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

Susanna Levantesi[1], **Matteo Lizzi**[1], **and Andrea Nigri**[2*]

[1]Department of Statistics, Sapienza University of Rome, Viale Regina Elena 295-G, 00161 Rome, Italy
[2]Department of Social and Political Sciences, Bocconi University, Milan, Italy
[*]andrea.nigri@unibocconi.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, but it is not clear when a method will perform best. Indeed, new mortality models will appear in the literature but may take years before they can be fully evaluated. As proposed by Booth and Tickle(2008), the accuracy of population estimates should be regularly tested establishing the improvement evidence. Researchers appear to be more focused on technical progress in methods rather than on the accuracy of the estimation they provide, in particular minimizing the bias.

Several approaches have been used to model 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 (for a review see Pollard(1987)). 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 both age patterns and trends over time. This approach includes the more complex stochastic models such as the Lee-Carter Lee and Carter(1992) and, more in general, 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 have suggested additional approaches that have 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, however. In this line, Cairns et al.(2006) 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 basically 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 estimation 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) the mortality rates estimated by the original mortality model. Those authors show that mortality modeling can benefit from machine learning that better captures patterns that traditional models do not identify.

The need for new models that can understand mortality improvements more accurately than canonical tools is evident. This paper contributes to the literature on mortality evaluation by introducing an innovative approach based on machine learning techniques that demographers have not yet explored, contributing to the undervalued field of models assessment. Specifically, we provide an application of the method proposed by[Friedman(2020)], namely Contrast Trees, for evaluating the accuracy of the mortality estimates given by different models that are not treatable with standard validation methods (e.g. those based on the likelihood function). The key significance of the Contrast Trees in mortality evaluation is the supply of a unified approach allowing for assessing and comparing the accuracy of both standard mortality models and machine learning algorithms involved in mortality estimates. 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 mortality rates estimates. Indeed, 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, to ensure the effectiveness of estimation comparing different methods leveraging the proposed approach

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 30-90, analyzing the age groups 30-60 and 61-90 separately to give greater evidence of the differences in mortality that characterizes the younger and older ages. We split the data set into training set and test set according to the common rule 70%-30%. The dataset partition is obtained by using the dissimilarity-based compound selection proposed in[Willett(1999)].

Mortality rate

We calculate the age-specific central death rates for each year t according yo the following formula:

$$m_{x,t} = \frac{D_{x,t}}{E_{x,t}} \tag{1}$$

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 this section, we describe the four models to which the Contrast Trees methodology is applied to evaluate their 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.

In the following, we provide the models' specifications and brief descriptions.

Lee-Carter (LC) model

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

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

We further assume that deaths are independent Poisson distributed[Brouhns et al.(2002)]. 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 trends, with κ_t the time index and β_x modifying the effect of κ_t across ages[1]. The LC model can be reformulated into a Generalized Nonlinear Models (GNM) framework according to the Poisson assumption for death counts.

Age-Period-Cohort (APC)

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

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

Where the regression coefficients $\beta_{1,x}$, $\beta_{2,t}$, $\beta_{3,x+t}$ are the age trend, the period trend and the cohort trend.

Gradient Boosting Machine (GBM)

A tree-based algorithm proposed by[Friedman(2001)] that uses fixed size decision trees as weak learners. The prediction is obtained by a sequential approach, where each decision tree uses the information from the previous one to improve the current 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 weak learner fitted on the model residuals $y - F_m(\mathbf{x})$ with y target variable.

eXtreme Gradient Boosting Machine (XGBM)

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

Contrast Trees

The goal of the Contrast Trees (CTs) method is to uncover regions in the predictor variables space that present very high values of the error rate quantified by a discrepancy measure[Friedman(2020)]. CTs are easy to be interpreted and can be used as diagnostic tools to detect the inaccuracies of models engendered by any learning method.

Suppose to have a set of predictor variables $x = (x_1, x_2, ..., x_p)$ 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, CTs 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:

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

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

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

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

The choice of the discrepancy measure depends on the problem to be solved, allowing CTs 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:

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

[1]The model is subject to the following constraints on κ_t and β_x: $\sum_{t \in \mathscr{T}} \kappa_t = 0 \quad \sum_{x \in \mathscr{X}} \beta_x = 1$. Future mortality rates are obtained by modeling the time index κ_t through an autoregressive integrated moving average (ARIMA) process. In general, a random walk with drift properly fits the data.

