A neural network analyzer for mortality forecast

Donatien Hainaut[†]

September 29, 2017

[†] *ISBA, Université Catholique de Louvain.*
Email: donatien.hainaut (@) uclouvain.be

Abstract

This article proposes a neural network approach to predict and simulate human mortality rates. This semi-parametric model is capable to detect and duplicate non-linearities observed in the evolution of log-forces of mortality. The method proceeds in two steps. During the first stage, a neural network based generalization of the principal component analysis summarizes the information carried by the surface of log-mortality rates in a small number of latent factors. In the second step, these latent factors are forecast with an econometric model. The term structure of log-forces of mortality is next reconstructed by an inverse transformation. The neural analyzer is adjusted to French, UK and US mortality rates, over the period 1946-2000 and validated with data from 2001 to 2014. Numerical experiments reveal that the neural approach has an excellent predictive power, compared to the Lee-Carter model with and without cohort effects.

Keywords : Longevity, mortality, Lee-Carter, neural network, perceptron.

1 Introduction

The improvement of longevity observed over the last century is a matter of concerns for the insurance industry. This growth is explained by the reduction of mortality caused by infectious and chronic diseases at older ages. This trend calls for more advanced techniques to manage the longevity risk. A popular framework for mortality rates is the model of Lee and Carter (1992). In their approach, the log-force of mortality is the sum of a fixed age component and of an age specific function multiplied by a time component. The robustness of this approach contributes to its success among practitioners. We refer the reader to Lee (2000), Pitacco (2004), Wong-Fupuy and Haberman (2004) or Cairns (2008) for a review of various extensions of the Lee-Carter (LC) model. Renshaw and Haberman (2003) propose a multi-factors version and Renshaw and Haberman (2006) augment the one factor LC model with a cohort effect.

Three approaches exist for estimating the LC model and its various extensions. The first one, pioneered by Lee and Carter (1992), counts two steps. In the first stage, age components and latent time processes are obtained by a PCA. In the second step, an autoregressive model or a random walk is fitted to forecast the time effect. Yang et al. (2010) compare the performance of this calibration procedure for various models. Toczydlowska et al. (2017) propose a probabilistic extension of PCA in a state space framework.

The second approach of calibration is based on generalised linear models (GLM). Brouhns et al. (2002) use a Poisson distribution and estimate parameters of a LC model by loglikelihood maximization. Renshaw and Haberman (2006) adapt this approach for estimating a LC model with cohort effects. O'Hare and Li (2012) study the mortality at young ages and Van Berkum et al. (2016) detect structural change in the evolution of mortality with GLM. The recent article of Currie (2016) provides a comprehensive survey on generalized linear and non-linear models of mortality.

The third method for estimating parameters consists to perform the joint inference of latent time processes and age parameters, in a single step by a Markov Chain Monte-Carlo (MCMC) method. Antonio et al. (2015) apply this Bayesian approach to predict the joint mortality of multiple populations. Fung et al. (2016, 2017) propose a state-space framework for mortality modelling with cohort effects. This approach is computationally intensive but remedies to the drawback of two steps procedures that are somewhat ad-hoc methods, without statistical foundations. In comparison to a state space estimation, Fung et al. (2015) show that estimating mortality models with a two steps procedure leads to underestimate annuity prices.

The first approach for estimating mortality models is based on PCA. The PCA can be regarded as an extraction method that attempts to characterize lower-dimensional structure in large multivariate datasets. When the data has a nonlinear structure, as it is the case for mortality rates, it will not be detected by a PCA. In the early 1990s, a neural network based generalization of PCA to the nonlinear feature extraction problem was introduced by Kramer (1991) in the chemical engineering literature, who referred to the resulting technique as nonlinear principal component analysis (NLPCA). Another solution to this problem comes from Hastie and Stuetzle (1989), who named their method principal curves and surfaces (PCS). Malthouse (1998) demonstrated that NLPCA and PCS are closely related. Kramer's NLPCA has been applied in various field: chemical engineering (Dong and McAvoy 1996) to psychology (Fotheringhame and Baddeley 1997) or climatic mathematics (Monahan, 2000). In this article, we use a NLPCA so as to summarize the surface of mortality rates.

Our work contributes to the literature in several directions. Firstly, the proposed model for mortality is non-linear and semi-parametric. Secondly, we use a neural analyzer for identifying latent time processes. To the best of our knowledge, only a few research

articles apply neural networks to forecast mortality and in existing studies, the neural net is substituted to an econometric model or to a linear regression. For example, Atsalakis et al. (2007) propose a neural network with fuzzy logic inference. Abdulkarim and Garko (2015) fit a feed-forward neural network with a particle swarm algorithm so as to forecast the maternal mortality in a region of Nigeria. Puddu and Menotti (2009) use a multilayer perceptron to predict the coronary heart disease mortality in seven countries. Puddu and Menotti (2012) extend this approach to predict to predict the 45-year all-cause mortality in Italian rural areas and they don't observe any difference between the performance of multilayer perceptrons or multiple logistic regressions. However, previous studies do not use a neural net to forecast and simulate the complete term structure of mortality rates and are then not adapted for actuarial applications. Furthermore, the neural network is not used directly to predict the mortality. Instead, it summarizes the evolution of mortality curves into latent factors that are extrapolated with a random walk. The term structure of mortality rates is next reconstructed with the neural network.

Notice that our approach to estimate parameters is a two steps procedure. In the first stage, we fit the neural net and filter latent processes representative of time effects. In the second stage, we fit a random walk to forecast the evolution of these latent variables. In our framework, the neural net defines the non linear function between latent processes and observed mortality rates. In theory, it is then possible to perform the joint inference of latent time processes and neural net parameters, in a single step by a Markov Chain Monte-Carlo (MCMC) method. However, this approach would be more computationally intensive.

In numerical applications, the core of our analysis focuses on mortality rates of the French population from 1946 to 2014. This sample of data is partitioned in two subsets. The first one contains observations from 1946 to 2000 and serve us to calibrate the neural analyzer. Whereas the second subset of mortality rates from 2001 to 2014 is used for validation. The neural net is benchmarked to several extensions of the Lee-Carter model. We compare it to the original and multi-factors Lee Carter model, estimated with a PCA. We also consider the Lee Carter model, fitted by log-likelihood maximization in a GLM framework. Finally, we compare the neural analyzer to the Lee Carter model with age specific cohort effect, as proposed by Renshaw and Haberman (2006). All numerical experiments conclude that the neural analyzer has an excellent predictive power compared to the LC model. Finally, the calibration of the neural analyzer to UK and US mortality data confirms the robustness of our conclusions.

2 The Lee-Carter model and its extensions

Lee and Carter (1992) proposed a pioneering model for mortality forecasting. They assumed that log-forces of mortality have a linear structure with respect to time and that

covariates depend upon the age. This model became rapidly a standard in the industry due to its robustness and easiness of implementation. Renshaw and Haberman (2003) extended this framework by proposing a multi-factor model which provides a better fit of mortality at old ages. Their approach is based on a principal component analysis and for most of populations, three factors explain at least three quarters of the variance of centered log-mortality rates. Renshaw and Haberman (2006) studied a model with age-specific cohort and period effects. They also adapted Wilmoth (1993) and Brouhns et al. (2002a) to estimate the Lee Carter model by log-likelihood maximization, under Gaussian and Poisson error structures. As we use these models as benchmark to measure the efficiency of the neural network approach, we briefly review them in this section.

Throughout the rest of this article, the time of decease of an individual of age x is assimilated to the first jump of a non-homogeneous Poisson process, that is denoted $(N_t^x)_{t\geq 0}$. The intensity of the mortality Poisson process is called the force of mortality or the mortality rate and depends upon time t and age x. It is denoted by $\mu(t,x)$ and may be interpreted as the instantaneous probability of death at time t, for a x year old human. The mortality rate is also related to the probability of survival till time $s \geq t$ by the following relation

$$\begin{aligned} {}_sp_x &:= P(\tau \geq s) \\ &= \mathbb{E}(N_s^x \geq 1 \,|\, N_t^x) \\ &= \exp - \int_t^s \mu(s, x+s-t)\, ds\,. \end{aligned}$$

On the other side, the probability of dying at age x during year t, is the complementary probability of ${}_tp_x$ defined by

$$q(t,x) := 1 - \exp\left(- \int_t^{t+1} \mu(s, x+s-t)\, ds \right).$$

In the original Lee Carter model, the log mortality rates are related to ages as follows:

$$\ln \mu(t,x) = \alpha_x + \beta_x \kappa_t\,. \tag{1}$$

where $\alpha_x \in \mathbb{R}^{x_{max}}$ is a constant vector representing the permanent impact of age on mortality. Whereas $\beta_x \in \mathbb{R}^{x_{max}}$ is a constant vector that quantify the marginal effect of the latent factor κ_t on mortality at each age. κ_t is a latent process that describes the evolution of mortality over time. Notice that the Lee-Carter model is itself a parametric version of the Cox's model (1972), in which covariates are replaced by time dependent latent factors[1].

[1]Let us consider T , a duration random variable of hazard rate: $h_\theta(t) = \theta h(t)$. If $h(t)$ is a deterministic function and $\theta = \exp\left(\beta^\top z\right)$ with z, the vector of covariates at time t, then $\ln h(t|z) = h(t) + \beta^\top z$ which is the LC model if covariates are time dependent latent processes.

Actuarial models may distort reality as they only use the surface of log-mortality rates as input. Authorizing insurers to have a broader access to individual's medical data, would allow a better segmentation of risks. This would permit to identify competing risks which are critical to calculate survival probabilities, as underlined in the paper of Dimitrova et al. (2013) or Puddu et al. (2017). Indirectly, this would also contribute to the development of new models with other covariates than just the insured's age. But for the moment, the lack of data prevents such an evolution.

