

Enhancing diagnostic of stochastic mortality models leveraging contrast trees. An application on Italian data

Susanna Levantesi¹, Matteo Lizzi¹, and Andrea Nigri^{2*}

¹Department of Statistics, Sapienza University of Rome, Viale Regina Elena 295-G, 00161 Rome, Italy

²Department of Economics, Management and Territory, University of Foggia, Foggia, Italy

*andrea.nigri@unifg.it

ABSTRACT

The rise in longevity in the twentieth century has led to a growing interest in modeling mortality, and new advanced techniques such as machine learning have recently joined to more traditional models, such as the Lee-Carter or the Age Period Cohort. However, the performances of these models, in terms of fitting to the observed data, are difficult to compare in a unified framework. The goodness-of-fit measures summarizing the discrepancy between the estimates from the model and the observed values are different for traditional mortality models and machine learning. We, therefore, employ a new technique, Contrast trees, which, leveraging on decision trees, provides a general approach for evaluating the quality of fit of different kinds of models by detecting the regions in the input space where models work poorly. Once the low-performance regions are detected, we use Contrast boosting to improve the inaccuracies of mortality estimates provided by each model. To verify the ability of this approach, we consider both standard stochastic mortality models and machine learning algorithms in the estimate of the Italian mortality rates from the Human Mortality Database. The results are discussed using both graphical and numerical tools, with particular attention to the high-error regions.

Introduction

Since 1980, innovative approaches and developments in mortality modeling have been constantly proposed. Mortality analysis has received a considerable contribution from statistical science, building solid foundations for the evolution of mortality methods. Estimating longevity is not straightforward; accuracy depends on the particular situation or trends, and it is not easy to comprehend when a method will good perform. Indeed, new mortality models will appear in the literature but may take years before they can be fully evaluated. As stated by [Booth and Tickle\(2008\)](#), the accuracy of mortality estimates should be regularly tested to set the improvement evidence. Researchers appear to be more focused on technical progress of a method rather than on the accuracy of the estimation provided, focusing on minimizing the bias.

Several approaches have been used to model the mortality surface, determining how death rates change over time. Until the 1980s, mortality models were relatively simple and involved a fair degree of subjective judgment (see [Pollard\(1987\)](#) for a detailed review on this aspect). The growing availability of reliable data, in lockstep with the improvement of statistical-mathematical methods, has allowed the creation of ever-finer mortality models. According to [Booth and Tickle\(2008\)](#), literature would suggest three approaches to demographic modeling. The first one (explanation) makes use of structural or epidemiological models from certain causes of death. A classic example is the dependence of lung cancer on tobacco smoking. The second one (expectation) is based on subjective expert opinion, involving varying degrees of formality. Finally, the third and most commonly used approach is extrapolative, using the regularity typically found in age patterns and trends over time. This approach includes the more complex stochastic mortality models such as the Lee-Carter [Lee and Carter\(1992\)](#) and, more generally, the Generalized Age Period Cohort (GAPC) model. Despite the Lee-Carter model being widely recognized as the cornerstone of mortality modeling and forecasting, over the last decade, scholars suggested additional approaches that also gained interest in the academic world [Brouhns et al.\(2002\)](#), [Renshaw and Haberman\(2006\)](#), [Cairns et al.\(2006\)](#), [Cairns et al.\(2009\)](#).

Despite models like the Lee-Carter and its variants having been widely used, becoming a benchmark for many newly proposed methodologies, they present several shortfalls. In this line, [Cairns et al.\(2008\)](#) tried to address the issue of what would be the best way to estimate mortality, exhibiting interesting criteria that a good mortality model should hold. They referred to good-practice guidelines such as the consistency with historical data and the long-term dynamics, biologically reasonable. Following this line of research, recent longevity literature stimulated the use of machine learning techniques in demographic research allowing the integration of stochastic models into a data-driven approach.

The significant reduction in the forecasting error reached by the application of machine learning techniques became partic-

ularly useful for both researchers and practitioners. The main contributions are from Deprez et al.(2017) Levantesi and Pizzorusso(2019) and Levantesi and Nigri(2020). The common idea behind all these works is to improve the fitting accuracy of canonical models using machine learning algorithms. In other words, to correct the mortality surface produced by standard stochastic mortality models. All of the proposed methods calibrate a machine learning estimator used to adjust (and improve) mortality rates estimated by the original mortality model. Those authors show that mortality modeling can benefit from machine learning as it better captures patterns that traditional models do not identify.

The need for new tools for comparing models' performances is evident to understand mortality evolution more accurately. This paper contributes to the literature on mortality modeling by introducing an innovative approach based on machine learning techniques that demographers have not yet explored, contributing to the undervalued field of model assessment. This approach, namely Contrast trees, recently proposed by Friedman(2020), and here applied to mortality data, helps evaluate the accuracy of the mortality estimates (fitted mortality rates) given by models that are not treatable with model selection criteria based on the likelihood function. Therefore, this technique provides a unified framework for assessing and comparing the goodness-of-fit to historical data of traditional mortality models with machine learning algorithms. Our paper highlights the ability of Contrast trees to identify the regions in the predictor variables space that show very high values of the error rate quantified by a discrepancy measure. The regions' width and shape change from model to model. Moreover, in addition to evaluating the accuracy of the models, the Contrast trees enables improving the performance of the models through a boosting procedure that reduces the inaccuracies. We use this methodology, namely Contrast boosting, to improve the fitting of historical mortality data. According to the demographic literature, the reliable estimation of mortality data may refer not only to the extrapolation but also to an accurate fitting of the historical mortality surface. For instance, in longevity analysis is common to deal with subpopulations i.e. regions or provinces, characterized by a high level of stochasticity often due to a small number of count data at single ages. This is the case in which specific ages or years are not covered with data information, making the mortality estimation challenging. Our approach is crucial to evaluate the mortality matrix estimation provided by a mortality model and to ensure estimation effectiveness by comparing different methods. To summarize, through this new technique based on Contrast trees, we aim to find the best model that fits observed mortality rates by grasping and detecting the inaccuracies of any model and boosting its predictive power.

The remainder of this paper is organized as follows: Section 2 introduces the model framework, both Contrast trees and Contrast boosting. In Section 3, we describe the numerical implementation, also providing an overview of the mortality models, expressed in a regression framework, which we assess by the Contrast trees approach. We devote a specific sub-section to explanation and discussion of the numerical results. Section 4 concludes the paper, providing other possible practical implementations of the method in mortality assessment and the limitations of our research.

