions.

2.4 Evaluation of models

In summary, the interpretations of the model parameters provide clarity and trend identification and facilitate comparative analysis, making it a valuable tool for demographic research, policy formulation, and forecasting. The additive model emphasizes the average shape of the age profile in population one and the trend of mortality during the specified period. It provides insights into overall demographic patterns and mortality trends but may not explain how mortality varies across age groups in different populations. On the other hand, the multiplicative model offers a more nuanced analysis by considering the dynamic response of age-specific mortality probabilities to changes in a critical factor (k_t). It also allows for comparative analysis across populations, which can be valuable for understanding differences in mortality patterns across demographic groups.

Therefore, the additive model may be more appropriate if the goal is to understand overall demographic trends and mortality patterns in population 1. However, suppose the analysis requires a detailed examination of how mortality varies across different populations and age groups and how it responds to changes in a specific factor. In that case, the multiplicative model may be preferable. On the other hand, if there are conditions of differences in the mean mortality of the populations that will be maintained but are not expected to diverge in the future, the models derived from Li and Lee's proposal will be the most appropriate. Ultimately, the choice between the models should be based on the specific research questions, objectives, and numerical results.

In general, there are three strategies for the validation of the numerical results of the model predictions:

1. evaluate the model in a test sample different to the fitting sample,
2. develop the model with around 70 % of the sample and calculate the predictive power with the remaining 30%, or
3. use the same sample, but calculate predictive indicators using bootstrap or resampling techniques.

In this paper, we use the second and the third as we only have one large sample. Specifically, we use the hold-out method for time series, which separates the data into two subsets in chronological order, one used to train the model and the other one to perform the validation test. We have used 60% (30 years) of the original periods to develop the models (training set) and calculated the predictive power with the remaining 40% (20 years) of the periods (validation set). The hold-out method is widely used in the mortality literature (e.g., Debón et al. (2010b), Ahcan et al. (2014), Danesi et al. (2015), Neves et al. (2017), Diaz et al. (2018), and Atance et al. (2020)) because it respects the inherent temporal ordering of the data, ensuring that predictions are genuinely forward-looking.

The steps in the hold-out were as follows:

1. Mortality models were fitted to the training dataset.
2. The index k_t was predicted using a time series model (ARIMA) for all years in the test period.
3. Probability of death predictions $(\hat{q}_{xti})$ were generated with the predicted indexes (obtained in the previous step) for the test period.
4. The model predictions $(\hat{q}_{xti})$ were compared with the observed mortality probabilities $(\dot{q}_{xti})$ in the validation period obtaining measures of goodness of fit.

Additionally, we have used resampling methods, such as k-fold cross-validation–implemented so that the folds respect the chronological ordering of the data and with fixed origin in year 1981–to provide a clear framework for evaluating both short-term and long-term forecast accuracy by simulating a realistic forecasting scenario.

Therefore, we are going to use measures of goodness of fit in training and test sets such as:

- Mean Squared Error (MSE) that measures the average of the squares of the errors (difference between the predictions and the actuals)

$$MSE(\hat{q}) = mean\left((\dot{q}_{xti} - \hat{q}_{xti})^2\right), \tag{9}$$

- Root Mean Squared Error (RMSE) that measures the average error

$$RMSE(\hat{q}) = \sqrt{MSE(\hat{q})}, \tag{10}$$

and,

- Mean absolute percentage error (MAPE) that measures the average absolute percentage error

$$MAPE(\hat{q}) = mean\left(\frac{|\dot{q}_{xti} - \hat{q}_{xti}|}{\dot{q}_{xti}}\right). \tag{11}$$

In the field of mortality modelling, cross-validation (CV) has only recently been introduced. Most existing applications have focused on single-population models (e.g., see Atance et al. (2020), Kessy et al. (2022), Lindholm and Palmborg (2022), and Barigou et al. (2023)), highlighting the limited use of resampling methods in multipopulation contexts. Our work extends this approach by employing the CvmortalityMult R-package (Atance & Debón, 2025) and some modified functions to implement CV for the proposed multipopulation mortality models.

3 Application of the models

All these models will be used to fit and predict mortality rates for the three countries to explore and understand their similarities and differences.

3.1 Description of the data

The data used in this analysis come from the Human Mortality Database (2024). In particular, we have worked with published male and female life tables for Spain, Italy and the UK mortality data corresponding to 1971–2020. The models described in Sect. 2 have been used to fit life tables from the three countries for 1971–2000 and a range of ages from 21 to 100 for each sex. Then, the fitted models have been used to obtain predictions for 2001–2020 and compare them with the observed values.

When forecasting mortality for pension and insurance purposes, it is common practice to focus on the age ranges for adults and the elderly rather than for the entire human life span. Life insurance and pension products primarily cover working-age individuals and retirees, and adult and elderly mortality trends significantly influence the financial risk associated with these products. In contrast, infant and adolescent mortality exhibits much lower variability and has a negligible financial impact in actuarial applications (Cairns et al., 2006;

Plat, 2009). Given these considerations, restricting the analysis to a relevant age range enhances the precision and practical applicability of the mortality predictions.

Figures 1 and 2 show the behaviour of the average logit of the probability of death for ages and years, respectively, by sex and country. This exploratory study shows the typical average mortality profile across ages, the decline in mortality rates over the years (Fig. 2), and the different patterns by sex and country with higher mortality rates for males and the UK for most ages, specifically intermediate and the oldest ages (Fig. 1) and a more pronounced decline for females (Fig. 2). The exploratory analysis in Fig. 2 reveals a partial convergence in overall mortality among Spain, Italy, and the United Kingdom, grouping the three within the same developed-country cluster under the criteria of Atance et al. (2024). However, a full-surface cluster analysis (Debón et al., 2017) still singles out the UK as distinct from its two Mediterranean counterparts. This residual difference is precisely what the country-specific index in our proposed multipopulation models is designed to capture.

3.2 Model fitting

The estimation of parameters has been obtained for the models for the three countries for each sex. According to Debón et al. (2011), which applies them to Spanish regions, these models can obtain life tables for the different populations that do not differ too much, eliminate the irregularities in those with smaller populations and finally, respect their peculiarities. However, in this new application, we intend to quantify and describe the differences between these three countries for each sex, taking full advantage of the interpretability of these models.

