CARLOS III UNIVERSITY OF MADRID

Longevity projections: Incorporating sample and population information through the modelization of differences in common sample points

by

Natalia Salazar Vesga

Student number: 100398920

Supervised by

José Miguel Rodríguez Pardo

and

Jesús R. S. Del Potro

A thesis submitted in partial fulfillment for the
Masters in Actuarial Science and Quantitative Finance

in the
Graduate School of Business

July 2020

Abstract

Projecting and understanding longevity has always been a major concern for both demographers and insurance companies. Having reliable projections of the mortality patterns at a country level allows governments to structure their pension schemes and public healthcare policies, and at a sample level, assists companies with the pricing of their life-insurance products as well as with the calculation of their Solvency Capital Requirement *(SCR)*. Understanding mortality at a company level is not an easy task, due to the lack of a large series of data, the most common mortality projection models can not be applied. This is the reason why insurers make use of more pragmatic approaches, generally at the cost of imposing safety margins on their estimations and overestimating death probabilities, which results in a higher cost of solvency capital and lower profitability for the business at hand. This document aims to explore a way in which population mortality models can be adapted to forecast the behavior of the insured population introducing a correction in the forecast of the general population that stems from the modelization of the differences between the two aforementioned groups.

Esta tesis es propiedad del autor. No está permitida la reproducción total o parcial de este documento sin mencionar su fuente. El contenido de este documento es de exclusiva responsabilidad del autor, quien declara que no se ha incurrido en plagio y que la totalidad de referencias a otros autores han sido expresadas en el texto.

En caso de obtener una calificación igual o superior a 9.0 (Sobresaliente), autorizo la publicación de este trabajo en el centro de Documentación de la Fundación Mapfre.

☑ Sí, autorizo a su publicación.

□ No, desestimo su publicación.

Contents

Chapter 1

Introduction

Predicting and studying the evolution of demographic variables, is a problem of paramount importance for both the private and the public sector. Being able to understand the size, status and behavior of populations is a determining factor when formulating both public policies and business plans. The age structure at different points in time is an element of special interest in the insurance and pensions industry. It is important to study not only how population size at each age has behaved in the past, but also how it will behave in the future, to make predictions that allow taking decisions based on future expectations. The most important variables underlying the composition of the population pyramid are the birth and mortality rates at every age. This research document focuses on the latter.

Mortality, being the measure of deaths in a population, serves as the counterbalance to fecundity. To visualize mortality and fecundity within a population, demographers create life tables to display age-specific statistical summaries of a population's survival patterns. Structuring life actuarial products depends on the evolution of said patterns. Mortality evolution is uncertain, this transforms insurance payments and their benefits into random variables that need to be understood, studied and forecasted. The improvement of mortality rates in the past decades represents a benefit for the human species, while simultaneously creating a challenge for the governments and the private sector when designing and maintaining support systems for the elderly, such as healthcare and pension provisions.

To be able to correctly execute the pricing and reserving of actuarial life contracts, in some cases, the insurance industry and the legislation that ensures the insurer's solvency, requires forecasts up to 50 years ahead. Patterns of human mortality so far away in time depend on unknown factors such as the evolution of healthcare systems, the appearance of viruses and pandemics, the occurrence of natural disasters, among others. Taking these elements into consideration when predicting the behavior of mortality is a task of extreme difficulty, as it involves predicting the behavior of a random variable using other random unknown variables as explanatory factors. Hence, the most common techniques for forecasting future mortality, revolve around making use of historical data to extrapolate past trends.

Several methods have been developed for forecasting. Probably one of the most used formulations in the actuarial realm is the two-factor Lee-Carter model (1992). This simple model sets the observed log mortality rate as the dependent variable while reducing the time-dependent component of mortality to a single index that is fitted by ordinary least squares (OLS) and then forecasts it using time series methods. Some other extensions on the Lee-Carter model such as the one proposed by Renshaw and Haberman [2006], take into account cohort effects as well as age-specific period effects. Other approaches, such as the one by Bell and Monsell [1991] use a principal component analysis to include higher-order effects, which results in an improvement of the fit within the sample.

Moreover, Bell [1997] compares the goodness to fit of several methods, among them: fitting parametric curves to age-specific rates and the use of principal components to obtain a linear transformation of the data organized according to a simplified structure. He finds that in terms of forecasts, the most accurate procedure is to apply a simple random walk model with a drift to the rates for each age separately. A very successful extension, which produces highly accurate forecasts, was proposed by Cairns et al. [2006] Also known as the CBD model, their approach aims to describe the future evolution of longevity risk, by including two main factors. Firstly, one that affects equally mortality rates at all ages. Secondly, one that affects at a larger extent mortality at higher ages. Finally, some non-parametric approaches have also been developed. The most notorious work in this matter was developed by Currie et al. [2004]. They use a penalized generalized linear model with Poisson errors to build a regression and penalty matrices, that are later used for smoothing and forecasting two-dimensional mortality tables. This last approach, along with other relevant models (parametric and non-parametric), will be further discussed in the literature review chapter.

As discussed earlier, understanding and forecasting mortality rates is a subject of interest for both the public and the private sector. The public sector has an easier job at hand when it comes to modeling and predicting, as they are able to get robust estimates based on very long and large sets of panel data made available to them by the national statistics agencies. On the other hand, the task is somewhat more challenging for insurance companies as they usually do not possess data sets for more than a couple of years and with a relatively small number of observations at best. The situation is even more challenging if the products are new and there are no data points to characterize the insured population in a reliable manner.

The discrepancies between the behavior of the general and insured population are not a trivial matter. It has been shown in several studies that there are statistically significant differences in the mortality rates of both groups. These arise as a consequence of the industry's extensive underwriting process, which has the tendency of cream skimming risks by insuring individuals that are generally in much better health than the average citizen within the general population. It is also the case, that policyholders in the insured population tend to have a higher socio-economic status which grants access to better health care and living conditions, which results in a longer lifespan. A comparison between the general and the insured population mortality carried out in Mexico by Ornelas and Guillen [2013], showed that *"Members of the insured subpopulation are believed to invest more in prevention, resulting in their presenting lower mortality rates than those of the general population for the same age and gender groups."*

As a result of the significance of the differences between both populations, along with the lack of data at the insurer's level to make a robust forecast, a method of integrating both sources of information needs to be considered. Therefore, once the in-sample general population mortality has been modeled and the indicators for future years have been forecasted, it is necessary to make certain adjustments so that the forecasts based on the general population adjust better to those of the insured population. The challenge is to find a way to include the scarce information about the insured population into the general population forecasts, so that they can be somewhat tailored to the past experience of the company and are more appropriate to make calculations for the contracts at hand.

1.1 Motivation

To guarantee the profitability and solvency of the insurance contracts, companies need to be able to forecast in a reliable way the random variables that give the products they commercialize their stochastic nature. In the case of the life insurance business, a trustworthy prediction of the mortality rates will result in a truthful prediction of the claim sizes, allowing the insurer a better control of their business as well as providing finer tools for profitability analysis. Moreover, the regulators are also interested in the capability of the companies to make use of reliable models so that their solvency is guaranteed at a grater confidence level via the capital requirements.

The need for better forecasting tools is evident, the problem arises when in reality the data that reflects the insurer's experience and helps characterize the insured population is not sufficient to make a robust estimation. As discussed earlier, there are significant differences between the mortality of the general and insured population, so that forecasts based solely on the general public are not adequate. Arising from this problem, the need to integrate the information coming both from the general and the insurer's experience appears as conspicuous. Using the mortality data of the country's statistical agencies is useful for creating a solid estimation for the yearly death rates, for each age group, within the specific geographic area. Then, the data reflecting the insurer's experience can be used to introduce a correction factor that reconciles the future population forecasts with the experience of the industry.

1.2 Objectives

Once the importance of the task at hand has been understood, the main objectives of this work will be discussed. This document aims to study the problem of incorporating sample and population information together and to develop a procedure so that this process can be done in the most accurate manner possible. To do so, four main research questions will be considered:

1. What would be an optimal methodology to produce an in-sample fit of the yearly population mortality rates by age?

2. What methodology produces the most accurate forecast of the yearly population mortality rates by age?

3. What would be a way to incorporate the insurer's experience to improve the forecast of the population mortality?

4. How does the use of corrected mortality estimates compare to the use of the ones calculated with the general population?

After resolving these questions, the goal is to have a good understanding of what the process of incorporating sample and population data requires, and to develop a methodology that is straightforward yet sound so that it is at the disposition of the companies for accurately executing their mortality forecasts.

1.3 Brief Description of the Document

In the second chapter, a literature review will be carried out. It will focus around revising some of the most prominent models for in-sample fitting and forecasting, pointing out their advantages and difficulties from both a theoretical and practical point of view. Later, this section will also explore how the insured population's mortality patterns are forecasted within the insurance industry. It will focus on the goodness to fit of the forecasts as well as the practical implications of said approaches.

Later on, in the third chapter, the data set will be introduced. To be able to revise the accuracy of the model proposed, the mortality estimates will be constructed incorporating the historical data from an insurance company. Some of the main aspects of the databases for both the population and the insurer's sample will be discussed. The focus will mainly revolve around the collection methods and the traceability of the final data. Moreover, this chapter will also discuss preliminary hypotheses and the expected results before performing the numerical analysis.

The fourth chapter will focus mainly on the model and methodology behind the integration of the population and sample data once the best in-sample fitting and later forecasting method have been chosen. Based on the existing literature a model will be formulated, applied and tested. Finally, in the fifth and sixth chapters, the results of the numerical analysis will be presented and the conclusions will be drawn along with the answers to the research questions aforementioned.

1.4 Summarized Findings and Conclusions

In general terms, the findings of this research can be synthesized as follows:

1. For the particular data-set employed to perform the population-level analysis, all the models employed (The Lee-Carter, CBD, and P-splines) result in a good in-sample fit that explains more than 90% of the data variation in all cases when applied individually.
2. For the population data-set at hand, the forecast with the three models and their combinations is also accurate predicting in all cases more than 96% of the data behavior when employing a 10/90 back-testing method on the individual models.
3. Despite the small data-set available at the insured population level, the methodology proposed to correct the population mortality forecast, so that its pattern resembles the one experienced by the insurance company, proves to be satisfactory. The modelization of the differences between the common data points for both samples appears to be a proper tool to generate reliable population-level forecasts.
4. In general, it is difficult to model and forecast the insured population on its own due to the short span these time series usually have. By implementing the model suggested in this document, it is possible to take advantage of the larger extension of the population-level data and pool the accuracy of the forecasts stemming from those longer series to generate a reasonable and reliable forecast at the sample-level.

Chapter 2

Literature Review

As mentioned in the first chapter, both the in-sample fitting and forecasting of the mortality rates, along with the ways of ameliorating population forecasts with information from the insurer's experience, are crucial problems within the life-insurance business. Due to their significance, these topics have appealed to the interest of many researchers and have been amply studied by academics in both the actuarial and demographic disciplines. Earlier, during the introductory section, the state of the art for the in-sample fitting and mortality forecasting has been discussed. In the first section of this chapter, three of the most influential models for this purpose will be discussed. Their benefits and detriments will be analyzed and finally, according to the conclusions drawn from this analysis, the best model for in-sample fitting and forecasting will be selected. Subsequently, in the second section of this chapter, some methods for incorporating together sample and population information will be discussed. It is important to note that this last inquiry, albeit important, has not been sufficiently studied. Here the main existing methodologies used within the insurance industry will be briefly introduced, while the chosen model will be further explained and developed during the fourth chapter.

2.1 Models for in-sample fitting and forecasting

2.1.1 The Lee-Carter Model

The focus of this model revolves around extrapolating mortality trends without incorporating knowledge about medical, behavioral or social influences on mortality change.

Unlike previous methods, the Lee-Carter model does not put a limit to the gains in life expectancy, allowing death rates to decline exponentially. This feature is an advantage, as a more realistic forecast mortality trends can be obtained and there is no need to artificially impose the deceleration of gains in life expectancy.

In their paper, the authors work with annual age-specific death rates for the US population from 1900 to 1987. They forecast and fit mortality for the whole population, without differentiating between sexes, they also work with yearly data grouped in five-year age intervals. When analyzing the quality of the data, they point out that mortality measures at older ages are less reliable. This occurs as a consequence of sample size diminishing substantially for such age groups. Regardless of this issue, the largest gains in life expectancy are expected (from data experience) to occur at older ages. Meaning that there is a need for older age groups to be treated in an especially careful manner.

The main purpose of the authors is to derive a parsimonious model that captures the predominant outlines of the mortality pattern through the variations in time of a single parameter. To have a better understanding of the "raw" mortality change, patterns that diverge from the long-run trends (such as changes in mortality due to historical circumstances) are not captured by the model. The movements of the mortality rate are described by the following equations:

$$\begin{aligned} ln(m(x,t)) &= a_x + b_x k_t + \varepsilon_{x,t} \\ m(x,t) &= e^{a_x + b_x k_t + \varepsilon_{x,t}} \end{aligned} \tag{2.1}$$

Where $m(x,t)$ represents the central mortality rate for age x in year t, k_t is a time-varying index of the level of mortality, b_x is a constant that indicates which rates of mortality decline faster or slower with respect to changes in k_t. In other words:

$$k_t = \frac{\partial ln(m(x,t))}{\partial t} = \frac{b_x \partial k_t}{\partial t} \tag{2.2}$$

In principle, b_x could be negative for some ages (meaning that mortality rises at those ages while decreasing at others). Nonetheless, this is not the case seen in practice over the long run. Moreover, if k_t is a linear function of time, it means that mortality at each age changes at its own constant and exponential rate. As k_t approaches $-\infty$, the death rate goes to zero, meaning that negative death rates can't be an outcome of the model.

The error term, $\varepsilon_{x,t}$ has mean 0 and variance σ^2. This term reflects particular age-specific historical influences that, as mentioned earlier, are not captured by the model. The model allows deriving a one-parameter family of life tables from two observed life tables by expressing death rates as a function of k rather instead of time. Therefore, for any value of k, a set of central death rates is defined, hence it is possible to obtain a life table.

The authors also mention the possibility to work inversely, by finding analytically, the given set of life tables in a family that result in an observed number of deaths, $D(t)$, given a population age distribution, $N(x,t)$. In other words, finding the values of $k(t)$ so that:

$$D(t) = \sum [N(x,t)] e^{a_x + b_x k_t} \tag{2.3}$$

As they are estimations, the original least-squares solutions for $k(t), a_x$ and bx do not result in life tables that imply the exact observed historical death rates observed for the population age distributions. Nonetheless, the differences between the estimates and the real data can be eliminated by keeping the estimated values of a_x and bx and calculating $k(t)$ according to equation 2.3. Moreover, for periods in which the age-specific rates are unknown, despite population age distributions and the total number of deaths being available; the model can be calibrated with the aid of equation 2.3. Hence, in cases where there is a difference between the years of publication of total deaths and age-specific death rates, the forecast can still be done, using for its base year the last period for which total deaths are known.

When forecasting, the method proposed by Lee and Carter does not predict each age-specific rate independently. By doing joint predictions, it avoids calculating $\frac{n(n-1)}{2}$ covariances of errors that are later needed to find the confidence intervals (when n age groups are considered). Instead, the forecasting method makes use of the high level of inter-temporal correlation across ages by making the death rates dependant on a time-changing parameter. The variances and covariances of the death rate functions follow an ARIMA model for $k(t)$. The model for each $k(t)$ is correct if each death rate is well explained by a random walk with drift.

A downside of this methodology is that the forecasts can only fit historical data, so if the realized mortality rates were unlikely when looking at the historical data, the model will not be able to predict them. Another important aspect is that $k(t)$ is fitted as a single parameter, so it resumes all the movements for the individual age-specific rates. Meaning that the forecast would be different if each age-specific rate was modeled independently.

The model fitting finds the minimum least squares solution to equation 2.1, for a given matrix of death rates, $m_{x,t}$. Since the procedure does not yield a single solution, some additional conditions are set:

$$\begin{gathered}\sum_{x=0}^{X} b_x = 1 \\ \sum_{t=0}^{\prime} b_t = 0\end{gathered} \tag{2.4}$$

The previous conditions imply that a_x is an additive constant, equal to the average of the logarithm of the mortality rate, $ln(m_{x,t})$.

Since k_t is unknown, the model can not be fitted using ordinary methods. Therefore, the Singular Value Decomposition (SVD) method is used to find a solution to the minimum least-squares problem. Moreover, the procedure results in fitted values that correspond to the minimized error of the logarithms of the death rates. Afterwards, k_t needs to be re-estimated in a second step using the model equation 2.1 taking now the vectors for a_x and b_x found form the SVD as given. The re-estimation on the second step ensures that each year given a population-age distribution, the implied and actual number of deaths are the same.

The authors call attention to the fact that the data available for death rates are divided by age groups of four years represents a problem. The last age group is 85 and over, which is not useful since the interest of the forecast revolves around higher ages as well. Not taking into account the population distribution for these ages leads to distortions in the predictions. To divide the death rates in this interval up to the age group 105-109, the method suggested by Coale and Guo [1989] is used. They use they assume that mortality rates increase at a linearly decreasing rate with age (Unlike the traditional Gompertz curve that assumes it occurs at a constant rate). Furthermore, the logarithm of the mortality rate ratio $\frac{{}_5m_x}{{}_5m_{x-5}}$ is assumed to decline by a constant increment as x rises above 80. The decline in the increase of mortality from a five-year interval to the next is noted as R and is found based on the following equation:

$$ln\left(\frac{{}_5m_{105}}{{}_5m_{75}}\right) = 6k_{80} - 15R \tag{2.5}$$

Where $k_{80} = \frac{{}_5m_{80}}{{}_5m_{7}5}$, $k_{85} = k_{80} - R$, $k_{90} = k_{80} - 2R$ and so on. To find a solution, an arbitrarily high value is assigned to ${}_5m_{105}$ which ensures that only two percent of the population reaching 105 years of age survive to the age of 110. Based on these solutions, it is also possible to find the mortality rates up to the age interval 105-110.

Regarding the in-sample fit, Lee and Carter find a linear decreasing pattern to describe K_t. The variable declined at a similar pace across all the sample. The short term fluctuations are not that large across the modeled series, so K_t has rather constant variance. Additionally, the fit for most of the age groups for the logarithm of the mortality rate (the mortality rate), is good when compared to the logarithm of actual death rates (the actual death rates), except for ages 20-24. The error for this age segment is low relative to the others, so it doesn't impact the life expectancy forecast a lot. All in all, the fit of the model in-sample is exceptionally good. It is relevant to note that the advantage of forecasting death rates instead of life expectancy directly, is that k_t is linear with respect to time, while life expectancy is not.

Finally, the forecast of the mortality index, k_t, is done using an ARIMA time series model. The authors find (via the Box-Jenkins method) that k_t is better described as a random walk with drift. To decrease the width of the confidence intervals for the forecasted variable, events that increase mortality way beyond its normal values such as the influenza epidemic of 1918 are treated as anomalies utilizing an intervention model that uses a dummy variable for this event, removing its influence.

Forecasts of the 95% confidence interval based on a model based on the whole period and another model fitted with data between 1933-1989 are similar both in expected values and in confidence bands, which somewhat proves the structural homogeneity of the later period. A small investigation on how the base period affects the forecasts shows that fitted models and forecasts display a certain degree of instability when the base period is reduced to 10 or 20 years. Hence, it becomes evident that it is better to use longer base periods to forecast, as they are more stable than shorter ones. In general, when the quality of the forecast is evaluated employing back-testing it proves to be successful.

All in all, the Lee-Carter model is very useful as it provides a very good in-sample fit as well as an adequate forecast when evaluated with the 95% confidence level. It extrapolates mortality trends without artificially imposing the deceleration of gains in life expectancy, making them more realistic. The model does not rely on unknown factors to explain mortality, decreasing the likelihood of omitting variables. Moreover, it does not predict each age-specific rate independently, which allows it to take advantage of the high level of inter-temporal correlation across ages. Finally, by forecasting death rates instead of life expectancy, it exploits the advantage of k_t being is linear with respect to time.

Regarding the disadvantages, it is important to note that mortality measures at older ages are less reliable, as a consequence of a smaller sample size for such age groups. Additionally, only the predominant outlines of the mortality pattern are captured. Therefore, changes that diverge from the long-run trends are not captured. As a consequence, the parsimonious model is not able to predict realizations of the mortality rate that seemed unlikely when looking at the historical data. Furthermore, the quality of the forecast is affected when shorter base periods used for fitting.

2.1.2 The CBD model

Another model to fit and forecast mortality trends is the one proposed by Andrew Cairns, David Blake, and Kevin Dowd 2006, also known as the CBD model due to its authors' initials. Their paper sketches a two factor stochastic model for mortality improvements, also known as longevity risk. The first factor affects mortality rates equally across all ages, while the second factor affects mortality at higher ages more. Larger effects for older individuals make sense, just as Lee-Carter mention, because the historical data shows that longevity improvements are higher at older ages in comparison to younger ones.

To assess the longevity risk, the paper analyzes the pricing of longevity bonds with different terms to maturity referenced to different cohorts. These bonds allow companies to hedge longevity risks. Said bonds have coupons that vary according to a given survivor index. The more individuals form a specific age group die every year, the lower the coupon payment (this scheme makes sense as the company only needs to make payments to those who remain alive). The methodology review here will not focus on the bonds, it will be centered around the mortality fitting and forecasting. Altogether, the authors' main finding is that the longevity risk for shorter time horizons is lower, while for longer periods (more than 10 years) it increases significantly.

Death rates evolve in a stochastic manner, historical data exhibits some unpredictable improvements in mortality, which appear to be more significant at higher ages. Understanding longevity is of crucial importance for policy purposes, social security planning and structuring and underwriting insurance contracts. There is a need to have an accurate estimate of how much people will live so that the contracts and public policies are built based on expectations that are aligned with reality. This allows public systems to be solvent (e.g. pensions) and also for companies not to have to use up their profit margins to cover for unexpected demographic events.

The paper identifies three main types of mortality risk that life insurers and annuity providers are exposed to:

1. **Mortality risk**: Related to uncertainty about the movements (in any direction) of the mortality rates.
2. **Longevity risk**: Related to the long term survival rates being in reality higher than expected.
3. **Short-term, catastrophic mortality risk**: It happens if, over short periods, mortality rates differ significantly from their normal levels (according to experience). It is expected, for these mortality shocks to be transitory so that the rates return to normal after it has passed.

Several approaches have been contemplated for modeling the randomness in aggregate mortality rates through time. Influenced by the Lee-Carter model as well as some other works in the literature, Cairns et al. work in discrete time with annual aggregate mortality rates. Likewise, they use time series models to capture the random elements present in the stochastic development of the mortality rates. As mentioned earlier, the authors introduce a two-factor model. From the data, it is seen that the two factors are successful in modeling historical mortality trends at different ages. Besides, The model allows simulating longevity risk employing cohort survival rates.

The model is defined by the following equation:

$$p(t, T_0, T_1, x) = p[I(T_1) = 1|I(T_0) = 1, \mu_t] \tag{2.6}$$

Where the forward survival probabilities, $p(t, T_0, T_1, x)$, represent the probability at time t that a person of age x at time zero and still alive at T_0, survives until time $T_1 > T_0$. $I(u)$ corresponds to an indicator function equal to 1 if when $t = \mu$ the individual of age x is alive and is equal to 0 if not. Finally, μ_u refers to the evolution of the mortality curve up to $t = u$. The observed period coincides with the interval $(T_0, T_1]$. Therefore, $\forall t \geq T_1$ there is no uncertainty. In other words, when $p(t, T_0, T_1, x) = p(t = T_1, T_0, T_1, x)$.

The model for the mortality curve is outlined by:

$$\tilde{q}(t, x) + 1 - \tilde{p}(t, x) = 1 - p(t + 1, t, t + 1, x) = \frac{e^{A_1(t+1)+A_2(t+1)(x+t)}}{1 + e^{A_1(t+1)+A_2(t+1)(x+t)}} \tag{2.7}$$

Here, $\tilde{p}(t, x)$ is the realized survival probability for the cohort of age x at time 0. By analogy, $\tilde{q}(t, x)$ represents the realized mortality rate. Additionally, $A_1(u)$ and $A_2(u)$ are stochastic processes at time $t = u$ and whose values are estimated from the data applying minimum least squares to equation 2.7. Unlike some other simpler curves, the fit for this parametric model is outstanding, especially at higher ages.

To perform the in-sample fit, the authors use mortality data for England and Wales from 1961 to 2002. After finding the estimated values, they observe that $A_1(u)$ is downward sloping, meaning improvement of mortality over the years. They also see that $A_2(u)$ is upward sloping, which confirms the hypothesis that mortality improvements are greater at higher ages as the curve gets steeper over time. To forecast $A_1(u)$ and $A_2(u)$ ($A(t)$ in general) they model them as two dimensional random walks with drift ass seen un equation 2.8:

$$A(t + 1) = A(t) + \vartheta + CZ(t + 1) \tag{2.8}$$

Where ϑ and C are a constant 2x1 vector and constant 2x2 upper triangular matrix respectively, while $Z(t)$ is a two dimensional standard normal random variable.

A criterion for evaluating a model's fit is to analyze if it is biologically reasonable, meaning that the model is in line with how experts think mortality rates evolve during time[1]. After fitting the data, when analyzing the vector for ϑ, since $\vartheta_1 < 0$ it reflects that mortality rates improve each year, while $\vartheta_2 > 0$ shows how mortality improves at a slower pace for higher ages. Nonetheless, at very high ages (larger than 113) mortality raises over time. This last observation is counter-intuitive to the model (as mortality is not expected to deteriorate), but it is not regarded as a serious issue by the authors. The fact that 113 presents a very unlikely survival age after all and the small sample size at these ages, might induce an estimation error.

