Point and Interval Forecasts of Death Rates Using Neural Networks

Simon Schnürch, Ralf Korn
IAALS Webinar, 12 April 2022
Longevity risk has a long history.

“People always live forever when there is any annuity to be paid them. [...] An annuity is a very serious business; it comes over and over every year, and there is no getting rid of it.”

— Jane Austen, Sense and Sensibility
All over the world, mortality has decreased a lot – often more than expected.

![Heat maps of log death rates for males over age and time (blue = low, red = high).](image)
Existing stochastic mortality models are not optimal in every situation.

Some standard multi-population mortality models

\[
\log m_{x,t}^i = \alpha_x^i + \beta_x^i \kappa_t^i, \quad \text{(Lee and Carter [1992], LC)}
\]

\[
\log m_{x,t}^i = \alpha_x^i + \beta_x \kappa_t + \beta_x^i \kappa_t^i, \quad \text{(Li and Lee [2005], ACF)}
\]

\[
\log m_{x,t}^i = \alpha_x + \beta_1^i \kappa_t^{i,1} + \cdots + \beta_k^i \kappa_t^{i,k}. \quad \text{(Wen et al. [2021])}
\]

with

- death rate \( m_{x,t}^i = \frac{\text{#(Deaths)}}{\text{Exposure}} \) for age \( x \) in population \( i \) and year \( t \),
- basic age structure of mortality \( \alpha_x^i \),
- period effects \( \kappa_t, \kappa_t^i \),
- age effects \( \beta_x, \beta_x^i \).
Existing stochastic mortality models are not optimal in every situation.

Some standard multi-population mortality models

\[
\log m_{x,t}^i = \alpha_x^i + \beta_x \kappa_t^i, \quad \text{(Lee and Carter [1992], LC)}
\]

\[
\log m_{x,t}^i = \alpha_x^i + \beta_x \kappa_t^i + \beta_x^i \kappa_t^i, \quad \text{(Li and Lee [2005], ACF)}
\]

\[
\log m_{x,t}^i = \alpha_x + \beta_1 \kappa_t^1 + \cdots + \beta_k \kappa_t^k. \quad \text{(Wen et al. [2021])}
\]

with

- death rate \( m_{x,t}^i = \frac{\#(\text{Deaths})}{\text{Exposure}} \) for age \( x \) in population \( i \) and year \( t \),
- basic age structure of mortality \( \alpha_x^i \),
- period effects \( \kappa_t^i, \kappa_t^i \),
- age effects \( \beta_x, \beta_x^i \).

Figure: Death rate forecasts for years 2007–2016, ages 60, 65, 71, 77, 83, 89, female population of England & Wales of LC models (trained on 10 or 20 years of data).
Previous research shows that machine learning can improve forecasts.

Previous applications of machine learning to mortality modeling and forecasting include

- Poisson regression tree boosting [Deprez et al., 2017, Levantesi and Pizzorusso, 2019, Levantesi and Nigri, 2019],
- cluster analysis [Hatzopoulos and Haberman, 2013, Debón et al., 2017, Wen et al., 2020, Schnürch et al., 2021],
- feed-forward neural networks (FFNNs) [Shah and Guez, 2009, Richman and Wüthrich, 2021],
- recurrent neural networks (RNNs) [Nigri et al., 2019, Richman and Wüthrich, 2019, Petneházi and Gáll, 2019].

→ We consider FFNNs [Richman and Wüthrich, 2021] and RNNs [Richman and Wüthrich, 2019] as benchmarks.
We propose a neural network architecture for point and interval death rate forecasts.

Accurate mortality forecasts

- are important for pension plans, insurance companies, governments, ...
- are not always achieved by classical methods,
- should give an impression of the possible distribution of future mortality rates.

We propose

- a convolutional neural network (CNN) for mortality forecasting
- along with a reliable method for quantifying its prediction uncertainty, and
- compare its performance to other neural network architectures and the LC and ACF models.
We propose a neural network architecture for point and interval death rate forecasts.

