A random forest algorithm to improve the Lee–Carter mortality forecasting: impact on q-forward

Susanna Levantesi & Andrea Nigri
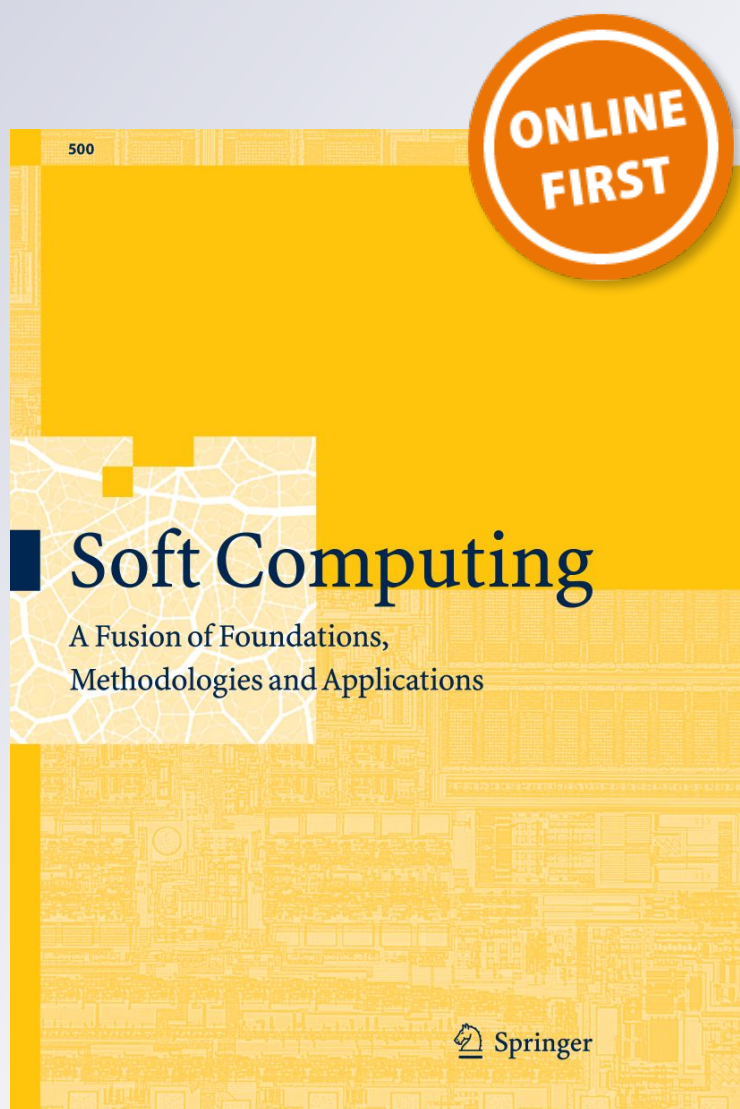
500

Soft Computing

A Fusion of Foundations, Methodologies and Applications

ONLINE FIRST

✑ Springer

🐎 Springer

Springer

FOCUS

A random forest algorithm to improve the Lee–Carter mortality forecasting: impact on q-forward

Susanna Levantesi[1] · **Andrea Nigri[1]**

Abstract

Increased life expectancy in developed countries has led researchers to pay more attention to mortality projection to anticipate changes in mortality rates. Following the scheme proposed in Deprez et al. (Eur Actuar J 7(2):337–352, 2017) and extended by Levantesi and Pizzorusso (Risks 7(1):26, 2019), we propose a novel approach based on the combination of random forest and two-dimensional P-spline, allowing for accurate mortality forecasting. This approach firstly provides a diagnosis of the limits of the Lee–Carter mortality model through the application of the random forest estimator to the ratio between the observed deaths and their estimated values given by a certain model, while the two-dimensional P-spline are used to smooth and project the random forest estimator in the forecasting phase. Further considerations are devoted to assessing the demographic consistency of the results. The model accuracy is evaluated by an out-of-sample test. Finally, we analyze the impact of our model on the pricing of q-forward contracts. All the analyses have been carried out on several countries by using data from the Human Mortality Database and considering the Lee–Carter model.

Keywords Mortality · Machine learning · Two-dimensional P-spline · q-Forward

1 Introduction

The downward trends in mortality rates experienced in the last century by developed countries engender economic and financial challenge for annuity providers at both public and private level. To manage longevity risk, annuity providers use projected life tables in order to include future mortality trends in the present value of liabilities. The accuracy of mortality forecasting is then considered an important issue for longevity risk management.

Recently, some authors applied machine learning techniques to improve the fitting accuracy of mortality models. Deprez et al. (2017) use a regression tree boosting machine to enhance the quality of mortality estimates provided by two stochastic mortality models, namely Lee–Carter model and Renshaw–Haberman model. This approach allows to identify strengths and weaknesses of each model respect to the individual features (age, year of birth, gender,…).

Levantesi and Pizzorusso (2019) extend this approach to improve the projected mortality rates provided by the Lee–Carter model using machine learning techniques that better capture not identifiable patterns. They apply three different supervised learning methods, decision tree, random forest and gradient boosting to predict the ratio between observed and estimated deaths (according to a specified model). The machine learning estimator of this ratio is then forecasted using the Lee–Carter formulation, moving away from the machine learning framework.

Other authors use deep learning to detect the future mortality trends. Hainaut (2018) applies neural network algorithms to identify the latent factors of mortality that are then forecasted by a random walk with drift. Richman and Wüthrich (2018) use neural networks to estimate a multi-population version of the Lee–Carter model. Nigri et al. (2019) use recurrent neural network with long short-term memory to forecast the future evolution of the time index of the Lee–Carter model, providing more accurate forecasting than the traditional ARIMA processes. Finally, Piscopo (2017, 2018a, b) has implemented an integrated dynamic evolving neuro-fuzzy inference system for mortality forecasting by combining neural networks and fuzzy inference.

Communicated by Philippe de Peretti.

✉ Susanna Levantesi
  susanna.levantesi@uniroma1.it

[1] Department of Statistics, Sapienza University of Rome, Viale Regina Elena, 295/G, 00161 Rome, Italy

As suggested by Levantesi and Pizzorusso (2019), the use of a machine learning estimator of the ratio between observed and estimated deaths allows to significantly improve the mortality estimation provided by a mortality model, even more in the forecasting. Following this line of research, we propose a novel approach for the machine learning estimator forecasting, exploiting the two-dimensional P-splines to smooth and extrapolate their future values on the mortality surface.

We develop a case study applying the random forest algorithm on mortality data of a set of developed countries. Moreover, we analyze the impact of the mortality rates obtained by both the traditional Lee–Carter model and its improved version obtained by the combination of a random forest algorithm and two-dimensional P-splines on the pricing of q-forward contracts. Compared to other longevity-linked securities traded on the market, the q-forward contract has some important advantages: it is built on standardized mortality indices reflecting the experience of a large population, is easier to price, less expensive and more liquid.

The paper is organized as follows. Section 2 provides a description of the model, including a brief introduction of the Lee–Carter model, an illustration of the machine learning estimator and P-splines. Section 3 shows a numerical application based on the population of six countries worldwide. An out-of-sample test is also provided to check the model ability to predict mortality. Conclusions follow.

2 The model

The generally available mortality data contains the following features, identifying an individual: age (x), calendar year (t), cohort (c) and gender (g). We define a feature space $\mathcal{S} = \mathcal{X} \times \mathcal{T} \times \mathcal{C} \times \mathcal{G}$ where: $\mathcal{X} = \{x_1, \ldots, x_m\}$, $\mathcal{T} = \{t_1, \ldots, t_n\}$, $\mathcal{C} = \{c_1, \ldots, c_k\}$ and $\mathcal{G} = \{$males, females$\}$. The number of mortality observations is then obtained by multiplying the range of the three variables x, t and g. We assign to each individual the feature $\mathbf{s} = (x, t, c, g) \in \mathcal{S}$ that can be used as input of a regression model as well as for machine learning and deep learning regression algorithms.

We consider demographic data available in the Human Mortality Database (HMD) that provides the number of deaths, $D_{x,t}$ and the exposure to risk, $E_{x,t}$ for individuals age x in the calendar year t. In the proposed framework, the data are organized so that the features of each individual x, t, c and g are recorded in a single row. Therefore, deaths and the exposure to risk are defined by $D_{\mathbf{s}}$ and $E_{\mathbf{s}}$, respectively.

The number of deaths, $D_{\mathbf{s}}$, is assumed independent in $\{\mathbf{s} \in \mathcal{S}\}$ and Poisson-distributed: $D_{\mathbf{s}} \sim \text{Pois}(m_{\mathbf{s}} \cdot E_{\mathbf{s}})$ for all $\{\mathbf{s} \in \mathcal{S}\}$, where $m_{\mathbf{s}} = \frac{D_{\mathbf{s}}}{E_{\mathbf{s}}}$ is the central death rate.

2.1 Lee–Carter model

We present the Lee–Carter (LC) model in the Poisson framework as proposed by Brouhns et al. (2002) as a function of the feature $\mathbf{s}$:

$$\log(m_{\mathbf{s}}) = \alpha_x + \beta_x \kappa_t, \quad \forall \mathbf{s} = (x, t, c, g) \in \mathcal{S} \tag{1}$$

where α_x is the age-specific parameter providing the average age profile of mortality; $\beta_x \cdot \kappa_t$ is the age-period term describing the mortality trends. (κ_t is the time index and β_x modifies the effect of κ_t across ages.) The following constraints on κ_t and β_x avoid identifiability problems with the parameters:

$$\sum_{t \in \mathcal{T}} \kappa_t = 0 \quad \sum_{x \in \mathcal{X}} \beta_x = 1$$

Mortality forecasting is obtained by modeling the time index κ_t by an autoregressive integrated moving average (ARIMA) process. In general, a random walk with drift properly fits the data:

$$\kappa_t = \kappa_{t-1} + \delta + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_k^2) \tag{2}$$

where δ is the drift parameter and ϵ_t are the error terms, normally distributed with null mean and variance σ_k^2.

