NEIGHBOURING PREDICTION FOR MORTALITY

BY

CHOU-WEN WANG, JINGGONG ZHANG AND WENJUN ZHU

ABSTRACT

We propose a new neighbouring prediction model for mortality forecasting. For each mortality rate at age x in year t, $m_{x,t}$, we construct an image of neighbourhood mortality data around $m_{x,t}$, that is, $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$, which includes mortality information for ages in $[x - x_1, x + x_2]$, lagging k years $(1 \leq k \leq s)$. Combined with the deep learning model – convolutional neural network, this framework is able to capture the intricate nonlinear structure in the mortality data: the neighbourhood effect, which can go beyond the directions of period, age, and cohort as in classic mortality models. By performing an extensive empirical analysis on all the 41 countries and regions in the Human Mortality Database, we find that the proposed models achieve superior forecasting performance. This framework can be further enhanced to capture the patterns and interactions between multiple populations.

KEYWORDS

Mortality forecasting, neighbourhood effect, artificial intelligence, deep learning, convolutional neural networks, longevity risk.

JEL codes: C45, C51, C52, C53, G22, J11.

*We thank Mario V. Wüthrich (editor) and three anonymous referees for helpful comments. Wang acknowledges the support of MOST (107-2410-H-110-010-MY3). Zhang thanks the research funding support from the Nanyang Technological University Start-up Grant (04INS000509C300) and the Ministry of Education Academic Research Fund Tier 1 Grant (RG55/20). Zhu also thanks the research funding support from the Nanyang Technological University Start-Up Grant (04INS000384C300), Singapore Ministry of Education Academic Research Fund Tier 1 (RG143/19), and the Society of Actuaries Education Institution Grant.

Astin Bulletin 51(3), 689-718. doi:10.1017/asb.2021.13 © The Author(s), 2021. Published by Cambridge University Press on behalf of The International Actuarial Association.

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

1. Introduction

As the populations of the world's leading economies age, longevity risk has become a high-profile risk in recent years, which threatens the stability of the global financial system and brings unexpected economic challenges to individuals, corporations, and governments. Meanwhile, rapid advances in Artificial Intelligence (AI) are changing the environment in which we live. Advanced AI techniques, such as deep learning, can stimulate better solutions for longevity risk management through providing more sophisticated and accurate mortality models (Richman, 2018). This paper proposes a new neighbouring prediction model powered by convolution neural network (CNN) to improve mortality modelling and forecasting. Achieving enhanced mortality models with better predictability, AI techniques will act as a catalyst to help facilitate the development of the new *Life Market* – the traded market in longevity and mortality linked assets and liabilities (Blake *et al.*, 2013).

An accurate mortality model is essential for seeking effective solutions for longevity risk. Starting from the seminal paper by Lee and Carter, various mortality models have been developed (see, e.g., Lee and Carter, 1992; Cairns *et al.*, 2006, 2009; Renshaw and Haberman, 2006). For most mortality models in the literature, they share the assumptions that mortality rates are composed of a few additive latent factors and these factors can be decomposed along three directions: period, age, and cohort. In this paper, we propose a neighbouring prediction model which pushes the boundary further by capturing the neighbourhood effect. The key technical innovation in this framework is that we construct 2-dimensional *images* of neighbourhood mortality data, combined with the deep learning algorithm – CNN, which is specialised to process grid-like data such as images – to build mortality forecasting models. More specifically, for each mortality rate at age x in year t, denoted as $m_{x,t}$, we construct the 2-dimensional "image" of neighbourhood mortality data around $m_{x,t}$, that is, $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$, which includes mortality information for ages in $[x - x_1, x + x_2]$, lagging k years $(1 \leq k \leq s)$. Figure 1 illustrates an example of neighbourhood mortality data of $\mathcal{E}_{m_{x,t}}(3, 3, 5)$, where $x_1 = x_2 = 3, s = 5$. In this construction, the neighbourhood mortality image $\mathcal{E}_{m_{x,t}}(3, 3, 5)$ has a size of (5×7) (i.e., $(s \times (x_1 + x_2 + 1))$). Each observation of $m_{x,t}$ is treated as the response variable, and this neighbourhood data will then be used to predict $m_{x,t}$.

The proposed neighbouring prediction framework is able to capture the intricate, highly nonlinear structure in the mortality data, which can go beyond the directions of period, age, and cohort. We call this the neighbourhood effect. By zooming in on the mortality table, this framework models the mortality of age x in year t as a function of not only the age of x and the cohort of $t - x$ but also ages and cohorts in the neighbourhood of x. In addition, the CNN algorithm utilised in this framework improves the prediction performance, as CNN provides a more computationally efficient way to specialise neural networks for data with a clear grid-structured

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

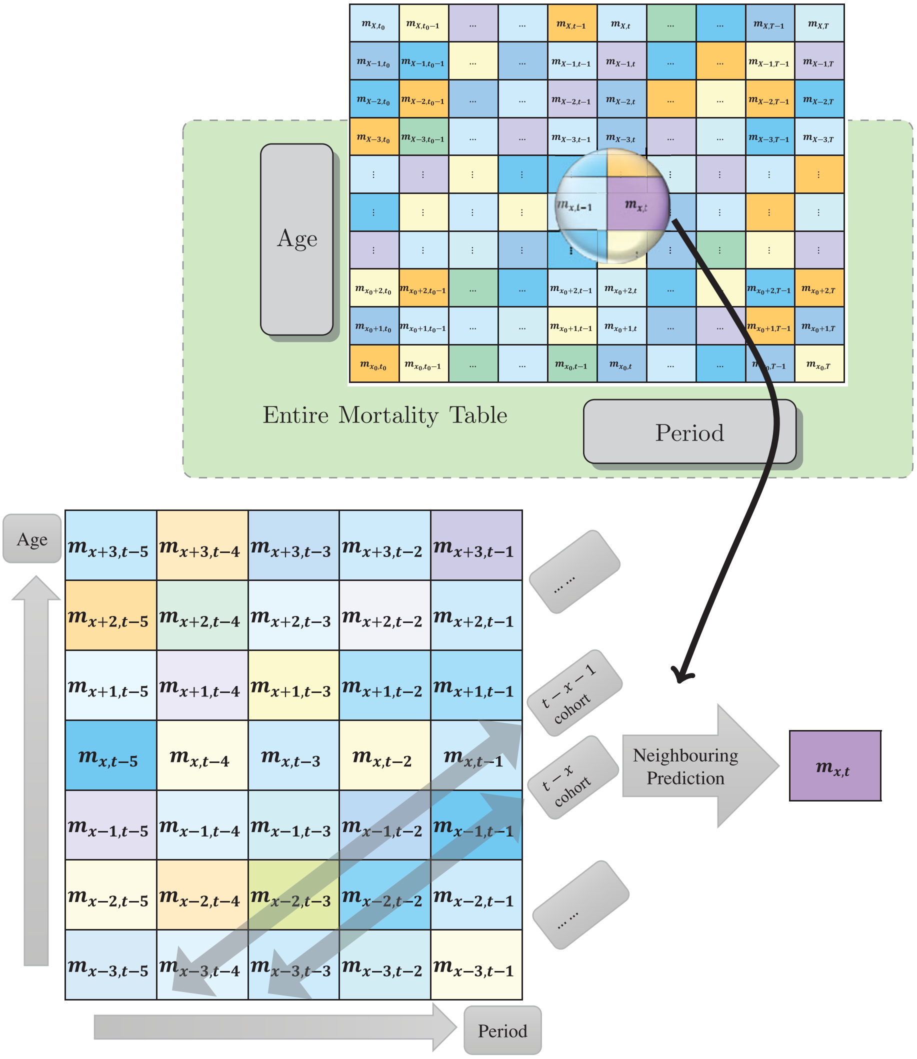
Age
Entire Mortality Table
Period
$m_{x,t-1}$ $m_{x,t}$
Age
$m_{x+3,t-5}$ $m_{x+3,t-4}$ $m_{x+3,t-3}$ $m_{x+3,t-2}$ $m_{x+3,t-1}$
$m_{x+2,t-5}$ $m_{x+2,t-4}$ $m_{x+2,t-3}$ $m_{x+2,t-2}$ $m_{x+2,t-1}$
$m_{x+1,t-5}$ $m_{x+1,t-4}$ $m_{x+1,t-3}$ $m_{x+1,t-2}$ $m_{x+1,t-1}$
$m_{x,t-5}$ $m_{x,t-4}$ $m_{x,t-3}$ $m_{x,t-2}$ $m_{x,t-1}$
$m_{x-1,t-5}$ $m_{x-1,t-4}$ $m_{x-1,t-3}$ $m_{x-1,t-2}$ $m_{x-1,t-1}$
$m_{x-2,t-5}$ $m_{x-2,t-4}$ $m_{x-2,t-3}$ $m_{x-2,t-2}$ $m_{x-2,t-1}$
$m_{x-3,t-5}$ $m_{x-3,t-4}$ $m_{x-3,t-3}$ $m_{x-3,t-2}$ $m_{x-3,t-1}$
……
$t-x-1$ cohort
$t-x$ cohort
……
Neighbouring Prediction
$m_{x,t}$
Period

FIGURE 1: Neighbouring prediction for $m_{x,t}$. This figure illustrates $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$ neighbourhood of $m_{x,t}$, with $x_1 = x_2 = 3, s = 5$.

topology and to discover intricate relations that are highly nonlinear in the mortality data. Moreover, reshaping the original mortality table into the neighbourhood mortality data could significantly enlarge the sample, making it possible to use big data to enhance the mortality forecasting. Finally, $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$ is $\mathcal{F}_{t-1}$ measurable, that is, it only uses information up to year $t-1$. This means that the neighbouring prediction model by construction has a one-year-ahead forecasting nature, in both the training set and test set.