Figures 3 to 7 illustrate the estimated parameters obtained from the models. Figs. 3 and 4 reveal similar values for parameter a_x across all models. In fact, it is hard to distinguish

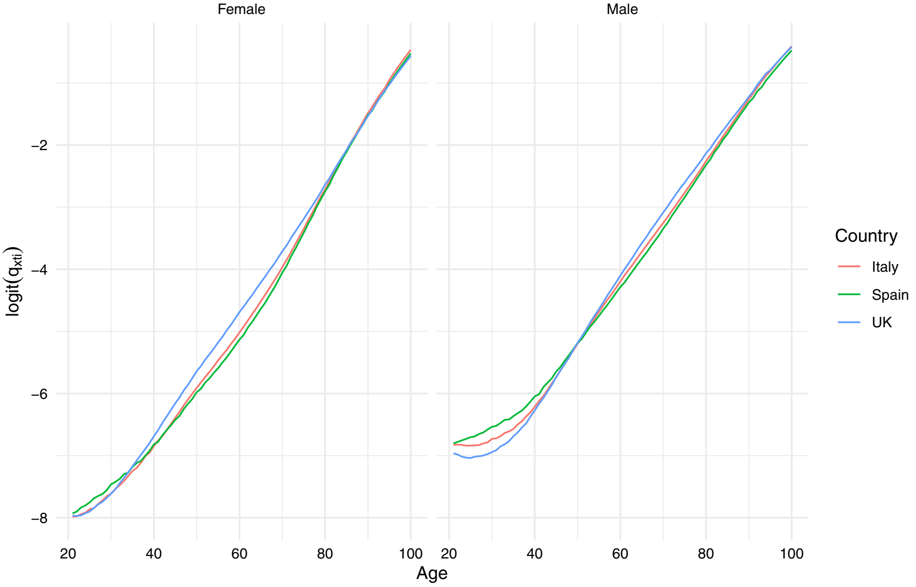

Fig. 1 Behaviour of the average across all years of the logit probability of death for ages by sex and country

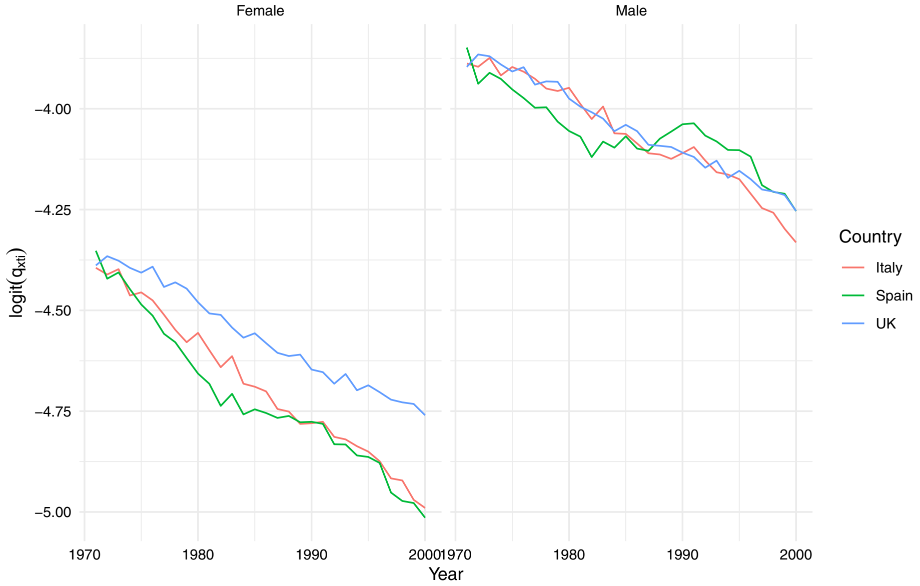

Fig. 2 Behaviour of the average across all ages of the logit probability of death for years by sex and country

the results for the 4 different models but they indicate consistently higher mortality rates for males across all age groups. Moving on to Fig. 5, we see positive values for b_x in females for all ages, suggesting a decrease in mortality. Males, however, exhibit a different pattern. Their b_x values are negative between ages 23 and 37, indicating an increase in mortality for that specific age range, followed by a decrease. In both sexes, the model with the different behaviour is multiplicative; the other three show a very similar evolution for the b_x parameter. These negative b_x values for young adult males are plausibly linked to transient, cause-specific mortality shocks (e.g., AIDS, drug use, violent deaths or war casualties) that have since diminished; not modelling such exogenous events explicitly may bias forecasts for this cohort and therefore constitutes a limitation that could be addressed in future work by incorporating external covariates or event indicators.

Figure 6 depicts a general decline in mortality over the years for both genders, with a steeper decrease observed for females and the Li-Lee-multiplicative model. Finally, Fig. 7 focuses on the country index. Italy[2] is the reference country, so its index is 0 in the additive model and 1 in the multiplicative model by construction. Spain displays consistently lower death probabilities than Italy, whereas the United Kingdom records the highest mortality of the three countries in every specification except the Liâ€“Lee multiplicative model. In the Liâ€“Lee additive formulation, all I_i indices are zero because the country-specific terms a_{xi} already capture national differences. The index for Spain is smaller than that of the UK for the Li-Lee-multiplicative model for males, indicating a smaller decrease in mortality rates but relative to the country-specific a_{xi}.

Table 1 shows the MSE, RMSE and MAPE for both models in each sex. The Li-Lee versions improve the fitting, and the Li-Lee-multiplicative model shows the best global result

[2] The reference population is selected by the first in alphabetical order by default in R.

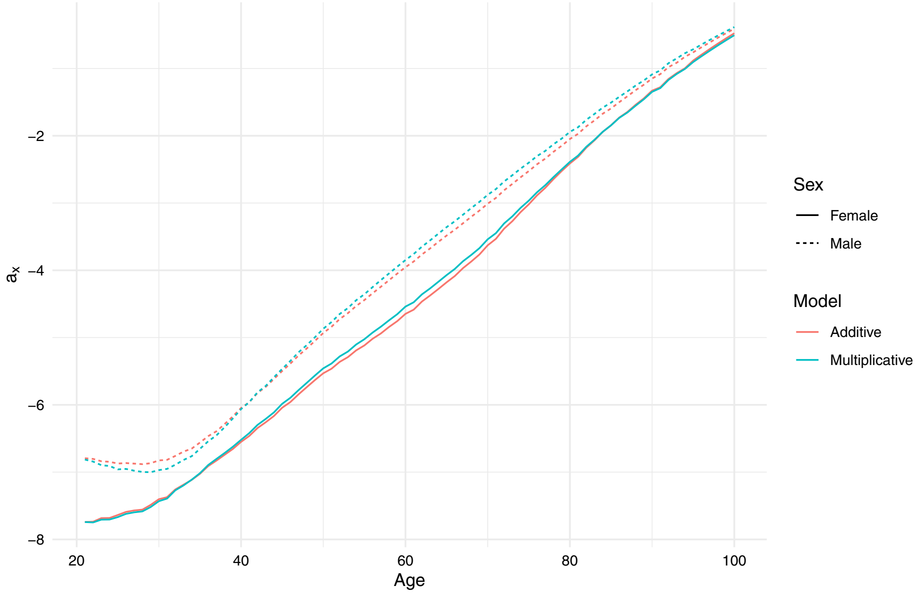

Fig. 3 Estimated a_x parameter for additive and multiplicative models for each sex