The Lee Carter model is estimated by a two-stage procedure looking first at the observation equation as a regression (ignoring the latent factor structure explicitly). In the second stage time-series models are adjusted to latent factors. In Lee and Carter (1992), this regression is performed by a singular value decomposition (SVD). This approach being well documented in the literature, we refer to e.g. Pitacco et al. (2009) for details. Renshaw and Haberman (2006) adapted Wilmoth (1993) and Brouhns et al. (2002a) to estimate the LC model by log-likelihood maximization, in a GLM framework with a Gaussian error structure. We will compare these two methods of calibration in numerical applications. To ensure the identifiability of the model, two constraints are imposed during the calibration:

$$\sum_x \beta_x = 1 \qquad \sum_t \kappa_t = 0\,. \tag{2}$$

In multi-factors extensions of the Lee-Carter model proposed by Renshaw and Haberman (2003), the log-force of mortality is a linear combination of d time latent factors noted $\kappa_t^{i=1,\dots,d}$, with covariates that depend on the age as follows:

$$\ln \mu(t,x) \;=\; \alpha_x + \sum_{i=1}^{d} \beta_x^i \kappa_t^i\,. \tag{3}$$

Where the $\beta_x^{i=1\dots d} \in \mathbb{R}^{x_{max}}$ are constant vectors such that $\sum_x \beta_x^i = 1$. $\kappa_t = (\kappa_t^i)_{i=1\dots d}$ are d latent processes satisfying the constraint $\sum_t \kappa_t^i = 0$ for $i = 1\dots d$, to ensure the identifiability. This model is estimated by a SVD. The last model that we consider, adds a cohort effect in the dynamic of log-force of mortality:

$$\ln \mu(t,x) = \alpha_x + \beta_x \kappa_t + \beta_x^g \gamma_{t-x}\,, \tag{4}$$

where $\beta_x^g \in \mathbb{R}^{x_{max}}$ represents the marginal effect of a generation factor, γ_{t-x} , on mortality. Renshaw and Haberman (2006) estimate this model by log-likelihood maximization, in a general linear model (GLM) framework. We refer the interested reader to their article for details about the estimation procedure. Table 1 summarizes the models to which our neural network analyzer is compared in the sequel. It also presents methods of calibration used for each approach.

	Calibration	symbol	log mortality
Multifactor Lee Carter	SVD	LC SVD	$\ln \mu(t,x) = \alpha_x + \sum_{i=1}^{d} \beta_x^i \kappa_t^i$
1D Lee Carter	Loglikelihood maximization	LC GLM	$\ln \mu(t,x) = \alpha_x + \beta_x \kappa_t$
Lee Carter with cohort effects	Loglikelihood maximization	LC COH	$\ln \mu(t,x) = \alpha_x + \beta_x \kappa_t + \beta_x^g \gamma_{t-x}$

Table 1: Summary of models to which the neural net approach is compared.

In the second stage of the calibration procedure, a time-series model is specified for the latent processes. Most of authors use an AR(1) model or a random walk with drift. In this paper, we opt for the second choice and assume that increments of κ_t^i are Gaussian random variables with a mean γ_i and a variance σ_i^2:

$$\kappa_t^i - \kappa_{t-1}^i \quad = \quad \gamma_i + \sigma_i \, \epsilon_t \quad i = 1, ..., d \tag{5}$$

where ϵ_t is a standard normal random variable. Other dynamics, like the switching regime diffusion in Hainaut (2012) have been proposed so as to detect a change of trends in the evolution of mortality. But as the random walk model became the standard in the industry, we adopt it as reference to forecast future mortality rates by simulations. In the numerical illustration, we use a Jarque-Bera test to validate the hypothesis that increments of κ_t^i are normally distributed, at least during the most recent decades.

3 The neural net analyzer

The main assumption underlying the LC model and its extensions is the linear dynamic of log-forces of mortality. This specification justifies to apply the PCA to fit latent stochastic processes and age effects. PCA can be regarded as an extraction method that attempts to characterize lower-dimensional structure in large multivariate datasets. If the underlying distribution is Gaussian, then PCA is an optimal feature extraction algorithm. However, if the data has a non-linear structure, as it could be the case for mortality rates, the PCA fails to detect it.

In the early 1990s, a neural-network-based generalization of PCA was introduced by Kramer (1991) in the chemical engineering literature, who referred to the resulting technique as the nonlinear principal component analysis (NLPCA). Directly inspired from the literature on neural networks, we propose here a neural net analyzer that detects the nonlinearities in the lower-dimensional structure of the log-forces of mortality.

In our datasets, the available mortality forces range from year t_{min} to t_{max} and from age

x_{min} to x_{max}. The number of observations for a given year is noted $n_x = x_{max} - x_{min}$. Available demographic data contains the number of deaths aged x per year, $d_{x,t}$, and the exposure to risk, $E_{x,t}$. Notice that $E_{x,t}$ is measured by the size of the population aged x last birthday in the middle of the observation year t. The death probability is then approached by $q_x = \frac{d_{x,t}}{E_{x,t}}$. Under the assumption that the force of mortality is a stepwise constant function on $[t, t+1[\times[x, x+1[$, we calculate it as follows:

$$\mu(s,y) \;=\; -\ln\left(1 - q(t,x)\right) \quad \forall\, s \in [t,\, t+1[\; y \in [x,\, x+1[\,.$$

To compare our results with these yield by other models, we use as input for the neural net the centered log-forces of mortality, denoted by:

$$X(t) := \begin{pmatrix} \ln \mu(t, x_{min}) - \alpha_{x_{min}} \\ \vdots \\ \ln \mu(t, x) - \alpha_x \\ \vdots \\ \ln \mu(t, x_{max}) - \alpha_{x_{max}} \end{pmatrix} \quad t = t_{min}, ..., t_{max}\,.$$

$X(t)$ is a vector of dimensions $n_x = x_{min} - x_{max}$ and α_x is the vector of average log-mortality rates:

$$\alpha_x \;=\; \frac{1}{t_{max} - t_{min} + 1} \sum_{t=t_{min}}^{t_{max}} \ln \mu(t,x) \qquad x = x_{min}, \ldots, x_{max}. \tag{6}$$

We aim to determine two functions: an encoding and a decoding function. We denote these functions by $f^{enc} : \mathbb{R}^{n_x} \to \mathbb{R}^d$ and $f^{dec} : \mathbb{R}^d \to \mathbb{R}^{n_x}$. The encoding function, $f^{enc}(.)$, is nonlinear and projects curves of mortality rates at time t, $X(t) \in \mathbb{R}^{n_x}$ on a hyperplan of lower dimensions, in $\mathbb{R}^d$. As in the multi-factor LC model, the coordinates of the projection in $\mathbb{R}^d$ are contained in a d-plet $\kappa_t^{nn} := (\kappa_t^{nn,1}, ..., \kappa_t^{nn,d})$ such that

$$\kappa_t^{nn} := f^{enc}(X(t)) \quad t = t_{min}, ..., t_{max}\,.$$

The decoding function $f^{dec}(.)$ uses this summarized information to build an approximation $\hat{X}(t) \in \mathbb{R}^{n_x}$ of the initial curve of log-mortality rates:

$$\hat{X}(t) \;:=\; f^{dec}\left(\kappa_t^{nn}\right).$$

Compared to the original LC model, the linear relation $\beta_x \kappa_t$ is replaced by a non-linear function and log-mortality forces are ruled by the following relation:

$$\begin{aligned} \ln \mu(t,x) &= \alpha_x + f^{dec}\left(x, \kappa_t^{nn}\right) \\ &= \alpha_x + f^{dec}\left(x, f^{enc}\left(X(t)\right)\right) \end{aligned} \tag{7}$$

The encoding and decoding functions are calibrated so as the minimize the sum of squared residuals between initial and reconstructed mortality curves:

$$\left(f^{enc}, f^{dec}\right) = \arg\min \sum_{t=t_{min}}^{t_{max}} \left\| X(t) - \hat{X}(t) \right\|_2^2 . \tag{8}$$

In the neural analyzer net, the functions $f^{enc}(.)$, $f^{dec}(.)$ are approximated by two feed-forward neural networks. A neural net is a series of parallel layers of interconnected neurons. A neuron in the i^{th} layer receives as input, the output of neurons located in the previous layer. Let n_j be the number of neurons in layer j. The output of the i^{th} neurons in layer j , denoted by $y_{i,j}$, is computed as follows:

$$y_{i,j} = \phi_{i,j}\left(\sum_{k=1}^{n_{j-1}} \omega_{i,k}^{j} y_{k,j-1}\right)$$

where $\omega_{i,k}^{j}$ are the weights and $\phi_{i,j}()$ is a transfer function. In our framework, two transfer functions are used. The first one is the hyperbolic tangent sigmoid function, $\phi_{sig}(z) : \mathbb{R} \to (-1, 1)$, defined by

$$\phi_{sig}(z) = \frac{2}{1 + \exp(-z)} - 1 .$$

The second transfer function is the identity function: $\phi_{id}(z) = z$. Cybenko (1989) demonstrates that finite linear combinations of fixed univariate functions with a set of affine functionals can uniformly approximate any continuous function with support in the unit hypercube. Hornik (1991) shows that it is not the specific choice of the activation function, but rather the multilayer feedforward architecture itself which gives neural networks the potential of being universal approximators. He also proves that a three layers neural network with n_1 input neurons, hyperbolic transfer functions in the second layer, and linear transfer functions in the third layer of n_2 neurons can approximate to arbitrary accuracy any continuous function from $\mathbb{R}^{n_1}$ to $\mathbb{R}^{n_2}$ at the condition that the number of neurons in the second layer is large enough. These fundamental results justify our approach that consists to define f^{enc} and f^{dec} by feed-forward neural networks. We test the architecture recommended in McNelis (2005) and presented in figure 1.