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

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 empirical cumulative distributions of y and z. See[Friedman(2020)] for further details about the tree split procedure.
In numerical applications, for sake of simplicity, we use the discrepancy measure $d_m^{[1]}$.

Lack-of-fit contrast curves

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

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

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 the average discrepancy.

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

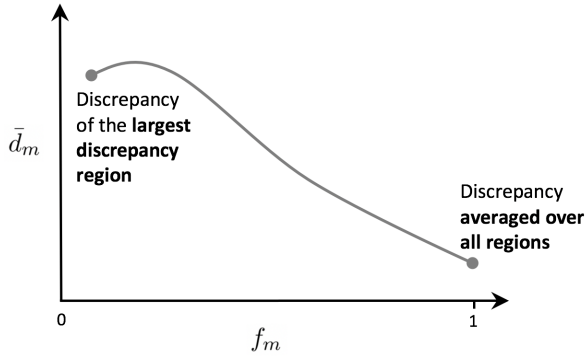


Figure 1. Example of a lack-of-fit contrast curve

Contrast Boosting

To improve the models,[Friedman(2020)] proposes a contrast-boosting strategy that, dealing with the uncovered errors, can enable the regression models to provide more accurate predictions. Contrast Boosting works by gradually modifying a starting value of z to reducing its discrepancy with y over the data. The resulting prediction is then affected by these modifications on the initial value of z. Specifically, we consider the estimation Contrast Boosting, which takes z as an estimate of a parameter of the full conditional distribution of a target variable given a set of predictor variables, $p_y(y|x)$. The procedure consists of modifying 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 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 $z^{(2)} = z + \delta_m^{(2)}$ 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 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 updates $\delta_m^{(k)}$.

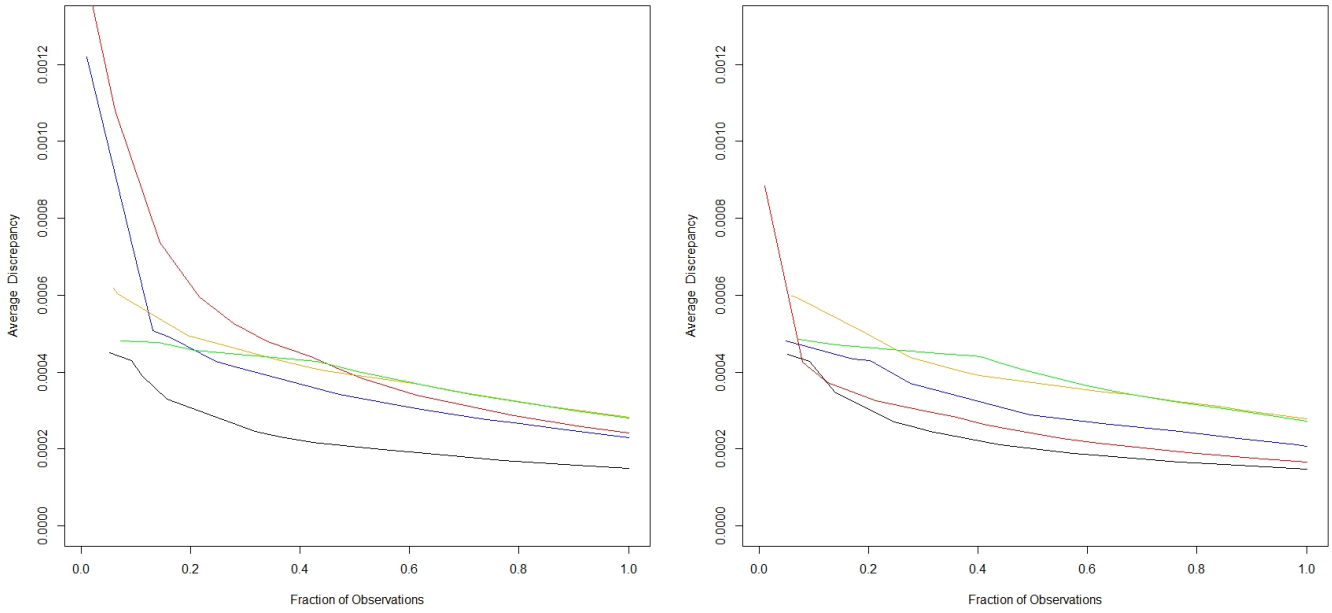
Results

The analyses have been implemented via the *conTree* R package developed by[Friedman and Narasimhan(2020)]. The maximum tree size that corresponds to the number of regions is set to 100. The choice of this parameter is not straightforward as it involves a

trade-off between discrepancy and interpretability. Smaller trees give rise to larger regions defined by simpler conjunctive rules and are thereby easier to interpret. Larger trees have the potential to uncover smaller regions of higher discrepancy defined by more complex rules.