2.2 Machine learning estimator

We denote $\hat{D}_{\mathbf{s}}$ the expected number of deaths estimated by a given mortality model and $\hat{m}_{\mathbf{s}}$ the corresponding central death rate. We define the ratio between the death observations and the estimated values as:

$$\psi_{\mathbf{s}} = \frac{D_{\mathbf{s}}}{\hat{D}_{\mathbf{s}}} \tag{3}$$

We estimate $\psi_{\mathbf{s}}$ by a regression algorithm depending on age, calendar years, year of birth (cohort) and gender:

$$\psi_{\mathbf{s}} \sim \text{gender} + \text{age} + \text{year} + \text{cohort} \tag{4}$$

We apply a regression algorithm to the ratio between death observations and estimated valued ($\psi_{\mathbf{s}}$) in order to improve the forecasting provided by standard stochastic mortality models, catching unidentifiable outlines. More in detail, our aim is to use a machine learning algorithm to complement the Lee–Carter mortality model and make the model's prediction more accurate. The application of the algorithm directly to the mortality rate is beyond the scope of this paper that do not consider machine learning as a substitute of mortality models, but benefits from artificial intelligence to detect

Fig. 1 Predictor space division .
(source: James et al. (2017))
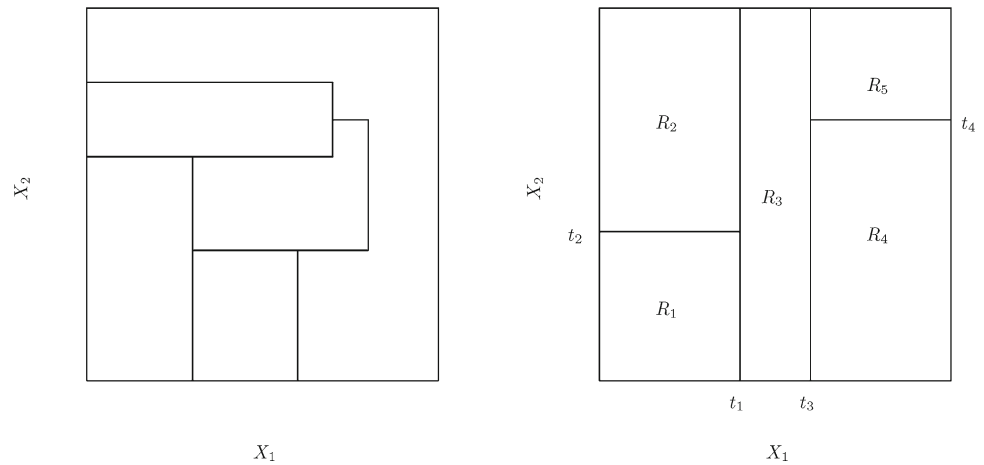
the mortality hidden features. Using only machine learning to model mortality can make results difficult to understand because the underlying decision process of a machine learning algorithm is often considered as unclear. However, the analysis of the model's features importance can offer more intuition into what the algorithm is learning.

We propose to solve Eq. 4 using the random forest algorithm. It belongs to the family of ensemble methods that are useful to get the error reduction pulling down the prediction variance, preserving the bias. The regression tree architecture is the basis behind the concept of random forest. Let us consider the need to approximate a function $y = f(x)$ (in our case, the function which describes $\psi_{\mathbf{s}}$). The regression trees allow to get the best function approximation (Loh 2011) by a procedure consisting in the following steps:

– We divide the predictor space that is, the set of possible values for $X_1, X_2, \ldots, X_p$, into J distinct and non-overlapping regions, $R_1, R_2, \ldots, R_J$ (Fig. 1).
– For each observation that falls into the region R_j, we make the same prediction, which is simply the mean of the response values for the training observations in R_j.

The basic idea is to divide the predictor space into high-dimensional rectangles, finding the boxes $R_1, \ldots, R_J$ that minimize the residual sum of squares (RSS): $\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$.

Once the regions $R_1, \ldots, R_J$ have been created, we predict the response for a given test observation using the mean of the training observations in the region to which that test observation belongs. It is of course computationally infeasible to consider every possible partition of the feature; therefore, we consider a "top-down" approach (Quinlan 1986) by using a recursive binary splitting. Considering a generic (upside down) tree structure, the algorithm (Breiman et al. 1984) starts at the top of the tree, where all observations belong to a single region, and then successively splits the pre-

dictor space (see Fig. 2). Thus, starting from the root node, which generates the subsequent nodes up to the final ones (or leaves) according the split rule (Morgan and Sonquist 1963; Breiman et al. 1984):

– If $X_1 < t$, then the unit belongs to the left branch, where t is the cut point such that the space predictor splitting leads to the greatest possible reduction in RSS.
– Otherwise, it belongs to the right branch.

To identify the best split, consider a purity measurement (homogeneity) for each node quantified by the entropy or the index of Gini. The maximum purity is reached when only one class of Y is present in the node.

The regression tree-based methods have some interesting properties (see Breiman (2001) for more details). For example:

– They are nonparametric and able to model complex relations between inputs and outputs, without any (probabilistic) a priori assumption;
– They handle heterogeneous data (ordered or categorical variables, or a mix of both);
– They intrinsically implement feature selection, making them robust to not significant or noisy variables;
– They are robust to outliers or missing values;
– They are easily interpretable.

Random forests basically consist in building an ensemble of decision trees grown from a randomized variant of the tree. Starting from a single learning set, the basic idea is to introduce a random perturbation into the learning procedure in order to introduce a differentiation among the trees and combine the predictions of all these trees using aggregation techniques. Breiman (1996) proposed a first aggregation method the so-called bagging in which the different trees are built by using random bootstrap copies of the original data.

Fig. 2 Top-down approach: splitting rule . (source: James et al. (2017))
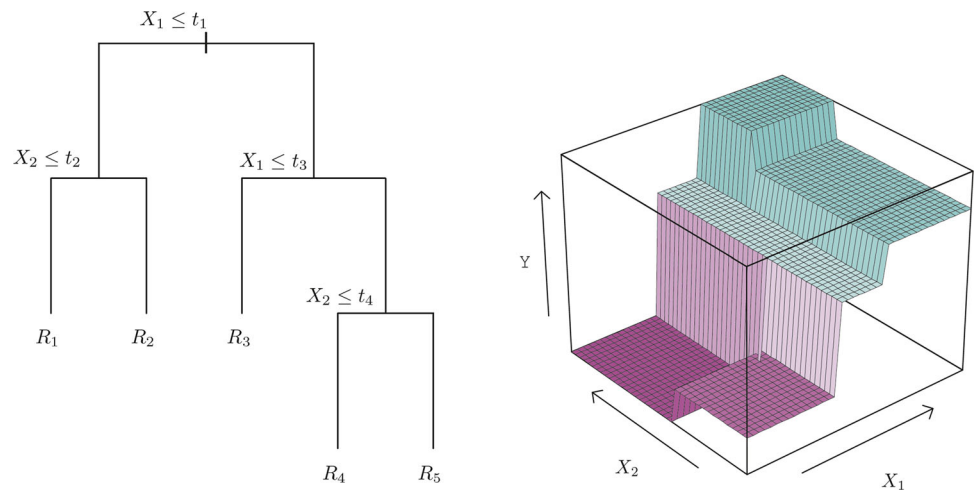
Its natural evolution, the random forests have been developed by the same author in 2001 Breiman (2001). In the random forests, the bagging approach has been extended and combined with randomization of the input variables that are used when considering candidate variables to split internal nodes t. In particular, instead of looking for the best split s^* among all variables, the algorithm chooses a random subset of K variables for each node and then determines the best split using these variables.

It is also important to consider the so-called variable importance. To this purpose, Breiman (2001) proposed a weighted impurity measure for evaluating the importance of a variable X_m in predicting Y, for all nodes t averaged over all N_T trees in the forest. If impurity $i(t)$ is the Gini index, the impurity measure is called the Gini importance or mean decrease Gini (here denoted by $IncNodePurity$):

$$IncNodePurity(X_m) = \frac{1}{N_T} \sum_{T} \sum_{t \in T : v(s_t) = X_m} p(t) \Delta i(s_t, t) \tag{5}$$

where $v(s_t)$ is the variable used in split s_t and $\Delta i(s_t, t)$ is the impurity decrease in a binary split s_t dividing node t into a left node t_l and a right node t_r:

$$\Delta i(s_t, t) = i(t) - \frac{N_{t_l}}{N_t} \cdot i(t_l) - \frac{N_{t_r}}{N_t} \cdot i(t_r) \tag{6}$$

where $p(t) = \frac{N_t}{N}$ is the proportion of samples reaching t, and $p(t_l) = \frac{N_{t_l}}{N}$ and $p(t_r) = \frac{N_{t_r}}{N}$ are the proportion of samples reaching the left node t_l and the right node t_r, respectively.

The mean decrease Gini in Eq. 5 calculates the importance of each variable X_m as the sum over the number of splits that include the variable, proportionally to the number of samples it splits.

We denote $\hat{\psi}_s$ the random forest estimator, obtained by applying the random forest algorithm from the R package randomForest (Liaw 2018). Since this procedure proved to be very costly from a computational point of view, the number of trees must be carefully chosen; it should not be too large but at the same time able to produce an adequate percentage of variance explained and a low mean of squared residuals, MSR.