The input and output layers count the same number of neurons, n_l, with a hyperbolic tangent sigmoid transfer function. Mortality log-forces, $X(t)$ are divided into n_l groups of $n_c = \frac{n_x}{n_l}$ elements. Each subgroup of data is sent exclusively to a single neuron of the input layer. The central layer is a bottleneck with d neurons that have a linear transfer function. The encoding phase is then summarized by the following two operations:

$$\begin{aligned} y_{i,1}(t) &= \phi_{sig}\left(\sum_{k=1}^{n_x} \omega_{i,k}^{1} X_k(t)\right) \quad i = 1, ..., n_l \\ \kappa_t^{nn,i} &= \sum_{k=1}^{n_l} \omega_{i,k}^{2} y_{k,1}(t) \quad i = 1, ..., n_d , \end{aligned}$$

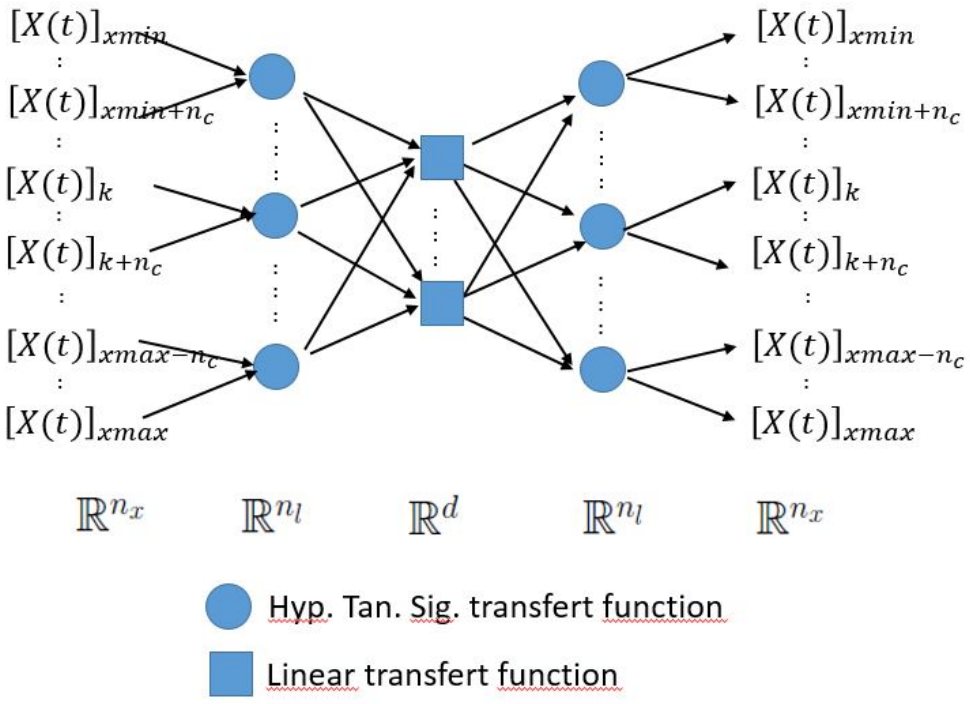

Figure 1: Architecture of the neural analyzer

where $\omega^1_{i,k} \neq 0$ if $x_{min} + (i-1)\, n_c \leq k < x_{min} + i\, n_c$ and $\omega^1_{i,k} = 0$ otherwise. Whereas the decoding phase is given by the following two steps:

$$\begin{aligned} y_{i,3}(t) &= \phi_{sig}\left(\sum_{k=1}^{n_d} \omega^3_{i,k} \kappa^{nn,k}_t\right) \quad i = 1, ..., n_l \\ \hat{X}_i(t) &= \sum_{k=1}^{n_l} \omega^4_{i,k} y_{k,3}(t) \quad i = 1, ..., n_x\,, \end{aligned}$$

where $\omega^4_{i,k} \neq 0$ if $x_{min} + (k-1)\, n_c \leq i < x_{min} + k\, n_c$ and $\omega^4_{i,k} = 0$ otherwise. The weights $\omega^j_{i,k}$ are calibrated by minimizing the quadratric spread between the input and the output, as defined in equation (8). The dimension of $X(t)$ being high, the number of parameters to calibrate is important. Applying a gradient method to adjust the network is then slow and the risk of staying eventually trapped in local minimum during the gradient descent is non negligible. For this reason, we fit the neural analyzer with a genetic algorithm that is described in the next section.

As for the LC model, we calibrate next n^d random walks each component of the d-plet $\kappa^{nn}_t := (\kappa^{nn,1}_t, ..., \kappa^{nn,d}_t)$:

$$\kappa^i_t - \kappa^i_{t-1} = \gamma^{nn}_i + \sigma^{nn}_i\, \epsilon_t \quad i = 1, ..., d \tag{9}$$

where ϵ_t is a standard normal random variable. We will see in the numerical application that despite its relative simplicity, a random walk process provides an excellent statistical fit to the time-series of $\kappa_t^{nn,i}$.

4 Genetic Algorithm (GA)

The main issue inherent to mortality models is the high dimension of the vector of parameters. For this reason, estimating the weights of the neural net with a gradient descent is time consuming. Instead, we adopt a two steps strategy. In the first stage, we start the search of optimal parameters with a genetic algorithm (GA), developed by McNelis (2005). We use next the solution found by this GA algorithm as starting point of a gradient descent. The GA algorithm is a powerful evolutionary search process that proceeds in five steps.

1) The first step consists to create a population of candidate parameters. Contrary to a gradient descent, a genetic algorithm does not start with one initial vector of parameters, but with an initial population of N^* (an even number) coefficient vectors, called the first generation. Letting n^ω be the total number of coefficients to estimate, the first generation is the set of n^ω by 1 random vectors: $P = \{P_1, P_2 \ldots, P_{N^*}\}$.

2) The second step is called the selection. We choose randomly two pairs of coefficients from the population, with replacement. We evaluate the goodness of fit for these four coefficient vectors, in two pair-wise combinations, according to the quadratic error function defined by equation (8). Coefficient vectors that come closer to minimizing the sum of errors receive better scores. This is a simple tournament between the two pairs of vectors: the winner of each tournament is the vector with the best scores. These two winning vectors P_1, $P_2 \in \Omega$ are retained for "breeding" purposes.

3) The third step is the crossover, in which the two parents, selected during the second stage, "breed" two children. The algorithm allows crossover to be performed on P_1 and P_2, with a fixed probability $p > 0$. The algorithm uses one of three different crossover operations, with each method having an equal 1/3 probability of being chosen:

a. Shuffle crossover. We draw n^ω random numbers from a binomial distribution. If the m^{th} draw is equal to 1, the coefficients $P_{1,m}$ and $P_{2,m}$ are swapped; otherwise, no change is made. The two vectors resulting from these swaps are the children, denoted by C_1 and C_2.

b. Arithmetic crossover. A random number is chosen, $\delta \in (0, 1)$. This number is used to create two children that are linear combinations of the two parent factors, $C_1 = \delta P_1 + (1 - \delta)P_2$ and $C_2 = (1 - \delta)P_1 + \delta P_2$.

c. Single-point crossover. For each pair of vectors, an integer I is randomly chosen from the set$[1, n^{\omega} - 1]$. The two vectors are then cut at integer I and the coefficients to the right of this cut point, $P_{1,1:I+1}$ and $P_{2,1:I+1}$ are swapped to produce C_1 and C_2.

Following the crossover operation, each pair of parent vectors gives birth to two children coefficient vectors. In case of no crossover, the children are copies of their parents: $C_1 = P_1$ and $C_2 = P_2$.

4) The fourth step is the mutation of children. With some small probability p^{mut} that decreases over time, each coefficient of the two children's vectors is subjected to a mutation. We can draw a parallel between this step and the simulated annealing method. Simulated annealing is a probabilistic technique, introduced by Khachaturyan et al. (1979) for approximating the global optimum of a given function. Here, the probability of each element is subject to mutation in generation $G = 1, 2, ..., G^*$ given by the probability $p^{mut} = 0.15 + \frac{0.33}{G}$. G is the generation number, G^* is the maximum number of generations. If mutation is to be performed on a vector element, we use a non-uniform mutation operation, due to Michalewicz (1996) and recommended by McNelis (2005). We draw two random numbers r_1 and r_2 from the [0, 1] interval and one random number s from a standard normal distribution. The mutated coefficient $\tilde{C}_{i,k}$ for $i = 1, 2$ and $k = 1$ to n^{ω} is given by the following formula:

$$\tilde{C}_{i,k} = \begin{cases} C_{i,k} + s\left(1 - r_2^{(1-\frac{G}{G^*})^b}\right) & if\ r_1 > 0.5 \\ C_{i,k} - s\left(1 - r_2^{(1-\frac{G}{G^*})^b}\right) & if\ r_1 \leq 0.5 \end{cases}$$

where b is a parameter that governs the degree to which the mutation operation is non-uniform. We set $b = 2$. With this approach, the probability of creating via mutation a new coefficient that is far from the current coefficient value diminishes as $G \to G^*$, where G^* is the number of generations. Thus, the mutation probability itself evolves through time. McNelis (2005) mentions that the mutation operation is non-uniform since, over time, the algorithm is sampling increasingly more intensively in a neighborhood of the existing coefficient values. This more localized search allows for some fine tuning of the coefficient vector in the later stages of the search, when the vectors should be approaching close to a global optimum.

5) The fifth and last step is the election tournament. Following the mutation operation, the four members of the "family" (P_1, P_2, C_1, C_2) engage in a tournament. The score of children and parents is measured by their quadratric errors, as defined by equation (8). The two vectors with the best goodness of fit, whether parents or children, survive and pass to the next generation, while the two with the worst score are extinguished.