Accurate mortality forecasts

- are important for pension plans, insurance companies, governments, ...
- are not always achieved by classical methods,
- should give an impression of the possible distribution of future mortality rates.

We propose

- a convolutional neural network (CNN) for mortality forecasting
- along with a reliable method for quantifying its prediction uncertainty, and
- compare its performance to other neural network architectures and the LC and ACF models.
Feed-forward neural networks have been applied successfully.

**Definition (FFNN)**

An FFNN is a collection of connected layers

\[ \varphi^l : \mathbb{R}^{n_{l-1}} \rightarrow \mathbb{R}^{n_l}, Z^{l-1} \mapsto Z^l := \sigma^l \left( W^l Z^{l-1} + b^l \right), \]

where \( l = 1, \ldots, L \).

- Trainable parameters in weight matrices \( W^l \), bias vectors \( b^l \)
- Non-linear activation function \( \sigma^l \)
- Inputs: Age, country, gender (categorical, embedded) and year (numerical)
- Adding to the contribution of Richman and Wüthrich [2021], we optimize over \( \sim 45000 \) combinations of hyperparameters via 3-fold cross validation.
- Bootstrap ensemble of 100 FFNNs to stabilize results
Recurrent neural networks are better suited for time series prediction.

- Main characteristic of RNNs: Internal state for remembering historical observations

- Long short-term memory [LSTM, Hochreiter and Schmidhuber, 1997] architecture avoids problems with vanishing gradients

- Nigri et al. [2019] combine LC and LSTM for mortality rate forecasting

- Richman and Wüthrich [2019] directly apply LSTM
  - Inputs: Age, country, gender, historical death rates
  - We re-implement their approach, comparing ~2000 hyperparameter combinations via cross validation
  - Bootstrap ensemble (size: 10) is used as well (see FFNNs)
We propose CNNs for capturing two-dimensional structure in mortality rates.

- Bai et al. [2018]: CNNs are better suited for certain sequential prediction tasks than RNNs.

- We propose to use a bootstrap ensemble of CNNs for forecasting based on two-dimensional patterns in mortality data.

- A similar approach has been successfully investigated by Perla et al. [2021].

Figure: Heat map displaying the death rates of the female population of England & Wales between 1991 and 2000 for ages 0 to 100. (blue = low, red = high).
CNNs contain different layers for different tasks.

### CNN layer types

**Convolutional:**
\[
Z^l_{k,i,j} = \sigma^l \left( \sum_{r=1}^{n_l-1} \sum_{p=1}^{f} \sum_{q=1}^{f} W^l_{k,r,p,q} Z^{l-1}_{r,i+p-1,j+q-1} + b_k \right).
\]

**Max pooling:**
\[
Z^l_{k,i,j} = \max_{p=(i-1)n_p+1,...,i\cdot n_p, q=(j-1)n_p+1,...,j\cdot n_p} Z^{l-1}_{k,p,q}.
\]

**Dense:**
\[
Z^l = \sigma^l \left( W^l Z^{l-1} + b^l \right).
\]

**Figure:** Schematic illustration of a CNN consisting of a convolutional, a pooling and a dense layer followed by another dense output layer.
Our CNN can be interpreted in terms of an LC-type modeling approach (see Perla et al. [2021]).

For fixed $x \in X_{\text{out}}, \ t \in T, \ i \in P$ our model reads

$$\log m_{x,t}^i = b_x + \langle (W_{x,j})^\top_{j=1,...,k}, Z_t^i \rangle,$$

where $Z_t^i \in \mathbb{R}^k$ is a non-linear function of $(m_{x,t}^i)_{x \in X_{\text{in}}, \ t=t-\tau,...,t-1}$. In this sense, our model generalizes the common age effect model considered by Wen et al. [2021],

$$\log m_{x,t}^i = \alpha_x + \sum_{j=1}^k \beta_{x}^{j} \kappa_{t}^{i,j},$$

with

$$\alpha_x \triangleq b_x,$$

$$\beta_{x}^{1}, ..., \beta_{x}^{k} \triangleq (W_{x,j})_{j=1,...,k},$$

$$\kappa_{t}^{i,1}, ..., \kappa_{t}^{i,k} \triangleq Z_t^i.$$
Our CNN can be interpreted in terms of an LC-type modeling approach (see Perla et al. [2021]).