Another canon for classifying an estimation as biologically reasonable is that, for a given future year, mortality for older cohorts is higher than that of younger cohorts. In Mathematical terms, for a fixed time $(\bar{t})$:

$$\frac{\partial \tilde{q}(\bar{t}, x)}{\partial x} > 0 \tag{2.9}$$

This implies that $\vartheta_2 > 0$ or that it is negative with very little significance.

To examine the cohort dynamics, a survivor index, $S(t)$ is introduced. Said index is built based on mortality rates for a specific cohort. Investigating it allows studying biological reasonableness. By studying a cohort it is possible to analyze the force of mortality and how it changes over time. To determine the evolution of the survivor index they use Monte Carlo simulation to characterize the evolution process of $A(t)$ using equation 2.8, and then use those results to simulate the realized mortality rates $(\tilde{q}(t, x))$ and the survivor index.

At a first stage, when ignoring parameter uncertainty, by taking estimates of ϑ and $V = CC'$ as the true parameter values they find that: (i) Ex-ante probabilities of survival for 0 to 65 or expected values of the survival curve, also called spot survival probabilities, appear higher when only incorporating data for later periods. This signals for better improvements in the future for $A(t)$. (ii) The percentiles of the survival curve are narrower in the beginning and wider in the end. (iii) The confidence interval for $S(t)$ is narrower when only using the later data. (iv) The variance of the projections is low in earlier years, meaning there is more confidence in the projections for the near future. (v) After the 10^{th} year, the variance grows exponentially, this occurs as there is a compound effect of the yearly mortality shocks over time since each shock affects the

[1] For example, mortality for a given year increases the older the individual, the probability of never dying is zero, among others.

survival rates of the following years. Later, when acknowledging parameter uncertainty via Bayesian methods, by using a non-informative prior distribution for what before was $Z(t) \sim N(\vartheta, V)$, the authors conclude that the further in time, the larger the uncertainty for the $S(t)$ projection due to uncertainty regarding its underlying parameters.

All things considered, the application of the CBD model results in both an adequate in-sample fit and a good forecast. The inclusion of the two factors seems solid when looking at past data allowing to improve the fit. The model also allows the simulation of the distribution of the survivor index over various time horizons and understanding the implications of accounting for parameter uncertainty. Additionally, the results of the estimation, seem to generate a biologically reasonable model. Regarding the downsides, as any parametric estimation, there is the risk of model error, meaning that the model's form is not entirely accurate, biasing the estimation and the forecasts. Likewise, there is the problem of the forecasts being less reliable as t increases. Nonetheless, this is a problem present in most, if not all parametric approaches.

2.1.3 The P-spline smoothing

Currie et al., in their paper *"Smoothing and forecasting mortality rates"*, explore a procedure to smooth and forecast mortality tables via the penalized Splines (P-splines) method. To do so, they use a penalized generalized linear model with Poisson errors. To test the model on real data, the authors use a mortality database for the UK.

2.1.3.1 Generalized linear models

Generalized Linear Models (GLMs) extend the linear model to accommodate dependent variables that do not follow a normal distribution. It is quite common to find in situations in which the dependent variable does not meet the standard hypothesis of the linear model (normal data, constant variance, etc.). There are two main elements present when thinking about generalized linear models: the distribution of the dependent variable, and how the model establishes a relationship between the mean of the said response variable and the other variables that are thought to be explanatory. In a standard GLM model:

$$\begin{aligned} y &= X\beta + \varepsilon \\ \varepsilon &\sim N(0, \sigma^2 I) \\ E(y) &= \mu = X\beta \end{aligned} \tag{2.10}$$

Where $X\beta$is the linear predictor, i.e. a linear combination of the predictor variables represented as η that is related to the mean (in the case of an ordinary regression $\eta = \mu$ so the link function is the identity). Here, y is a random vector that comes from a distribution of the exponential family and whose mean is μ, also known as the random component. The linear predictor $\eta = X\beta$ corresponds is also known as the systematic component. Finally, the link function is a monotonous and derivable function that establishes the relationship between the mean and the linear predictor:

$$\begin{aligned} \eta &= g(\mu) \\ E(y) = \mu &= g^{-1}(\eta) \end{aligned} \tag{2.11}$$

From this model it is possible to distinguish two parts: the probability function of the response variable and the linear structure of the model.

Despite GLM models being somewhat flexible, since they can be adjusted for a large variety of distributions, they assume that there is a linear influence of the explanatory variables on the dependent vector, meaning that:

$$\eta = \beta_0 + \beta_1 x \tag{2.12}$$

Nonetheless, there are many cases where the effect of x is not linear and can have an unknown form given by $f(x)$ that can take different functional forms to improve the fit:

$$\eta = \beta_0 + f(x) \tag{2.13}$$

2.1.3.2 Smoothing methods: Splines

Another option to increase the goodness-to-fit is to use a smoother. This tool represents the trend of the dependent variable and as a function of one or more predictors that are linear in x. This method implies that this new estimation has less variability than the original y, smoothing the result.

All non-parametric regressions have as an advantage that they are based on the data itself to specify the shape of the model. This means that the curve at a given point depends solely on the observations at that point as well as the neighboring ones. Among the non-parametric regression techniques, the Splines and Splines with penalties (P-splines) can be found.

Splines are polynomial piecewise-defined functions. Certain constraints are imposed on each joint, also called nodes, which divide the dominion of the function into regions. The splines are defined by three main elements: its polynomial degree, the number of nodes and the location of the nodes.

Although there are many possible combinations, a popular choice consists of polynomials of third-degree that smoothly join at the nodes (meaning that their first and second derivatives are continuous at these points). A spline of polynomial degree m with k nodes $(C_1, C_2, ..., C_n)$ is defined as follows:

$$y = \sum_{i=0}^{m} \beta_i x^i + \sum_{i=m+1}^{m+k} \beta_i (x - c_{(i-m)})_+^m \tag{2.14}$$

Where:

$$(u)_+ = \begin{cases} u & if : u > 0 \\ 0 & otherwise \end{cases} \tag{2.15}$$

Natural cubic splines are another version of the spline that is linear beyond the limit nodes. For this version of the model fewer parameters need to be estimated, as now the condition for the derivative of the estimation to be continuous at all the nodes is eliminated. The model is now expressed according to equation 2.16:

$$y = \beta_0 + \beta_1 x + \sum_{i=2}^{m+1} \beta_i (x - c_{(i-1)})_+^m \tag{2.16}$$

Usually, $3 \leq n \leq 7$. Moreover, if the sample size is large ($n \geq 100$) and the dependent variable is continuous, $k = 5$ is a good balance between flexibility and precision trade-off. If the sample is small ($n \leq 30$), $k = 3$ is a good starting point. The Akaike Information

Criteria (AIC) can be used to choose k. Finally, a reasonable default placing for the nodes is along the quantiles of the dominion (x).

Another approach for fitting the sample is to use smoothing splines, this procedure consists of minimizing the residual sum of squares (RSS) for the fitted model:

$$RSS(f,\lambda) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_{x_1}^{x_n} f^{''}(x)^2 dx \tag{2.17}$$

The first term of the equation measures the differences between the fitted and real value, while the second term penalizes the curvature of $f(x)$. By penalizing the curvature (i.e. the second derivative) it is ensired that the fit does not have too many abrupt changes in its slope so that the curve has a parsimonious movement. Here, λ is the smoothing parameter and controls the balance between the estimation bias and variance of the fitted curve. If $\lambda = 0$, the curve interpolates the data, and if $\lambda \to \infty$, the second derivative becomes 0, resulting in a linear fit.

Smoothing splines are natural cubic splines where there are as many nodes as unique observations of the dominion, x. The penalty term on equation 2.17 ensures that the coefficients are reduced towards linearity, limiting the number of degrees of freedom used, helping to avoid the overparameterization of the model.

The smoothing splines are linear, meaning that for each unique value of x_i there is a base of functions $h(x_i)$ where:

$$f_\lambda(x) = \sum_{i=1}^{n} h(x_i) y_i \tag{2.18}$$

Taking equations 2.18, the expression 2.17 can be rewritten an solved to find the optimal version of $f(x)$ or θ in this case:

$$\begin{aligned} RSS(\theta,\lambda) &= (y - h\theta)^{'}(y - h\theta) + \lambda\theta^{'}\Omega\theta \\ \hat{\theta} &= (h^{'}h + \lambda\Omega)^{-1}h^{'}y \end{aligned} \tag{2.19}$$

Here, the larger the value of λ, the smaller the estimation coefficients. After the smoothing process is completed, the adjusted model is given by the following expression:

$$\hat{f}(x) = \sum_{j=1}^{h} h_j(x)\hat{\theta}_j \tag{2.20}$$

The problem related to this optimization process is how to determine the correct value for λ, the smoothing parameter, for a particular data-set. This can be done either through cross validation[2] or through generalized cross validation[3].

2.1.3.3 Smoothing methods: P-splines

As mentioned previously, there are two main approaches when smoothing with splines. Firstly, smoothing splines use as many parameters as observations, which makes their implementation not efficient when the number of data is very high. Secondly, regression splines can be adjusted by least-squares once the number of nodes has been selected. Nonetheless, deciding the number of nodes can be a complicated process. Penalty splines solve the difficulties of both approaches: they use fewer parameters than smoothing splines, and selecting correctly the number of nodes is not as crucial as in the regression splines.

The P-Splines model has several advantages: The splines have a low range, meaning that the size of the base is smaller than the dimension of the data; contrary to smoothing splines, where there are as many nodes as data. This allows to work with matrices of smaller dimensions and makes the process more efficient in terms of computation, as for P-Spline the number of nodes can't be larger than 40. Furthermore, the inclusion of penalties, makes the number and location of the nodes less decisive.

For n data points (x_i, y_i) the model to be fitted is defined by:

$$\begin{gathered} y_i = f(x_i) + \varepsilon_i \\ \varepsilon_i \sim N(0, \sigma^2) \end{gathered} \tag{2.21}$$

[2] Where each time a data-point (x_i, y_i) is left out and the value of that point is estimated using the remaining observations. Then a sum of squares is built: $CV(\lambda) = n^{-1}\sum_{i=1}^{n}(y_i - \hat{f}_\lambda^{-i}(x_i))^2$where $\hat{f}_\lambda^{-i}(x_i)$ refers to the fit in x_i when leaving out observation i. Finally the optimal value of the smoothing parameter is found by optimizing $\frac{d(CV)}{d\lambda}$

[3] For an extension on this method see Durban [2003]

Where, $f(x_i)$ is a smooth function that relates each coordinate pair. The objective of the P-splines method is to find $f(x_i)$. The procedure consists on using a base, $B = B(x)$, for the regression and then modify the likelihood function introducing a penalty system based on differences between adjacent coefficients. For normally distributed data, the regression model in matrix notation is defined as:

$$\begin{aligned} y &= Ba + \varepsilon \\ \varepsilon &\sim N(0, \sigma^2 I) \end{aligned} \tag{2.22}$$

To find for the regression coefficients, the penalized minimum square function is solved:

$$S(a, y, \lambda) = (y - Ba)^{'}(y - Ba) + \lambda a^{'} Pa \tag{2.23}$$

Where P is the matrix that penalizes the coefficients and λ is the smoothing parameter. If $\lambda = 0$ the problem is equivalent to regressing y as the dependent variable and B as the matrix of independent factors. It is important to point out that the system of equations that gives solutions to this problem depends on the size of the base and not on the number of observations. The base, B, for the regression can be calculated using B-splines.

A B-spline of grade p consists of $(p+1)$ segments of a p degree polynomial, that meet at p nodes. Moreover, at each node, the derivatives up to the $(p-1)^{th}$ degree are continuous. Besides, the B-spline is positive in the domain extended by $(p+2)$ nodes and equal to 0 for the remaining part. Except for the ends, where the domain overlaps with $2p$ segments of its neighboring polynomials. Finally, for each value of x, there are $p+1$ B-splines that are non-null. Examples of first and second degree B-splines can be seen in figures 2.1 and 2.2 respectively.

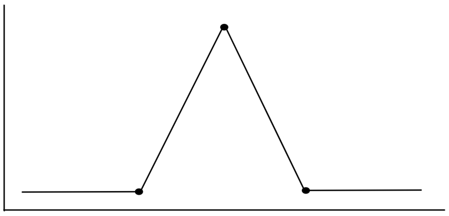


FIGURE 2.1: Example of a first degree B-spline: It is made up of 2 pieces of linear polynomials joined together at three nodes.

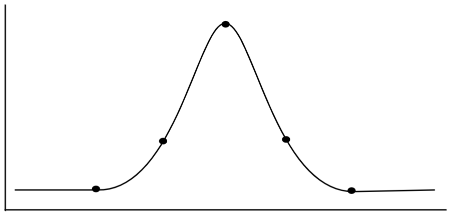


FIGURE 2.2: Example of a third degree B-spline: It is made up of 4 pieces of cubic polynomials joined together at four nodes.

It is important to note that the size of the base has an impact on the fitted curve, the larger the base, the curve is less smooth. When the number of nodes is equal to the number of data points, the curve ends up interpolating the data. To solve this issue, it is possible to introduce a penalty in the second derivative of the curve, once the base and the number of nodes have been selected. This new penalization system transforms the problem in equation 2.23 in:

$$S(a, y, \lambda) = (y - Ba)^{'}(y - Ba) + \lambda \int_x (B^{''}a)^2 dx \tag{2.24}$$

This penalty transformation is common and it is what is used for smoothing splines, nonetheless it is possible to penalize any derivatives of any order.

However, the main aspect of employing P-splines is that the penalty is discrete. Consequently, the coefficients are penalized directly, instead of penalizing the curve and reducing the problem's dimensions. Another advantage of P-splines is that if the curve is polynomial, a P-spline will retrieve it exactly. It is also relevant to mention that the mean and variance of the adjusted values will be the same as the ones of the data regardless of the smoothing parameter value.

Another way to introduce the penalties is to penalize d-order differences between the adjacent coefficients of the bases of B-splines, this is a good approximation to the penalty portrayed in equation 2.24. The differences penalty when added to the least-squares function, leads to the penalized least squares equation: 2.23 in:

$$S(a, y, \lambda) = (y - Ba)^{'}(y - Ba) + \lambda a^{'} P_d a \Rightarrow \hat{a} = \frac{B^{'}y}{B^{'}B + \lambda P_d} \tag{2.25}$$

Where $P_d = (\Delta^d)'\Delta^d$. The penalization can have any degree, for instance if $d = 2$ there is a quadratic penalty on the differences with the adjacent data points complied in a difference upper triangular matrix, D:

$$(a_1 - 2a_2 + a_3)^2 + ... + (a_{k-2} - 2a_{k-1} + a_k)^2 = a^{'}D^{'}Da \tag{2.26}$$

This procedure helps avoiding undersmoothing by penalizing the erratic behavior of $\hat{a_k}$ and forcing the splines coefficients to follow a smooth pattern.

As discussed earlier, the role of λ is to control the smoothness of the curve, but when a penalization on differences is included, this parameter penalizes the coefficients that are widely separated from each other. Therefore, if $\lambda \to \infty$, the coefficients get closer to zero and the fit becomes polynomial. On the contrary, if $\lambda \to 0$ the process is now the same as employing ordinary least squares. Just like other smoothing methods, there are several criteria to select the value of lambda: AIC and BIC among others[4].

All in all, the objective of the P-Splines is, once a base B and a penalty P, have been chosen to find a set of parameters that optimize the penalized likelihood given in equation 2.25.

2.1.3.4 Currie et al. P-Spline extension

As discussed previously, the aim of the authors in this paper is to apply the techniques discussed in subsections 2.1.3.1 to 2.1.3.3 to forecast future mortality by extrapolating past trends. Their model has two main advantages, first of all, that it is non-parametric so unlike its parametric counterparts (such as the Lee-Carter and CBD models) does not assume anything about the functional form of the mortality rates. Another advantage is that the model is built in such a way that the forecast occurs as a natural consequence of the smoothing procedure.

The proposed methodology uses two-dimensional B-splines with penalties regression (known as P-splines) and extends it to generate forecasts. Once the bi-variate P-splines is completed, a family of fitted mortality surfaces is obtained. Afterward, the authors take the future values as missing values, which through the penalization allows estimating future data-points at the same time that the mortality surface is being fitted. In

[4] For an extension on these methods see Durban [2003]

the original P-Splines method, the choice of the penalty function is not as crucial to the outcome. Nonetheless, this issue is now critical as the penalty function will now determine the forecast's form.

The fitting and forecasting are done using data sets of two insurance companies in the UK. For each calendar year and age, the authors have the deaths (claims matrix denoted as Y) and exposure (years lived matrix noted as E). With this data, the matrix of raw hazards $R = \frac{Y}{E}$ is defined and this is the variable that is projected in time.

As mentioned in section 2.1.3.3 the method of P-splines consists of two steps, first to use B-splines as the basis for the regression, and secondly to modify the log-likelihood by a difference penalty on the regression coefficients. This base method is extended by the authors to two dimensions by adapting regression and penalty matrices and making them appropriate for two-dimensional modeling.

To briefly set the problem on section 2.1.3.3 in the notation of the mortality trends context the following variables are defined: For each age i, there is a set of data-points (y_i, e_i, x_i) for $i = (1, ..., n)$, where y_i and e_i corresponds to the total deaths and exposures respectively in year x_i. It is assumed that y_i is described by a Poisson distribution where $E(y_i) = \mu_i = e_i\theta_i$. The aim is to find a smooth estimate $\theta = (\theta_i)$ for the observed forces of mortality $\hat{\theta}_i = \frac{y_i}{e_i}$.

One way of fitting the force of mortality is to employ a polynomial GLM, taking into account the offset of the exposure: $log(\mu) = log(e) + log(\theta) = log(e) + Xa$, such a model uses $1, x, x^2, ..., x^n$ as the basis functions. A more flexible option is to use a set of polynomial B-splines $B_1(x), ..., B_k(x)$, where each spline is a polynomial segment joined at the knots. Now, a matrix B is defined; its rows correspond to the B-splines in the basis for each sample year. The fitted values $B\hat{a}$, correspond to the weighted averages of the coefficient subsets. To avoid under smoothing it is possible to penalize the erratic behavior of $\hat{a_k}$ by introducing a quadratic penalty like the one described in equation 2.26. When the penalty function is included in the log-likelihood equation, the optimization problem is expressed as follows:

$$\ell_p = \ell(a, y) - \frac{1}{2}a^{'}Pa \tag{2.27}$$

Where $\ell(a, y)$ is the usual log likelihood for a GLM, and $P = \lambda D^{'} D$ is the penalty matrix and λ is the smoothing parameter.

The two dimensional regression matrix proposed by Currie et al. is structured as follows: A regression problem has x_1 and x_2 as regressors and an $m \times n$ matrix (Y) containing the data. The matrix Y is indexed by x_1 on its rows (age in this case), and by x_2 on its columns (year for the mortality problem). Let B_a be an $m \times c_a$ matrix of B-splines for smoothing along x_1 and B_y an $n \times c_y$ matrix of B-splines for smoothing along x_2. Hence, the two-dimensional base matrix is defined as its tensor product, more specifically the Kronecker product, of both matrices[5]:

$$B = B_y \otimes B_a \tag{2.28}$$

At this point, the model to be fitted is given by $y = f(x_1, x_2) + \varepsilon$, where y is a vector of length mn. Under this framework[6] $E(Y) = B_a A B_y$, where A is a $c_a \times c_y$ matrix that contains all the regression coefficients, a. The rows and columns of A are given by $A = (a_1, ..., a_{c_y})$ and $A' = (a_1, ..., a_{c_y})$ respectively. Once matrix A is defined, a penalization system for its rows and columns is defined. In this case, a roughness penalty across all the c_y columns of A is applied, and D_a refers to the difference matrix acting on the columns. Analogously, rows are penalizing considering the linear predictor corresponding to each row of Y, and D_y refers to the difference matrix acting on the rows of A. Rows are penalized by means of the expression $a^{'}(I_{c_y} \otimes D_a^{'} D_a)a$ while columns are penalized using $a^{'}(D_y^{'} D_y \otimes I_{c_a})a$. Consequently, the penalty matrix is given by the expression:

$$P = \lambda_a I_{c_y} \otimes D_a^{'} D_a + \lambda_y D_y^{'} D_y \otimes I_{c_a} \tag{2.29}$$

Where λ_a and λ_y are the smoothing parameters for age and year, respectively. The regression coefficients $\hat{a}$ are estimated by maximizing the penalized log likelihood given in equation 2.27, employing B and P as defined in equations 2.28 and 2.29.

Once the two dimensional B-splines base and penalty matrix are defined, the forecasting procedure can be explained. As mentioned earlier, the authors take the future values as missing, which allows performing the in-sample fitting and the forecast jointly.

[5] For an illustrative example of the graphic representation of the bi-dimensional base see appendix A

[6] In regular notation $E(y) = Ba$

For a given age, there is data for y_1 and e_1 available for n_1 years. Moreover, B_1 corresponds to the B-spline regression matrix in a P-spline mortality rates model. To forecast n_2 years, the set of nodes used during the computation of B_1 is extended and the regression matrix B is computed for $n_1 + n_2$ years, therefore:

$$B = \begin{bmatrix} B_1 & 0 \\ B_2 & B_3 \end{bmatrix} \tag{2.30}$$

Moving further, the future values y_2 and e_2 are included into vectors that hold both the historical and the missing data: $y^{'} = (y_1^{'}, y_2^{'})$ and $e^{'} = (e_1^{'}, e_2^{'})$. Now, the regression coefficients are estimated optimizing the log-likelihood function given in equation 2.27 employing the observed data y_1 only[7] From this equation, it is important to note that the penalty function is what allows the forecast to be done. Hence, it is the form of the penalty function what determines the form of the forecast. Moreover, penalizing the elements of a guarantees that the coefficients and the forecasted data-points are smooth.

After maximizing the log-likelihood function and employing the matrix defined in equation 2.30 the penalized likelihood equations are obtained. These equations can be solved through the penalized version of the scoring algorithm:

$$(B^{'}V\tilde{W}B + P)\hat{a} = B^{'}V\tilde{W}B\tilde{a} + B^{'}V(y - \tilde{\mu}) \tag{2.31}$$

Here, B and P are the regression and penalty matrices respectively, $\tilde{W}$ corresponds to the diagonal matrix of weights. Matrix B is defined according to the expression in equation 2.30. While all the elements with a tilde correspond to the current estimates and $\hat{a}$ to the updated estimate of a. Moreover, V is a block diagonal matrix $(I, 0)$[8]. Employing this version of the scoring algorithm, the fitting and forecasting are done synchronously.

After fitting the data for $n_1 = 52$ and forecasting for $n_2 = 50$ the authors find that setting a first-order penalty does not affect majorly the regression coefficients. The authors also point out that penalties of higher-order affect the extrapolated regression coefficients more notoriously, and that these projections correspond with the order of the

[7] Consequently, equation 2.27 changes to: $\ell_p = \ell(a, y_1) - \frac{1}{2}a^{'}Pa$.

[8] Where I is an identity matrix of size n_1 and 0 refers to a square null matrix of size n_2.

penalty. The projections are approximately constant for first order penalties, linear for second-order, or quadratic for third order. This means that mortality rates continue at a constant level, improve at a constant rate or improve at an accelerating (quadratic) rate, respectively. These functional forms in the year direction occur since the age penalty keeps the age structure across age groups. The choice of the penalty degree is done by a trial and error process to find the trend that suits the past behavior of the series best. In their case, Currie et al. choose a second-order penalty that results in a linear extrapolation.

To see how the suggested model behaves compared to the Lee-Carter model, the authors employ back-testing, fitting on an earlier segment of the data, and forecasting on a later segment, to compare the forecasted values with the realized ones. They find that the Lee-Carter method predicts larger falls in mortality than the P-splines method (making the mortality rate curve less parsimonious with the parametric model). Nonetheless, as both models forecast within the 95% confidence interval, they are consistent with each other. Regarding the goodness-to-fit, the P-spline model seems to be better, as it has lower deviance. This lower deviance of the P-splines occurs because the model is local and two-dimensional, which allows the mortality surface to react to changes in the observed mortality that occur locally. Finally, the P-Spline achieves lower deviance than the Lee-Carter approach while utilizing fewer parameters.

All things considered, the benefits of the paper's suggested methodology revolve around its lower computational complexity, as it estimates fewer parameters. The improvement in computational speed becomes more relevant, the larger the data-set used. Besides, the model allows extrapolating missing values in any direction of the data-grid using equation 2.31. Another benefit is that forecasting is a natural consequence of the in-sample fitting and that the model is non-parametric. A downside of this approach is that its correctness relies on the ability to choose correctly the penalty degree, and the fact that there is not a standardized procedure to optimize this parameter's value.

2.1.4 Overall comparison

The discussion in the previous sections has brought attention to the arguments for and against the usage of each model for in-sample fitting and forecasting. There are advantages and disadvantages of each method, nonetheless, there is not a model that appears superior. The differences between the forecasts obtained using each method draw attention to the difficulty of forecasting so far ahead in time. All methods ensure that the projected results are within the 95% confidence interval, and their width reflects the level of uncertainty, that increases the further the forecast is from the last data point. This is observed in all papers, as the confidence intervals get larger for years that are farther away in the future.

As discussed earlier, the first step for incorporating population and sample information is to perform an in-sample fit and to forecast the mortality rate, based on the population information and using stochastic modeling. To reduce model risk, and since no methodology proves to have a superior fit across all the age and time spectrum, the suggestion is to use all three models and compare their performance when averaging the results and also by themselves. It is important to remember that the three models discussed are among the most popular ones and have proven to behave well for different data-sets, countries and year intervals (CF: Lee [2000], Yang et al. [2010], and Goicoa et al. [2019]).