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

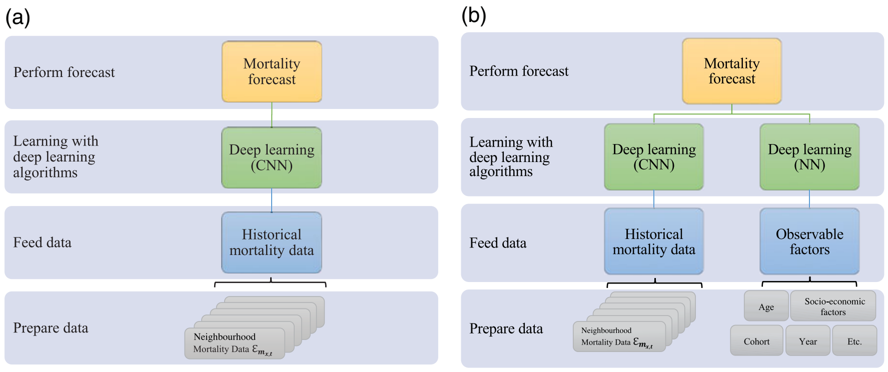
(a)
Perform forecast
Mortality forecast
Learning with deep learning algorithms
Deep learning (CNN)
Feed data
Historical mortality data
Prepare data
Neighbourhood Mortality Data $\mathcal{E}_{m_{x,t}}$
(b)
Perform forecast
Mortality forecast
Learning with deep learning algorithms
Deep learning (CNN)
Deep learning (NN)
Feed data
Historical mortality data
Observable factors
Prepare data
Neighbourhood Mortality Data $\mathcal{E}_{m_{x,t}}$
Age
Socio-economic factors
Cohort
Year
Etc.

FIGURE 2: Conceptual modelling framework for neighbouring prediction models. (a) Pure model. (b) Hybrid model.

Figure 2 displays the conceptual frameworks for the neighbouring prediction models, including the *Pure Model* and *Hybrid Model*. In the Pure Model (Figure 8(a)), the historical mortality data are reshaped as neighbourhood mortality data, $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$, and then fed into a deep learning algorithm to perform mortality prediction. In the Hybrid Model (Figure 8(b)), besides neighbourhood mortality data, $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$, external observable factors (e.g., age, year, cohort, region, socio-economic variables, etc.) are fed into another deep learning model. These two parts are joined then together with a concatenating function. By performing an extensive empirical analysis on 41 countries and regions around the world, we find that the proposed model achieves superior fitness of historical data and better out-of-sample forecast performance. For example, the overall test error of the Pure Model (Hybrid Model) reduces by 1.97% (2.01%) compared to the best performed benchmark model and reduces by 7.02% (7.06%) compared to the worst performed benchmark model.

We further propose two enhanced neighbouring prediction models by including multi-population data. These models are able to take mortality dependence among different populations into consideration. We use all available mortality data from all regions/countries in the sample to construct a universal model that are applicable for all populations. In particular, the *Pure Model* is enhanced with all multi-population data that are available in our sample space, which we call Pure-All Model, or *PALL Model*. Similarly, the enhanced *Hybrid Model* with mortality data from all populations is called Hybrid All Model, or *HALL Model*. We find that the prediction performance of the PALL Model and HALL Model is significantly improved, especially for regions with very limited mortality data. These enhanced models not only benefit from an increased sample size but also are capable to capture the co-movement and dependence in mortality between different populations.

We perform various robustness checks for the proposed neighbouring prediction models. First, we test the robustness of the four neighbouring mortality

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

models to different CNN structures. We find that the prediction errors are stable for CNN architecture parameters. Second, we test different methods to handle missing data. Our prediction results are robust to different methods. Next, we perform a *completely* missing data analysis – a leave-one-population-out (LOPO) analysis. For each country/region, we pretend that the mortality data for this particular population are completely missing. We construct a neighbouring mortality model using mortality data from other countries, predict the mortality rates for this population, and finally evaluate the performance with the true mortality data from this population that is left out of model estimation. This completely missing data problem cannot be tackled by traditional mortality models. Interestingly, we find that the results from neighbouring prediction model produce competitively well-prediction results in the completely missing data case as the benchmarks using full information. These results demonstrate that our proposed approach is able to learn representation and commonality of data and generalise the discovered intricate structure from the data to guarantee good predicting performance. This is of particular interest in situations where data scarcity is a problem. A good example where data may be completely missing is when life insurance companies plan to investigate a new business line, no historical data are readily available for risk modelling. In the final robustness check, we estimate the neighbouring mortality models with an extending window estimation, while the main results are based on splitting the data into training set (80%) and test set (20%). We show that the proposed models have robust prediction results in both estimation procedures.

This paper makes an important contribution to the mortality modelling and forecasting literature. In their seminal paper, Lee and Carter (1992) proposed a linear extrapolation model for stochastic mortality modelling and forecasting, which has become the standard model in the mortality forecast literature. Since then, various extensions of the Lee–Carter model have been proposed for single population modelling (see, e.g., Cairns *et al*., 2006; Renshaw and Haberman, 2006; Cairns *et al*., 2009). For recent development of mortality modelling and forecasting, see Blake *et al*. (2018) and the references therein. Our paper provides a neighbouring prediction model for mortality, which is able to capture the neighbourhood effect and achieves more accurate out-of-sample forecasting. We also propose enhanced prediction models that include mortality data from multiple countries. To this end, our paper also taps the literature of multi-population mortality modelling. Started by Li and Lee (2005), a number of two- and multi-population mortality models have been proposed to enhance mortality modelling and reduce basis risk (see, e.g., Dowd *et al*., 2011; Jarner and Kryger, 2011; Zhou *et al*., 2013; Chen *et al*., 2015; Wang *et al*., 2015, 2018; Li *et al*., 2017; Zhu *et al*., 2017).

In recent years, there are also burgeoning literature on employing machine learning in the area of mortality modelling. For example, Hainaut (2018) uses autoencoder networks to estimate latent factors for mortality. Richman and Wüthrich (*forthcoming*) extend the Lee–Carter model to multiple populations

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

based on neural networks. Dong *et al.* (*forthcoming*) use tensor decomposition to construct multi-population mortality models. Based on similar neighbourhood idea, Perla *et al.* (*forthcoming*) perform time series forecasting for mortality using deep learning algorithms such as recurrent neural network (RNN) and CNN. They show that a simple CNN model can be interpreted as a generalisation of the Lee–Carter model. In particular, they show that CNNs work better than RNNs though the neighbourhood has a time-causal property. The key contribution of our paper is to design the mortality modelling problem as a computer vision task for capturing the neighbourhood effect of mortality. In line with this idea, we crop mortality tables into neighbourhood images, which preserve the intricate local relationship in the mortality data. Consequently, this paper employs 2-dimensional CNN specialised for computer vision tasks to represent and generalise the relationship.

The rest of the paper proceeds as follows. Section 2 proposes the neighbouring prediction models. Section 3 presents the main results, showing that the proposed models have superior prediction performance in an extensive empirical analysis. Section 4 provides various robustness checks. Finally, Section 5 concludes.

2. Neighbouring prediction models

2.1. The neighbourhood effect

Most existing mortality models share the same general form, in which the central death rate at age x and year t, $m_{x,t}$, can be expressed as follows (Cairns *et al.*, 2009):

$$g(m_{x,t}) = \sum_{i}^{I} \beta_x^{(i)} \cdot \kappa_t^{(i)} \cdot \gamma_{t-x}^{(i)}, \tag{2.1}$$

where $\beta_x^{(i)}$ represents age-related effect; $\kappa_t^{(i)}$ represents period-related effect; $\gamma_{t-x}^{(i)}$ represents cohort-related effect, $i = 1, 2, \ldots, I$; and the function $g(\cdot)$ is a transformation function. For example, in Lee–Carter model (Lee and Carter, 1992), the logarithmic transformation of $m_{x,t}$ is specified as a two-factor model:

$$\log(m_{x,t}) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)}. \tag{2.2}$$

In this specification, $I = 2$, $\kappa_t^{(1)} = 1$, $\gamma_{t-x}^{(1)} = \gamma_{t-x}^{(2)} = 1$. The time-varying mortality index, $\kappa_t^{(2)}$, represents the overall level of mortality improvement. In another widely used mortality model, Cairns–Blake–Dowd (CBD) model (Cairns *et al.*, 2006), the logit transformation of the death rate, $q_{x,t} \approx 1 - \exp(-m_{x,t})$, is modelled as:

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

$$\log\left(\frac{q_{x,t}}{1-q_{x,t}}\right)=\kappa_t^{(1)}+\kappa_t^{(2)}(x-\bar{x}), \tag{2.3}$$

where $\bar{x}$ is the average age in the sample span, and we can see that in this specification, $\beta_x^{(1)}=1, \beta_x^{(2)}=x-\bar{x}, \gamma_{t-x}^{(1)}=\gamma_{t-x}^{(2)}=1$. $\kappa_t^{(1)}$ represents the overall mortality improvement, and $\kappa_t^{(2)}$ is the steepness of the logit transformation of the mortality curve.

Note that a mortality model specification as in Equation (2.1) has the following two assumptions:

(a) The mortality rate (upon g- transformation) can be decomposed into I additive latent factors.

(b) Each factor can be decomposed into three multiplicative effects: age, period, and cohort.

The above two assumptions achieve great prediction accuracy and model interpretability in a parsimonious framework; hence, these models expressed in Equation (2.1) have been the preferred methodology in the literature and the official forecasts (e.g., U.S. Census Bureau). Assumption (a) implies a linear relationship of different latent factors. Assumption (b) indicates that each factor is decomposed along three directions: age, period, and cohort.