For fixed $x \in X_{\text{out}}$, $t \in T$, $i \in P$ our model reads

$$\log m^i_{x,t} = b_x + \langle (W_{x,j})_{j=1,...,k}^\top, Z^i_t \rangle,$$

where $Z^i_t \in \mathbb{R}^k$ is a non-linear function of $(m^i_{\tilde{x},\tilde{t}})_{\tilde{x} \in \chi_m, \tilde{t} = t-\tau,...,t-1}$. In this sense, our model generalizes the common age effect model considered by Wen et al. [2021],

$$\log m^i_{x,t} = \alpha_x + \sum_{j=1}^k \beta^i_{x,k} \kappa^i_{t,j},$$

with

$$\alpha_x \triangleq b_x,$$
$$\beta^1_{x}, \ldots, \beta^k_{x} \triangleq (W_{x,j})_{j=1,...,k},$$
$$\kappa^i_{t,1}, \ldots, \kappa^i_{t,k} \triangleq Z^i_t.$$
In contrast to previous neural network applications in mortality forecasting, we include prediction uncertainty.

Goal:

\[
P \left( \hat{m}_{x,t}^{lower} \leq m_{x,t} \leq \hat{m}_{x,t}^{upper} \right) \geq a \text{ for some large } a \in [0, 1].
\]

Assumption:

\[
\log m_{x,t} = \log m_{x,t}^{true} + \varepsilon_{x,t}.
\]

Bias-variance decomposition:

\[
\mathbb{E} \left( (\log m_{x,t} - \log \hat{m}_{x,t})^2 \right) = \text{Bias} \left( \log \hat{m}_{x,t} \right)^2 + \text{Var} \left( \log \hat{m}_{x,t} \right) + \text{Var} \left( \varepsilon_{x,t} \right).
\]

We follow the approach proposed by Heskes [1997] for general FFNNs:

- Assume \( \text{Bias}(\log \hat{m}) \equiv 0. \)
- Estimate model uncertainty \( \text{Var} \left( \log \hat{m}_{x,t} \right) \) based on ensemble variance.
- Train an additional neural network to estimate noise variance \( \text{Var} \left( \varepsilon_{x,t} \right). \)
- This yields a variance estimator \( (\hat{s}_{x,t}^i)^2 \), which leads under a normal assumption to the interval bounds

\[
\log \hat{m}_{x,t}^{lower|upper} := \log \hat{m}_{x,t} \pm \phi^{-1} \left( \frac{1 + a}{2} \right) \hat{s}_{x,t}^i.
\]
In contrast to previous neural network applications in mortality forecasting, we include prediction uncertainty.

Goal:
\[
P\left(\hat{m}_{x,t}^{i,\text{lower}} \leq m_{x,t}^{i} \leq \hat{m}_{x,t}^{i,\text{upper}}\right) \geq a \text{ for some large } a \in [0, 1].
\]

Assumption:
\[
\log m_{x,t}^{i} = \log m_{x,t}^{i,\text{true}} + \varepsilon_{x,t}^{i}.
\]

Bias-variance decomposition:
\[
\mathbb{E}\left(\left(\log m_{x,t}^{i} - \log \hat{m}_{x,t}^{i}\right)^2\right) = \text{Bias}\left(\log \hat{m}_{x,t}^{i}\right)^2 + \text{Var}\left(\log \hat{m}_{x,t}^{i}\right) + \text{Var}\left(\varepsilon_{x,t}^{i}\right).
\]