2.2 Methods for incorporating together sample and population information

As mentioned in the previous chapter, the insured population does not behave in the same way as the general population, meaning that the former can not be exclusively modeled by fitting and forecasting the latter. The problem with the data at hand for the insured population is that usually, the time series data available are short, therefore, applying the models discussed earlier for the insured population is not feasible. Short series produce poor forecasts due to a higher estimation error. In general, for short series, every single observation could influence the forecast, therefore the method selected should provide a cautious estimate of errors and possible variability connected to the forecast. In case the sample contains outliers, these are not easily identifiable in periods shorter than 20 years and can't be dismissed easily as doing so will reduce the sample size, even more, aggravating the sample size problems. Due to the limitations of modeling methods for insured data, actuaries within insurance companies have developed more pragmatic approaches to model mortality for the groups at hand.

In most countries, governmental agencies collect information regarding similar products across different insurance companies and publish mortality tables for the industry. Since there is a low number of exposure for younger ages and also for advanced ones, the table is calculated separately for three age tiers: young ages (around 20-33), central ages (around 24-67), and old ages (around 68 and over). Each tier is treated differently, the methodologies and age ranges also varying per country. For the central ages depending on the quantity of information available Generalized additive models (GAM) can be used, this approach consists on an extension of the GLMs where the independent variable is not necessarily a linear function of the dependent variables, but instead, it consists on local regressions weighted by age, which allows using the exact distribution of the dependent variable and avoiding imposing assumptions of normality (which is a usual practice in other parametric methods).

For the central ages, the splines method can also be utilized, and different underlying distributions for the death probabilities can be assumed. For old and young ages it is a usual practice to employ the Coale–Kisker methodology. This is done due to the fact that at said ages the number of exposures is low, causing the volatility of the observed mortality to be high, which makes the usual graduation methods inappropriate. This model assumes that an exponential increase of the central mortality rate at advanced and young ages is not constant and that it increases following a linear pattern.

When the markets for a given country exhibit high volatility for the mortality rates it is common to apply a security loading factor so that the insured mortality table can account for possible deviations from the past experience. There are plenty of functional forms that could be used to define the loading factor, but they usually take into account the variance estimated at each age in the fit and adjust accordingly. It is natural that ages with higher volatility would perceive a higher loading factor than the ones presenting a lower variance. Usually, these types of loading factors aim to place the realized deaths underneath the estimated ones with an almost certain probability (90% or higher).

Additional security loading factors can be set in place to cover for variations that can not be predicted solely from the sample's behavior. Other reasons for loading factors, as suggested by Torres and Mayorga [2017], are:

1. **Confidence intervals**: The mortality table is constructed using information that in some cases can be limited, meaning that it does not include the total behavior

of the sector or population to be modeled. Hence, a safety margin that allows covering the mortality levels of the population of interest should be added. It covers for the limited experience within the industry.

2. **Variation between companies**: The insured mortality table should cover the claims that occur for all companies in the sector. The fact that different companies have different market shares and the number of exposures should be accounted for.

3. **Random fluctuations**: The calculated table is expected to cover most of the random mortality deviations that can be experimented by the different insurance companies. This security loading factor covers information deficiencies of the companies that have a small pool of insured individuals.

4. **Unknown variations**: The mortality table should cover not only the expected events but also the unexpected ones. Tables from a set of data can only represent in an adequate manner the current circumstances or the ones in the very near future (one or two years). Nonetheless, due to practical reasons these tables are not calculated yearly, therefore, this loading factor has a greater relevance depending on the number of years the table would be used for in the future.

Due to solvency requirements and the fact that the main use of the tables is the calculation of life insurance product prices and liabilities, it is important to note that most of the time the statistical agencies and regulating entities will have a tendency to overestimate mortality. This is done to avoid liquidity problems in the future, meaning that the estimated probabilities are not an exact representation of the expected death rates realizations, but instead a modified one that mainly serves regulatory purposes. Overestimating death probabilities diminishes insolvency risks, becoming an interest for the regulating entity, nonetheless, it also implies a higher solvency capital requirement which may represent an additional financial burden for the insurer, situation that is not in line with the company's interests.

Only companies that are large enough are approved by the regulator to build their own mortality tables. Nonetheless, due to the lack of large amounts of data per product, the regulator tends to pool together entire portfolios that belong to different companies, regardless of age, coverage, and characteristics of the products at hand. This pool of risks affects the ability of the companies to use life tables that reflect their own risk structures and do not adapt to their own experience. As pointed out by Lledó et al. [2018], Solvency II allows for insurance companies to use as a *Best Estimate* life table a percentage factor, however, this sometimes proves to be insufficient and inaccurate, given that, although the life insurance products marketed by the different companies are similar, the mortality and composition of each insurance portfolio differ among them.

Another approach to address the insurance mortality difficulties is to model claim sizes instead of death frequencies. This approach reveals a difficulty known as the problem of duplicates, which originates because claims do not correspond precisely to deaths as an individual might hold more than one policy and therefore be the source of multiple claims. This problem ends up causing statistical overdispersion which is by all accounts undesirable. Moreover, to this situation it is added that mortality by lives is heavier than by amounts, this finding is in line with industry experience, but it is not accounted for in most claim-size modeling methods.

As it is shown here, due to the difficulties of modeling the death rates of the insured population, the industry and the legislators have recurred to more pragmatic approaches that do not necessarily represent the reality of each company in an accurate way. The actual procedures promote the regulator's interests at the expense of those of the companies. The research on this matter is pretty limited, despite it being a subject of paramount importance to the insurance industry, In the coming sections a model that tries to overcome these difficulties will be presented.

Chapter 3

Data Description and Preliminary Hypotheses

3.1 Data-sets and sources

This section will focus on the data collection methods and the traceability of the final databases. Here, both the population and the insurer data-sets will be explained. It is important to note that the population data has been obtained from public data sources, which makes it easier to audit its quality. On the other hand, the insurer data was obtained from a private source under a confidentiality agreement, which makes it more difficult to inquire about data verifiability. Nonetheless, a great effort will be made to explore and evaluate both data sets in the most meticulous way possible.

In the following sections for each data set, its sources, reliability, and collection methods, will be appraised. An exploratory data analysis will be conducted as an initial screening of the information and describing it through summary statistics. This first step is useful, as it will help identify patterns in the data that can give light to the formulation of initial hypotheses and expected a priori results. All the data will correspond to the population, either total or insured, from the Netherlands.

3.1.1 Population data

3.1.1.1 Data sources, traceability, and variables

The historical data from the total population has been obtained from the Human Mortality Database (2020). This website is a project created in conjunction with the University of California, Berkeley (USA), and the Max Planck Institute for Demographic Research (Germany). It contains detailed mortality and population data from 41 different countries. For a nation to be included, its death registration and census data must be considered as complete, so that the uniform method can be employed to reconstruct the original data. Consequently, the countries and areas included are relatively wealthy and for the most part highly industrialized. These databases on the website are updated continuously as new data become available for each country.

Particularly, the data available for the Netherlands originally contains period data from 1850 to 2016. For this study the death counts ($D_{x,t}$), the population size on January 1^{st} ($l_{x,t}$) and the population exposed to risk of death on each period ($E_{x,t}$), all given per year and age, are taken into account. To perform the fit for the historical time series data, only the information between 1957 and 2016 is considered. The first part of the sample is left out, as a 60 year series is considered long enough to produce a reliable time series forecast [1]. Moreover, the data collection methods are most likely more precarious the farther away in time, resulting in more variability and dispersion which is avoided by excluding the earliest observations.

Deaths, population estimates, and risk exposures are provided by single years of age up to 109, with an open age interval for 110+. However, these data are sometimes the product of aggregate raw data (e.g., 5-year age groups, open age intervals), which have been split into single years of age through interpolation methods[2]. An advantage of using this data source is that all the actions performed on the data as well as the methodologies employed are very well documented so that the users can rest assured that there is not faulty manipulation in any part of the data preparation process.

[1] Although mortality patterns might change over 50 to 60 years, due to improvements in medicine, science, access to public health care, etc., it is necessary to use an input period that is long enough so that the time series analysis can lead to reliable outcomes. Authors such as McNown and Rogers [1999] and Denton et al. [2005] advocated for this approach.

[2] For more information on said methods, consult the Methods Protocol at `https://www.mortality.org/Public/Docs/MethodsProtocol.pdf`

For the Netherlands, the original and official data on deaths, and population for earlier years have been taken from the country's statistical agency (Centraal Bureau voor de Statistiek). Moreover, data for the most recent years has been obtained from the online database of Statistics Netherlands (CBS StatLine). These numbers collect statistics for the entire territory of the country without differentiating between rural or urban areas.

The calculations done in this paper are based on the data for the total population. Before calculating the total estimates, raw data for women and men are pooled. In other words, death rates and other quantities do not correspond to the average of the separate values for females and males. Meaning that the total values are affected by the relative size of the two sexes at a given age and time. Moreover, it is relevant to point out that the data in its raw form is given in a period format (by the year of occurrence rather than by year of birth).

Just as some of the authors discussed during the literature review point out, at older ages, the number of deaths and the exposure-to-risk end up becoming quite small. This phenomenon causes a considerable random variation in the observed death rates. Because of this, estimations and forecasts at older ages need to be treated carefully, and the faults on this portion of the data need to be acknowledged.

One of the variables employed is exposure-to-risk ($E_{x,t}$), which refers to estimates of the population exposed to the risk of death during a given age-time interval. These estimates are based on annual population estimates calculated at the beginning of the year. Some small corrections are made to reflect the timing of deaths during the interval. In general, the exposure estimations are calculated based on assumptions of uniformity in the distribution of events. The uniformity assumption is usually not that detrimental to the exactitude of the data, nonetheless, in some cases, it can mask the occurrence of historical events that otherwise would have been reflected on the time series (See: Lledó et al. [2020]).

3.1.1.2 Exploratory data analysis

Although the fit is done using data between 1957 and 2016, a larger data set from years 1910 to 2016 is used in this section to portray a more comprehensive picture of the evaluation of mortality across the years. The data-set contains the observed values for death counts, population size at the beginning of the year, and the exposed population. For some years the population size for ages larger than 100 was zero, therefore the data points for ages 101 until 110+ were not taken into account as this led in some cases to divisions by a null coefficient, which resulted in indeterminate results. Ultimately, 107 years and 101 ages (0 to 100) are available for the analysis, leading to a total of $10,807$ observations per variable and $32,421$ data points.

As a first approach, a graphic analysis will be conducted to determine patterns in the data. Figure 3.1 represents the number of deaths with respect to the age of the individuals. Each line describes mortality in a given year, while lines of the same color correspond to the same decade.

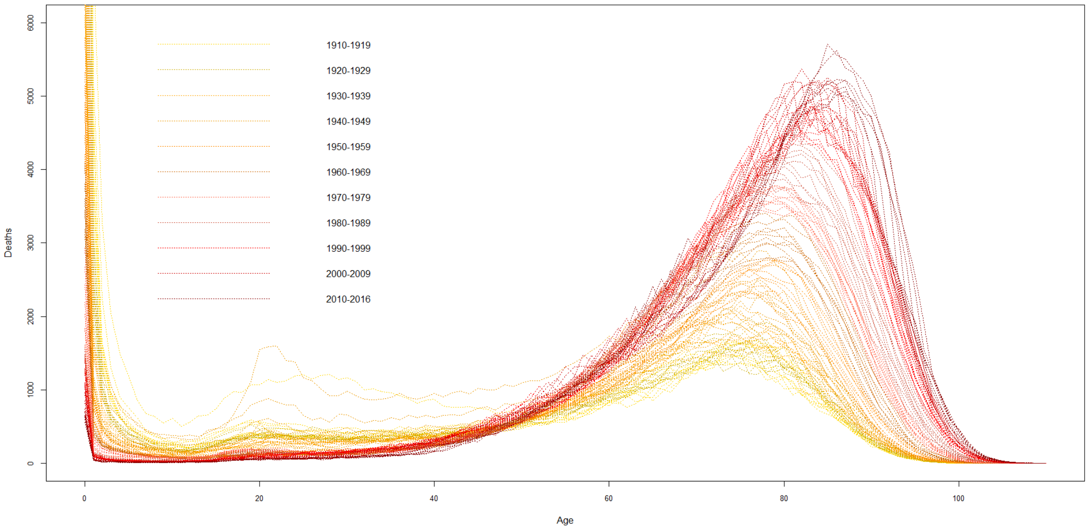


FIGURE 3.1: Total number of deaths per age and year.

Just as it happens with most populations, during the first year of life the death probabilities reach their maximum. This occurs because the first months of children are critical since their poorly developed immune systems must suffice them to survive against the adverse conditions of the external world (environmental, nutritional, medical, etc.). The modal age has gone from 75 to approximately 85 years. In the years after these ages, there is a decrease in the number of deaths. However, it should be kept in mind that this does not mean a decrease in death probabilities, but a decrease in the number of deaths of individuals with 90 years as a proportion of the exposed population. The ratio decreases since there is a diminution on the number of people left alive at said ages.

Furthermore, it is also noticeable that the distance between the mean and the modal ages has diminished.

Two phenomena regarding the dynamics of longevity are observed (common to most developed populations):

1. **The continuous displacement of the number of deaths towards older ages:** The modal age shifts to the right, also showing slight increases over time for the number of deaths at that age. This phenomenon may occur as a consequence of a shift in the average age at which chronic diseases begin.
2. **The compression of mortality:** Meaning the gains in life expectancy are not perpetual, but finite instead.

On the other hand, there are important differences in the number of deaths depending on the decade analyzed. For the ages between 20 and 30 years approximately, a higher level of mortality in comparison to other age ranges, especially for the years between 1940 and 1949.

Another possible way way to represent the data are the raw death probabilities per year and age as follows:

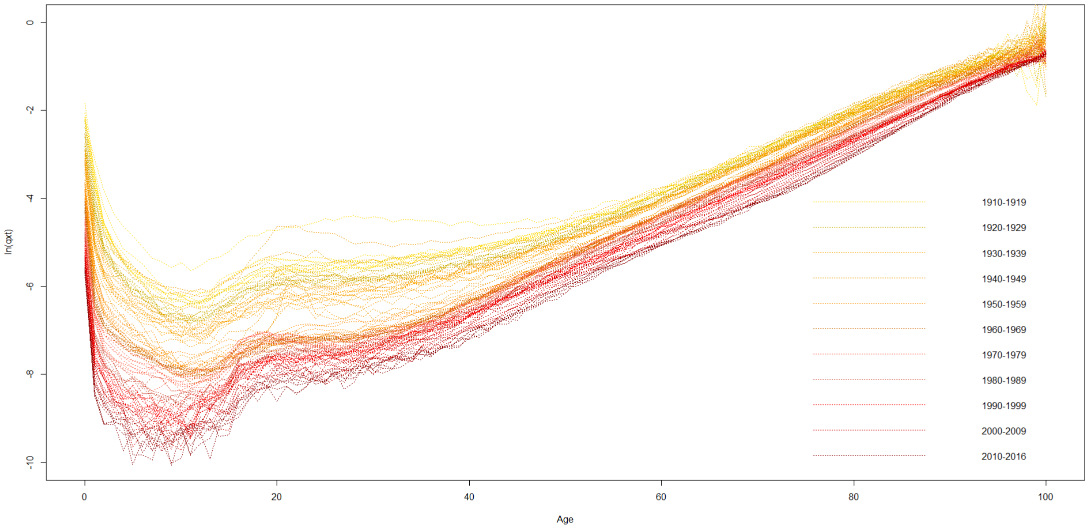


FIGURE 3.2: Death probabilities per age and year (log scale).

It is noticeable that mortality rates despite increasing with age, have been decreasing over the years. There is also a greater variability for the death counts of the oldest and youngest individuals (ages 0-40 and 95-110 respectively). Figure 3.2, represents the data on a new scale, nonetheless, it ratifies the conclusions expressed in the previous paragraphs. In general, after surpassing the first months of life, where death probabilities are high as those of an 80-year-old person, the probability increases progressively along with the individuals' age. However, there is a steeper climb between ages 15 and 20. The highest mortality rate for all ages corresponds to the years of 1944 and 1945, the years of the Dutch Famine, an event that took place in the German-occupied Netherlands, near the end of World War II. A German blockade cut off food and fuel shipments from farm towns, affecting around 4.5 million people and taking the lives of 22,000 individuals of all ages.

The average death probabilities across all years in the sample at age 0, 50, and 100 are 0.03195, 0.00547, and 0.61951 respectively. As expected, the maximum standard deviation is observed at the age of 100 (0.28759 percentage points), while the minimum occurs at age 12 (0.00065 percentage points). During the first year of life, the standard deviation is 0.03490 and it decreases quite substantially for age 1 at 0.00896 percentage points. As a conclusion from this analysis, it is possible to affirm that considering its sources, the transparency in the data treatment methods, and the ability of the data to represent historical events and known demographic patterns, the population data can be taken to be reliable.

3.1.2 Sample data

3.1.2.1 Data sources, traceability, and variables

The sample data has been obtained directly from a Dutch insurance company. It corresponds to a product that is mainly commercialized via bancassurance distribution channels, meaning that as a consequence of an arrangement between a given bank and an insurance company, the latter can sell its products to the bank's client base. The products hereby analyzed are life-risk insurance contracts under the format of Annual Renewable Term (or ART) insurance, a term life insurance that offers a guarantee of future insurability for a given period of years. For the population object of the insurance contract and sum insured discussed here, there are no medical examinations required, meaning that as long as they meet the mortgage requirements from the bank, the access to the insurance product is a priori guaranteed. Moreover, during the stated period, the policyholder will be able to renew the coverage each year without reapplying or passing any filter, such as medical exams, to reaffirm eligibility. The insurance products in this portfolio are mainly linked to mortgages. In the case of mortgage indexation, the insurance covers the risk of death, so that if the insured individual dies, the company will cover the rest of the payments left instead of passing the debt to the deceased's heirs.

The products in the portfolio have been commercialized for 15 years. The available dataset contains information for ages 23 until 94 and for the period between 2007-2016. At earlier ages, the portfolio is substantially small and unstable, as it is not usual for young individuals to get mortgages or annuities. For the interest the analysis the insurer's information will be considered only form age 40 onward, this subset is also taken to match the sample information with the population information analyzed at the country level, which has been described in the previous subsections. To protect the identity of the insured individuals, the data provided has been anonymized and summarized into two matrices, one for exposures ($E_{x,t}$) and another one for deaths ($D_{x,t}$), for both of them each row corresponds to the age, while each column refers to a given year.

As a final note, it is worthy to mention that the information is captured by both the affiliated bank as well as the insurance company from start to finish, which makes the data traceable and trustable, as it is reported by the source and heavily monitored by both parties in the interest of the good development of the business. One last remark is that 2016 is not the last year of data available to the company, this is important because it allows the information that has been reported to be fully updated, so that the Incurred But Not Reported claims (or IBNR) are fully developed and the information truly reflects the occurrences in a given year.

3.1.2.2 Exploratory data analysis

As mentioned earlier, the data base contains a total of 10 years of observations across ages 40 to 94. The portfolio size and composition is explored in the subsequent plots:

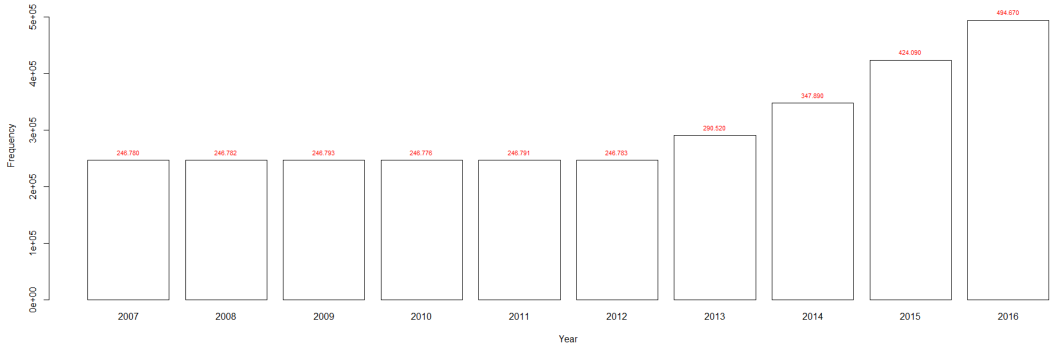


FIGURE 3.3: Portfolio size per year.

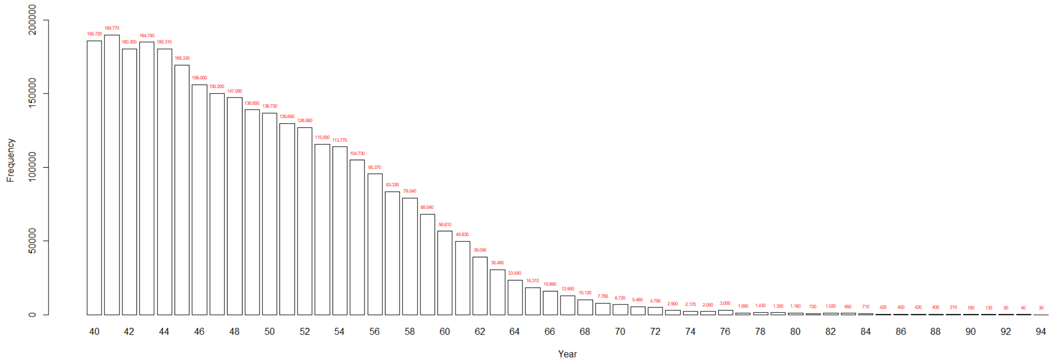


FIGURE 3.4: Portfolio composition by age.

As seen in graph 3.3, the portfolio is rather stable among the first 6 years, with an average of around 246.780 clients per year. The size of the portfolio starts increasing as of the year 2013 and it continues to do so for the next 4 years. On average there is an increase in the portfolio size of 70.580 customers in the last 4 years. The maximum increase is when passing form 2014 to 2015 with a total of 76.200 new insured individuals. Moreover, in graph 3.4 the age composition of the portfolio can be observed. Form this representation, it is seen that the majority of the portfolio is made up of individuals between the ages of 40 to 60. After said age, the amount of individuals per age starts to drop further at each age, until the portfolio ceases entirely at age 95.

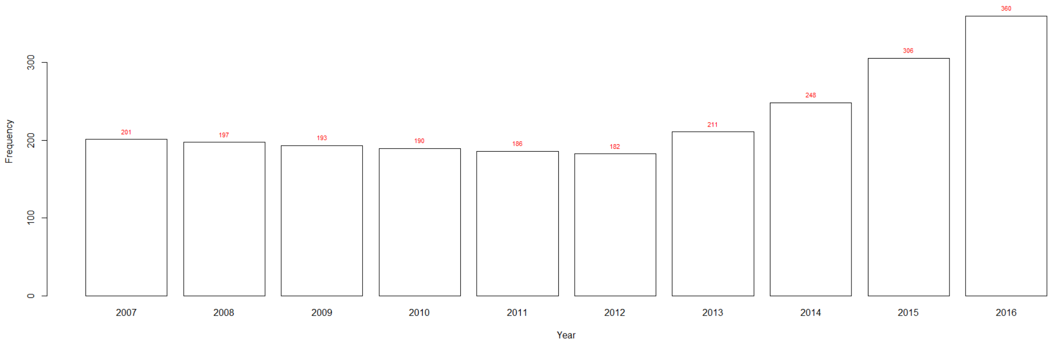


Figure 3.5: Deaths per year.

It is visible in figure 3.5, the number of deaths has a direct relationship with the portfolio size. In the first 6 years, deaths are lower, compared to those in the last 4 years. This occurs as there are more exposures in the later years. Nonetheless, from 2007 until 2012 deaths decrease relative to the rather constant portfolio size. This occurrence could potentially indicate a good risk-selection strategy that improves as the company gets more experience with the product in the market.

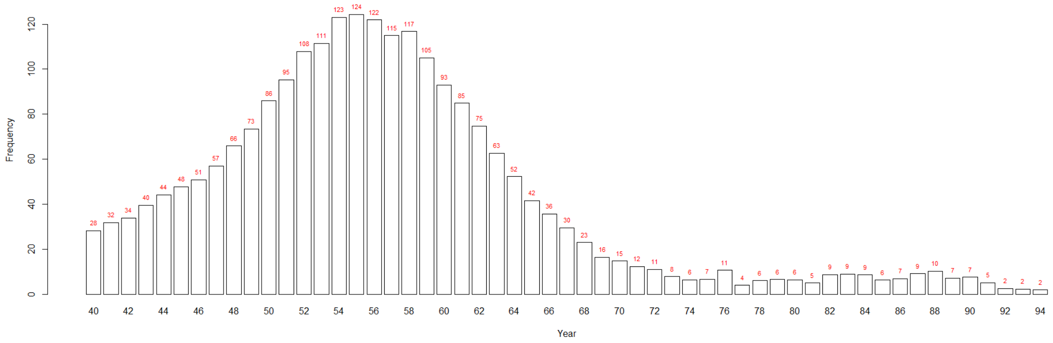


Figure 3.6: Deaths per age.

Figure 3.6 depicts how deaths per age behave following a Gaussian bell. The largest density of deaths occurs from ages 46 to 64. The low amount of claims from ages 70 and onward, occur due to the low level of exposures at these ages, which was earlier shown figure 3.4.