The above process is repeated, with parents returning to the population pool for pos-

sible selection again, until the next generation is populated by N^* vectors.

Once the next generation is populated, we introduce elitism. It consists to evaluate all the members of new and past generations according to the score. If the best member of the older generation performs better than the best member of the new generation, this member replaces the worst member of the new generation. One continues this process for G^* generations. The literature gives us little guidance about selecting a value for G^*. Since we evaluate convergence by the score of the best member of each generation, G^* should be large enough so that we see no changes in the fitness values of the best for several generations.

5 Application to the French population

This section focuses on mortality rates observed for the French population over the period 1946 to 2014. The data set is provided by the Human Mortaty Database [2]. Years before 1946 are excluded from the scope of the study given the perturbations on mortality caused by the first and second world wars. The ages considered range from 20 to 109 years. The LC models and the neural networks are calibrated with mortality curves from year 1946 up to 2000. To compare the predictive capability of models, log-forces of mortality are projected by simulations over fourteen years (10 000 simulations) and their average is compared with the observed mortality during the period 2001-2014.

The table 2 reports the calibration errors of LC models with one to three latent factors fitted with a SVD (LC SVD), of the LC model fitted by loglikelihood maximization (LC GLM) and of the LC model with cohort effect (LC COH). We present the sum of squared errors and the average of errors between observed and modeled log-forces of mortality. The table also provides the maximum and minimum spreads and the number of fitted parameters. An analysis of these figures reveals that calibrating the LC model with a SVD leads to a higher quadratic error than the one obtained with statistical approaches. The best fit is obtained with the model that includes a cohort effect.

[2] www.mortality.org

French population, 1946-2000					
Model	Coef.	$\sum \left\|X - \hat{X}\right\|_2^2$	Avg. $\left\|X - \hat{X}\right\|_2$	$\max\left(X - \hat{X}\right)$	$\min\left(X - \hat{X}\right)$
LC SVD 1	180	152.50	0.0024	0.8567	-0.4330
LC SVD 2	270	139.35	0.0023	0.6398	-0.4352
LC SVD 3	360	134.61	0.0023	0.6364	-0.4350
LC GLM	180	29.31	0.0010	0.47017	-0.5515
LC COH	270	10.78	0.0006	0.28573	-0.2220

Table 2: Goodness of fit for variants of the LC model. The first and second columns report the number of latent factors and fitted coefficients. The third and fourth columns present the sum of squared errors and the average errors. The two last columns contain the maximum and minimum errors.

We mention in section 2 that it is common to assume that increments of latent processes κ_t^i follow a random walk with drift. This hypothesis of normality is tested in table 3 with the Jarque-Bera (JB) test, for the increments of a 3 dimensions LC model observed over the period 1970-2000. The JB statistics clearly confirm this assumption. However, the same test applied to the sample of increments over the whole period of calibration (1946-2000) leads to the rejection of normality for the second latent factor. We can draw a parallel with the conclusions of Hainaut (2012) who fits a switching regime process to latent processes. This analysis clearly reveals a change of regime between 1960 and 1970. The same conclusions apply to latent processes of LC GLM and LC COH models. This change of trend may be explained by the reduction of mortality caused by coronary heart diseases, following two vast prevention campaigns launched during the sixties. For this reason, the random walks used in simulations to predict the evolution of log-forces of mortality are fitted to increments of κ_t^i observed only between 1970 and 2000.

		Jarque Bera statistics for 3D Lee Carter		
factors	Normality	p-value	JB statistic	Critical Value 5%
$\kappa_t^1 - \kappa_{t-1}^1$	Accept	0.1679	2.0286	4.4466
$\kappa_t^2 - \kappa_{t-1}^2$	Accept	0.3555	1.2713	4.4466
$\kappa_t^3 - \kappa_{t-1}^3$	Accept	0.3355	1.3327	4.4466

Table 3: Jarque Bera test applied to increments of latent factors over the period 1970-2000, for the LC SVD 3 model.

French population, 2001-2014					
Dim.	Coef.	$\sum \left\|X - \hat{X}\right\|_2^2$	Avg. $\left\|X - \hat{X}\right\|_2$	$\max\left(X - \hat{X}\right)$	$\min\left(X - \hat{X}\right)$
LC SVD 1	180	38.37	0.0049	0.7006	-0.1297
LC SVD 2	270	38.89	0.0049	0.7088	-0.1293
LC SVD 3	360	38.50	0.0049	0.6618	-0.1287
LC GLM	180	11.68	0.0027	0.5532	-0.1724
LC COH	270	17.65	0.0026	0.5329	-0.1377

Table 4: Predictive goodness of fit for variants of the LC model. The first and second columns report the number of latent factors and fitted coefficients. The third and fourth columns present the sum of squared errors and the average errors. The two last columns contain the maximum and minimum errors.

The results about the predictive capability of LC models are reported in table 4. An analysis of the sum of squared errors emphasizes that the performance of models fitted by SVD with one to three factors are nearly identical. The predictive capability of the model with a cohort effect is slightly less good than the one of a LC model estimated by loglikelihood maximization. These figures will be compared to these obtained with the neural net analyzer in the next paragraphs.

The neural network is fitted to the same dataset of log-forces of mortality from 1946 to 2000. Several neural architectures are tested: from 3 to 8 neurons for the input/output layers and 2 to 3 neurons for the intermediate layer. The size of populations in the genetic algorithm is set to 100 vectors of candidate parameters and we consider 500 generations. The time to calibrate the neural net on a personal computer varies between five to fifteen minutes, depending on the processor.

The calibration errors are reported in table 5. A comparison with errors presented in table 2 confirms that the neural analyzer outperforms LC models fitted by SVD and provides a comparable or better fit than LC GLM and LC COH, depending upon the configuration of neurons. Increasing the number of neurons in the input/output layer improves the goodness of fit. The quadratic error obtained with a 8-3-8 neural net (8 input/output and 3 intermediate neurons) is lower than the one for the LC COH model. This confirms that the neural net approach captures age-specific cohort effects.

French population, 1946-2000.						
n_l	n_d	Coef.	$\sum \left\|X - \hat{X}\right\|_2^2$	Avg. $\left\|X - \hat{X}\right\|_2$	$\max\left(X - \hat{X}\right)$	$\min\left(X - \hat{X}\right)$
3	2	552	15.04	0.0008	0.2988	-0.3953
4	2	736	13.37	0.0007	0.3066	-0.3862
5	2	920	12.64	0.0007	0.3088	-0.3678
6	2	1104	12.18	0.0007	0.3047	-0.3603
7	2	1288	12.15	0.0007	0.3142	-0.3678
8	2	1472	11.97	0.0007	0.3085	-0.3648
3	3	558	14.83	0.0008	0.3027	-0.3961
4	3	744	12.64	0.0007	0.3081	-0.3894
5	3	930	11.85	0.0007	0.3072	-0.3721
6	3	1116	11.56	0.0007	0.3101	-0.3658
7	3	1302	10.68	0.0007	0.2896	-0.3336
8	**3**	**1488**	**9.71**	**0.0006**	**0.2844**	**-0.3141**

Table 5: Goodness of fit for the neural network model. The first, second and third columns report respectively the number of input/output neurons, of latent factors and of fitted coefficients. The fourth and fifth columns present the sum of squared errors and the average errors. The two last columns contain the maximum and minimum errors.

The figure 2 shows filtered latent factors by tested neural networks. For most of configurations, the latent processes $\kappa_t^{nn,i}$ exhibit a quasi-linear trend, either increasing or decreasing. As for the LC model, we assume that increments of latent factors follow a random walk with drift for the prediction. This hypothesis is checked with a Jarque Bera test for the 3-2-3 neural net, over the period 1970-2000. Statistics of this test, reported in table 6, confirm the reliability of this assumption. As for LC models, the same test applied to the sample of increments over the whole period of calibration (1946-2000) rejects the normality for the first latent factor. If we look to the evolution of this process (first graph of figure 2), we observe a change of trend between periods 1948-1960 and 1960-2000. As mentioned previously, This change of trend may be partly explained by the reduction of mortality caused by coronary heart diseases, following two prevention campaigns launched around the sixties.