The goodness of fit of the Lee–Carter model can be improved by the random forest estimator, $\hat{\psi}_s$, that is applied to the central death rate estimated by the Lee–Carter model, $\hat{m}_s$. Therefore, the random forest improved rates are calculated by:

$$\hat{m}_s^\psi = \hat{\psi}_s \cdot \hat{m}_s \tag{7}$$

To obtain a better performance from the Lee–Carter model forecasting, we need to estimate future values of the random forest estimator. To this aim, we propose a methodology based on the smoothing and extrapolation of $\hat{\psi}_s$ using P-splines (in particular, the two-dimensional version). The use of the P-splines is legitimized by some of their important properties partially deriving from the B-spline (De Boor 1978). Compared to other smoothing methods (in particular, Kernel-based methods), P-splines do not show the spread of a curve or density outside the data domain, indeed, as proved by Eilers and Marx (1996), they show no boundary effects. The authors also underline how average and variance of the fitted data (for any level of smoothing) are equal to the average and the variance of the observed data. This point is very important when working on density smoothing. The latter is not guaranteed in other smoothing methods, such as for Kernel methods that compromise the variance estimate, which increases proportionally to smoothing. Last but not least, the possibility of working in a Poisson setting (on the log-likelihood) that is the base of the most used mortality models (Lee–Carter included). As stated by Currie et al. (2006) (p. 280): "an important feature of this method is that forecasting is a natural consequence of the smoothing process. We con-

sider future values as missing values; the penalization then allows the estimation of future values simultaneously with the fitting of the mortality surface."

The form of the forecast is then determined by the penalty function that is more important in the extrapolation of future vales respect to the smoothing of data. Therefore, the penalty function must be chosen with careful.

2.3 P-splines

We introduce the P-spline model starting from its base, the B-spline. We will follow the line described by Eilers and Marx. The P-spline estimation is performed in two main steps, in order:

1. Use B-spline as the basis for the regression (for an extensive introduction to B-spline see Eilers and Marx (1996)),
2. Modify the likelihood by a penalty on the regression coefficients.

Let $B_j(x, q)$ denote the value at x of the jth B-spline of degree q. For a given equidistant grid of knots, we can define $\hat{y}(x) = \sum_{j=1}^{n} \hat{a}_j B_j(x, q)$. In the present study, the rows of the B are the values of the B-splines in the basis evaluated at each year. Therefore, aB are weighted averages of a local subset of the coefficients. In many cases, the application of B-spline leads to an over-smoothed behavior of the fitted function. At this purpose, O'Sullivan (1986, 1988) proposed a penalty on the second derivative:

$$\sum_{i=1}^{m} \left\{ y_i - \sum_{j=1}^{n} a_j B_j(x_i) \right\}^2 + \lambda \int_{x_{\min}}^{x_{\max}} \left\{ \sum_{j=1}^{n} a_j B_j''(x_i) \right\}^2 dx. \tag{8}$$

Eilers and Marx improved the O'Sullivan' model, proposing a penalty term on higher-order finite differences of the coefficients of adjacent B-splines:

$$\sum_{i=1}^{m} \left\{ y_i - \sum_{j=1}^{n} a_j B_j(x_i) \right\}^2 + \lambda \sum_{j=k+1}^{n} (\Delta^k a_j)^2 \tag{9}$$

where λ is smoothing parameter (see e.g., D'Amato et al. (2011) for the use of P-splines on the Lee–Carter model). It worth to notice that the penalty from Eq. 9 allows to preserve the moments of data.

As already mentioned, P-splines also allow to work in a Poisson setting. In this regard, in order to obtain the penalized log-likelihood, the penalty can be incorporated into the log-likelihood function, $\ell(a, y)$, estimated by a given mortality model.

$$\ell_p = \ell(a, y) - \frac{1}{2} a' P a \tag{10}$$

$$\ell(a, y) - \frac{1}{2} \lambda a' D' D a \tag{11}$$

where D is a difference matrix of order d. $P = \lambda D'D$ is the penalty matrix and λ is the smoothing parameter. Maximizing ℓ_p gives the penalized likelihood $B'(y - \mu) = Pa$, which can be solved by the penalized version of the scoring algorithm:

$$(B' \tilde{W} B + P)\hat{a} = B' \tilde{W} B \tilde{a} + B'(y - \tilde{\mu}) \tag{12}$$

where B is the regression matrix, P is the penalty matrix, $\tilde{a}$, $\tilde{\mu}$ and $\tilde{W}$, the diagonal matrix of weights and $\hat{a}$ denotes the updated estimate of a. As shown in Eilers and Marx (1996), it is possible to use penalized spline in a Poisson setting by using a penalized version of the iteratively re-weighted least squares (IRWLs used for generalized linear model estimation).

$$(B' \tilde{W} B + P)\hat{a} = B' \tilde{W} \tilde{z}; \tag{13}$$

$$\tilde{z} = B \tilde{a} + \tilde{W}^{-1}(y - \tilde{\mu}); \tag{14}$$

where $\tilde{\mu} = \text{diag}(\mu)$.

The P-splines application leads to face some difficulties about the right number of knots, the degree of the P-spline, the penalty order and the smoothing parameter. The latter could be negligible in the fitting phase, but when we will approach the forecasting estimation we will have to take into account that the penalty order greatly influences the forecasting values. Ruppert (2002) and Currie and Durban (2002) list some sufficient advices in choosing the P-spline parameters; with equally spaced data, use one knot for every four or five observations up to a maximum of forty knots. Anyway, the authors proposed the BIC criterion (Schwarz 1978) to choose the smoothing parameter. As pointed out in the previous section, the final aim of the paper is obtaining a smoothed surface of ψ_s estimation over time that is achieved through the two-dimensional P-spline approach. Since the Eilers and Marx improvement, the P-spline has been widely used, embracing different scientific fields. One of its major uses is certainly in the area of mortality in which there was a need to smooth the roughness of the raw data. In the analysis of mortality, we usually deal with two-dimensional surfaces so that some authors (Eilers and Marx 2002; Eilers et al. 2006; Eilers and Marx 2010; Currie et al. 2004, 2006; Camarda 2012) proposed a P-spline evolution: the two-dimensional P-splines. The first steps to approach the two-dimensional p-spline is to start the construction of the two-dimensional basis, analogously to the one-dimensional case. For the purposes of regression, we assume that the data are arranged as a column vector, $y = \text{vec}(Y)$. Let $B_a = B(x_a)$, $n_a \times c_a$, be a regression matrix of B-splines based on the explanatory variable

Table 1 RSS and MSR by the random forest algorithm for LC model. Ages 20–100 and years 1947–2014

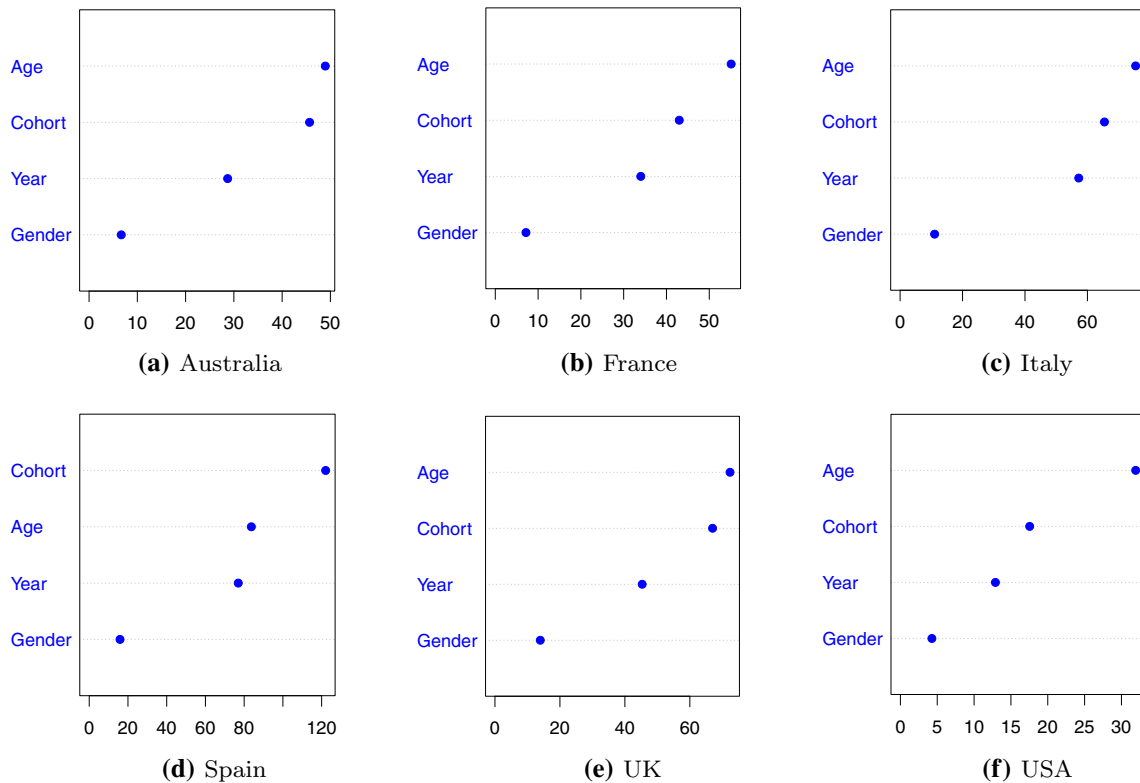| Indicator	Australia	France	Italy	Spain	UK	USA
RSS (%)	56.91	87.18	87.43	89.11	89.58	93.76
MSR	0.0057	0.0017	0.0025	0.0030	0.0019	0.0004

Fig. 3 Variable importance. *%IncNodePurity*. Ages 20–100 and years 1947–2014

for age x_a and $By = B(x_y)$, $n_a y \times c_y$, be a regression matrix of B-splines based on the explanatory variable for year x_y. The regression matrix for our two-dimensional model is the Kronecker product:

$$B = B_y \otimes B_a \tag{15}$$

From the Kronecker product definition, and by using IRWLS, we can independently penalize the age and year coefficients over the rows and columns of A that is a $c_a \times c_y$ matrix containing the arranged elements of $a = \text{vec}(A)$. Let D_a and D_y the difference matrices acting on the columns and rows of A, respectively. The penalty is:

$$P = \lambda_a I_{c_y} \otimes D'_a D_a + \lambda_y D'_y D_y \otimes I_{c_a} \tag{16}$$

where λ_a, λ_y are the smoothing parameter used for age and year, respectively. I_{c_a}, I_{c_y} are identity matrices of dimension c_a, c_y. The latter point allows to choose different smoothing parameters over age and years independently, giving an higher flexibility model.