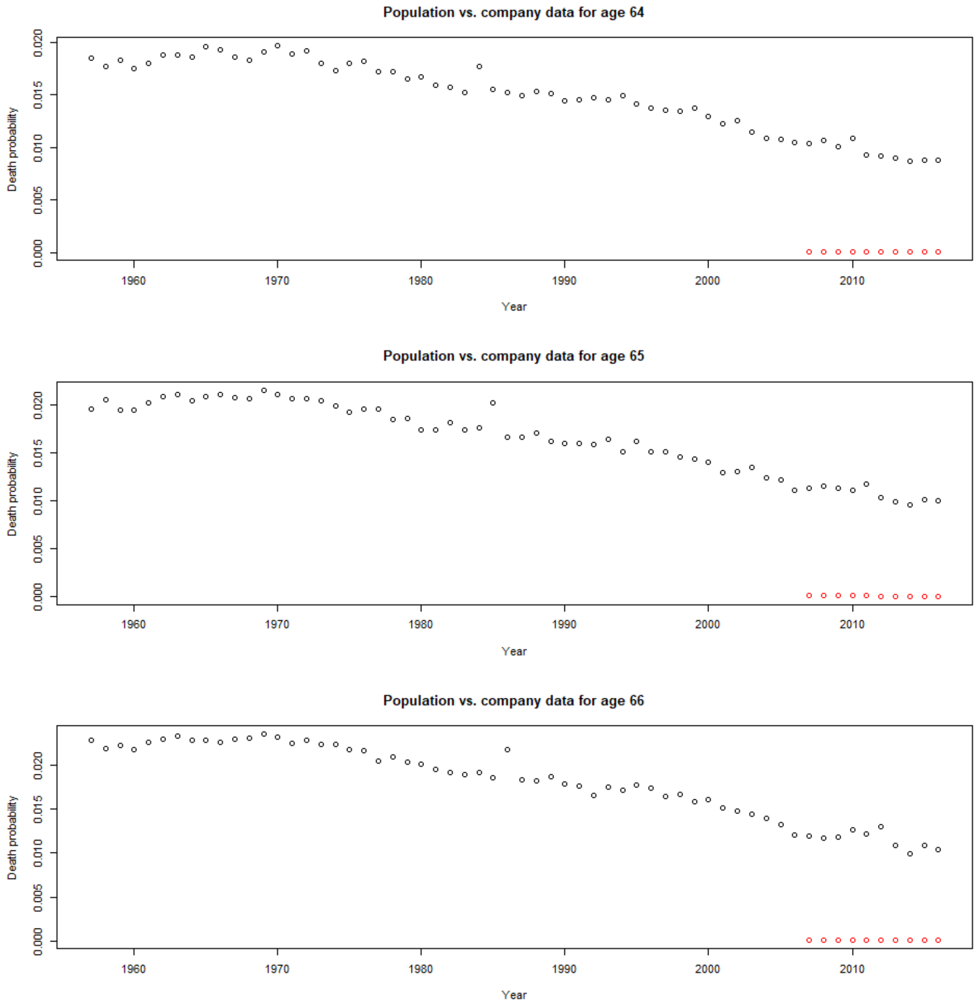


Figure 3.7: Population vs. sample observed death probabilities for ages 64-66

Lastly, figure 3.7 places on the same Cartesian space the population and the sample information, it portrays the observed death probabilities for ages 64 to 65. The population death probabilities per age at every year are shown in black, while the sample probabilities are shown in red. It is evident that the insured individuals are less likely to die at every point in time when compared to the gross population. This occurs as the insured people belong to a particular segment of society that has been chosen out of the total population, in this case via the selection process of the bank. The selection process evaluates if a mortgage should be conceded to a client, it takes into account the earnings of the individual, his credit score, education, among other variables that are positively correlated with a longer life expectancy. Moreover, the sample data has less variance in comparison to the population, this occurs as selected clients are more homogeneous in comparison to the total pool of the country's inhabitants. Finally, this group of figures serves to illustrate the fact that the behavior of the sample and the overall population are different, and that the former can not be explained only by modeling the latter.

3.2 Preliminary hypotheses and expected a priori results

Figure 3.7 already gives a very good impression of what the final results are expected to look like. After integrating the population and the sample data together, the forecast should follow the same trend of the insured data (shown by the red data-points). The final forecast should be able to take the trend stemming from the population data (black points) and correct it so that it resembles that of the insured population. Between the in-sample population models, a priori none of them is expected to be superior. In general most of their results should be similar and their in-sample fit and forecast goodness-to fit measures shouldn't differ much unless unexpected anomalies in the data are experienced.

Chapter 4

Methodology

Mortality tables are usually constructed taking as a base the survivors for each age. The usual practice when calculating a mortality table is to start from an initial population value, $l_{0,t}$ for the biometric variable $l_{x,t}$ which represents the number of individuals alive each year that have reached age x. Assuming death as the sole exiting factor (no migratory movements are possible), $l_{x,t}$ decreases for every age effect of mortality in the group. Said effect is represented by the probability that a person of age x will die before reaching age $x+1$ and is noted as q_x. It is also common to use the variable $m(x,t)$, also known as the raw mortality rate which is defined as the ratio between the number of people who die during the year t at age x and the population exposed to death risk that in year t have reached age x, as follows [1]:

$$m(x,t) = \frac{D_{x,t}}{E_{x,t}} \tag{4.1}$$

Moreover, the death probability ($q(x,t)$) refers to the likelihood that an individual aged x during year t dies in the period within t and $t+1$. In general $q(x,t)$ and $m(x,t)$ are linked by means of the following expression:

$$q(x,t) \approx 1 - e^{-m(x,t)} \tag{4.2}$$

[1] Please note that the notation here does not correspond to actuarial commutation symbols but demographic notation instead.

Finally, the mortality force, denoted as $\mu(t,x)$, refers to the instantaneous rate of mortality at age x at instant t in time.

4.1 In-sample fitting and forecasting

The first step of the process is to smooth and project the death probabilities, $q(x,t)$, using stochastic modeling. To reduce model risk, as none of the models fits optimally for all ages and years, three of the most popular models in this area have been selected and their results will be optimally combined. All three models have been used to smooth q (x,t) for ages 40 to 89. For ages 90 to 95, only the Lee-Carter has been used, since the CBD does not adjust well for advanced ages and the P-splines needs a broader age range so that it can be employed.

When analyzing the different approaches on stochastic mortality models, it is usual to see that some models fit and forecast the force of mortality, $\mu(t,x)$, while some others attempt to explain death probabilities, $q(x,t)$. This discrepancy on dependent variables is not a problem, as equation 4.2 provides an expression for finding the equivalence between both variables.

Depending on what distribution used, the number of deaths, $D_{x,t}$ is given by either a Poisson or a Binomial distribution according to the following statistical hypotheses:

$$D_{x,t} \sim Poisson(\lambda = e^c_{x,t}\mu(x,t)) \tag{4.3}$$

$$D_{x,t} \sim Binomial(n = l_{x,t}, p = q(x,t)) \tag{4.4}$$

Where $\lambda = e^c_{x,t}$ corresponds to the average population in year t and $l_{x,t}$ to the population at the beginning of the year.

4.1.1 The Lee-Carter procedure

Just as described in equation 2.1 the original Lee-Carter model was based on a Poisson distribution. In her paper, *Smoothing constrained generalized linear models with an application to the Lee-Carter model*, Currie [2013] points out that a better fit is obtained if the Binomial distribution is used. Nonetheless, in order to make all models comparable a Poisson distribution will be employed as follows:

$$log(m(x,t)) = \alpha_x + \beta_x k_t + \epsilon \tag{4.5}$$

Where ϵ is the error term. The estimation of the parameters in equation 4.5 is done employing the maximum likelihood procedure. Moreover, the projections are executed assuming an ARIMA process for k_t, while assuming that its future behaviors are described by means of a random walk with drift as follows:

$$\begin{gathered} k_{t+1} = k_t + drift + \varepsilon_{t+1} \\ \varepsilon_{t+1} \sim N(0, \sigma^2) \end{gathered} \tag{4.6}$$

4.1.2 The CBD procedure

The functional form of the two factor model proposed by Cairns et al. can be rewritten as follows when using a Poisson distribution:

$$log(m(x,t)) = k_t^1 + k_t^2(x - \bar{x}) + \epsilon \tag{4.7}$$

Where ϵ is the error term, x refers to the age of the individuals. Moreover, k_t^1 is the factor affects mortality rates equally across all ages, while k_t^2 is a factor that affects mortality at higher ages more. As mentioned in the literature review segment, larger effects for older individuals are reasonable. It can be seen in the data that longevity improvements are higher at older ages in comparison to younger ones. This approach employs a Poisson distribution and the projections for the time-dependent parameters k_t^1 and k_t^2 are done assuming these variables behave according to a random bi-variant walk with drift.

4.1.3 The P-spline procedure

The model proposed by Currie et al., encourage the use of splines with penalties to estimate the mortality force. This technique works as a smoothing method in the context of generalized linear models. The main characteristics of this approach are the use of a base of B-splines for the regression and modifying the likelihood function by penalizing the regression coefficients. The proposed model can be written as:

$$log(\mu(x,t)) = \sum_i \sum_j \theta_{ij} B_{ij}(x,t) \tag{4.8}$$

Where $B_{ij}(x,t)$ is the regression's base, which takes into account both the effect of the age and the year. This double account of effects is achieved by building the base using the Kronecker product. Moreover, θ_{ij} refers to the coefficients that need to be estimated employing the penalized likelihood maximization. The penalty imposed on the coefficients controls the smoothness of the fitted data. It depends on two parameters, one that controls the smoothness across the ages, and another one that controls it across the years. The optimum value of these parameters is selected using the BIC criteria. Predictions are obtained by extending the bases of B-splines (taking the forecasted years as missing values) and readjusting the model.

4.1.4 Overall final result integration

One aspect to take into account when evaluating the quality of the forecast is to estimate the uncertainty of the model's parameters. In the case of the Lee-Carter and CBD model, there is not an explicit and closed expression to estimate the model's parameters. Therefore, the uncertainty linked to these parameters can be quantified using bootstrap techniques.

Particularly for this paper, the semi-parametric bootstrap proposed by Brouhns et al. [2005] has been used. This method consists on generating an amount B of samples for the number of deaths $D^b_{x,t}$ where $b = 1, ..., B$. The sample is generated utilizing a Poisson distribution with mean $\lambda = D^{x,t}$ i.e. the number of deaths registered effectively per age x and year t. Each bootstrapped sample is used to estimate the model again, which results in B estimated parameters. With these estimates, B projected trajectories are generated, while acknowledging for the prediction and the model error. Here, $B = 1000$.

Finally, for the uncertainty linked to the P-splines, it is possible to find explicit expressions for estimating the parameters, therefore their uncertainty related to both adjustment and projection can be calculated by projecting and stressing the death probabilities $q(x,t)$ at 99.5% confidence level.

To evaluate the goodness to fit of the models before combining them and to verify that the average death probabilities estimated by each of the three models are a good approximation to the raw death probabilities observed in the population, the R^2 can be employed as a unit of measure. Here the R^2 is calculated by comparing the logarithms of the raw and estimated death probabilities. The logarithms are used instead of the non-transformed data, because as $q_x \in [0,1]$ calculating the R^2 in the regular scale wouldn't make much sense. The goodness-to-fit measure is defined according to the following expression:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_x \sum_t \left[log(q_{x,t}) - log(q^{\hat{i}}_{x,t})\right]^2}{\sum_x \sum_t \left[log(q_{x,t}) - log(q^{-}_{x,t})\right]^2} \tag{4.9}$$

When combining all models, the R^2 is no longer a good point of reference as knowing which model contributes less or more to the final fit is not that straight forward. Therefore, once it has been established at an individual level that each model is a good fit the comparison among them and the combinations will be done using the Squared Sum of Residuals, which refers only to the part equivalent to RSS in equation 4.9 as follows:

$$RSS = \sum_x \sum_t \left[log(q_{x,t}) - log(q^{\hat{i}}_{x,t})\right]^2 \tag{4.10}$$

To have a measure of the contribution of each model to the in-sample fitting, all the combinations of the chosen models will be taken into account for the fit between ages 40-89. As mentioned earlier, only the Lee carter is suitable for ages 90-99. All in all $q^{\hat{}}_{x,t}$ can take an array of values given by the model estimations individually or by the model combinations follows:

$$q^{\hat{}}_{x,t} = \begin{cases} \frac{q^{LC}_{x,t} + q^{CBD}_{x,t} + q^{P-spline}_{x,t}}{3} \\ \frac{q^{CBD}_{x,t} + q^{P-spline}_{x,t}}{2} \\ \frac{q^{LC}_{x,t} + q^{P-spline}_{x,t}}{2} \\ \frac{q^{LC}_{x,t} + q^{CBD}_{x,t}}{2} \end{cases} \tag{4.11}$$

Finally, to test the accuracy of the forecast, a 10/90 back-testing method will be employed using the same combinations described in equation 4.10 to compare the projected and actual values. In this case, the R^2 of the forecast will be calculated only when the models are evaluated individually, and the RSS for all cases to enable comparison.

4.2 Model for the integration of the population and sample data

As mentioned earlier, there is not much literature around mortality projection for short time series, such as the ones that are usually available at the company level; let alone on how to incorporate sample information to population mortality projections. The model presented here, formally and fully develops an idea initially suggested by the Heriot-Watt University professor Iain Currie, in her post on the Longevitas information matrix blog: *"Forecasting with limited portfolio data"*. This procedure, called the Piggy-Back model, aims to make corrections on the population forecast by modeling the differences between the population and the insured data at each age and year. The gap is modeled via a generalized linear model regression (GLM) and then added to the original population-level forecast, to obtain a forecast that approaches that of the insured portfolio.

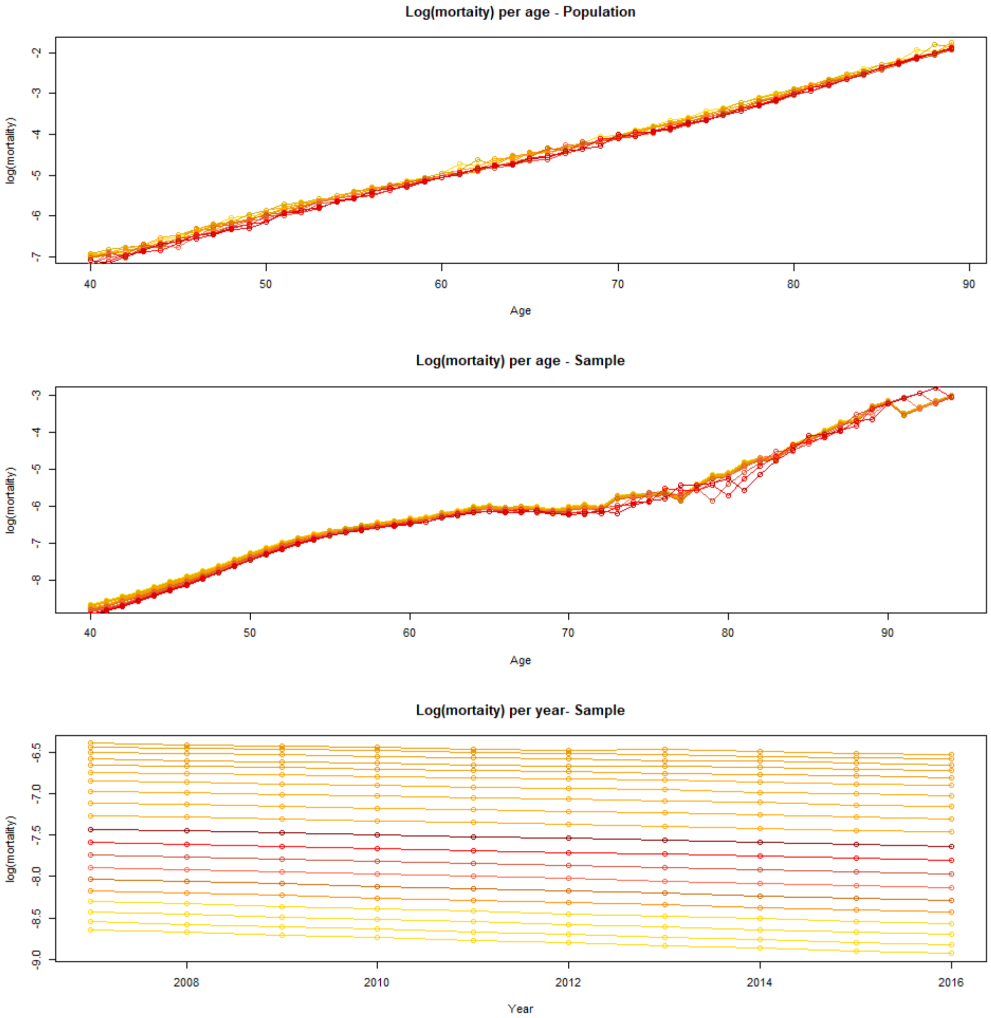


Figure 4.1: Population and sample mortality behavior

When observing figures compiled in 4.1, it becomes evident that the mortality level for both the population and sample data behave linearly with respect to age. Hence, it could be reasonable to deduct that the gap between both data-sets it behaves approximately linearly with age. Moreover, when observing the last plot of the compilation, it becomes clear that a reasonable and very simple assumption to describe the behavior of both series is that the gaps between them are constant in time. This is a crucial assumption because it enables the user to adjust the population forecast by the simple expedient of estimating the aforementioned yearly gaps. Just as in the plots, the size of the gaps will depend linearly on age and can be modeled using a GLM. These assumptions are rather strong, nonetheless, they are reasonable when considering the behavior of the data at hand. Some other assumptions might be needed to improve the outcome for other data-sets.

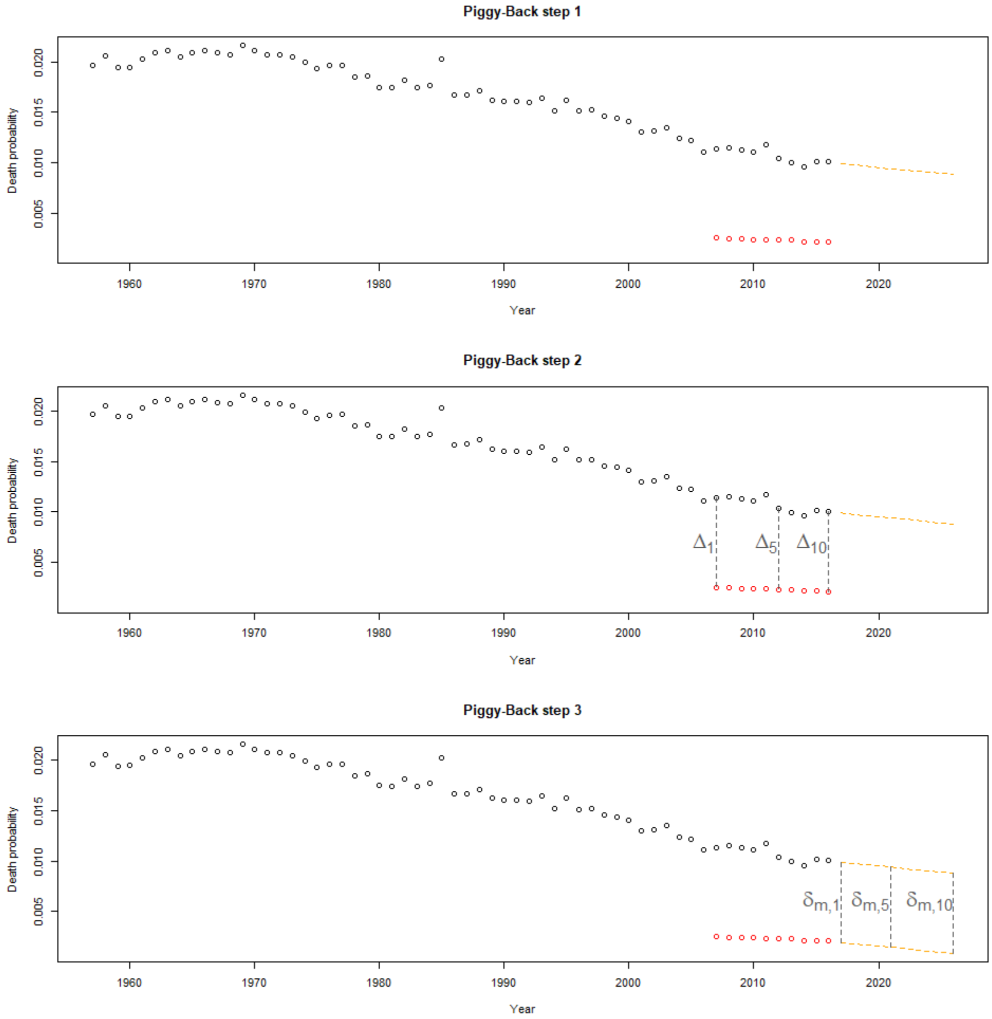


FIGURE 4.2: Piggy-Back procedure

In general terms, the Piggy-Back procedure can be described graphically with the following steps shown in figure 4.2. The first plot portrays the observed population mortality in black, the observed sample mortality in red, and the forecast (using any of the procedures described in section 2.1) in orange. For each year where there are observations for both the population and the sample data the differences Δ_i are taken as portrayed in the second plot. Finally, those differences are modeled (δ_i) and then used to correct the initial forecast so that one that follows the trend of the sample data is obtained.

Another assumption that has been maintained when modeling the behavior of the population data, is that the number of deaths follow a Poisson distribution. This is also further assumed for the behavior of the deaths within the insured portfolio, so that the outcome is consistent in terms of the underlying distribution. Once all these assumptions are set in place, the following model for the differences can be derived:

$$\begin{aligned} D_{x,t} &\sim Poisson(\lambda = e^{c}_{x,t}\mu(x,t)) \\ &40 \leq x \leq 89, 2007 \leq t \leq 2016 \\ log(m(x,t)) &= \hat{\alpha_x} + \hat{\beta_x}\hat{k_t} + a_0 + a_1 x_x \end{aligned} \tag{4.12}$$

$$D_{x,t} = a_0 + a_1 X + \varepsilon \tag{4.13}$$

As seen in equation 4.12, in the end, the company's mortality is explained by a combination of the in-sample fit obtained by a sum of the estimation given by the model at the population level (Lee-Carter, CBD, P-Splines or any combination) and the result when modeling the sample deaths via the GLM in equation 4.13. In the case of the Lee-Carter, the population fit is represented by the segment $log(m(x,t)) = \hat{\alpha_x} + \hat{\beta_x}\hat{k_t}$ of the equation[2].

Equation 4.13 corresponds to the GLM that regresses as the dependent variable, the deaths occurred in the insurance portfolio and as an independent variable the vector of ages 40 to 89. This GLM takes as offsets the level of exposures at each age and year in the portfolio, as well as the estimates given by the population data for those same ages and years. In simple terms, the role of the offset is to shift the intercept, so that the estimation is adapted to both the size of the portfolio and the population

[2] If the CBD model were to be employed, this segment would be equal to equation 4.7 and in case it was done for the P-Splines the segment would be equal to equation 4.8, or any of the estimates obtained with a given combination.

estimates. It is important to note that this GLM uses a Poisson distribution as the link function. Therefore, the remainder of the equation 4.12 corresponds to the intercept and age coefficient of the aforementioned liner model.

Once the modelization is finalized, the sample forecast is simply achieved by taking the forecasted values of the population model and correcting them with with the computations obtained when utilzing the intercept and slope obtained from the GLM as follows:

$$\begin{gathered} log(m(x,t)) = \hat{\alpha_x} + \hat{\beta_x}\hat{k_t} + \hat{a_0} + \hat{a_i}x_x \\ 40 \leq x \leq 89, 2007 \leq t \leq 2026 \end{gathered} \tag{4.14}$$

All in all, the model receives its name *Piggy-Back* because it is piggybacking the company's forecast on an existing in sample-fit and forecast at the population level. The advantage of this model is that even when portfolio data does not support a stand-alone forecast, a piggyback model should remove some of the basis risk. Even at a higher level of finesse, several adjustments can be fine-tuned to take account of the different risks associated with different classes of business without deviating too much of the real values and causing unnecessary capital costs for the insurer.

Chapter 5

Results

5.1 In-sample fitting and forecasting

The following subsections will show the results obtained for both the in-sample fitting and forecasting as well as the goodness-to-fit measures described in section 4.1.4. The results for each method will be shown separately and then combined according to the combinations described in equation 4.10.

5.1.1 Lee-Carter model results

To perform the fit the R package StMoMo is employed. The Lee-Carter model is fitted assuming a Poisson distribution, therefore the link is a log function. The mortality rates estimated by the model are obtained through the *fitted* function whose argument is an object in the workspace that contains the model fitted values. In the end all models will be fitted using a shorter time series than the one available described in section 3.1.1 to avoid outliers such as the high levels of mortality seen due to the Spanish flu and the Dutch Potato Riots. The period used for the fit gets reduced to 50 years: from 1957 to 2017.

For the forecast, a bi-variate random walk with drift is used to project the rates for the future years. In this case, the forecast is done for 10 years, corresponding to the period 2007-2016. For the statistical inference regarding the parameter uncertainty, a sample of 1000 projected trajectories are generated using a Poisson distribution with the parameters obtained when fitting the model to the data. Once the samples are generated, a Lee-Carter model is fitted to each one of them and extracting the model parameters. With the said parameters, a random walk with drift is used to make a projection the rates for the future 10-year period. To obtain the confidence intervals, among the curves simulated earlier, the ones that belong to the 00.5% and 99.5% quantiles are calculated.

Figure 5.1 shows the fit obtained from the Lee-Carter model for ages 64 to 65. The actual yearly death probabilities are represented by the circles while the fitted values are represented by the solid black line. Moreover, the forecasted death rates are depicted by the solid red line while the confidence interval due to parameter uncertainty is shown with the dashed red lines.

