MILLIMAN WHITE PAPER

Improved mortality rate forecasting using machine learning and open data

A multi-population approach to mortality rate forecasting using open data and interpretable neural networks.

Jan Thiemen Postema
Raymond van Es

Milliman

Longevity is a major factor in the profitability of (life) insurers throughout the world. More accurate mortality forecasts could therefore have a substantial impact on their financial results. In this paper we'll investigate using open data and advanced modelling approaches to improve mortality forecasts. We illustrate the performance of the method on mortality for France and the Netherlands.

Currently, the most popular method for evaluating longevity risks is through life tables that are created using statistical models, most notably the Lee-Carter (LC) method [1]. However, ever since such methods have become industry standard, research has continued to find suitable models that more accurately predict these types of risk.

One avenue that is actively being explored by researchers and practitioners throughout the world is that of multiple populations. As Lee & Li point out in their 2005 paper, (interconnected) populations with similar socioeconomic factors are likely to behave similarly and benefit from the same improvements in healthcare [2]. Also, when a developing country has "caught up" with the most advanced countries, it usually starts following the same patterns in terms of mortality rates. Therefore, including statistics from multiple populations can improve the quality of a mortality forecast. In this same paper Lee & Li proposed a novel method that incorporates multiple populations. This strategy has become popular, and many new incarnations of this approach have since been proposed. Jansen provides an extensive overview of advances in this area [3]. However, as Richman & Wüthrich point out, most of the proposed statistical methods for doing multi-population forecasting suffer from the issue of parameter estimation, involving a substantial number of expert judgements [4]. Instead of a statistical model, they propose to use a neural network to extend the LC model to multiple populations.

Such applications of machine learning (ML) algorithms have seen increasing interest in this area. Most of this research focusses on extending or improving statistical models through the use of ML. For example, Levantesi & Pizzorusso and Deprez et al. utilise machine learning methods such as gradient boosted trees to improve the goodness of fit of traditional models [5] [6]. However, there is very little research available that investigates the use of ML models on their own. Levantesi & Pizzorusso theorise that this is because those models are often seen as "black boxes" and are considered difficult to interpret.

More recently, though, some work has been done on exploring ML-only approaches. Some notable examples include Perla et al., who proposed using a shallow convolutional neural network (CNN) [7], and Bravo, who proposed using a long short-term memory (LSTM) neural network [8]. Both were shown to outperform traditional LC models on out-of-sample test sets. The companion Milliman paper to the present one by Elfassihi also explores using neural networks and random forests [9].

However, despite the growing interest, the concerns expressed by Levantesi & Pizzorusso remain regarding the lack of interpretability of ML models in general. In this paper we propose a method that relieves some of those concerns and provide a ML approach to the problem of mortality forecasting that outperforms traditional methods while being explainable. Our strategy consists of training a temporal fusion transformer (TFT) model on multi-population, age-specific mortality data that has been enriched with socioeconomic data collected by the World Bank.

Temporal Fusion Transformer

The TFT is a neural network architecture that has been specifically designed for time series forecasting. This makes it especially suitable for building mortality forecasts. Additionally, due to its use of so-called attention transformers, its predictions are interpretable, without the use of post hoc explainable artificial intelligence (XAI) techniques that usually don't deal well with sequential time series data. This concept, which was first introduced by Vaswani et al. has been the state of the art in natural language processing (NLP) for some time now [10]. It allows the model to learn relationships across multiple time steps.

Furthermore, this model supports past covariates. A major drawback of using covariates in most forecasting models is that the covariates need to be available in the past, as well as the future, which is often not the case without having recourse to an additional forecasting component for such variable. Nevertheless, covariates such as the aforementioned socioeconomic data can still be very valuable in teaching the network the interconnectedness that might exist between the different time series (or in our case countries). For conciseness we won't go into detail on the exact model architecture, and we refer the reader to the original paper by Lim et al. [11] for such detail.