In this paper, we propose a neighbouring prediction model which pushes the boundary further by predicting $m_{x,t}$ with a neighbourhood around $m_{x,t}$. More specifically, we construct neighbourhood data, $\mathcal{E}_{m_{x,t}}(x_1,x_2,s)$, as follows:

$$\mathcal{E}_{m_{x,t}}(x_1,x_2,s)=\begin{bmatrix} m_{x+x_1,t-s}, & \dots & m_{x+x_1,t-2}, & m_{x+x_1,t-1} \\ \vdots & \dots & \vdots & \vdots \\ m_{x+1,t-s}, & \dots & m_{x+1,t-2}, & m_{x+1,t-1} \\ m_{x,t-s}, & \dots & m_{x,t-2}, & m_{x,t-1} \\ m_{x-1,t-s}, & \dots & m_{x-1,t-2}, & m_{x-1,t-1} \\ \vdots & \dots & \vdots & \vdots \\ m_{x-x_2,t-s}, & \dots & m_{x-x_2,t-2}, & m_{x-x_2,t-1} \end{bmatrix}. \tag{2.4}$$

In other words, for each mortality rate at age x and year t, $m_{x,t}$, we can construct $\mathcal{E}_{m_{x,t}}(x_1,x_2,s)$, which is a two-dimensional neighbourhood around $m_{x,t}$, including mortality information for ages $x-x_1,\dots,x,\dots,x+x_2$, lagging k-years, $k=1,2,\dots,s$, up to the year $t-1$. The mortality data $\mathcal{E}_{m_{x,t}}(x_1,x_2,s)$ are cropped from the original entire mortality table and collected as rectangular, equal-sized submatrices of the full observed sample and will be used as our predictive model inputs. As such, these submatrices of neighbourhood mortality data can be treated as *images*. Importantly, note that in $\mathcal{E}_{m_{x,t}}(x_1,x_2,s)$ defined according to Equation (2.4) is $\mathcal{F}_{t-1}$ measurable, which means that

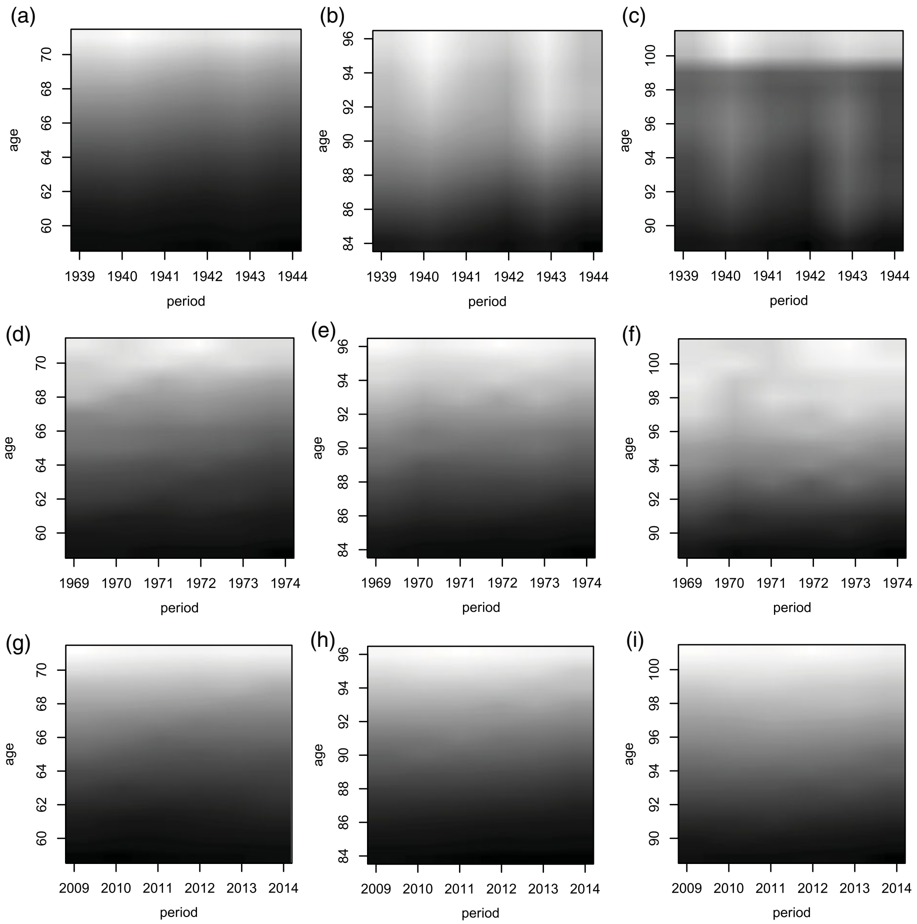

FIGURE 3: Examples of neighbourhood mortality data. This figure illustrates images of sample neighbourhood mortality data, $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$, for total population in the US. The window width of $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$ is set to be $x_1 = x_2 = s = 6$. Ages of 65, 90, 95 in the years 1945, 1975, and 2015 are displayed. In these images, the darker the color, the lower the mortality rates.

using $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$ to predict mortality $m_{x,t}$ is performing a one-step-ahead forecasting, even in training set (i.e., in-sample data).

Figure 3 displays images of sample neighbourhood mortality data for total population in the US. In these images, the darker the color, the lower the mortality rates. We can see that the patterns vary a lot in different neighbourhoods in the mortality table. For example, in image of $\mathcal{E}_{m_{90,1945}}(6, 6, 6)$, we can see two obvious light vertical lines for years 1940 and 1943: the mortality rates of these two years are higher than the other years. We also observe in this image that year 1942 has much darker color than the years close to it. For $\mathcal{E}_{m_{95,1945}}(6, 6, 6)$, we see a steep mortality increase from age 99 to age 100 in the image, which cannot be found from other neighbourhood mortality data. These various local

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

patterns cannot be fully represented by the three factors: age–year–cohort, in a linear model. We construct $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$ in such a way that it could be treated as 2D images; hence, not only mortality rate of age x or in year t is included in the model, but mortality of neighbouring ages or years and their interactions are also used to construct the prediction model. The 2D neighbourhood images lead us to use 2D CNN that is specialised for such computer vision task. Using these neighbourhood data as model inputs is important for mortality forecasting as it provides a more comprehensive picture of these intricate local nonlinear patterns.

2.2. Neighbouring prediction models

We introduce the following four neighbouring prediction models:

Pure Model:

$$m_{x,t} = f(\mathcal{E}_{m_{x,t}}(x_1, x_2, s)), \tag{2.5}$$

where $f(\cdot)$ represents a predictive model constructed from certain deep learning algorithm with image data as inputs. This model is named “Pure Model” because only historical mortality data are used as model inputs. The historical mortality data are reshaped as neighbourhood mortality data, $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$, which are then fed into a deep learning model to perform mortality prediction. The flow chart of the conceptual framework for the Pure Model is displayed in Figure 8(a).

Hybrid Model:

$$m_{x,t} = f_1 \circ c\left(f_2(\mathcal{E}_{m_{x,t}}(x_1, x_2, s)), w\left(X_t\right)\right), \tag{2.6}$$

where $f_2(\cdot)$ is a predictive model constructed from certain deep learning algorithm with image data as inputs and a densely connected neural layer as output; $f_1(\cdot)$ and $w(\cdot)$ are deep learning algorithms with vector data as inputs and a densely connected neural layer as output; X_t is a vector of observable factors; $c(\cdot, \cdot)$ is a concatenating function which links model $f_2(\cdot)$ and model $w(\cdot)$; and “$\circ$" denotes function composition. In this model, historical mortality data are reshaped as neighbourhood mortality data, $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$, and fed into deep learning model $f_2(\cdot)$. External observable factors (e.g., age, year, cohort, region, socio-economic variables, etc.) are fed into another deep learning model $w(\cdot)$. Finally, these two parts are joined together with a concatenating function and then learnt by deep learning algorithms. Hence, it is called Hybrid Model, the conceptual flow chart for which is displayed in Figure 8(b).

Pure All (PALL) Model:

The *Pure Model* can be further enhanced by including multi-population data. In other words, in the PALL Model, we will use all available mortality

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

data from all regions/countries in the sample, reshape them into neighbourhood mortality images, and construct a universal model that is applicable for all populations. This model will be able to take mortality dependence among different populations into consideration. Moreover, in some regions, the data availability and credibility may be a big issue in mortality modelling and forecasting. Hence, the single population mortality models may not have satisfying performance. As a result, including all data available from different populations into the neighbouring prediction model potentially can significantly increase the sample size and, hence, improve the model performance. For example, in the empirical analysis discussed in this paper, utilising neighbourhood mortality data from all populations increases the sample size to 100,608.

Hybrid All (HALL) Model:

Similarly, including data from different countries/regions can augment the *Hybrid Model* to a multi-population version. Again, the HALL model not only benefits from an increased sample size but also is capable to capture and model the dependence between different populations.

The advantages of the four proposed neighbouring prediction models are fivefold. First, this framework captures the neighbourhood effect. Using $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$ to construct the feature space, the model is capable of learning the representation of the neighbourhood through advanced abstraction and discover more general predictive relations, including age, period, cohort, and many more abstract neighbourhood representations and relationships. Because this model zooms in on the mortality table, it models $m_{x,t}$ as a function of not only the age of x but also ages in the neighbourhood of x, that is, $[x - x_1, x + x_2]$, together with their interactions. Similarly, instead of modelling $m_{x,t}$ as a function of $t - x$ cohort, this framework models $m_{x,t}$ as a function of the interactions of $t - x$ cohorts and other cohorts in the neighbourhood. Moreover, $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$ is lagging k years, where $1 \leq k \leq s$, which means that by construction, the proposed framework is performing one-year ahead forecast. Second, utilising deep learning algorithm that is specialised for computer vision (CNN that will be described in Subsection 2.3.2), the proposed neighbourhood model is able to discover intricate relations that are highly nonlinear, extending the linear specification in Equation (2.1). Third, reshaping the original entire mortality table into the neighbourhood mortality data could significantly increase the number of observations used to construct the prediction model (see, for example, the last column of Table 1 how the sample size is increased after reshaping the mortality data. In particular, in PALL and HALL models, the sample size is increased to 100,608.). Fourth, the Hybrid model in Figure 8(b) provides a convenient framework to study the relationship between socio-economic factors and

TABLE 1

SUMMARY OF COUNTRIES AND REGIONS IN THE HUMAN MORTALITY DATABASE.