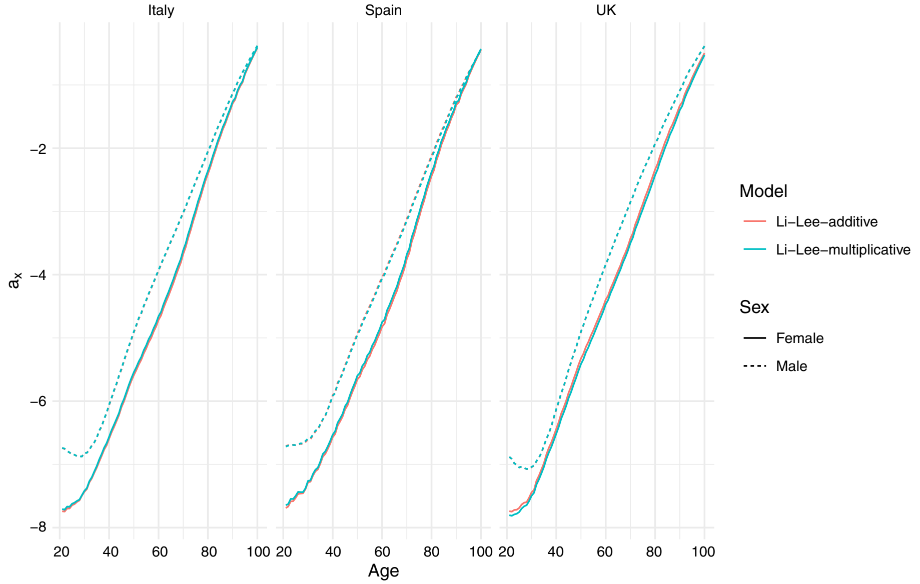

Fig. 4 Estimated a_x parameter for additive and multiplicative models for each sex

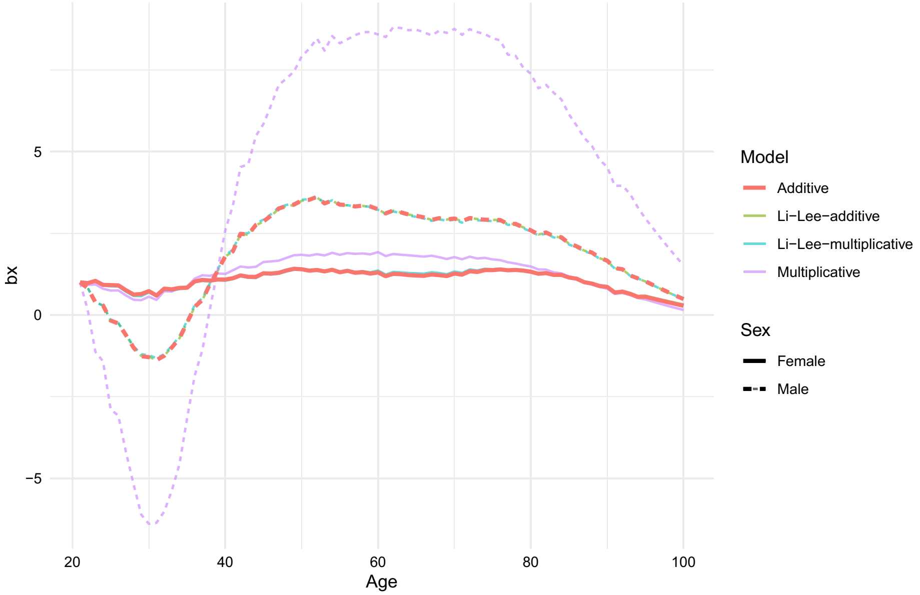

Fig. 5 Estimated b_x parameter for additive and multiplicative models for each sex

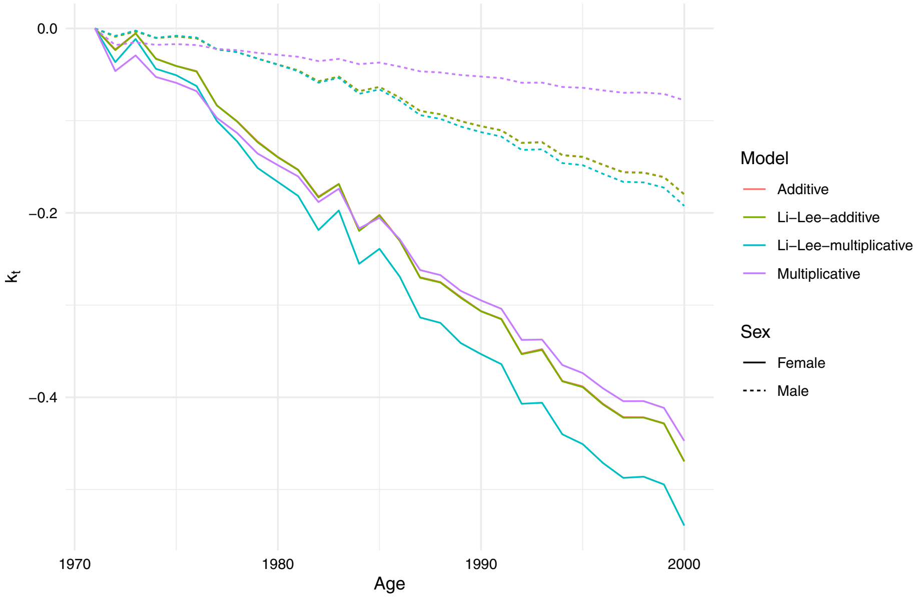

Fig. 6 Estimated k_t parameter for additive and multiplicative models for each sex

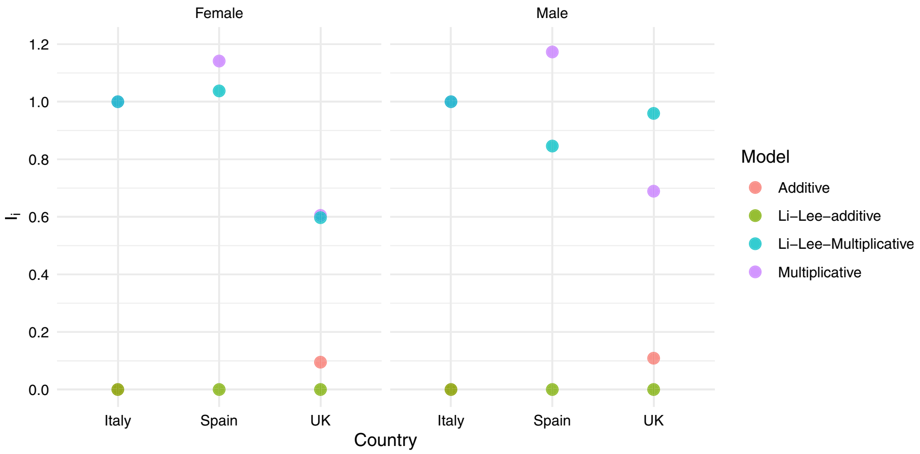

Fig. 7 Estimated I_i parameter for additive and multiplicative models in each sex

Table 1 Goodness-of-fit measures for the models in training data set

Sex	Model	MSE	RMSE	MAPE
Females	Additive	0.00005	0.007117	8.50
	Multiplicative	0.00003	0.005721	7.05
	Li-Lee-additive	0.00001	0.003500	5.30
	Li-Lee-multiplicative	0.00001	0.003309	4.82
Males	Additive	0.00003	0.005408	8.83
	Multiplicative	0.00002	0.004503	8.08
	Li-Lee-additive	0.00002	0.003980	5.33
	Li-Lee-multiplicative	0.00002	0.004029	5.12

for goodness-of-fit measures with the lowest MAPE for both sexes and lower RMSE in almost all cases except men where the additive Li-Lee model is lower by a small difference. In contrast, Debón et al. (2011) found the additive model shows the best global result for both sexes and the deviance goodness-of-fit measure in an application to Spanish regions. However, in this study, the prediction model's performance is evaluated with MSE and RMSE, measures of the distance between observed $\dot{q}_{xti}$ and adjusted values $\hat{q}_{xti}$ in each population i, whose expression measures the error of estimations without any correction and they are easier to interpret than the deviance. On the other hand, MAPE measures the relative error, giving less importance as the probability of death increases so that, for the high age probabilities, it would allow more error. Thus, MAPE is less suitable for measuring prediction error in this context, which is why the Li-Lee-multiplicative model is the best in relative terms but also globally.