Data

Our aim is to train a model that can accurately predict mortality rates for a variety of countries. To that end, we use the Human Mortality Database (HMD), which provides an extensive set of high-quality mortality data for 41 countries. In our experiments, we use the period from 1960 to 2000 as a training set and the 16-year period from 2000 to 2016 as an out-of-sample validation set. These periods are chosen such that they exclude major worldwide events that could have a substantial impact on the mortality rates, such as World War II and the 2020 COVID-19 pandemic. Even though this data is generally of high quality, there are some countries where the mortality rates are not available over the whole period we consider. Some examples of this include Germany, which maintained two different methodologies for measuring mortality before its reunification, and Israel, where data on the HMD is only available from 1983 onwards. In those cases, we remove the country from the data set, which leaves us with 24 countries. If there are only occasional data points missing for a country (e.g., because a mortality rate is missing for one age group in a certain year), then the mortality rate is interpolated linearly based on the mortality rate in the previous and next years for that same age group.

Finally, as Elfassihi showed, there's significant uncertainty on the mortality force for ages over 95, which is why they are excluded from the data set [9].

The set of mortality data is enriched with open data from the World Bank, which collects a range of socioeconomic factors from the national statistical agencies. We collect all 112 indicators that are available since 1960 and add them to the training data set. Then the pairwise Pearson correlation coefficient between those indicators and the mortality rate is calculated per time series for all countries. This coefficient is used, together with expert judgement, to select the seven most promising variables. In the table in Figure 1 we show the selected variables and their correlations with mortality. Those seven variables are added to the HMD data, which forms our complete training set. If an indicator is missing for a specific country or year it is interpolated linearly. These covariates are then standardised by scaling them to a range from 0 to 1, which is necessary for use in the model.

Model

Our aim is to estimate the future log crude mortality rate ($log(\hat{m}(x,t,g,i))$), at age x in calendar year t, gender g and country i. The crude mortality rate is defined as

$$\frac{\boldsymbol{D(x,t,g,i)}}{\boldsymbol{E(x,t,g,i)}}$$

where $D(x,t,g,i)$ is the number of deaths and $E(x,t,g,i)$ is the central exposure.

Using the log instead of the actual rate helps to diminish the effect of ages with high mortality rates compared to those with low mortality rates, as mortality rates show an exponential increase with age.

TRAINING

To find out what the added benefit is of including extraneous variables in the model, we trained a TFT model both with and without the use of past covariates (i.e., extraneous variables that are only available in the past). In both cases, we run a grid search to find the optimal parameters for the model and evaluate it based on the mean squared error (MSE). The model has been implemented using the open source PyTorch Forecasting package [12].