Country name	Abbreviation	Sample period	Sample size (Reshaped)	Country name	Abbreviation	Sample period	Sample size (Reshaped)
Australia	AUS	1921–2016	2790	Japan	JPN	1947–2017	2015
Austria	AUT	1947–2017	2015	Latvia	LVA	1959–2017	1643
Belarus	BLR	1959–2016	1612	Lithuania	LTU	1959–2017	1643
Belgium	BEL	1841–2018	4985	Luxembourg	LUX	1960–2015	1487
Bulgaria	BGR	1947–2017	1798	Netherlands	NLD	1850–2016	4946
Canada	CAN	1921–2016	2790	New Zealand	NZL	1948–2013	1860
Chile	CHL	1992–2008	341	Norway	NOR	1846–2018	5046
Croatia	HRV	2002–2017	310	Poland	POL	1958–2016	1643
Czechia	CZE	1950–2017	1922	Portugal	PRT	1940–2015	2170
Denmark	DNK	1835–2016	5423	Republic of Korea	KOR	2003–2018	248
Estonia	EST	1959–2017	1643	Russia	RUS	1959–2014	1550
Finland	FIN	1878–2015	4035	Slovakia	SVK	1950–2017	1922
France	FRA	1816–2017	6076	Slovenia	SVN	1983–2017	899
Germany	DEU	1990–2017	682	Spain	ESP	1908–2016	3193
Greece	GRC	1981–2013	837	Sweden	SWE	1751–2018	8060
Hong Kong	HKG	1986–2017	806	Switzerland	CHE	1876–2016	4108
Hungary	HUN	1950–2017	1922	Taiwan	TWN	1970–2014	1209
Iceland	ISL	1901–2016	3308	U.K.	GBR	1922–2016	2759
Ireland	IRL	1950–2017	1829	U.S.A.	USA	1933–2017	2449
Israel	ISR	1983–2016	868	Ukraine	UKR	1959–2013	1519
Italy	ITA	1872–2014	4247				

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

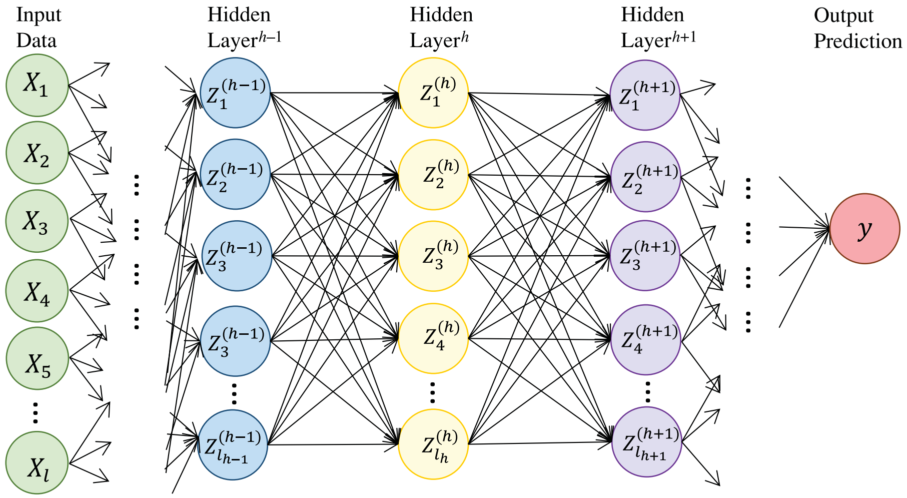

FIGURE 4: A fully connected neural network structure.

aggregate mortality changes, in addition to improving the prediction capacity of the model. Finally, the PALL Model and HALL Model are able to enhance the prediction and model the mortality dependence between different populations.

2.3. A Convolutional Neural Network (CNN) approach

The deep learning algorithms we propose to construct prediction functions are CNNs, a special class of neural networks (NNs) that use convolution in at least one of their layers. In this subsection, we first describe a fully connected NN in Subsection 2.3.1. Then in Subsection 2.3.2, we introduce the CNN approach that will be employed for estimating the four neighbouring prediction models in Subsection 2.2.

2.3.1. *Fully connected Neural Networks (NNs)*

NNs are models with multiple layers of representation (known as neurons) in the structure, obtained by transforming the representation at one level into a representation at a more abstract higher level. In other words, the outputs of neurons at the previous level become inputs to other neurons at the next level. The most common type of NNs is the fully connected architecture as illustrated in Figure 4. This model starts with the raw input of a l-dimensional vector, $X = (X_1, X_2, \ldots, X_l)^{'}$, and eventually produces an output prediction of y. In a *deep learning* NN architecture, besides an input layer and an output layer, the structure has $H(H > 1)$ additional hidden layers. The h-th ($h = 1, 2, ..., H$) hidden layer, $Z^{(h)} = \left(Z_1^{(h)}, Z_2^{(h)}, \ldots, Z_{l_h}^{(h)}\right)^{'}$, contains l_h neurons, obtained by

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

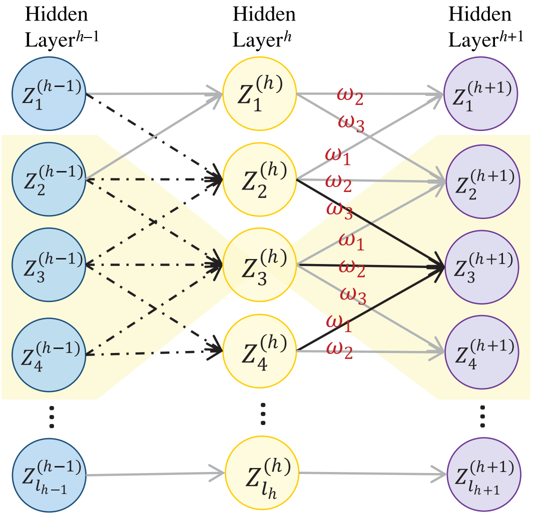

FIGURE 5: A sparsely connected neural network structure.

transforming the linear combination of the neurons from the previous layer through a nonlinear function, σ_{h-1}, elementwise. More specifically,

$$Z^{(h)} = \sigma_{h-1}\left(\boldsymbol{\alpha}^{(h-1)} + \boldsymbol{\omega}^{(h-1)} Z^{(h-1)}\right), \tag{2.7}$$

where $\boldsymbol{\alpha}^{(h-1)}$ is a $(l_h \times 1)$-vector of bias units that captures the intercepts in the model; $\boldsymbol{\omega}^{(h-1)}$ is a $(l_h \times l_{h-1})$-dimensional weight matrix; σ_{h-1} is called *activation* functions and is usually pre-assumed and nonlinear. Finally, this is a fully connected structure in which neurons between two adjacent layers are fully pairwise connected.

2.3.2. *Convolutional Neural Networks (CNNs)*

CNNs are special neural networks that use the convolutional kernels (a.k.a. filters) to derive features from the data in at least one of their layers. After the pioneering work by LeCun *et al.* (1998), CNNs fell out of favour for more than one decade until the breakthrough work by Krizhevsky *et al.* in 2012 and have become the workhorse of deep learning with larger volumes of data and more computing power becoming available. CNNs provide a more computationally efficient way to specialise neural networks for data with a clear grid-structured topology. This makes CNNs very suitable to construct our proposed neighbouring prediction models in Subsection 2.2, with the reshaped neighbourhood mortality images $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$ as inputs (see Figure 3). As a result, in this paper, we propose to use CNN to construct neighbouring prediction models.

CNNs leverage three key ideas to improve the computational efficiency and the predictive ability of the model: sparse connection, weight sharing, and the use of multiple layers (LeCun *et al.*, 2015). Figure 5 uses layers $h-1, h$, and $h+1$ in the NN architecture to illustrate the differences between a fully

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

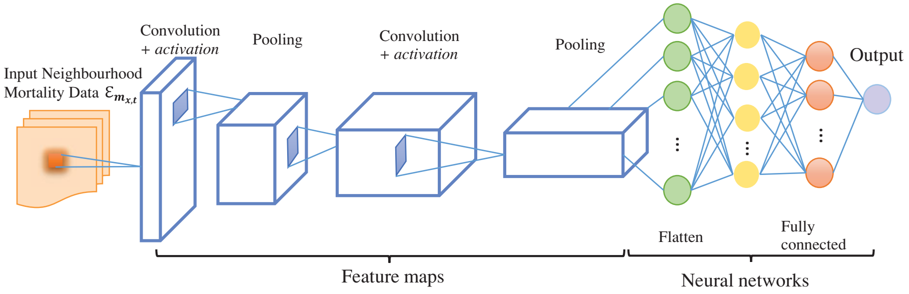

FIGURE 6: An illustrative CNN architecture in constructing neighbouring prediction models.

connected NN and a sparsely connected CNN. For example, the neuron $Z_3^{(h)}$ is impacted with only three neurons in layer $h-1$ (shaded in yellow). Similarly, only three neurons in the next layer, layer $h+1$, are affected by $Z_3^{(h)}$ (also shaded in yellow). The three weights, ω_1, ω_2, and ω_3, are shared by all neurons in layer h. Although each neuron is connected to the local conjunctions from the previous layers, the use of multiple layers connects most of the neurons in an indirect way. For example, while $Z_3^{(h+1)}$ is directly connected to three neurons in layer h (connections represented in solid arrows), it can be indirectly linked to many neurons in remote layers (connection represented in dashed arrows).

The local connection is achieved through the convolution operation, a fundamental building block in CNN model. Compared to regular fully connected neural networks, which learn global patterns in the input feature space, CNNs are only connected to small regions and learn local patterns, using convolution operation. This not only improves the computational efficiency of CNN models but also makes them very suitable to process image data with local patterns. Constructing neighbouring prediction models with CNNs involves four main components: (1) convolution operation; (2) nonlinear activation; (3) pooling; and (4) fully connected neural networks. Figure 6 illustrates the four components in a CNN architecture. The neighbourhood mortality data are fed into the CNN architecture and go through several layers of components (1)–(3), that is, convolution operation, nonlinear activation, and pooling. Units in the convolution layer are arranged in feature maps and passed through a nonlinear activation function. A pooling function is then used to aggressively down-sample feature maps. Pooling plays an important role in reducing the number of parameters to proceed and creating translation-invariance to small shifts. In this paper, we use Max pooling (Zhou and Chellappa, 1988), which is a typical pooling operation that extracts the maximum value from the input feature maps. Online Appendix A provides a detailed example to illustrate the 2D-convolution.

In the last stage of the architecture, these layers of convolution, activation, and pooling are flattened and stacked into fully connected NNs and the output is the model forecast for $m_{x,t}$. Finally, it is worth mentioning that mortality rate is obviously a non-negative value. This non-negative property can be easily guaranteed in a CNN framework, by simply choosing the output layer activation function with non-negative range. Throughout this paper, we use the Rectified Linear Unit (RELU), defined as $\sigma(x) = \max(x, 0)$, as the nonlinear activation function in our empirical analysis.

