Institut de Statistique, Biostatistique et Actuariat

Dimension-Reduction with Feed-Forward Neural Network Applied to Mortality.

Membres du jury:

Prof. Hainaut Donatien *Promoteur*
Prof. Devolder Pierre

Mémoire présenté en vue de
l'obtention du mastère
en sciences actuarielles
(orientation sciences actuarielles)
par:

Haddi Naïm

Louvain-La-Neuve

Juin 2021

Acknowledgment.

Thanks to my family.

Contents

1 Introduction **1**

2 Aims and objectives **2**

3 Data Pre-Processing **3**
3.1 Smoothing: Whittaker-Henderson 4

4 Modelling **7**
4.1 Actuarial Model: Poisson Model 7
4.1.1 Probability of Dying, Mortality Rate and Lifetime Expectancy . . . 7
4.1.2 Prospective Tables & Exposure to Risk 10
4.1.3 Lee Carter to Poisson Model 11
4.1.4 Poisson Model 12
4.2 Towards Dimension-Reduction 14
4.3 Machine Learning Model: Feed-Forward Neural Network 15
4.3.1 Loss Function: Deviance 17
4.4 Training, validation and forecasting 19
4.4.1 Model validation 19
4.4.2 Forecasting 20

5 Results **22**
5.1 The Swedish population case 22
5.1.1 Results of validation 22
5.1.2 Goodness of fit & Predictive Power 24
5.1.3 Analysis of trends 26
5.1.4 Comparison with market benchmarks: LC model and its cohort variant 29
5.2 Application to several countries 37
5.2.1 Results of validation 37
5.2.2 Goodness of Fit & Predictive Power 38
5.2.3 Analysis of trends 40
5.2.4 Comparison of lifetime expectancies 41
5.2.5 Insurance Product: Life annuity 43

6 Perspectives & Conclusion **47**

7 Appendix **50**

1 Introduction

Since private property exists, people are afraid to loss what they own. This phenomenon is called the risk aversion. As a mean to avoid this kind of feeling, insurance industry and actuarial sciences has emerged. It is in the context of shipping transactions in Grece that first form of insurance was created. The vessels full of foods and other commodities were shuttling between Europe and Africa, bringing to their recipient what they did not have in their country. At that time, wealthy businessmen suggested to invest in theses shipping by repaying the cargo if the delivery went wrong against a high interest rate if the boat arrives at its destination. With the essor of world trading, theses individuals were not willing to assume this risk anymore. It is in the 14th century that professionals risk takers decided to take the lead to establish insurance contracts [Smith, 2021]. At the beginning, insurance were mainly based on statics and simple models. Nowadays, a tremendous amount of financial data are generated all over the world due to the globalization of the economic world and the rapid spread of information. With these changes, risk became increasingly complex leading the complexification of the models used as well. Now that computers replace humans for calculation, all models are implemented in softwares based on excel or SAS in the industry. With the efficience of machine learning in terms of forecasting, certain undertakings began to insert it in their models in order to help them in their decision making. However, a lot of small insurance companies are still using classical softwares given the cost of changing to other models. Sometimes, enterprises have recourse to outsourcing for their computations with consulting firms. For instance, Addactis is a French consulting company that delivers advanced technical solution using machine learning. Even though machine learning registers better results than classical models, insurance companies are a bit afraid of the blackbox related to theses modelling techniques. Indeed, humans can not know what exactly the computer is doing when the learning operates. As the insurance industry knows well their classical models, one can adapt its figures each year based on its own statistical results. If a problem of machine learning happens in a company entirely based on this kind of model, the loss may be more consequent and difficult to adjust. The preference of the classical model instead of machine learning one may be in order to smooth the yearly result of the company. From an academic point of view, machine learning became a real matter of interest in actuarial sciences. An increasing number of researchs, for example among others the papers "Machine Learning in PC Insurance: A Review for Pricing and Reserving" and "An Individual Claims Reserving Model", are done on this promising subject. In 2019, the emergence of new courses in the master of actuarial sciences at UCLouvain as, for instance "Data sciences for finance and insurance" and "Modélisation prédictive et apprentissage statistique en assurance", has turned the master towards machine learning applications.

2 Aims and objectives

In this master thesis, a machine learning technique is implemented in order to predict mortality with an actuarial model. The main purpose is to implement a feed-forward neural network to perform a dimension-reduction on log-forces of mortality. Before explaining concepts used to do so, it is important to discuss about the framework of this master thesis and literature reviewed before it, from a machine learning and an actuarial percepective.

At the inception, this deep-learning method was created to model the processing of information by biological neural networks located in the mammalian cortex. In 1957, the first artificial perceptron is built by Rosenblatt. A repesentation of a single neuron perceptron is available in the "Machine Learning Model" section with Figure 2. Since that time, a lot of variant and extension has been made and this applied to several fields and industries. For instance, convolution neural network, currently used for imagery recognition, has been introduced in the 60's to differenciate simple from complex cells in the cortexes of cats and monkeys. In the 90's, long short term memory emerge as artificial neural net that can make feedback connections. This kind of technique is used for handwritting or speech recognition. In 1991, Kramer suggested to use a feed-forward neural net in order to perfom a non linear principal component analysis in dimension reduction context for chemistry purposes [Kramer, 1991]. The aim is to detect non linear connections between variables as generalization of the principal component analysis, a method used for dimensionalty reduction. This paper has inspired Professor D. Hainaut in 2018 to propose a feed-forward neural network that recognize non-linearities in the lower-dimensional structure of the log-forces of mortality [Hainaut, 2018]. The calibration of its model was done with a genetic algorithm and mean square error as a loss function on the french population. In this master thesis, the aim is to suggest a different loss function and algorithm based on mortality assumptions. An extension to several countries is suggested as well to challenge the model. This may help to understand how the model reacts in an other context. Finally, a life insurance product is calculated to show the financial impact of mortality given the related country.

The outline of the remainder of this master thesis is as follows. Firstly, a Data Pre-Processing section to clarify how the outliers problem is overpassed. Secondly, a modelling section that explains actuarial and machine learning concepts used to build the model. Thirdly, a results section to discuss about the output of model. Finally, a perspectives and conclusion section to conclude the present work.

3 Data Pre-Processing

This section is devoted to data pre-processing. Before discussing about the management of outliers, it is important to define the type of data used and the problems related to it. Data are taken from the Human Mortality Database, gathering 41 countries demographic data [HMD, 2021].
The variables needed in this framework are the number of death and the exposure to risk[1]. All theses variables are presented as a matrix with ages for the row and the period of interest for the columns. By dividing the number of death with the size of the population, the instantaneous mortality rate is obtained, namely the variable of interest[2]. The focus is not done on the other variables, as the number of death is not used anymore after applying this relation and the exposure to risk seems to register plausible data.

Instantaneous mortality rates represent probabilities[3] and are noted $\mu_x(t)$, where x is the age of the people concerned and t the year in consideration. Given that, $\mu_x(t)$ follows probability theory and are subject to the three Kolmogorov axioms [Wackerly et al., 2002]:

" Suppose S is a sample space associated with an experiment. To every event A in S (A is a subset of S), we assign a number, P(A), called the probability of A, so that the following axioms hold:

Axiom 1 (Kolmogorov). *P(A)* ≥ 0

Axiom 2 (Kolmogorov). *P(S) = 1*

Axiom 3 (Kolmogorov). *If A1, A2, A3,. . . form a sequence of pairwise mutually exclusive events in S (that is,* $A_i \cap A_j = \emptyset$ *if* $i \neq j$ *), then P(A1* $\cup$ *A2* $\cup$ *A3* $\cup$ *...)=* $\sum_{i=1}^{\infty} P(A_i)$."

A first problem of outliers is that $\mu_x(t)$ are sometimes overestimated violating Axiom 2. For instance, in Sweden 1947, a mortality rate of 102,6% is registered for a 101 years person. A second problem is that $\mu_x(t)$ are sometimes wrong or missing. Why wrong? Because mortality rates are assumed to increase until the last age where there are still survivors. For example, in Sweden 1955, rates at 97, 98, 99 and 100 years old are respectively equal to 45.18%, 38.38%, 32.61% and 41.92%.
Theses problems are more present at older ages due to the few number of people still alive. This creates a lot of variance for theses ages and it is a well known problem in the life

[1] Theses variables are explained in details in section 4.1.1 and 4.1.2

[2] This relation is given by expression (9) and is explained in section 4.1.2

[3] In fact, it is an approximation of a conditional probability but, here, the simplification is made for the outliers management

insurance industry. Once 100 years old is overpassed, it is really difficult to find realistic figures.
To overpass this problem of outliers, the inconsistent morality rates have been replaced by a rate which constitutes the upper bound. After that, a Whittaker-Henderson method is used to smooth rates at oldest ages and reduce the range of age in order to obtain an entirely increasing curve.

3.1 Smoothing: Whittaker-Henderson

This section is dedicated to the explanations of the smoothing process implemented in this framework. Before in 1899, methods of smoothing were linked to a moving average filter that cannot smooth the upper end of the mortality curve and integrate a smoothing parameter. To answer these two problems, G. Bohlmann implemented the Whittaker-Henderson (W-H) method which is nowadays considered as a well known procedure in the actuarial world [Weinert, 2007].
The purpose of this method is to define $\widehat{\mu}_x^s(\mathrm{t})(\mathrm{h})$[4] as the estimator of the empirical mortality curve $\mu_x(\mathrm{t})$ which minimizes the following formula:

$$WH_h[\mu_x(t)] = F[\mu_x(t)] + hS[\mu_x(t)] \tag{1}$$

On one hand, equation (2) gives the quality of fit component $F[\mu_x(t)]$ reconciling the true curve and the estimated one together.

$$F[\mu_x(t)] = [\mu_x(t) - \widehat{\mu}_x^s(t)(h)]^T \mathbf{W} [\mu_x(t) - \widehat{\mu}_x^s(t)(h)] \tag{2}$$

$$\mathbf{W} = \begin{bmatrix} \mathrm{N}_1 & 0 & ... & 0 \\ 0 & \mathrm{N}_2 & ... & .. \\ .. & 0 & .. & 0 \\ 0 & ... & 0 & \mathrm{N}_m \end{bmatrix}$$

Where $\mathbf{W}$ is a diagonal matrix of weights N_i[5] with a dimension of $m \times m$ depending on the chosen age's range of $\mu_x(t)$ and $\mathbf{h}$ represents the smoothing parameter given the opportunity to increase the smoothness of the curve. On the other hand, equation (3) gives the smoothness component suggesting that *a priori* the true curve is smoothed.

$$S[\mu_x(t)] = \mu_x(t)^T K_Z^T K_Z \mu_x(t) \tag{3}$$

[4]In fact, the method is applied on mortality rates $q_x(\mathrm{t})$ but, here, the simplification is made for an outliers management purpose

[5]Here, the weights are all equal to 1 to keep a constant smoothing along the curve.

The point is to establish a compromise between these two components (2) and (3) to obtain $\widehat{\mu}_x^s$(t)(h). Before developing the construction of K_Z, the difference operator Δ need to be defined. This operator allows to evaluate the smoothness of the curve in discrete times through finite differences.

$$(\Delta^z q_x) = \Delta(\Delta^{z-1} q_x) \tag{4}$$

The exponent **z** gives a constant, linear or polynomial character to the curve. Here, the value of **z** is equal to 2. By integrating the previous information, the formula (4) becomes:

$$\begin{aligned}\Delta^2 \mu_x(t) &= \Delta[\mu_{x+1}(t) - \mu_x(t)] = [\mu_{x+2}(t) - \mu_{x+1}(t)] - [\mu_{x+1}(t) - \mu_x(t)] \\ &= \mathbf{1}\mu_x(t)\text{-}\mathbf{2}\mu_{x+1}(t) + \mathbf{1}\mu_{x+2}(t)\end{aligned}$$

$\Delta^z \mu_x(t)$ being a linear function, the following diagonal matrix K_Z of size $(m-z)\times m$ is obtained :

$$K_Z = \begin{bmatrix} \mathbf{1} & \mathbf{-2} & \mathbf{1} & 0 & ... & 0 \\ 0 & \mathbf{1} & \mathbf{-2} & \mathbf{1} & .. & .. \\ .. & 0 & .. & .. & .. & 0 \\ 0 & ... & 0 & \mathbf{1} & \mathbf{-2} & \mathbf{1} \end{bmatrix}$$

By extracting $\widehat{\mu}_x^s$(t)(h) from the expression (1), the W-H mortality rate estimator obtained is stated below:

$$\widehat{\mu}_x^s(t)(h) = [W + hK_Z^T K_Z]^{-1} W \mu_x(t)$$

In order to have a visual perception of this smoothing method and the operation made on the data, the figure 1 represents the surface of mortality of Swedish women from age 20 to 100 years old since 1946.

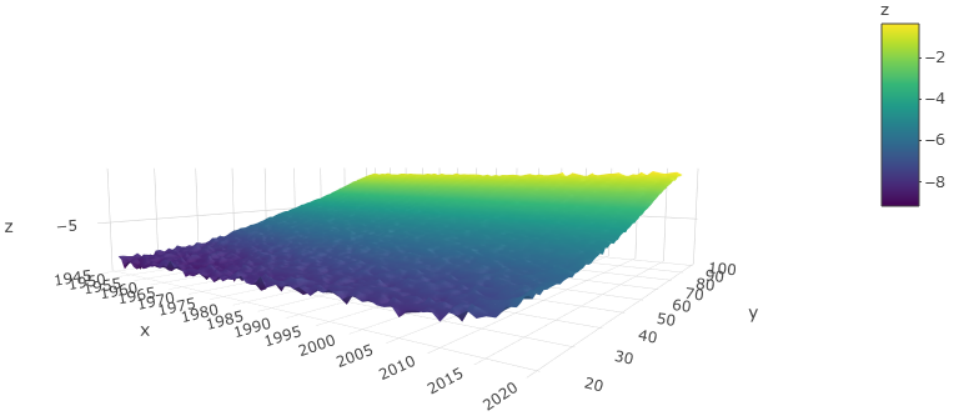


Figure 1: Surface of mortality - Sweden F

As it is observable, noise is still present for younger ages. However, this is not a problem in this case, as the purpose is managing outliers and not properly smoothing the curve at each age.

4 Modelling

This section is devoted to the explanations of concepts and models used in this framework. The first part of this section explains the actuarial notions involved in the mortality model. The second part covers the deep learning part of this master thesis explaining what exactly a feed-forward neural network is. A link is done between these parts to clarify how the models are combined in a dimension reduction context.

4.1 Actuarial Model: Poisson Model

The outline of this section is as following. First, the instantaneous mortality rate and lifetime expectancy is defined through basic concepts of life insurance. Lifetime expectancy is a key element to analyse longevity. Then, a link between classical mortality and prospective tables is done to define the exposure to risk. After that, Lee Carter model and the Age Period models are explained as it is used as a benchmark in the results section. Finally, the actuarial model, on which this framework is based, namely the Poisson model, is explained.

4.1.1 Probability of Dying, Mortality Rate and Lifetime Expectancy

T_X constitutes a continuous random variable corresponding to the remaining lifetime of an individual. The probability of dying at time t[6] given that the individual is alive at age x is defined by expression (5) and could be rewritten, with the following equation, as the inverse probability of living until age $x+t$ [Dickson et al., 2009]:

$$P(T_X \leq t) = {}_tq_x = 1 - {}_tp_x = 1 - P(T_X > t) \tag{5}$$

As a real-valued random variable, X has its cumulative distribution function (CDF) defined by $F_X(x) = P(X \leq x)$ and its survival function defined by $S_X(x) = 1 - F_X(x) = P(X > x)$[7]. Therefore, expression (5) is considered as the CDF of T_X and $P(T_X > t)$ its survival function being the probability of living. This probability is intuitively defined as the ratio of the number of survivors L_x between two years such as:

$$P(T_X > t) = \frac{L_{x+t}}{L_x} \tag{6}$$

In a given population, equation (5) is calculated with $\sum_{i=0}^{t-1} D_{x+i}$[8], the number of death during a certain period t, divided by the number of people alive L_x. As the difference of

[6]Previously, t was considered as a calendar year. In this section, it is a number of years.