FIGURE 1: EXTRANEOUS VARIABLES AND THEIR AVERAGE CORRELATIONS TO THE MORTALITY RATE

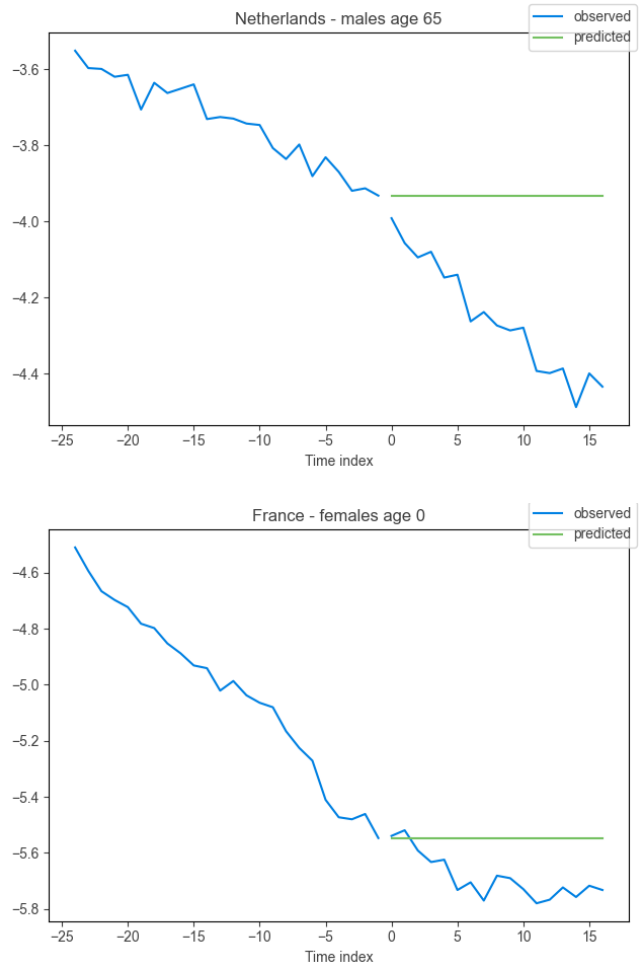
VARIABLE	DESCRIPTION	CORRELATION WITH MORTALITY
FI.RES.TOTL.CD	Total reserves (includes gold, current US$ per 1.000 people)	-0.65
SP.POP.TOTL.MA.ZS	Population, male (% of total population)	0.10
EN.ATM.CO2E.SF.KT.POP	CO2 emissions from solid fuel consumption (kt per 1.000 people)	0.12
SH.MED.PHYS.ZS	Physicians (per 1.000 people)	-0.65
FP.CPI.TOTL.ZG	Inflation, consumer prices (annual %)	0.41
AG.PRD.FOOD.XD	Food production index (2014-2016 = 100)	-0.31
SP.POP.DPND.YG	Age dependency ratio, young (% of working-age population)	0.70

EVALUATION

The performance of the model is evaluated using the MSE, which is calculated on a 16-year out-of-sample period. We calculate this score for four models, a baseline, a LC model, a TFT without open data and a TFT with open data.

The baseline model simply repeats the last observation. The baseline model has a mean MSE of 0.1613 on the evaluation set, which serves as a benchmark for high-order MSE in the comparison. Figure2 shows two examples of the predictions produced by this baseline model. In the first graph, which shows mortality for Dutch males aged 65, the predictions are substantially off, as the mortality rate continues to decline. However, in the second graph, for French female newborns, the predictions are reasonable, as the mortality rate has more or less stopped declining in recent years. In the table in Figure 3 we show the mean and median MSEs of the baseline model, as well as the MSE for two specific countries, France and the Netherlands.

FIGURE 2: THE LOG MORTALITY RATE OBSERVED VS. PREDICTED BY THE BASELINE MODEL



Note: On this time index t=0 refers to the year 2000.

The values found in Figure 3 for the LC model are courtesy of Elfassihi. For more background on how the fitting of this model was performed, please refer to their paper [9]. Because we use a smaller set of countries than Elfassihi, the scores have been updated to reflect this.

When we look at the scores for the TFT models we find that in each situation the TFT model outperforms the baseline and LC models. We also find that adding the open data makes for a substantial improvement on the scores. Even though the standard TFT model already outperforms the traditional LC model, the added socioeconomic data helps the model to predict future mortality rates more accurately.

FIGURE 3: MSE ON THE EVALUATION SET (2000-2016)

MSE	BASELINE	LEE-CARTER	TFT WITHOUT COVARIATES	TFT WITH COVARIATES
Mean	0.1613	0.1540	0.1092	0.0883
Median	0.1536	0.1437	0.1007	0.0886
France	0.0951	0.0686	0.0519	0.0407
The Netherlands	0.1409	0.1129	0.0649	0.0629

INTERPRETATION

Being able to explain what is happening in a mortality forecast is paramount to its successful implementation. One of the main components of a TFT model is the attention transformer. The TFT model uses a modified version of the Multi-Head Attention framework that aggregates the weights across multiple heads to give an indication of a feature's importance.

We can look at the attention on different levels. Figure 4 shows the aggregated attention over all predictions. Here we find that, in general, the most recent years carry the most weight, which is in line with what would be expected.