Materials and Methods

Data source

We consider the Italian mortality data available in the Human Mortality Database (HMD) over the period 1950-2018. We refer to the male population aged 0-90, analyzing the age groups 0-29, 30-60, and 61-90 separately to provide further evidence of the differences in mortality that characterizes the younger ages, the adult ages, and the older ages. We split the data set into a training set and a test set according to the common splitting rule 70%-30%. We use the training set to obtain the parameters' estimate of each model. We apply the parameters' estimate in the test set to evaluate the out-of-sample performance. Finally, we will calculate the out-of-sample errors using data from the test set. The dataset partition is obtained by using the dissimilarity-based compound selection proposed in Willett(1999).

Mortality rate

We calculate the central death rates $m_{x,t}$ for each age x and year t according to the following formula:

$$m_{x,t} = \frac{D_{x,t}}{E_{x,t}} \quad (1)$$

Where $D_{x,t}$ is the number of deaths aged x in year t , and $E_{x,t}$ are the exposures-to-risk aged x in year t .

Mortality models

In the following, we briefly describe the four models to which the Contrast trees methodology is applied. The scope is to evaluate the models' quality of fit. The first two models belong to the family of generalized age-period-cohort (GAPC) that are expressed in a regression framework to be suitable for applying Contrast trees, which requires data organized in columns. The last two are well-known machine learning techniques also used for regression tasks.

83

84 **Lee-Carter (LC) model**

The original LC model [Lee and Carter\(1992\)](#) assumes that:

$$\log(m_{x,t}) = \alpha_x + \beta_x \kappa_t \quad (2)$$

85 The age-specific parameter α_x provides the average age profile of mortality, the age-period term $\beta_x \cdot \kappa_t$ describes the mortality
86 trends, with κ_t the time index and β_x modifying the effect of κ_t across ages. The model is subject to the following constraints on
87 κ_t and β_x : $\sum_t \kappa_t = 0$ and $\sum_x \beta_x = 1$. The LC model can be reformulated into a Generalized Non-linear Model (GNM) framework,
88 as in [Villegas et al.\(2018\)](#), following the approach proposed by [Brouhns et al.\(2002\)](#), which assume that deaths are independent Poisson
89 distributed. The authors use a GNM and apply the maximum likelihood method to fit the model to historical data. Under this
90 specification, the LC model can be seen as a non-linear regression model where mortality rates are the target variable, predicted
91 using features (age and time) [Richman and Wüthrich\(2021\)](#).

92 **Age-Period-Cohort (APC)**

We use the model's version reformulated into a Generalized Linear Models (GLM) framework [Alai and Sherris\(2014\)](#):

$$\log(m_{x,t}) = \beta_0 + \beta_{1,x} + \beta_{2,t} + \beta_{3,t-x} \quad (3)$$

93 Where the regression coefficients $\beta_{1,x}$, $\beta_{2,t}$, $\beta_{3,t-x}$ are the age trend, the period trend and the cohort trend ($t - x$ represents the
94 year of birth).

95 **Gradient Boosting Machine (GBM)**

96 GBM is a tree-based algorithm proposed by [Friedman\(2001\)](#) that uses fixed-size decision trees as weak learners. The prediction is
97 obtained by a sequential approach, where each decision tree uses the information from the previous one to improve the current
98 fit. Given a current model fit, $F_m(\mathbf{x})$, the algorithm provides a new estimate, $F_{m+1}(\mathbf{x}) = F_m(\mathbf{x}) + h_m(\mathbf{x})$, where $h_m(\mathbf{x})$ is the
99 weak learner fitted on the model residuals $y - F_m(\mathbf{x})$ with y target variable.

100 **eXtreme Gradient Boosting Machine (XGBM)**

101 XGBM is an efficient implementation of gradient boosting decision trees proposed by [Chen et al.\(2015\)](#), and designed to be fast to
102 execute and highly effective. To verify if a simple data preprocessing has some meaningful effect on the quality of models, we
103 apply XGBM to both raw and preprocessed data: the latter is obtained by centering and scaling the raw data using mean and
104 standard deviation.

105 **Traditional diagnostic tools**

106 In the following, we briefly mention some traditional diagnostic tools that are often used in the literature to assess the
107 goodness-of-fit of a mortality model.

- 108 • Analysis of mortality residuals (or standardized mortality residuals) calculated as the difference between the crude
109 estimate of mortality rate by age and year based on observed data and the corresponding estimated mortality rate using
110 a specified mortality model. For example, [Cairns et al.\(2010\)](#) verified that they are consistent with the hypothesis of i.i.d.
111 $N(0, 1)$ and have zero correlation both across adjacent ages and across adjacent years.
- 112 • Proportion of variance explained (R^2) by the model or the parameters of the model (see, e.g., [Bongaarts\(2005\)](#))
- 113 • Model selection criteria that penalize the log-likelihood with the increase in number of parameters: Akaike Information
114 Criterion (AIC), Schwarz-Bayes Criterion (SBC) (or Bayes Information Criterion (BIC)) and Likelihood-ratio test
115 (LRT) [Li et al.\(2009\)](#). Note that in this case the evaluation of the goodness-of-fit is given on the basis of the log-likelihood.
- 116 • Qualitative model selection criteria: [Cairns et al.\(2008\)](#) provide a list of criteria that might be considered desirable in a
117 mortality model, such as, e.g., ease of implementation, parsimony, and transparency. Relating to the fitting ability to the
118 observed data, the model should be consistent with historical data, and parameter estimates should be robust relative
119 to the range of data used. For example, [Djoundje et al.\(2022\)](#) consider consistency, stability, and parsimony in addition to
120 standard goodness-of-fit indices (deviance residual, BIC, and residual patterns).
- 121 • Checking for the absence of autocorrelation in the residuals of the model by the Portmanteau test (see, e.g., [Torri\(2011\)](#)).

122 **Contrast trees**

123 Contrast trees is an innovative approach that, leveraging tree-based machine learning techniques, allows for deeply assessing the
 124 goodness-of-fit of a model by identifying where the model performs worse. Specifically, the goal of the Contrast trees method
 125 is to uncover regions in the predictor variables space presenting very high values of the error rate quantified by a discrepancy
 126 measure [Friedman\(2020\)](#). In the context of mortality modeling, the main feature that distinguishes this method from the traditional
 127 diagnostic methods mentioned above is the ability to automatically identify the regions in which a given model provides a high
 128 error for certain combinations of ages and calendar years. Furthermore, Contrast trees have the advantage of being easy to
 129 interpret and can be used as a diagnostic tool to detect the inaccuracies of every kind of model, for example, both those whose
 130 parameters estimate is based on a likelihood function and those based on machine learning algorithms. Our analysis shows how
 131 Contrast trees can be used for assessing the goodness-of-fit of different mortality models to observed data.