Remark 1. *The parameters in the CNN architecture are learnt through stochastic gradient descent* (*SGD; see, e.g., Bottou and Bousquet,* 2008)*, where their gradients are computed with backpropagation, which is implemented in the TensorFlow* (*Abadi et al.,* 2016) *and accessible via the Keras library in R* (*Chollet et al.,* 2018)*. Backpropagation is a procedure to apply chain rule for calculating derivatives of an objective function. For backpropagation in detail, see, for example, Hastie et al.* (2005)*.*

3. Empirical analysis

3.1. The data

In this section, we apply the proposed neighbouring prediction models to all the populations in the Human Mortality Database (HMD; http://www.mortality.org). The HMD collects a rich set of mortality data for 41 countries and regions, including deaths, exposures, etc. We use total population mortality data between age 65 and 95. Robustness tests show that the model proposed in this paper is robust to age intervals and genders. For Hybrid Model, we consider year and cohort factors for all the 41 countries and regions as observable factors. For HALL model, we further add country name indicators as observable variables. Table 1 summarises the mortality data. For both left and right panels, the last column reports the new sample size after reshaping the data into the neighbourhood mortality images according to Equation (2.4). Also note that the sample size after reshaping reported in the table is that after removing samples with "NA" values. We also train the models using other methods to handle missing data, and we find consistent results.

Following the machine learning convention, we split the sample into two subsamples: training set and test set. Training set contains 80% of the data and is used to estimate the model, and test set contains the remaining 20% data to evaluate the out-of-sample performance. Such a procedure treats data as random sample. However, it is important to note that this does not imply that our model assumes observations are independent. As an illustration example, suppose we construct a 3 × 3 neighbourhood mortality data (i.e., images), $\mathcal{E}_{m_{x,t}}(1, 1, 3)$ and $\mathcal{E}_{m_{x+1,t}}(1, 1, 3)$ for $m_{x,t}$ and $m_{x+1,t}$, respectively. Then, there will be an overlapped neighbourhood which includes $m_{x,t-3}, m_{x,t-2}, m_{x,t-1}, m_{x+1,t-3}, m_{x+1,t-2}$, and $m_{x+1,t-1}$. This means both $m_{x,t}$ and $m_{x+1,t}$ will be impacted and predicted by the overlapped neighbourhood. In other words, although the samples $(\mathcal{E}_{m_{x,t}}(1, 1, 3), m_{x,t})$ and $(\mathcal{E}_{m_{x+1,t}}(1, 1, 3), m_{x+1,t})$

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

are treated as random sample in the learning procedure, it does not mean that the CNN model we train assumes $m_{x,t}$ and $m_{x+1,t}$ are independent. As a robustness check, in Subsection 4.4, we perform an extending window prediction procedure, which splits data by preserving the temporal order.

We evaluate and compare the prediction power of different models based on one-step-ahead forecasts, using mean absolute percentage error (MAPE) defined as follows:

$$MAPE = \frac{100\%}{N} \sum_t \sum_x \left| \frac{m_{x,t} - \hat{m}_{x,t}}{m_{x,t}} \right|, \tag{3.1}$$

where $\hat{m}_{x,t}$ is the predicted value of $m_{x,t}$ based on different predictive models; N is the number of observations in the test set. MAPE is a commonly used measure to evaluate forecasting for mortality. To improve the robustness of the prediction results, throughout the empirical analysis of this paper, we adopt the ensemble idea and neighbouring prediction models are averaged over 20 training runs (see, e.g., Dietterich, 2000). In particular, multiple different random seeds are used to initialise neural networks and predictions are constructed by averaging forecasts from all models trained.

Preliminary analysis indicates a trade-off in the neighbourhood window size: Using a larger (smaller) window size to construct the neighbourhood mortality data captures more (less) neighbourhood information that can be abstracted by the CNN model, but produces more (less) noises. Therefore, in order to strike a balance between accuracy and stability, in this section, we estimate the neighbouring prediction models with a median window width of $x_1 = x_2 = s = 6$. Given the size of the neighbourhood data, we use a kernel size of 3×3. In addition, for the depth of the feature maps in the CNN architecture, we use depths of 8, 16, and 16 in the first, second, and ending output, respectively. We call this a [8-16-16] structure. The number of neurons in the fully connected layer is set to be 16. In the CNN-Hybrid model, for the function $\omega(\cdot)$, we use a fully connected neural network model that has 1 hidden layer with 4 neurons and we call this a [8-16-16] + [4] structure. More details of the CNN structure used for these two models are illustrated in Listings 1 and 2 in Online Appendix B.

We use four mortality models that have been widely used in the mortality modelling literature as our benchmark models. Table 2 summarises the specifications of the four models, including the Lee–Carter Model (LC; Lee and Carter, 1992), Renshaw–Haberman Model (RH; Renshaw and Haberman, 2006), Cairns–Blake–Dowd Model (CBD; Cairns *et al.*, 2006), and CBD Model with Quadratic Term and Cohort Effect (M7; Cairns *et al.*, 2009). These four models are representative benchmarks. In particular, LC and CBD contain age effect and period effect only; RH and M7 are models with age effect, period effect, and cohort effect. In addition, LC, CBD, RH, and M7 models have 1, 2, 3, and 4 latent factors, respectively.

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

TABLE 2

SPECIFICATIONS OF FOUR STOCHASTIC MORTALITY MODELS AS BENCHMARK MODELS.

Mortality model 1: Lee–Carter model $\log(m_{x,t}) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)}$
Mortality model 2: Renshaw–Haberman model $\log(m_{x,t}) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\gamma_{t-x}^{(3)}$
Mortality model 3: Cairns–Blake–Dowd model $\log\left(\frac{q_{x,t}}{1-q_{x,t}}\right) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x})$
Mortality model 4: Cairns–Blake–Dowd model with quadratic term and cohort effect $\log\left(\frac{q_{x,t}}{1-q_{x,t}}\right) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x}) + \kappa_t^{(3)}((x-\bar{x})^2 - \hat{\sigma}_x^2) + \gamma_{t-x}^{(4)}$

3.2. Pure model and hybrid model prediction results

We first discuss the prediction results with the models that utilise mortality data from individual country only, that is, Pure Model and Hybrid Model, which are summarised in Table 3. For both left and right panels, columns 1–4 display prediction errors from the four benchmark models; the fourth and third last columns show prediction errors from Pure Model for training set and test set, respectively; the last two columns show prediction errors from the Hybrid Model for training and test sets. It is important to note that in the proposed neighbouring prediction model, the neighbourhood mortality data, $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$, contain information up to time $t-1$ (see Equation (2.4)) and are used to predict $m_{x,t}$, which is the mortality information at time t. This indicates that the neighbouring prediction model, by the construction, is performing a one-step-ahead forecasting even with training set. Therefore, MAPE of the neighbouring prediction results with both the training set and the test set is comparable to the benchmark MAPEs. The second part of Table 3 summarises the overall average prediction errors of different models. We can see that on average, Pure (Hybrid) Model produces 4.22% (4.26%) prediction error on the training set and 4.87% (4.83%) prediction error on the test set. In contrast, benchmark models have 6.84%, 7.63%, 10.87%, and 11.89% prediction errors for LC, CBD, M7, and RH, respectively. In addition, Pure and Hybrid models have very close training errors and test errors, indicating that these models are not overfitting the data.

It is also interesting to compare the results between the Pure Model and Hybrid Model. We see that the Hybrid Model has smaller average test error (4.83%) than the Pure Model (4.87%). Moreover, Hybrid produces more stable prediction performance. This can be shown from the standard deviations of the prediction errors among the 41 populations. For example, the standard deviations of the prediction errors from the Pure Model are 1.55% and 1.87% for

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

TABLE 3

SUMMARY OF PREDICTION RESULTS FOR SINGLE POPULATION PURE MODEL AND HYBRID MODEL.

	LC (%)	CBD (%)	M7 (%)	RH (%)	Pure (%) (training)	Pure (%) (test)	Hybrid (%) (training)	Hybrid (%) (test)		LC (%)	CBD (%)	M7 (%)	RH (%)	Pure (%) (training)	Pure (%) (test)	Hybrid (%) (training)	Hybrid (%) (test)
AUS	4.66	5.78	9.80	4.07	3.57	4.21	3.58	4.07	JPN	5.48	6.86	10.42	24.02	2.77	2.99	3.06	3.24
AUT	5.70	10.55	7.03	4.53	2.94	3.18	2.97	3.10	LVA	9.93	9.10	5.22	16.10	4.01	4.78	3.90	4.51
BLR	8.06	7.30	8.79	18.51	3.97	5.04	3.94	4.95	LTU	6.98	11.42	8.81	13.22	4.24	5.33	4.31	5.27
BEL	6.47	6.72	19.68	11.07	4.18	4.83	4.20	4.81	LUX	12.50	13.44	11.45	20.96	8.06	10.04	8.26	9.81
BGR	6.56	7.13	13.49	10.00	4.10	5.00	3.95	5.04	NLD	5.26	4.57	10.44	6.29	4.38	5.60	4.43	5.57
CAN	4.60	5.22	2.55	12.82	2.79	3.17	2.74	3.06	NZL	6.58	6.60	5.46	7.94	4.92	5.74	4.78	5.56
CHL	5.94	5.85	4.97	10.28	3.69	4.10	4.62	4.22	NOR	7.19	5.55	11.52	6.34	4.32	4.71	4.33	4.85
HRV	6.22	10.29	5.11	4.76	3.97	4.54	4.03	4.69	POL	4.59	8.52	3.25	16.45	2.76	3.03	2.72	2.93
CZE	5.37	8.23	3.31	5.36	3.11	3.39	3.10	3.40	PRT	6.40	7.56	17.21	24.42	4.08	4.70	4.14	4.83
DNK	7.73	4.79	10.23	5.97	4.66	5.54	4.73	5.49	KOR	3.40	5.68	3.18	3.13	8.44	5.89	6.48	4.08
EST	10.38	9.62	5.73	61.80	4.81	5.84	4.74	5.74	RUS	10.53	7.17	6.86	15.18	2.80	3.29	2.86	3.33
FIN	5.78	7.19	16.91	5.57	5.17	7.21	5.46	7.20	SVK	6.97	7.28	4.34	8.59	4.14	5.07	4.11	4.91
FRA	6.56	8.69	51.01	18.27	3.98	4.26	3.98	4.21	SVN	4.99	10.85	4.81	6.66	4.59	5.54	4.67	5.58
DEU	5.21	11.61	3.33	4.79	2.40	2.46	2.91	2.92	ESP	10.33	7.50	30.54	6.68	4.26	5.12	4.05	4.90
GRC	6.90	9.65	5.58	6.09	2.95	2.99	3.22	3.14	SWE	8.16	4.94	55.79	6.55	5.54	6.47	5.63	6.58
HKG	6.36	5.82	4.73	3.59	5.24	5.85	5.38	5.81	CHE	5.61	8.12	7.45	26.44	4.57	5.66	4.47	5.48
HUN	6.84	8.72	5.20	6.44	2.94	3.49	2.94	3.39	TWN	4.08	4.47	3.49	5.76	3.24	3.83	3.44	4.04
ISL	15.00	15.12	15.13	16.39	9.96	12.51	10.46	12.73	GBR	6.03	4.54	5.02	4.62	3.41	3.61	3.50	3.60
IRL	8.23	5.86	5.29	24.30	4.35	5.24	4.47	5.25	USA	4.06	5.96	3.26	4.88	2.13	2.23	2.13	2.23
ISR	4.83	6.48	3.64	9.89	4.00	4.37	4.29	4.91	UKR	8.33	6.28	6.51	10.44	3.19	3.88	3.18	3.57
ITA	5.64	5.61	25.60	8.29	4.28	5.05	4.30	5.01									

Overall average errors for 41 countries and regions

LC (%)	CBD (%)	M7 (%)	RH (%)	Pure (%) (training)	Pure (%) (test)	Hybrid (%) (training)	Hybrid (%) (test)
6.84	7.63	10.78	11.89	4.22	4.87	4.26	4.83

training set and test set, respectively. In contrast, the corresponding standard deviations for Hybrid are 1.49% and 1.86%. These results suggest the benefit of using the Hybrid Model by including more information.