Renshaw and Haberman (2006) suggest carrying out diagnostic checks on the fitted model by plotting residuals. In this study, we are going to model them with RF and see their importance and then their behavior with PDPs: this will allow us to improve predictions and understand the relationships. Fig. 8 summarises the relative contribution of the quantitative variables age and year, and of the categorical variable country (Spain and the United Kingdom, with Italy as the reference). In both models and across sexes, predictor importance declines in the order Age, Year and Country.

Next, PDPs are shown in Figs. 9, 10, and 11 with logit residuals (Eq. 8). Figs. 9, 10, and 11 show which ages, years and countries values of the residuals are predicted closest to zero and therefore where the models work best.

3.3 Model forecasting

The predictions beyond the last time period are carried out by the projection of time series previously adjusted to the time parameters k_t. The corresponding *ARIMA* models are obtained using the functions *auto.arima* and *forecast* from the forecast R-package (Hyndman et al., 2024; Hyndman & Khandakar, 2008). Fitted *ARIMA* models for models k_t index for females and males are shown in Table 2 and drawn in Figs. 12, 13, 14,and 15, respectively.

Then, $\operatorname{logit}(\hat{q}_{xti})$ are obtained by substituting the corresponding predicted $\hat{k}_t$ in the fitted expression of the corresponding model.

Additionally, we propose to obtain predicted logit residuals $\hat{r}_{xti}$ using RF models for obtaining better predictions as follows,

$$\log\left(\frac{\hat{q}_{xti}}{1-\hat{q}_{xti}}\right)+\hat{r}_{xti}.$$

Table 3 shows MSE, RMSE and MAPE for all models for predictions in the 2011-2020 period, with slightly better results for the RMSE measure for the additive model than for the multiplicative model for both sexes. However, the multiplicative version of the Li-Lee framework demonstrates comparable or superior performance. Table 3 also shows the predictions by adding residual modelling which improves the results in all cases.