Suppose to have a set of predictor variables $x = (x_1, x_2, \dots, x_p)$ and two outcome variables y and z for each x . We aim to find those values of x for which the respective distributions of $y|x$ and $z|x$, or some statistics such as mean or quantiles, are most different. In summary, Contrast trees provide a lack-of-fit measure for the conditional distribution $p_y(y|x)$, or some statistics. Consider the M^{th} iteration, where the tree splits the space of the predictor variables into M disjoint regions $\{R_m\}_{m=1}^M$, each one containing a subset of the data. We denote $f_m^{(l)}$ and $f_m^{(r)}$ the fraction of observations in the left and right region with respect to R_m , respectively. While, the quantities $d_m^{(l)}, d_m^{(r)}$ respectively represent the discrepancy measures associated to the fractions $f_m^{(l)}$ and $f_m^{(r)}$. Given a specified subset of the data $\{x_i, y_i, z_i\}_{x_i \in R_m}$, a discrepancy measure between y and z values can be generally defined as:

$$d_m = D(\{y_i\}_{x_i \in R_m}, \{z_i\}_{x_i \in R_m}) \quad (4)$$

The quality of a split is quantified by the following measure:

$$Q_m(l, r) = \left(f_m^{(l)} \cdot f_m^{(r)} \right) \cdot \max \left(d_m^{(l)}, d_m^{(r)} \right)^\beta \quad (5)$$

132 The factor $\left(f_m^{(l)} \cdot f_m^{(r)} \right)$ discourages highly asymmetric splits in anticipation of further splitting, while the other factor
 133 $\max \left(d_m^{(l)}, d_m^{(r)} \right)^\beta$ attempts to isolate the $R_m^{(l)}$ and $R_m^{(r)}$ regions with high discrepancy. The parameter β regulates the relative
 134 influence of the two factors but, as stated by [Friedman\(2020\)](#), results are insensitive to its value. We will use $\beta = 2$ in our
 135 analysis.

The choice of the discrepancy measure depends on the problem to be solved, allowing Contrast trees to be applied to a variety of problems [Friedman\(2020\)](#). They are similar to loss criteria in prediction problems. The discrepancy measures that could be appropriate to represent the problem under investigation are the following:

$$d_m^{[1]} = \frac{1}{N_m} \sum_{x_i \in R_m} |y_i - z_i| \quad (6)$$

$$d_m^{[2]} = \frac{1}{2N_m - 1} \sum_{i=1}^{2N_m - 1} \frac{|\hat{F}_y(t_{(i)}) - \hat{F}_z(t_{(i)})|}{\sqrt{i \cdot (2N_m - i)}} \quad (7)$$

136 where N_m is the number of observations in the region R_m , $t_{(i)}$ is the i^{th} value of t in sorted order, and \hat{F}_y and \hat{F}_z are the respective
 137 empirical cumulative distributions of y and z . See [Friedman\(2020\)](#) for further details about the tree split procedure.

138 In numerical applications, for sake of simplicity, we use the discrepancy measure $d_m^{[1]}$.

139 **Lack-of-fit contrast curves**

The results obtained by applying the Contrast trees to different models can be summarized in the lack-of-fit contrast curves, which have point coordinates

$$[f_m, \bar{d}_m]$$

140 where $f_m = \frac{1}{N} \sum_{d_j \geq d_m} N_j$ is the fraction of observations in the region R_m containing N_m observations, and $\bar{d}_m = \frac{\sum_{d_j \geq d_m} d_j N_j}{\sum_{d_j \geq d_m} N_j}$ is
 141 the average discrepancy.

142 From the above expressions, we can deduce that the lack-of-fit curves by construction are decreasing. By way of example,
 143 we show a typical pattern of this curve in Fig. 1, where the leftmost point on the abscissa-axis provides the fractions of
 144 observations that fall into the regions with the higher discrepancy, while the rightmost point corresponds to all the observations
 145 ($f_m = 1$). Looking at the ordinate-axis, the leftmost point on each curve represents the \bar{d}_m value of the largest discrepancy
 146 region of its corresponding tree; the rightmost point provides the \bar{d}_m value across all regions. Points in between give a \bar{d}_m value
 147 over the regions with the highest discrepancy that contain the corresponding fraction of observations [Friedman\(2020\)](#).

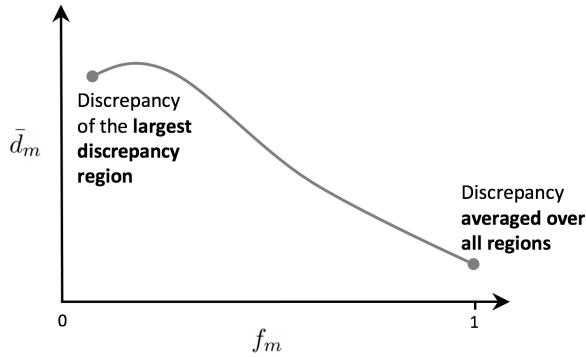


Figure 1. Example of a lack-of-fit contrast curve

148 Contrast boosting

149 To improve the models accuracy, [Friedman\(2020\)](#) proposes a contrast-boosting strategy that, dealing with the uncovered errors, can
 150 enable the regression models to provide more accurate predictions. Contrast boosting works by gradually modifying a starting
 151 value of z to reducing its discrepancy with y over the data. The resulting prediction is then affected by these modifications on
 152 the initial value of z . We consider the estimation Contrast boosting, which takes z as an estimate of a parameter of the full
 153 conditional distribution of a target variable given a set of predictor variables, $p_y(y|x)$. The procedure consists of modifying
 154 the z values within a certain region $R_m^{(1)}$ of a CT, so that its discrepancy with y is zero, i.e. to set $d_m = 0$ in Eq. 4. This is an
 155 iterative procedure, where the first modification is from z to $z^{(1)} = z + \delta_m^{(1)}$ for $x \in R_m^{(1)}$, the second from $z^{(1)}$ to $z^{(2)} = z + \delta_m^{(2)}$
 156 for $x \in R_m^{(2)}$, and so on. The z values final estimate is then $\tilde{z}(x) = z(x) + \sum_{k=1}^K \delta_m^{(k)}$, where K are the maximum number of
 157 iterations. In practice, each updated value of z is contrasted with y producing new regions $R_m^{(k)}$ ($1 \leq k \leq K$) with corresponding
 158 updates $\delta_m^{(k)}$.

159 Results

160 We implement the analyses using the *conTree* R package developed by [Friedman and Narasimhan\(2020\)](#). We set to 100 the maximum
 161 tree size corresponding to the number of regions. It is worth noting that the choice of this parameter is not straightforward
 162 because it involves a trade-off between discrepancy and interpretability. The smaller the trees, the larger the regions (defined by
 163 simple rules and easy to be interpreted). The larger the trees, the higher the potential to uncover small high discrepancy regions
 164 (defined by complex rules).