3 Numerical analysis

The model presented in Sect. 2 is applied to the population of six countries worldwide: Australia, France, Italy, Spain, UK and the USA. Data are downloaded from the Human Mortality Database (www.mortality.org), while model fitting is performed with *StMoMo* package provided by Villegas et al. (2015). We analyze the following set of explanatory variables:

$$\mathcal{X} = \{20, \ldots, 100\}, \mathcal{T} = \{1947, \ldots, 2014\},$$
$$\mathcal{C} = \{1847, \ldots, 1994\}, \mathcal{G} = \{\text{males, females}\}$$

where $m_\mathbf{s}$ is the observed central death rate for the feature **s**, $\hat{m}_\mathbf{s}$ the estimated value and N the number of observations.

The first step of the model's application consists in the estimation of the Lee–Carter model parameters, separately for male and female population. The second step in modeling $\psi_\mathbf{s}$, the ratio between the death observations and the estimated values from the Lee–Carter model by the random forest algorithm, while the final step consists in smoothing

Table 2 MAPE. Ages 20–100 and years 1947–2014

Gender	Model	Australia (%)	France (%)	Italy (%)	Spain (%)	UK (%)	USA (%)
Males	LC	10.31	9.76	11.09	12.32	10.52	7.71
	LC ψ	5.13	2.69	3.19	3.65	2.86	1.42
Females	LC	9.31	9.36	10.29	13.29	10.27	6.53
	LC ψ	5.91	3.01	3.37	4.18	3.02	1.56

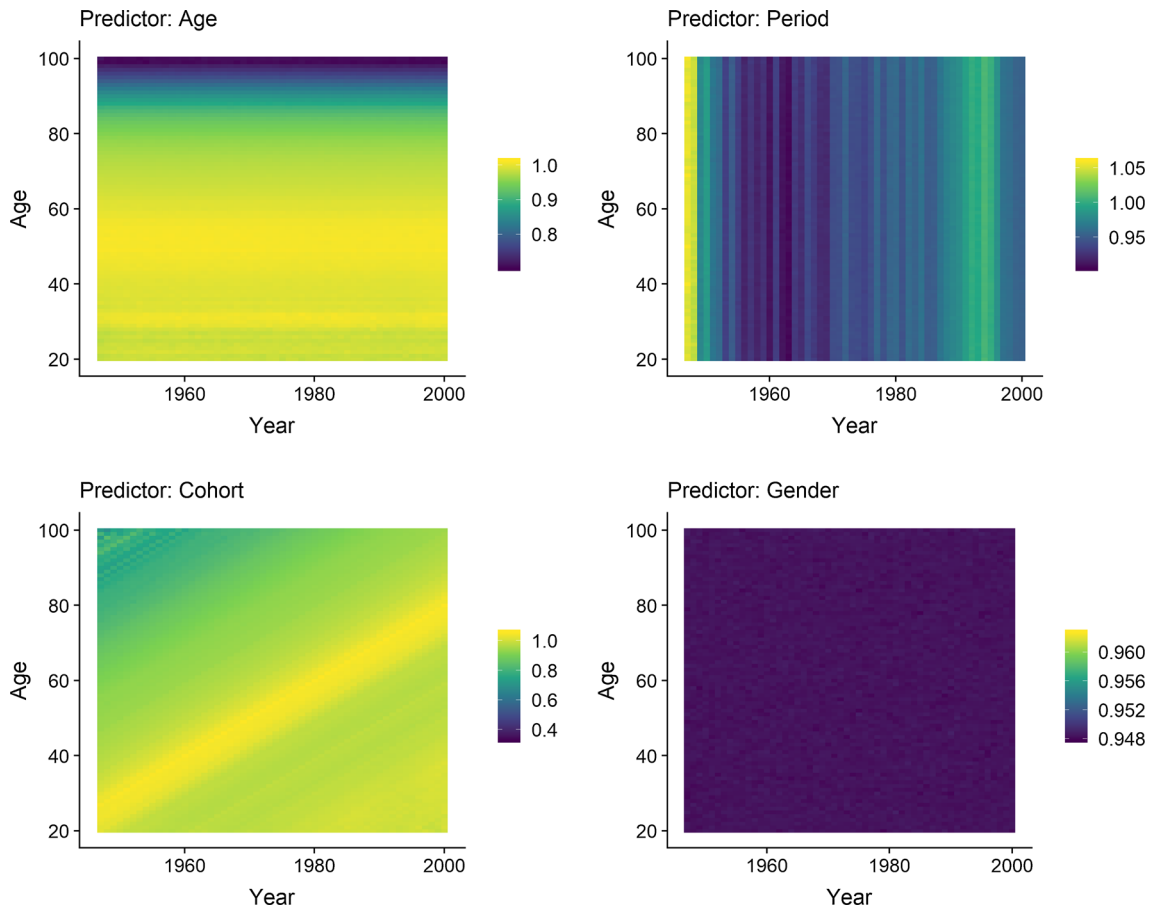**Fig. 4** $\hat{\psi}_{\mathbf{s}}$ sensitivity to predictor. Ages 20–100 and years 1947–2014. Italy

and extrapolating the machine learning estimator, and then applied to the Lee–Carter forecasted mortality rates.

3.1 Model fitting

We model $\psi_{\mathbf{s}}$ using the random forest algorithm implemented in the R package `randomForest`. We set the parameters of the algorithm as follows: to avoid model's over-fitting the number of trees ($ntrees$) has been set to 200 and the number of input variables to be used in each node ($mtry$) to 2. Table 1 shows the percentage of variance explained by the random forest algorithm, RSS, and the level of the mean of squared residuals MSR by each country. The explained variance is calculated on the central death rate as: $\sum_{\mathbf{s}}(\hat{m}_{\mathbf{s}}-m_{\mathbf{s}})^2$. Its level

is higher than 80% for all the countries, except for Australia (57%).

We show in Fig. 3 the values of the mean decrease Gini of attributes assigned by the random forest (sorted decreasingly from top to bottom). The lessen importance of the gender predictor in explaining $\hat{\psi}_s$ points out the ability of the Lee–Carter model to catch well the gender difference in mortality in the analyzed countries. Conversely, the higher importance of the age predictor leads to the conclusion that this feature is not well-detected by the Lee–Carter model, mostly in the old ages. In fact, if we consider ages 20–90, the age predictor is no longer the most important in regressing $\hat{\psi}_s$ (see Fig. 7 in "Appendix"). In this case, the cohort predictor rises to the top of the importance as the Lee–Carter model does not include a cohort effect.
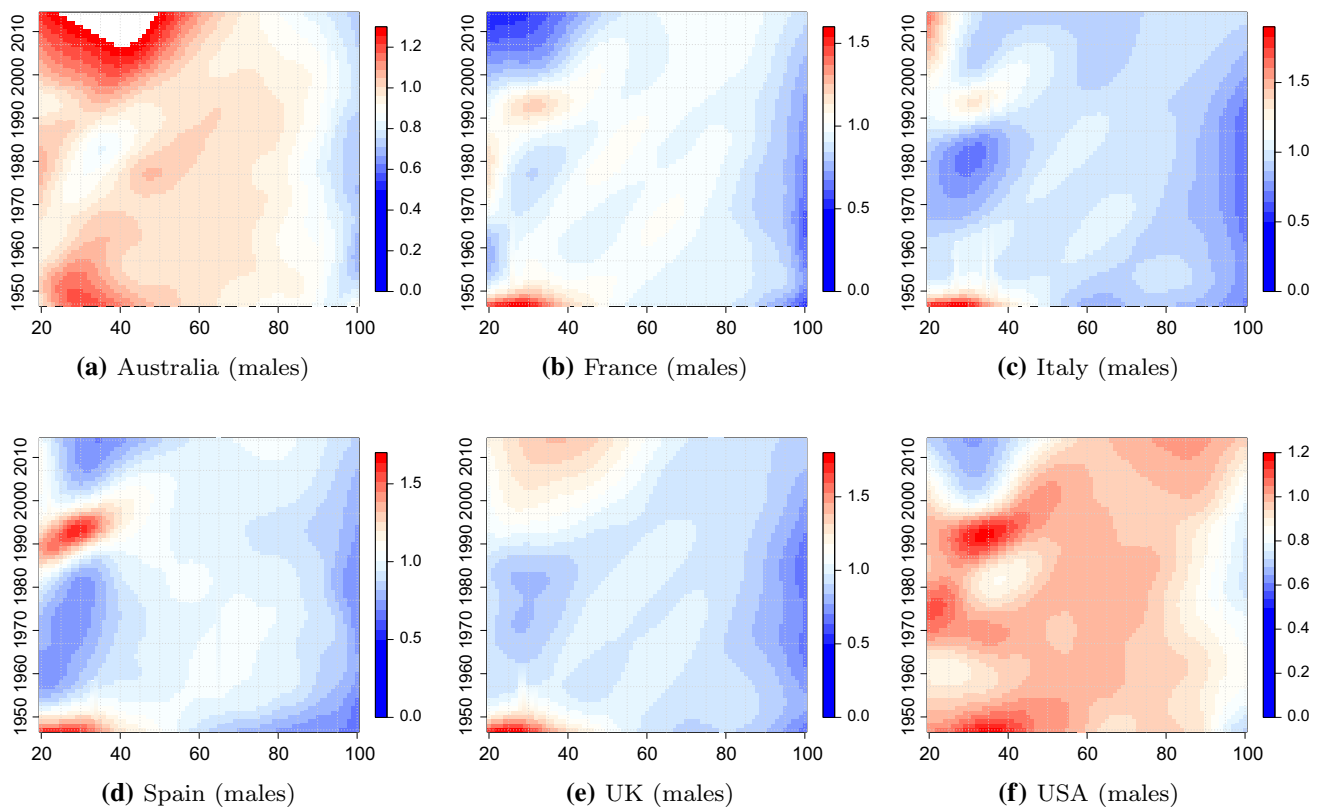
Fig. 5 $\hat{\psi}_\mathbf{s}$ smoothed values (1947–2000) and extrapolated values (2000–2014). Results from *MortalitySmooth* package. Male population. Age 20–100

The goodness of fit is measured by the mean absolute percent error (MAPE), defined as:

$$\text{MAPE} = \frac{100}{N} \sum_\mathbf{x} \left| \frac{m_\mathbf{s} - \hat{m}_\mathbf{s}}{m_\mathbf{s}} \right| \qquad (17)$$

The resulting MAPE is reported in Table 2 by country. In summary, the random forest algorithm improves the fitting of the Lee–Carter model for all the countries. The lower level of improvement is recorded in Australia, where the variance explained by the algorithm is the lowest.