We follow the approach proposed by Heskes [1997] for general FFNNs:
- Assume \text{Bias}\left(\log \hat{m}\right) \equiv 0.
- Estimate model uncertainty \text{Var}\left(\log \hat{m}_{x,t}^{i}\right) based on ensemble variance.
- Train an additional neural network to estimate noise variance \text{Var}\left(\varepsilon_{x,t}^{i}\right).
- This yields a variance estimator \left(\hat{s}_{x,t}^{i}\right)^2, which leads under a normal assumption to the interval bounds
\[
\log \hat{m}_{x,t}^{i,\text{lower}|upper} := \log \hat{m}_{x,t}^{i} \pm \Phi^{-1}\left(\frac{1 + a}{2}\right)\hat{s}_{x,t}^{i}.
\]
In contrast to previous neural network applications in mortality forecasting, we include prediction uncertainty.

**Goal:**

\[ \mathbb{P} \left( \hat{m}_{x,t}^{i,\text{lower}} \leq m_{x,t}^{i} \leq \hat{m}_{x,t}^{i,\text{upper}} \right) \geq a \text{ for some large } a \in [0, 1]. \]

**Assumption:**

\[ \log m_{x,t}^{i} = \log m_{x,t}^{i,\text{true}} + \varepsilon_{x,t}^{i}. \]

**Bias-variance decomposition:**

\[ \mathbb{E} \left( \left( \log m_{x,t}^{i} - \log \hat{m}_{x,t}^{i} \right)^2 \right) = \text{Bias} \left( \log \hat{m}_{x,t}^{i} \right)^2 + \text{Var} \left( \log \hat{m}_{x,t}^{i} \right) + \text{Var} \left( \varepsilon_{x,t}^{i} \right). \]

We follow the approach proposed by Heskes [1997] for general FFNNs:

- Assume \( \text{Bias}(\log \hat{m}) \equiv 0 \).
- Estimate model uncertainty \( \text{Var} \left( \log \hat{m}_{x,t}^{i} \right) \) based on ensemble variance.
- Train an additional neural network to estimate noise variance \( \text{Var} \left( \varepsilon_{x,t}^{i} \right) \).
- This yields a variance estimator \( \left( \hat{s}_{x,t}^{i} \right)^2 \), which leads under a normal assumption to the interval bounds

\[ \log \hat{m}_{x,t}^{i,\text{lower}|\text{upper}} := \log \hat{m}_{x,t}^{i} \pm \Phi^{-1} \left( \frac{1 + a}{2} \right) \hat{s}_{x,t}^{i}. \]
We evaluate the forecasts on real data based on different error measures.

- Data: Human Mortality Database [2019]

- Models (LC10, LC20, ACF, FFNN, RNN, CNN) calibrated on years up to 2006, ages 0 to 100

- Out-of-sample evaluation on the years 2007 to 2016, ages 60 to 89, 54 populations

- Error measures: MSE, MAE, MdAPE for point forecasts

\[
\text{MSE} := \frac{1}{N} \sum_{x,t,i} \left( \hat{m}_{x,t}^i - m_{x,t}^i \right)^2, \quad \text{MAE} := \frac{1}{N} \sum_{x,t,i} \left| \hat{m}_{x,t}^i - m_{x,t}^i \right|, \quad \text{MdAPE} := \text{median}_{x,t,i} \left\{ \frac{\left| \hat{m}_{x,t}^i - m_{x,t}^i \right|}{m_{x,t}^i} \right\} \cdot 100\%,
\]

and PICP, MPIW for interval forecasts

\[
\text{PICP} := \frac{1}{N} \sum_{x,t,i} \mathbb{1}\{m_{x,t}^i \in [\hat{m}_{x,t}^{i,\text{lower}}, \hat{m}_{x,t}^{i,\text{upper}}]\}, \quad \text{MPIW} := \frac{1}{N} \sum_{x,t,i} \left( \hat{m}_{x,t}^{i,\text{upper}} - \hat{m}_{x,t}^{i,\text{lower}} \right).
\]
The CNN yields good point forecasts and reliable interval forecasts.