165 The models' performance results on the test set are summarized in the lack-of-fit contrast curves, deduced by contrasting
 166 the observed mortality data to the estimates provided by each model. These curves are shown in Fig. 2 for the three different
 167 ages groups analyzed. The panels (a)-(c) of these figures refer to the lack-of-fit curves obtained without applying the Contrast
 168 boosting (Base models), while panels (d)-(f) refer to the lack-of-fit curves obtained after applying Contrast boosting to the
 169 output of the models (Boosted models). For the 0-29 age group (Fig. 2, panel (a) and (d)), both APC and XGBM model
 170 have the lowest discrepancy values for each fraction of observations, providing the best fitting. The average discrepancy for
 171 this age group is higher than for the 30-60 age group. The 0-29 age group is known to be characterized by high accidental
 172 mortality, the so-called "accident hump" around age 20-25, due to accidental deaths or suicides caused by increased risk-taking
 173 behavior. Mortality at age 0-29 is therefore hard to predict, and Contrast boosting is not able to actually reduce the average
 174 discrepancy. For the 30-60 age group (Fig. 2, panel (b) and (e)), the APC model seems to best perform across all regions since
 175 the discrepancy values are consistently lower than those of the other models. For the XGBM models, we can observe that the
 176 model applied to preprocessed data (XGBM prep) performs better in the regions with the highest average discrepancy with
 177 respect to the model applied to raw data. From the scale of the plots, we can see that Contrast boosting reduces discrepancy
 178 across almost all regions for the GBM and LC models, where the relative effect of boosting is particularly evident. For the
 179 61-90 age group (Fig. 2, panel (c) and (f)), the GBM model seems by far the worst performing model. Albeit the application of
 180 Contrast boosting significantly reduces the discrepancy, the GBM continues to be less accurate than the other models. It should
 181 also be noted that the effect of Contrast boosting in high-discrepancy regions for the other models is negligible, except for the
 182 APC.

183 Table 1 reports the values of the average discrepancy measure for both the base and the boosted models considered in the
 184 analysis. The APC and the XGBM base models provide the lowest average discrepancy values (0.000410 and 0.000417,

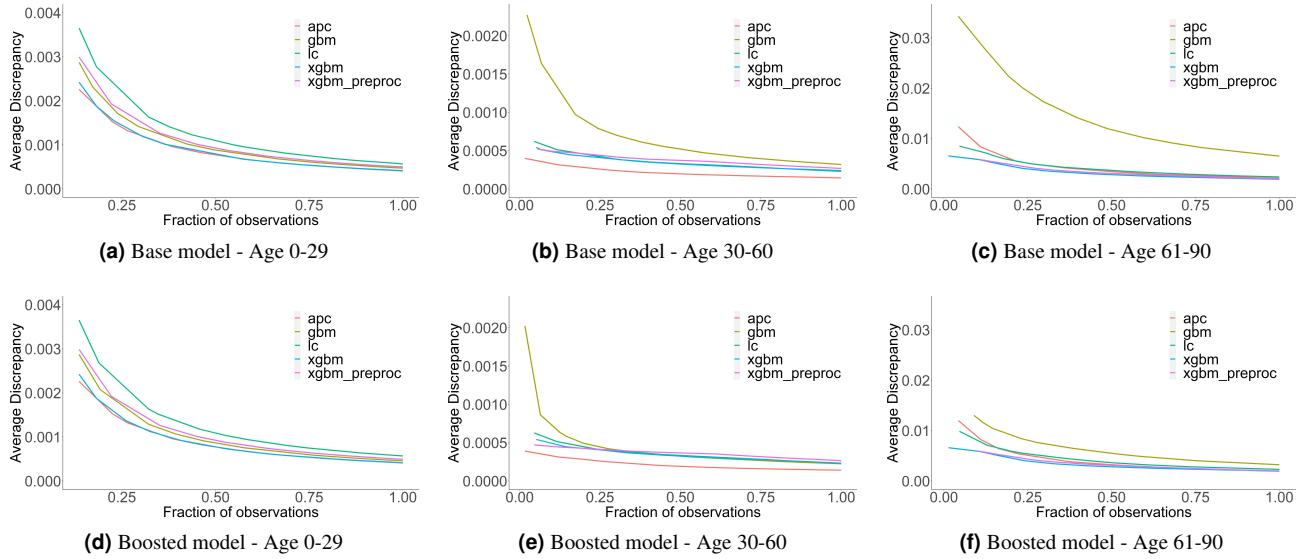


Figure 2. Lack-of-fit contrast curves for APC, LC, GBM, XGBM and XGBM prep.

185 respectively), which remain substantially unchanged after the Contrast boosting procedure. The APC model shows the lowest
 186 value of \bar{d}_m also for the age group 30-60, in line with the dynamics of the lack-of-fit curves depicted in panels (b) and (e) of Fig.
 187 2. However, the lack-of-fit curves provide more structured information than the average discrepancy, in particular, regarding
 188 how and how much \bar{d}_m varies across the input space. For example, for the age group 61-90 in the base model (panel (c)),
 189 we can appreciate that the main difference among models (except for GBM, which is out of range) measured by the average
 190 discrepancy is caused by the high discrepancy regions (where the fraction of observation is less than about 0.20). For ages
 191 61-90, the GBM base model shows the worst fitting to the observed mortality data. Although Contrast boosting produces a
 192 strong improvement in the discrepancy measure, GBM remains the worst model in terms of discrepancy. Contrast boosting is
 193 very effective also for the GBM model in the age group 30-60, as it heavily lowers (-30%) the average discrepancy between
 observed and estimated values.

Model	Age 0-29			Age 30-60			Age 61-90		
	Base	Boosted	% Change	Base	Boosted	% Change	Base	Boosted	% Change
APC	0.000410	0.000409	0%	0.000145	0.000141	-3%	0.002142	0.001948	-9%
LC	0.000571	0.000568	0%	0.000231	0.000232	0%	0.002395	0.002314	-3%
GBM	0.000473	0.000459	-3%	0.000320	0.000225	-30%	0.006525	0.003238	-50%
XGBM	0.000417	0.000415	0%	0.000240	0.000233	-3%	0.001916	0.001940	1%
XGBM prep	0.000500	0.000493	-1%	0.000268	0.000265	-1%	0.002003	0.002005	0%

Table 1. Values of the average discrepancy \bar{d}_m calculated on $m_{x,t}$ in the test set.

194 For a comparison with the average discrepancy, we also calculate the Root Mean Square Error (RMSE) and Mean Absolute
 195 Percentage Error (MAPE) on the base model and the boosted one. Intuitively, the three measures \bar{d}_m , RMSE, and MAPE
 196 quantify the "distance" between the estimates and the actual observations. However, the average discrepancy is an innovative
 197 measure summarizing the discrepancy over all the regions identified by the Contrast trees, while RMSE and MAPE are
 198 commonly used error measures calculated on the overall input space without distinguishing by region.