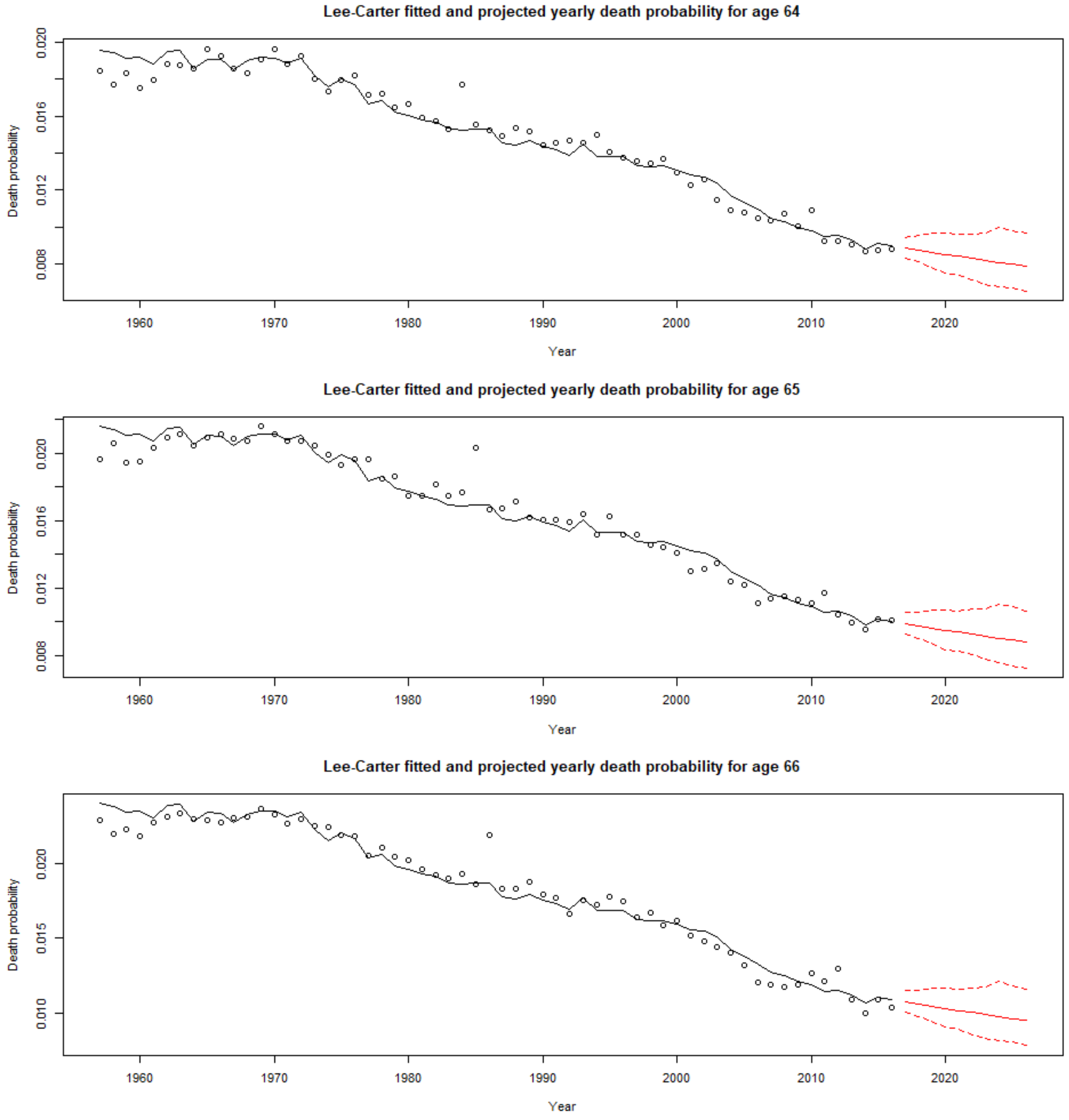


FIGURE 5.1: Lee-Carter death probabilities ages 64-66

From the data, it is visible that at the beginning of the series (years before 1975) the mortality has a higher variance. This observation could be explained by a lack of refinement in the data collection methods, which got perfected over time with the appearance of new technologies. Or due to a larger vulnerability of the population to external circumstances, as medicine and technologies were less advanced during these times. All in all, the downward trend of mortality seems to be consistent across all years, there is a raise in mortality that stands out in 1984. According to Mackenbach et al. [1991]:

> *"The relatively high mortality rates in the southern part of The Netherlands, which dominate the geographical mortality pattern of the country as a whole, appear to be largely due to cardiovascular diseases. Four cardiovascular diseases, which together account for only 43% of all deaths, account for 86% of the excess all cause mortality in the South in 1984."*

Regarding the fit, it is quite accurate along with all the series. In the beginning, it is visible how the fitted values accommodate to the changes in mortality, capturing up to a certain extent the wider movements of the death rates. For the first part of the data, the fit is similarly accurate for all ages displayed, it is important to remember that the goodness-to-fit will vary for every age. When considering the R^2 of the model, the differences between the actual and fitted values across all ages and years are taken into account. Therefore, better a fit for the younger age groups may compensate for the lower accuracy at older ages, not penalizing that much the goodness-to-fit of the sample as a whole.

It is also visible in the confidence intervals the prediction is less trustworthy the further away in time. This is concluded as the confidence intervals are narrower in the beginning and get wider in the end yet symmetrical around the prediction. This conclusion is the usual one. It is natural to expect that the further away from the less ability the historical data has to predict the events of interest. Finally, the predicted death probabilities are expected to decrease in the future, meaning that the forecast captures the improvements in mortality, implying that each cohort will live for more years than the precedent one.

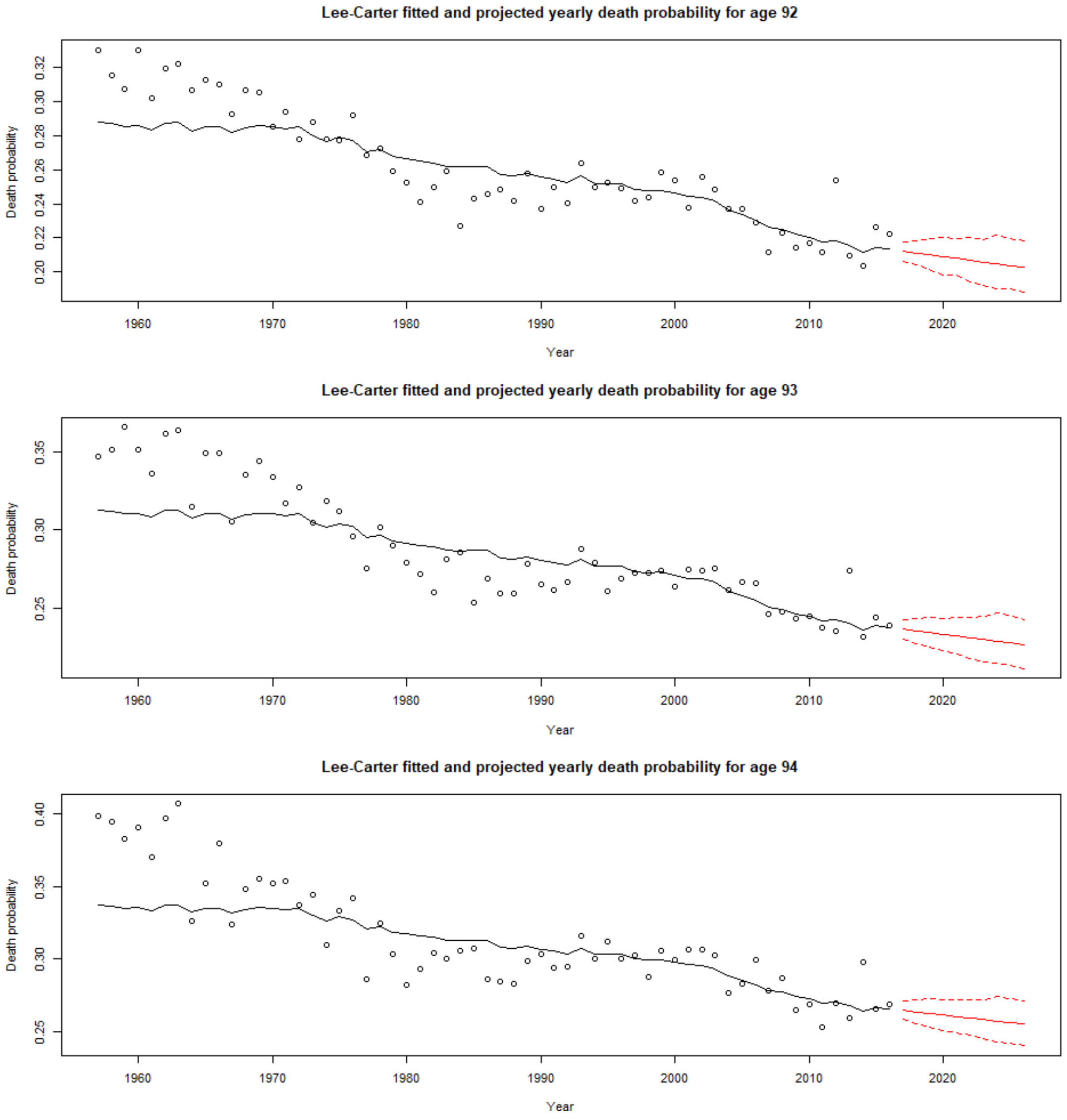


FIGURE 5.2: Lee-Carter death probabilities ages 92-94

Figure 5.2 depicts the fit obtained from the Lee-Carter model for ages 92 to 94. Just as discussed in earlier chapters, data at older ages have a higher variance, making the fit less accurate. This larger variance occurs as fewer individuals make it to these ages making the data less reliable in terms of variability. Nonetheless, just as with younger ages at each age the first part of the time series is more sparse, and then it becomes less variable towards the end. Moreover, the confidence intervals are somewhat wider in comparison to those of younger ages, which makes sense as the information for older ages is more difficult to predict. Also, the slope of the forecast is less steep for older ages, meaning that the improvements in mortality are predicted to be smaller in comparison to younger ages.

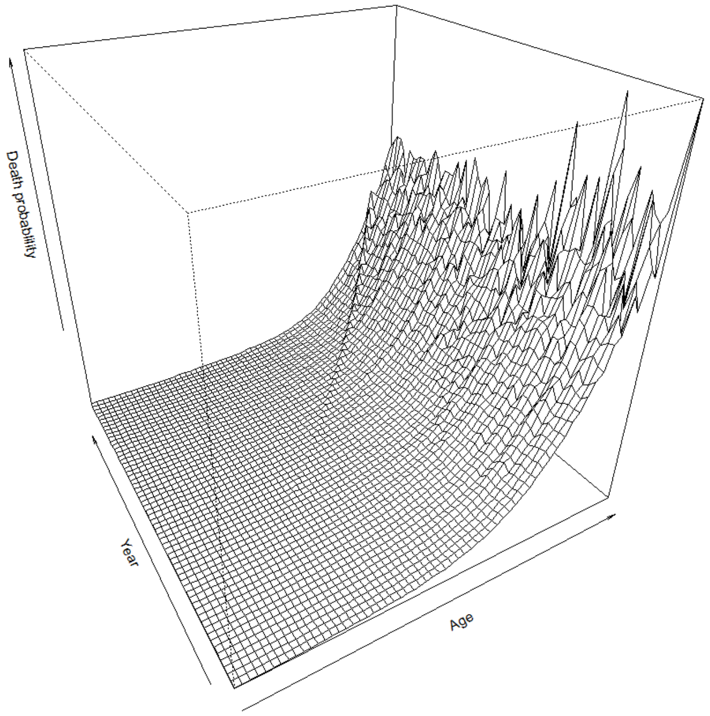


Figure 5.3: Observed death probabilities surface per year and ages 40 to 100

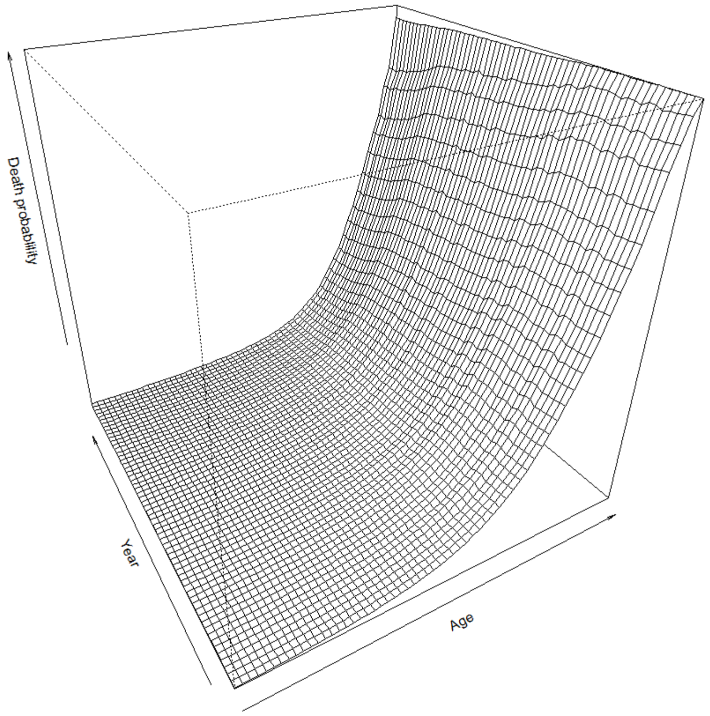


Figure 5.4: Lee-Carter fitted death probabilities surface per year and ages 40 to 100

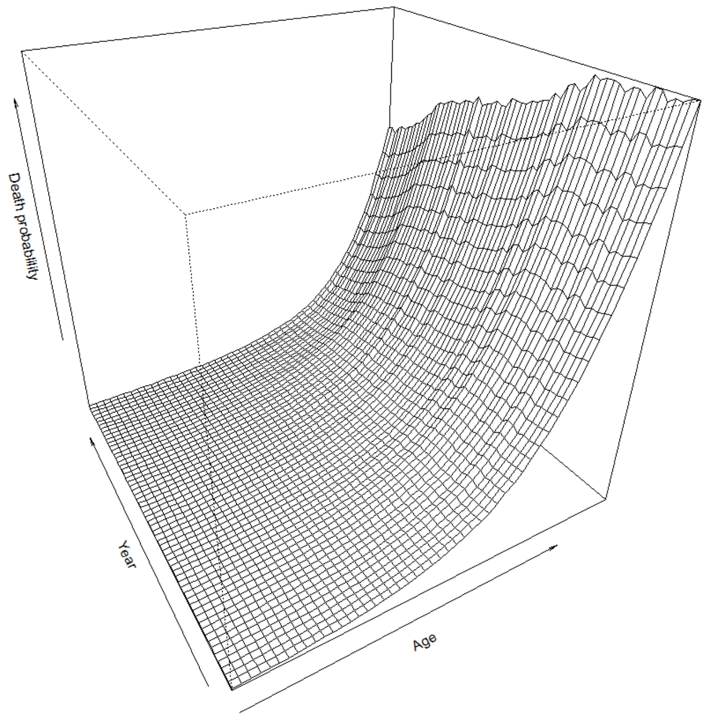


Figure 5.5: Lee-Carter fitted death probabilities surface per year and ages 40 to 89

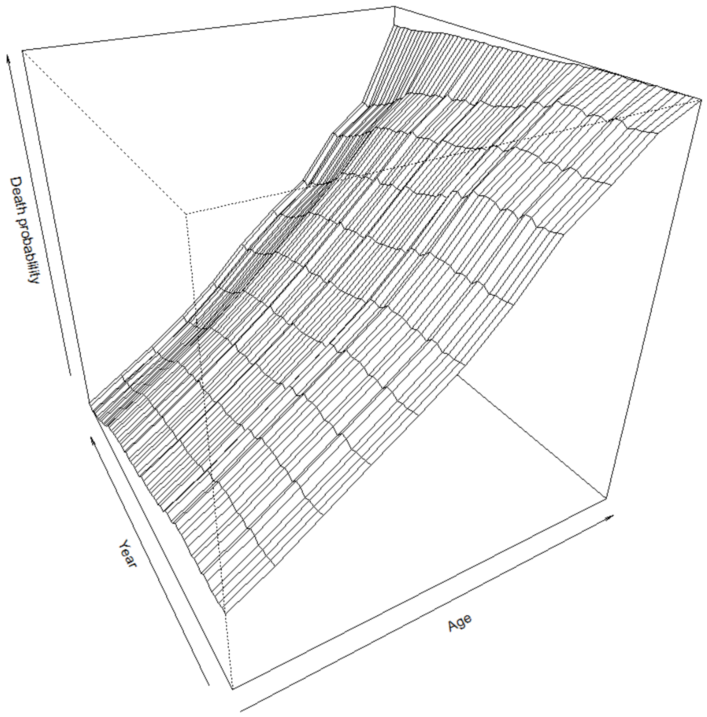


Figure 5.6: Lee-Carter fitted death probabilities surface per year and ages 90 to 100

Figure 5.4 shows the full fitted mortality surface for all years and ages. Then, figures 5.5 and 5.6 break the same surface but for ages 40 to 89 and 90 to 100 respectively. The solid black lines represented in the figures 5.1 and 5.2 correspond to cross-sections of this surface, while the dots on the same figures correspond to the surface represented in figure 5.3 that represents the whole observed mortality surface for all years and ages. It is visible that mortality increases across ages in a rather stable pattern between ages 40 to 90. Moreover, at older ages the pattern is less clear and from age 100 death probabilities increase at a higher pace.

When comparing the observed death probabilities with the fitted values in figures 5.3 and 5.4 respectively, it is visible that the fitted mortality is smoother, in other words, has fewer jumps, than the real values, especially in the second half of the age vector. This occurs since not all the outliers are captured by the fitted function. In general, a smoother function is preferable over one that captures every occurrence, otherwise it will indicate overfitting.

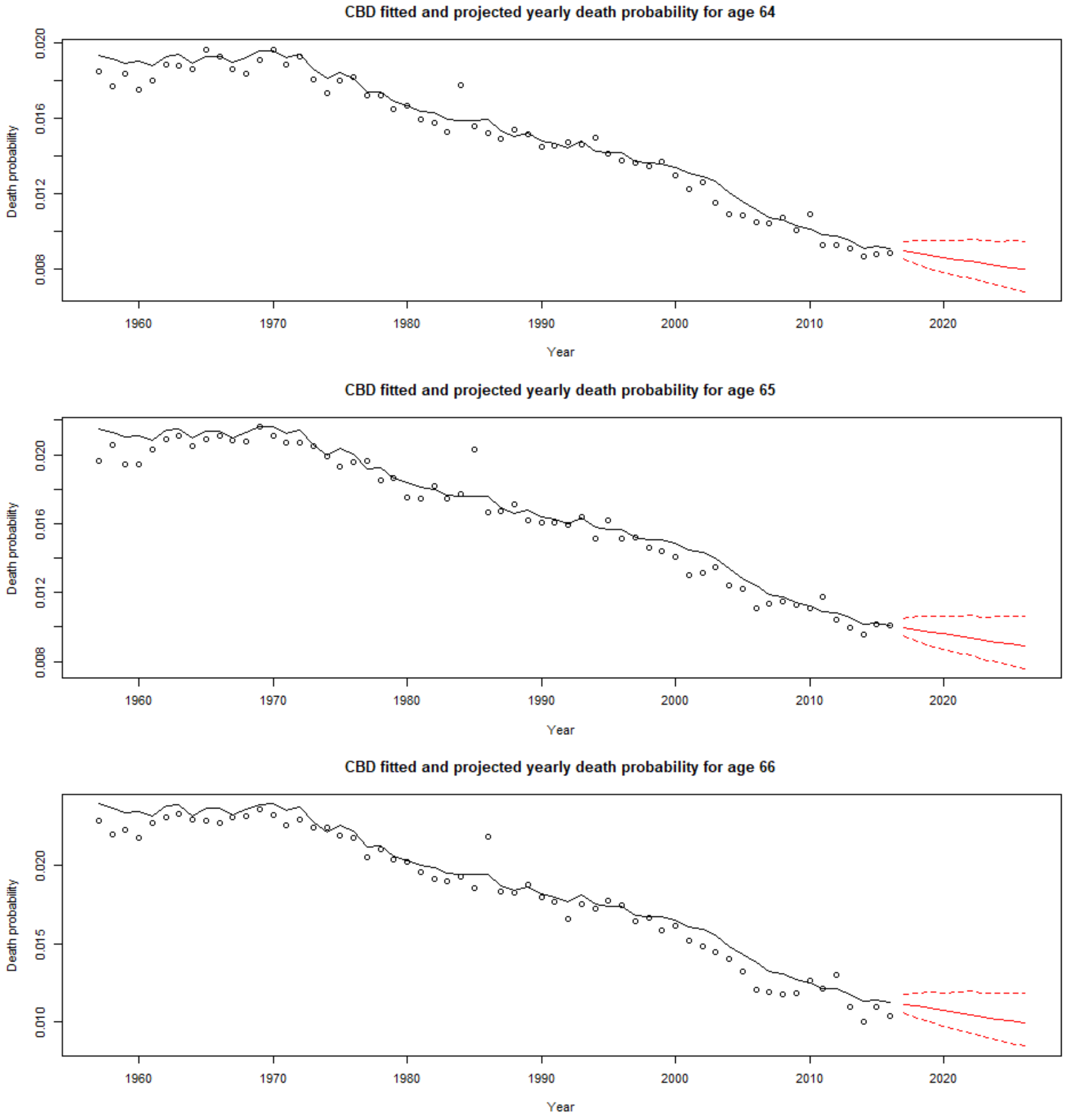
5.1.2 CBD model results

Just as with the previous model, the fit is done using the R package *StMoMo*. The CBD model is fitted assuming a Poisson distribution, therefore the link is a log function. The mortality rates estimated by the model are obtained through the *fitted*. For the forecast, a bi-variate random walk with drift is used to project the rates for 10 years in the future. A bootstrapping technique is used to perform the statistical inference regarding the parameter uncertainty, with the results of the projected trajectories and their parameters the 00.5%, and 99.5% quantiles are calculated.

Figure 5.7 represents the fit obtained from the CBD model for ages 40 to 42. Following the same conventions as in previous graphs, the actual yearly death probabilities are represented by the circles, the fitted values by the solid black line, the forecasted death rates by the solid red line, and the confidence interval due to parameter uncertainty by the dashed red lines. As the observed death probabilities are the same for all models, figures 5.1 and 5.7 are on the same scale, which is useful to compare both fits visually.

Although all models result in a good fit across all ages and years, for ages 64 to 65, between 1998 and the end of the sample the CBD model overestimates death probabilities slightly more in comparison to the Lee-Carter fit. Nonetheless, both fits behave similarly.

In general, the CBD fitted values tends to overestimate death probabilities across all years, but this bias is not statistically significant.



CBD fitted and projected yearly death probability for age 64

CBD fitted and projected yearly death probability for age 65

CBD fitted and projected yearly death probability for age 66

FIGURE 5.7: CBD death probabilities ages 64-66

Just as the Lee-Carter model, the CDB fit captures the changes in mortality, even portraying up to a certain extent most of the changes in the death rates. The confidence intervals are of similar width for both models, meaning that there is no significant difference in terms of parameter uncertainty. Moreover, there is a predicted improvement in mortality but it is slightly higher than the one predicted by the Lee-Carter fit. All in all, the model has an outstanding in-sample fit, as well as an excellent forecast accuracy when calculating a goodness-to-fit statistic based on a 10-year back-testing.

Figure 5.9 shows the fitted mortality surface for all years and ages 40 to 89, while figure 5.8 represents the whole observed mortality surface for that same age-year space as follows:

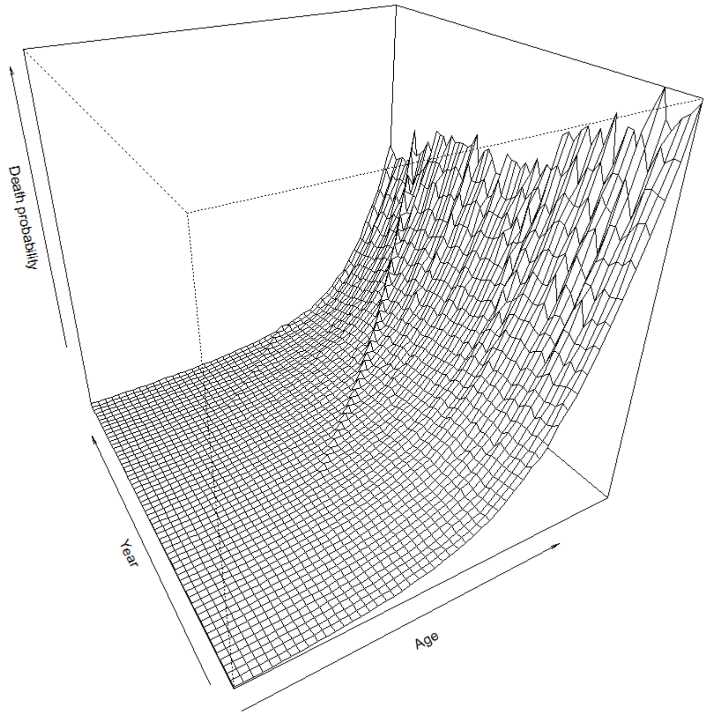


Figure 5.8: Observed death probabilities surface per year and ages 40 to 89

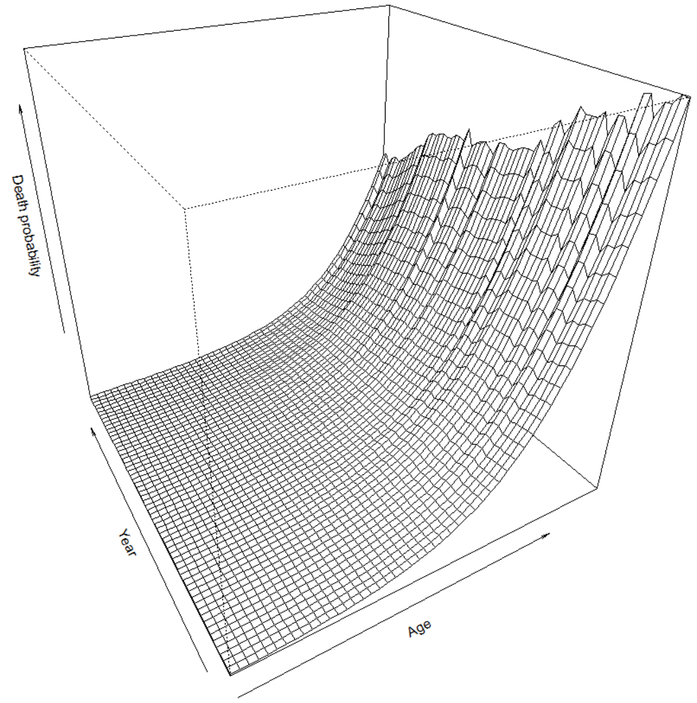


Figure 5.9: CBD fitted death probabilities surface per year and ages 40 to 89

The figures show how mortality increases across the observed ages. When comparing the observed death probabilities with the fitted values in both figures, once again, the fitted mortality is smoother and has less variability across the whole surface. All in all, the fitted surface mimics quite accurately the observed one.