[7][Wackerly et al., 2002] and [Dickson et al., 2009]

[8]D_x being the number of death at age x

survivors between two years represents the number of deaths for this period, from equation (6), the following equation is infered:

$$P(T_X \leq t) = 1 - \frac{L_{x+t}}{L_x} = \frac{L_x - L_{x+t}}{L_x} = \frac{\sum_{i=0}^{t-1} D_{x+i}}{L_x}$$

A first assumption is made concerning D_x claiming that each death is considered to happen in the middle of the year.

By applying the Bayes Law[9], the probability of dying between two moments could be written as follow:

$$\begin{aligned} P(t \leq T_X \leq t+s | T_X \geq t) &= \frac{P(t \leq T_X \leq t+s)}{P(T_X \geq t)} = \frac{{}_{t+s}q_x - {}_tq_x}{{}_tp_x} = \frac{{}_tp_x - {}_{t+s}p_x}{{}_tp_x} \\ &= 1 - \frac{{}_{t+s}p_x}{{}_tp_x} = \frac{{}_{t|s}q_x}{{}_tp_x} \end{aligned}$$

As the time is a continuous variable, the previous expression is used to show the link between the probability of dying and the instantaneous mortality rate using limits such as:

$$lim_{\Delta \to 0} \frac{P(t \leq T_X \leq t+\Delta | T_X \geq t)}{\Delta} = lim_{\Delta \to 0} \frac{P(t \leq T_X \leq t+\Delta)}{\Delta P(T_X \geq t)} = \frac{-{}_tp'_x \Delta t}{{}_tp_x}$$

As ${}_tp_x$ is considered as a derivable function, the instantaneous mortality rate μ_{x+t} is obtained by the following equation:

$$\mu_{x+t} = -\frac{l'_{x+t}/l_x}{l_{x+t}/l_x}\Delta t = -\frac{l'_{x+t}}{l_{x+t}}\Delta t = \frac{d}{d_{x+t}}[ln(l_{x+t})]$$

The exponential relation between the probability of living and the instantaneous mortality rate is derived from the previous expression for age ξ by integrating it over [x;x+t] such as:

$$\begin{gathered} -\int_x^{x+t} \mu_\xi d\xi = ln(l_{x+t}) - ln(l_x) = ln({}_tp_x) \\ \Longleftrightarrow \\ {}_tp_x = exp(\int_x^{x+t} \mu_\xi d\xi) \end{gathered}$$

Therefore, by assuming that these rates are piecewise constant over [t-1;t] with $t \in \mathbb{N}$, the biased estimator $\widehat{\mu}_x$ is approximated by the following equation:

$$p_x \approx exp(-\mu_x) \Longleftrightarrow \widehat{\mu}_x \approx -ln(p_x) \approx -ln(1-q_x) \tag{7}$$

[9] P(A|B)=P(A∩B)/P(B) - [Wackerly et al., 2002]

As this relation only depends on one factor, this estimator is widely used in the actuarial industry. By the way, it is the estimator used by Professor D. Hainaut in the benchmark paper of this master thesis [Hainaut, 2018].

By assuming that the time before a death and the number of deaths are independent, it is derivated that death time can be interpreted as the moment of the first jump of an non-homogeneous Poisson process.

$$ {}_tp_x = P(N_t = 0) $$

Where N_t represents a non-homogeneous Poisson process with intensity $\mu(\xi) = \mu_{x+\xi}$

$$ P(N_t = k) = \frac{(\int_0^t \mu(\xi)d\xi)^k}{k!} exp(-\int_0^t \mu(\xi)d\xi) $$

$$ P(N_t = 0) = exp(-\int_0^t \mu(\xi)d\xi) $$

A longevity measure is required to compare the benchmark models with the one implemented in this framework. In order to evaluate if the models effectively capture longevity, the lifetime expectancy or remaining expected lifetime at a certain age is needed. Expression (5) has been previously defined as the CDF of T_X and is the starting point to infer the longevity measure. As the probability distribution function (PDF) of a continuous random variable X is $f(x) = \frac{dF_X(x)}{d(x)}$ as long as the derivative exists[10], the PDF of T_X is quickly derived by the following equation:

$$ \frac{\delta P(T_X \leq t)}{\delta t} = \frac{\delta(1 - {}_tp_x)}{\delta t} = -\frac{\delta_t p_x}{\delta t} = {}_tp_x\mu_{x+t} $$

Thus, the remaining expected lifetime at age x, noted $E(T_X)$ or e_x(t), is defined as:

$$ E(T_X) = \int_0^{\omega-x} P(T_X > u)du = \int_0^{\omega-x} {}_up_x du $$

The conditional expected lifetime at age x is written such as:

$$ E(T_X) = E(T_X - x|T_X > x) = \int_0^{\omega-x} {}_up_x\mu_{x+t}du $$

As mentioned in expression (7), an approximation is made linking probability of dying with the instantaneous mortality rates. Given that, the expression hereabove can be rewritten as follows:

$$ E(T_X) = \sum_{t=1}^{\omega-x} {}_tp_x \tag{8} $$

with ${}_tp_x \approx exp(\sum_{y=x}^{x+t-1} -\mu_y)$

[10] [Wackerly et al., 2002]

4.1.2 Prospective Tables & Exposure to Risk

In the previous section, actuarial notions have been given on classical mortality table basis. It means that t represented a number of years and x the age on a certain calendar year. The purpose to this section is to extend the notion needed to prospective tables[11].

Prospective tables are interesting because they follow population with its associated cohorts. It means that mortality is registered by generations rather than yearly aggregated. Classical mortality tables have turned obsolete because those were not able to capture mortality improvements. Thus, the year of birth becomes an important parameter.
$q_x(t)$ corresponds to the death probability at age x during the calendar year t and is estimated as follows:

$$\widehat{q_x}(t) = \frac{D_x(t)}{L_x(t)}$$

Where $L_x(t)$ and $D_x(t)$ being respectively, in year t, the number of survivors and the number of deaths from a x years old population. The instantaneous mortality rates are assumed piecewise constant (PC assumption), so that it is uniformly distributed over a period of one year. As it has been mentioned previously, the year of birth is become a key element for prospective tables. That is why the exposure to risk $ETR_x(t)$ is an important notion that is used in several models. At age x with the last birthday happening during year t, $ETR_x(t)$ is the total time lived by people with the same age x and during the same year t.

As previously established in expression (7), the estimator of $\widehat{\mu}_x$ is approximated by making the PC assumption on the mortality force but this estimator is biased. Via the $ETR_x(t)$ and through the bias of the likelihood function, an unbiased estimator of $\widehat{\mu}_x$ is computable. The proof is derived as follows:

By considering a survival indicator $X_i(t)$ for each individuals in $L_x(t)$ population that follows a Bernoulli:

$$X_i(t) = \begin{cases} 0 & \text{if the individual died at age x} \\ 1 & \text{otherwise} \end{cases} \quad with : i = 1, .., L_x(t)$$

There is:

$$\sum_{i=1}^{L_x(t)} X_i = D_x(t)$$

[11] This section is inspired by [Antoine, 2006]

The part[12] of the year lived by the i^{th} person in this population is noted τ_i. Given the previously mentioned definition of $ETR_x(t)$, there is:

$$\sum_{i=1}^{L_x(t)} \tau_i = ETR_x(t)$$

Under the PC assumption made on the mortality force, the contribution of the i^{th} individual to the likelihood is given such as:

$$\begin{cases} \text{If the individual survives:} & {}_tp_x = exp(-\mu_x(t)) \\ \text{If he dies during year t:} & \tau_{it}p_x\mu_{x+\tau_i}(t+\tau_i) = exp(-\mu_x(t)\tau_i)\mu_x(t) \end{cases}$$

The likelihood is then:

$$L(\mu_x(t)) = \prod_{i=1}^{L_x(t)} exp(-\mu_x(t)\tau_i)(\mu_x(t))^{X_i} = exp(-\mu_x(t)ETR_x(t))(\mu_x(t))^{D_x(t)} \tag{9}$$

The estimator is found by setting the derivative of $ln\ L(\mu_x(t))$ to zero:

$$\mathcal{L}(\mu_x(t)) = -\mu_x(t)ETR_x(t) + D_x(t)ln(\mu_x(t))$$

$$\frac{\delta\mathcal{L}(\mu_x(t))}{\delta\mu_x(t)} = -ETR_x(t) + \frac{D_x(t)}{\mu_x(t)} = 0$$

$$\Longleftrightarrow$$

$$\widehat{\mu}_x(t) = \frac{D_x(t)}{ETR_x(t)} \tag{10}$$

Therefore, equation (10) corresponds to the unbiased estimator that is used to estimate $\widehat{\mu}_x(t)$ the instantaneous mortality rate.

4.1.3 Lee Carter to Poisson Model

Initially fitted on 1933-1987 American data, the Lee Carter model (LC model) has been implemented in 1992 by R. Lee and L. Carter to extrapolate past trends to the period 1990-2065 [Lee and Carter, 1992]. Being a widely used benchmark in the insurance industry, the LC force of mortality is driven by the following relation:

$$ln\mu_x(t) = \alpha_x + \beta_x\kappa_t \tag{11}$$

[12] expressed in percent

Where α_x, β_x are defined for $x_0 = x_1$, ..., x_m and κ_t is a random process.

This model has a variant, called the Age Period Cohort model (APC model), with a cohort dimension given by the γ_{t-x} component such as:

$$ln\mu_x(t) = \alpha_x + \kappa_t + \gamma_{t-x} \tag{12}$$

Those models are both used further to forecast and simulate mortality rates and by extension lifetime expectancies. This allows to have a benchmark to compare forecast simulated mortality rates and lifetime expectancies implemented by the feed-forward neural networks built in this framework.
α_x corresponds to the mean of log rates representing the accident component. $\beta_x\kappa_t$ represents the aging component separated with β_x being the marginal reduction of mortality at each age and κ_t capturing the evolution of mortality. The last component is really important in this framework. This is discussed later in the "Towards Dimension-Reduction" section.
The following strong assumption is made in this model concerning the noise $e_{x,t}$ in expression (13):

$$ln\widehat{\mu}_x(t) = \alpha_x + \beta_x\kappa_t + e_{x,t}$$

$$ln\widehat{\mu}_x(t) - \alpha_x - \beta_x\kappa_t \sim N(0, \sigma) \tag{13}$$

This process follows a normal distribution with null mean and a constant volatility. However, it is seen as a big weakness in this model because it supposes that the volatility is homoscedastic. A model is considered homoscedastic when the variance of the residuals stays constant for each observation [Davidson and MacKinnon, 1993]. As it has been previously discussed for the management of outliers, there is a lot of variance in mortality rates at the oldest ages due to the smaller number of people still alive.

4.1.4 Poisson Model

The Poisson regression model is used to model a counting variable Y which occurs during a given time interval or on a given space interval. In this context, the counting variable could be related to the number of deaths. By combining it with the expression (10), the statistical model is turned in a mortality model such as:

$$Y = D_x(t) \sim Poi(\lambda = ETR_x(t)\mu_x(t))$$

Where $\mu_x(t) = exp(\alpha_x + \beta_x\kappa_t)$ directly refered to the Lee-Carter model.

The probability to have d deaths for people born in year t at age x is defined such as:

$$P(D_x(t) = d) = \frac{exp(-\mu_x(t)ETR_x(t))(\mu_x(t)ETR_x(t))^d}{d!}$$

The Poisson likelihood is proportional to the true likelihood and is given by the following equation:

$$L(\mu_x(t)) = exp(-\mu_x(t)ETR_x(t))(\mu_x(t)ETR_x(t))^{D_x(t)} \tag{14}$$

This expression will be important in the section explaining the development of the Poisson deviance.

As the purpose is to model κ(t), the fit is done by the maximum likelihood estimation for the remaining parameter $\widehat{\alpha}_x$, μ_x having the same structure as the LC model (cf. equation (11)). The following equations correspond to the constraints set in order to avoid an identification problem:

$$\sum_{x_1}^{x_m} \beta_x = 1$$

$$\sum_{t_1}^{t_n} \kappa_t = 0$$

Given these constraints, $\widehat{\alpha}_x$ is derived from the observations such as:

$$argmin \sum_{t=t_1}^{t_n} \sum_{x=x_1}^{x_m} (ln\widehat{\mu}_x(t) - \alpha_x - \beta_x\kappa_t)^2 \tag{15}$$

As expressed by equation (13), the noise is gaussian with a constant variance. Therefore, the least squares criterion (15) corresponds to the maximum likelihood estimation used to fit $\widehat{\alpha}_x$. Its estimation is derived by the following development:

$$\frac{\delta}{\delta\alpha_x} \sum_{t=t_1}^{t_n} \sum_{x=x_1}^{x_m} (ln\widehat{\mu}_x(t) - \alpha_x - \beta_x\kappa_t)^2 = 0$$

$$\sum_{t=t_1}^{t_n} ln\widehat{\mu}_x(t) - n\alpha_x - \beta_x \sum_{t=t_1}^{t_n} \kappa_t = 0$$

$$\widehat{\alpha}_x = \frac{1}{n} \sum_{t=t_1}^{t_n} ln\widehat{\mu}_x(t) \tag{16}$$

Where **n** represents the number of year sample on which data are fitted.

Once that $\widehat{\mu}_x$ and $\widehat{\alpha}_x$ are estimated (respectively expressions (10) and (16)) , one needs to set data as the input of the neural network in order to implement the dimension-reduction. This topic is covered in the next section.

4.2 Towards Dimension-Reduction

This section is devoted to the modelling approach considered in this framework. In other words, the purpose is to explain how a non linear principal component analysis (NLPCA) and a singular value decomposition (SVD) are combined to perform a dimension-reduction in the context of a feed-forward Neural Network. The main idea behind dimensionalty reduction is reducing the dimension of the data on a intrinsic dimensionalty of these data. This intrinsic dimensionalty is generally unknown and subject to assumptions. Thus, dimension-reduction is considered as an ill-posed problem. The general purpose of dimension reduction is to reduce the size of data for a gain of space or to reduce the time of calculations [Maaten et al., 2009]. In this framework, the dimension reduction is used to model κ_t^i from LC model.

Once $\widehat{\alpha}_x$ is derived, one needs to center on 0 in order to set up a non linear principal component analysis. The aim is to decorrelate the variables to obtain a new set of variables called the principal components [Shlens, 2014].

The first step is subtracting $\widehat{\alpha}_x$ from $ln\mu_x(t)$ in order to establish a singular value decomposition on $\beta_x\kappa_t$, being the principal components. The following matrix $\mathbf{M}$ of dimensions $\mathbf{m}\times\mathbf{n}$, respectively being the range of age and the number of year, is the input of the neural network:

$$\mathbf{M} = ln\mu_x(t) - \alpha_x \tag{17}$$

Mathematically, the NLPCA executes a non linear transformation translating the initial set of features, such as the matrix of residual observations $\mathbf{M}$, to a new space composed by principal components, being the eigenvectors of $\mathbf{M}^T\mathbf{M}$ [Hainaut, 2018]. To do so, it is necessary to deduce the SVD of $\mathbf{M}$ to separate the eigenvectors $\mathbf{v}_i$ from the normed eigenvectors $\mathbf{u}_i$ of $\mathbf{M}^T\mathbf{M}$:

$$M = \sum_{i\geq 1} \sqrt{\lambda_i v_i^T u_i}$$

Where parameters $\lambda_1 \geq \lambda_2 \geq ... \geq 0$ represent the eigenvalues of the product of matrix $\mathbf{M}^T\mathbf{M}$.

$$\kappa_t^i = \sqrt{\lambda_i} \sum_{j=x_1}^{x_m} v_{i,j}\mathbf{u}_i$$

$$where\ i \in [1, d]$$

$$\beta_x^i = \frac{1}{\sum_{j=x_1}^{x_m} v_{i,j}}\mathbf{v}_i$$

$$where\ i \in [1, r]$$

A dimensionality reduction is thus established by projecting points at time t representing mortality curves from a space of size **n** to a lower dimension hyperplan of size **d**. The result is a vector of d-plet $\kappa_t^i = (\kappa_t^1, \kappa_t^2.., \kappa_t^d)$ containing the coordinates of the mortality curve. At this step, the non linear part of the principal component analysis has not been explained yet because it is performed by the neural networks. Moreover, in the context of the dimensionalty reduction with a neural network, the ill-posed problem is avoided by the cross-validation defining the best number of d for the model. More details are given in the next section once the feed-forward network is explained.