199 By comparing Table 2 showing the values of RMSE and MAPE with Table 1 reporting the values of the average discrepancy,
 200 we note a greater convergence of the error measures in the boosted models rather than in the base models. This result is
 201 intuitively straightforward since the boosted models are obtained by just reducing the discrepancy measure.

202 We also calculate average discrepancy, RMSE, and MAPE on the logarithm of the central death rates (Tables 3-4). These
 203 measures assign a relatively large weight to errors at young ages, while error measures calculated on the central death rates
 204 assign a large weight to errors at older ages. Indeed, for the age group 0-29, all the errors reported in Tables 3-4 are significantly
 205 higher than those in Tables 1-2. The errors calculated on the logarithm of the central death rates highlight the ability of Contrast
 206 boosting to reduce the inaccuracy of GBM and XGBM-prep in fitting observed mortality at ages 0-29.

Error	Model	Age 0-29			Age 30-60			Age 61-90		
		Base	Boosted	% Change	Base	Boosted	% Change	Base	Boosted	% Change
RMSE	APC	0.002040	0.002039	0%	0.000264	0.000263	0%	0.004260	0.004139	-3%
	LC	0.003471	0.003471	0%	0.000491	0.000496	1%	0.004258	0.004363	2%
	GBM	0.001648	0.001647	0%	0.000640	0.000455	-29%	0.012248	0.005439	-56%
	XGBM	0.001517	0.001515	0%	0.000342	0.000338	-1%	0.003260	0.003278	1%
	XGBM prep	0.001939	0.001935	0%	0.000391	0.000386	-1%	0.003339	0.003345	0%
MAPE	Age 0-29			Age 30-60			Age 61-90			
	Base	Boosted	% Change	Base	Boosted	% Change	Base	Boosted	% Change	
	APC	14.7%	14.5%	-1%	4.5%	4.3%	-3%	3.9%	3.4%	-14%
	LC	14.2%	13.8%	-3%	7.2%	7.1%	-1%	4.9%	4.9%	0%
	GBM	23.4%	18.8%	-20%	13.0%	7.6%	-41%	18.3%	9.2%	-50%
	XGBM	15.9%	15.3%	-3%	6.9%	6.2%	-10%	3.7%	3.8%	2%
	XGBM prep	20.0%	18.2%	-9%	7.3%	7.2%	-1%	3.6%	3.6%	0%

Table 2. Values of the RMSE and MAPE calculated on $m_{x,t}$ in the test set.

Model	Age 0-29			Age 30-60			Age 61-90		
	Base	Boosted	% Change	Base	Boosted	% Change	Base	Boosted	% Change
APC	0.149906	0.148218	-1%	0.040837	0.040276	-1%	0.036633	0.035584	-3%
LC	0.151968	0.149051	-2%	0.066676	0.070489	6%	0.042757	0.039114	-9%
GBM	0.292233	0.260784	-11%	0.109899	0.052510	-52%	0.118491	0.052240	-56%
XGBM	0.195720	0.191478	-2%	0.066986	0.062600	-7%	0.036712	0.036779	0%
XGBM prep	0.207129	0.186137	-10%	0.072703	0.072571	0%	0.035729	0.035505	-1%

Table 3. Values of the average discrepancy \bar{d}_m calculated on $\log(m_{x,t})$ in the test set.

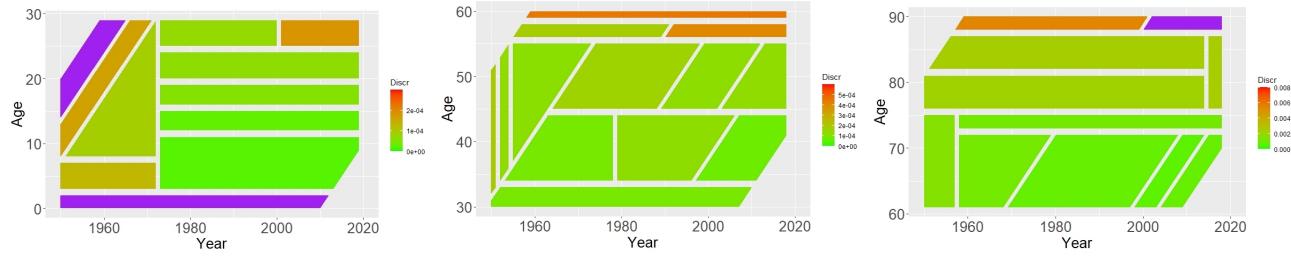
209 The most interesting feature of the application of Contrast trees to the field of mortality estimate is the automatic identification
210 of the regions of the predictors' space where a given model provides high discrepancy values for certain combinations of
211 ages-years obtained by comparing the model estimates with the observed mortality rates. These regions can be easily detected
212 and possibly interpreted, providing a further explanation of the model performances as well as helping to assess whether a
213 model can be reliable or not. Fig. 3 and Fig. 4 show the heatmap of all the error regions for the base model and the boosted one,
214 respectively. Low discrepancy regions are painted in green, while high discrepancy regions are painted in red. For the sake of
215 plot readability, we colored in purple the regions presenting a discrepancy value exceeding 3e-04, 6e-04, and 0.008 for the age
216 groups 0-29, 30-60, and 61-90, respectively.

217 We can generally observe that the regions' width and shape change from model to model. Some regions show remarkable
218 mortality estimation errors in specific age groups, others in specific intervals of years, others in a specific range of cohorts. All
219 the models considered show high discrepancy values in the first year of age (Fig. 3, age group 0-29, left panels), confirming
220 the difficulty of adequately estimating the mortality of newborns. This situation remains unchanged after the application of
221 Contrast boosting, which, in this case, seems to be not effective (Fig. 4, age group 0-29, left panels). For the age group 30-60 in
222 the base model (Fig. 3, central panels), the two XGBM models show high discrepancy values after age 45-46, while GBM in
223 the years 2000-2018. The LC model instead evidences high errors in estimating the mortality of cohorts born between 1920 and
224 1932. Considering the 61-90 age group (Fig. 3, right panels), we notice that the GBM model continues to fail in estimating
225 mortality rates in the years 2000-2018, while the LC model (and also APC) mortality rates in the cohorts born between 1920
226 and 1932. By comparing the results for the base models (Fig. 3) with those for the boosted ones (Fig. 4), we observe a clear
227 effect of boosting on the GBM model for the 30-60 and 61-90 age groups and the XGBM for the 30-60 age group.

