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.

Reviewer 1

Referee:
Based on the aims and scope of the journal, an in-depth revision of the language writing, applications, and conclusions of the paper is needed before possible publication.
As a general comment, this paper investigates a new technique, Contrast Trees, which, leveraging decision trees, provides a general approach for evaluating the quality of fit of different models by detecting the regions in the input space where models work poorly. To verify the ability of this approach, the authors consider standard stochastic mortality models and machine learning algorithms in estimating the Italian mortality rates from the Human Mortality Database. Once the low-performance regions are detected, the authors use Contrast Boosting to improve the inaccuracies of mortality estimates provided by each model. The results are discussed using graphical and numerical tools, particularly for the high-error regions.

1) First, it would be interesting to know how to implement the model in R; maybe the authors can add an Appendix. Implementation details are fascinating.

Authors: The replication scripts, written in the R statistical programming language (R. Core Team, 2013), are hosted on the Open Science Framework (OSF) at https://osf.io/2drhc/?view_only=56b1a1e27238423d8294ee58ca38c4b1

2) One of the principal challenges for mortality models is robustness and residuals with good behaviour; the goodness-of-fit measures are not enough to recommend a model and test the prediction. Therefore, these tools should help improve the models and choose the best one.
In addition, the authors do state the main advantages of this "new technique" compared with other diagnostic tools.

Authors: Thank you for this comment. First, we have mentioned some diagnostic tools that are often used in the literature to assess the goodness-of-fit of a mortality model:

- Analysis of mortality residuals (or standardised mortality residuals) calculated as the difference between the crude estimate of mortality rate by age and year based on observed data and the corresponding estimated mortality rate using a specified mortality model. For example, Cairns et al. (20109 verified that they are consistent with the hypothesis of i.i.d. N(0,1) and have zero correlation both across adjacent ages and across adjacent years.
- Proportion of variance explained (R^2) by the model or the parameters of the model (see, e.g. Bongaarts, 2005)
- Model selection criteria that penalize the log-likelihood with the increase in number of parameters: Akaike Information Criterion (AIC), Schwarz-Bayes Criterion (SBC) (or Bayes Information Criterion (BIC)) and Likelihood-ratio test (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.
- Qualitative model selection criteria: Cairns et al., 2008 provide a list of criteria that might be considered desirable in a mortality model, such as, e.g., ease of implementation,

parsimony, and transparency. Relating to the fitting ability to the observed data, the model should be consistent with historical data, and parameter estimates should be robust relative to the range of data used. For example, Djeundje et al. 2022 consider consistency, stability, and parsimony in addition to standard goodness-of-fit indices (deviance residual, BIC, and residual patterns).

- Checking for the absence of autocorrelation in the residuals of the model by the Portmanteau test (see, e.g., Torri, 2011).

We have included this new part in a specific subsection called “Traditional diagnostic tools ” in the section Materials and Methods.
Secondly, we have added some sentences in the text (see 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, 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.

References

Bongaarts, J. (2005). Long-Range Trends in Adult Mortality: Models and Projection Methods. Demography, 42(1): 23-49.

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.

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

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.

Torri, T. (2011). Building blocks for a mortality index: an international context. Eur. Actuar. J. 1 (Suppl 1): S127-S141

3) Finally, the author(s) should discuss the real advantage of their technique compared to the existing ones. I think the results should be revised in more detail, and conclusions or practical interpretations about them should be indicated in the Conclusions. For example, how do these diagnostic tools help to improve the model projections? How does this result help actuaries in pricing? Do the results depend on the country? Are there different results for different age ranges?

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.

Concerning results, they almost certainly depend on the country (we focus the analysis on the Italian mortality data). However, our paper aims to highlight 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. Therefore, presumably, from country to country as well.
Regarding the age range, we have extended the analysis to the 0-29 age group.

4) Furthermore, please, provide future manuscripts with continuous page and line numbers!

Authors: We have provided the manuscript with continuous page and line numbers.

Particular comments

5) Page 2, the age range should be justified. The challenge is to fit models with an extensive age range with ancient ages. On the other hand, if authors analyze the age groups 30-60 and 61-90 separately, they will obtain very optimistic results. Why not three age groups?

Authors: 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.

6) Page 3, Please explain how x+t represents the cohort.

Authors: We mean "t-x" instead of "x+t". It was a typo that we have corrected in the new version of the paper.

7) Page 5, How is measured accuracy in terms of mxt or log(mxt)? Do results depend on those measures?

Authors: In the first version of the paper, the accuracy of each model was measured by the discrepancy, RMSE, and MAPE using the central mortality rates, mxt. In the revised version of the paper, we have also calculated the error measures on the log(mxt). We thank the referee for this comment that allows us to better evidence the errors at young ages rather than at older ages. Indeed, the error measures on the log(mxt) assign a relatively large weight to errors at young ages, while error measures calculated on the central death rates mxt assign a large weight to errors at older ages. We have discussed this feature in the section Results.

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.