3.3. Enhancing prediction with multi-population data

In this section, we discuss the results of the proposed neighbouring prediction models with multi-population data, using PALL and HALL models. Each of the enhanced models constructs a single prediction model applicable for all 41 populations in the sample. This universal model not only benefits from more data, but it is also enhanced, being able to model the interactions, comovements, and dependence among multiple populations.

For each country/region, we divide its sample into 80%-training set and 20%-test set. Next, we collect training data from all populations to fit a Pure Model or a Hybrid Model that incorporates mortality dependence and is applicable to all populations. Finally, we evaluate the out-of-sample prediction performance of each country/region individually using its corresponding test sets, based on the MAPE in Equation (3.1). The prediction results are collected in Table 4. We can see that PALL and HALL not only have better performance than the benchmarks, but they also improve the individual Pure and Hybrid models. In particular, PALL in general has slightly larger training errors but much smaller test errors. The overall average test error of the PALL Model is 4.61%, improved by 0.26% compared to the Pure Model. This indicates that enhanced with bigger data sets, the proposed Pure Model strikes a better bias-variance trade-off, improving the predicativity by sacrificing a little goodness-of-fit. Similarly, compared to individual Hybrid Model, HALL strikes a better balance between bias and variance, leading to better out-of-sample performance. Comparing HALL and Pure Models, we see that HALL has average test error of 4.49%, which is 0.38% lower than the average test error of Pure Model.

It is also interesting to note that for populations that suffer more from the data scarcity problem, they tend to benefit more significantly from the enhanced model. The best example to explain this point is the mortality data from Republic of Korea (KOR). It has sample period from 2003 to 2018, only 16 years. Even after reshaping into neighbouring mortality data set, the sample size becomes 248, which is still very small. Hence, the individual Pure Model based on KOR data only has test error of 5.89%, which is not satisfying compared to the benchmark models. However, enhanced with more data from other populations, the PALL Model has a test error of 2.78%. The performance of PALL improves 3.11% (i.e., error reduction) compared to the Pure Model and improves 0.62%, 2.90%, 0.40%, and 0.35%, compared to the benchmark models LC, CBD, M7, and RH, respectively. This indicates that PALL is particular useful when data availability and credibility are primary issues for mortality modelling of certain populations. Similar to PALL,

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

TABLE 4

SUMMARY OF PREDICTION RESULTS FOR PALL MODEL AND HALL MODEL.

	LC (%)	CBD (%)	M7 (%)	RH (%)	PALL (%) (training)	PALL (%) (test)	HALL (%) (training)	HALL (%) (test)		LC (%)	CBD (%)	M7 (%)	RH (%)	PALL (%) (training)	PALL (%) (test)	HALL (%) (training)	HALL (%) (test)
AUS	4.66	5.78	9.80	4.07	3.79	3.98	3.69	3.97	JPN	5.48	6.86	10.42	24.02	3.05	3.08	2.81	3.08
AUT	5.70	10.55	7.03	4.53	3.09	3.10	2.94	2.82	LVA	9.93	9.10	5.22	16.10	4.39	4.29	4.12	4.06
BLR	8.06	7.30	8.79	18.51	4.45	4.63	4.02	4.14	LTU	6.98	11.42	8.81	13.22	4.70	4.77	4.58	4.53
BEL	6.47	6.72	19.68	11.07	4.65	4.88	4.62	4.69	LUX	12.50	13.44	11.45	20.96	9.27	9.80	9.17	8.60
BGR	6.56	7.13	13.49	10.00	5.02	5.22	4.65	4.77	NLD	5.26	4.57	10.44	6.29	5.09	5.37	4.95	5.80
CAN	4.60	5.22	2.55	12.82	2.87	2.95	2.76	2.96	NZL	6.58	6.60	5.46	7.94	4.95	5.16	4.82	4.96
CHL	5.94	5.85	4.97	10.28	4.05	4.31	3.78	3.82	NOR	7.19	5.55	11.52	6.34	4.60	4.63	4.46	4.64
HRV	6.22	10.29	5.11	4.76	3.56	3.52	3.60	3.03	POL	4.59	8.52	3.25	16.45	3.06	3.09	2.80	2.92
CZE	5.37	8.23	3.31	5.36	3.15	3.25	2.85	2.94	PRT	6.40	7.56	17.21	24.42	4.25	4.37	4.09	4.04
DNK	7.73	4.79	10.23	5.97	5.23	5.57	5.11	5.37	KOR	3.40	5.68	3.18	3.13	3.38	2.78	3.23	2.45
EST	10.38	9.62	5.73	61.80	5.36	5.46	5.00	5.33	RUS	10.53	7.17	6.86	15.18	3.58	3.52	2.98	3.13
FIN	5.78	7.19	16.91	5.57	5.84	6.79	5.67	6.97	SVK	6.97	7.28	4.34	8.59	4.25	4.49	4.05	4.09
FRA	6.56	8.69	51.01	18.27	4.42	4.48	4.37	4.52	SVN	4.99	10.85	4.81	6.66	5.06	5.00	4.77	4.94
DEU	5.21	11.61	3.33	4.79	2.51	2.51	2.29	2.17	ESP	10.33	7.50	30.54	6.68	4.99	5.11	4.73	4.86
GRC	6.90	9.65	5.58	6.09	3.09	2.90	3.05	2.94	SWE	8.16	4.94	55.79	6.55	6.85	6.81	7.30	6.74
HKG	6.36	5.82	4.73	3.59	4.95	4.99	4.72	4.26	CHE	5.61	8.12	7.45	26.44	4.72	4.96	5.01	5.00
HUN	6.84	8.72	5.20	6.44	3.48	3.54	3.22	3.30	TWN	4.08	4.47	3.49	5.76	3.28	3.63	4.44	4.54
ISL	15.00	15.12	15.13	16.39	11.69	12.11	11.43	13.64	GBR	6.03	4.54	5.02	4.62	3.37	3.51	3.53	3.69
IRL	8.23	5.86	5.29	24.30	4.92	4.99	4.71	4.60	USA	4.06	5.96	3.26	4.88	2.32	2.41	2.93	2.74
ISR	4.83	6.48	3.64	9.89	4.10	4.03	3.88	3.94	UKR	8.33	6.28	6.51	10.44	3.66	3.65	4.68	3.89
ITA	5.64	5.61	25.60	8.29	5.24	5.50	5.17	5.31									

Overall average errors for 41 countries and regions

	LC (%)	CBD (%)	M7 (%)	RH (%)	PALL (%) (training)	PALL (%) (test)	HALL (%) (training)	HALL (%) (test)
	6.84	7.63	10.78	11.89	4.49	4.61	4.41	4.49

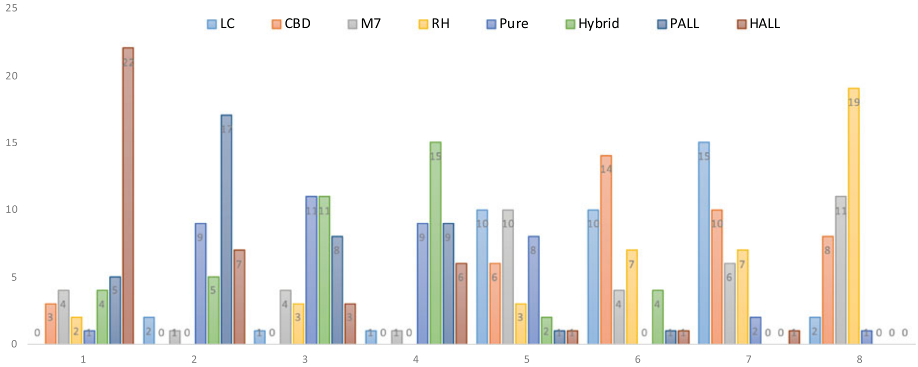

FIGURE 7: Ranking of prediction performances. This figure displays the ranking of model performances based on out-of-sample prediction errors.

the HALL Model also tends to improve the performance more for population with restricted data availability. For example, for populations with reshaped sample sizes smaller than 1000 (e.g., CHL, HRV, KOR), HALL has significant prediction improvement in both training and test sets.

Comparing the prediction results from the four neighbouring prediction models, that is, Pure, Hybrid, PALL, and HALL, it is clear that the model performance is enhanced by including bigger data sets from multiple populations. In particular, compared to the benchmark models, the proposed neighbouring prediction models all have lower average prediction errors, with HALL model producing the smallest average test errors among all models (4.49%). The neighbouring prediction models also have more stable performance, with smaller standard deviations of the prediction errors among different populations. Figure 7 further demonstrates the superiority of the proposed neighbouring prediction model, which displays the ranking of model performances. The eight models are ranked based on the prediction errors. HALL ranks as the best performing model in 22 out of 41 countries/regions, followed by PALL (5 out of 41) and Hybrid (4 out of 41). The PALL Model has the second-best performance in 17 out of 41 populations. In addition, both Pure and Hybrid models are the second-runner up models in 11 out of 41 countries/regions. Finally, Hybrid, PALL, and HALL are ranked last in none of the countries/regions.