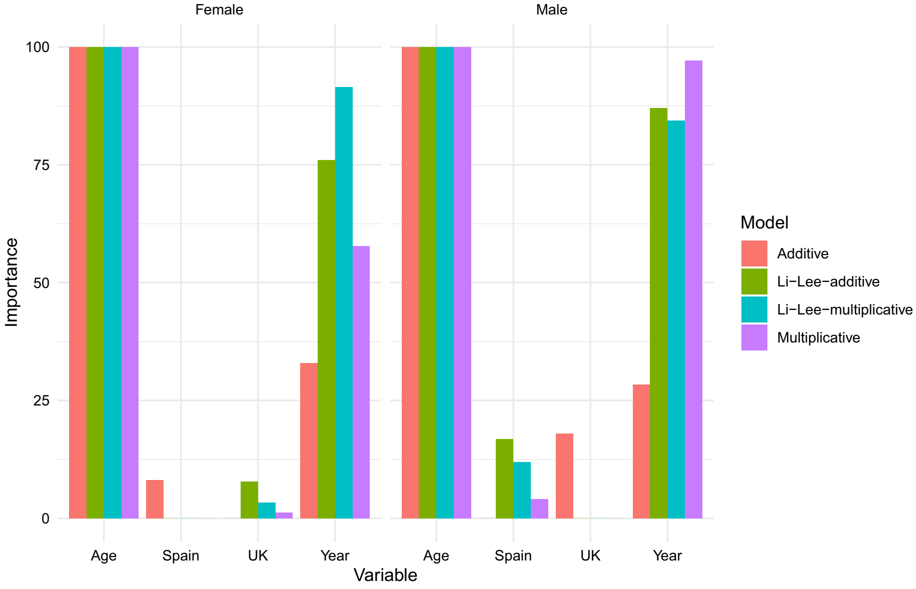

Fig. 8 Importance of the factors in the model Random Forest for residuals

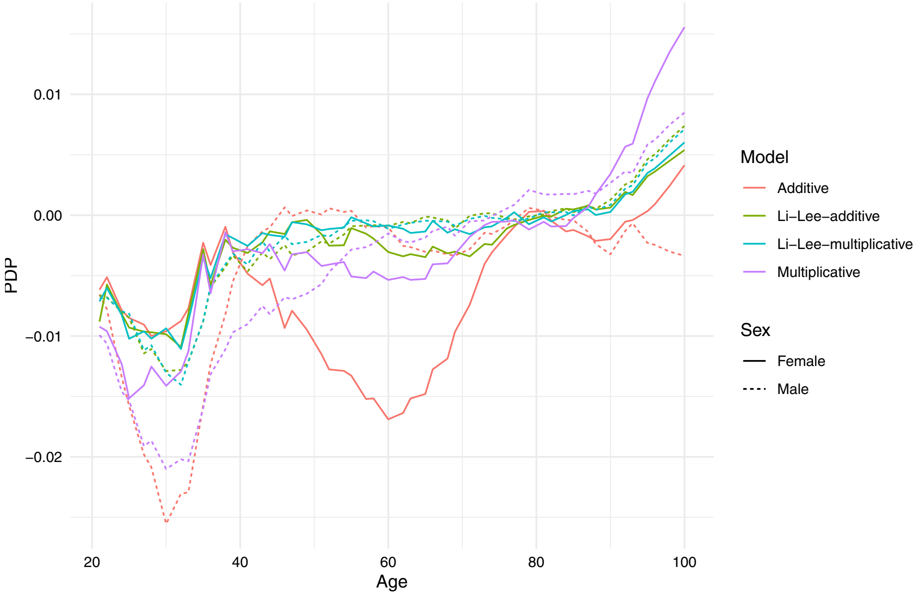

Fig. 9 Partial dependent plot for age in the model Random Forest for residuals

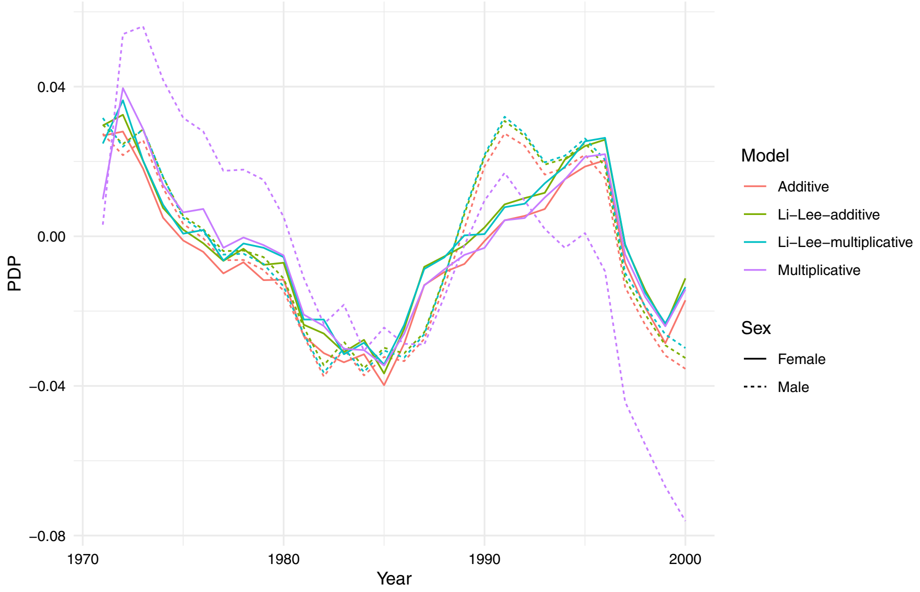

Fig. 10 Partial dependent plot for year in the model Random Forest for residuals

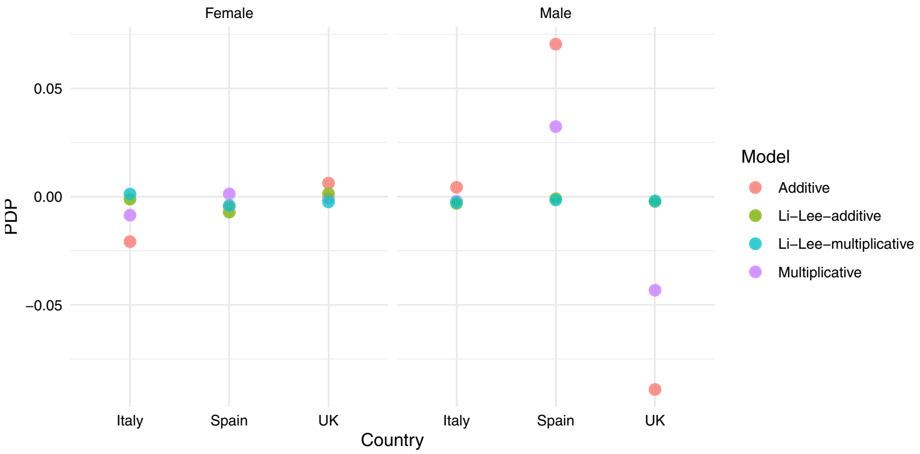

Fig. 11 Partial dependent plot for country in the model Random Forest for residuals

Table 2 ARIMA models for k_t of each model and sex

Sex	Additive	Multiplicative	Li-Lee-additive	Li-Lee-multiplicative
Females	ARIMA(1,1,0)	ARIMA(2,1,0)	ARIMA(1,1,0)	ARIMA(1,1,0)
Males	ARIMA(1,1,0)	ARIMA(1,1,0)	ARIMA(1,1,0)	ARIMA(1,1,0)

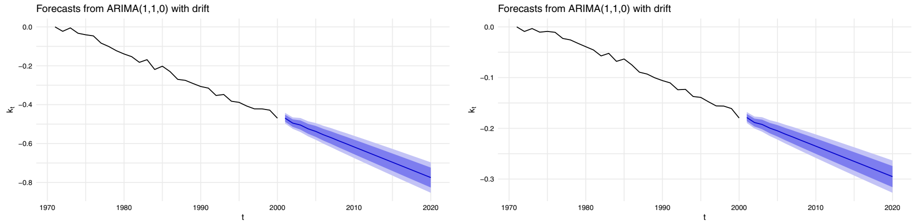

Fig. 12 *ARIMA* models from the functions *auto.arima* and *forecast* for Additive model mortality index for females (left) and males (right)

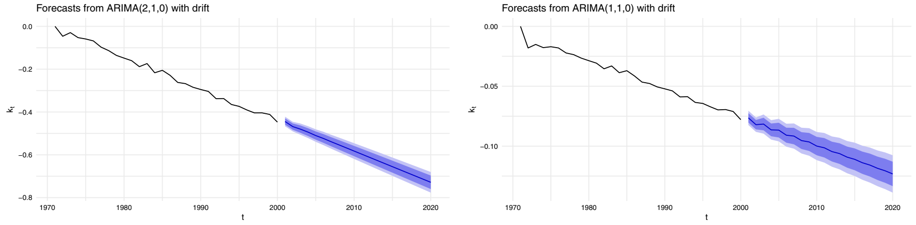

Fig. 13 *ARIMA* models from the functions *auto.arima* and *forecast* for Multiplicative model mortality index for females (left) and males (right)

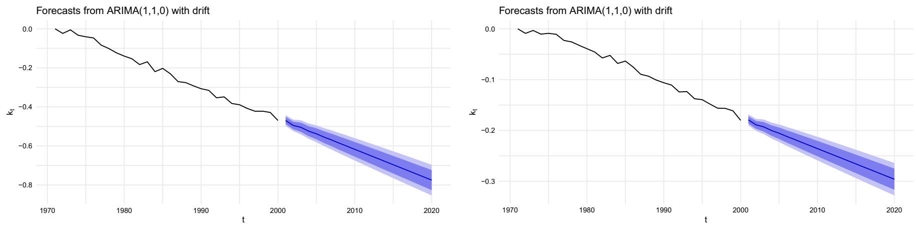

Fig. 14 *ARIMA* models from the functions *auto.arima* and *forecast* for Li-Lee additive model mortality index for females (left) and males (right)

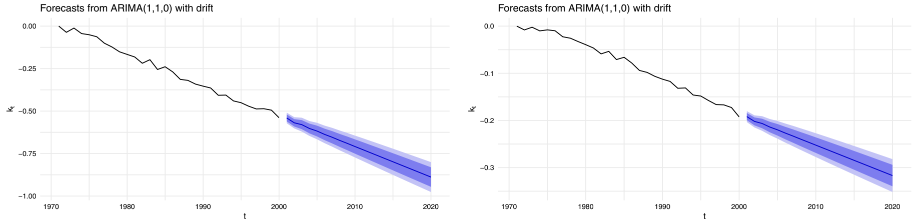

Fig. 15 *ARIMA* models from the functions *auto.arima* and *forecast* for Li-Lee multiplicative model mortality index for females (left) and males (right)

Table 3 Goodness-of-fit measures for predictions in test set of each model

Sex	Model	MSE	RMSE	MAPE
Females	Additive	0.000029	0.005357	15.52
	Additive+RF	0.000017	0.004148	10.47
	Multiplicative	0.000039	0.006287	16.40
	Multiplicative+RF	0.000031	0.005588	13.40
	Li-Lee-additive	0.000025	0.004968	14.71
	Li-Lee-additive+RF	0.000019	0.004417	10.52
	Li-Lee-multiplicative	0.000020	0.004482	13.97
	Li-Lee-multiplicative+RF	0.000019	0.004449	10.72
Males	Additive	0.000040	0.006398	30.67
	Additive+RF	0.000025	0.005043	22.58
	Multiplicative	0.000053	0.007281	42.75
	Multiplicative+RF	0.000028	0.005269	19.33
	Li-Lee-additive	0.000027	0.005196	32.25
	Li-Lee-additive+RF	0.000025	0.005006	22.55
	Li-Lee-multiplicative	0.000026	0.005084	31.19
	Li-Lee-multiplicative+RF	0.000025	0.005045	22.43

Furthermore, residuals are modelled primarily to analyze their behaviour rather than for predictive purposes. The forecasting process relies on the projection of the models, ensuring that predictions are based on each approach's underlying structure and assumptions rather than direct extrapolation of residual patterns. Therefore, we obtain goodness-of-fit measures for predictions using 10-fold cross-validation for each model, with an initial training period spanning ten years (1971–1980) and forecasts covering the period from 1981 to 2020. Table 4 presents the MSE, RMSE, and MAPE values, averaged over 10 folds with size 4 years: 1981–1984, 1985–1988, 1989–1992, 1993–1996, 1997–2000, 2001–2004, 2005–2008, 2009–2012, 2013–2016, and 2017-2020. Again, the results indicate that the RMSE values are slightly lower for the additive model than for the multiplicative model for both sexes. However, the multiplicative version demonstrates comparable or slightly superior performance in the Li-Lee framework.