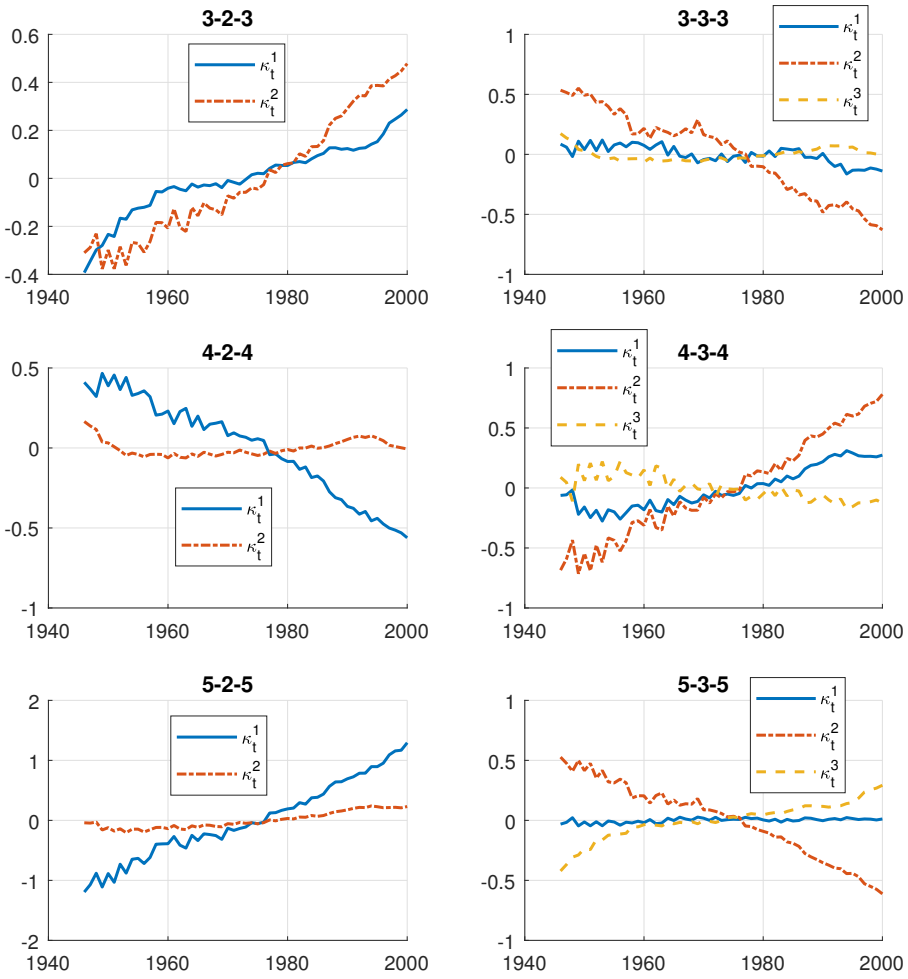

Figure 2: Latent processes, $\kappa_t^{nn,i}$, filtered with different configurations of neural networks. The title of each graph reports respectively the number of neurons in the input / intermediate / output layers.

	Jarque Bera statistics for a 3-2-3 Neural Net			
factors	Normality	p-value	JB statistic	Critical Value 5%
$\kappa_t^{nn,1} - \kappa_{t-1}^{nn,1}$	Accept	0.5000	0.0762	4.4496
$\kappa_t^{nn,2} - \kappa_{t-1}^{nn,2}$	Accept	0.3698	1.2296	4.4496

Table 6: Jarque Bera test applied to increments of latent factors over the period 1970-2000, for neural analyzer with three neurons in input / output layers and two neurons in the intermediate layer.

French population, 2001-2014, forecast.						
n_l	n_d	Coef.	$\sum \left\|X - \hat{X}\right\|_2^2$	Avg.$\left\|X - \hat{X}\right\|_2$	$\max\left(X - \hat{X}\right)$	$\min\left(X - \hat{X}\right)$
3	**2**	**552**	**8.17**	**0.0023**	**0.4922**	**-0.1474**
4	2	736	9.60	0.0025	0.4607	-0.1587
5	2	920	10.84	0.0026	0.4567	-0.1788
6	2	1104	14.20	0.0030	0.4549	-0.1705
7	2	1288	16.51	0.0032	0.4498	-0.1647
8	2	1472	16.62	0.0032	0.4408	-0.1754
3	3	558	9.09	0.0024	0.5021	-0.1470
4	3	744	9.71	0.0025	0.4701	-0.1546
5	3	930	10.57	0.0026	0.4603	-0.1628
6	3	1116	10.81	0.0026	0.4333	-0.1875
7	3	1302	14.24	0.0031	0.4084	-0.1902
8	**3**	**1488**	**17.43**	**0.0033**	**0.5053**	**-0.1914**

Table 7: Predictive goodness of fit for the neural network model. The first and second columns report the number of latent factors and fitted coefficients. The third and fourth columns present the sum of squared errors and the average errors. The two last columns contain the maximum and minimum errors.

To validate the predictive capability of the neural model, we forecast log-forces of mortality over fourteen years and compare them to the real rates observed over the period 2001-2014. 10 000 simulations are performed and we consider as forecast, the yearly average of simulated log-mortality rates. The table 7 presents the errors of estimation. A comparison with errors of LC models confirms the excellent predictive power of the neural network: the sum of squared errors falls to 8.17, for the 3-2-3 configuration whereas the predictive error of the LC model with cohort effects has a predictive error of 17.65. The figure 3 compares predicted and real log-forces of mortality for years 2001 and 2014, with this configuration of neurons. A deeper analysis of figures in table

7 reveals that increasing the number of neurons deteriorates the predictive power of networks. In particular, the 8-3-8 neural net yields the highest prediction error, despite having the lowest calibration error. This phenomenon is related to the mechanism of overfitting. Overfitting occurs when the model is excessively complex, such as having too many parameters relative to the number of observations. An overfitted model has poor predictive performance and it overreacts to minor fluctuations in the training data. Overfitting may easily be avoided by choosing the neural network architecture that offers the best trade-off between calibration and prediction errors. In our case, the predictive power of the 3-2-3 configuration (3 input/output and 2 intermediate neurons) being excellent and its calibration error being close to the one of the LC COH model, the remainder of this section focuses on this network.

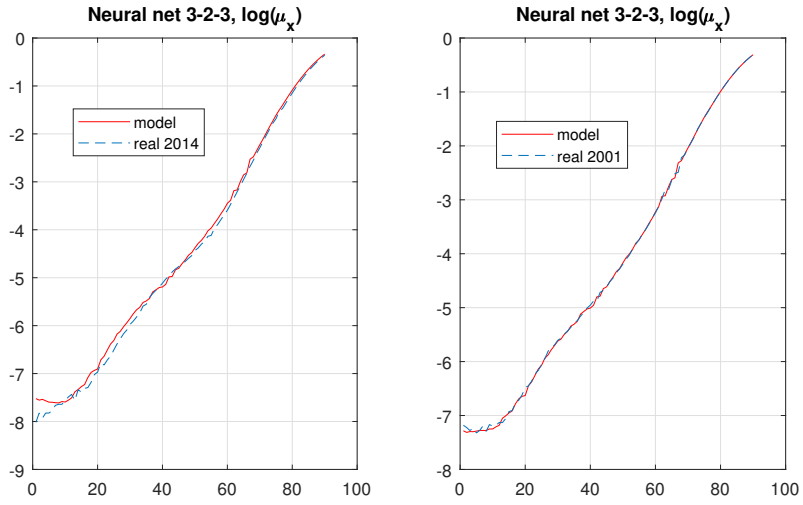

Figure 3: The left and right plots compare the real log-mortality rates in 2014/2001 to the average log-forces of mortality simulated with the 3-2-3 neural analyzer. 10 000 simulations are realized.

The left and right plots of figure 4 respectively present a breakdown of means and standard deviations of relative average errors of calibration, per age, and for the period 2001-2014. The 3-2-3 network is here compared to LC models with (LC GLM) and without cohort effects (LC COH). The fact that standard deviations of these errors for all models continuously increase with age, is inherent to the hypothesis of linearity of log-forces of mortality with respect to latent factors. Excepted for ages above 80 years, the deviation of relative errors for the neural net is nearly constant and may then be attributed to measurement errors, which is a desirable quality for a model. Before 50 years, the average of relative errors for the 3-2-3 net is close to zero For the age group 50 to 80, average relative errors and their deviations computed with the neural net are lower in absolute value than these obtained by other approaches.

If we look to the left plot of figure 4, we observe a clear cut in the evolution of rel-

ative average errors at the age of 50 years. This cut comes from the configuration of the neural analyzer: we have three input / output neurons affected respectively and exclusively to three age groups. The first input neuron receives only information about the mortality between the ages of 20 and 50 years and this deteriates the goodness of fit around the age of 50. It is probably possible to improve the calibration by sharing some information between adjacent neurons.

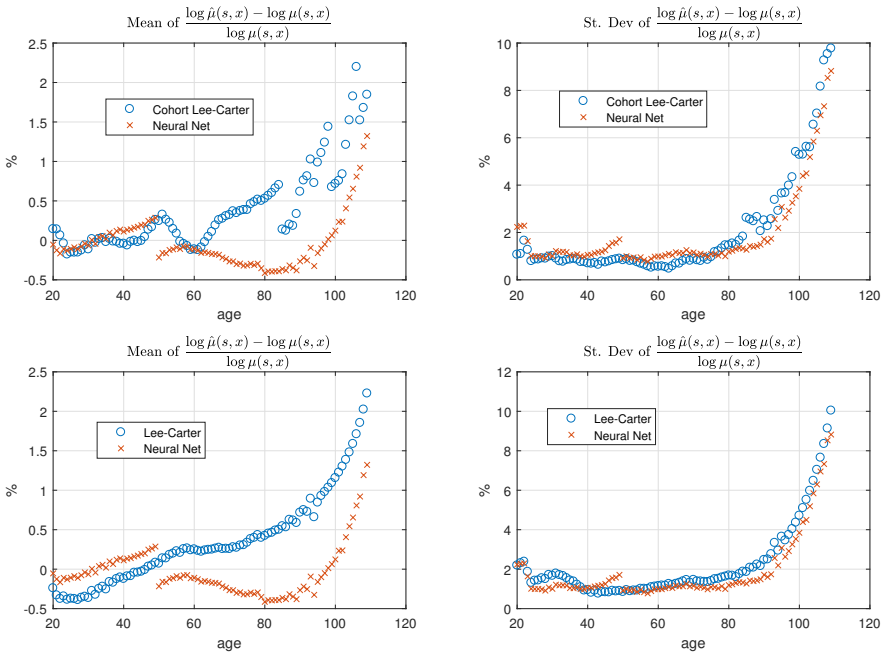

Figure 4: Breakdown of means and standard deviations of relative average errors of calibration, per age. We compare the LC GLM and LC COH models to the 3-2-3 Neural network.