4. ROBUSTNESS ANALYSES

4.1. Robustness: CNN structure

Previously, all empirical results are based on the baseline CNN architecture of [8-16-16] for the Pure Model and [8-16-16]+[4] for the Hybrid Model. In this section, we perform robustness tests for different CNN structure parameters, using US mortality data as an example. The results are shown in Table 5. Panel

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

TABLE 5

ROBUSTNESS CHECK FOR CNN STRUCTURES.

Panel A: Prediction errors for the Pure Model with different CNN structures

	Pure [8-16]		Pure [16-16]		Pure [8-16-16]		Pure [8-16-32]	
$x_1/x_2/s$	Training (%)	Test (%)	Training (%)	Test (%)	Training (%)	Test (%)	Training (%)	Test (%)
2	2.29	2.40	2.12	2.21	2.18	2.24	2.37	2.49
3	2.21	2.28	2.11	2.16	2.20	2.25	2.20	2.28
4	2.31	2.49	2.07	2.17	2.41	2.52	2.43	2.55
5	2.38	2.48	2.28	2.34	2.42	2.46	2.51	2.55
6	2.15	2.26	**1.99**	**2.08**	2.13	2.23	2.14	2.29
8	2.19	2.33	2.05	2.17	2.20	2.35	2.09	2.18
9	2.87	3.08	2.48	2.57	2.72	2.84	6.70	4.85

Panel B: Prediction errors for the Hybrid Model with different CNN structures

	Hybrid [8-16-16]+[4]		Hybrid [8-16-16]+[8]		Hybrid [8-16-16]+[4-8]		Hybrid [8-16-16]+[64-16]	
$x_1/x_2/s$	Training (%)	Test (%)	Training (%)	Test (%)	Training (%)	Test (%)	Training (%)	Test (%)
2	2.29	2.40	2.19	2.28	2.16	2.24	2.27	2.39
3	2.21	2.28	2.25	2.32	2.08	2.10	2.13	2.18
4	2.31	2.49	2.51	2.66	2.38	2.52	2.24	2.30
5	2.38	2.48	2.40	2.50	2.28	2.33	2.09	2.14
6	2.13	2.23	2.18	2.29	2.10	2.14	**1.95**	**2.06**
8	2.19	2.33	2.07	2.16	2.10	2.23	2.04	2.10
9	2.87	3.08	2.86	3.07	2.86	3.02	2.62	2.74

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

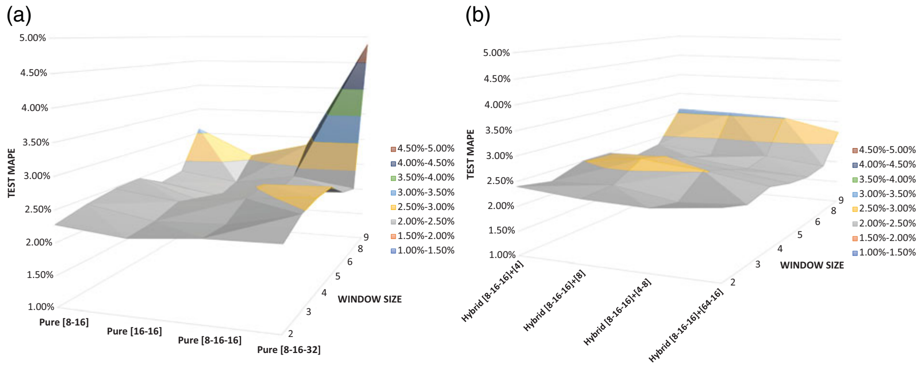

FIGURE 8: 3D illustration of robustness check for CNN structures. (a) Test error of Pure models. (b) Test error of Hybrid models.

1 shows the robustness test results for the Pure Model, and Panel 2 for the Hybrid Model. Each row shows the results for different window sizes of the neighbourhood mortality data, that is, different values of x_1, x_2, and s. Figure 8 illustrates the test errors of different models with 3D plots.

We are relieved to find that the prediction errors are stable for different combinations of CNN architecture parameters. Moreover, we also observe some patterns of trade-off. For example, in general, when window length increases from 2 to 9 years, both training and testing errors first decrease and then increase for all CNN structures. This is because increasing window size will capture more neighbourhood information that can be abstracted by the CNN model and hence obtain more accurate prediction results. However, too large window size will produce more noises. In addition, too large window size will also reduce the effective sample size used to train the model. In practical application of this model, window sizes of 4–6 years in general seem to be reasonable choices. In addition, more complicated CNN structures may help improve the goodness-of-fit of the model; however, they tend to underperform in test sets. For example, moving from left columns to right columns, the complexity of CNN structures is increasing and the prediction errors generally first decrease and then increase. For Pure Models, the structure with the best performance is [16-16], and for Hybrid model, it is [8-16-16]+[64-16].

4.2. Robustness: Methods to handle missing data

In the main results of this paper, we delete missing mortality data. As another robustness check, it is also interesting to see whether different methods to handle missing data will impact the mortality prediction of our proposed model. In this subsection, using Sweden's mortality data as an example (as the U.S. data set does not have missing data issue), we compare three methods: (1) linear interpolation with the non-missing mortality data of the two nearest ages; (2) interpolation according to the Gompertz law ($\log(\mu_x) = a + bx$); and

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

(3) interpolation according to the Gompertz–Makeham law ($\mu_x = a + be^{\gamma x}$). Results are summarised in Table 6. We find that prediction results are robust to different methods.

4.3. Robustness: Leave-One-Population-Out (LOPO) analysis

In this subsection, we perform another interesting robustness check: a completely missing data analysis. For each country/region, we pretend that the mortality data for this particular population are *completely* missing. We construct a neighbouring mortality model using mortality data from other countries, predict the mortality rates for this population, and finally evaluate the performance with the true mortality data from this population that is left out of model estimation. In other words, we are performing a LOPO analysis. More specifically, the LOPO analysis is performed in two steps. First, for each population, we train the model using all data from the other countries/regions, leaving all of its own information out, and calculate corresponding training error. Second, we evaluate the out-of-sample performance by calculating the test error of the trained model using the data of this population that is left out in the first step. It is important to note that traditional mortality models cannot address the completely missing case. We find that the LOPO have satisfying prediction performance. This LOPO exercise further demonstrates the prediction power of the neighbouring prediction models. Our proposed approach is able to learn representation and commonality of data and generalise the discovered intricate structure from the data to guarantee good predicting performance. This is of particular interest in countries where data scarcity is a problem.

We use the HALL model to perform LOPO exercise, and the results are summarised in Table 7. Benchmark models could not address the issue of completely missing data; hence, a LOPO exercise is not possible for them. Interestingly, the neighbouring model produces competitively well-prediction results in the completely missing data case as the benchmarks using full information, with overall test error of 5.05%. It is important to note that in LOPO analysis, training errors are calculated with a large sample that includes data of 40 populations. In contrast, test errors are calculated for each tested population whose information is left out in the model training stage. As a result, in Table 7, we observe that errors are stable for training but not test samples and test errors are generally smaller than training errors. Since the training errors and test errors in the LOPO analysis are calculated on different data sets, they are no longer comparable. The results demonstrate the power of big data in mortality modelling using deep learning algorithms – when data are sufficiently big, the neighbouring prediction model can be general enough to learn the intricate structure and pattern from the data and make reliable predictions. The LOPO exercise results are particularly useful in the case when data are completely unavailable, for example, when life insurance companies plan to investigate a new business line, but no historical data are available yet.

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

TABLE 6

ROBUSTNESS CHECK FOR METHODS TO HANDLE MISSING DATA.

Panel A: Missing value robustness check of Pure model

	Pure [8-16]		Pure [16-16]		Pure [8-16-16]		Pure [8-16-32]	
	Training (%)	Test (%)	Training (%)	Test (%)	Training (%)	Test (%)	Training (%)	Test (%)
Remove	5.99	6.85	6.01	6.94	5.54	6.47	6.11	6.82
Interpolation	5.93	6.68	5.24	6.29	6.08	6.72	5.70	6.42
Gompertz	5.83	6.55	5.56	6.56	6.11	6.62	5.83	6.58
Gompertz–Makeham	5.70	6.51	5.83	6.93	5.82	6.76	5.92	6.71

Panel B: Missing value robustness check of Hybrid model

	Hybrid [8-16-16]+[4]		Hybrid [8-16-16]+[8]		Hybrid [8-16-16]+[4-8]		Hybrid [8-16-16]+[64-16]	
	Training (%)	Test (%)	Training (%)	Test (%)	Training (%)	Test (%)	Training (%)	Test (%)
Remove	5.63	6.58	6.18	6.57	6.11	6.79	5.97	6.46
Interpolation	5.54	6.34	5.50	6.32	5.76	6.43	5.69	6.39
Gompertz	5.94	6.34	5.67	6.34	5.75	6.36	5.66	6.41
Gompertz–Makeham	5.63	6.39	5.65	6.38	5.59	6.33	5.73	6.54

TABLE 7

SUMMARY OF THE COMPLETELY MISSING DATA ANALYSIS.

	LOPO (%)	LOPO (%)		LOPO (%)	LOPO (%)		LOPO (%)	LOPO (%)
	(training)	(test)		(training)	(test)		(training)	(test)
AUS	5.28	4.11	GRC	5.06	3.26	POL	5.63	2.91
AUT	4.99	3.07	HKG	5.20	5.10	PRT	5.46	4.25
BLR	5.32	4.52	HUN	4.98	3.24	KOR	5.29	3.77
BEL	5.44	4.96	ISL	5.72	25.42	RUS	5.52	3.77
BGR	5.03	4.86	IRL	6.54	6.22	SVK	5.41	4.22
CAN	5.02	2.87	ISR	5.31	3.87	SVN	5.55	5.40
CHL	5.50	4.41	ITA	5.36	5.06	ESP	5.60	6.90
HRV	5.52	4.08	JPN	5.33	3.36	SWE	5.41	7.33
CZE	5.06	2.96	LVA	5.15	4.32	CHE	5.57	5.58
DNK	5.23	5.36	LTU	5.46	4.87	TWN	5.41	4.80
EST	5.20	5.13	LUX	5.28	9.27	GBR	5.39	3.33
FIN	5.07	6.12	NLD	5.97	5.49	USA	5.41	2.15
FRA	5.36	4.60	NZL	5.79	5.25	UKR	5.93	3.71
DEU	5.33	2.49	NOR	5.43	4.56			

Overall average errors for 41 countries and regions

LOPO (%)	LOPO (%)
(training)	(test)
5.40	5.05

4.4. Robustness: Extending window prediction

Our main results with the neighbouring prediction are based on splitting the data sample into 80% training set and 20% test set. In this robustness analysis, we evaluate the neighbouring prediction models with an extending window estimation procedure. We start estimating the models with the data in the first 80% years. We then perform one-year-ahead forecast based on the obtained prediction models. We then extend the estimation window by one year, re-estimate the model, and perform one-year-ahead forecast again. This extending window procedure continues until all data are exhausted. The extending window forecasting procedure splits data by preserving temporal order, which avoids look-ahead bias. The results are summarised in Table 8. We can see that the prediction results are similar to the main results presented in previous sections. In fact, the prediction accuracies from the extending window forecasting procedure are slightly higher than using the 80–20% splitting procedure. For example, for HALL model, the average MAPE is 3.91%. This is 2.93% better than the best performed benchmark model and 7.98% better than the worst performed benchmark model. The results demonstrate that the neighbouring prediction models are robust with different prediction procedures. Robustness analyses also show that the prediction performance is robust to different training-test splittings.