To illustrate the variability in model performance across different ages and validation folds, the MSE values for models on average for each age and each fold for females and males are presented in Figs. 16 and 17, respectively. Fig. 16 reveals that the MSE varies between ages, with the multiplicative model performing worse in most specific age groups. Higher MSE values at extreme ages suggest potential challenges in capturing mortality trends among the oldest ages. Meanwhile, the trends observed in Fig. 17 suggest that MSE fluctuates over time, with variations in predictive accuracy depending on the model specification and sex, providing insight into the consistency and robustness of each approach. In particular, the Li-Lee models exhibit MSE values lower than their additive and multiplicative counterparts in certain validation periods, indicating an improved fit in those cases. We also find that the prediction error is generally smaller when the test fold is shorter than the corresponding training set, with the single exception of the final fold, which includes the excess mortality produced by the COVID-19 pandemic.

Dong et al. (2020) found that the multiplicative model's out-of-sample forecasting performance is significantly improved for individual populations and the aggregate population compared with using the single-population mortality model based on rank-1 singular value decomposition (SVD) which corresponds to the Lee-Carter model. Their results also shed light on the similarities and differences in mortality among 10 European countries (Denmark, United Kingdom, Finland, France, Italy, the Netherlands, Norway, Spain, Sweden,and Switzerland) and the 2 genders.

4 Conclusions

To understand forecasting error, evaluating error in specific-age death probabilities is essential (Booth et al., 2006); therefore, we focused on predicting mortality probabilities instead of using mortality indicators for life expectancy. Additionally, Santolino (2023) has concluded that models that provide a good fit or a good prediction performance on the log scale might be inadequate in the original scale, and vice versa, and using one selection measure or another ultimately depends on the decision-maker's preferences. Therefore, in this paper, we have included absolute and relative goodness-of-fit measures but calculated on the scale of the original probabilities because those are used in actuarial and related calculations and,

Table 4 Average goodness-of-fit measures obtained from 10-fold cross-validation predictions, based on the initial training set (1971–1980), for all models and sexes

Sex	Model	MSE	RMSE	MAPE
Females	Additive	0.000033	0.005745	11.43
	Multiplicative	0.000041	0.006403	10.03
	Li-Lee-additive	0.000026	0.005099	11.33
	Li-Lee-multiplicative	0.000021	0.004583	9.02
Males	Additive	0.000035	0.005916	13.08
	Multiplicative	0.000039	0.006245	15.80
	Li-Lee-additive	0.000031	0.005568	13.54
	Li-Lee-multiplicative	0.000030	0.005477	13.78

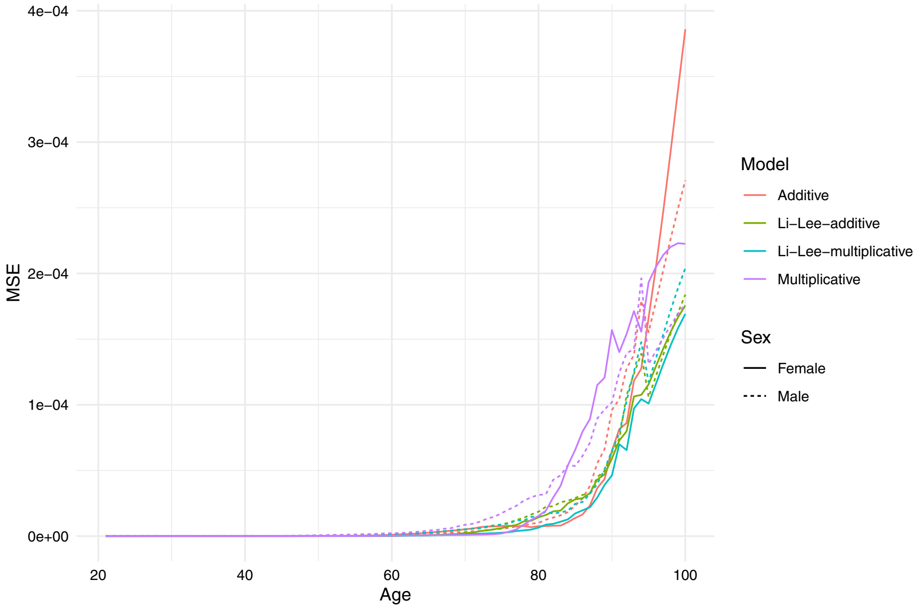

Fig. 16 Average *MSE* by age from 10-fold cross-validation predictions, using the initial training set (1971-1980), for all models and sexes

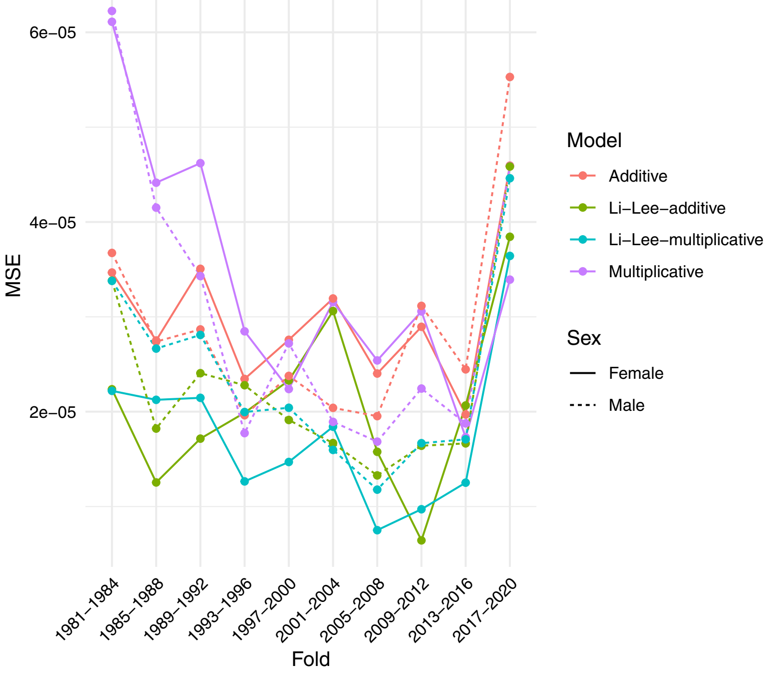

Fig. 17 Average *MSE* by fold from 10-fold cross-validation predictions, using the initial training set (1971-1980), for all models and sexes