In the following, we will show how to use CTs for assessing the accuracy of the mortality estimates provided by different models. The results obtained on the test set are summarized in the lack-of-fit contrast curves, deduced by contrasting the observed data to the estimates provided by each model. These curves are shown in Fig. 2 and Fig. 3 for the 30-60 and 61-90 age group, respectively. The left panels of these figures refer to the lack-of-fit curves obtained without applying the Contrast Boosting (Base CTs), while the right panels refer to the lack-of-fit curves obtained after applying Contrast Boosting to the output of the models (Boosted CTs).

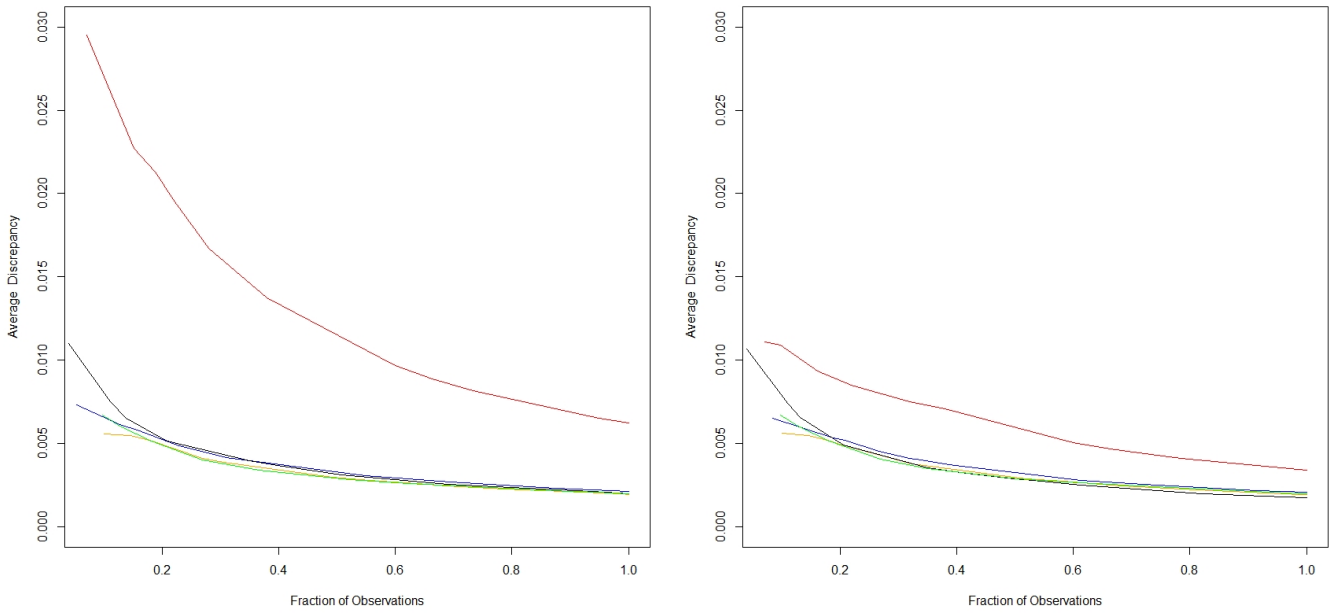
For the 30-60 age group (Fig. 2), the APC model seems to best perform across all regions, as its discrepancy is consistently lower than that of other models. Looking at the XGBM models, as for the effect of preprocessing, it can be observed that the model applied to preprocessed data performs better in the regions with highest average discrepancy in respect to the model applied to raw data. From the scale of the graphs, we can observe that the Contrast Boosting reduces discrepancy across almost all regions for the GBM and LC models, where the relative effect of boosting is particularly evident. For the 61-90 age



(a) Base CTs **(b)** Boosted CTs

Figure 2. Lack-of-fit contrast curves for APC (black), LC (blue), GBM (red), XGBM (orange) and XGBM prep (green). Age 30-60.