3.2 Sensitivity to predictor

As shown in Sect. 2.2, the random forest algorithm provides information about the importance of the variables. Since our investigation falls into a population study, some demographic aspects need to be considered. First of all, we have to investigate if the algorithm predictors evaluation is reasonable in a demographic perspective. To this aim, we proceed with a predictors sensitivity analysis considering the case of Italy. According to the *IncNodePurity* measure in Fig. 3 (panel c), the most important variable (with a higher node purity) in Italy is the age. It is quite close to the second and the third one,

respectively cohort and calendar year. The gender predictor instead seems to be very far from the others, showing the lowest level of purity. This picture appears to be consistent with the age structure transformation and the remarkable cohort effect characterizing the Italian mortality dynamic over time. It could be confirmed (or discredited) by a sensitivity analysis of the predictors. We progressively add a single predictor in the regression model of $\psi_\mathbf{s}$, in order to check the contribution of each predictor to improve the mortality surface estimation given by the Lee–Carter model. Figure 4 shows the values of $\hat{\psi}_\mathbf{s}$ obtained by the random forest algorithm in a regression model based on a single predictor. The contribution of the age predictor is quite age homogeneous except for the very old ages (90–100). The period predictor shows high levels immediately after the Second World War that represents one of the most important historical shocks in mortality dynamic. In this case, the Lee–Carter model fails to capture this mortality peak. Similarly, the cohort predictor has high values for cohorts born during the Second World War. Lastly, the gender predictor shows similar values, confirming that gender is not important in predicting $\hat{\psi}_\mathbf{s}$. Therefore, we deduce that the gender difference in the Italian mortality is well-detected by the Lee–Carter model.
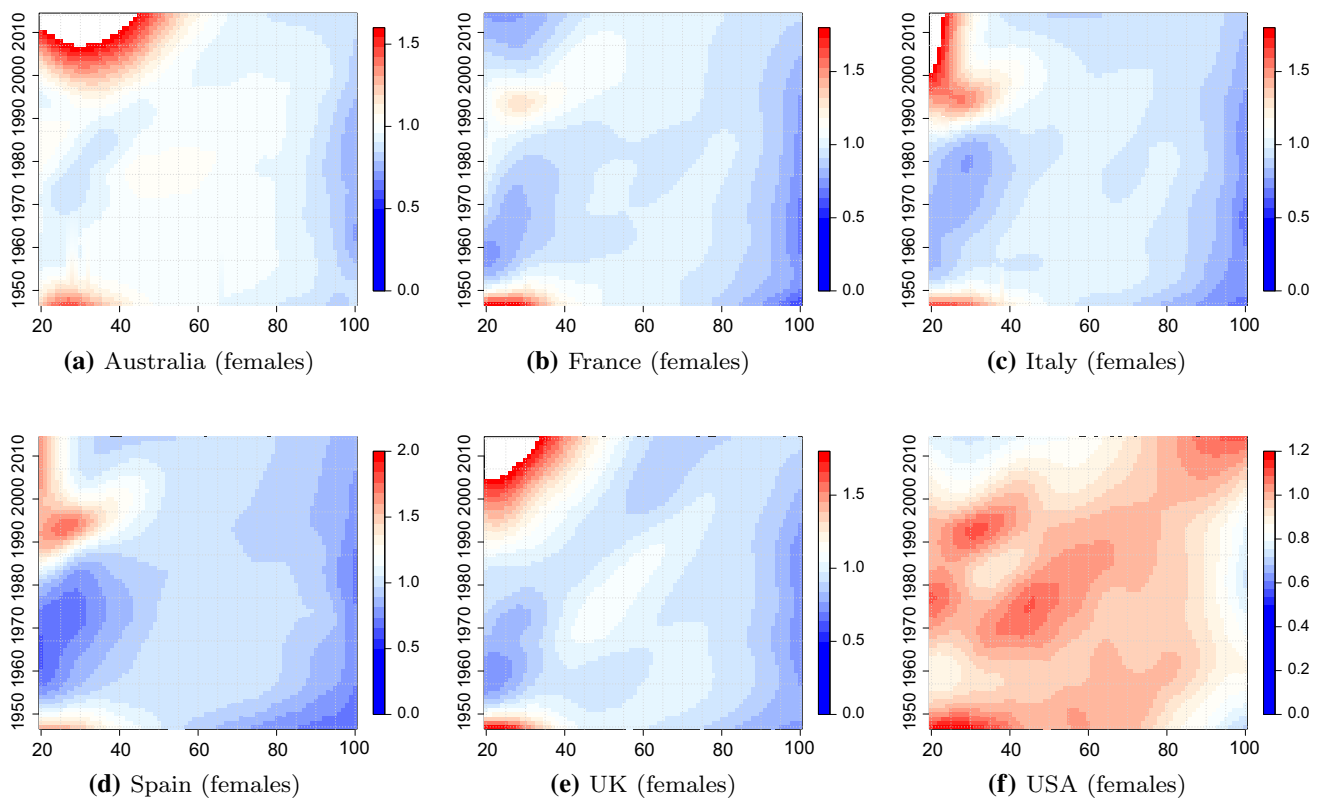
Fig. 6 $\hat{\psi}_{\mathbf{s}}$ smoothed values (1947–2000) and extrapolated values (2000–2014). Results from *MortalitySmooth* package. Female population. Age 20–100

Table 3 Out-of-sample test results: RMSE for the LC model without and with machine learning. Ages 20–100 and years 2001–2014

Gender	Model	Australia	France	Italy	Spain	UK	USA
Males	LC	0.0269	0.0330	0.0327	0.0381	0.0341	0.0358
	LC ψ	0.0161	0.0158	0.0212	0.0207	0.0145	0.0255
Females	LC	0.0123	0.0203	0.0203	0.0268	0.0192	0.0149
	LC ψ	0.0085	0.0057	0.0057	0.0109	0.0079	0.0179

3.3 P-splines smoothing and extrapolation

To improve the performance of the Lee–Carter model in mortality forecasting, we have to estimate the future values of $\hat{\psi}_{\mathbf{s}}$. Firstly, we smooth the values of $\hat{\psi}_{\mathbf{s}}$ on two dimensions, age and time, simultaneously using the two-dimensional smoothing algorithm based on P-splines proposed in Camarda (2012). This algorithm also allows to extrapolate future values. It is implemented in the R package *MortalitySmooth*.

We perform an out-of-sample test splitting the dataset in two parts: data from 1947 to 2000 constitutes the fitting period and data from 2001 to 2014 the forecasting period. The resulting values of $\hat{\psi}_{\mathbf{s}}$ are illustrated in Figs. 5 and 6 for males and females respectively. Results for all the countries (except USA) indicate that the Lee–Carter model mostly overestimates mortality in the ages 90–100 for both genders ($\hat{\psi}_{\mathbf{s}} < 1$), while it underestimates mortality especially in the ages 20–40 for both genders ($\hat{\psi}_{\mathbf{s}} > 1$).

To evaluate the goodness of the out-of-sample test results, we use the root-mean-squared error of $m_{\mathbf{x}}$ (RMSE), defined as:

$$\text{RMSE} = \sqrt{\frac{\sum_{\mathbf{s}} \left(m_{\mathbf{s}} - \hat{m}_{\mathbf{s}}\right)^2}{N}} \tag{18}$$

Table 3 shows the results of the out-of-sample test. The random forest algorithm produces a significant improvement in forecasting respect to the Lee–Carter model without machine learning.

3.4 Sensitivity to the age range

We develop a sensitivity analysis on the age range to study the level of the improvement provided by machine learning algorithms on a reduced dataset. In addition to the base case (age range 20–100), we also consider the following range: 40–100

Table 4 MAPE. Ages 40–100 and years 1947–2014

Gender	Model	Australia (%)	France (%)	Italy (%)	Spain (%)	UK (%)	USA (%)
Males	LC	10.14	9.26	9.43	9.63	10.06	7.96
	LC ψ	4.50	2.26	2.74	3.16	2.34	1.29
Females	LC	7.98	8.32	7.86	9.76	7.81	6.62
	LC ψ	4.39	3.01	2.41	3.26	2.06	1.31

Table 5 MAPE. Ages 60–100 and years 1947–2014

Gender	Model	Australia (%)	France (%)	Italy (%)	Spain (%)	UK (%)	USA (%)
Males	LC	11.21	11.40	11.13	11.13	11.52	9.22
	LC ψ	4.52	2.38	2.93	3.21	2.41	1.35
Females	LC	8.37	9.46	9.40	10.05	8.90	7.50
	LC ψ	3.82	1.95	2.15	2.88	1.73	1.26

Table 6 Out-of-sample test results: RMSE for the LC model without and with machine learning. Ages 40–100 and years 2001–2014

Gender	Model	Australia	France	Italy	Spain	UK	USA
Males	LC	0.0310	0.0385	0.0384	0.0429	0.0393	0.0410
	LC ψ	0.0197	0.0165	0.0228	0.0184	0.0173	0.0272
Females	LC	0.0140	0.0232	0.0234	0.0303	0.0217	0.0172
	LC ψ	0.0094	0.0091	0.0059	0.0105	0.0092	0.0207

Table 7 Out-of-sample test results: RMSE for the LC model without and with machine learning. Ages 60–100 and years 2001–2014