5.1.3 P-splines model results

For the P-splines in-sample fit the R library *MortalitySmooth* is employed, here the Poisson distribution is utilized. In the end, a two-dimensional P-spline model is fitted with 6 nodes for age and 13 for the year. For the model adjustment, adjust the *Mort2Dsmooth* function is used, whose arguments are the ages, the years, a matrix summarizing the observed deaths, a matrix for the initial exposure to risk as well as the number of nodes.

There is not a standardized way to choose the number of nodes for age and years, choosing them can be tricky, because if done incorrectly the resulting projections could be odd. Literature indicates that employing between 10 and 15 nodes usually yields correct approximations. Frequently the higher the amount of nodes yields a better the in-sample fit, as there are more individual sections to accommodate each part of the data more exactly. Nonetheless, a good in-sample fit can lead to overfitting causing trouble in the forecast. To choose the number of nodes, the in-sample fit is measured through the AIC criterion; the objective is to minimize it.[1]

After choosing the number of nodes, the range of years to predict is defined between 1957 to 2026, the algorithm will take the last 10 years of data as missing values, thus creating a prediction as a result of the smoothing process. The predictions along with their standard errors are calculated and from that the confidence intervals. Figure 5.10 represents the fit obtained from the P-splines model for ages 64 to 66, following the same conventions as in previous graphs. Unlike the previous models, this fit is smoother, it is a differentiable function across all of its dominion. Unlike the previous fits, it manages to capture the general trend without paying too much attention to the outliers. In general, if there are not that may extreme values this might be a positive attribute to the fit. Moreover, the confidence intervals of the prediction are somewhat wider towards the end than the ones given by the other models.

[1] To see all the node combinations and their respective goodness to fit measures refer to Appendix B.

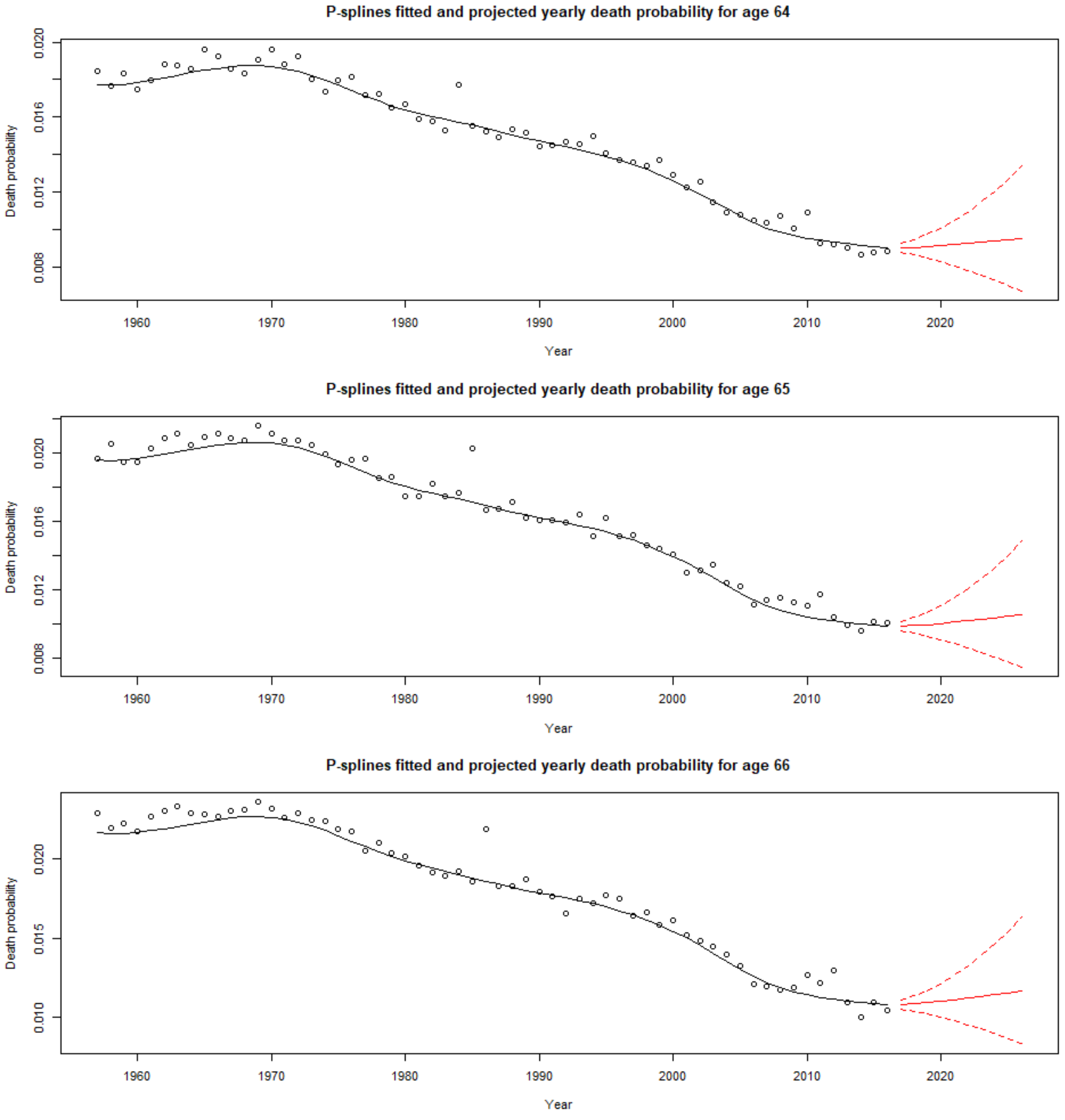


FIGURE 5.10: P-splines death probabilities ages 64-66

Figure 5.11 shows the fitted mortality surface for all years and ages 40 to 89, when compared to the observed values in figure 5.8 the smoothness of the fit becomes apparent. Differentiability and continuity are both desirable properties of the function which tan be attained when using these types of splines modeling. The fitted surface reflects the fact that mortality increases across the observed ages.

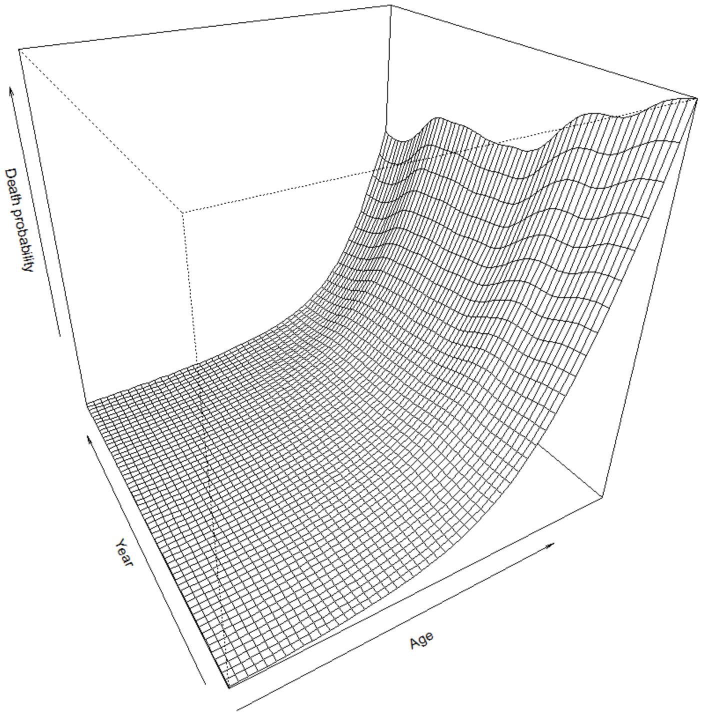


FIGURE 5.11: P-splines fitted death probabilities surface per year and ages 40 to 89

5.1.4 Overall final fitting and forecast

The following figure shows all three in-sample fits for the same age across all years along with their forecasted confidence intervals, as well as the projections for the model combinations:

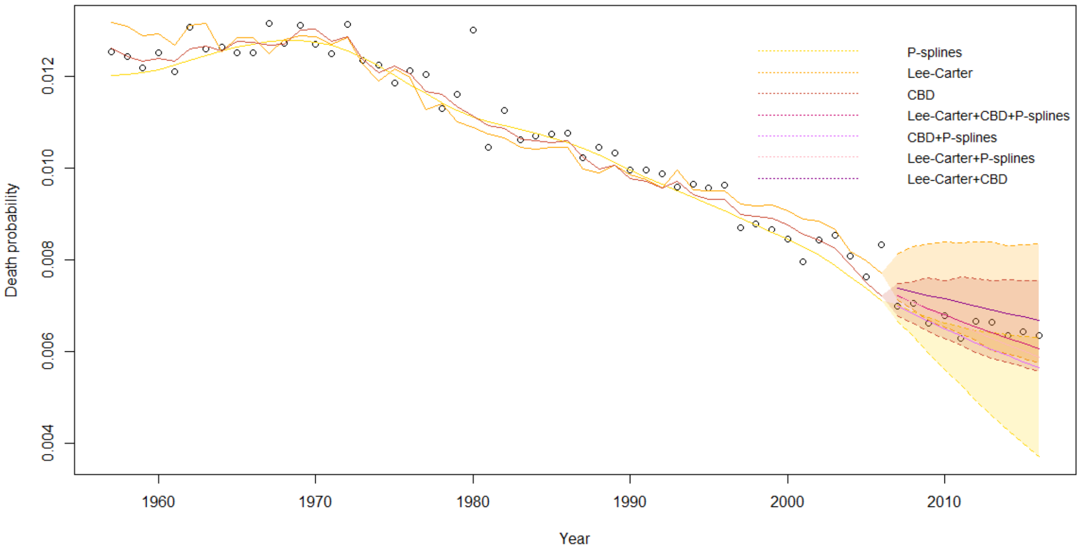


FIGURE 5.12: Lee-Carter, CBD and P-splines fitted death probability and backtesting for age 65

It is important to point out that the confidence intervals can not be easily derived when combining models, therefore here only the projection is plotted. Here the fit is performed only until 2015 leaving the last 10 years of data out of the sample to see how they behave in comparison to the forecast

When observing all three fits back to back, all the models have a relatively good fit throughout the whole series. Unlike its two counterparts, the P-Splines model captures the sudden jumps in death probabilities in a softer (more smooth) manner. The predictions made by all three models result in confidence intervals that overlap. The projections that have the largest area of the intervals in common are the Lee-Carter and the CBD models. While the Lee-Carter projection in general predicts higher mortality rates, the P-splines predicts the lowest ones. This occurs to the fact that the smoothing technique tries to follow the slope of the downward-curved pattern given by the last 15 years of observations. In general most of the observed mortality data points happen to be inside the forecasted intervals that were left outside for the modelization depicted in this figure. It is important to point out that all things considered, all the models and their combinations behave well and are close to the forecasted values.

The following table shows the goodness-to-fit statistics for the in-sample modelling as well as the forecast for the individual models, calculated according to equation 4.9:

Model	***In-sample*** R^2	***Forecast*** R^2
LC_O	0.9232655	0.9677311
LC_Y	0.9989670	0.9947542
$LC_{(Y+O)}$	0.9990588	0.9969759
CBD_Y	0.9970984	0.9943769
PS_Y	0.9985459	0.9905307

Table 5.1: Goodness-to-fit for the different individual models. The subscript Y means that the fit has been done for the young ages (40-89), while the subscript O means that the fit has been done for old ages (90-100). LC and PS refer to the Lee-Carter and P-Splines models respectively.

Lastly, the following table lists the RSS for the in-sample modelling as well as the forecasted series for the individual models and their combinations as described in equation 4.10. The residual sum of squares is calculated according to equation 4.10:

Model(s)	***In-sample*** RSS	***Forecast*** RSS
LC_O	3.769604	0.2699137
LC_Y	6.094333	4.653584
PS_Y	9.423546	7.53237
CBD_Y	18.80440	6.25437
$PS_Y + CBD_Y$	9.723201	5.834666
$LC_Y + CBD_Y$	8.532539	5.327389
$LC_Y + PS_Y$	6.156271	4.270949
$LC_Y + PS_Y + CBD_Y$	7.008102	4.728295

Table 5.2: Residual sum of squares for the different individual models.

As discussed earlier, all three models can be used to smooth $q(x;t)$ for ages 40 to 89 years, while for ages 90 to 95 years, only the Lee-Carter can be used, since the CBD and P-splines can not be employed for advanced ages. Therefore the individual and combined fits and forecasts are tested individually for young ages only (40 to 89) and in the case of the Lee-Cater for the whole series (40 to 95). Due to the higher variability of data at old ages, the Lee-Carter in sample fit can explain 96% of the changes in the data. In general terms, all fits are excellent. For the individual models, at young ages, the highest in-sample fit is achieved by the Lee-Carter model (LC_Y). Finally, for the forecast, the best fit across young ages is achieved by the Lee-Carter model alone (LC_Y). Nonetheless, as seen by the R^2 the goodness-to-fit differs by a matter of centesimal places, meaning that all explain more of the 95% of the changes in the data.

When comparing the RSS between the models and their combinations, the Lee-Carter alone appears to be superior in terms of the in-sample fit, while the Combination of the Lee-Cater and the P-Splines seems to yield a better forecast, although its quadratic residual is only slightly higher than that of the Lee-Carter model. The RSS for older ages is significantly lower compared to those of younger ages, but this occurs simply because there are fewer observations in the older segment, therefore less adding terms. Due to its high performance, the Lee-Carter forecast and in-sample estimation will be chosen as the base to perform the integration of the sample and population data.

5.2 Results when integrating population and sample data

After performing the procedure described in section 4.2 when selecting the Lee-Carter in-sample fitting coefficients and forecast, the following plots are obtained:

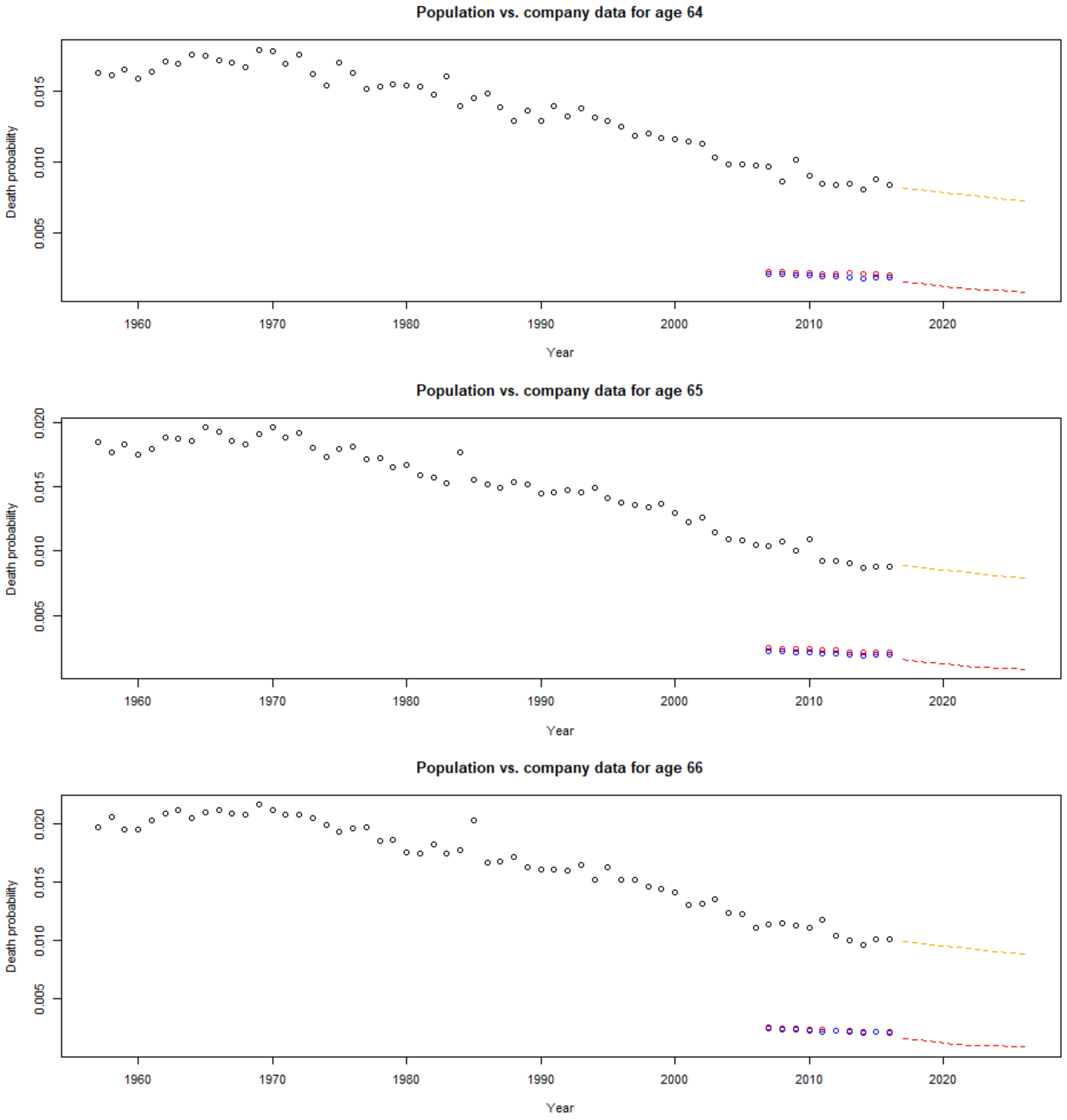


Figure 5.13: Sample forecast ages 64-66

Here, the observed population mortality is shown in black, the observed sample mortality in red, and the forecast using the Lee-Carter model in orange. The data points shown in blue, correspond to the estimations using the coefficients obtained when fitting the GLM in equation 4.13. Both sets of data points, that refer to the observed and the fitted values (red and blue) overlap almost perfectly, meaning that the GLM manages to explain the differences between the population and the sample series very well. Finally, the dashed red line corresponds to the forecast of the insured sample shown in equation 4.14. The forecast for the sample is slightly more steep than the one of the general population, meaning that the insured clients are expected to have larger improvements in their mortality level when compared to the total population. Moreover, the sample forecast represents a reasonable outcome when seen in the context of both the population and the sample data. It takes into account the sample size of the insured portfolio via the offset parameter and also the trends of both the insured population series as well as the Lee-Carter prediction. In general, the outcome of the integration proves to be satisfacotry

Chapter 6

Conclusions

Evaluating and forecasting stochastic variables such as the mortality rates is not a simple task. There is always uncertainty in the forecasts, therefore insurers seek prudence and despite allowing for this uncertainty, resort to other methods to ensure confidence in their outcomes. For instance, they discount the predicted mortality weighed payments employing lower interest rates or set surcharges on the mortality predictions. These actions naturally have financial implications for the pricing and reserving of annuities and pensions. In most cases, these procedures promote the regulator's interests at the expense of those of the companies and do not show the real situation of the insurer and its business portfolio.

Moreover, predicting the behavior of the insured population is difficult due to the reduced length of the data-sets at hand, so the pragmatic approaches described earlier are employed by the insurers. Research on how to accurately forecast the mortality trends of the insured population is pretty limited, despite it being a subject of exceptional importance. This document has presented an alternative way of projecting insured mortality by piggybacking the company's forecast on an existing in sample-fit and forecast at the population level. This approach results in a reasonable and trustworthy forecast that could mean an alternative to the insurers avoiding them unnecessary capital costs when calculating their solvency capital requirements.

All in all, it is shown that when analyzing historical mortality patterns form the Netherlands, all the models employed (The Lee-Carter, CBD, and P-splines) result in a good in-sample fit and forecast. Moreover, the methodology proposed to correct the population mortality forecast, based on the differences of common sample points, so that it resembles the trend of the insured portfolio, appears as adequate. By implementing the model suggested in this document, it is possible to take advantage of the larger extension of the population-level data and pool the accuracy of the forecasts stemming from those longer series to generate a reasonable and reliable forecast at the sample-level.

Appendix A

Two-dimensional Kronecker product cubic B-spline basis

This figure is taken form Currie et al. [2004]. It shows an example of a bi-dimensional B-splines base. This figure is analogous to the process illustrated previously in one dimension in figures 2.1 and 2.2.

> *"The age–year grid is populated by a set of overlapping hills which are placed at regular intervals over the region. Each hill is the Kronecker product of two one-dimensional hills (B-splines), one in age and one in year. For clarity, only a subset of hills from a small basis is shown in [the figure], but in practice there are about 200 such hills which give a dense covering of the age–year region, and this results in a flexible basis for two-dimensional regression."* (Currie, Durban, and Eilers [2004]; pages 8-9).

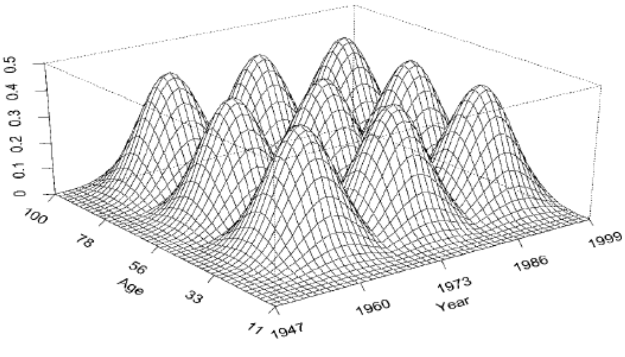


FIGURE A.1: Example of a third degree bi-dimensional B-spline: It is made up of pieces of cubic polynomial surfaces. The y axis corresponds to the mortality rate.

Appendix B

P-Splines node selection

This table depicts all the in-sample goodness-to-fit as well as the forecast R^2 form the backtesting procedure. All the possible combinations between 2 and 15 nodes for both age and year were tested. To choose the optimal number of nodes the aim is to maximize the BIC while having a relatively high R^2 for the forecast. In the end the chosen combination was 4 nodes for the age dimension and 15 for the year dimension.

Age Nodes	***Year Nodes***	***In-sample BIC***
2	2	8087.204
3	2	7831.392
4	2	7838.926
5	2	7855.739
6	2	7870.809
7	2	7887.129
8	2	7903.234
9	2	7918.751
10	2	7925.654
11	2	7937.392
12	2	7949.696
13	2	7960.847
14	2	7972.937
15	2	7976.975
2	3	7229.560

Age Nodes	***Year Nodes***	***In-sample BIC***
3	3	6978.473
4	3	6952.902
5	3	6976.183
6	3	6993.589
7	3	7020.617
8	3	7045.936
9	3	7072.457
10	3	7085.389
11	3	7105.143
12	3	7122.973
13	3	7140.968
14	3	7155.483
15	3	7167.930
2	4	5061.206
3	4	4738.994
4	4	4725.748
5	4	4743.726
6	4	4758.990
7	4	4791.990
8	4	4819.271
9	4	4844.970
10	4	4861.891
11	4	4881.155
12	4	4903.416
13	4	4916.635
14	4	4932.792
15	4	4941.501
2	5	5036.770
3	5	4677.186
4	5	4696.006
5	5	4693.806
6	5	4713.969

Age Nodes	***Year Nodes***	***In-sample BIC***
7	5	4738.276
8	5	4754.183
9	5	4766.440
10	5	4770.618
11	5	4774.403
12	5	4784.261
13	5	4784.791
14	5	4787.234
15	5	4791.334
2	6	5153.044
3	6	4826.065
4	6	4795.663
5	6	4805.898
6	6	4789.207
7	6	4805.782
8	6	4813.563
9	6	4826.726
10	6	4826.390
11	6	4834.367
12	6	4838.834
13	6	4838.023
14	6	4842.991
15	6	4850.023
2	7	5042.262
3	7	4716.478
4	7	4689.197
5	7	4671.675
6	7	4666.627
7	7	4676.771
8	7	4687.985
9	7	4695.425
10	7	4699.759
11	7	4708.355

Age Nodes	***Year Nodes***	***In-sample BIC***
12	7	4706.530
13	7	4709.123
14	7	4717.085
15	7	4721.690
2	8	4768.422
3	8	4439.611
4	8	4411.054
5	8	4393.442
6	8	4380.428
7	8	4391.788
8	8	4404.079
9	8	4408.108
10	8	4413.867
11	8	4419.515
12	8	4419.226
13	8	4423.075
14	8	4431.924
15	8	4431.344
2	9	4868.332
3	9	4546.049
4	9	4515.569
5	9	4494.023
6	9	4475.148
7	9	4489.640
8	9	4502.268
9	9	4504.793
10	9	4512.613
11	9	4512.210
12	9	4513.990
13	9	4520.171
14	9	4529.217
15	9	4523.935
2	10	4901.413

Age Nodes	***Year Nodes***	***In-sample BIC***
3	10	4580.767
4	10	4532.847
5	10	4506.330
6	10	4486.599
7	10	4497.254
8	10	4511.556
9	10	4512.946
10	10	4520.686
11	10	4523.520
12	10	4525.239
13	10	4530.500
14	10	4538.892
15	10	4533.531
2	11	4825.742
3	11	4508.142
4	11	4453.606
5	11	4430.945
6	11	4411.936
7	11	4420.726
8	11	4433.894
9	11	4435.161
10	11	4444.416
11	11	4445.031
12	11	4446.501
13	11	4452.469
14	11	4462.747
15	11	4458.193
2	12	4748.511
3	12	4434.104
4	12	4391.660
5	12	4373.216
6	12	4354.537
7	12	4362.389

Age Nodes	***Year Nodes***	***In-sample BIC***
8	12	4374.180
9	12	4376.974
10	12	4388.075
11	12	4386.104
12	12	4388.617
13	12	4395.493
14	12	4403.266
15	12	4399.418
2	13	4695.891
3	13	4382.127
4	13	4358.371
5	13	4344.639
6	13	4327.145
7	13	4333.962
8	13	4343.226
9	13	4349.380
10	13	4362.496
11	13	4355.204
12	13	4361.126
13	13	4369.267
14	13	4368.924
15	13	4366.774
2	14	4764.321
3	14	4458.377
4	14	4430.924
5	14	4406.335
6	14	4388.460
7	14	4398.418
8	14	4413.169
9	14	4413.574
10	14	4423.517
11	14	4425.532
12	14	4426.654

Age Nodes	***Year Nodes***	***In-sample BIC***
13	14	4432.346
14	14	4442.935
15	14	4438.239
2	15	4777.333
3	15	4476.214
4	15	4418.124
5	15	4398.097
6	15	4381.028
7	15	4389.767
8	15	4403.324
9	15	4404.060
10	15	4414.576
11	15	4414.265
12	15	4415.618
13	15	4421.439
14	15	4432.076
15	15	4429.694

Table B.1: In-sample goodness-to-fit for all possible combinations between 2 and 15 nodes for both year and age.