group (Fig. 3), the GBM model seems by far the worst performing model. Although the application of the Contrast Boosting significantly reduces the discrepancy, the GBM remains less accurate than the other models. It should also be noted that the effect of Contrast Boosting in high-discrepancy regions for the other models is negligible, except for the APC. Table 1 reports the values of the average discrepancy measure for both the base and the boosted models considered in the analysis. The APC model shows the lowest value of $\bar{d}_m$ for the age group 30-60, in line with the dynamics of the lack-of-fit curves depicted in the left panel of Fig. 2. However, the lack-of-fit curves provide more structured information than the average discrepancy, in particular, regarding how and how much $\bar{d}_m$ varies across the input space. This is clearer when comparing XGBM and APC models for the age group 61-90 in the base model: whereas the average discrepancy of these models differs for 2.6% (0.002010 compared to 0.001960), looking at the lack-of-fit curve depicted in Fig. 3 the XGBM model (orange curve) results to perform better than the APC model (black curve) in the high discrepancy regions. Looking at the base model at age 61-90, the GBM model shows the worst fitting to the observed data. Although the Contrast Boosting produces a strong improvement of the discrepancy, the GBM remains the worst model in terms of discrepancy. For the 30-60 age group in the base model, the values of $\bar{d}_m$ are quite close to each other for all the models, except for the APC model that definitely shows the lowest value. The Contrast Boosting strongly lowers the discrepancy between observed and estimated values for the GBM model. For



(a) Base CTs

(b) Boosted CTs

Figure 3. Lack-of-fit contrast curves for APC (black), LC (blue), GBM (red), XGBM (orange) and XGBM prep (green). Age 61-90.

Model	Age 30-60			Age 61-90		
	Base model	Boosted model	% Change	Base model	Boosted model	% Change
APC	0.000150 (1)	0.000147 (1)	-2%	0.002010 (3)	0.001740 (1)	-13%
LC	0.000229 (2)	0.000208 (3)	-9%	0.002130 (4)	0.002070 (4)	-3%
GBM	0.000241 (3)	0.000167 (2)	-31%	0.006230 (5)	0.003400 (5)	-45%
XGBM	0.000283 (5)	0.000278 (5)	-2%	0.001960 (1)	0.001950 (2)	-1%
XGBM prep	0.000281 (4)	0.000272 (4)	-3%	0.001970 (2)	0.001970 (3)	0%

Table 1. Values of the average discrepancy $\bar{d}_m$

sake of comparison with the average discrepancy, we also calculate the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) on both the base and boosted models. Intuitively, all the three measures, $\bar{d}_m$, RMSE, MAPE, quantify the "distance" between the estimates and the actual observations. However, the average discrepancy whose values are reported in Table 1 is an innovative measure summarizing the discrepancy over all the regions identified by the CTs, while RMSE and MAPE are commonly used error measures calculated on the overall input space without distinguishing by region. We calculate all of them out-of-sample.

By comparing Table 2 showing the values of RMSE and Table 3 showing the values of MAPE with Table 1 reporting the values of the average discrepancy, we note a greater convergence of the error measures in the boosted models rather than in the base models. This result is intuitively straightforward since the boosted models are obtained by just reducing the discrepancy measure. Results of the CTs, comparing the estimates of the models with the real mortality rates, provide the discrepancy measure for each region detected by the CTs. These regions can easily be identified and, if possible, interpreted, providing a further explanation to the model performances. Moreover, high-discrepancy regions can be used for assessing whether or where a model should be trusted or not. We show the heatmap of the error regions in Fig. 4 and Fig. 5 for ages 30-60 and 61-90, respectively. Low discrepancy regions are painted in green, while high discrepancy regions are painted in red. Purple is used, for the sake of image readability, for regions whose discrepancy value exceeds 6e-04 and 0.008 respectively.

For the age group 30-60, the APC model fails to predict mortality rates for advanced ages, as do the two XGBM models. On the contrary, GBM models fail for the most recent cohorts, while the high discrepancy regions relative to the LC model present a more complicated structure. If one compares the results of CTs for base and boosted GBM, the effect of boosting is quite noticeable: while the model performance is still wanting for most recent cohorts, the discrepancy is clearly reduced. A

Model	Age 30-60			Age 61-90		
	Base model	Boosted model	% Change	Base model	Boosted model	% Change
APC	0.000249 (1)	0.000246 (1)	-1%	0.003796 (4)	0.003631 (3)	-4%
LC	0.000352 (2)	0.000332 (3)	-6%	0.003779 (3)	0.003851 (4)	2%
GBM	0.000465 (5)	0.000294 (2)	-37%	0.012344 (5)	0.006111 (5)	-50%
XGBM	0.000379 (4)	0.000379 (5)	0%	0.003070 (2)	0.003063 (2)	0%
XGBM prep	0.000373 (3)	0.000370 (4)	-1%	0.003047 (1)	0.003047 (1)	0%