Gender	Model	Australia	France	Italy	Spain	UK	USA
Males	LC	0.0366	0.0470	0.0465	0.0516	0.0467	0.0506
	LC ψ	0.0220	0.0219	0.0289	0.0226	0.0239	0.0365
Males	LC	0.0166	0.0278	0.0284	0.0364	0.0260	0.0209
	LC ψ	0.0098	0.0119	0.0076	0.0124	0.0125	0.0232

Table 8 q-forward price (in thousands) for the LC model without and with machine learning. Ages 60, 70, 80. Maturity $T = 10$. Males

Age	Model	Australia	France	Italy	Spain	UK	USA
60	LC	6.06	10.65	8.63	8.46	8.51	10.66
	LC ψ	6.15	8.96	7.18	8.46	8.57	9.51
70	LC	20.18	22.48	23.90	20.74	26.51	27.66
	LC ψ	20.10	21.08	21.73	20.28	25.63	26.07
80	LC	62.13	58.39	62.96	60.54	75.11	74.05
	LC ψ	59.41	56.93	61.90	58.80	72.64	73.20

and 60–100. The aim is to check if the change of the calibration period could have an important impact on the results. The values of $\hat{\psi}_\mathbf{s}$ are shown in appendix (Fig. 8 (males)–Fig. 9 (females) for ages 40–100 and Fig. 10 (males)–Fig. 11 (females) for ages 60–100). The MAPE values providing a measure of the quality of fitting are shown in Table 4 (males) and Table 5 (females) for ages 40–100 and in Table 6 (males) and Table 7 (females) for ages 60–100. Also for a shorter age

range, the random forest algorithm improves the estimates provided by Lee–Carter model.

3.5 Pricing q-forward

This section focuses on studying the impact of the mortality rates obtained from our model, based on the combination of a random forest algorithm and two-dimensional P-splines, on the pricing of q-forward contracts. A comparison with the traditional Lee–Carter model is also provided.

A q-forward contract is a longevity-based instrument that can be interpreted as a zero-coupon swap. According to the definition provided by The Life & Longevity MarketsAssociation (2010), "A q-forward is an agreement between two counterparties to exchange at a future date (the maturity of the contract) an amount equal to the realized mortality rate of a given population at that future date (the floating leg), in return for a fixed mortality rate agreed upon at the inception of the contract (the fixed leg)".

Different pricing methods have been proposed in the literature. The most common are the risk-neutral approach and the Sharpe ratio approach (see Barrieu and Veraart (2016) for a review of some classical financial and actuarial formulae that can be applied in pricing q-forwards and see e.g., Loeys et al. (2007) for the application of the Sharpe ratio approach). Moreover, some authors propose a pricing methodology based on Solvency II, using the Cost of Capital approach (see Levantesi and Menzietti 2017 and Zeddouk and Devolder 2019).

Table 9 q-forward price (in thousands) for the LC model without and with machine learning. Ages 60, 70, 80. Maturity $T = 10$. Females

Age	Model	Australia	France	Italy	Spain	UK	USA
60	LC	3.93	3.56	3.62	2.45	6.58	7.79
	LC ψ	3.83	3.52	3.47	2.53	5.48	6.25
70	LC	10.92	8.13	9.59	7.08	16.44	18.22
	LC ψ	10.49	7.93	9.20	7.08	14.54	16.08
80	LC	35.47	30.18	34.36	32.86	45.07	48.25
	LC ψ	34.31	29.15	33.44	31.62	43.17	46.31

Let define $q_{\mathbf{s},T}^{E}$ as the expected mortality rate (the fixed leg of the contract), function of the feature $\mathbf{s}$ and the maturity date T. It is calculated from the corresponding central death rate $m_{\mathbf{s},T}^{E}$ according to the well-known equation $q_{\mathbf{s},T}^{E} = 1 - e^{-m_{\mathbf{s},T}^{E}}$.

We evaluate the q-forward contracts according to the Sharpe ratio approach as in Loeys et al. (2007). Therefore, the q-forward price is calculated as:

$$q_{\mathbf{s},T}^{F} = (1 - \text{SR} \cdot T \cdot \sigma_q)q_{\mathbf{s},T}^{E} \tag{19}$$

where T is the time to maturity, SR is the Sharpe ratio for the q-forward and $\sigma_q = \sqrt{\text{var}\left(\frac{\Delta q_{s,T}}{q_{s,T}}\right)}$ is the historical standard deviation of changes in the mortality rate. The value of SR is set to 0.1 according to the results obtained by Barrieu and Veraart (2016). We consider q-forwards with a ten-year maturity ($T = 10$) and three different underlying ages $x \in \{60, 70, 80\}$. Results are shown in Tables 8 and 9 for males and females, respectively. The q-forwards price level increases with age because the mortality models provide age increasing mortality rates. Our model affects the price of q-forwards, producing lower values than those obtained with the Lee–Carter model, with the exception of Australia and UK (age 60) for males and Spain (age 60) for females.

4 Conclusions

In recent decades, different approaches to modeling and forecasting mortality have been proposed in the literature. These models were intended to portray the mortality surface, and on the other hand, they aimed to anticipate future changes in mortality shape as accurately as possible. Since the first Lee–Carter formulation, several stochastic models were introduced, albeit most of them still show the same critical issues. Some of these drawbacks are well-known among scholars and they have been extensively treated. One of the most important concerns the fixed structure of the β_x index over time (Lee and Miller 2001; Girosi and King 2008).

In developed countries, mortality deceleration is lead by a decline at younger ages (infant and child mortality) as well as an accelerating at old ages. This phenomenon is called "rotation". To overcome this problem, Li et al. (2013) proposed a rotation of the β_x for long-term mortality projections.

Our analysis supports the basic finding that mortality rises at older ages, where the Lee–Carter model fails to provide an appropriate estimate showing overestimated mortality rates. Machine learning allows to take into consideration this important aspect, overcoming this limitation without adjusting the model parameters.

Following the idea from Deprez et al. (2017) about the use of machine learning models as a diagnostic tool, we therefore retrace the Levantesi and Pizzorusso (2019) approach aimed to forecast the machine learning estimator of the ratio between the observed deaths and the estimated ones from the Lee–Carter model. Following this line of investigation, we introduce a novel scheme based on the random forest algorithm and exploiting the two-dimensional P-spline. We estimate, smooth and forecast a machine learning estimator contributing to improve the quality of the Lee–Carter model estimates.

Using data of several countries from the HMD, this paper aims to demonstrate how machine learning algorithms can bring effective benefits to the study of mortality, both in the fitting and forecast phases. We have shown that an ensemble method can strongly reinforce a mortality model, preserving the "desirable features" that a good mortality model should have in accord with the criteria proposed by Cairns et al. (2008), basically concerning some common-sense guidelines, such as the consistency with historical data, the long-term dynamics biologically reasonable and the parsimony. On the other hand, it is important to point out that machine learning algorithms must be used with caution, being careful to keep the demographic interpretability of the phenomenon. However, the combination of a good algorithm and a correct demographic interpretation can lead to significant improvements in the study of mortality and can have important implications on the evaluation of mortality-linked securities, such as q-forward contracts.

A Appendix

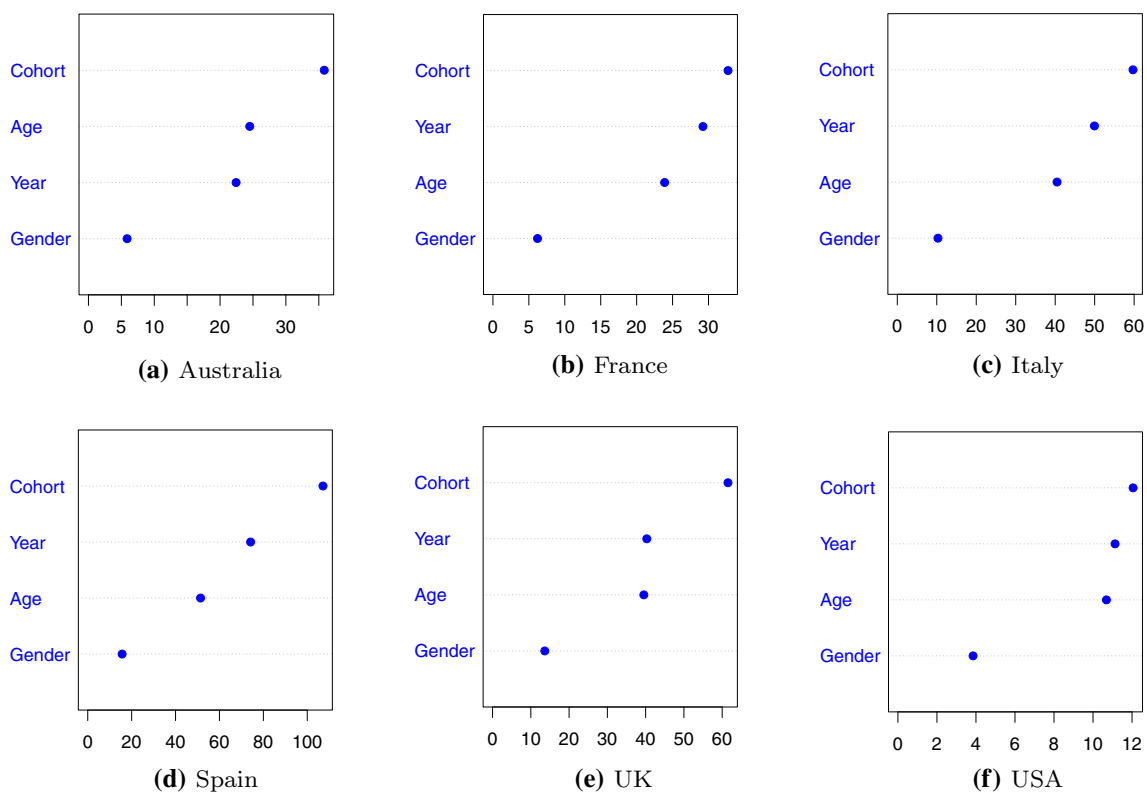See Figs. 7, 8, 9, 10 and 11.