Figure 5 illustrates the influence of latent factors filtered by a 3-2-3 neural net on the term structure of log-forces of mortality (forecast, year 2014). The left plot emphasizes that $\kappa_t^{nn,1}$ mainly influences log-mortality rates between 20 and 50 years old. Increasing $\kappa_t^{nn,1}$ reduces log-mortality rates for this age group and slightly increases log-forces of mortality for ages 50 and above. The right graph shows that the second latent factor $\kappa_t^{nn,2}$ mainly concerns individuals aged between 51 and 109 years. Increasing $\kappa_t^{nn,2}$ reduces log-mortality rates for this age range and slightly increases rates for ages before 50. Compared to LC models, latent factors yield by the neural net offer then the same ease of interpretation.

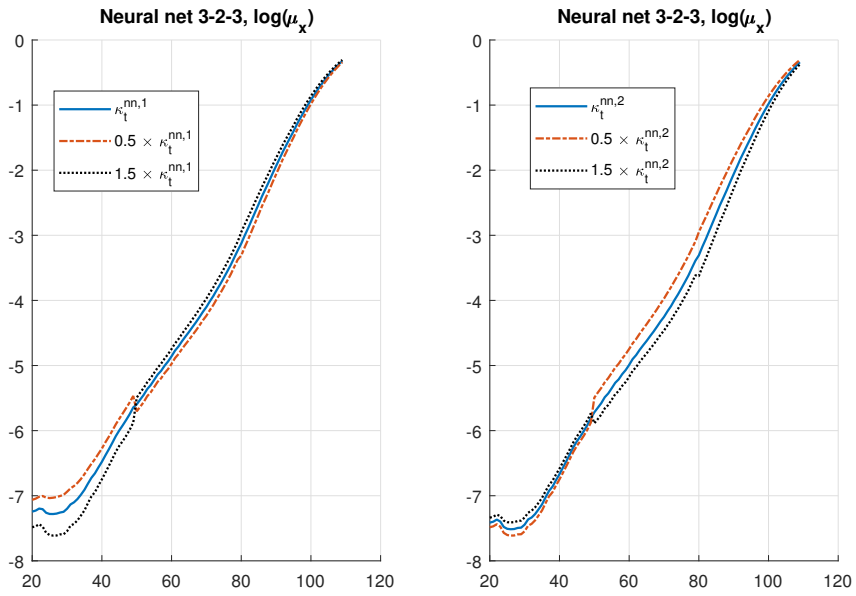

Figure 5: 3-2-3 Neural network: sensitivity of log-mortality rates to variations of latent factors

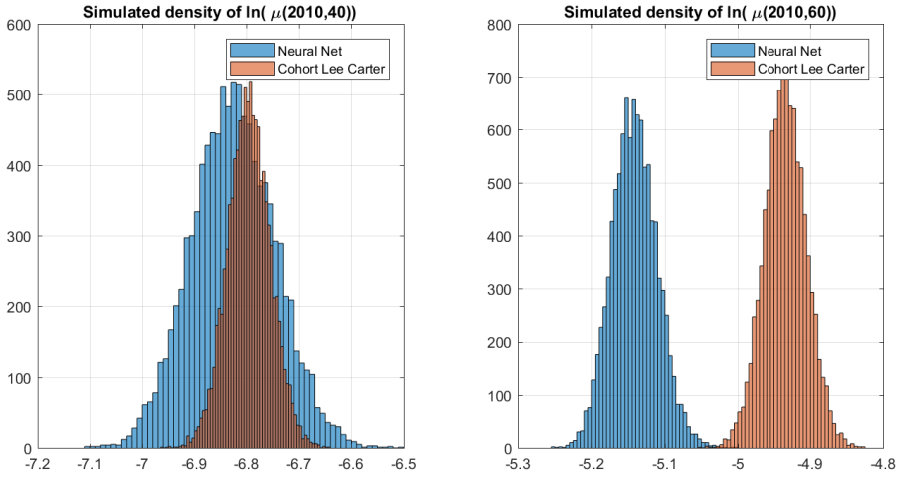

Figure 6: Comparison of simulated densities for $\ln \mu(2010, 40)$, yield by the LC with cohort effect and the 3-2-3 Neural model.

3-2-3 Neural analyzer				
	20 y.	40 y.	60 y.	80 y.
$\mathbb{E}(\ln \mu(t,x))$	-7.4532	-6.8255	-5.1414	-3.3868
$std(\ln \mu(t,x))$	0.0690	0.07972	0.031125	0.037625
$\mathbb{S}(\ln \mu(t,x))$	0.0714	0.07142	0.16981	0.16981
$\mathbb{K}(\ln \mu(t,x))$	2.9881	2.9881	3.1085	3.1088
Cohort LC model (LC COH)				
	20 y.	40 y.	60 y.	80 y.
$\mathbb{E}(\ln \mu(t,x))$	-7.3365	-6.7959	-4.9346	-3.4342
$std(\ln \mu(t,x))$	0.0853	0.04085	0.0291	0.0623
$\mathbb{S}(\ln \mu(t,x))$	-0.0416	-0.0024	-0.0023	-0.0023
$\mathbb{K}(\ln \mu(t,x))$	2.9772	2.9660	2.9282	2.9791
Historical log-mortality rates (1946-2010)				
	20 y.	40 y.	60 y.	80 y.
$\mathbb{E}(\ln \mu(t,x))$	-6.5021	-5.8614	-4.3754	-2.604
$std(\ln \mu(t,x))$	0.7993	0.6400	0.3858	0.3814
$\mathbb{S}(\ln \mu(t,x))$	1.2508	1.1690	0.4596	0.0881
$\mathbb{K}(\ln \mu(t,x))$	3.1277	2.9206	2.3744	2.0588

Table 8: This table compares moments of log-forces of mortality simulated by the Neural analyzer (3 input/output neurons and 2 intermediate neurons) and the cohort LC model. 10 000 simulations are performed and rates are computed age 20, 40, 60 and 80, for the year 2010. The third sub-table reports the moments of observed log-forces of mortality over the period 1946-2010.

The table 8 compares the moments of simulated log-forces of mortality, predicted by the cohort LC model and the 3-2-3 neural analyzer. The forecasts are computed for the year 2010, with models fitted to data from 1946-2000. The figure 6 shows the densities of $\ln \mu(2010, 40)$ obtained by simulations. These statistics and the graph emphasize that distributions of simulated log-forces of mortality display visible differences depending on the model. With the neural analyzer, the distribution exhibits a higher variance than for the LC model and right asymmetry. Whereas log-mortality rates predicted by the neural net are slightly leptokurtic, they are strictly Gaussian in the cohort LC model[3]. It is interesting to compare simulated moments to these calculated with past mortality rates, over the period 1946-2010. We observe that the empirical historical distribution also displays a right asymmetry that is not present in the LC model. The historical variance is also much higher than the one predicted by the LC and neural models. The distribution is leptokurtic at 20 years old, and the kurtosis

[3]Notice that in the LC model, the log-mortality rates are normally distributed: their skewness and kurtosis are then respectively equal to 0 and 3. Skewness and kurtosis reported in table 8 are not exactly equal to these figures because they are computed with simulated log-forces of mortality.

decreases next with age. However, these statements must be nuanced given the limited number of observations available to calculate these statistics.

We pursue our analysis of LC and neural networks by a comparison of cross-sectional lifetime expectancies predicted by models, over the period 2001-2014. The lifetime expectancy for a x years old individual on year t, is defined as follows

$$e_x(t) \quad := \quad \sum_{s=1}^{x_{max}} {}_sp_x(t)\,,$$

where ${}_sp_x(t)$ is the survival probability from age x to age $x+s$, calculated with cross-sectional mortality rates:

$$\begin{aligned} {}_sp_x(t) &= \exp\left(-\int_0^s \mu(t, x+u)\, du\right) \\ &\approx \exp\left(-\sum_{k=0}^{s-1} \mu(t, x+k)\right) \end{aligned}$$

The table 9 presents information about cross-sectional lifetime expectancies at 20, 40, 60, 80 years old obtained with the cohort LC model (LC COH). 10 000 simulations are performed and lifetime expectancies are computed scenario per scenario. Averages of predicted expectancies are reported in the first third of the table. These figures forecast that the maximum improvement of longevity concerns the 20 years individuals who gain 2.2 years of lifetime expectancy between 2001 and 2014. This improvement is slightly lower than the real one observed over this period (2.99 years). The LC model underestimates the improvement of longevity by 0.20 years for an 80 years old person to 0.76 years for a 20 years old individual, in 2014.

	$e_{20}(t)$	$e_{40}(t)$	$e_{60}(t)$	$e_{80}(t)$
2001	60.394	41.251	23.665	8.998
2005	61.111	41.899	24.312	9.382
2010	61.943	42.650	25.071	9.857
2014	62.592	43.241	25.637	10.236
	$e_{20}^{Obs}(t) - e_{20}(t)$	$e_{40}^{Obs}(t) - e_{40}(t)$	$e_{60}^{Obs}(t) - e_{60}(t)$	$e_{80}^{Obs}(t) - e_{80}(t)$
2001	-0.0115	0.0137	0.0143	0.0076
2005	0.2509	0.2087	0.0804	0.0026
2010	0.4723	0.4355	0.1648	0.1001
2014	0.7631	0.6995	0.2337	0.1959

Table 9: Cohort Lee-Carter model (LC COH): average cross-sectional lifetime expectancies and their spread with real expectancies.

The table 10 presents information about cross-sectional lifetime expectancies at 20, 40, 60, 80 years old computed with the 3-2-3 neural analyzer. As for the LC model, the maximum improvement of longevity concerns the 20 years old generation who gains on average 2.19 years of lifetime expectancy between 2001 and 2014. In a similar way to the LC-COH model, the neural net underestimates the real improvement of longevity observed over this period.