4.3 Machine Learning Model: Feed-Forward Neural Network

This section is dedicated to the explanation of the multi-layer perceptron (MLP) and in particular the feed-forward neural network (NN). This machine learning method is composed by a succession of neuron layers which are connected layers by layers [Azencott, 2018]. A single neuron is called a perceptron and is represented by the following illustration (Figure 2):

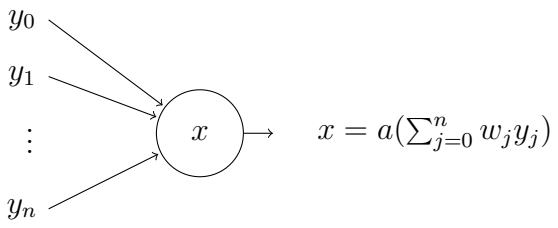


Figure 2: Single neuron perceptron

An input vector $\vec{\boldsymbol{y}}$ of dimension $n+1$ is connected through an activation function $a()$ to the next layer, being the single neuron x. Each connection has an attributed weight w_j which represents the parameters of the model. Mathematically, the output signal of the neuron x is the result of the linear combination of the inputs vector's elements and those weights to which an activation function is applied. A multi-layer perceptron is then obtained by assembling the perceptrons in a succession of layers. Each layer has an assiociated activation function. A 3-2-3 MLP is represented in the hereafter Figure 3.

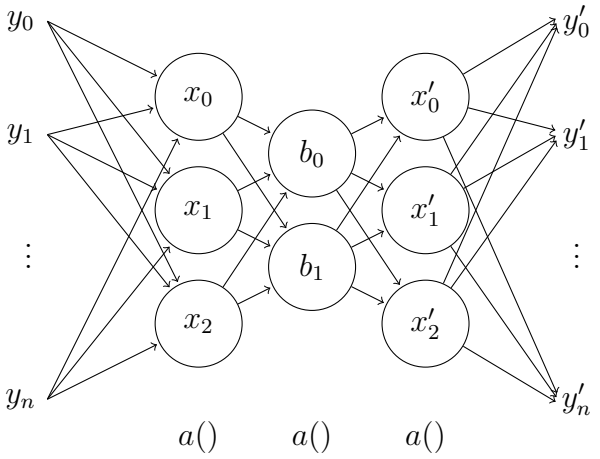


Figure 3: Multi-layer perceptron (3-2-3)

The muti-layer perceptron corresponds to a feed-forward neural networks when the information flows in one direction as described in Figure 3. The NN of Figure 3 is composed of 3 layers and a respective number of neurons or nodes equal to 3, 2 and 3. The NN's architecture is a bottleneck as central layers ($\vec{\boldsymbol{x}}$, $\vec{\boldsymbol{b}}$, $\vec{\boldsymbol{x'}}$), called hidden layers, contain less elements than the extreme ones. The advantage of the neural network is that its architecture is easy to manipulate. Indeed, in R, implementing a NN with a certain structure is easily done and its shape and activation function between layers can be changed at will. The R packages used to implement the model are called "tensorflow" and "keras" [13].
The dimensionality reduction happens in those hidden layers to end in the bottleneck. Depending on the activation function attributed to the layers, the reduction could be linear or non linear. In this framework, a 3 hidden layers NN is built with an identity function for the central layer and non linear functions (hyperbolic tangent sigmoid function) for the first and the third one. Both functions are respectively represented in equations (18) and (19):

$$a(x) = x \tag{18}$$

$$a(x) = \frac{2}{1 + exp(-x)} - 1 \tag{19}$$

The hyperbolic tangent sigmoid function (19) is used to perform the non linearity of the principal component analysis. This activation function has been selected because of its domain of definition is defined between -1 and 1 centered on 0 to be in concordance with

[13][Abadi et al., 2015] and [Chollet et al., 2015]

the NLPCA. The following Figure 4 shows the shape of the function:

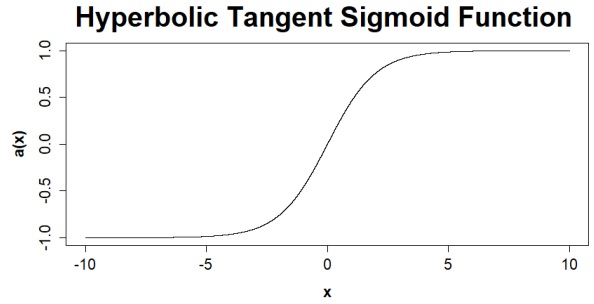


Figure 4: Hyperbolic Tangent Sigmoid Function

4.3.1 Loss Function: Deviance

A MLP is built so that the distance between input $\vec{\boldsymbol{y}}$ and output signal $\vec{\boldsymbol{y'}}$ is minimized for regression or classification purposes. Concerning the learning part of the model, a loss function combined with an algorithm is used to set weights w_j which, as previously mentioned, represents the parameters of the model. The loss function could be seen as the optimisation criterion to train the model converging towards the estimated parameters [Denuit et al., 2019]. The function to minimize is defined as follows:

$$\Omega = argmin \frac{1}{n} \sum_{i=1}^{n} L(y_i, y_i')$$

Ω corresponds to the matrix of weights $w_{i,j}$ which are associated to the neuron (i,j) depending on the MLP's configuration. This criterion is a penalty function that may be, for instance among others a deviance, a quadratic function, *etc.* depending on the purpose of the model.

In this framework, the loss function is a Poisson deviance as the number of deaths is related to a Poisson distribution. This random variable Y_i is divided by the exposure v_i to obtain the key quantity $Y_i/v_i = y_i$. Indeed, the instantaneous mortality rate can be derived by dividing the number of deaths by the exposure to risk as in the equation (10). A parallel is done during the development of the Poisson deviance. The unscaled deviance is therefore taken as loss function and is equal to:

$$L(y_i, \widehat{y}_i) = D(y_i, \widehat{y}_i) = \phi D(y_i, \widehat{y}_i) = \phi 2[l_s(y_i) - l(\widehat{y}_i)] \tag{20}$$

This function is obtained by multiplying by 2 ϕ the difference of log-likelihood of the saturated model (y_i) and the model considered $(\widehat{y}_i)$. As the definition's domain of a Poisson distribution is defined on $\mathcal{R}^+$, an exponential transformation of the output signal is

required as it has been centered on 0 for the NLPCA. By starting from the likelihood implemented at an individual level previously defined by expression (9), the Poisson deviance is infered in the following way:

$$L(\mu_x(t)) = \prod_{i=1}^{L_x(t)} exp(-\mu_x(t)\tau_i)(\mu_x(t))^{X_i} = exp(-\mu_x(t)ETR_x(t))(\mu_x(t))^{D_x(t)}$$

By applying the log on the likelihoods:

$$l_s(\mu_x(t)) = -\mu_x(t)ETR_x(t) + D_x(t)ln(\mu_x(t))$$
$$l(\widehat{\mu}_x(t)) = -\widehat{\mu}_x(t)ETR_x(t) + D_x(t)ln(\widehat{\mu}_x(t))$$

The equation (20) becomes:

$$D(\mu_i, \widehat{\mu}_i) = 2\phi[-\mu_x(t)ETR_x(t) + D_x(t)ln(\mu_x(t)) - (-\widehat{\mu}_x(t)ETR_x(t) + D_x(t)ln(\widehat{\mu}_x(t)))]$$

As ϕ=1, for poisson and binomial distributions:

$$D(\mu_i, \widehat{\mu}_i) = 2[-\mu_x(t)ETR_x(t) + D_x(t)ln(\mu_x(t)) + \widehat{\mu}_x(t)ETR_x(t) - D_x(t)ln(\widehat{\mu}_x(t)))]$$

With relation (10):

$$D(\mu_i, \widehat{\mu}_i) = 2[-\mu_x(t)ETR_x(t) + \mu_x(t)ETR_x(t)(t)ln(\mu_x(t)) + \widehat{\mu}_x(t)ETR_x(t) - \mu_x(t)ETR_x(t)ln(\widehat{\mu}_x(t)))$$

$$D(\mu_i, \widehat{\mu}_i) = 2ETR_x(t)[-\mu_x(t) + \mu_x(t)(t)ln(\mu_x(t)) + \widehat{\mu}_x(t) - \mu_x(t)ln(\widehat{\mu}_x(t)))]$$

By respecting the Poisson's domain of definition, the equation finally becomes:

$$\mathrm{D}(\mu_x(t), \widehat{\mu}_x(t)) = \begin{cases} 2ETR_x(t)[-\mu_x(t) + \mu_x(t)ln(\mu_x(t)) + \widehat{\mu}_x(t) - \mu_x(t)ln(\widehat{\mu}_x(t))] & \mu_x(t) > 0 \\ 2ETR_x(t)\widehat{\mu}_x(t)) & \mu_x(t) = 0 \end{cases}$$

This equation represents the unscaled deviance of a Poisson distribution used as optimisation criterion in the model. The NN trains by minimizing the difference of the log-likelihood of the saturated model, being the observed one and the estimated model on a Poisson law basis.

As previously mentioned in section 4.2, the matrix $\mathbf{M}$ in expression (17) is the input of the neural network. This matrix could be seen as a sample of $\mathbf{n}$ year vector with a size equal to the chosen range of age $\mathbf{m}$. Each vector enters in the feed-forward neural network as the inputs signal $\vec{\boldsymbol{y}}$. Then, the learning operates in order to converge to estimated κ_t^i located in the bottleneck $\vec{\boldsymbol{b}}$ of the NN. The dimension of the lower hyperplan $\mathbf{d}$ represents the number of κ_t^i which is the size of the bottleneck. The determination of the size of the bottleneck is explained in the next section.

4.4 Training, validation and forecasting

As explained in the previous part, the model learns with the observed data to be able to predict after it is done. However, in order to know if the predictive power of a model is good, it is important to confront it to new data. A back-testing of the model can be performed by spliting the initial dataset in a training test and a test set as shown in Figure 5. However, a validation of the model should previously be done to define the best model among all machine learning models differentiated by their hyperparameters. Once the best model is defined in terms of hyperparameters, a forecasting is done on the test set to compare with observed data.

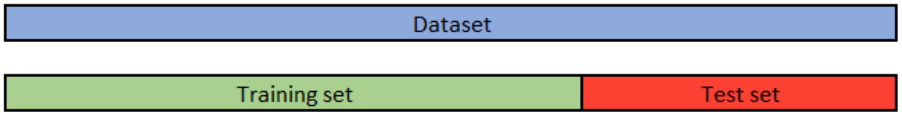


Figure 5: Training vs. Test set

4.4.1 Model validation

The validation of the model is an important step as it allows to determine the related hyperparameters. These parameters are set to define the best model in terms of goodness of fit and especially in terms of predictive power. For the concerned model, the hyperparameters targeted are the number of epochs and the configuration of the neural network. The first validation is done on the number of epochs on a chosen NN configuration. It

represents the number of passes of the training dataset run by the machine learning algorithm [Burkov, 2019]. The configuration of the NN represents the number of neurons in the bottleneck and in the other hidden layers. It is important in this framework as it matches to the hyperparameter that defines the number of κ_t^i. As mentioned in section 4.3.1, the dimension of the lower hyperplan constitutes an ill-posed problem. However, this problem is bypassed by establishing a validation to determine the size of the bottleneck. The poisson deviance[14] has been chosen as a score to evaluate the performance of the validation.

The technique used to perform the validations is called the k-fold cross-validation. This constitutes a robust method as the principle to establish k cross-validations to make a prediction on each k part of the data called fold [Azencott, 2018]. Then, the error of prediction ($\mathbf{e}_i$ in Figure 6) are evaluated for these folds in order to apply an average on it. The data from the validation set are at least once considered as unseen. This prevent the model from overfitting which occurs when the model is too close from the observed data. A 4-fold cross validation is represented in the hereafter Figure 6.

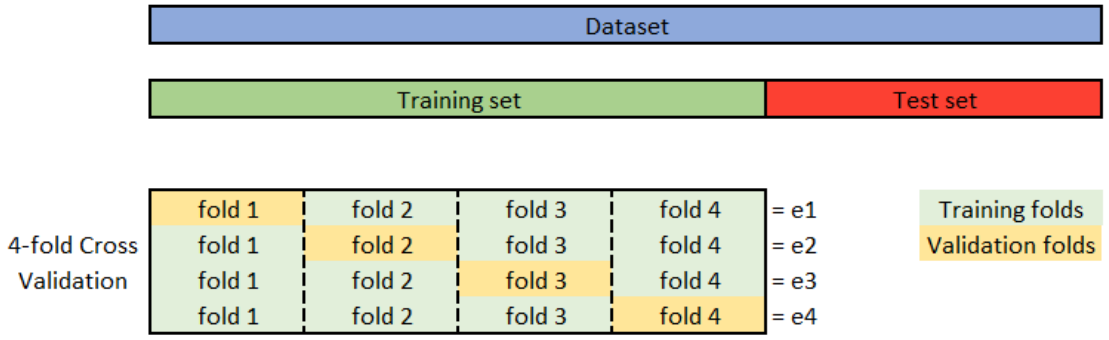


Figure 6: Validation set

4.4.2 Forecasting

In the section 4.3.1 related to the NN loss funtion, the model trains by equalizing the input and the output signal through the NN's structure. That action creates weights that are saved to go on with the forecasting step. A part of theses weights are retrieved from the training model to inject it in a new half NN, represented by the left side of the Figure 7, to reproduce the κ_t^i located in the bottleneck.

At this step, a linear regression is performed on theses κ_t^i with respect to the training set years. Then, the linear model extrapolates the κ_t^i for the test set years. Finally, the new $\kappa_t^{'}$(i) and weights are reinjected in a second half NN to forecast mortality rates.

[14] cf. section 4.3.1

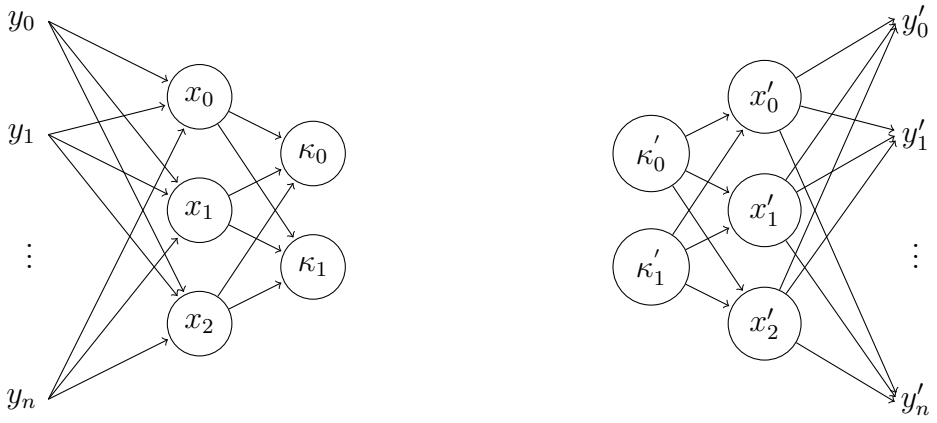


Figure 7: Half neural networks between which a linear regression is performed

5 Results

This section of the master thesis is devoted to the description of the results. First, the analysis is performed on the Swedish population in order to compare the models given the loss function used. Then, a quick comparison is made between male and female. [15] For the sake of clarity, male results are presented in the appendix. In the second part, the model is extended to 3 countries, and a comparison is made between the southern and northern countries of the European Union.

5.1 The Swedish population case

It is important to set the outline of this analysis. Firstly, the results of validation are given to set the hyperparameters. Secondly, the goodness of fit and predictive power are reviewed to assess the performance of the model with different loss functions. In the third section, an analysis of κ_t^i trends is established to understand their process pattern. Finally, a comparison is made between the NN model with deviance as loss function and a benchmark model, to check for any possible difference with the insurance market.

The models are compared with the Swedish population as the country have started counting deaths since 1751. The model is trained and validated on the period 1946-2001. The predictions are made for the period 2002-2018. This choice was made as mortality before 1946 is not representative due to the World War II. The split of the dataset is done to get a validation set divisible by 4 in order to perform a 4-fold cross validation, as shown in Figure 6 in section 4.4.1.