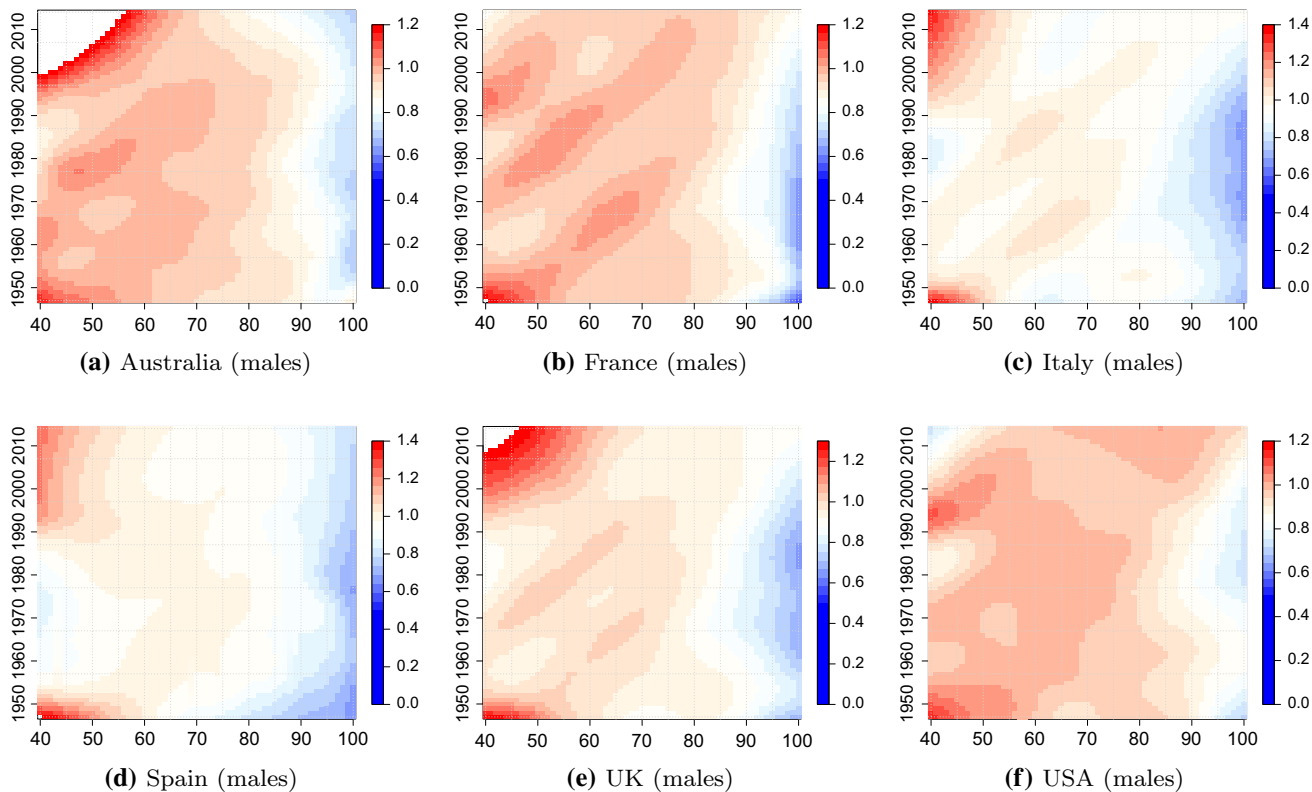**Fig. 7** Variable importance. $\%IncNodePurity$. Ages 20–90 and years 1947–2014

Fig. 8 $\hat{\psi}_s$ smoothed values (1947–2000) and extrapolated values (2000–2014). Results from *MortalitySmooth* package. Male population. Age 40–100
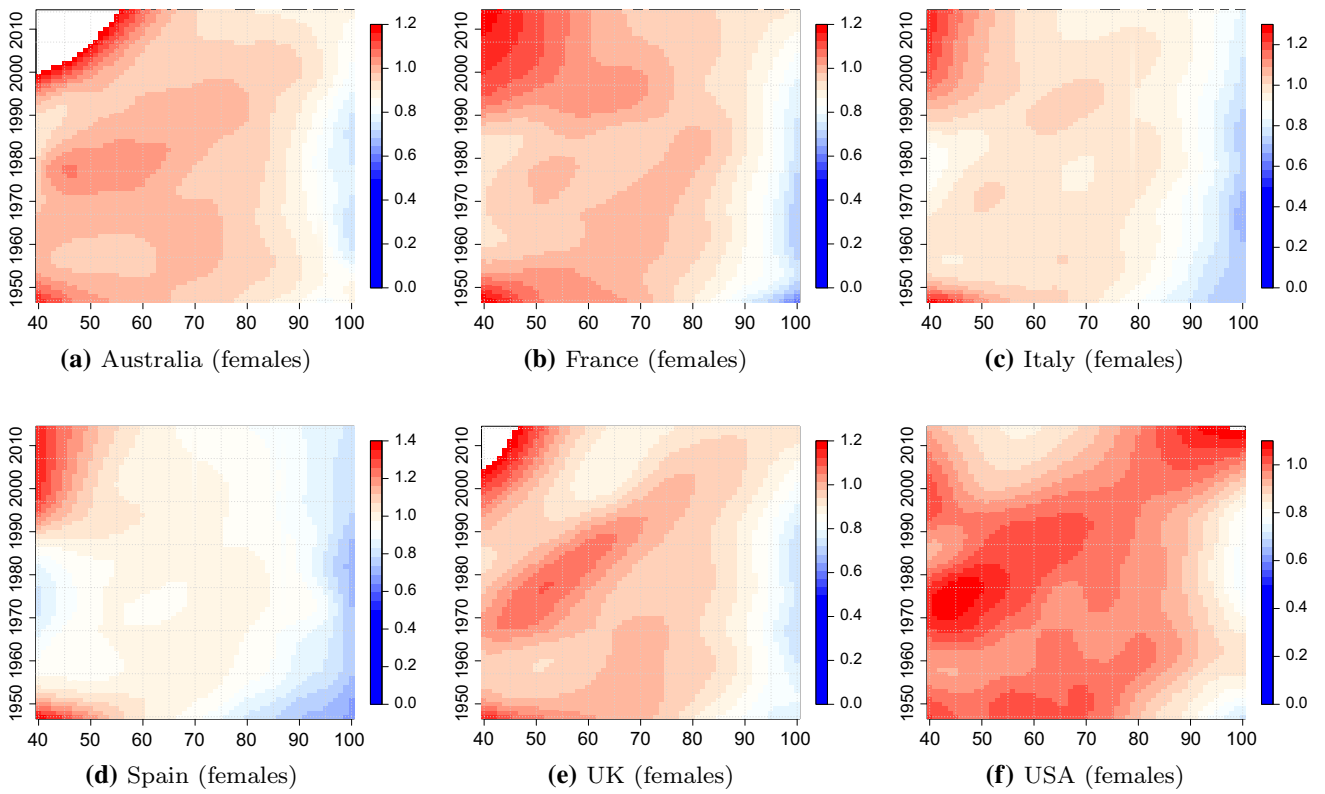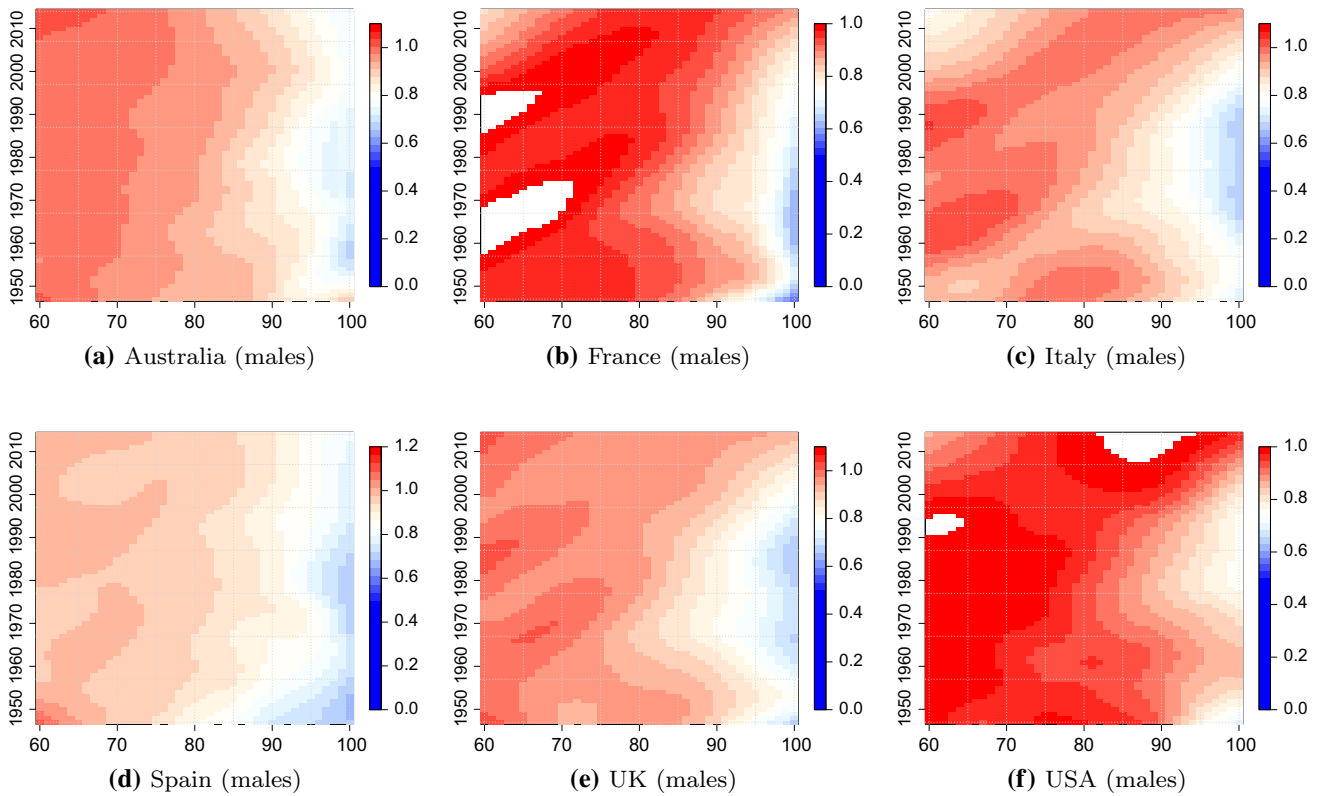
Fig. 9 $\hat{\psi}_\mathbf{s}$ smoothed values (1947–2000) and extrapolated values (2000–2014). Results from *MortalitySmooth* package. Female population. Age 40–100