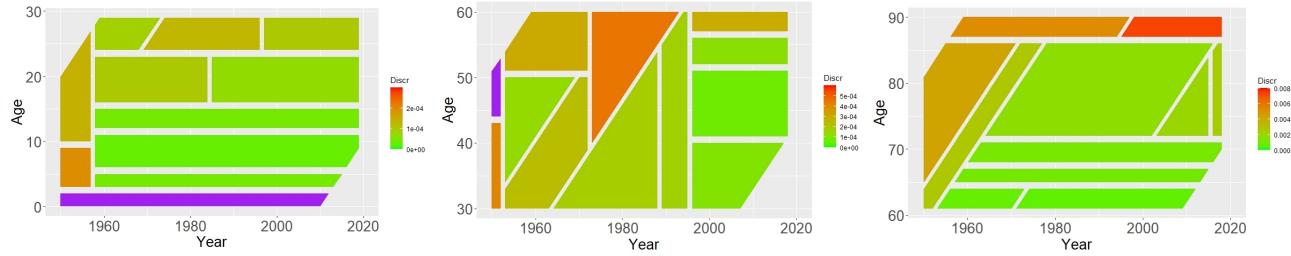
228 Discussion

229 Evaluating, and thus eventually improving, the fit of mortality models is crucial for both demographers and actuaries. Indeed,
230 in particular situations, common in actuarial practice, data quality can turn the mortality estimate difficult. A prime example is
231 the case of small subpopulations where a common method such as the Lee-Carter may not guarantee reliable estimation. In this
232 sense, our proposal fills the gap between mortality modeling and model diagnostics, particularly for nontraditional modeling as
233 a machine learning framework.

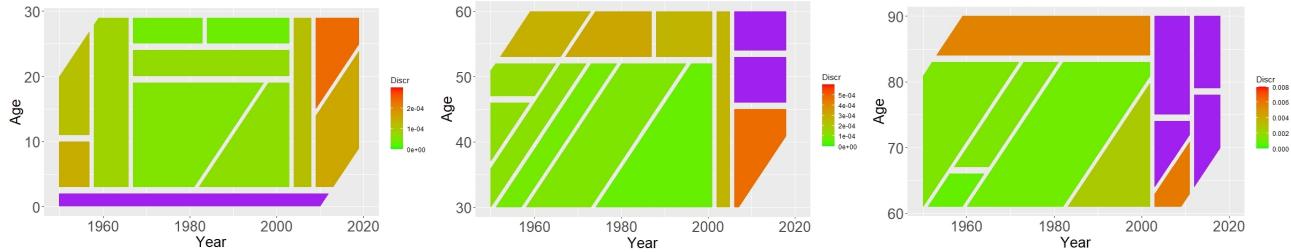
234 Contrast trees consist of a general method based on machine learning that can be applied to any model, expressed as a regression
235 model, to evaluate the goodness of fit and identify the worst-performing regions in the input space. The main characteristic that
236 discriminates this method from traditional diagnostic tools is automatically identifying the regions in which a given model
237 produces a high error for certain combinations of ages and calendar years. Well-known diagnostic tools often used in the
238 literature to assess the goodness-of-fit of a mortality model, such as BIC and AIC, require the likelihood function, which is not
239 available for machine learning models. Therefore, Contrast trees provide a unified approach for assessing and comparing the



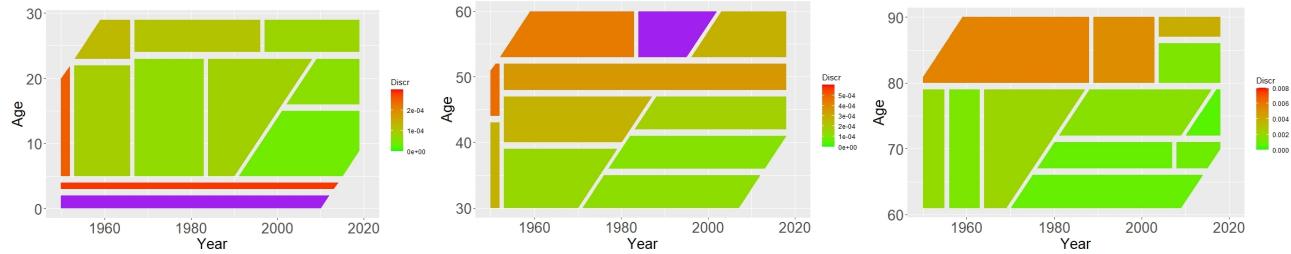
(a) APC (left: age 0-29; centre: age 30-60 right: age 61-90)



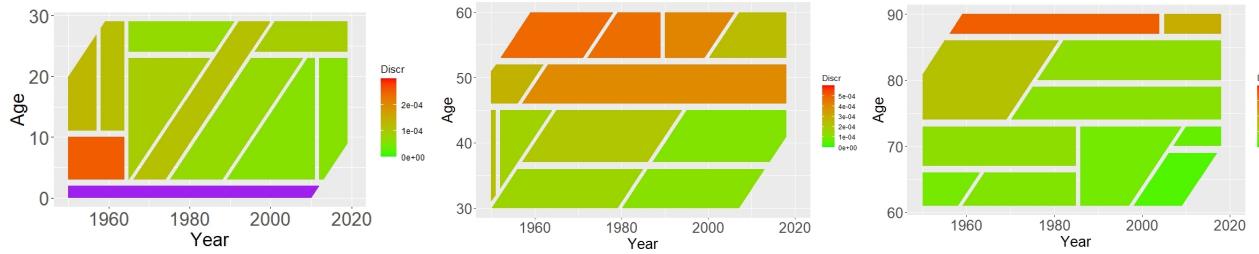
(b) LC (left: age 0-29; centre: age 30-60 right: age 61-90)



(c) GBM (left: age 0-29; centre: age 30-60 right: age 61-90)

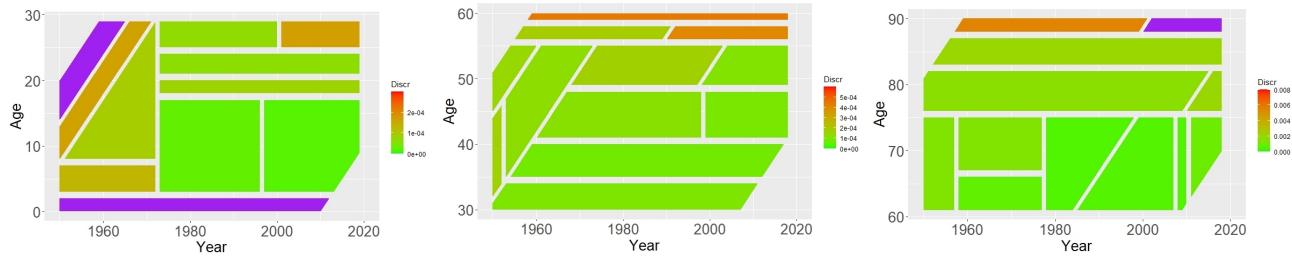


(d) XGBM (left: age 0-29; centre: age 30-60 right: age 61-90)