- The FFNN performs well with respect to squared and absolute error but rather weakly in terms of the relative error. Its PICP is above 95%, as required.

- The CNN minimizes MAE and MdAPE and has a high PICP.

- In a robustness check on a 20-year forecasting period, the CNN clearly outperforms all other models.

### Table: Out-of-sample error measures for 54 populations, ages 60 to 89 and years 2007 to 2016 (models trained on data up to 2006).

| Model | MSE×10^5 | MAE×10^3 | MdAPE[%] | PICP[%] | MPIW |
|-------|----------|-----------|-----------|---------|------|
| LC10  | 4.9      | 3.7       | 5.7       | 74.0    | 0.012|
| LC20  | 5.5      | 4.0       | 5.8       | 74.3    | 0.011|
| ACF   | 3.4      | 3.3       | 5.5       | 77.2    | 0.010|
| FFNN  | 2.6      | 3.0       | 5.9       | 97.0    | 0.017|
| RNN   | 5.9      | 4.2       | 6.1       | 86.0    | 0.015|
| CNN   | 2.9      | 3.0       | 4.9       | 98.0    | 0.019|
Looking at specific errors reveals additional information.

Figure: MdAPE by age / year.

Figure: MdAPE by population.
The CNN achieves better point forecasts than the LC models for England & Wales.

Figure: Model forecasts and CNN prediction intervals compared to ground truth for English & Welsh females from 2007 to 2016 for ages 60, 65, 71, 77, 83, 89.

© Fraunhofer ITWM
12 April 2022 16
A surrogate model helps visualizing the global model structure.

We observe a generally sensible behavior of the CNN’s fitted death rates:

- Log-linear increase with age,
- mostly reasonable country-specific effects,
- higher fitted death rates for males,
- slow decrease over time.

Figure: Coefficients of a linear global surrogate model trained on the logarithmized fitted values of the CNN, $R^2 = 0.9698$. 
Long-term forecasts of the CNN look plausible.

Figure: Forecasts for the male (left) and female (right) populations of England & Wales for the years 2007, 2016, 2026, 2036 (including prediction intervals) and ages 60–89.
Calculating annuity values gives an impression of model risk.

Present values of 30-year temporary life annuities starting in 2007 calculated with different models, interest rate 0.9%.

| Population          | LC30 | ACF | CNN | FFNN | RNN | DAV2004R |
|---------------------|------|-----|-----|------|-----|----------|
| Lithuania (M)       | 13.3 | 13.6| 14.1| 15.4 | 13.3| –        |
| England & Wales (F) | 21.1 | 21.4| 21.1| 22.0 | 21.6| –        |
| West Germany (F)    | 21.7 | 21.8| 21.3| 22.1 | 22.3| 23.0     |
| West Germany (M)    | 19.2 | 19.3| 19.5| 19.7 | 19.2| 21.2     |
There is potential for further improvement.

Possible extensions and improvement ideas for our model include

- check assumptions made for prediction uncertainty evaluation (normal distribution assumption; out-of-sample bias $\approx 0$)
- stacking, i.e., setting up combinations of model architectures (e.g., FFNN and CNN),
- further investigating explainability of the CNN (e.g., via SHAP).

For more details, see the article at doi:10.1017/asb.2021.34 or contact me at simon.schnuerch@itwm.fraunhofer.de.
References (1)

S. Bai, J. Z. Kolter, and V. Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. Preprint, 2018. URL http://arxiv.org/pdf/1803.01271v2.

A. Debón, L. Chaves, S. Haberman, and F. Villa. Characterization of between-group inequality of longevity in European Union countries. Insurance: Mathematics and Economics, 75:151–165, 2017. doi:10.1016/j.insmatheco.2017.05.005.

P. Deprez, P. V. Shevchenko, and M. V. Wüthrich. Machine learning techniques for mortality modeling. European Actuarial Journal, 7(2):337–352, 2017. ISSN 2190-9733. doi:10.1007/s13385-017-0152-4.