Table 2. Values of the RMSE

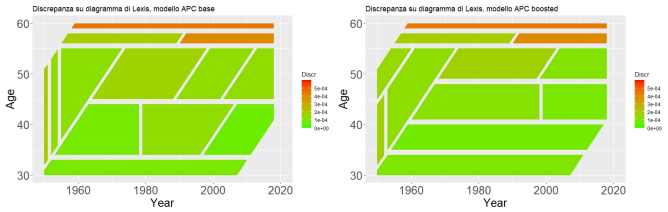
Model	Age 30-60			Age 61-90		
	Base model	Boosted model	% Change	Base model	Boosted model	% Change
APC	4.4% (1)	4.3% (1)	-2%	3.7% (3)	3.1% (1)	-18%
LC	6.9% (3)	5.9% (3)	-15%	4.7% (4)	4.6% (4)	-3%
GBM	10.3% (5)	5.7% (2)	-44%	17.7% (5)	11.6% (5)	-35%
XGBM	7.2% (4)	6.9% (5)	-5%	3.6% (2)	3.5% (3)	-1%
XGBM prep	6.9% (2)	6.5% (4)	-6%	3.5% (1)	3.5% (2)	0%

Table 3. Values of the MAPE

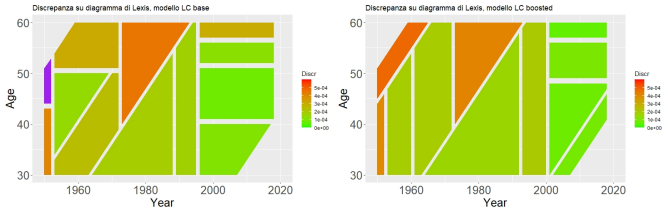
similar effect, though not as evident, is present for LC models.

The main features of the highest discrepancy regions, interestingly, seem to remain similar when looking at the 61-90 group, except for the LC model, whose region structure is now easily understandable: the model fails for advanced ages.

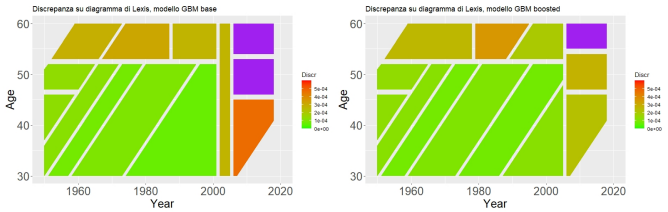
As for preprocessing, it seems from lack-of-fit curves and heatmaps that reduce the amount of discrepancy in low performing regions is reduced for both 30-60 and 61-90 age groups.



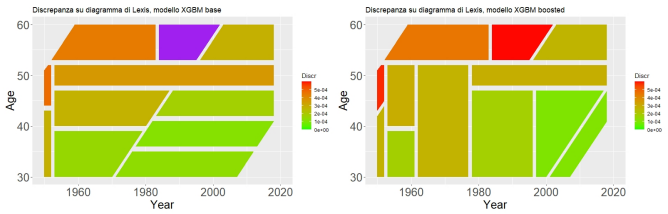
(a) APC (left: base; right: boosted)



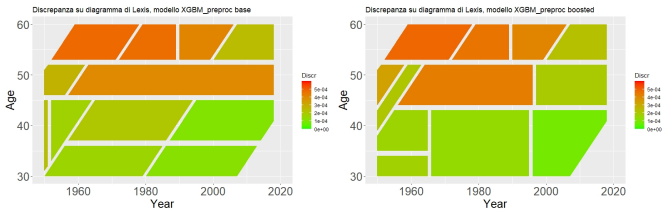
(b) LC(left: base; right: boosted)



(c) GBM (left: base; right: boosted)

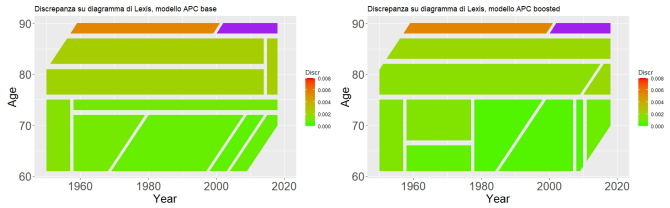


(d) XGBM (left: base; right: boosted)