5.1.1 Results of validation

As stated earlier, the validation allows to define the structure of the NN in terms of number of neurons in layers. The purpose of actuarial sciences lies on the prediction of future numbers. Therefore, the rule of thumb is to keep the NN's configuration providing the best predictive power. The metric used as a score to evaluate models is the poisson deviance (DEV). The minimum deviance is retained as the best score[16].
To have a point of comparison, another loss function is used to challenge the one implemented in this framework. This loss function is based on the paper of Professor D. Hainaut [Hainaut, 2018]. The function is a mean square error (MSE), $\Omega = argmin \sum_{i=1}^{n} \frac{(y_i - y_i')^2}{n}$.

[15] In the appendix, Figure 28 and 29 gives a visualization of the log-mortality rates for both gender
[16] cf. section 4.3.1

A cross-validation is performed to define the best number of epochs. **360** is the number retained for the epochs and the batches size is set on the number of years used to train the model. The algorithm chosen to perform the convergence to the weights is the command "rmsprop" in R, being the root mean squared propagation. Table 1 and 14[17] show respectively the results of the validation test concerning the configuration of the NN for women and for men.

NN	Goodness of Fit		Predictive Power	
	Deviance	MSE	Deviance	MSE
3-2-3	5144.667	6497.266	4013.424	3704.305
4-2-4	5050.647	5276.860	3987.610	3810.989
5-2-5	4746.936	5509.214	3944.053	3800.326
6-2-6	4992.425	5079.214	3954.005	**3703.652**
7-2-7	4471.976	5697.414	**3807.043**	3865.772
8-2-8	4481.609	4494.290	3964.431	3837.166
3-3-3	5131.544	5607.163	3968.401	3765.459
4-3-4	5051.182	5857.930	4006.179	**3662.538**
5-3-5	4755.558	5283.624	3991.368	3949.709
6-3-6	4924.625	4392.230	3899.578	3998.922
7-3-7	4610.198	4561.092	**3875.501**	3986.124
8-3-8	4544.856	4378.114	3942.189	3892.611

Table 1: Validation results for NN's configuration - F. Sweden

As it is observable for both gender, the best loss function to predict mortality is the MSE. If a focus is done on its performance, the best structure retained is the 4-3-4 for women and the 5-3-5 for men. Concerning the deviance, respectively, it is 7-2-7 and 5-3-5 as for the MSE. Therefore, the rest of the analysis is done with these configurations.

A remark is important to make about the gender. It seems to be more difficult to predict males death given the high differences with women in terms of predictive power. However, the goodness of fit is rather the same, so that the model trains in the same way for both gender.

[17] cf. appendix

5.1.2 Goodness of fit & Predictive Power

The hyperparameters being set, the best models can be run to forecast mortality rates. The figures trained on 1946-2001 are replicated by the models to the predicted part of the data 2002-2018 as explained in the Forecasting section. Then a back testing is performed in order to challenge observed with forecast data. The goodness of fit in 1970 is represented in Figure 8 and 9 and the predictive power is plotted in Figures 10 and 11 for each loss function.

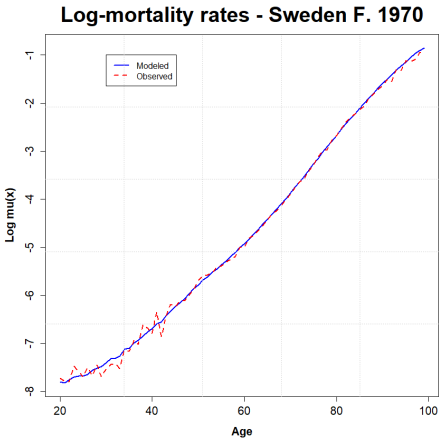


Figure 8: Goodness of fit - MSE

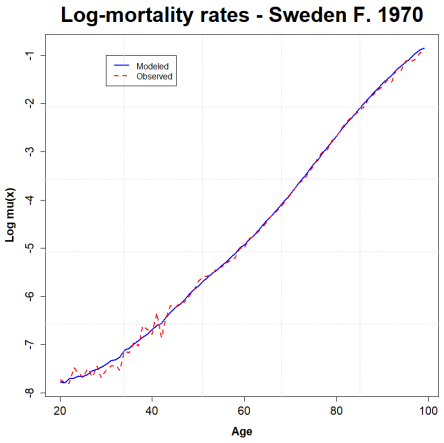


Figure 9: Goodness of fit - DEV

As can be seen in Figure 8 and 9, the noise at younger ages is smoothed by the NN which is a good point for insurers. Indeed, their purpose is to model the mortality from rough data to produce a curve without noise in particular to define contracts prices. An insurance contract is presented in the last part of this master thesis. As it refers to the training part, this result of goodness of fit is not surprising because the model adapt to fit to the observed output signal.
For men[18], there is a little difference concerning the smoothness of the curve. The curves have a slight saw-tooth shape.

[18] cf. Figure 21 and 22 in the appendix

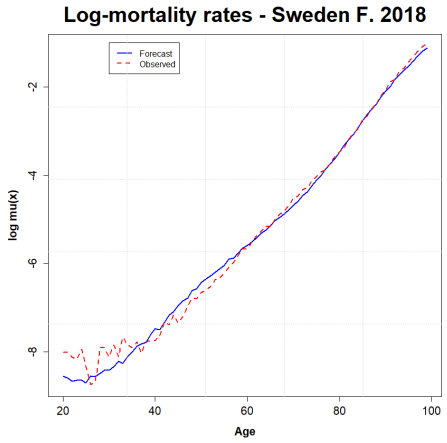


Figure 10: Predictive power - MSE

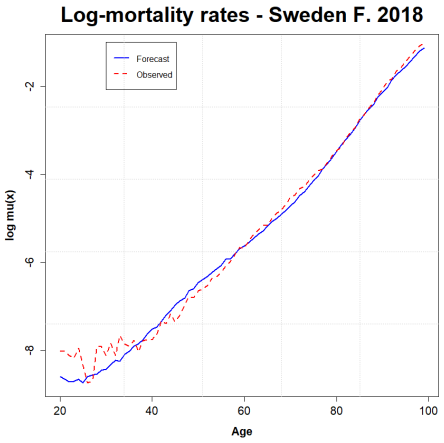


Figure 11: Predictive power - DEV

Figure 10 and 11 show the production by the model of mortality rates on unseen data. It is surprising to see that MSE and NN curves have nearly the same shape. As it is observable, the curves obtained for both functions are quite closed to historical mortality curve except at younger ages. Indeed, between 20 and 40 years old, the forecast curves underestimate mortality. Between 40 and 58 years old, the predicted curves overestimate mortality. However, a good point is that the forecast curves are quite smoothed.
Concerning the Swedish men population[19], the model for both loss function fails to predict mortality as it tends to underestimate it. There seems to have a lot of variance for men in terms of death in Sweden. This gives indications to stop the analysis with this gender for the forecasting and simulating part. However, analysis of κ_t^i trends for men are maintained to confirm theses indications as it relies on the training part.
The Table 2 reports the goodness of fit and the predictive power of the model for both gender in terms of poisson deviance.

Gender	Goodness of Fit		Predictive Power	
	Deviance	MSE	Deviance	MSE
Female	7161.688	6592.015	4507.074	**3651.395**
Male	5944.437	5684.782	**52077.39**	53284.58

Table 2: Goodness of Fit and Predictive Power - Sweden

As it is noticeable, the loss function that produces the best prediction is the MSE for women and the deviance for men. As expected, the precision of the model in terms of predictive power for men is too low with a score 10 times higher than that of women.
Figure 12 represents a graph of the log-mortality rates for 2002, 2010 and 2018 with deviance as loss function. As one can observe, there is a downward shift of the same curve

[19] cf. Figure 23 and 24 in the appendix

over time. This movement of the curve is the traduction of the linear regression applied on κ_t^i and that mortality is reducing over time. The same effect is observable for both loss functions applied on men population [20].

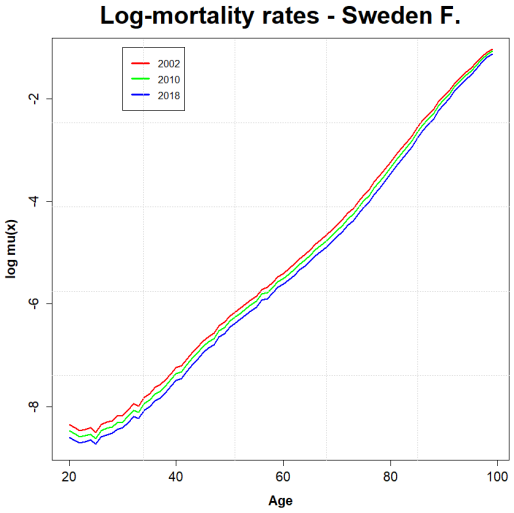


Figure 12: Log μ_x(t) in 2002, 2010 and 2018 produced by 7-2-7 NN with deviance

5.1.3 Analysis of trends

An examination of the κ_t^i trends, the longevity component of the LC model, is important to understand how κ_t^i are evolving and thus trying to simulate given their process pattern. In order to do this, the 2 best NN configurations, respectively 7-2-7 and 4-3-4, are retained to have a better visualization of the possible trends.

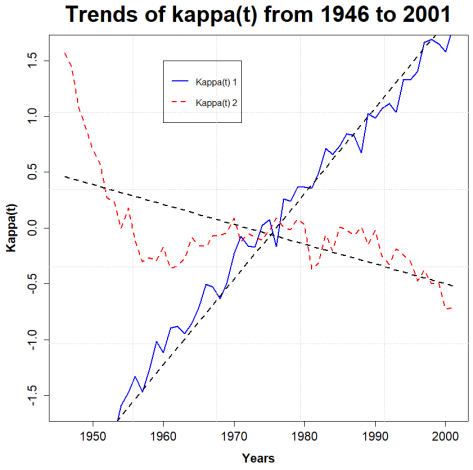


Figure 13: 7-2-7 NN - MSE

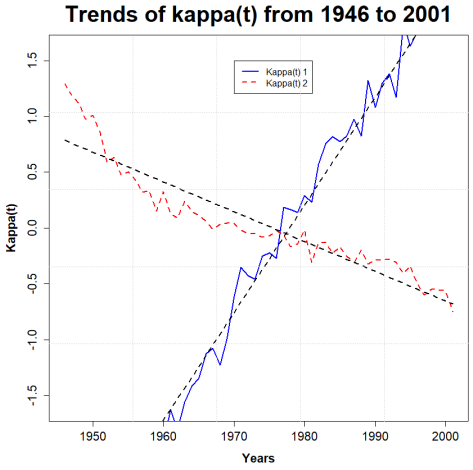


Figure 14: 7-2-7 NN - DEV

[20] cf. Figure 25 in the appendix

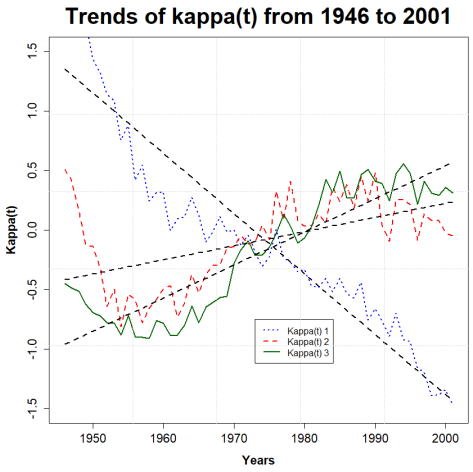


Figure 15: 4-3-4 NN - MSE

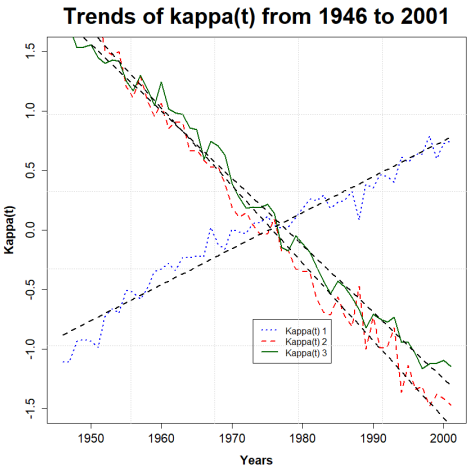


Figure 16: 4-3-4 NN - DEV

As it is observable, there is a big difference between the two loss functions. For the deviance (Figure 14 and 16), the trends seem to follow a random walk with a certain drift. For the MSE (Figure 13 and 15), it is more complicated to make the same assumption given the shape of the κ_t^i. Indeed, by comparing Figure 13 and 14, the κ_t^2 produced with MSE model seem to be more unstable than the one produced by model using the deviance. Therefore, a Jarque Bera test is performed to verify if the increments follow a normal distribution. The null assumption is that the process follows a normal distribution and the condition is given by the following equation:

$$JB = \frac{n-k}{6}(S^2 + \frac{(K-3)^2}{4})$$

With n , k , S and K representing respectively the number of observation, the number of explanatory variables that come from residuals of linear regression, skewness factor and the kurtosis factor [Jarque and Bera, 1987]. The two last components are explained more in details in the following section.

The statistics of the Jarque Bera test and the results of it are compiled in Table 3.

	NN	Increments	JB statistic	Critical value 5%	P-value	$H_0 : Normal$
Deviance	7-2-7	$\kappa_t^1 - \kappa_{t-1}^1$	2.7573	5.004	0.2519	Non-rejected
		$\kappa_t^2 - \kappa_{t-1}^2$	0.19741	5.004	0.906	Non-rejected
	4-3-4	$\kappa_t^1 - \kappa_{t-1}^1$	1.1652	5.004	0.5584	Non-rejected
		$\kappa_t^2 - \kappa_{t-1}^2$	2.6162	5.004	0.2703	Non-rejected
		$\kappa_t^3 - \kappa_{t-1}^3$	0.02691	5.004	0.9866	Non-rejected
MSE	7-2-7	$\kappa_t^1 - \kappa_{t-1}^1$	1.798	5.004	0.407	Non-rejected
		$\kappa_t^2 - \kappa_{t-1}^2$	1.3268	5.004	0.5151	Non-rejected
	4-3-4	$\kappa_t^1 - \kappa_{t-1}^1$	1.2591	5.004	0.5328	Non-rejected
		$\kappa_t^2 - \kappa_{t-1}^2$	0.98468	5.004	0.6112	Non-rejected
		$\kappa_t^3 - \kappa_{t-1}^3$	1.2591	5.004	0.5328	Non-rejected

Table 3: Jarque Bera Test - F. Sweden

As an approximation, the number of observations used for the critical value at 5% is 50 given a critical value equal 5.004.
As expected, the normality of the κ_t^i increments with the deviance cannot be rejected. The same conclusion is made for a configuration that is not the best given the validation results. Surprisingly, it is the case for the MSE as well where, even for the worst of the two model, the normality of the κ_t^i increments it produced is non-rejected.
Given that the increments of the process follow a normal distribution, it is possible to conclude that the model using the deviance or MSE produces κ_t following a random walk with a drift. This is particularly interesting as the simulation of this kind of process is easy to implement. The details of the simulation of a random walk with drift are given in the next section.
Concerning the analysis of κ_t^i trends for Swedish men, the results are nearly similar. Figure 26 and 27 located in the appendix represents respectively the κ_t^i trends of the NN using MSE and poisson deviance. The hereafter Table 4 represents the statistics of the Jarque Bera test and the results of it.

	NN	Increments	JB statistic	Critical value 5%	P-value	$H_0 : Normal$
Deviance	5-3-5	$\kappa_t^1 - \kappa_{t-1}^1$	0.33274	5.004	0.8467	Non-rejected
		$\kappa_t^2 - \kappa_{t-1}^2$	1.7909	5.004	0.4084	Non-rejected
		$\kappa_t^3 - \kappa_{t-1}^3$	0.19652	5.004	0.9064	Non-rejected
MSE	5-3-5	$\kappa_t^1 - \kappa_{t-1}^1$	0.46408	5.004	0.7929	Non-rejected
		$\kappa_t^2 - \kappa_{t-1}^2$	7.4712	5.004	0.023886	Rejected
		$\kappa_t^3 - \kappa_{t-1}^3$	0.25252	5.004	0.8814	Non-rejected

Table 4: Jarque Bera Test - M. Sweden

The same conclusion are made for the NN model using a poisson deviance as loss function. Therefore, model using deviance produce random walk with drift as κ_t^i. However, this is different for model using MSE. Indeed, the assumption of normality for κ_t^2 increments is rejected at a significant threshold of 5%. This conclusion is interesting to note. Even if men data are difficult to forecast, the model using deviance produces random walk with drift when the other does not. Therefore, the model using deviance seem to produce κ_t^i trends easier to interpret as theses are less unstable. As mentioned in section 5.1.2, men's data give a poor efficiency of prediction. Moreover, one of the two loss functions fail to produce interpretable κ_t^i. Therefore, this ends analysis on men's data.