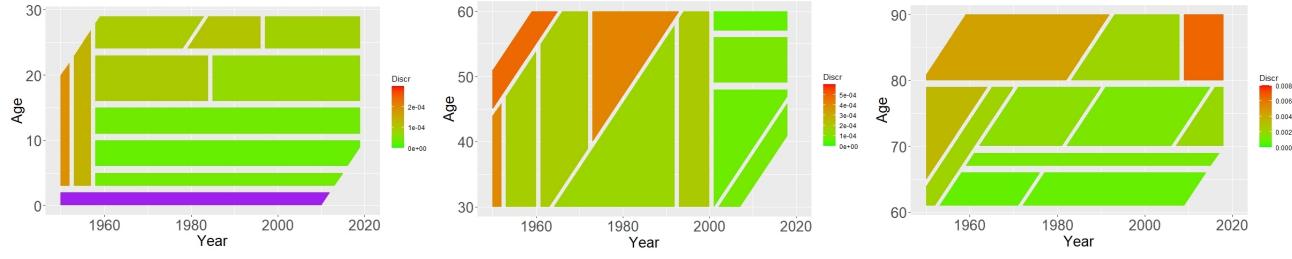


(e) XGM prep (left: age 0-29; centre: age 30-60 right: age 61-90)

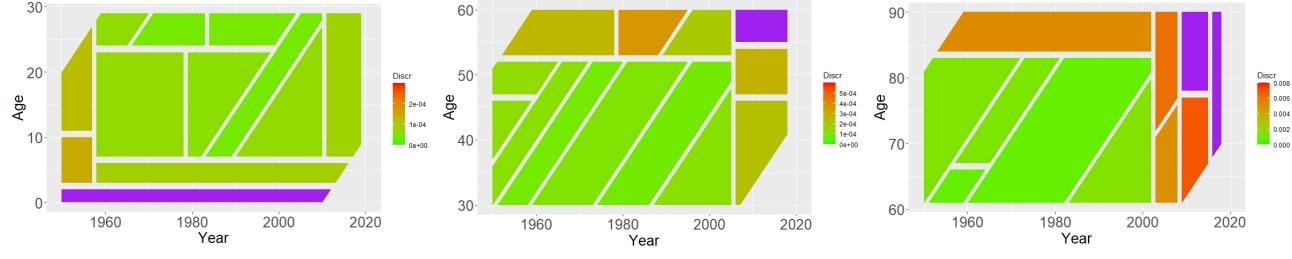
Figure 3. Contrast trees regions, Base model. Years 1950-2018. Regions presenting a discrepancy value exceeding 3e-04 (age 0-29), 6e-04 (age 30-60), and 0.008 (61-90) are colored in purple.



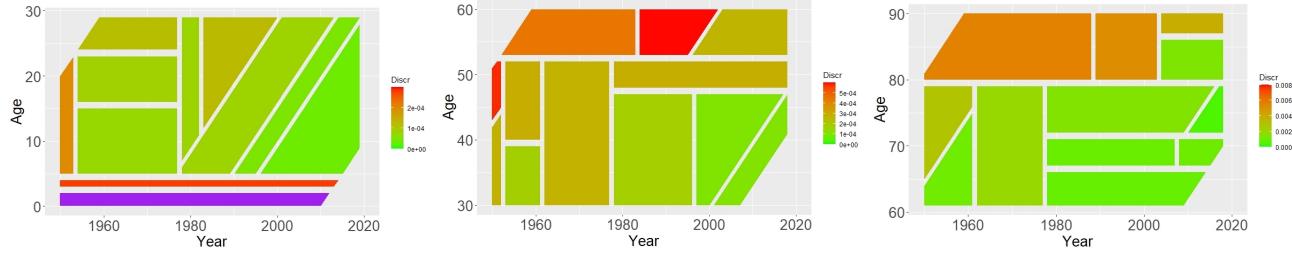
(a) APC (left: age 0-29; centre: age 30-60 right: age 61-90)



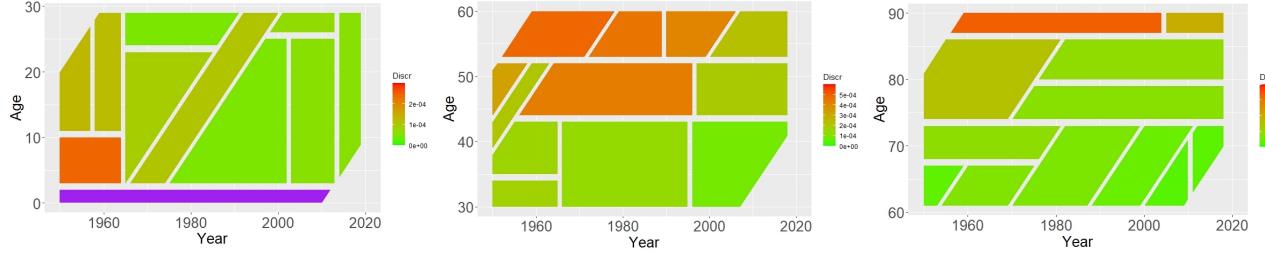
(b) LC (left: age 0-29; centre: age 30-60 right: age 61-90)



(c) GBM (left: age 0-29; centre: age 30-60 right: age 61-90)



(d) XGBM (left: age 0-29; centre: age 30-60 right: age 61-90)



(e) XGM prep (left: age 0-29; centre: age 30-60 right: age 61-90)

Figure 4. Contrast trees regions, Boosted model. Years 1950-2018. Regions presenting a discrepancy value exceeding 3e-04 (age 0-29), 6e-04 (age 30-60), and 0.008 (61-90) are colored in purple.