Fig. 10 $\hat{\psi}_\mathbf{s}$ smoothed values (1947–2000) and extrapolated values (2000–2014). Results from *MortalitySmooth* package. Male population. Age 60–100

(a) Australia (females)   **(b)** France (females)   **(c)** Italy (females)

(d) Spain (females)   **(e)** UK (females)   **(f)** USA (females)
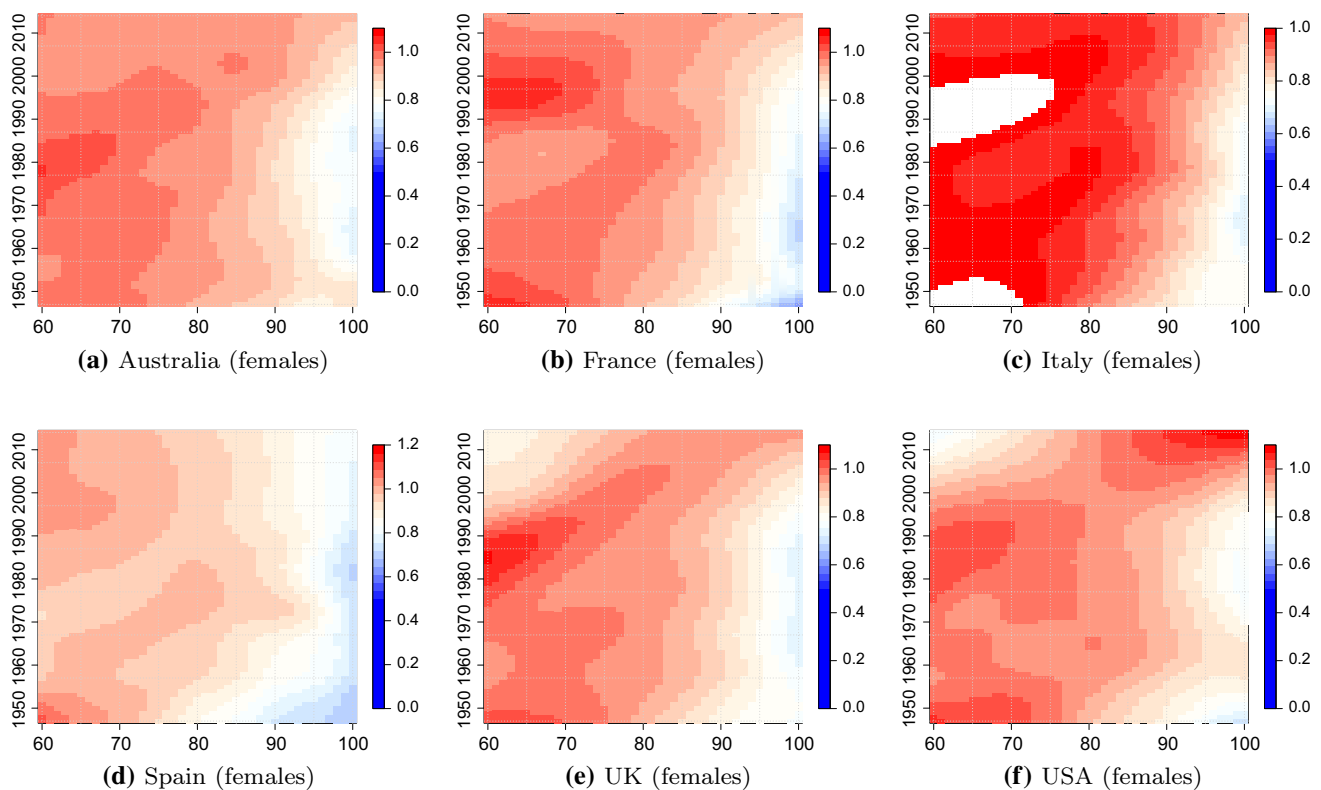
Fig. 11 $\hat{\psi}_s$ smoothed values (1947–2000) and extrapolated values (2000–2014). Results from *MortalitySmooth* package. Female population. Age 60–100

References

Barrieu PM, Veraart LAM (2016) Pricing q-forward contracts: an evaluation of estimation window and pricing method under different mortality models. Scand Actuar J 2:146–166

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. Chapman and Hall/CRC

Brouhns N, Denuit M, Vermunt JK (2002) A Poisson log-bilinear regression approach to the construction of projected lifetables. Insur Math Econ 31(3):373–393

Cairns AJG, Blake D, Dowd K (2008) Modelling and management of mortality risk: a review. Scand Actuar J 2–3:79–113 Pensions Institute Discussion Paper No. PI-0814

Camarda CG (2012) *MortalitySmooth*: an R package for smoothing Poisson counts with P-splines. J Stat Softw 50(1):1–24. http://cran.r-project.org/package=MortalitySmooth

Currie ID, Durban M (2002) Flexible smoothing with P-splines: a unified approach. Stat Model 2:333–49

Currie ID, Durban M, Eilers PHC (2004) Smoothing and forecasting mortality rates. Stat Model 4:279–298

Currie ID, Durban M, Eilers PHC (2006) Generalized linear array models with applications to multidimensional smoothing. J R Stat Soc B 68:259–280

D'Amato V, Piscopo G, Russolillo M (2011) The mortality of the Italian population: smoothing techniques on the Lee–Carter model. Ann Appl Stat 5(2A):705–724

De Boor C (1978) A Practical guide to splines. Springer, New York

Deprez P, Shevchenko PV, Wüthrich M (2017) Machine learning techniques for mortality modeling. Eur Actuar J 7(2):337–352

Eilers PHC, Marx BD (1996) Flexible smoothing with b-splines and penalties. Stat Sci 11:89–102

Eilers PHC, Marx BD (2002) Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. Chemom Intell Lab Syst 66:159–174

Eilers PHC, Currie ID, Durban M (2006) Fast and compact smoothing on large multidimensional grids. Comput Stat Data Anal 50:61–76

Eilers PHC, Marx BD (2010) Splines, knots, and penalties. Wiley Interdiscip Rev Comput Stat 2:637–653

Girosi F, King G (2008) Demographic forecasting. Princeton University Press, Princeton

Hainaut D (2018) A neural-network analyzer for mortality forecast. Astin Bull 48(2):481–508

James G, Witten D, Hastie T, Tibshirani R (2017) An introduction to statistical learning: with applications in R. Springer texts in statistics. Springer, Berlin. ISBN-10: 1461471370

Lee RD, Carter RL (1992) Modeling and forecasting US mortality. J Am Stat Assoc 87(419):659–671

Lee R, Miller T (2001) Evaluating the performance of the Lee–Carter method for forecasting mortality. Demography 38(4):537–549

Levantesi S, Menzietti M (2017) Maximum market price of longevity risk under solvency regimes: the case of solvency II. Risks 5(2):29

Levantesi S, Pizzorusso V (2019) Application of machine learning to mortality modeling and forecasting. Risks 7(1):26

Li N, Lee R, Gerland P (2013) Extending the Lee–Carter method to model the rotation of age patterns of mortality decline for longterm projections. Demography 50(6):2037–2051

Liaw A (2018) Package *randomforest*. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Loeys J, Panigirtzoglou N, Ribeiro R (2007) Longevity: a market in the making. J.P. Morgan's Global Market Strategy, London

Loh WY (2011) Classification and regression trees. Wiley Interdiscip Rev Data Min Knowl Discov 1:14–23

Morgan JN, Sonquist JA (1963) Problems in the analysis of survey data, and a proposal. J Am Stat Assoc 58:415–434

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

O'Sullivan F (1986) A statistical perspective on ill-posed inverse problems (with discussion). Stat Sci 1:505–527

O'Sullivan F (1988) Fast computation of fully automated logdensity and log-hazard estimators. SIAM J Sci Stat Comput 9:363–379

Piscopo G (2017) Dynamic evolving neuro-fuzzy inference system for mortality prediction. Int J Eng Res Appl 7(3):26–29

Piscopo G (2018a) AR dynamic evolving neuro-fuzzy inference system for mortality data. In: Skiadas CH, Skiadas C (eds) Demography and health issues. Population aging, mortality and data analysis. Springer, Berlin

Piscopo G (2018b) A comparative analysis of neuro fuzzy inference systems for mortality prediction. In: Corazza M, Durbán M, Grané A, Perna C, Sibillo M (eds) Mathematical and statistical methods for actuarial sciences and finance. Springer, Berlin

Quinlan JR (1986) Induction of decision trees. Mach Learn 1:81–106

Richman R, Wüthrich M (2018) A neural network extension of the Lee–Carter model to multiple populations. SSRN manuscript, ID 3270877

Ruppert D (2002) Selecting the number of knots for penalized splines. J Comput Graph Stat 11:735–57

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

The Life and Longevity Markets Association (2010) Technical note: q-forward. http://www.llma.org

Villegas AM, Kaishev VK, Millossovich P (2015) Stmomo: an r package for stochastic mortality modelling. J Stat Softw 84(3). https://cran.r-project.org/web/packages/StMoMo/vignettes/StMoMoVignette.pdf

Zeddouk F, Devolder P (2019) Pricing of longevity derivatives and cost of capital. Risks 7:41