5.1.4 Comparison with market benchmarks: LC model and its cohort variant

This section is devoted to the comparison of the NN model using deviance with the LC model and its cohort variant, the Age Period Cohort Model (APC model). As introduced in the "Lee Carter to Poisson model" section, the two models are respectively given by expressions 11 and 12. It allows to have a reference with a widely used model that predicts mortality in the industry.

As risk management is a key function in the actuarial sciences, it is interesting to simulate (*i.e*: 10 000 times) the projections of κ_t^i in order to capture the distribution of the modeled mortality rates. In a second time, average cross-sectional lifetime expectancies with theses simulations to compare models in terms of longevity. 10000 simulations are done.
For the NN simulation, one must simulate κ_t^i as random walk with drift as previously mentioned in the "analysis of trends" section for the period 2002-2100. The drift μ and the volatility σ are directly inferred from the κ_t^i modeled by the NN in Figure 14. As κ_t^i seems to be more unstable before 1970, the empirical variance for σ^{i2} is taken after this moment. The process is given by the following equation [Shumway and Stoffer, 2005]:

$$\kappa_t^i = \kappa_{t-1}^i + \mu^i \times \Delta(t) + \sigma^i \times W(t)$$

Where W(t) is a normal distribution $\sim$ N(0,1)

For the LC and APC models, the package "St MoMo" is used to fit the models on period 1946-2001 and simulate rates from 2002 until 2100 [Villegas et al., 2018]. For the forecasting of the component γ_{t-x} of the APC model in expression (12), an ARIMA(1,1,0) model with 0 mean is implemented.

First, a comparison of the log-mortality rates' moments is made between the NN, the LC model and the APC model for 2010. In parallel, the moments of the historical rates over

the 1946-2010 are calculated to compare theses moments with past data. The expectancy $\mathbf{E}(\ln(\mu_x(t))$, the standard deviation $\mathbf{std}(\ln(\mu_x(t))$, the skewness $\mathbf{S}(\ln(\mu_x(t))$ and the kurtosis $\mathbf{K}(\ln(\mu_x(t))$ of log-mortality rates at 20, 40, 60 and 80 years old are displayed in Table 4.

The skewness is the third moment of a distribution that represents its asymmetric character. If S=0 it is a symmetric distribution, if S>0 the distribution is right-skewed and for S<0 it is a left-skewed density. Depending on the sign of the skewness, the queue of the distribution is extended presenting an asymmetry.
The kurtosis, being the fourth moment, refers to the flatten nature of the density curve. When K=3, the distribution follows a normal law. When K>3 and K<3, the distribution is said respectively leptokurtic and mesokurtic. The lower the kurtosis parameter, the flatter the distribution. [Joanes and Gill, 1998].

Table 5 shows that the 3 models produce similar results in terms of $\mathbf{E}(\ln(\mu_x(t))$. At 20 years old, the NN model exacerbates mortality compared to the LC model. However, at 40, 60 and 80 years old, it is the opposite. At 20 and 40 years old, the NN model is situated between the LC and APC models. These two models produce a mortality respectively more and less optimistic. At 60 and 80 years old, the two benchmark models are slightly equal and the NN model generates smaller mortality rates. As the moments of observed log-mortality rates are taken on period 1946-2010, it is not surprising that there is a huge difference with the simulated data from 2010. The slight difference between the NN model and the other models may be due to the "StMoMo" package which uses the biased instead of the unbiased estimator of $\mu_x(t)$. For instance, in 1946 for male population, the biased and unbiased estimator of $\mu_x(t)$ are respectively equal to 0.00251917 and 0.002515048 and the related $\log(\mu_x(t))$ are -5.983826 and -5.985463. [21].
Concerning the standard deviation and, by extension, the variance, it is a bit different given the model. For the LC model, the standard deviation is decreasing with age except at 80 years old where it increases slowly. The NN model registers the highest variance of the 3 models with an unstable evolution over ages, being more than the double of the variance of the two other models. The APC model displays a peak of standard deviation at 20 years old and a constant one equal to 0.08689399 at other ages. The standard deviation of observed log-mortality rates confirms that there is globally more variance in younger ages and that there is a second peak at older ages. Given these information, the LC model seems to follow in better way the evolution of the standard deviation given the age. The variance in rates for NN model depends a lot on the choice of the sample variance σ^{i2} used for the random walks with drift simulation. This choice can be argued

[21] Cf. relation (7) in section 4.1.1 and relation (10) in section 4.1.2

	7-2-7 NN (DEV) 2010			
Years	20	40	60	80
E(ln (μ_x(t))	-8.497261	-7.387299	-5.546537	-3.371299
std(ln (μ_x(t))	0.2125229	0.2480037	0.1724455	0.2082204
S(ln (μ_x(t))	0.8443526	0.741623	0.5907117	0.6954808
K(ln (μ_x(t))	3.603209	3.372476	3.426111	3.276108
	LC model			
Years	20	40	60	80
E(ln (μ_x(t))	-8.552356	-7.303192	-5.395471	-3.188609
std(ln (μ_x(t))	0.1070532	0.09493897	0.0772386	0.08431916
S(ln (μ_x(t))	0.01407564	0.01407564	0.01407564	0.01407564
K(ln (μ_x(t))	2.92881	2.92881	2.92881	2.92881
	APC model 2010			
Years	20	40	60	80
E(ln (μ_x(t))	-8.455195	-7.432752	-5.391484	-3.114615
std(ln (μ_x(t))	0.1197854	0.08689399	0.08689399	0.08689399
S(ln (μ_x(t))	0.01439141	0.006708721	0.006708721	0.006708721
K(ln (μ_x(t))	3.074388	3.08409	3.08409	3.08409
	Obs. log-mortality rates 1946-2010			
Years	20	40	60	80
E(ln (μ_x(t))	-7.881095	-6.839157	-5.014519	-2.80483
std(ln (μ_x(t))	0.4054658	0.3864469	0.2813184	0.3349214
S(ln (μ_x(t))	0.642908	-0.04227362	0.456794	-0.06682649
K(ln (μ_x(t))	4.168016	2.731596	2.360034	1.681965

Table 5: Moments of log-mortality rates with 10000 simulations in 2010

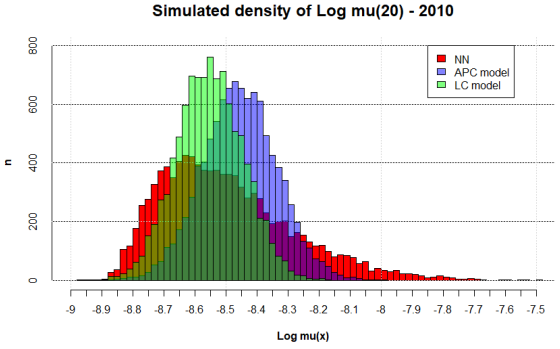
according to the stability of the trends of κ_t^i by defining a representative time period on which to base it.

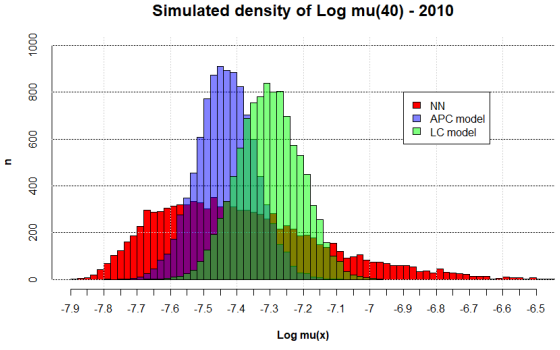
In Professor D. Hainaut paper [Hainaut, 2018], the standard deviation of the NN with MSE as loss function is approaching the one from the cohort LC model implemented on the French population. However, a remark on population size is to precise. In 2010, the French and Swedish population size is respectively equal to 65.03 and 9.341 millions people. In this framework, the models are fit on female population which is even a narrower population of interest counting 4 706 949 women. This could have an effect on variance given that the unbiased estimator of $\widehat{\mu}_x(t)$ is used and depends on the number of deaths and the exposure to risk [22].

[22] cf. equation (10) in section 4.1.2

Regarding the skewness, the NN model displays a highly right-skewed distribution that decreases with the age and soars at 80 years old. The LC model presents a constant skewness over the different ages but its value is nearly null as a normal distribution as the model provides. The APC model has a right asymmetry at 20 years old and a constant one, nearly null as well, for the other ages. The skewness of historical data is more difficult to interpret as it changes its sign at each age.

As for the skewness, the kurtosis of the LC model's rates is constant and approaches the kurtosis of a normal distibution, being slightly mesokurtic. The APC model registers a constant kurtosis over ages expect for 20 years old. The NN's rates kurtosis is leptokurtic and have an unstable evolution. Historical data register a decreasing kurtosis with a peak at 20 years old, moving the distribution from a leptokurtic to a mesokurtic one. The combination of a high kurtosis and a highly right-skewed distribution at 20 years old could be interpret as the accident hump at this age for past data.
To have have a better visualization of the distributions of each model, Figure 17 is plotted hereafter. Theses histograms represent the density of the simulated log-mortality rate in 2010 at 20, 40, 60 and 80 years old.





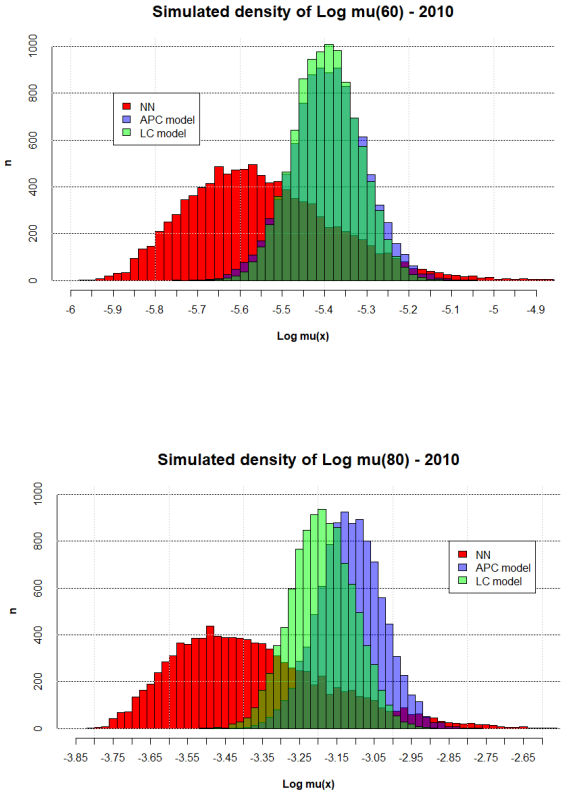


Figure 17: Comparison of the NN, LC and APC model in terms of simulated density log-mortality at 20, 40, 60 and 80 years old - 2010

As one can notice, the benchmark models present the same shape of distribution but react differently given the age. There is a huge difference concerning the spread of the log-mortality rates between the NN model and the two others.

Secondly, the simulated log-mortality rates are used to calculate the average cross-sectional lifetime expectancies. As done before, these are computed at 20, 40, 60 and 80 years old with the expression (8). The Table 5 displays lifetime expectancies for 2002, 2007, 2013 and 2018. The fourth part of this table reports the observed lifetime expectancies extrapolated by the linear regression explained in the "Forecasting" section.

	7-2-7 NN (DEV)			
	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)
2002	67.62645	46.99884	26.95658	9.646426
2007	67.80762	47.2852	27.33564	9.948878
2013	67.9912	47.58003	27.73023	10.28187
2018	68.1206	47.79018	28.01385	10.53416
	LC model			
	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)
2002	68.419	46.60523	25.87252	8.754344
2007	68.93691	47.17348	26.4255	9.06927
2013	69.52778	47.8281	27.06829	9.443993
2018	69.9956	48.35043	27.58587	9.752718
	APC model			
	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)
2002	71.31004	48.82048	25.72209	9.200709
2007	71.70708	49.71852	26.68567	9.482261
2013	72.35903	51.12826	27.52193	9.79657
2018	72.86629	51.21332	29.01503	10.26951
	Obs. lifetime expectancies			
	e_{20}^{Obs}(t)	e_{40}^{Obs}(t)	e_{60}^{Obs}(t)	e_{80}^{Obs}(t)
2002	67.60409	46.95189	26.88284	9.587876
2007	67.79015	47.24645	27.27273	9.895656
2013	67.97855	47.54963	27.67885	10.23514
2018	68.11099	47.76559	27.97081	10.49272

Table 6: Cross-sectional lifetime expectancies produced with 10000 simulations

The cross-sectional lifetime expectancies are very different given the models. On the period 2002-2018, the 20 years old population gains 0.5 year of lifetime expectancy according to the NN model. The LC and APC models predict respectively a gain of 1.6 and 1.55 year for the longevity of the same age of population. This two benchmark models seem to be more realistic than the NN. This observation is not necessarily true for other ages. For the NN model, the gap of lifetime expectancies between 2002 and 2018 is more important with ages. For the LC and the APC models, it is the opposite leading to nearly the same gap for each model at 80 years old. The NN model seems to be not really representative for the youngest group of age.

Table 7 represents the difference between real lifetime expectancies with simulated ones

at 20, 40, 60, and 80 years in 2002, 2007, 2013 and 2018. According to the two benchmark models, the NN fails to predict realistic longevity for young age. Indeed, there is a difference of -4.76 years between the APC and the NN models in 2018. The differences between LC and NN are comparatively less important. This gives indications that NN does not globally capture cohort effect for younger ages. However, this is the opposite for older ages. The NN model is even more optimistic concerning longevity than the APC model. It is relevant to note that simulated lifetime expectancies are always higher than the observed ones for the NN model. This is probably due to the fact that the distribution is a bit right-skewed.

	7-2-7 NN (DEV)			
	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)
2002	-0.02236394	-0.04695	-0.07374	-0.05855
2007	-0.01746965	-0.03875	-0.06291	-0.053222
2013	-0.0127498	-0.0304	-0.05138	-0.04673
2018	-0.009617079	-0.02459	-0.04304	-0.04144
	LC model			
	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)
2002	-0.81491	0.34666	1.01032	0.833532
2007	-1.14676	0.07297	0.84723	0.826386
2013	-1.54923	-0.27847	0.61056	0.791147
2018	-1.88461	-0.58484	0.38494	0.740002
	APC model			
	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)
2002	-3.70595	-1.86859	1.16075	0.387167
2007	-3.91693	-2.47207	0.58706	0.413395
2013	-4.38048	-3.57863	0.15692	0.43857
2018	-4.7553	-3.44773	-1.04422	0.22321

Table 7: Cross-sectional lifetime expectancies produced with 10000 simulations against observed one

Table 8 gives the simulated cross-sectional lifetime expectancies as in Table 6 but for years 2003, 2025, 2050 and 2100. It is with a larger time scale that the distortion of longevity is observable in the NN model and especially for younger ages. Between 2003 and 2100, e_{20}^{NN}(t), e_{20}^{LC}(t) and e_{20}^{APC}(t) respectively rise from 1.21, 6.65 and 6.72 years. For the same period, e_{40}^{NN}(t), e_{40}^{LC}(t) and e_{40}^{APC}(t) respectively increase from 1.97, 7.61 and 8.80 years. The global augmentation of the expected remaining lifetime between 20 and 40 years old is due to the fact that at 40 years old, the accident hump is overpassed leading to

an extension of e_{40}(t). From 2003 until 2100, e_{60}^{NN}(t), e_{60}^{LC}(t), e_{60}^{APC}(t) respectively rise from 2.68, 7.78, 10.82 years. For period 2003-2018, e_{80}^{NN}(t), e_{80}^{LC}(t), e_{80}^{APC}(t) respectively increase from 2.70, 5.18 and 7.82 years.