FIGURE 4: THE ATTENTION PER HISTORICAL TIME STEP

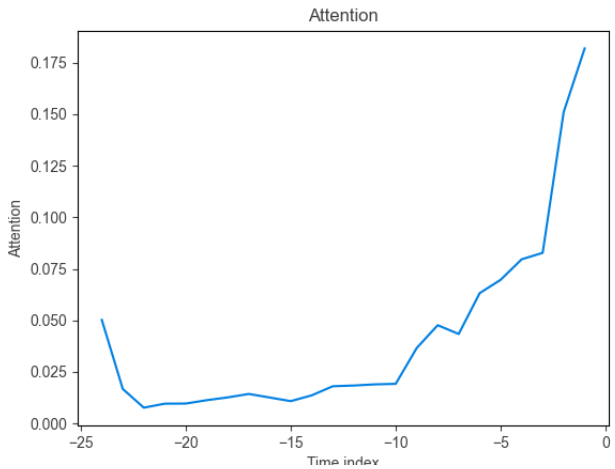


Figure 5 and Figure 6 show predictions for individual time series based on the TFT model with covariates. These graphs consist of the observed values in blue and the predicted values in green. We note that the TFT outputs a likelihood, with (in our case) four quantiles. Calculating the error measures is done on the central estimate, which is depicted by the green line in the plots. These figures also show the attention per time step in grey. Here we can see that the attention differs per time series.

FIGURE 5: PREDICTIONS FOR FRENCH NEWBORN FEMALES

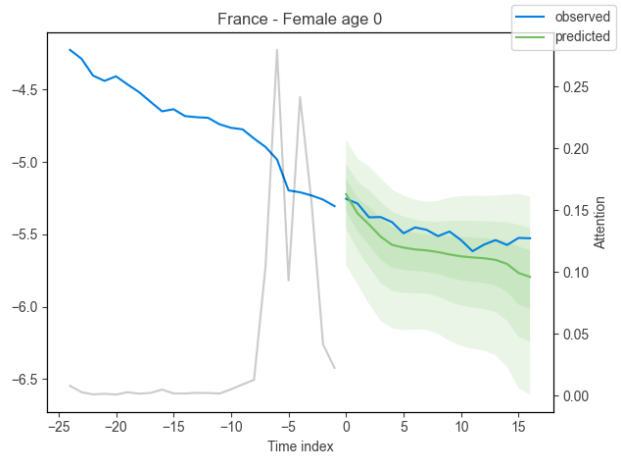
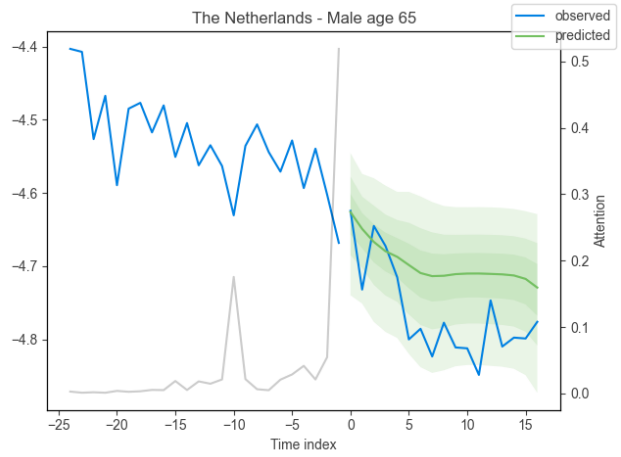


FIGURE 6: PREDICTIONS FOR DUTCH MALES AGED 65



Another aspect of the TFT that improves the interpretability is the feature selection network. Using this network, we can get some insight into the feature importance. In

Figure the importance of the static variables is shown. Here we see that country is the most important static indicator. In Figure 8 the feature importance of the encoder variables is shown. As one would expect, the most important feature in the training set is the “logmx” (the log of the crude mortality rate) in the past for all countries. The feature is followed in importance by the number of physicians per 1.000 people and the CO2 emissions from solid fuel consumption. This indicates that these features might be important when comparing and predicting the mortality rates in different countries.