	$e_{20}^{NN}(t)$	$e_{40}^{NN}(t)$	$e_{60}^{NN}(t)$	$e_{80}^{NN}(t)$
2001	60.392	41.247	23.635	8.9842
2005	61.125	41.913	24.19	9.3227
2010	61.971	42.685	24.839	9.7315
2014	62.588	43.25	25.316	10.043
	$e_{20}^{Obs}(t) - e_{20}^{NN}(t)$	$e_{40}^{Obs}(t) - e_{40}^{NN}(t)$	$e_{60}^{Obs}(t) - e_{60}^{NN}(t)$	$e_{80}^{Obs}(t) - e_{80}^{NN}(t)$
2001	-0.0092	0.0174	0.0446	0.0217
2005	0.2367	0.1947	0.2018	0.0628
2010	0.4444	0.4012	0.3975	0.2264
2014	0.7676	0.6909	0.5548	0.3891

Table 10: 3-2-3 Neural net: average cross-sectional lifetime expectancies and their spread with real expectancies.

The neural analyzer predicts realistic log-mortality rates over a short period of time, following the last year of calibration. However, does it remains reliable for long term forecasting? To answer this question, we calculate the cross-sectional lifetime expectancies of a 20, 40, 60 and 80 years old individual, from 2001 to 2100. The evolution of expectancies at 20 and 60 years old are shown in figure 7. The lifetime expectancies, computed with a 3D LC model fitted by SVD grow respectively linearly from 60 to 65 years and from 23 to 27 years. The same expectancies forecasted by the LC GLM model respectively increase from 60 to 74 and from 23 to 35 years. Whereas the LC model with cohort effects predicts a rise from 60 up to 74 and 23 up to 35 years. Life expectancies computed with the neural net display a concave growth. They dominate these yield by the LC model but are below the forecasts of LC GLM and LC COH models, excepted over the period 2000-2020. Table 11 compares life expectancies predicted by LC COH and Neural net models, for different ages and years. According to the neural analyzer, the average lifetime will respectively increase of 8 and 5 years for a 20 and 80 years old individual, over the next century. Whereas the LC COH forecasts an increase of 12 and 8 years for persons aged 20 and 80 years. We cannot say which model is the most reliable for long term forecast of log-forces of mortality. However, the neural net approach predicts realistic projections.

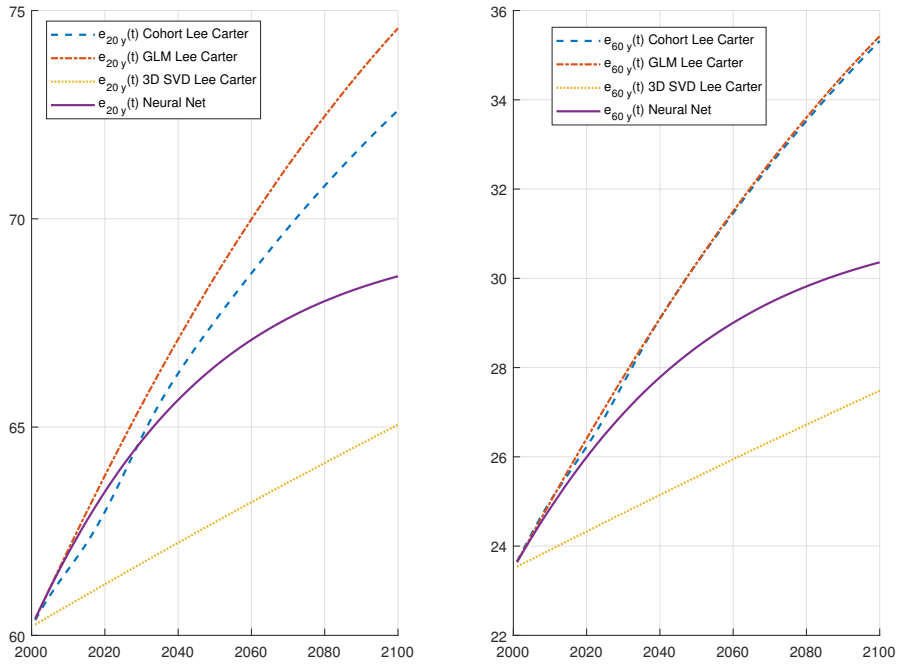

Figure 7: Predicted cross-sectional lifetime expectancies with LC and neural models, over the period 2001-2100.

	$e_{20}^{NN}(t)$	$e_{40}^{NN}(t)$	$e_{60}^{NN}(t)$	$e_{80}^{NN}(t)$
2001	60.400	41.255	23.642	8.9887
2025	64.088	44.636	26.505	10.863
2050	66.453	46.851	28.454	12.369
2100	68.621	48.919	30.358	14.098
	$e_{20}^{LC\,COH}(t)$	$e_{40}^{LC\,COH}(t)$	$e_{60}^{LC\,COH}(t)$	$e_{80}^{LC\,COH}(t)$
2001	60.367	41.223	23.656	8.9891
2025	63.842	44.485	26.899	11.123
2050	67.542	48.068	30.33	13.232
2100	72.603	52.942	35.318	16.895

Table 11: Long term predictions of cross-sectional life expectancies, computed by simulations with the cohort LC and Neural models, over the period 2001-2100.

6 Comparison with the UK and US populations

In this section, we check that the neural net approach efficiently explain the US and UK mortality. Table 12 reports calibration errors for LC models with and without a cohort effect and for neural nets, fitted to UK log-forces of mortality, from 1946 up to 2000. As for the French population, the LC model with a cohort effect yields a low calibration error. However, neural nets achieve similar or better performances, depending upon the configuration. We also observe that increasing the number of neurons systematically reduces the calibration error. According to figures of table 13, the predictive power of the LC COH over the period 2001-2013 is lower than the one of the LC model, fitted by log-likelihood maximization. For the UK population, the 3-2-3 network displays an excellent predictive capability compared to other neural configurations and to competing models. Table 14 and 15 reports the calibration and prediction errors for models adjusted to US mortality rates. The lowest calibration errors are obtained with neural nets and three intermediate neurons. Despite that the LC COH model has an excellent explanatory power for the period 1946-2000, its predictive capability is clearly lower than these of neural nets, whatever their configuration. Contrary to French and UK cases, we don't observe any deterioration of predictive errors when we increase the number of neurons. We conclude form this analysis that the efficiency of the neural net analyzer does not depend upon the reference dataset.

UK population, 1946-2000				
Model	$\sum \left\|X - \hat{X}\right\|_2^2$	Avg.$\left\|X - \hat{X}\right\|_2$	$\max\left(X - \hat{X}\right)$	$\min\left(X - \hat{X}\right)$
LC SVD 1	131.4	0.0023	0.9390	-0.4606
LC SVD 2	114.08	0.0021	0.7845	-0.3871
LC SVD 3	111.18	0.0021	0.7509	-0.3748
LC GLM	32.85	0.0011	0.3484	-0.4444
LC COH	8.946	0.0005	0.2478	-0.1962
NN 3-2-3	15.16	0.0008	0.2104	-0.2927
NN 4-2-4	11.01	0.0007	0.2119	-0.2789
NN 5-2-5	10.09	0.0006	0.2076	-0.2655
NN 4-3-4	9.04	0.0006	0.1934	-0.2751
NN 5-3-5	7.74	0.0006	0.2153	-0.2673
NN 6-3-6	7.79	0.0006	0.2329	-0.2577

Table 12: Goodness of fit for extension of the LC model and neural nets. The second and third columns present the sum of squared errors and the average errors. The two last columns contain the maximum and minimum errors.

UK population, 2001-2013				
Model	$\sum \left\|X - \hat{X}\right\|_2^2$	Avg.$\left\|X - \hat{X}\right\|_2$	$\max\left(X - \hat{X}\right)$	$\min\left(X - \hat{X}\right)$
LC SVD 1	27.43	0.0044	0.5337	-0.0974
LC SVD 2	26.90	0.0044	0.5865	-0.1022
LC SVD 3	26.68	0.0044	0.5648	-0.1035
LC GLM	11.68	0.0027	0.5532	-0.1723
LC COH	13.38	0.0026	0.3631	-0.2311
NN 3-2-3	12.83	0.0028	0.5179	-0.1410
NN 4-2-4	13.13	0.0029	0.5228	-0.1750
NN 5-2-5	14.18	0.0030	0.5409	-0.1382
NN 4-3-4	13.54	0.0029	0.5215	-0.1300
NN 5-3-5	14.38	0.0030	0.5310	-0.1043
NN 6-3-6	14.55	0.0030	0.5517	-0.1358

Table 13: Predictive goodness of fit for LC models and neural networks. The second and third columns present the sum of squared errors and the average errors. The two last columns contain the maximum and minimum errors.

US population, 1946-2000				
Model	$\sum \left\|X - \hat{X}\right\|_2^2$	Avg.$\left\|X - \hat{X}\right\|_2$	$\max\left(X - \hat{X}\right)$	$\min\left(X - \hat{X}\right)$
LC SVD 1	81.69	0.0018	0.4293	-0.3642
LC SVD 2	76.02	0.0017	0.3108	-0.3756
LC SVD 3	73.18	0.0017	0.3222	-0.3560
LC GLM	12.26	0.0007	0.1863	-0.2456
LC COH	6.23	0.0004	0.1403	-0.1966
NN 3-2-3	8.19	0.0006	0.1441	-0.1801
NN 4-2-4	6.55	0.0005	0.1363	-0.1675
NN 5-2-5	6.40	0.0005	0.1852	-0.1980
NN 4-3-4	5.61	0.0005	0.1452	-0.1574
NN 5-3-5	5.46	0.0005	0.1713	-0.1831
NN 6-3-6	4.75	0.0004	0.1639	-0.1796

Table 14: Goodness of fit for extensions of the LC model and neural nets. The first and second columns report the number of latent factors and fitted coefficients. The third and fourth columns present the sum of squared errors and the average errors. The two last columns contain the maximum and minimum errors.