Appendix C

R code and documentation

C.1 Loading the data

```
#Set the directory.

setwd("~/UC3M/TFM/Datos_Poblacionales")

# Environment until line 693
load("~/UC3M/TFM/Datos_Poblacionales/Environment_Thesis_script.RData")

#Sets the bottom, left, top and right margins respectively.

par(mar = c(4, 4.5, 4, 2))
par(mfrow=c(1,1))

#Read the entire data base corresponding to years 1910-2016.

Dx_T_NL <- read.delim("Dx_NL.txt", header=F)
Exp_T_NL <- read.delim("Ex_NL.txt", header=F)
lx_T_NL <- read.delim("lx_NL.txt", header=F)

Dx_T_NL2<-Dx_T_NL[!is.na(Dx_T_NL)]
Exp_T_NL2<-Exp_T_NL[!is.na(Exp_T_NL)]
lx_T_NL2<-lx_T_NL[!is.na(lx_T_NL)]

Dth_1<-matrix(Dx_T_NL2,111,107)
Exp_1<-matrix(Exp_T_NL2,111,107)
lx_1<-matrix(lx_T_NL2,111,107)

#Define the age interval. All three models are applied simultaneously to ages
```

```
#40-90 and only the Lee Carter can be used to adjust advanced ages.

Dth=Dth_1[41:111,]
Exp=Exp_1[41:111,]
lx=lx_1[41:111,]

death=Dth
exposure=Exp
Age <- 40:110
Year <- 1910:2016

ages <-40:110
years <-1910:2016
```

C.2 EDA

```
#Graph: Number of deaths.

matplot(0:110,Dx_T_NL,type="l",lty=3,col=c(rep("gold",10),rep("gold3",10),
rep("orange",10),rep("orange2",10),rep("darkorange",10),rep("darkorange3",10),
rep('tomato',10),rep("tomato3",10),rep("red",10),rep("red3",10),rep("red4",7)),
ylim=c(0,6000),cex.axis=0.8,main="Total number of deaths",ylab="Deaths",xlab="Age")

legend(7,6000,legend=c("1910-1919","1920-1929","1930-1939","1940-1949","1950-1959",
"1960-1969","1970-1979","1980-1989","1990-1999","2000-2009","2010,2016"),cex=0.7,
lty=3,col=c("gold","gold3","orange","orange2","darkorange","darkorange3",'tomato',
"tomato3","red","red3","red4"))

#Graph: Mortality rates

matplot(0:100,log(Dx_T_NL[1:101,]/lx_T_NL[1:101,]),type="l",lty=3,
col=c(rep("gold",10),rep("gold3",10),rep("orange",10),rep("orange2",10),
rep("darkorange",10), rep("darkorange3",10),rep('tomato',10),
rep("tomato3",10),rep("red",10),rep("red3",10),rep("red4",7)),xlab="Age",
ylim=c(-10,0), main="Death probablities (log scale)",ylab="ln(qxt)")

legend(70,-6,legend=c("1910-1919","1920-1929","1930-1939","1940-1949","1950-1959",
"1960-1969","1970-1979","1980-1989","1990-1999","2000-2009","2010,2016"),cex=0.7,
lty=3,col=c("gold","gold3","orange","orange2","darkorange","darkorange3",'tomato',
"tomato3","red","red3","red4"))

#Average raw mortality per age.

raw_mort_EDA<-Dx_T_NL[1:101,]/lx_T_NL[1:101,]
```

```
raw_mort_EDA<-t(raw_mort_EDA)
colMeans(raw_mort_EDA)

#Variance of the raw mortality per age.

var<-c()
for (i in 1:101){
  var[length(var)+1]  = var(raw_mort_EDA[,i])
}
plot(0:100,sqrt(var))
```

C.3 Lee Carter

```
#Modify the data matrices, the series as it is is too long and it might
#affect the fit, therefore it is limited to the past 60 years of information
#selecting only the time frame 1957-2016.

Dth=Dth_1[41:111,48:107]
Exp=Exp_1[41:111,48:107]
lx=lx_1[41:111,48:107]

death=Dth
exposure=Exp
Age <- 40:110
Year <- 1957:2016

ages <-40:110
years <-1957:2016

#The data is introduced so that it can be recognized by the StMoMo library,
#making use of the EWMaleData object which is already built in.

library(StMoMo)

NLData=EWMaleData
NLData$years=years
NLData$ages=ages
NLData$Dxt=Dth
NLData$Ext=lx
NLData$series="total"
NLData$label="Netherlands"

#MODEL FITTING
```

```
#The Lee-Carter model is fitted assuming a Poisson distribution,
#so the link is log.

LCfit_NL <- fit(link="log",lc(), data = NLData)

#The mortality rates estimated by the Lee-Carter model are obtained using the
#"fitted" function whose argument is an object that contains the model fit.
#Since the type = "rates" we get directly the fitted probablities.
#If the type = "link" we would get directly the outcome of the alpha + (beta*Kappa)

mxtHat_NL <- fitted(LCfit_NL, type = "rates")

#We use a bi-variate random walk with drift to project the rates for
#the future years (in this case 10). The argument h = 10 indicates the
#number of years we want to predict. For the forecast function, when the
#argument is an object of class "fitStMoMo" by default adjusts a random
#walk with drift.

LCfor_NL=forecast(LCfit_NL,h=10)

#The projected mortality is obtained.

mxtCentral_NL <- LCfor_NL$rates


#INFERENCE INCLUDING UNCERTAINTY ABOUT THE PARAMETERS

#1000 values of a Poisson distribution are generated with the parameters obtained
#when fitting the model to the data.
#Once the samples are generated, a Lee-Carter model is fitted to each simulation,
#obtaining the model parameters for each.

LCboot_NL=bootstrap(LCfit_NL,nBoot=1000,type="semiparametric")

#With the parameters that have been obtained, a random walk with drift is used to
#make a porjection the rates for the future years (10 in this case).

LCsimPU_NL=simulate(LCboot_NL,h=10)


#CALCULATION OF CONFIDENCE INTERVALS AT 99.5%

#Among the simulations above we calculate the curves such that the results are
#between 99.5% of them.

#Calculation of the intervals for the sample.

mxtHatPU0.05_NL <- apply(LCsimPU_NL$fitted, c(1, 2), quantile, probs = 0.005)
```

```
mxtHatPU99.5_NL <- apply(LCsimPU_NL$fitted, c(1, 2), quantile, probs = 0.995)

#Calculation of the intervals for the predictions.

mxtPredPU0.05_NL <- apply(LCsimPU_NL$rates, c(1, 2), quantile, probs = 0.005)
mxtPredPU99.5_NL <- apply(LCsimPU_NL$rates, c(1, 2), quantile, probs = 0.995)

raw_mort <- NLData$Dxt/NLData$Ext
colnames(raw_mort) <- years
rownames(raw_mort) <- ages


#GRAPHS

#40-49 -> for(i in 41:50) | 50-59 -> for(i in 51:60) | 60-69 -> for(i in 61:70) |
#70-79 -> for(i in 71:80) | 80-89 -> for(i in 81:90) | 90-99 (old) ->
#for(i in 91:100)

par(mfrow=c(5,2))
for(i in 41:50) {
  fitted_mxt_x<-mxtHat_NL[i-40,]
  projected_mxt_x<-mxtCentral_NL[i-40,]
  projected_mxt_x_0.05<-mxtPredPU0.05_NL[i-40,]
  projected_mxt_x_99.5<-mxtPredPU99.5_NL[i-40,]
  plot(1957:2016,fitted_mxt_x,type = "l",
  ylim=c(min(fitted_mxt_x,projected_mxt_x_0.05,
  projected_mxt_x_99.5, projected_mxt_x,raw_mort[i-40,]),
  max(fitted_mxt_x,projected_mxt_x_0.05,
  projected_mxt_x_99.5,projected_mxt_x,raw_mort[i-40,])),xlim=c(1957, 2026),
  xlab="Year", ylab="Death probability",
  main = paste ("Lee-Carter fitted and projected yearly death probability for age",
  i-1))
  points(1957:2016,raw_mort[i-40,])
  lines(2017:2026,projected_mxt_x, col="red")
  lines(2017:2026,projected_mxt_x_0.05, col="red", lty="dashed")
  lines(2017:2026,projected_mxt_x_99.5, col="red", lty="dashed")
}


#GRAPHS (SNIPS) TO INCLUDE IN THE DOCUMENT.

#Fit for ages 64-66
par(mfrow=c(3,1))
for(i in 65:67) {
  fitted_mxt_x<-mxtHat_NL[i-40,]
  projected_mxt_x<-mxtCentral_NL[i-40,]
  projected_mxt_x_0.05<-mxtPredPU0.05_NL[i-40,]
  projected_mxt_x_99.5<-mxtPredPU99.5_NL[i-40,]
  plot(1957:2016,fitted_mxt_x,type = "l",
  ylim=c(min(fitted_mxt_x,projected_mxt_x_0.05,projected_mxt_x_99.5,
  projected_mxt_x,raw_mort[i-40,]),max(fitted_mxt_x,projected_mxt_x_0.05,
  projected_mxt_x_99.5,projected_mxt_x,raw_mort[i-40,])),
  xlim=c(1957, 2026),xlab="Year", ylab="Death probability",
  main = paste ("Lee-Carter fitted and projected yearly death probability for age",
  i-1))
  points(1957:2016,raw_mort[i-40,])
  lines(2017:2026,projected_mxt_x, col="red")
  lines(2017:2026,projected_mxt_x_0.05, col="red", lty="dashed")
  lines(2017:2026,projected_mxt_x_99.5, col="red", lty="dashed")
}

#Fit for ages 93-94.
par(mfrow=c(3,1))
for(i in 93:95) {
  fitted_mxt_x<-mxtHat_NL[i-40,]
  projected_mxt_x<-mxtCentral_NL[i-40,]
  projected_mxt_x_0.05<-mxtPredPU0.05_NL[i-40,]
  projected_mxt_x_99.5<-mxtPredPU99.5_NL[i-40,]
  plot(1957:2016,fitted_mxt_x,type = "l",
  ylim=c(min(fitted_mxt_x,projected_mxt_x_0.05,projected_mxt_x_99.5,
  projected_mxt_x,raw_mort[i-40,]),max(fitted_mxt_x,
  projected_mxt_x_0.05,projected_mxt_x_99.5,projected_mxt_x,raw_mort[i-40,])),
  xlim=c(1957, 2026),xlab="Year", ylab="Death probability",
  main = paste ("Lee-Carter fitted and projected yearly death probability for age",
  i-1))
  points(1957:2016,raw_mort[i-40,])
  lines(2017:2026,projected_mxt_x, col="red")
  lines(2017:2026,projected_mxt_x_0.05, col="red", lty="dashed")
  lines(2017:2026,projected_mxt_x_99.5, col="red", lty="dashed")
}

#3D surface Full (40:100) actual values.
par(mfrow=c(1,1))
persp(ages[1:61], years, raw_mort[1:61,], phi = 30, theta = -30, col = "white",
xlab = "Age", ylab = "Year" , zlab ="Death probablility",
main = "Lee-Carter observed death probabilities surface per year and ages 40 to 100")

#3D surface Full (40:100) fitted values.
par(mfrow=c(1,1))
persp(ages[1:61], years, mxtHat_NL[1:61,], phi = 30, theta = -30, col = "white",
xlab = "Age", ylab = "Year" , zlab ="Death probablility",
main = "Lee-Carter fitted death probabilities surface per year and ages 40 to 100")

#3D surface 40:89 fitted values.
par(mfrow=c(1,1))
persp(ages[1:50], years, mxtHat_NL[1:50,], phi = 30, theta = -30, col = "white",
xlab = "Age", ylab = "Year" , zlab ="Death probablility",
main = "Lee-Carter fitted death probabilities surface per year and ages 40 to 89")

#3D surface 90:111 (old) fitted values.
par(mfrow=c(1,1))
persp(ages[51:61], years, mxtHat_NL[51:61,], phi = 30, theta = -30, col = "white",
xlab = "Age", ylab = "Year" , zlab ="Death probablility" ,
main = "Lee-Carter fitted death probabilities surface per year and ages 90 to 100")
```

C.4 CBD

```
#Ages are restricted for the range between 40-90, since the CBD model
#will only be applied in this range and also only the time frame
#between 1957-2016.

Dth_2=Dth_1[41:90,48:107]
Exp_2=Exp_1[41:90,48:107]
lx_2=lx_1[41:90,48:107]

death_2=Dth_2
exposure_2=Exp_2
Age_2 <- 40:89
Year_2 <- 1957:2016

ages_2 <- 40:89
years_2 <- 1957:2016

#The data is introduced so that it can be recognized by the StMoMo library,
#making use of the EWMaleData object which is already built in.

NLData_2=NLData
NLData_2$years=years_2
NLData_2$ages=ages_2
NLData_2$Dxt=Dth_2
NLData_2$Ext=lx_2
NLData_2$series="total"
NLData_2$label="Netherlands"


#MODEL FITTING

#The CBD model is fitted assuming a Poisson dustribution, so the link is log.
#We use the "fit" function of the StMoMo library.
#The only argument to specify is the link function and the data base.
```

```
CBDfit_NL <- fit(link="log",cbd(), data =NLData_2)

#The mortality rates estimated by the Lee-Carter model are obtained
#using the "fitted" function whose argument is an object that contains
#the model fit.
#Since the type = "rates" we get directly the fitted probablities.
#If the type = "link" we would get directly the outcome of the alpha + (beta*Kappa).

CBDmxtHat_NL <- fitted(CBDfit_NL, type = "rates")

#We use a bi-variate random walk with drift to project the rates for
#the future years (in this case 10). The argument h = 10 indicates the
#number of years we want to predict.
#For the forecast function, when the argument is an object of class
#"fitStMoMo" by default adjusts a random wal with drift.

CBDfor_NL=forecast(CBDfit_NL,h=10)

#The projected mortality is obtained.

CBDmxtCentral_NL <- CBDfor_NL$rates


#INFERENCE INCLUDING UNCERTAINTY ABOUT THE PARAMETERS

#1000 values of a Poisson distribution are generated with the
#parameters obtained when fitting the model to the data.
#Once the samples are generated, a Lee-Carter model is fitted to
#each simulation, obtaining the model parameters for each.

CBDboot_NL=bootstrap(CBDfit_NL,nBoot=1000,type="semiparametric")

#With the parameters that have been obtained, a random walk with drift
#is used to make a porjection the rates for the future years (10 in this case).

CBDsimPU_NL=simulate(CBDboot_NL,h=10)


#CALCULATION OF CONFIDENCE INTERVALS AT 99.5%

#Among the simulations above we calculate the curves such that
#the results are between 99.5% of them.

#Calculation of the intervals for the sample.

CBDmxtHatPU0.05_NL <- apply(CBDsimPU_NL$fitted, c(1, 2), quantile, probs = 0.005)
CBDmxtHatPU99.5_NL <- apply(CBDsimPU_NL$fitted, c(1, 2), quantile, probs = 0.995)
```

```
#Calculation of the intervals for the predictions.

CBDmxtPredPU0.05_NL <- apply(CBDsimPU_NL$rates, c(1, 2), quantile, probs = 0.005)
CBDmxtPredPU99.5_NL <- apply(CBDsimPU_NL$rates, c(1, 2), quantile, probs = 0.995)

raw_mort_2 <- NLData_2$Dxt/NLData_2$Ext
colnames(raw_mort_2) <- years_2
rownames(raw_mort_2) <- ages_2


#GRAPHS

#40-49 -> for(i in 41:50) | 50-59 -> for(i in 51:60) | 60-69 -> for(i in 61:70) |
#70-79 -> for(i in 71:80) | 80-89 -> for(i in 81:90) | 90-99 (old) ->
#for(i in 91:100)

par(mfrow=c(5,2))
for(i in 41:50) {
  fitted_mxt_x_CBD<-CBDmxtHat_NL[i-40,]
  projected_mxt_x_CBD<-CBDmxtCentral_NL[i-40,]
  projected_mxt_x_0.05_CBD<-CBDmxtPredPU0.05_NL[i-40,]
  projected_mxt_x_99.5_CBD<-CBDmxtPredPU99.5_NL[i-40,]
  plot(1957:2016,fitted_mxt_x_CBD,type = "l",
  ylim=c(min(fitted_mxt_x_CBD,projected_mxt_x_0.05_CBD,
  projected_mxt_x_99.5_CBD,projected_mxt_x_CBD,raw_mort_2[i-40,]),
  max(fitted_mxt_x_CBD,projected_mxt_x_0.05_CBD,projected_mxt_x_99.5_CBD,
  projected_mxt_x_CBD,raw_mort_2[i-40,])),xlim=c(1957, 2026),
  xlab="Year", ylab="Death probability",
  main = paste ("CBD fitted and projected yearly death probability for age", i-1))
  points(1957:2016,raw_mort_2[i-40,])
  lines(2017:2026,projected_mxt_x_CBD, col="red")
  lines(2017:2026,projected_mxt_x_0.05_CBD, col="red", lty="dashed")
  lines(2017:2026,projected_mxt_x_99.5_CBD, col="red", lty="dashed")
}

#GRAPHS (SNIPS) TO INCLUDE IN THE DOCUMENT.

#Fit for ages 64-66
par(mfrow=c(3,1))
for(i in 65:67) {
  fitted_mxt_x_CBD<-CBDmxtHat_NL[i-40,]
  projected_mxt_x_CBD<-CBDmxtCentral_NL[i-40,]
  projected_mxt_x_0.05_CBD<-CBDmxtPredPU0.05_NL[i-40,]
  projected_mxt_x_99.5_CBD<-CBDmxtPredPU99.5_NL[i-40,]
  plot(1957:2016,fitted_mxt_x_CBD,type = "l",
  ylim=c(min(fitted_mxt_x_CBD,projected_mxt_x_0.05_CBD,
  projected_mxt_x_99.5_CBD,projected_mxt_x_CBD,raw_mort_2[i-40,]),
  max(fitted_mxt_x_CBD,projected_mxt_x_0.05_CBD,projected_mxt_x_99.5_CBD,
  projected_mxt_x_CBD,raw_mort_2[i-40,])),xlim=c(1957, 2026),
  xlab="Year", ylab="Death probability",
  main = paste ("CBD fitted and projected yearly death probability for age", i-1))
  points(1957:2016,raw_mort_2[i-40,])
  lines(2017:2026,projected_mxt_x_CBD, col="red")
  lines(2017:2026,projected_mxt_x_0.05_CBD, col="red", lty="dashed")
  lines(2017:2026,projected_mxt_x_99.5_CBD, col="red", lty="dashed")
}

#3D surface Full (40:89) actual values
par(mfrow=c(1,1))
persp(ages_2, years_2, raw_mort_2, phi = 30, theta = -30, col = "white",
xlab = "Age", ylab = "Year" , zlab ="Death probability",
main = "Observed death probabilities surface per year and ages 40 to 89")

#3D surface 40:89
par(mfrow=c(1,1))
persp(ages_2, years_2, CBDmxtHat_NL, phi = 30, theta = -30, col = "white",
xlab = "Age", ylab = "Year" , zlab ="Death probability",
main = "CBD fitted death probabilities surface per year and ages 40 to 89")
```

C.5 P-splines

```
library(MortalitySmooth)

# We call the function my.predict which is a slight modification of the
#predict function that is included in R, which leads to tighter
#confidence intervals.

source("my.predict.R")

#Find the optimal combination of nodes minimizing the BIC.

model_bic = c()
nodos_edad = c()
nodos_yrs = c()

for(j in 2:15){
  for(i in 2:15){
    fitBIC3 <- Mort2Dsmooth(x=ages_2, y=years_2, Z=death_2 ,
    offset=log(exposure_2),ndx=c(i,j))
    model_bic[length(model_bic)+1]  = fitBIC3$bic
    nodos_edad[length(nodos_edad)+1]  = i
    nodos_yrs[length(nodos_yrs)+1]  = j
  }
}
results <- cbind(nodos_edad,nodos_yrs,model_bic)
min(model_bic) #Corresponds to 6 for the age and 13 for the years.

#We fit a two-dimensional P-spline model with 3 nodes for age and 13 for year.
#In this case we are working with a Poisson distribution
#To fit the model we use the function "Mort2Dsmooth" whose arguments are
#the ages, the years, the matrix of deaths and initial risk exposure.
#It is necessary to choose correclty the number of nodes for the age and years.
#This could be something complex because the resulting projections could be odd.
#Between 10 and 15 generally yields a good forecast.

fitBIC3 <- Mort2Dsmooth(x=ages_2, y=years_2, Z=death_2 ,offset=log(exposure_2),
ndx=c(6,13))

#We define the range of years to predict.

newyears <- 1957:2026

#We create the new dataset on which we want to make predictions.

newdata <- list(x=ages_2, y=newyears)

# We calculate predictions employing the predict and my.predict functions,
#the first one is used to obtain the prediction itself while the second one
#gives the standard error of the prediction.
# The se.fit argument allows us to object to standard errors for the
#linear predictor.

pre.for3 <- my.predict(fitBIC3, newdata=newdata, se.fit=TRUE)
pre.for2=predict(fitBIC3, newdata=newdata, se.fit=TRUE)

#mx of the P-Spline and its transformaci n to qx

PSqxtHat_NL <-exp(pre.for3$fit)
PSqxtHat_NL=1-exp(-PSqxtHat_NL)


#CALCULATION OF CONFIDENCE INTERVALS AT 99.5%

pre.for3_0.05_NL=exp(pre.for3$fit-2.57*pre.for3$se.fit)
pre.for3_0.05_NL=1-exp(-pre.for3_0.05_NL)
pre.for3_99.5_NL=exp(pre.for3$fit+2.57*pre.for3$se.fit)
pre.for3_99.5_NL=1-exp(-pre.for3_99.5_NL)
```

```
#GRAPHS

#40-49 -> for(i in 41:50) | 50-59 -> for(i in 51:60) | 60-69 -> for(i in 61:70) |
#70-79 -> for(i in 71:80) | 80-89 -> for(i in 81:90) | 90-99 (old) ->
#for(i in 91:100)

par(mfrow=c(5,2))
for(i in 41:50) {
  plot(years_2,raw_mort_2[i-40,], xlim = c(1957,2026),
  ylim = c(min(PSqxtHat_NL[i-40,],pre.for3_0.05_NL[i-40,61:70]),
  max(raw_mort_2[i-40,])),
  xlab="Year", ylab="Death probability",
  main = paste ("P-splines fitted and projected yearly death probability for age",
  i-1))
  lines(years_2, PSqxtHat_NL[i-40,1:60])
  lines(2017:2026,PSqxtHat_NL[i-40, 61:70], col="red")
  lines(2017:2026,pre.for3_0.05_NL[i-40,61:70], col="red", lty="dashed")
  lines(2017:2026,pre.for3_99.5_NL[i-40,61:70], col="red", lty="dashed")
}


#GRAPHS (SNIPS) TO INCLUDE IN THE DOCUMENT.

#Fit for ages 64-66
par(mfrow=c(3,1))
for(i in 65:67) {
  plot(years_2,raw_mort_2[i-40,], xlim = c(1957,2026),
  ylim = c(min(PSqxtHat_NL[i-40,],pre.for3_0.05_NL[i-40,61:70]),
  max(raw_mort_2[i-40,])),
  xlab="Year", ylab="Death probability",
  main = paste ("P-splines fitted and projected yearly death probability for age",
  i-1))
  lines(years_2, PSqxtHat_NL[i-40,1:60])
  lines(2017:2026,PSqxtHat_NL[i-40, 61:70], col="red")
  lines(2017:2026,pre.for3_0.05_NL[i-40,61:70], col="red", lty="dashed")
  lines(2017:2026,pre.for3_99.5_NL[i-40,61:70], col="red", lty="dashed")
}

#3D surface fitted values 40:89
par(mfrow=c(1,1))
persp(ages_2, years_2, PSqxtHat_NL[,1:60], phi = 30, theta = -30, col = "white",
xlab = "Age", ylab = "Year" , zlab ="Death probability",
main = "P-splines fitted mortality rate surface per year and ages 40 to 89")
```