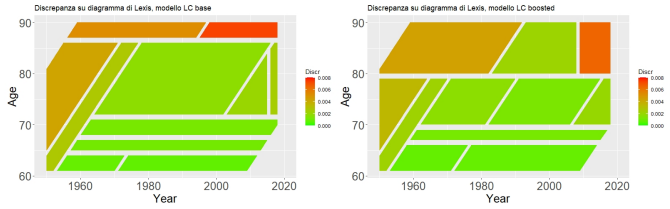


(e) XGM prep (left: base; right: boosted)

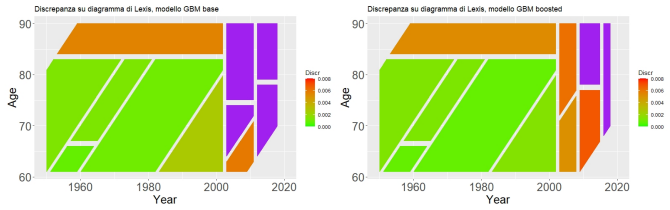
Figure 4. CTs regions, age 30-60.



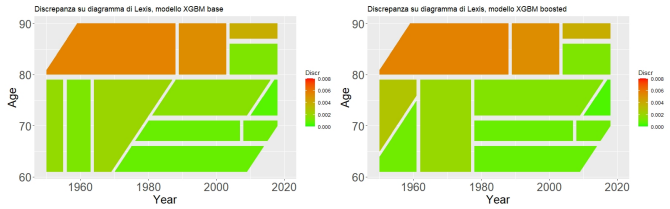
(a) APC (left: base; right: boosted)



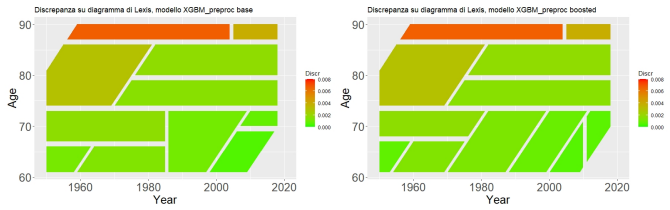
(b) LC (left: base; right: boosted)



(c) GBM (left: base; right: boosted)



(d) XGBM (left: base; right: boosted)



(e) XGM prep (left: base; right: boosted)

Figure 5. CTs regions, age 61-90.

Discussion

Evaluate and thus eventually improve the mortality models fit is crucial for both demographers and actuaries. Indeed, in particular situations, common in actuarial practice, data quality can turn the mortality estimate difficult. A prime example is the case of small subpopulations where the common method such as LC may not guarantee reliable estimation. In this sense, our proposal fills the gap between mortality modeling and model diagnostics, in particular for nontraditional modeling as a machine learning framework CTs consist of a general method based on machine learning that can be applied to any model, expressed as a regression model, to evaluate the goodness of fit and identify the worst-performing regions in the input space. While, other well-known goodness of fit evaluation criteria, such as Bayesian Information Criterion, Akaike Information Criterion, Schwartz Information Criterion, require the likelihood function, which is not available for machine learning models. Therefore, it provides a unified approach for assessing and comparing the accuracy of both traditional models and machine learning algorithms.

The CTs 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 trees structure of CTs enables quantifying the discrepancy between the estimates provided by a model and the actual observations in each region identified by CTs.

References

Alai and Sherris(2014). Alai, D.H., Sherris, M. (2014). Rethinking age-period-cohort mortality trend models. Scandinavian Actuarial Journal, 3: 208-227.

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.(2006). 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.

Deprez et al.(2017). Deprez, P., Shevchenko, P.V., Wüthrich, M.V (2017). Machine learning techniques for mortality modeling. European Actuarial Journal, 7: 337-352. https://doi.org/10.1007/s13385-017-0152-4.

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

Friedman and Narasimhan(2020). Friedman, J.H., Narasimhan, B. (2020). conTree: Contrast Trees and Distribution Boosting. R package version 0.2-8.

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.

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

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

Pollard(1987). Pollard, J.H. (1987). Projection of age-specific mortality rates. In: Population Bulletin of the United Nations 21/22: 55-69.

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.

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.

Willett(1999). Willett, P. (1999). Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. Journal of Computational Biology, 6 (3-4): 447-457.

Author contributions

Authors equally contributed to this work

Data availability

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

Competing interests

The authors declare no competing interests.

Acknowledgements

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