US population, 2000-2015				
Model	$\sum \left\|X - \hat{X}\right\|_2^2$	Avg. $\left\|X - \hat{X}\right\|_2$	$\max\left(X - \hat{X}\right)$	$\min\left(X - \hat{X}\right)$
LC SVD 1	15.61	0.0029	0.2998	-0.2311
LC SVD 2	19.31	0.0032	0.3082	-0.3035
LC SVD 3	21.26	0.0034	0.2945	-0.2866
LC GLM	10.73	0.0024	0.2020	-0.3567
LC COH	25.73	0.0032	0.1931	-0.6505
NN 3-2-3	11.86	0.0027	0.2152	-0.3409
NN 4-2-4	13.58	0.0029	0.2223	-0.3700
NN 5-2-5	14.75	0.0030	0.2170	-0.3615
NN 4-3-4	15.28	0.0031	0.3003	-0.3520
NN 5-3-5	16.62	0.0032	0.3223	-0.3786
NN 6-3-6	18.87	0.0034	0.2911	-0.4110

Table 15: Predictive goodness of fit for LC models and neural networks. The first and second columns report the number of latent factors and fitted coefficients. The third and fourth columns present the sum of squared errors and the average errors. The two last columns contain the maximum and minimum errors.

7 Conclusions

This study proposes a new method based on a neural network to predict and simulate the future human mortality. Contrary to previous attempts in the literature, the neural network is not used as a substitute to an econometric model. Instead, it summarizes the information carried by the surface of log-forces of mortality in a limited number of latent factors. These factors are next extrapolated and future term structures of mortality rates are obtained by an inverse transform. Given the important number of parameters, a genetic algorithm combined to a gradient descent, is used to calibrate the network.

Numerical tests performed on the French, UK and US log-forces of mortality, emphasizes that the neural analyzer outperforms LC model and its multi-factor extensions, fitted by SVD or log-likelihood maximization. The neural net approach has an explanatory power that is comparable or even better the LC model with age specific cohort effects. On the other hand, the latent factors filtered by the neural network exhibit a clear linear trend and Jarque Bera tests confirm that it is not absurd to forecast them with a simple random walk.

A comparison of average of predicted mortality rates with observed rates over the period 2001-2014 underlines the excellent predictive power of the neural approach, compared

to competing models. However, the number of neurons must be chosen carefully to avoid over-parameterization. In particular for French and UK population, the predictive power of the neural net worsens if there are too many neurons.

The average relative errors for neural nets and their standard deviations are lower than these obtained with a LC model with cohort effects. Excepted for ages above 80 years, the deviation of relative errors is nearly constant, which is a desirable quality for a model. The probability density function of future log-forces of mortality forecast by the neural net differs from the one obtained with the cohort LC model. More precisely, the density displays a higher variance and leptokurticity and a small right asymmetry. Finally, the neural analyzer predicts believable long term life expectancies.

Neural network models are promising for applications in actuarial sciences and there are many tracks for further research. In particular, it would be interesting to extend our approach to explain the joint evolution of mortality of several populations. Another way to explore is the optimality of the structure of the neural net. It is probably possible to enhance the predictive power by modifying the architecture of the neural network.

References

[1] Abdulkarim SA, Garko AB, 2015 Forecasting maternal mortality rate using particle Swarm optimization based artificial neural network. Dutse Journal of Pure and Applied Sciences 1(1): 55 - 59.

[2] Antonio K., Bardoutsos A., Ouburg W. 2015. Bayesian Poisson log-bilinear models for mortality projections with multiple populations. European Actuarial Journal 5:245–281

[3] Atsalakis G, Nezis D, Matalliotakis G, Ucenic CI, Skiadas C, 2007. Forecasting Mortality Rate Using a Neural Network with Fuzzy Inference System. Working paper 0806. University of Crete.

[4] Brouhns N, Denuit M, Vermunt JK 2002. A Poisson log-bilinear regression approach to the construction of projected lifetables. Insurance: Mathematics and Economics 31(3):373-393.

[5] Cairns AJC, 2008. Modelling and management of mortality risk: a review. Scandinavian Actuarial Journal 2-3, p79-113.

[6] Currie I. D., 2016. On fitting generalized linear and non-linear models of mortality. Scandinavian Actuarial Journal (4), 356–383.

[7] Cox DR. 1972. Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B. 34:187–220.

[8] Cybenko G. 1989. Approximation by Superpositions of a Sigmoidal Function. Math. Control Signals Systems 2:303-314.

[9] Dimitrova D.S., Haberman S., Kaishev V.K. 2013. Dependent competing risks: Cause elimination and its impact on survival. Insurance: Mathematics and Economics 53= 464-477.

[10] Dong D, McAvoy TJ, 1996. Nonlinear principal component analysis—Based on principal curves and neural networks. Comp. Chem. Eng., 20, 65–78.

[11] Fung M.C., Peters G., Shevchenko P. 2015. A State-Space Estimation of the Lee-Carter Mortality Model and Implications for Annuity Pricing. Available at SSRN: https://ssrn.com/abstract=2699624

[12] Fung M.C., Peters G., Shevchenko P, 2016. A unified approach to mortality modelling using state-space framework: characterisation, identification, estimation and forecasting. Available at SSRN: https://ssrn.com/abstract=2786559

[13] Fung M.C., Peters G., Shevchenko P. 2017. Cohort Effects in Mortality Modelling: A Bayesian State-Space Approach. Available at SSRN: https://ssrn.com/abstract=2907868

[14] Fotheringhame D, Baddeley R, 1997. Nonlinear principal components analysis of neuronal spike train data. Biol. Cybernetics, 77, 282–288.

[15] Hainaut D, 2012. Multidimensional Lee–Carter model with switching mortality processes. Insurance: Mathematics and Economics 5(2), 236-246.

[16] O'Hare C., Li Y., 2012. Explaining young mortality. Insurance, Mathematics and Economics 50, 12–25.

[17] Hornik K, 1991. Approximation Capabilities of Multilayer Feedforward Networks. Neural Networks, 4(2), 251–257.

[18] Khachaturyan A, Semenovskaya S, Vainshtein B, 1979. "Statistical-Thermodynamic Approach to Determination of Structure Amplitude Phases". Sov.Phys. Crystallography. 24 (5): 519–524.

[19] Kramer MA, 1991. Nonlinear principal component analysis using autoassociative neural networks. AIChE J., 37, 233–243.

[20] Lee RD , Carter L, 1992. Modelling and forecasting the time series of US mortality. Journal of the American Statistical Association, 87 659-671.

[21] Lee RD, 2000. The Lee-Carter Method for Forecasting Mortality, with Various Extensions and Applications. North American Actuarial Journal 4(1), 80-91.

[22] Malthouse EC, 1998. Limitations of nonlinear PCA as performed with generic neural networks. IEEE Trans. Neural Networks, 9, 165–173.

[23] McNelis PD, 2005. Neural networks in finance: gaining predictive edge in the market. Elsevier academic press.

[24] Monahan HA, 2000. Nonlinear Principal Component Analysis by Neural Networks: Theory and Application to the Lorenz System. Journal of Climate 13, 821-835.

[25] Pitacco E, 2004. Survival models in a dynamic context: a survey. Insurance: Mathematics and Economics, 35 p279-298.

[26] Pitacco E, Denuit M, Haberman S, Olivieri A, 2009. Modeling longevity dynamics for pensions and annuity business. Oxford University Press, London

[27] Puddu PE, Menotti A, 2009. Artificial neural network versus multiple logistic function to predict 25-year coronary heart disease mortality in the Seven Countries. European Journal of Preventive Cardiology.16(5):583-591.

[28] Puddu P.E., Menotti A. 2012. Artificial neural networks versus proportional hazards Cox models to predict 45-year all-cause mortality in the Italian Rural Areas of the Seven Countries Study. BMC Medical Research Methodology, 12:100

[29] Puddu P.E., Piras P., Menotti A. 2017. Lifetime competing risks between coronary heart disease mortality and other causes of death during 50 years of follow-up. International Journal of Cardiology. 228: 359-363.

[30] Renshaw AE, Haberman S, 2003. Lee-Carter mortality forecasting with age-specific enhancement. Insurance: Mathematics and Economics, 33 p255-272

[31] Renshaw A, Haberman S, 2006. A cohort-based extension to the Lee–Carter model for mortality reduction factors. Insurance: Mathematics and Economics 38:556–570.

[32] Toczydlowska D. Peters G., Fung M.C., Shevchenko P.V. 2017. Stochastic Period and Cohort Effect State-Space Mortality Models Incorporating Demographic Factors via Probabilistic Robust Principle Components Available at SSRN: https://ssrn.com/abstract=2977306

[33] Van Berkum F., Antonio K., Vellekoop M., 2016. The impact of multiple structural changes on mortality predictions. Scandinavian Actuarial Journal 2016(7), 581–603.

[34] Wilmoth J.R. 1993. Computational methods for fitting and extrapolating the Lee-Carter model of mortality change. Technical Report; Department of Demography, University of California, Berkeley.

[35] Wong-Fupuy C, Haberman, 2004. Projecting mortality trends: recent developments in the UK and the US. North American Actuarial Journal, 8, p56-83.

[36] Yang, S. S., Yue, J. C., Huang, H., 2010. Modeling longevity risks using a principal component approach: A comparison with existing stochastic mortality models. Insurance: Mathematics and Economics 46(1), 254–270.

Acknowledgment.

I gratefully acknowledges the BNP Cardiff Chair "Data Analytics and Models for Insurance" for its financial support. I also thank Michel Denuit from the UCL and the two anonymous referees for their constructive advices.