therefore, have direct economic consequences. We have applied four multipopulation models to Italy, Spain and UK data for both male and female and have implemented a RF on the residuals of each models in order to improve the accuracy of the projections. Table 1 shows the goodness-of-fit performance values for the analysed models in the training data set. In general, including the multiplicative country effect improves the model's fitting relative to the additive one for males and for females. On the other hand, the multiplicative model shows the best global results for both sexes and all performance measures. The explanation for this can be found in introducing the country effect as a multiplicative term, which better adapts the model for the countries involved in the study and for intermediate and advanced ages. Note that the additive model has a simpler structure because it only considers the main effects, assuming that differences in the mortality of specific populations are age and time-independent (Debón et al., 2011). It has to be acknowledged that mortality across countries could undergo different phases and differences in mortality are often time-dependent (Raftery et al., 2013). However, as highlighted by Léger and Mazzuco (2021), these differences change over time, especially for infants, while our dataset is composed of adults and the elderly (ages 20–100). These authors group countries with similar evolution of mortality over time into a homogeneous cluster. In particular, they observe for the years 1960-2018 that Italy and Spain are in the same clusters during the early years, and they together undergo a rapid transition from one cluster to another, while such a transition is slower in the United Kingdom, but essentially the difference between cluster is due to higher infant mortality of Southern countries in the first period. In the second half of the period, the disparities seemed to be reduced, and all the countries followed the shifting and compression process of the mortality curves previously described. In any case, even if time-dependent differences emerge, it is interesting to observe from our analysis that the limitation of not considering the factor I_i dependent on age and time is overcome by the fact that the RF is applied to residuals that depend on age, time and countries. In this way, the residuals of each model are adjusted through the RF, taking into account also the time and age components. And empirically, the model that works better in terms of accuracy on the test set is the Additive one plus the RF (see Table 3).

Table 3 shows the prediction model's performance values for all models. A first conclusion, common to all models, is that prediction performs better for females than males. Further, the multiplicative model shows a worse global result for the prediction performance evaluation measures for both sexes. Taking into account the results from Table 1 relative to Tables 3 and 4, the multiplicative model can lead to over-fitting. The additive model can be more robust, and its parameters are better estimated with less erratic behaviour than the multiplicative model. In addition, residual modelling improves predictions and helps us to understand the relationships between the residuals and the underlying factors of age, year and country.

We note that the models have some advantages: easily interpretable parameters in as much as they describe the evolution of mortality over age, period and country; their computational cost is meagre as they only need an ARIMA model for forecasting; the comparison between countries is reduced to a unique index, and they are robust models considering the outliers. One comment must be made about the additive model; this type of model must be considered for future development as an alternative to similar models to predict a group of related populations.

Concerning the work of other authors, we should highlight two distinctive features of the methodology presented here: first, that models are fitted using maximum-likelihood, and second, the projections of different models are compared in terms of error in probabilities of death, and not in terms of the error in logarithmic transformation of mortality measures as in Dong et al. (2020). In addition, the new Li-Lee multiplicative model offers the best overall performance, and the residuals are modeled, obtaining interpretable results and improved predictions, unlike other works that only use the models to fit and predict the mortality rates for countries such as Li and Lee (2005), Debón et al. (2011), or Russolillo et al. (2011). Furthermore, the higher MSE values at extreme ages suggest challenges in capturing mortality trends among older age groups, highlighting areas for potential model refinement.

Although the conclusions about comparing the models are based on the three countries that we have chosen, we propose statistical tools which provide a clear framework for supporting decisions in geographically disaggregated information about mortality trends. However, this study has limitations because the results from the analysis of one dataset cannot lead to general conclusions. Future research should explore additional machine learning techniques and alternative statistical approaches to further refine mortality models, particularly in addressing age-specific prediction challenges and enhancing forecasting robustness in multipopulation settings. It would be useful to examine other datasets and research whether the conclusions are consistent for different studies.

A promising extension of our framework is to incorporate exogenous covariates into the RF residual-modelling stage, thereby capturing abrupt yet transitory shocks–e.g., a medical breakthrough that converts a lethal disease into a chronic condition. Embedding such context-specific information directly in the RF component, rather than solely in the deterministic trend, could help disentangle structural shifts from short-lived anomalies and, in turn, sharpen forecast accuracy and interpretability. We leave a systematic exploration of this extension for future work.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work was partially supported by grants PID2023-152106OB-I00 and CIAICO/2023/272, funded by MCIN/AEI/10.13039/501100011033 and the Generalitat Valenciana, respectively.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Atance, D., Balbás, A., & Navarro, E. (2020). Constructing dynamic life tables with a single-factor model. *Decisions in Economics and Finance, 43*(2), 787–825.

Atance, D., Claramunt, M. M., Varea, X., & Aburto, J. M. (2024). Convergence and divergence in mortality: A global study from 1990 to 2030. *PloS one, 19*(1), 0295842.

Atance, D., & Debón, A. (2025). Cvmortalitymult: Cross-validation for multi-population mortality models. The R journal **forthcoming**. R package version 1.0.9

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

Atance, D., Debón, A., & Navarro, E. (2020). A comparison of forecasting mortality models using resampling methods. *Mathematics, 8*(9), 1550.

Alonso-García, J. (2023). AAS thematic issue:"mortality: from Lee-Carter to AI". *Annals of Actuarial Science, 17*(1), 212–214.

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

Alexander, M., Zagheni, E., & Barbieri, M. (2017). A flexible bayesian model for estimating subnational mortality. *Demography, 54*(6), 2025–2041.

Bergeron-Boucher, M.-P., Simonacci, V., Oeppen, J., & Gallo, M. (2018). Coherent modeling and forecasting of mortality patterns for subpopulations using multiway analysis of compositions: An application to canadian provinces and territories. *North American Actuarial Journal, 22*(1), 92–118.

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

Barigou, K., Goffard, P.-O., Loisel, S., & Salhi, Y. (2023). Bayesian model averaging for mortality forecasting using leave-future-out validation. *International Journal of Forecasting, 39*(2), 674–690.

Booth, H., Hyndman, R. J., Tickle, L., & Jong, P. (2006). Lee-Carter mortality forecasting: A multi-country comparison of variants and extensions. *Demographic Research, 15*(9), 289–310.

Bjerre, D. S. (2022). Tree-based machine learning methods for modeling and forecasting mortality. *ASTIN Bulletin: The Journal of the IAA, 52*(3), 765–787.

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Bégin, J. F., Sanders, B., & Xu, X. (2023). Modeling and forecasting subnational mortality in the presence of aggregated data. *North American Actuarial Journal*. https://doi.org/10.1080/10920277.2023.2231996

Booth, H., & Tickle, L. (2003). The future aged: New projections of Australia's ederly population. *Population Studies, 22*(4), 38–44.

Cairns, A. J. G., Blake, D., & Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk & Insurance, 73*(4), 687–718. https://doi.org/10.1111/j.1539-6975.2006.00195.x

Cossette, H., Delwarde, A., Denuit, M., Guillot, F., & Marceau. E. (2007). Pension plan valuation and mortality projection: A case study with mortality data. *North American Actuarial Journal, 11*(2), 1–34.

Cairns, A. J., Kallestrup-Lamb, M., Rosenskjold, C., Blake, D., & Dowd, K. (2019). Modelling socio-economic differences in mortality using a new affluence index. *ASTIN Bulletin: The Journal of the IAA, 49*(3), 555–590.