FIGURE 7: THE FEATURE IMPORTANCE FOR THE STATIC VARIABLES

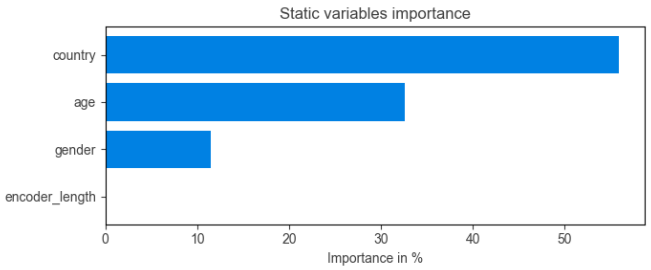
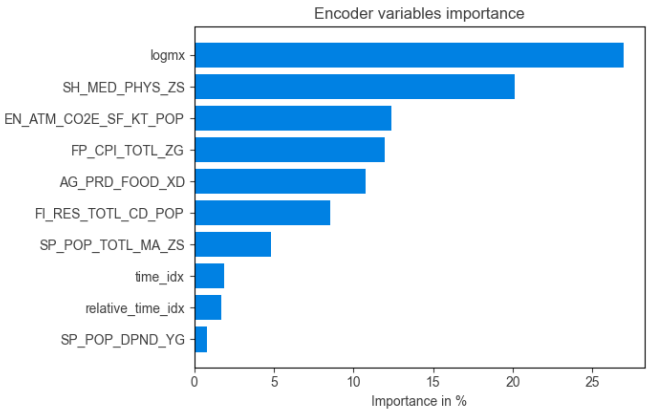


FIGURE 8: THE FEATURE IMPORTANCE OF THE COVARIATES



Conclusion

Using state-of-the-art machine learning algorithms can substantially improve the forecasts of mortality rates, when compared to traditional techniques such as the Lee-Carter model, while still being interpretable. Using such techniques can further help to challenge or refine best estimate mortality tables by including exogenous factors in the modelling, as well as improving risk modelling for economic capital by better taking into account dependencies between countries and/or between mortality and other risk factors, such as financial risks.

CONTACT

Jan Thiemen Postema
janthiemen.postema@milliman.com

Raymond van Es
raymond.vanes@milliman.com

REFERENCES

[1] R. D. Lee and L. R. Carter, "Modeling and Forecasting U.S. Mortality," *Journal of the American Statistical Association,* vol. 87, no. 419, p. 659, 1992.

[2] N. Li and R. Lee, "Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method," *Demography,* 2005.

[3] F. Janssen, "Advances in mortality forecasting: introduction," *Genus,* vol. 74, no. 1, p. 21, 2018.

[4] R. Richman and M. V. Wüthrich, "A neural network extension of the Lee–Carter model to multiple populations," *Annals of Actuarial Science,* vol. 15, no. 2, p. 346, 2021.

[5] S. Levantesi and V. Pizzorusso, "Application of Machine Learning to Mortality Modeling and Forecasting," *Risks,* vol. 7, no. 1, p. 26, 2019.

[6] P. Deprez, P. V. Shevchenko and M. V. Wüthrich, "Machine Learning Techniques for Mortality Modeling," *European Actuarial Journal,* vol. 7, no. 2, p. 337, 2017.

[7] F. Perla, R. Richman, S. Scognamiglio and M. V. Wüthrich, "Time-series forecasting of mortality rates using deep learning," *Scandinavian Actuarial Journal,* vol. 2021, no. 7, p. 572, 2021.

[8] J. M. Bravo, "Forecasting mortality rates with Recurrent Neural Networks: A preliminary investigation using Portuguese data," in *CAPSI 2021*, 2021.

[9] A. Elfassihi, "Longevity trend prediction using Machine Learning," Milliman White Paper, 2020.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All you Need," in *NeurIPS*, 2017.

[11] B. Lim, S. Ö. Arik, N. Loeff and T. Pfister, "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting,* vol. 37, no. 4, p. 1748, 2021.

[12] Beitner, J. (19 September 2020). Introducing PyTorch Forecasting, Retrieved 18 November 2022 from https://towardsdatascience.com/introducing-pytorch-forecasting-64de99b9ef46.