The evolution of the gap of e_x(t) between 2003 and 2100 is constantly increasing with age for the NN model when it is a bell-shaped for the APC and the LC models. In the paper of Professor D. Hainaut [Hainaut, 2018], the NN model fitted on the French population reacts differently as the NN model on the Swedish women producing more realistic mortality. This could be due to the sophisticated calibration of the model performed using an advanced algorithm to optimize the model. Indeed, on period 2001-2100, e_{20}^{NN}(t), e_{40}^{NN}(t), e_{60}^{NN}(t) and e_{80}^{NN}(t) rise from 8.221, 7.664, 6.716 and 5.1093 for the French population. Figure 30 located in the appendix represent the evolution of e_{20}^{NN}(t) on period 2003-2100. The NN model introduced in this framework seems to be plausible given that e_x(t) are increasing and concave but less realistic given the poor gain of longevity in 97 years. This problem was not perceptible on period 2002-2018 because of the concavity of the function.

	7-2-7 NN (DEV)			
	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)
2003	67.66489	47.05893	27.03541	9.709005
2025	68.27155	48.03791	28.35007	10.85018
2050	68.61422	48.60805	29.13173	11.66975
2100	68.86995	49.03177	29.71718	12.41035
	LC model			
	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)
2003	68.5245	46.72052	25.98437	8.817634
2025	70.61371	49.04656	28.28242	10.17704
2050	72.5154	51.21663	30.49821	11.61149
2100	75.17885	54.33315	33.76552	13.99646
	APC model			
	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)
2003	71.1884	48.46338	25.82959	9.20598
2025	73.52768	51.85002	29.95827	10.93625
2050	75.48874	54.22684	33.18143	14.43365
2100	77.90614	57.26006	36.65433	17.02462

Table 8: Cross-sectional lifetime expectancies produced with 10000 simulations

5.2 Application to several countries

This section is dedicated to the extension of the NN model to several countries. In a first time, the purpose is to compare mortality between countries located in the south and in the North of Europe. The chosen countries are Spain (SP), Portugal (PT) and Italy (IT) for the South and Sweden (SW), Finland (FN) and Norway (NOR) for the North. This choice was made to gather close countries in terms of distance and culture with available data. The analyses are established on female populations. The size of the female population is given per country to visualize the relative importance of these populations. In 2010, the number of women in Sweden, Finland and Norway is respectively equal to 4.71, 2.73 and 2.45 millions. In 2010, the number of women in Spain, Portugal and Italy is respectively equal to 22.56, 5.51 and 30.6 millions [LBM, 2021]. There is a non-negligible difference between the size of the population in the north and the south of Europe.
The results are globally presented in the same way as in the previous section (Validation, goodness of fit and predictive power, analysis of trends, comparison of lifetimes expectancies). In a second time, life annuities are calculated for each countries to observe the differences of this kind of insurance products caused by the mortality of the country.

As given by the expression (17) in section 4.2, the input of the "single country" model is the matrix $\mathbf{M}$. To extend it to several countries, one must stack the 3 $\mathbf{M}$ matrices vertically to obtain a matrix of size $3\mathbf{m}\times\mathbf{n}$. The matrices are concatenated vertically to increase connections between input and output layers with respectively the first and the third hidden layers. The loss function used in this section is the deviance.

5.2.1 Results of validation

In this section, a 4-fold cross-validations[23] is completed to set new hyperparameters. This validation is done on period 1946-2001. The results of the first validation for the number of epochs is **550**. This validation has been applied on the 3 countries from the North. This number of epochs is used for the training the both region. The model registers better results by training with less then 3×360 which is the number of epochs necessary for the training on one country. Concerning the configuration of the NN model, a second cross-validation is established given the error on predictive power. The metric used to evaluate the score of the model is the deviance of poisson. The hereafter Table 8 represents this score in terms of predictive power based on populations of the north:

[23] cf. Figure 6 in section 4.4.1

NN	Predictive Power
3-2-3	13921.01
4-2-4	13722.41
5-2-5	13739.97
6-2-6	14039.55
7-2-7	13865.62
8-2-8	14100.56
3-3-3	13876.38
4-3-4	13957.61
5-3-5	14139.68
6-3-6	13743.53
7-3-7	14114.85
8-3-8	13944.09

Table 9: Validation results for women in north countries

The best configuration retained after cross-validation is 4-2-4. This is the structure used to pursue the analysis for both models.

5.2.2 Goodness of Fit & Predictive Power

The model is trained on the period 1946-2001 and predictions are done on the period 2002-2017. Data was not available in 2018 for all countries. The following Table 9 displays the goodness of fit and the predictive power in terms of deviance for the 3 countries of the North and South together and separately:

Countries	Goodness of Fit	Predictive Power
North	18351.53	18462.19
Sweden	6068.033	4317.068
Finland	8385.063	9368.311
Norway	3898.432	4776.811
South	2118.739	11009.583
Spain	1029.959	5042.735
Portugal	316.774	**2805.158**
Italy	772.006	3161.690

Table 10: Goodness of Fit and Predictive Power - North and south of Europe

The results are very different depending on the country. For instance, the predictive power

of Finland is more than three times higher than the one of the Portugal. Given theses results, one could believe that the model forecasts mortality in a better way for countries from the South. However, this is not necessarily true. The two extremes countries in terms of predictive power results, which are Finland and Portugal, are compared given the mortality curve. The Figure 18 represents plots of log-mortality rates for theses countries in 2017 as follows:

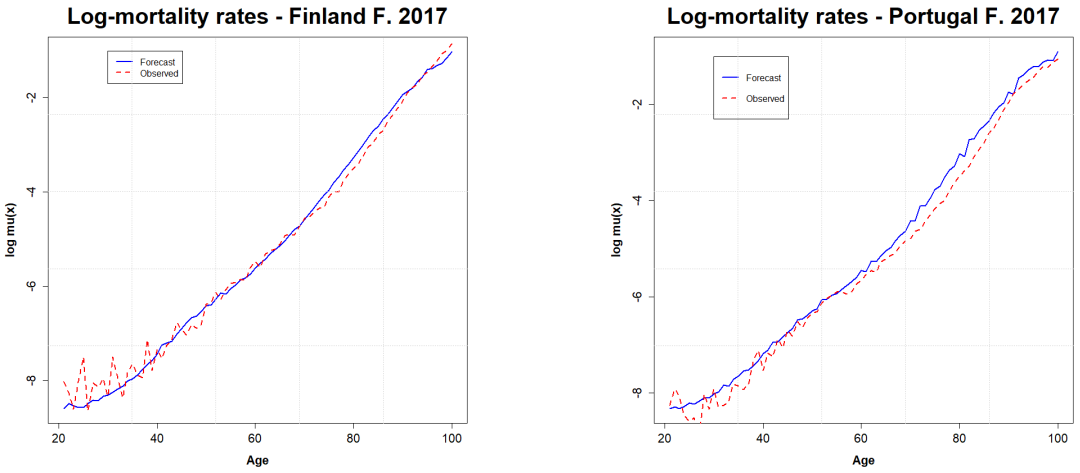


Figure 18: Back-testing on mortality 2017 - Finland/Portugal

For Portugal on the right side of Figure 18, the curve has an unstable shape in particular between 60 and 90 years old. There is also a non-negligible gap between the observed and forecast curve for this country. The forecast curve globally exacerbated mortality for Portugal in 2017. Concerning Finland, its mortality curve is smoother and the gap is smaller. Between 70 and 90 years old, the mortality is exacerbated by the forecast curve. However, the opposite effect is observed between 20 and 40 years old. The Figure 19 reports the log-mortality rates in 2017 of each country per region:

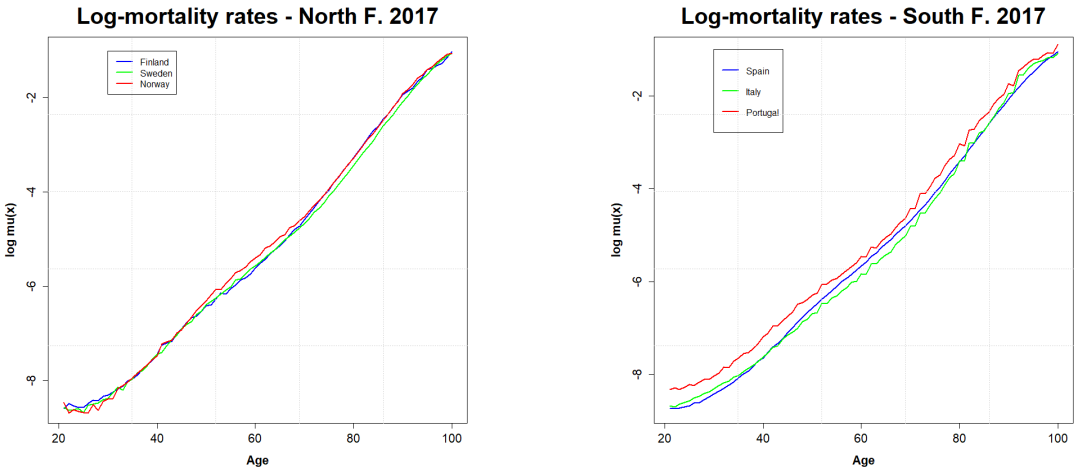


Figure 19: Log-mortality curves 2017 - North/South

As one can observe, some similarities of shape are visible per region. On the South side, the shape of Portugal's curve has identical irregularities as the Italian one. Theses two curves are a bit delayed from each other. The Spanish curve is smoother than the two others. On the North side, the curves are really closed together. As for Portugal and Italy, Finland and Sweden curves present the same irregularities. The curve of Norway is smoother than the two others as well.
It is also important to remark that the score of the predictive power for Sweden is improved by adding the two others countries in the model. Indeed, the deviance passes from 4507.074 to 4317.068. The NN model seems to recognize similarities in mortality of the 3 countries as a sort of common learning effect.

5.2.3 Analysis of trends

In this section, the κ_t^i of the NN model are analyzed. The Figure 20 shows the trends of κ_t^i on period 1946-2001 for both regions. At first sight on Figure 20, all increments of theses κ_t^i do not seem to follow a normal distribution. To verify it, a Jarque Bera test is established in Table 11.

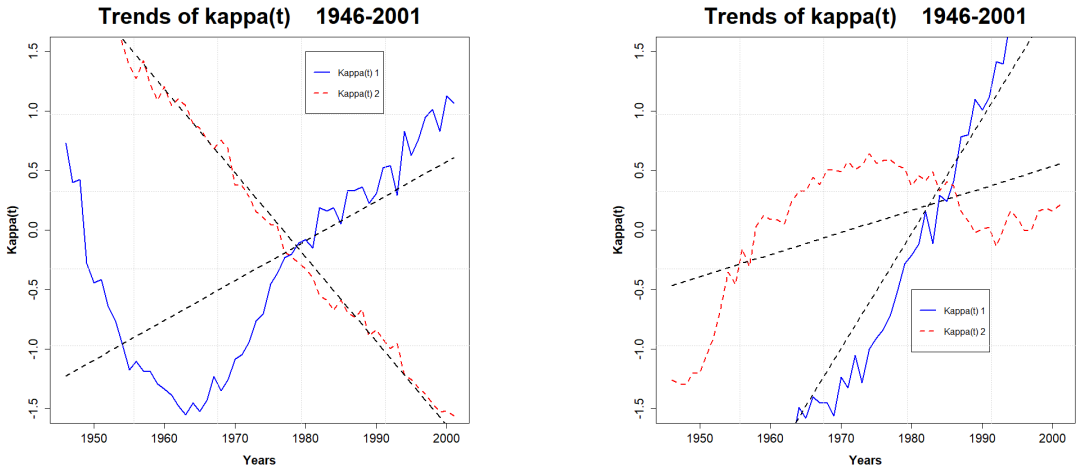


Figure 20: 4-2-4 NN - North/South

NN	Location	Trends	JB statistic	Critical value 5%	P-value	$H_0 : Normal$
4-2-4	North	$\kappa_t^1 - \kappa_{t-1}^1$	19.52	5.004	5.772e-05	Rejected
		$\kappa_t^2 - \kappa_{t-1}^2$	0.57681	5.004	0.7495	Non-rejected
	South	$\kappa_t^1 - \kappa_{t-1}^1$	1.0348	5.004	0.5961	Non-rejected
		$\kappa_t^2 - \kappa_{t-1}^2$	2.5764	5.004	0.2758	Non-rejected

Table 11: Jarque Bera Test

The Jarque Bera test rejects normality for κ_t^i increments of the North of Europe. This result makes it impossible to simulate log mortality rates for northern countries in a reliable way. Therefore, it is difficult to produce reliable simulations for the projection of these data. However, it is surprising that for one country κ_t^i clearly follow random walks and for 3 countries, it is not the case. The others κ_t^i are therefore considered as random walks with a drift given the results of Table 11.

5.2.4 Comparison of lifetime expectancies

This section is devoted to the comparison of lifetime expectancies e_x(t) per country over the period 2003-2100. The Table 16 located in the appendix gives the expected remaining lifetime at 20, 40, 60 and 80 years old forecast for year 2003, 2025, 2050 and 2100. The hereafter Table 12 displays the differences between e_x(t) over period 2003-2100 at 20, 40, 60 and 80 years old for each country.

Sweden				**Spain**			
e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)
1.45	2.27	2.96	2.79	0.84	1.43	2.16	2.18
Finland				**Portugal**			
e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)
1.12	7.61	7.78	2.99	1.62	2.48	3.09	2.28
Norway				**Italy**			
e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)
2.40	3.19	3.61	2.73	1.12	1.97	2.87	2.87

Table 12: Difference of e_x(t) per country over period 2003-2100 - North/South

Concerning e_x(t) evolution over 2003-2100, a difference is remarkable between North and South.

For northern countries, there is a sort of bell-shaped in e_x(t) evolution. It starts low with e_{20}(t) to pass a peak at 60 years old and decreases for the last age group. For example, e_x^{FN}(t) registers a peak of 7.78 at 60 years old which is comparable to e_{60}^{SW}(t) simulated with the LC model in section 5.1.4.

For southern countries, this bell-shaped is mitigated except for Portugal. The evolution of e_x(t) is globally increasing. If Tables 11 is compared with the size of population, one can see that a link may exist with the size of the population. The lower is the population, the more the bell-shaped is present in the evolution of e_x(t) over ages. Indeed, as one can see, Finland and Norway count respectively 2.73 and 2.45 millions of women in 2010 and registers the best amelioration in mortality comparing to the others countries. For Spain an Italy, it is the opposite case with a female population respectively equal to 22.56 and 30.6 millions in 2010.

For younger age, the same observations can be made about the evolution of e_{20}(t) as for Sweden female population in the section 5.1.4. Indeed, the improvement of mortality at 20 years old are very low. For instance, 20 years old Spanish women gain 0.84 in a century according to the NN model which seems to be non realistic.

Globally, the cross-sectional lifetime expectancies are higher for northern than for southern countries. Women in the North of Europe globally seem to live longer than women in the South. This is also true for the differences per country of e_x(t). The countries from the North registers better improvement in mortality than the other region. This disparity may be due to the assumption of the common learning effect mentioned in the section 5.2.2. To compare it, Table 13 is reported hereafter. It represents forecast cross-sectional lifetime expectancies in Sweden with the single country model implemented in section 5.1.4.

	7-2-7 NN (DEV)			
	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)
2003	67.64364	47.0141	26.96477	9.651187
2025	68.26541	48.02017	28.3175	10.81547
2050	68.61432	48.60367	29.12041	11.65389
2100	68.87202	49.03351	29.71758	12.40947

Table 13: Forecast cross-sectional lifetime expectancies with 7-2-7 NN (DEV) - Sweden

Finland and Norway have globally higher improvements of lifetime expectancy compared to Sweden, especially for e_{40}(t) and e_{60}(t). It seems that theses countries pull Sweden upwards in terms of longevity. Indeed, when the models trained only on Swedish Data in section 5.1.4, differences of e_x(t) at 20, 40, 60 and 80 years old were respectively equal to 1.23, 2.02, 2.75 and 2.76 [24]. By integrating Sweden and Norway in the NN model, theses figures pass to 1.45, 2.27, 2.96 and 2.79 according to Table 12. Moreover, there is another common learning effect pulling downwards lifetime expectancies. e_x(t) of Finland and Norway seem to have a negative impact on Sweden lifetime expectancies. By comparing the Sweden part of Table 17 and Table 13, one can see that e_x(t) are less important with other countries integrated in the model.