Carter, L.R., & Prkawetz, A. (2001). Examining structural shifs in mortality using the Lee-Carter method. Mpidr wp 2001-007, Center for Demography and Ecology Information, University of Wisconsin-Madison

Currie, I. D. (2016). On fitting generalized linear and non-linear models of mortality. *Scandinavian Actuarial Journal, 2016*(4), 356–383.

Debón, A., Chaves, L., Haberman, S., & Villa, F. (2017). Characterization of between-group inequality of longevity in European Union countries. *Insurance Mathematics and Economics, 75*, 151–165. https://doi.org/10.1016/j.insmatheco.2017.05.005

Diaz, G., Debón, A., & Giner-Bosch, V. (2018). Mortality forecasting in Colombia from abridged life tables by sex. *Genus, 74*, 1–23.

Delwarde, A., Denuit, M., Guillén, M., & 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.

Danesi, I.L., Haberman, S., & Millossovich, P. (2015). Forecasting mortality in subpopulations using Lee-Carter type models.: *A comparison. Insurance: Mathematics and Economics, 62*, 151–161.

Dong, Y., Huang, F., Yu, H., & Haberman, S. (2020). Multi-population mortality forecasting using tensor decomposition. *Scandinavian Actuarial Journal, 2020*(8), 754–775.

De Mori, L., Haberman, S., Millossovich, P., & Zhu, R. (2025). Mortality forecasting via multi-task neural networks. ASTIN Bulletin, 1–19 https://doi.org/10.1017/asb.2025.10

Debón, A., Montes, F., & 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*, 291–308.

Debón, A., Montes, F., & Puig, F. (2008). Modelling and forecasting mortality in Spain. *European Journal of Operation Research, 189*(3), 624–637.

Debón, A., Martínez-Ruiz, F., & 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., Martínez-Ruiz, F., & 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.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science, 11*(2), 89–121.

Euthum, M., Scherer, M., & Ungolo, F. (2024). A neural network approach for the mortality analysis of multiple populations: A case study on data of the Italian population. *European Actuarial Journal, 14*(2), 495–524.

Greenwell, B. M. (2017). pdp: An R package for constructing partial dependence plots. *The R Journal, 9*(1), 421–436.

Grigoriev, P., Shkolnikov, V., Andreev, E., Jasilionis, D., Meslé, F., & Vallin, J. (2010). Mortality in Belarus, Lithuania, and Russia: Divergence in recent trends and possible explanations. *European Journal of Population, 26*, 245–274.

Garrido, J., Shang, Y., & Xu, R. (2024). LSTM-based coherent mortality forecasting for developing countries. *Risks, 12*(2), 27.

Guillen, M., & Vidiella-i-Anguera, A. (2005). Forecasting Spanish natural life expectancy. *Risk Analysis, 25*(5), 1161–1170.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2024). forecast: Forecasting Functions for Time Series and Linear Models. R package version 8.22.0. https://pkg.robjhyndman.com/forecast/

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software, 27*(3), 1–22.

Haberman, S., & Renshaw, A. (2008). On simulation-based approaches to risk measurement in mortality with specific reference to Binomial Lee-Carter modelling. In: Society of Actuaries Living to 100 Symposium.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

Human Mortality Database: University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). http://www.mortality.org/. Accessed: April 8, 2024 (2024).

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(5), 1–26.

Kessy, S. R., Sherris, M., Villegas, A. M., & Ziveyi, J. (2022). Mortality forecasting using stacked regression ensembles. *Scandinavian Actuarial Journal, 2022*(7), 591–626.

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

Li, N., & Lee, R. (2005). Coherent mortality forecast for a group of populations: An extension of the Lee-Carter method. *Demography, 42*(3), 575–593.

Léger, A. E., & Mazzuco, S. (2021). What can we learn from the functional clustering of mortality data? an application to the human mortality database. *European Journal of Population, 37*, 769–798.

Lee, R.D., & Nault, F. (1993). Modeling and Forecasting Provincial Mortality in Canada. In: World Congress of the IUSSP, Montreal, Canada, Montreal. World Congress of the International Union for Scientific Study of Population

Levantesi, S., & Nigri, A. (2020). A random forest algorithm to improve the Lee-Carter mortality forecasting: Impact on q-forward. *Soft Computing, 24*(12), 8553–8567.

Lindholm, M., & Palmborg, L. (2022). Efficient use of data for LSTM mortality forecasting. *European Actuarial Journal, 12*(2), 749–778.

Lee, R. D., & Rofman, R. (1994). Modelación y proyección de la mortalidad en Chile. *Notas Población, 22*(59), 182–213.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News, 2*(3), 18–22.

Neves, C., Fernandes, C., & Hoeltgebaum, H. (2017). Five different distributions for the Lee-Carter model of mortality forecasting: A comparison using GAS models. *Insurance Mathematics and Economics, 75*, 48–57.

Nigri, A., Levantesi, S., Marino, M., Scognamiglio, S., & Perla, F. (2019). A deep learning integrated Lee-Carter model. *Risks, 7*(1), 33.

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

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2024). R Foundation for Statistical Computing. https://www.R-project.org/

Raftery, A. E., Chunn, J. L., Gerland, P., & Ševčíková, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography, 50*(3), 777–801.

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

Renshaw, A., & Haberman, S. (2003). Lee-Carter mortality forecasting: A parallel generalized linear modelling aproach for England and Wales mortality projections. *Journal of the Royal Statistical Society C, 52*(1), 119–137.

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

Richman, R. (2021). AI in actuarial science-a review of recent advances-part 1. *Annals of Actuarial Science, 15*(2), 207–229.

Richman, R., & Wuthrich, M.V.: Lee and Carter go machine learning: recurrent neural networks. Available at SSRN 3441030 (2019)

Santolino, M.: Should selection of the optimum stochastic mortality model be based on the original or the logarithmic scale of the mortality rate? Risks **11**(10) (2023) https://doi.org/10.3390/risks11100170

Therneau, T., & Atkinson, B. (2023). rpart: Recursive Partitioning and Regression Trees. R package version 4.1.23. https://CRAN.R-project.org/package=rpart

Turner, H., & Firth, D. (2023). gnm: Generalized Nonlinear Models. R package version 1.1-5. https://CRAN.R-project.org/package=gnm

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.

Wen, J., Cairns, A. J., & Kleinow, T. (2021). Fitting multi-population mortality models to socio-economic groups. *Annals of Actuarial Science, 15*(1), 144–172.

Wilmoth, J.R. (1996). Mortality projections for Japan: A comparison of four methods. Health and mortality among elderly populations, 266–287

Authors and Affiliations

Ana Debón[1] · Steven Haberman[2] · Gabriella Piscopo[3]

✉ Ana Debón
andeau@upv.edu.es

Steven Haberman
s.haberman@city.ac.uk

Gabriella Piscopo
gabriella.piscopo@unina.it

[1] Centro de Gestión de la Calidad y del Cambio, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Valencia, Spain

[2] Bayes Business School, City St George's, University of London, 106 Bunhill Row, London EC1Y 8TZ, United Kingdom

[3] Department of Economic and Statistical Science, University of Naples Federico II, Via Cintia 21, 80138 Naples, Italy