P. Hatzopoulos and S. Haberman. Common mortality modeling and coherent forecasts. An empirical analysis of worldwide mortality data. Insurance: Mathematics and Economics, 52(2):320–337, 2013. doi:10.1016/j.insmatheco.2012.12.009.

T. Heskes. Practical confidence and prediction intervals. In M. I. Jordan, M. C. Mozer, and T. Petsche, editors, Advances in Neural Information Processing Systems, volume 9, pages 176–182, Cambridge, MA, 1997. MIT Press. ISBN 0262100677.

S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. Neural Computation, 9(8):1735–1780, 1997. ISSN 0899-7667. doi:10.1162/neco.1997.9.8.1735.

Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research, Rostock (Germany). Data downloaded on July 2, 2019. URL www.mortality.org.

R. D. Lee and L. R. Carter. Modeling and Forecasting U.S. Mortality. Journal of the American Statistical Association, 87(419):659–671, 1992. ISSN 0162-1459. doi:10.1080/01621459.1992.10475265.

S. Levantesi and A. Nigri. A random forest algorithm to improve the Lee–Carter mortality forecasting: impact on q-forward. Soft Computing, 2(2):146, 2019. ISSN 1432-7643. doi:10.1007/s00500-019-04427-z.

S. Levantesi and V. Pizzorusso. Application of Machine Learning to Mortality Modeling and Forecasting. Risks, 7(1):26, 2019. doi:10.3390/risks7010026.

N. Li and R. D. Lee. Coherent Mortality Forecasts for a Group of Populations: An Extension of the Lee-Carter Method. Demography, 42(3):575–594, 2005. doi:10.1353/dem.2005.0021.
A. Nigri, S. Levantesi, M. Marino, S. Scognamiglio, and F. Perla. A Deep Learning Integrated Lee–Carter Model. *Risks*, 7(1):33, 2019. doi:10.3390/risks7010033.

F. Perla, R. Richman, S. Scognamiglio, and M. V. Wüthrich. Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal*, 7:572–598, 2021. ISSN 0346-1238. doi:10.1080/03461238.2020.1867232.

G. Petneházi and J. Gáll. Mortality rate forecasting: can recurrent neural networks beat the Lee-Carter model? Preprint, 2019. URL http://arxiv.org/pdf/1909.05501v2.

R. Richman and M. V. Wüthrich. Lee and Carter go Machine Learning: Recurrent Neural Networks. Tutorial, 2019. URL https://ssrn.com/abstract=3441030.

R. Richman and M. V. Wüthrich. A Neural Network Extension of the Lee-Carter Model to Multiple Populations. *Annals of Actuarial Science*, 15(2):346–366, 2021. ISSN 1748-4995. doi:10.1017/S1748499519000071.

S. Schnürch and R. Korn. Point and Interval Forecasts of Death Rates Using Neural Networks. *ASTIN Bulletin*, 52(1):333–360, 2022. ISSN 0515-0361. doi:10.1017/asb.2021.34.

S. Schnürch, T. Kleinow, and R. Korn. Clustering-Based Extensions of the Common Age Effect Multi-Population Mortality Model. *Risks*, 9(3):45, 2021. doi:10.3390/risks9030045.

P. Shah and A. Guez. Mortality forecasting using neural networks and an application to cause-specific data for insurance purposes. *Journal of Forecasting*, 28(6):535–548, 2009. ISSN 0277-6693. doi:10.1002/for.1111.

J. Wen, T. Kleinow, and A. J. G. Cairns. Trends in Canadian Mortality by Pension Level: Evidence from the CPP and QPP. *North American Actuarial Journal*, 41:1–29, 2020. doi:10.1080/10920777.2019.1679190.

J. Wen, A. J. G. Cairns, and T. Kleinow. Fitting Multi-Population Mortality Models to Socio-Economic Groups. *Annals of Actuarial Science*, 15(1):144–172, 2021. ISSN 1748-4995. doi:10.1017/S1748499520000184.