TABLE 8

SUMMARY OF EXTENDING WINDOW FORECASTING RESULTS.

	Pure (%)	Hybrid (%)	PALL (%)	HALL (%)		Pure (%)	Hybrid (%)	PALL (%)	HALL (%)
AUS	3.16	3.78	3.78	3.48	JPN	3.67	4.22	3.50	3.13
AUT	4.73	5.19	4.24	3.74	LVA	4.07	4.06	4.57	4.03
BLR	4.67	3.70	3.71	3.15	LTU	5.04	4.82	4.53	4.19
BEL	4.04	3.93	3.63	3.05	LUX	8.90	8.41	9.76	9.11
BGR	5.96	4.73	3.93	3.17	NLD	4.00	4.04	3.48	2.74
CAN	2.94	2.77	2.88	2.26	NZL	4.34	4.60	4.49	4.53
CHL	3.74	7.48	6.24	6.16	NOR	4.02	4.31	3.97	3.53
HRV	4.43	5.21	4.01	3.84	POL	3.37	3.68	2.97	2.56
CZE	3.67	4.07	3.29	3.04	PRT	3.82	4.11	3.72	3.12
DNK	4.15	4.29	3.62	3.17	KOR	3.91	4.37	3.96	4.35
EST	4.81	5.39	5.66	4.89	RUS	4.57	4.12	4.23	3.16
FIN	4.09	4.68	3.81	3.18	SVK	5.01	4.45	4.04	3.55%
FRA	3.89	4.32	3.65	2.94	SVN	4.82	5.17	5.26	5.01
DEU	3.41	3.25	3.66	3.08	ESP	4.45	4.26	3.70	3.24
GRC	4.55	5.81	4.56	4.25	SWE	4.17	4.03	3.50	3.29
HKG	6.08	5.52	5.04	4.55	CHE	4.73	3.32	4.54	3.83
HUN	3.09	3.41	3.23	2.81	TWN	3.64	3.30	3.10	4.27
ISL	16.06	15.61	13.96	13.74	GBR	4.43	4.90	3.17	2.56
IRL	5.69	5.99	4.78	4.32	USA	2.10	2.23	2.43	2.04
ISR	5.40	4.82	4.59	4.24	UKR	4.53	3.55	3.39	2.28
ITA	4.19	4.07	3.42	2.95					

Overall average errors for 41 countries and regions

Pure (%)	Hybrid (%)	PALL (%)	HALL (%)
4.64	4.73	4.34	3.91

5. CONCLUSION

In this paper, we propose a new neighbouring prediction framework for mortality forecasting. The proposed models make use of the neighbourhood mortality data, $\mathcal{E}_{m_{x,t}}(x_1, x_2, s)$, combined with the deep learning algorithm that is suitable for computer vision tasks - 2D CNN. This framework is able to capture the intricate *nonlinear* structure in the mortality data: the neighbourhood effect, which can go beyond the directions of period, age, and cohort. An extensive empirical analysis is conducted using mortality data from all the 41 countries and regions in HMD. We find that the proposed models achieve superior forecasting performance, and they can be further enhanced to model mortality comovements and dependence, by including multi-population data from different countries/regions. More interestingly, in a completely missing data exercise, we find the neighbouring prediction models can produce very satisfying performance even if the model for each particular country/region is trained with data from the other countries/regions, treating all of its own information as missing. These results demonstrate the power of the proposed neighbouring prediction models using big data and deep learning algorithms in mortality forecasting. The neighbouring prediction models proposed in this paper can stimulate better solutions for longevity risk management through providing more sophisticated and accurate mortality models.

Since the focus of this paper is to introduce the new neighbouring prediction models, we do not discuss applying the proposed models in longevity risk management. As a result, an interesting future work path is to investigate pricing mortality- or longevity-linked securities as well as reducing population basis risk with these models. Some interesting questions that follow, among others, include how to perform simulation study and how to evaluate forecast uncertainty for the neighbouring prediction models. We leave the discussion of these questions for future research.

Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/asb.2021.13

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016) Tensorflow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI'16)*, pp. 265–283.

Blake, D., Cairns, A., Coughlan, G., Dowd, K. and MacMinn, R. (2013) The new life market. *Journal of Risk and Insurance*, **80**(3), 501–558.

Blake, D., MacMinn, R., Tsai, J.C. and Wang, J. (2018) Longevity risk and capital markets: The 2017-18 update. Pension Institute Discussion Paper PI-1908.

Bottou, L. and Bousquet, O. (2008) The tradeoffs of large scale learning. *Advances in Neural Information Processing Systems* (eds. M. Jordan, Y. LeCun and S. Solla), pp. 161–168.

Cairns, A.J., Blake, D. and Dowd, K. (2006) A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance*, **73**(4), 687–718.

Cairns, A.J., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A. and 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), 1–35.

Chen, H., MacMinn, R. and Sun, T. (2015) Multi-population mortality model: A factor copula approach. *Insurance: Mathematics and Economics*, **63**, 135–146.

Chollet, F. et al. (2018) Keras: The Python deep learning library. Astrophysics Source Code Library.

Dietterich, T.G. (2000) Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, pp. 1–15. Springer.

Dong, Y., Huang, F., Yu, H. and Haberman, S. (forthcoming) Multi-population mortality forecasting using tensor decomposition. *Scandinavian Actuarial Journal*, pp. 334–356.

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

DOWD, K., CAIRNS, A.J.G., BLAKE, D., COUGHLAN, G.D. and KHALAF-ALLAH, M. (2011) A gravity model of mortality rates for two related populations. *North American Actuarial Journal*, **15**(2), 334–356.

HAINAUT, D. (2018) A neural-network analyzer for mortality forecast. *ASTIN Bulletin: The Journal of the IAA*, **48**(2), 481–508.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. and FRANKLIN, J. (2005) The elements of statistical learning: Data mining, inference and prediction. *The Mathematical Intelligencer*, **27**(2), 83–85.

JARNER, S.F. and KRYGER, E.M. (2011) Modelling adult mortality in small populations: The SAINT model. *ASTIN Bulletin*, **41**(2), 377–418.

KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G.E. (2012) Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pp. 1097–1105.

LECUN, Y., BENGIO, Y. and HINTON, G. (2015) Deep learning. *Nature*, **521**(7553), 436–444.

LECUN, Y., BOTTOU, L., BENGIO, Y. and HAFFNER, P. (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**(11), 2278–2324.

LEE, R.D. and CARTER, L.R. (1992) Modeling and forecasting US mortality. *Journal of the American Statistical Association*, **87**(419), 659–671.

LI, J.S.-H., CHAN, W.-S. and ZHOU, R. (2017) Semicoherent multipopulation mortality modeling: The impact on longevity risk securitization. *Journal of Risk and Insurance*, **84**(3), 1025–1065.

LI, N. and LEE, R. (2005) Coherent mortality forecasts for a group of population: An extension to the classical Lee-Carter approach. *Demography*, **42**(3), 575–594.

PERLA, F., RICHMAN, R., SCOGNAMIGLIO, S. and WUTHRICH, M.V. (2021) Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal*, pp. 1–27.

RENSHAW, A.E. and HABERMAN, S. (2006) A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, **38**(3), 556–570.

RICHMAN, R. (2018) AI in actuarial science. Available at SSRN 3218082.

RICHMAN, R. and WÜTHRICH, M.V. (forthcoming) A neural network extension of the Lee–Carter model to multiple populations. *Annals of Actuarial Science*, pp. 1–21.

WANG, C.-W., YANG, S.S. and HUANG, H.-C. (2015) Modeling multi-country mortality dependence and its application in pricing survivor index swaps—a dynamic copula approach. *Insurance: Mathematics and Economics*, **63**, 30–39.

WANG, H.-C., YUE, C.-S.J. and CHONG, C.-T. (2018) Mortality models and longevity risk for small populations. *Insurance: Mathematics and Economics*, **78**, 351–359.

ZHOU, R., LI, J.S.-H. and TAN, K.S. (2013) Pricing standardized mortality securitizations: A two-population model with transitory jump effects. *Journal of Risk and Insurance*, **80**(3), 733–774.

ZHOU, Y.-T. and CHELLAPPA, R. (1988) Computation of optical flow using a neural network. *IEEE International Conference on Neural Networks*, vol. 1998, pp. 71–78.

ZHU, W., TAN, K.S. and WANG, C.-W. (2017) Modeling multicountry longevity risk with mortality dependence: A Lévy subordinated hierarchical Archimedean copulas approach. *Journal of Risk and Insurance*, **84**(S1), 477–493.

CHOU-WEN WANG
Department of Finance
National Sun Yat-Sen University
Kaohsiung, Taiwan
Risk and Insurance Research Center
College of Commerce, National Chengchi University
Taipei, Taiwan
E-Mail: chouwenwang@mail.nsysu.edu.tw

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press

JINGGONG ZHANG
Nanyang Business School
Nanyang Technological University
Singapore
E-Mail: jgzhang@ntu.edu.sg

WENJUN ZHU (Corresponding author)
Nanyang Business School
Nanyang Technological University
Singapore
E-Mail: wjzhu@ntu.edu.sg

https://doi.org/10.1017/asb.2021.13 Published online by Cambridge University Press