5.2.5 Insurance Product: Life annuity

This section is dedicated to calculation of life annuities per country and discuss about the eventual differences of pricing. In a first time, it is important to explain the concept of life annuity and the related formula.

A life annuity is the payment of rent at a certain time as long as the client is alive against a unique premium or annual premiums depending on the choice of the insured. This payment can be in arrears or in advance, deferred or immediate, and during the whole or on a defined period. In this framework, the payment are considered done immediately in arrears during a fixed period **n**. This kind of insurance contract is called a Single Premium Immediate Annuity (SPIA) [Dickson et al., 2009]. The price of an annuity is given by the following equation:

$$ {}_{n}a_x = C \sum_{j=1}^{n} \frac{{}_{j}p_x}{(1+i)^j} $$

Where **i** is the guaranteed interest rate, ${}_{j}p_x$ the probability of living at age x[25] and **C** is

[24] theses figures are directly infered from Table 13

[25] cf. section 4.1.1

the payment received by the insured.

Two contracts are established given both a payment of C=1000 € each year. The first annuity is a contract for 60 years old people paying this amount during 20 years, noted ${}_{20}a_{60}$. It is particularly interesting if the policyholder combines it by investing its pension in this kind of contract. The second annuity is a contract for 80 years old people paying C during 10 years, noted ${}_{10}a_{80}$. If an old person is still healthy and does not have money to live 10 years, this sort of contract could suits him/her well.
The hereafter Table 14 displays the price of ${}_{20}a_{60}$ and ${}_{10}a_{80}$ per country for a guaranteed interest rate of 0.00, 0.01, 0.02 and 0.03% in 2017:

	Sweden				**Spain**			
	i=0.00	i=0.01	i=0.02	i=0.03	i=0.00	i=0.01	i=0.02	i=0.03
$\mathbf{{}_{20}a_{60}}$	18449.79	16705.37	15188.78	13865.34	18715.25	16937.66	15392.76	14045.1
$\mathbf{{}_{10}a_{80}}$	7491.032	7135.125	6804.544	6497.073	7394.524	7045.694	6721.575	6420.011
	Finland				**Portugal**			
	i=0.00	i=0.01	i=0.02	i=0.03	i=0.00	i=0.01	i=0.02	i=0.03
$\mathbf{{}_{20}a_{60}}$	18350.76	16622.18	15118.77	13806.32	18183.19	16478.46	14995.09	13699.52
$\mathbf{{}_{10}a_{80}}$	7138.795	6805.79	6496.235	6208.092	6770.653	6462.077	6174.932	5907.373
	Norway				**Italy**			
	i=0.00	i=0.01	i=0.02	i=0.03	i=0.00	i=0.01	i=0.02	i=0.03
$\mathbf{{}_{20}a_{60}}$	18195.34	16485.05	14997.33	13698.4	18504.05	16753.47	15231.58	13903.54
$\mathbf{{}_{10}a_{80}}$	7153.944	6820.407	6510.337	6221.698	7459.215	7105.483	6776.895	6471.253

Table 14: Unique premium price of SPIA in 2017 - North/South

The contracts are calculated with several interest rates to visualize the related impacts. As the financial world is in a prolonged period of low interest rates, the effective guaranteed interest rate on the market should be around 2%.
To evaluate the leverage of mortality of life annuity, the analysis needs to be done by isolating the financial leverage. Therefore, a first analysis is done without annual yield.

In northern countries, for a null guaranteed interest rate, the cheapest and the most expensive $\mathbf{{}_{20}a_{60}}$ contracts are respectively in Norway and Sweden. The most risky contract is considered as the cheapest as possible gains are the highest. In other words, the cheapest contract is the one with the highest probability of dying before its end. Therefore, Norway is the most risky $\mathbf{{}_{20}a_{60}}$ contract of North in terms of mortality and Sweden is the least risky one. There is a difference in $\mathbf{{}_{20}a_{60}}$ contracts of 254.45€ given the mortality of the country.

For a null guaranteed interest rate, the cheapest and the most expensive ${}_{10}\mathbf{a}_{80}$ contracts are respectively in Finland and Sweden. Therefore, Finland is the most risky ${}_{10}\mathbf{a}_{80}$ contract of North in terms of mortality and Sweden is still the least risky one. There is a difference in ${}_{10}\mathbf{a}_{80}$ contracts of 352.24€ given the mortality of the country.

In southern countries, for a null guaranteed interest rate, the cheapest and the most expensive ${}_{20}\mathbf{a}_{60}$ contracts are respectively in Portugal and Spain. Thus, Portugal is the most risky ${}_{20}\mathbf{a}_{60}$ contract of South and Spain is the least risky one. There is a difference in ${}_{20}\mathbf{a}_{60}$ contracts of 532.06€ given the mortality of the country.
For a null guaranteed interest rate, the cheapest and the most expensive ${}_{10}\mathbf{a}_{80}$ contracts are respectively in Portugal and Italy. Thus, Portugal is the most risky ${}_{10}\mathbf{a}_{80}$ contract of South and Sweden is the least risky one. There is a difference in ${}_{10}\mathbf{a}_{80}$ contracts of 688.56€ given the mortality of the country.
If all countries are considered at the same time, the cheapest country for both contracts is Portugal. The most expensive country for ${}_{20}\mathbf{a}_{60}$ and ${}_{10}\mathbf{a}_{80}$ are respectively Spain and Sweden. Theses countries with the related contracts are taken to pursue the analysis.

The hereafter Table 15 displays the difference of price of life annuities contracts produced by a difference of interest rate Δi:

	$\Delta i = 0.01$	$\Delta i = 0.02$	$\Delta i = 0.03$
${}_{20}\mathbf{a}^{P}_{60}$	1704.73	3188.1	4483.67
${}_{20}\mathbf{a}^{SP}_{60}$	1777.59	3322.49	4670.15
${}_{10}\mathbf{a}^{P}_{80}$	308.576	595.721	863.28
${}_{10}\mathbf{a}^{SW}_{80}$	355.907	686.488	993.959

Table 15: Price's differences of life annuities contracts produced by a difference of interest rate Δi

Regarding the guaranteed interest rate, the two cheapest contracts in terms of mortality (${}_{20}\mathbf{a}^{P}_{60}$ and ${}_{10}\mathbf{a}^{P}_{80}$) are the contracts that deliver the lowest gain given the interest rate. The mortality leverage eats the gain of interest and thus reduces the financial leverage. With the same logic, the most expensive contracts (${}_{20}\mathbf{a}^{SP}_{60}$ and ${}_{10}\mathbf{a}^{SW}_{80}$) recover the best financial leverage.

The Table 18 located in the appendix represents prices per country of the SPIA (${}_{20}\mathbf{a}_{60}$ and ${}_{10}\mathbf{a}_{80}$) in 2010 for both regions.
In seven years, ${}_{20}\mathbf{a}^{P}_{60}$ and ${}_{20}\mathbf{a}^{SW}_{60}$ contracts for a guaranteed interest rate equal to 2% pass respectively from 14888.87 and 15104.01 to 14995.09 and 15188.78. In seven years, ${}_{10}\mathbf{a}^{P}_{80}$

and ${}_{\mathbf{10}}\mathbf{a}^{SW}_{\mathbf{80}}$ contracts for a guaranteed interest rate equal to 2% pass respectively from 6037.865 and 6657.128 to 6174.932 and 6804.544.

6 Perspectives & Conclusion

This section is dedicated to the conclusion of this master thesis and the eventual perspectives to go furthurer.
In the present work, a feed-forward neural network is implemented for a dimensionalty reduction purpose applied to mortality. Based on Professor D. Hainaut paper, this model is built suggesting a deviance instead of a mean square error as a loss function. In particular, it is a Poisson deviance that is implemented as the number of death follows a Poisson law. The Poisson model has the structure of log-mortality rates as the Lee Carter model. The dimension-reduction is performed on log-mortality rates with a non linear principal component analysis to model $\kappa_t(i)$ located in the bottleneck of the neural nets with $\kappa_t(i)$ representing the aging component of the Lee Carter model.

A comparison of the two loss function is established on the Swedish population for both gender. The model is fit for the period 1946-2001 and data from 2002 to 2018 represent the test set. Cross-validations are performed to define, in terms of predictive power, the number of epochs and the ideal number of neurons in the hidden layers and especially in the bottleneck. In dimension reduction, there is an ill-posed concerning the definition of the size of the hyperplan. By performing a validation on the configuration of the neural network, this problem is avoided.
The best number of epochs to train the model is set to 360 given the results of cross-validation. The best configuration of neural network using the deviance and the MSE are respectively 7-2-7 and 4-3-4 for female population. Even though MSE loss function registers a better predictive power than the deviance, the log-mortality rate curves predicted by the model have nearly the same shape and display a smooth curve as desired. For male population, the best configuration of neural network using both loss functions is 5-3-5. However, the model underestimates mortality for this gender.

The trends of $\kappa_t(i)$ are analysed to define the type of process in order to simulate projections of theses trends.
For female, the Jarque Bera test confirms for the both loss functions that $\kappa_t(i)$ trends are assimilated to random walks with drift given its normal increments. For male, the Jarque Bera test confirms it only for the deviance. Despite of the poor predictive power with male data, the model using deviance as loss function still produce random walks with drift as $\kappa_t(i)$ trends.
Simulations of log-mortality rates are performed for the neural network in order to compare it to the Lee carter and the Age Period Cohort model from "St MoMo" package. The density for neural network simulated rates is a right-skewed leptokurtic distribution flatter and with a higher spread than the other models. This distribution is oriented by the

choice of the most representative period for the sample variance. A difference may exist as "St MoMo" package models and neural networks use respectively biased and unbiased estimator for log-mortality rates.
The simulated lifetime expectancies at 20, 40, 60 and 80 years old are calculated for each model. From 2002 to 2018, at 20 and 40 years old, the neural nets is closed to the Lee Carter model as it has no cohort effect. But, on the same period, at 60 and 80 years old, the neural nets is closed to the Age Period Cohort model as it has a cohort effect. However, on period 2003-2100, the neural nets is not comparable to LC and APC model anymore. Lifetime expectancies of the neural nets are plausible as it is concave and increasing over time but non realistic given the differences with LC and APC model concerning the gain of lifetime in 97 years.

An extension of the feed-forward neural network to several countries is suggested in this master thesis. A comparison on female population is established between northern and southern countries of Europe. The concerned countries are Sweden Norway and Finland for the north and Spain, Portugal and Italy for the south.
The model is fit on period 1946-2001 and data from 2002 to 2017 constitutes the test set. The results of cross-validations gives a number of epochs equal to 550 and 4-2-4 as the configuration of the neural networks. A "three countries" model trains faster than 3 times a "single country".
The log-mortality curves predicted are different given the region but present some similarities. In each region, the model forecast two same shape curves with irregularities and a smoother one. Concerning the analysis of trends, it is not as conclusive as with a single country. The trends of κ_t^i are not all linear, and thus intrepretable as random walks with drift. The forecast lifetime expectancies at 20, 40, 60 and 80 years old are calculated for each country. According to the neural networks, the north is globally living longer than the south. Over period 2003-2100 model registers better improvements of lifetime expectancies for northern than for southern countries. Small population countries register better enhancements at 40 and 60 years old than big one. Finally, an application of the model to an insurance contract has been presented to show financially the impact of mortality on those populations.

A comparison is made between the results of Sweden in the single country model and in the three countries model. First, the predictive power for Sweden is enhanced by adding Norway and Finland in the model. In a second time, lifetime expectancies are reduced in the three countries model. Finally, longevity of Swedish women on period 2003-2100 is improved. Therefore, a sort of common learning effect appears when 2 others countries are added as the model could recognize some similarities in mortality.

This could be something interesting to investigate in furthurer. One could isolate each country and establish cross analyses on it in order to exploit this track. By adding countries in the neural network, one can also check the limit of stacking countries in the same model. To have better results with the single country model, it could be worth to spend more time on the calibration of the model using a sophisticated algorithm as done in Professor D. Hainaut paper.

7 Appendix

NN	Goodness of Fit		Predictive Power	
	Deviance	MSE	Deviance	MSE
3-2-3	4431.032	4649.865	11091.82	11720.24
4-2-4	4514.585	4471.665	10943.54	11761.90
5-2-5	4101.371	4548.714	**10830.40**	**10883.87**
6-2-6	3942.246	4330.228	10978.72	11079.01
7-2-7	4056.966	4355.759	11506.61	11589.54
8-2-8	3858.249	3985.027	11181.21	11039.25
3-3-3	4519.308	4598.535	11195.41	11113.26
4-3-4	4038.925	4259.911	10790.16	11487.47
5-3-5	3908.430	4278.541	**10778.13**	**10741.15**
6-3-6	3981.121	4142.784	10872.54	11458.32
7-3-7	3842.975	4031.165	10891.28	11866.78
8-3-8	3899.035	3749.499	11302.87	11697.77

Table 16: Validation results - Sweden (Male)

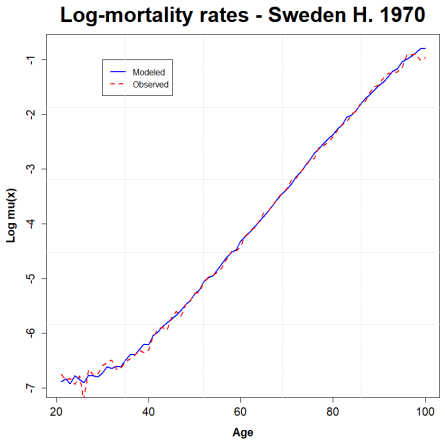


Figure 21: Goodness of fit - MSE (Male)

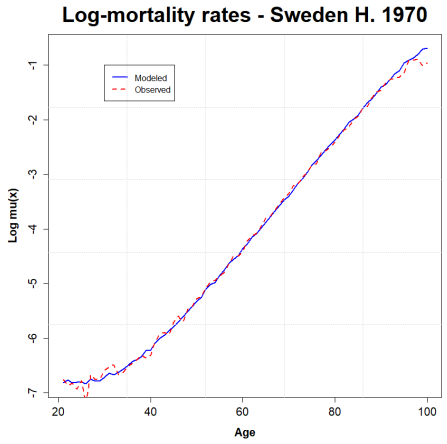


Figure 22: Goodness of fit - DEV (Male)

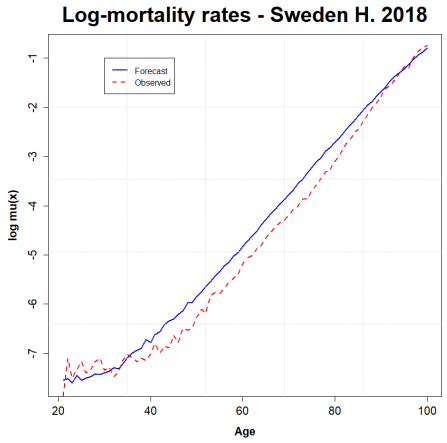


Figure 23: Goodness of fit - MSE (Male)

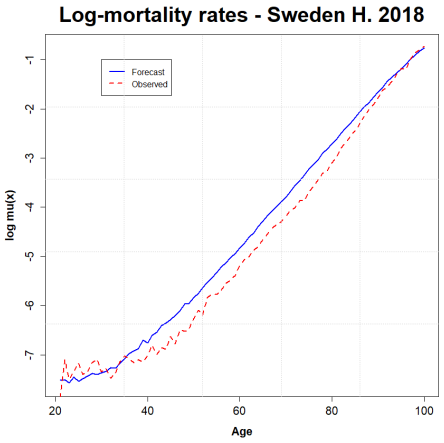
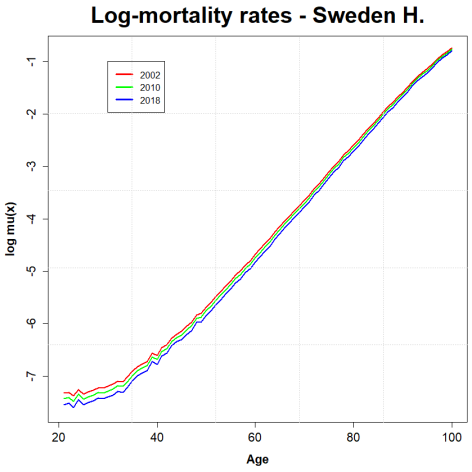