Error	Model	Age 0-29			Age 30-60			Age 61-90			
		Base	Boosted	% Change	Base	Boosted	% Change	Base	Boosted	% Change	
RMSE	APC	0.197670	0.197502	0%	0.062659	0.062708	0%	0.049847	0.049678	0%	
	LC	0.237064	0.232791	-2%	0.101946	0.106096	4%	0.060798	0.056322	-7%	
	GBM	0.707561	0.694538	-2%	0.182743	0.071172	-61%	0.233815	0.081485	-65%	
	XGBM	0.503457	0.496244	-1%	0.089719	0.084808	-5%	0.053659	0.053511	0%	
	XGBM prep	0.306517	0.272596	-11%	0.096897	0.097073	0%	0.048344	0.048111	0%	
MAPE	Age 0-29			Age 30-60			Age 61-90				
	MAPE	Base	Boosted	% Change	Base	Boosted	% Change	Base	Boosted	% Change	
		2.1%	2.2%	0%	0.8%	0.8%	0%	1.4%	1.2%	1%	
		LC	2.1%	2.1%	-2%	1.2%	1.3%	5%	1.6%	1.6%	1%
		GBM	3.5%	3.1%	-12%	1.9%	0.9%	-52%	4.7%	5.6%	2%
		XGBM	2.6%	2.5%	-1%	1.2%	1.1%	-4%	1.2%	1.3%	1%
		XGBM prep	2.8%	2.7%	-3%	1.2%	1.2%	0%	1.2%	1.2%	1%

Table 4. Values of the RMSE and MAPE calculated on $\log(m_{x,t})$ in the test set.

accuracy of traditional mortality models with machine learning algorithms.

In Contrast trees, the detection of the regions in which a model worst performs can be considered an evolution of the standard analysis on residuals, in which the detection of the highest residuals is typically assigned to graphical analyzes using heatmaps and scatter plots [Cairns et al.\(2009\)](#), [Villegas et al.\(2018\)](#), and to summary measures like RMSE and MAPE calculated on the overall input space and not by region. Conversely, the decision tree structure of Contrast trees enables quantifying the discrepancy between the estimates provided by a model and the actual observations in each region identified by Contrast trees.

References

- Alai and Sherris(2014).** Alai, D.H., Sherris, M. (2014). Rethinking age-period-cohort mortality trend models. Scandinavian Actuarial Journal, 3: 208-227.
- Bongaarts(2005).** Bongaarts, J. (2005). Long-Range Trends in Adult Mortality: Models and Projection Methods. Demography, 42(1): 23-49.
- Booth and Tickle(2008).** Booth, H., Tickle, L. (2008). Mortality modelling and forecasting: A review of methods. Annals of Actuarial Science, 3(1-2): 3-43. DOI: 10.1017/S1748499500000440.
- Brouhns et al.(2002).** Brouhns, N., Denuit, M., Vermunt, J. (2002). A Poisson log-bilinear approach to the construction of projected life tables, Insurance: Mathematics and Economics, 31: 373-393.
- Chen et al.(2015).** Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.
- Cairns et al.(2006).** 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 and Insurance, 73: 687-718.
- Cairns et al.(2008).** Cairns, A.J.G., Blake, D., Dowd, K. (2008). Modelling and management of mortality risk: a review. Scandinavian Actuarial Journal, 73 (2-3): 79-113.
- Cairns et al.(2009).** Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A., Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. North American Actuarial Journal, 13: 1-35.
- Cairns et al.(2010).** Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Khalaf-Allah, M. (2010). Evaluating the goodness of fit of stochastic mortality models. Insurance: Mathematics and Economics, 47(3): 255-265.
- Deprez et al.(2017).** Deprez, P., Shevchenko, P.V., Wúthrlich, M.V (2017). Machine learning techniques for mortality modeling. European Actuarial Journal, 7: 337-352. <https://doi.org/10.1007/s13385-017-0152-4>
- Djeundje et al.(2022).** Djeundje, V.B., Haberman, S., Bajekal, M. et al. (2022). The slowdown in mortality improvement rates 2011-2017: a multi-country analysis. European Actuarial Journal. DOI: 10.1007/s13385-022-00318-0
- Friedman(2001).** Friedman, J.H. (2001). Greedy function approximation: A Gradient Boosting Machine. Annals of Statistics 29: 1189-1232.
- Friedman(2020).** Friedman, J.H. (2020). Contrast trees and distribution boosting. Proceedings of the National Academy of Sciences, 117 (35): 21175-21184. DOI: 10.1073/pnas.1921562117

- 274 **Friedman and Narasimhan(2020).** Friedman, J.H., Narasimhan, B. (2020). conTree: Contrast Trees and Distribution Boosting. R package version 0.2-8.
- 275
- 276 **Lee and Carter(1992).** Lee, R.D., Carter, L.R. (1992). Modeling and forecasting US mortality. Journal of the American statistical association, 87 (419): 659-671.
- 277
- 278 **Levantesi and Nigri(2020).** Levantesi, S., Nigri, A. (2020). A random forest algorithm to improve the Lee-Carter mortality forecasting: impact on q-forward. Soft Computing, 24: 8553-8567. DOI: 10.1007/s00500-019-04427-z
- 279
- 280 **Levantesi and Pizzorusso(2019).** Levantesi S., Pizzorusso, V. (2019). Application of Machine Learning to Mortality Modeling and Forecasting. Risks, 7(1), 26. ISSN: 2227-9091. DOI:10.3390/risk7010026
- 281
- 282 **Li et al.(2009).** Li, J. S.H., Hardy, M. R., Tan, K. S. (2009). Uncertainty in mortality forecasting: an extension to the classical Lee-Carter approach. Astin Bulletin 39(1), 137-164.
- 283
- 284 **Pollard(1987).** Pollard, J.H. (1987). Projection of age-specific mortality rates. In: Population Bulletin of the United Nations 21/22: 55-69.
- 285
- 286 **Renshaw and Haberman(2006).** Renshaw, A.E. , Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. Insurance: Mathematics and Economics, 38 (3): 556-570.
- 287
- 288 **Richman and Wüthrich(2021).** Richman, R. and Wüthrich, M. (2021). A neural network extension of the Lee-Carter model to multiple populations. Annals of Actuarial Science, 15: 346-366.
- 289
- 290 **Torri(2011).** Torri, T. (2011). Building blocks for a mortality index: an international context. Eur. Actuar. J. 1 (Suppl 1): S127-S141
- 291
- 292 **Villegas et al.(2018).** Villegas, A.M., Kaishev, V. and Millossovich, P. (2018). StMoMo: An R Package for Stochastic Mortality Modelling. Journal of Statistical Software, 84 (3): 1-38.
- 293
- 294 **Willett(1999).** Willett, P. (1999). Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. Journal of Computational Biology, 6 (3-4): 447-457.
- 295

296 **Author contributions**

297 Authors equally contributed to this work.

298 **Data availability**

299 The dataset analyzed during the current study, referred to Human Mortality Database (HMD), is available at [https://www.
300 mortality.org/](https://www.mortality.org/)

301 **Competing interests**

302 The authors declare no competing interests.

303 **Acknowledgements**

304 A preliminary version of this paper was presented at the “10th International Conference IES 2022 Innovation & Society 5.0:
305 Statistical and Economic Methodologies for Quality Assessment”. An extended previous version was published in the Book of
306 short papers of the conference, edited by Rosaria Lombardo, Ida Camminatiello and Violetta Simonacci.