**Revision checklist - Reviewer #1**

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

Submission ID: 51a8fa45-cb85-4c01-ad48-a73538820020

We would like to thank the anonymous reviewer for the helpful comments and valuable suggestions that contributed to improve our paper. We have revised the paper accordingly.

All of the reviewer’s comments and suggestions have been incorporated into the revised paper.

We have provided a response to each of the comments below in blue text.

Referee:

Thanks to the authors for the efforts, explanations, and R-code. This proposal is an exciting extension for modeling mortality, but the main drawbacks remain in the paper.   
On the one hand, the division into three age groups could be more or less, but it helps the interest in machine learning methods. The classical Lee-Carter method was proposed to adjust and predict all ages, which is a strong advantage.  
On the other hand, authors advocate that contrast trees automatically identify the regions in which a given model provides a high error for certain combinations of ages and calendar years, which is easy to interpret. However, using traditional diagnostic methods such as residual diagnosis allows for obtaining the same.  
Finally, the authors should compare models in the log scale (for example, Figure 2).

Authors:

We thank the Reviewer for his positive comment on our paper.

We also thank the reviewer for opening the floor to an interesting discussion on traditional diagnostic methods in mortality modeling. This is a crucial issue useful to mention but with any doubt too wide to address in our work and thus beyond the scope of this paper. Nevertheless, we take this chance to mention the canonical tools, such as the residual diagnosis, and the main difference with respect to the proposed method.

The following text has been included in the final manuscript:

*"In mortality modeling, the objective of diagnostic checking is to ascertain whether the model fits the historical data by obeying an underlying probabilistic hypothesis. This procedure is carried out using residuals diagnosis checking with a Gaussian or more often a Poisson assumption (see, e.g., Renshaw, S. Haberman; Insurance: Mathematics and Economics 38 (2006) 556–570).*

*Leveraging Friedman (2020), who introduces contrast trees to estimate the full conditional probability distribution without any parametric assumptions, we propose a prominent alternative, with particular regard to the intersection of Machine Learning and Mortality modeling fields. In this sense, our proposal fills the gap between mortality modeling and model diagnostics, particularly for nontraditional modeling as a machine learning framework."*

We have now compared the model in the log scale (see Figure 2).

**Revision checklist - Reviewer #2**

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

Submission ID: 51a8fa45-cb85-4c01-ad48-a73538820020

Referee:

The manuscript applies contrast boosting technique to mortality projection models. While this is a valuable task, I am not convinced that this work in its current state is substantial enough to warrant publication for the following reason.

1) The chief goal of mortality projection models is to forecast mortality rates into future. Any assessment of this type of models must go beyond goodness of fit and assess the predictive performance because adherence to past data does not necessarily translate into good prediction. Various illustrations of this can be found in Djeundje et al (2022). For model comparison in this area, goodness of fit assessment must be carried out alongside the resulting predictive performance of the models and underlying uncertainty.  
  
Reference:  
Djeundje et al. (2022) The slowdown in mortality improvement rates 2011–2017: A multi-country analysis.  European Actuarial Journal

Authors: Thank you for this comment. We have provided a better picture of the framework we are proposing. Specifically, our approach is crucial to evaluate the mortality matrix estimation provided by a mortality model and to ensure estimation effectiveness by comparing different methods.

Furthermore, evaluating, and thus eventually improving, the fit of mortality models 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 a common method such as the Lee-Carter may not guarantee reliable estimation. In this sense, our proposal fills the gap between mortality modeling and model diagnostics, particularly for nontraditional modeling as a machine learning framework.

We have added a new sub-section in Materials and Methods to mention the main “Traditional diagnostic tools” used in the literature. We have also mentioned the paper of Djeundje et al. (2022) suggested by the referee.

We have also improved introduction, results, and discussion to better explain the main advantages of Contrast Trees and Contrast Boosting in mortality modelling.

We have clarified that the Contrast Trees technique 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, such as AIC, BIC, and LTR. 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.

Then, we have better specified that the other purpose of our paper is to leverage Contrast Boosting to improve the model’s performance in fitting historical mortality data. To summarize, through this new technique based on Contrast Trees that identify the regions in the predictor variables space that show very high values of the error rate (quantified by a discrepancy measure), we aim to find the best model that fits historical mortality rates by grasping and detecting the inaccuracies of any model and boosting its predictive power.

Finally, following the request of the other reviewer, we have extended the analysis to the 0-29 age group. In downloading the mortality data of this age group, we noticed that the data on the HMD website has been updated. Therefore, we downloaded the updated data also for the 30-60 and 61-90 age groups and redid the models' application entirely. Consequently, the results in the updated version of the paper do not perfectly coincide with those of the original version.