Figure 24: Goodness of fit - DEV (Male)



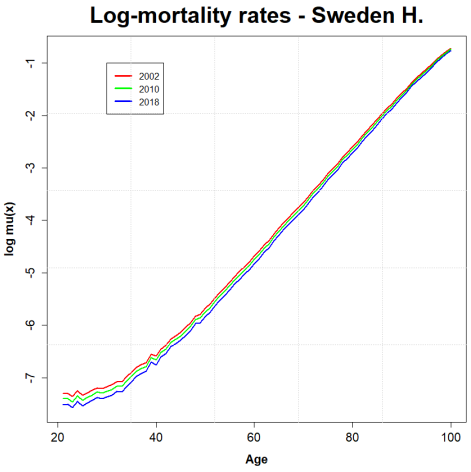


Figure 25: Log μ_x(t) in 2002, 2010 and 2018 - MSE/DEV (Male)

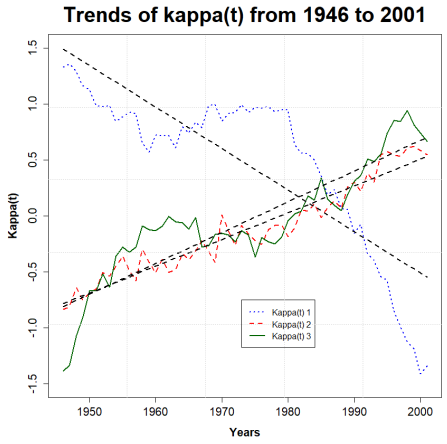


Figure 26: 5-3-5 - MSE (Male)

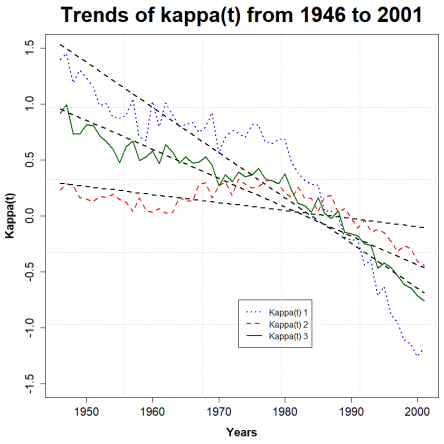


Figure 27: 5-3-5 - DEV (Male)

	Sweden				**Spain**			
	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)
2003	66.49053	45.82048	25.86202	8.932603	66.36162	45.99783	25.98243	8.596938
2025	67.19354	46.90161	27.25262	10.04061	66.7593	46.72068	27.07417	9.463547
2050	67.61003	47.56512	28.12418	10.88953	66.98837	47.10727	27.66727	10.1167
2100	67.94034	48.09211	28.82209	11.72708	67.20277	47.42798	28.14425	10.78142
	Finland				**Portugal**			
	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)
2003	65.76079	46.72052	25.98437	8.22387	64.21122	43.59422	23.76392	7.581181
2025	66.34649	49.04656	28.28242	9.425333	64.94262	44.71806	25.10935	8.342507
2050	66.65785	51.21663	30.49821	10.34511	65.39787	45.43115	26.01358	9.017431
2100	66.88303	54.33315	33.76552	11.2092	65.8301	46.07156	26.85389	9.865712
	Norway				**Italy**			
	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)	e_{20}(t)	e_{40}(t)	e_{60}(t)	e_{80}(t)
2003	65.25572	44.46163	24.67623	8.356974	66.68535	46.03339	25.87569	8.821333
2025	66.19847	45.67309	26.01268	9.211498	67.25558	47.04599	27.30588	9.993468
2050	66.90312	46.62053	27.09002	10.01254	67.567	47.59999	28.13167	10.86509
2100	67.6558	47.65427	28.28692	11.09062	67.80799	48.00422	28.74683	11.68962

Table 17: Forecast cross-sectional lifetime expectancies - North/South

	Sweden				**Spain**			
	i=0.01	i=0.02	i=0.03	i=0.04	i=0.01	i=0.02	i=0.03	i=0.04
$_{20}\mathbf{a}_{60}$	16608.11	15104.01	13791.18	12641.09	16864.13	15328.77	13989.21	12816.16
$_{10}\mathbf{a}_{80}$	6977.651	6657.128	6358.901	6081.054	6892.437	6578.055	6285.445	6012.741
	Finland				**Portugal**			
	i=0.01	i=0.02	i=0.03	i=0.04	i=0.01	i=0.02	i=0.03	i=0.04
$_{20}\mathbf{a}_{60}$	16516.19	15026.48	13725.68	12585.63	16356.37	14888.87	13606.8	12482.59
$_{10}\mathbf{a}_{80}$	6624.154	6326.005	6048.353	5789.452	6316.021	6037.865	5778.584	5536.582
	Norway				**Italy**			
	i=0.01	i=0.02	i=0.03	i=0.04	i=0.01	i=0.02	i=0.03	i=0.04
$_{20}\mathbf{a}_{60}$	16383.1	14908.72	13621.11	12492.46	16654.14	15144.98	13827.78	12673.89
$_{10}\mathbf{a}_{80}$	6684.538	6382.993	6102.197	5840.377	6937.314	6619.449	6323.663	6048.061

Table 18: Unique premium price of SPIA in 2010 - North/South

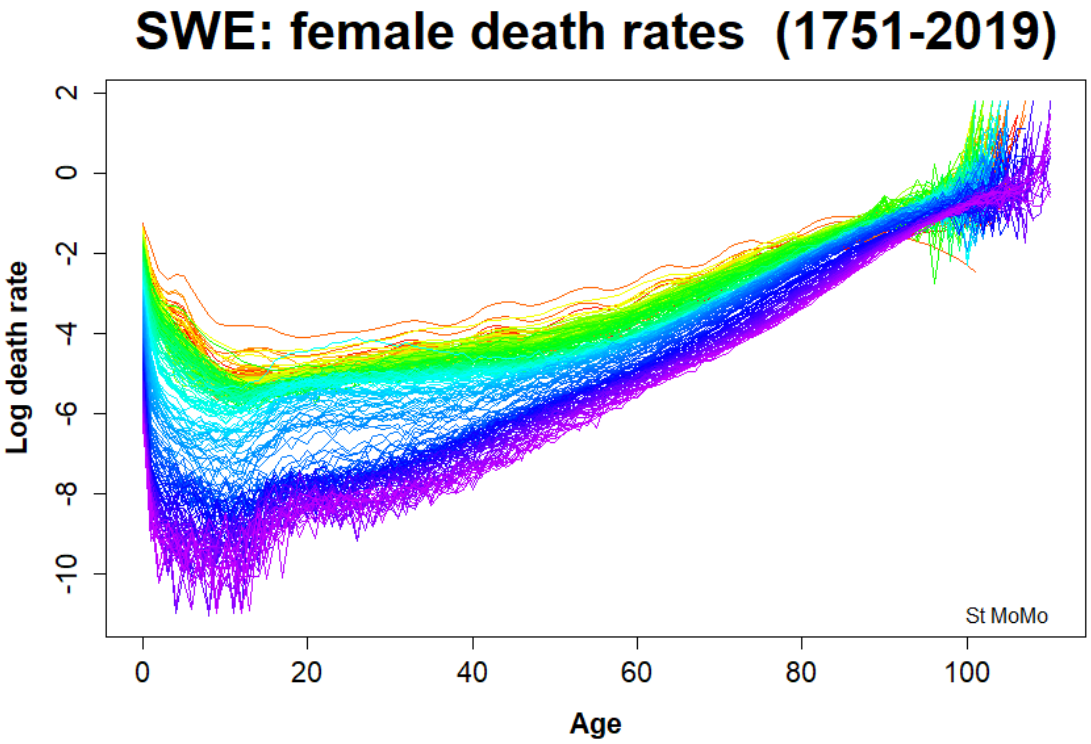


Figure 28: Female log-mortality rates in Sweden 1751-2019

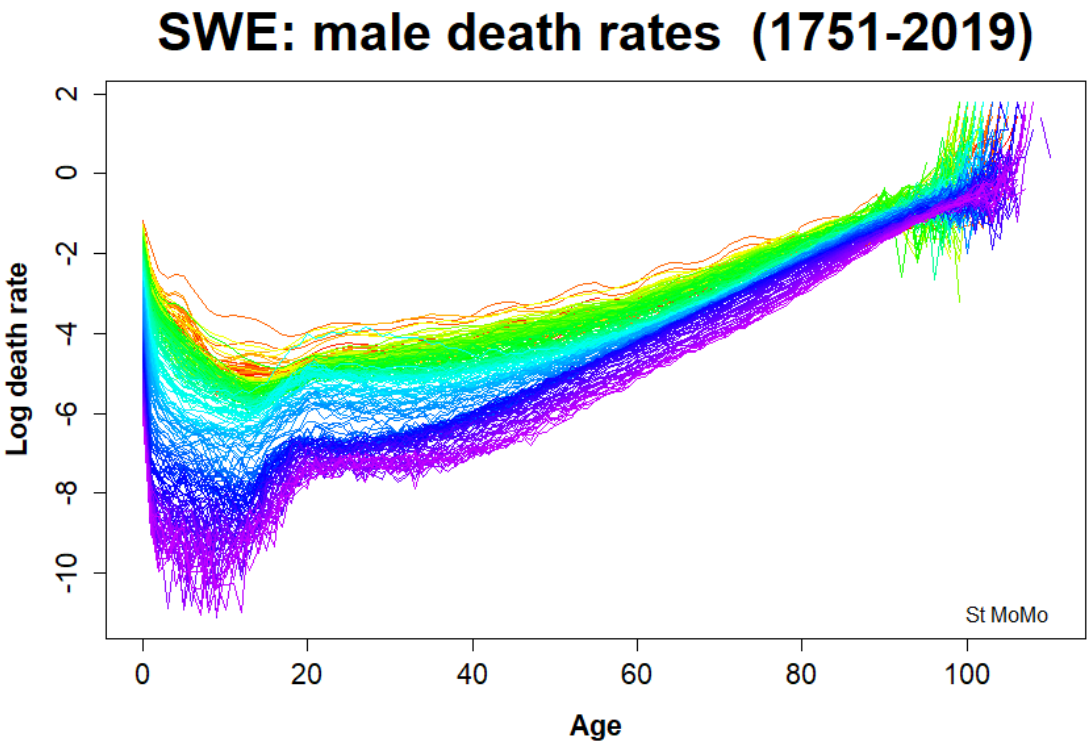


Figure 29: Male log-mortality rates in Sweden 1751-2019

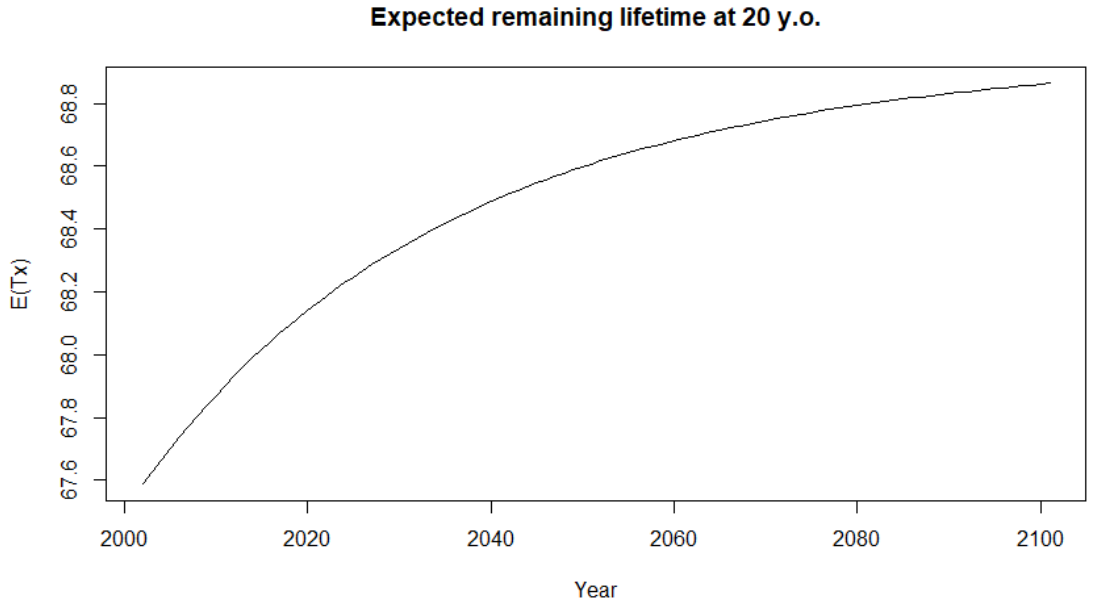


Figure 30: Evolution of the lifetime expectancy at 20 years old on period 2003-2100 - F. Sweden

References

[HMD, 2021] (2021). Human mortality database. Available at `https://mortality.org/`.

[LBM, 2021] (2021). Les données ouvertes de la banque mondiale. Available at `https://donnees.banquemondiale.org/`.

[Abadi et al., 2015] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[Antoine, 2006] Antoine, D. (2006). *Construction de tables de mortalité périodiques et prospectives [Texte imprimé] / Antoine Delwarde, Michel Denuit ; préface de Daniel Serant avec le concours du CERDALM, Centre RD sur l'Assurance de la Longévité et la Mortalité de SCOR Vie.* Assurance Audit Actuariat. Economica, Paris.

[Azencott, 2018] Azencott, C. (2018). *Introduction au Machine Learning.* Dunod.

[Burkov, 2019] Burkov, A. (2019). *The Hundred-Page Machine Learning Book.* Andriy Burkov.

[Chollet et al., 2015] Chollet, F. et al. (2015). Keras.

[Davidson and MacKinnon, 1993] Davidson, R. and MacKinnon, J. (1993). Estimation and inference in econometrics.

[Denuit et al., 2019] Denuit, M. M., Hainaut, D., and Trufin, J. (2019). *Effective statistical learning methods for actuaries. III, Neural networks and extensions / Michel Denuit, Donatien Hainaut, Julien Trufin.* Springer actuarial. Springer, Cham.

[Dickson et al., 2009] Dickson, D. C. M., Hardy, M. R., and Waters, H. R. (2009). *Actuarial Mathematics for Life Contingent Risks.* International Series on Actuarial Science. Cambridge University Press.

[Hainaut, 2018] Hainaut, D. (2018). A neural-network analyzer for mortality forecast. *ASTIN Bulletin*, 48(2):481–508.

[Jarque and Bera, 1987] Jarque, C. and Bera, A. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, 55:163–172.

[Joanes and Gill, 1998] Joanes, D. N. and Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):183–189.

[Kramer, 1991] Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243.

[Lee and Carter, 1992] Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting u. s. mortality. *Journal of the American Statistical Association*, 87(419):659–671.

[Maaten et al., 2009] Maaten, L. V. D., Postma, E., and Herik, J. (2009). Dimensionality reduction: A comparative review.

[Shlens, 2014] Shlens, J. (2014). A tutorial on principal component analysis.

[Shumway and Stoffer, 2005] Shumway, R. H. and Stoffer, D. S. (2005). *Time Series Analysis and Its Applications (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[Smith, 2021] Smith, D. (2021). Actuarial science history (what do actuaries do?). Available at `https://dave4math.com/actuarial-science-history/`.

[Villegas et al., 2018] Villegas, A. M., Kaishev, V. K., and Millossovich, P. (2018). StMoMo: An R package for stochastic mortality modeling. *Journal of Statistical Software*, 84(3):1–38.

[Wackerly et al., 2002] Wackerly, D. D., III, W. M., and Scheaffer, R. L. (2002). *Mathematical Statistics with Applications*. Duxbury Advanced Series, sixth edition edition.

[Weinert, 2007] Weinert, H. L. (2007). Efficient computation for whittaker–henderson smoothing. *Computational Statistics Data Analysis*, 52(2):959–974.

.