C.6 Comparative Graphs

```
par(mfrow=c(1,1))
i <- 61
fitted_mxt_x<-mxtHat_NL[i-40,]
projected_mxt_x<-mxtCentral_NL[i-40,]
projected_mxt_x_0.05<-mxtPredPU0.05_NL[i-40,]
projected_mxt_x_99.5<-mxtPredPU99.5_NL[i-40,]
plot(1957:2016,fitted_mxt_x,type = "l",
col="gold",ylim=c(min(fitted_mxt_x,projected_mxt_x_0.05,
projected_mxt_x_0.05_CBD,pre.for3_0.05_NL[i-40,61:70],
projected_mxt_x,raw_mort[i-40,]),
max(fitted_mxt_x,projected_mxt_x_0.05,projected_mxt_x_99.5,
projected_mxt_x,raw_mort[i-40,])),xlim=c(1957, 2026),
xlab="Year", ylab="Death probability",
main = paste ("Lee-Carter, CBD and P-splines fitted and projected yearly
death probability for age", i-1))
points(1957:2016,raw_mort[i-40,])
lines(2017:2026,projected_mxt_x, col="gold")
lines(2017:2026,projected_mxt_x_0.05, col="gold", lty="dashed")
lines(2017:2026,projected_mxt_x_99.5, col="gold", lty="dashed")

fitted_mxt_x_CBD<-CBDmxtHat_NL[i-40,]
projected_mxt_x_CBD<-CBDmxtCentral_NL[i-40,]
projected_mxt_x_0.05_CBD<-CBDmxtPredPU0.05_NL[i-40,]
projected_mxt_x_99.5_CBD<-CBDmxtPredPU99.5_NL[i-40,]
lines(1957:2016,fitted_mxt_x_CBD, col="orange")
lines(2017:2026,projected_mxt_x_CBD, col="orange")
lines(2017:2026,projected_mxt_x_0.05_CBD, col="orange", lty="dashed")
lines(2017:2026,projected_mxt_x_99.5_CBD, col="orange", lty="dashed")

lines(years_2, PSqxtHat_NL[i-40,1:60], col="tomato3")
lines(2017:2026,PSqxtHat_NL[i-40, 61:70], col="tomato3")
lines(2017:2026,pre.for3_0.05_NL[i-40,61:70], col="tomato3", lty="dashed")
lines(2017:2026,pre.for3_99.5_NL[i-40,61:70], col="tomato3", lty="dashed")

legend(2005,0.012,legend=c("Lee-Carter","CBD","P-splines"),cex=0.9,lty=3,
col=c("gold","orange","tomato3"),bty = "n")
```

C.7 In sample R^2

```
#Lee-Carter Qx.
Hat_LC=mxtHat_NL[1:50,]
```

```
#CBD Qx.
Hat_CBD=CBDmxtHat_NL

#P-Spline mx and its transformation to Qx.
Hat_SP=exp(pre.for3$fit[,1:60])
Hat_SP=1-exp(-Hat_SP)


#R2 AGES 40-89

#Average Qx ages 40-89.

MEAN_qx= (Hat_LC+ Hat_SP+ Hat_CBD)/3 #Change depending on the combinations
to test: Hat_LC | Hat_CBD | Hat_SP
MEAN_Qx=c(MEAN_qx)

#Qx from the observed data.

Qx=(Dth/lx)
rownames(Qx)<-ages
colnames(Qx)<-years
Qx_young=Qx[1:50,]
Qx_young=c(Qx_young)

#We calculate the R^2 between the logarithms of the raw qx and the average qx.
#To do so we calculate the Squared sum of residuals.

RSS_young=sum((c(log(Qx_young))-c(log(MEAN_Qx)))^2)

#The squared sum of the differences between the estimations and the mean real value.

TSS_young=sum((c(log(Qx_young))-mean(c(log(Qx_young))))^2)

#Calculation of the R^2

R2_young=1-RSS_young/TSS_young

##R2 AGES 89-99

Hat_LC_old=mxtHat_NL[51:60,]
Hat_LC_old=c(Hat_LC_old)

Qx_old=Qx[51:60,]
Qx_old=c(Qx_old)

RSS_old=sum((c(log(Qx_old))-c(log(Hat_LC_old)))^2)
TSS_old=sum((c(log(Qx_old))-mean(c(log(Hat_LC_old))))^2)
```

```
R2_old=1-RSS_old/TSS_old


##R2 ALL AGES (40-99)

#Average Qx for all ages.

MEAN_Qx_all=rbind(MEAN_qx,mxtHat_NL[51:60,])
MEAN_Qx_all=c(MEAN_Qx_all)

#Vectorized Qx from the observed data.

Qx_all=Qx[1:60,]
Qx_all=c(Qx_all)

RSS_all=sum((c(log(Qx_all))-c(log(MEAN_Qx_all)))^2)
TSS_all=sum((c(log(Qx_all))-mean(c(log(Qx_all))))^2)
R2_all=1-RSS_all/TSS_all
```

C.8 Backtesting

```
#R2 P-SPLINE BACKTESTING.

#P-spline ommitting the last 10 years.

library(MortalitySmooth)

fitBIC4 <- Mort2Dsmooth(x=ages_2, y=years_2[1:50], Z=death_2[,1:50] ,
offset=log(exposure_2[,1:50]),ndx=c(6,13))
newyears2 <- 1957:2016
newdata2 <- list(x=ages_2, y=newyears2)
pre.for5 <- my.predict(fitBIC4, newdata=newdata2, se.fit=TRUE)
pre.for4=predict(fitBIC4, newdata=newdata2, se.fit=TRUE)

#P-Spline mx transformed to Qx.

PSqxtHat_NL2 <-exp(pre.for5$fit)
PSqxtHat_NL2=1-exp(-PSqxtHat_NL2)

#Confidence intervals

pre.for5_0.05_NL=exp(pre.for5$fit-2.57*pre.for5$se.fit)
pre.for5_0.05_NL=1-exp(-pre.for5_0.05_NL)
pre.for5_99.5_NL=exp(pre.for5$fit+2.57*pre.for5$se.fit)
pre.for5_99.5_NL=1-exp(-pre.for5_99.5_NL)

#R2 backtesting.

Pred_Qx=c(PSqxtHat_NL2[,51:60])
Obs_Qx=c(raw_mort_2[,51:60])

RSS_fore_PS=sum((c(log(Obs_Qx))-c(log(Pred_Qx)))^2)
TSS_fore_PS=sum((c(log(Obs_Qx))-mean(c(log(Obs_Qx))))^2)
R2_fore_PS=1-RSS_fore_PS/TSS_fore_PS


#R2 LEE-CARTER BACKTESTING.

library(StMoMo)

NLData_3=EWMaleData
NLData_3$years=years[1:50]
NLData_3$ages=ages[1:60]
NLData_3$Dxt=Dth[1:60,1:50]
NLData_3$Ext=lx[1:60,1:50]
NLData_3$series="total"
NLData_3$label="Netherlands"

LCfit_NL_2 <- fit(link="log",lc(), data = NLData_3)
mxtHat_NL_2 <- fitted(LCfit_NL_2, type = "rates")
LCfor_NL_2=forecast(LCfit_NL_2,h=10)
mxtCentral_NL_2 <- LCfor_NL_2$rates

#Confidence intervals

LCboot_NL_2=bootstrap(LCfit_NL_2,nBoot=1000,type="semiparametric")
LCsimPU_NL_2=simulate(LCboot_NL_2,h=10)
mxtPredPU0.05_NL <- apply(LCsimPU_NL_2$rates, c(1, 2), quantile, probs = 0.005)
mxtPredPU99.5_NL <- apply(LCsimPU_NL_2$rates, c(1, 2), quantile, probs = 0.995)

#R2 LC All.

MEAN_Qx_LC=c(mxtCentral_NL_2)
Qx_LC=c(raw_mort[1:60,51:60])

RSS_fore_LC_all=sum((c(log(Qx_LC))-c(log(MEAN_Qx_LC)))^2)
TSS_fore_LC_all=sum((c(log(Qx_LC))-mean(c(log(Qx_LC))))^2)
R2_fore_LC_all=1-RSS_fore_LC_all/TSS_fore_LC_all

#R2 LC young.

MEAN_Qx_LC=c(mxtCentral_NL_2[1:50,])
```

```
Qx_LC=c(raw_mort[1:50,51:60])

RSS_fore_LC_young=sum((c(log(Qx_LC))-c(log(MEAN_Qx_LC)))^2)
TSS_fore_LC_young=sum((c(log(Qx_LC))-mean(c(log(Qx_LC))))^2)
R2_fore_LC_young=1-RSS_fore_LC_young/TSS_fore_LC_young

#R2 LC old.

MEAN_Qx_LC=c(mxtCentral_NL_2[51:60,])
Qx_LC=c(raw_mort[51:60,51:60])

RSS_LC_fore_old=sum((c(log(Qx_LC))-c(log(MEAN_Qx_LC)))^2)
TSS_LC_fore_old=sum((c(log(Qx_LC))-mean(c(log(Qx_LC))))^2)
R2_LC_fore_old=1-RSS_LC_fore_old/TSS_LC_fore_old


#R2 CBD BACKTESTING.

NLData_4=NLData
NLData_4$years=years_2[1:50]
NLData_4$ages=ages_2
NLData_4$Dxt=Dth_2[,1:50]
NLData_4$Ext=lx_2 [,1:50]
NLData_4$series="total"
NLData_4$label="Netherlands"

CBDfit_NL_2 <- fit(link="log",cbd(), data =NLData_4)
CBDmxtHat_NL_2 <- fitted(CBDfit_NL_2, type = "rates")
CBDfor_NL_2=forecast(CBDfit_NL_2,h=10)
CBDmxtCentral_NL_2 <- CBDfor_NL_2$rates

#Confidence intervals

CBDboot_NL_2=bootstrap(CBDfit_NL_2,nBoot=1000,type="semiparametric")
CBDsimPU_NL_2=simulate(CBDboot_NL_2,h=10)
CBDmxtPredPU0.05_NL <- apply(CBDsimPU_NL_2$rates, c(1, 2), quantile, probs = 0.005)
CBDmxtPredPU99.5_NL <- apply(CBDsimPU_NL_2$rates, c(1, 2), quantile, probs = 0.995)

#R2 CBD young.

MEAN_Qx_CBD=c(CBDmxtCentral_NL_2)
Qx_CBD=c(raw_mort[1:50,51:60])

RSS_fore_CBD=sum((c(log(Qx_CBD))-c(log(MEAN_Qx_CBD)))^2)
TSS_fore_CBD=sum((c(log(Qx_CBD))-mean(c(log(Qx_CBD))))^2)
R2_fore_CBD=1-RSS_fore_CBD/TSS_fore_CBD
```

```
#R2 COMBINED BACKTESTING

Hat_LC=mxtCentral_NL_2[1:50,]
Hat_CBD=CBDmxtCentral_NL_2
Hat_SP=PSqxtHat_NL2[,51:60]

#R2 ages 40-89.

MEAN_qx= (Hat_CBD+Hat_SP+Hat_LC)/3 #Change depending on the combinations
to test: (Hat_LC+Hat_CBD+Hat_SP)/3 | (Hat_LC+Hat_CBD)/2 | (Hat_LC+Hat_SP)/2 |
(Hat_CBD+Hat_SP)/2 | Hat_LC | Hat_CBD | Hat_SP
MEAN_Qx=c(MEAN_qx)

Qx_young=c(raw_mort[1:50,51:60])
RSS_young=sum((c(log(Qx_young))-c(log(MEAN_Qx)))^2)
TSS_young=sum((c(log(Qx_young))-mean(c(log(Qx_young))))^2)
R2_young=1-RSS_young/TSS_young

##R2 ages 40-99

MEAN_Qx_all=rbind(MEAN_qx,mxtCentral_NL_2[51:60,])
MEAN_Qx_all=c(MEAN_Qx_all)

Qx_all=raw_mort[1:60,51:60]
Qx_all=c(Qx_all)

RSS_all=sum((c(log(Qx_all))-c(log(MEAN_Qx_all)))^2)
TSS_all=sum((c(log(Qx_all))-mean(c(log(Qx_all))))^2)
R2_all=1-RSS_all/TSS_all


#GRAPH TO COMPARE FORECASTS

library(yarrr)

par(mfrow=c(5,2))
for(i in 61:70) {
  plot(years_2,raw_mort_2[i-40,], xlim = c(1957,2016),
  ylim = c(min(pre.for5_0.05_NL[i-40,51:60]),max(PSqxtHat_NL2[i-40,],
  raw_mort_2[i-40,])),xlab="Year", ylab="Death probability",
  main = paste ("Lee-Carter, CBD and P-splines fitted and projected yearly
  fit and backtesting for age", i-1))

  lines(years_2[1:50], PSqxtHat_NL2[i-40,1:50], col="gold")
  #lines(2007:2016,PSqxtHat_NL2[i-40,51:60], col="gold")
  lines(2007:2016,pre.for5_0.05_NL[i-40,51:60], col="gold",lty="dashed")
  lines(2007:2016,pre.for5_99.5_NL[i-40,51:60], col="gold",lty="dashed")
  polygon(c(2006,2007:2016,rev(2007:2016)),c(PSqxtHat_NL2[i-40,50],
  pre.for5_0.05_NL[i-40,51:60],rev(pre.for5_99.5_NL[i-40,51:60])),
  col=transparent(orig.col = "gold", trans.val = 0.8), border = NA)

  lines(years_2[1:50], mxtHat_NL_2[i-40,1:50], col="orange")
  #lines(2007:2016,mxtCentral_NL_2[i-40,], col="orange")
  lines(2007:2016,mxtPredPU0.05_NL[i-40,], col="orange",lty="dashed")
  lines(2007:2016,mxtPredPU99.5_NL[i-40,], col="orange",lty="dashed")
  polygon(c(2006,2007:2016,rev(2007:2016)),c(mxtHat_NL_2[i-40,50],
  mxtPredPU0.05_NL[i-40,],rev(mxtPredPU99.5_NL[i-40,])),
  col=transparent(orig.col = "orange", trans.val = 0.8), border = NA)

  lines(years_2[1:50], CBDmxtHat_NL_2[i-40,1:50], col="tomato3")
  #lines(2007:2016,CBDmxtCentral_NL_2[i-40,], col="tomato3")
  lines(2007:2016,CBDmxtPredPU0.05_NL[i-40,], col="tomato3",lty="dashed")
  lines(2007:2016,CBDmxtPredPU99.5_NL[i-40,], col="tomato3",lty="dashed")
  polygon(c(2006,2007:2016,rev(2007:2016)),c(CBDmxtHat_NL_2[i-40,50],
  CBDmxtPredPU0.05_NL[i-40,],rev(CBDmxtPredPU99.5_NL[i-40,])),
  col=transparent(orig.col = "tomato3", trans.val = 0.8), border = NA)

  lines(2007:2016,(PSqxtHat_NL2[i-40,51:60]+
  mxtCentral_NL_2[i-40,]+CBDmxtCentral_NL_2[i-40,])/3, col="deeppink3")

  lines(2007:2016,(PSqxtHat_NL2[i-40,51:60]+
  CBDmxtCentral_NL_2[i-40,])/2, col="mediumorchid1")

  lines(2007:2016,(PSqxtHat_NL2[i-40,51:60]+
  mxtCentral_NL_2[i-40,])/2, col="lightpink")

  lines(2007:2016,(mxtCentral_NL_2[i-40,]+
  CBDmxtCentral_NL_2[i-40,])/2, col="darkmagenta")

  legend(1995,0.013,legend=c("P-splines","Lee-Carter","CBD",
  "Lee-Carter+CBD+P-splines","CBD+P-splines", "Lee-Carter+P-splines",
  "Lee-Carter+CBD"),cex=0.9,lty=3,col=c("gold","orange","tomato3","deeppink3",
  "mediumorchid1","lightpink","darkmagenta"),bty = "n")
}

#GRAPHS (SNIPS) TO INCLUDE IN THE DOCUMENT.

#Fit for age 57-59
par(mfrow=c(3,1))
for(i in 58:60) {
  plot(years_2,raw_mort_2[i-40,], xlim = c(1957,2016),
  ylim = c(min(PSqxtHat_NL2[i-40,]),max(PSqxtHat_NL2[i-40,],raw_mort_2[i-40,])),
  xlab="Year", ylab="Death probability",
  main = paste ("Lee-Carter, CBD and P-splines fitted and projected yearly
  fit and backtesting for age", i-1))
  lines(years_2[1:50], PSqxtHat_NL2[i-40,1:50], col="gold")
  #lines(2007:2016,PSqxtHat_NL2[i-40,51:60], col="gold")
  lines(2007:2016,pre.for5_0.05_NL[i-40,51:60], col="gold",lty="dashed")
  lines(2007:2016,pre.for5_99.5_NL[i-40,51:60], col="gold",lty="dashed")
  polygon(c(2006,2007:2016,rev(2007:2016)),c(PSqxtHat_NL2[i-40,50],
  pre.for5_0.05_NL[i-40,51:60],rev(pre.for5_99.5_NL[i-40,51:60])),
  col=transparent(orig.col = "gold", trans.val = 0.8), border = NA)

  lines(years_2[1:50], mxtHat_NL_2[i-40,1:50], col="orange")
  #lines(2007:2016,mxtCentral_NL_2[i-40,], col="orange")
  lines(2007:2016,mxtPredPU0.05_NL[i-40,], col="orange",lty="dashed")
  lines(2007:2016,mxtPredPU99.5_NL[i-40,], col="orange",lty="dashed")
  polygon(c(2006,2007:2016,rev(2007:2016)),c(mxtHat_NL_2[i-40,50],
  mxtPredPU0.05_NL[i-40,],rev(mxtPredPU99.5_NL[i-40,])),
  col=transparent(orig.col = "orange", trans.val = 0.8), border = NA)

  lines(years_2[1:50], CBDmxtHat_NL_2[i-40,1:50], col="tomato3")
  #lines(2007:2016,CBDmxtCentral_NL_2[i-40,], col="tomato3")
  lines(2007:2016,CBDmxtPredPU0.05_NL[i-40,], col="tomato3",lty="dashed")
  lines(2007:2016,CBDmxtPredPU99.5_NL[i-40,], col="tomato3",lty="dashed")
  polygon(c(2006,2007:2016,rev(2007:2016)),c(CBDmxtHat_NL_2[i-40,50],
  CBDmxtPredPU0.05_NL[i-40,],rev(CBDmxtPredPU99.5_NL[i-40,])),
  col=transparent(orig.col = "tomato3", trans.val = 0.8), border = NA)

  legend(1997,0.022,legend=c("Lee-Carter","CBD","P-splines"),
  cex=0.9,lty=3,col=c("gold","orange","tomato3"),bty = "n")
}
```

C.9 Sample Information

```
#Import sample data

qx_todos <-as.data.frame(read.delim("Todos.txt", header=T))
qx_todos<-qx_todos[!is.na(qx_todos)]
qx_todos<-matrix(qx_todos,109,10)
row.names(qx_todos)<-0:108
qx_todos <- qx_todos[41:95,]
colnames(qx_todos) <- c("2007", "2008", "2009", "2010", "2011", "2012",
"2013", "2014", "2015", "2016")

exp_mujeres <-as.data.frame(read.delim("ExpM.txt", header=T))
exp_mujeres<-exp_mujeres[!is.na(exp_mujeres)]
exp_mujeres<-matrix(exp_mujeres,109,10)
row.names(exp_mujeres)<-0:108
```

```
exp_mujeres <- exp_mujeres[41:95,]

exp_hombres <-as.data.frame(read.delim("ExpH.txt", header=T))
exp_hombres<-exp_hombres[!is.na(exp_hombres)]
exp_hombres<-matrix(exp_hombres,109,10)
row.names(exp_hombres)<-0:108
exp_hombres <- exp_hombres[41:95,]  #41:66

Exposures <- exp_mujeres + exp_hombres
colnames(Exposures)<-c("2007", "2008", "2009", "2010", "2011", "2012",
"2013", "2014", "2015", "2016")

Claims <- Exposures*qx_todos
```

C.10 Sample EDA

```
#GRAPHS (SNIPS) TO INCLUDE IN THE DOCUMENT.

#EDA 1
par(mfrow=c(1,1))
a<- colSums(Exposures)
b<- barplot(a, main="Portfolio size per year",xlab="Year",
ylab = "Frequency", col = "white", ylim=c(0, 1.1*max(a)))
text(x = b, y = a, label = c("246.780", "246.782", "246.793", "246.776",
"246.791", "246.783", "290.520", "347.890", "424.090", "494.670"),pos = 3,
cex = 0.7, col = "red")

mean(494670-424090,424090-347890,347890-290520)

#EDA 2
par(mfrow=c(1,1))
a<- rowSums(Exposures)
b<- barplot(a, main="Portfolio composition by age",xlab="Year",
ylab = "Frequency", col = "white", ylim=c(0, 1.1*max(a)))
text(x = b, y = a, label = c("185.720","189.770","180.300","184.740","180.310",
"169.330","156.000","150.200","147.080","138.850","136.730","129.690","126.950",
"115.550","113.770","104.730","95.370", "83.330","79.040","68.040","56.610",
"49.830","39.090","30.480","23.540","18.310","15.890","12.950","10.120",
"7.780","6.720","5.460","4.790","2.900","2.170","2.050","3.000","1.090",
"1.430", "1.300","1.160","720","1.020","950","710","420","400","420","400",
"210","180","130","50","40","30"),pos = 3, cex = 0.5, col = "red")

#EDA 3
par(mfrow=c(1,1))
```

```
a<- colSums(Claims)
b<- barplot(a, main="Deaths per year",xlab="Year",ylab = "Frequency",
col = "white", ylim=c(0, 1.1*max(round(a,0))))
text(x = b, y = a, label = round(a,0),pos = 3, cex = 0.7, col = "red")

#EDA 4
par(mfrow=c(1,1))
a<- rowSums(Claims)
b<- barplot(a, main="Deaths by age",xlab="Year",ylab = "Frequency",
col = "white", ylim=c(0, 1.1*max(a)))
text(x = b, y = a, label = round(a,0),pos = 3, cex = 0.7, col = "red")

#Population Vs Company data ages 64-66
par(mfrow=c(3,1))
for(i in 65:67) {
  plot(years_2,raw_mort_2[i-40,], xlim = c(1957,2016),
  ylim = c(min(qx_todos[i-41,]),max(raw_mort_2[i-40,])),xlab="Year",
  ylab="Death probability", main = paste ("Population vs. company data for age",
  i-1))
  lines(2007:2016,qx_todos[i-40,], col="red", type = "p")
}
```

C.11 Piggy-Back

```
#GRAPHS (SNIPS) TO INCLUDE IN THE DOCUMENT.

#Theory
par(mfrow=c(3,1))
  i=66
  plot(years_2,raw_mort_2[i-40,], xlim = c(1957,2026),
  ylim = c(min(mxtCentral_NL[i-40,]-(raw_mort_2[i-40,51+9]-qx_todos[i-40,10])),
  max(raw_mort_2[i-40,])),xlab="Year", ylab="Death probability",
  main = "Piggy-Back step 1")
  lines(2017:2026,mxtCentral_NL[i-40,], col="orange",
  type = "l", lty="dashed")
  lines(2007:2016,qx_todos[i-40,], col="red", type = "p")

  plot(years_2,raw_mort_2[i-40,], xlim = c(1957,2026),
  ylim = c(min(mxtCentral_NL[i-40,]-(raw_mort_2[i-40,51+9]-qx_todos[i-40,10])),
  max(raw_mort_2[i-40,])),xlab="Year", ylab="Death probability",
  main = "Piggy-Back step 2")
  lines(2017:2026,mxtCentral_NL[i-40,], col="orange", type = "l",
  lty="dashed")
  lines(2007:2016,qx_todos[i-40,], col="red", type = "p")
  lines(c(2007,2007),c(raw_mort_2[i-40,51],qx_todos[i-40,1]),
  lty="dashed", col="gray36")
  text(2006,0.007, expression(Delta[1]), cex=2, col="gray36")
  lines(c(2012,2012),c(raw_mort_2[i-40,51+5],qx_todos[i-40,6]),
  lty="dashed", col="gray36")
  text(2011,0.007, expression(Delta[5]), cex=2, col="gray36")
  lines(c(2016,2016),c(raw_mort_2[i-40,51+9],qx_todos[i-40,10]),
  lty="dashed", col="gray36")
  text(2014.7,0.007, expression(Delta[10]), cex=2, col="gray36")

  plot(years_2,raw_mort_2[i-40,], xlim = c(1957,2026),
  ylim = c(min(mxtCentral_NL[i-40,]-(raw_mort_2[i-40,51+9]-qx_todos[i-40,10])),
  max(raw_mort_2[i-40,])),xlab="Year", ylab="Death probability",
  main = "Piggy-Back step 3")
  lines(2017:2026,mxtCentral_NL[i-40,], col="orange", type = "l",
  lty="dashed")
  lines(2007:2016,qx_todos[i-40,], col="red", type = "p")
  lines(2017:2026,mxtCentral_NL[i-40,]-(raw_mort_2[i-40,51+9]-qx_todos[i-40,10]),
  col="orange", type = "l", lty="dashed")
  lines(c(2017,2017),c(mxtCentral_NL[i-40,1],
  mxtCentral_NL[i-40,1]-(raw_mort_2[i-40,51+9]-qx_todos[i-40,10])),
  lty="dashed", col="gray36")
  text(2015.5,0.006, expression(delta["m,1"]), cex=2, col="gray36")
  lines(c(2021,2021),c(mxtCentral_NL[i-40,5],
  mxtCentral_NL[i-40,5]-(raw_mort_2[i-40,51+9]-qx_todos[i-40,10])),
  lty="dashed", col="gray36")
  text(2019.4,0.006, expression(delta["m,5"]), cex=2, col="gray36")
  lines(c(2026,2026),c(mxtCentral_NL[i-40,10],
  mxtCentral_NL[i-40,10]-(raw_mort_2[i-40,51+9]-qx_todos[i-40,10])),
  lty="dashed", col="gray36")
  text(2024.1,0.006, expression(delta["m,10"]), cex=2, col="gray36")



#Assumptions
par(mfrow=c(3,1))
  plot(40:89,log(raw_mort_2[,51]),xlab="Age", ylab="log(mortality)",
  main = "Log(mortaity) per age - Population", type="o", col="gold")
  lines(40:89,log(raw_mort_2[,52]),xlab="Age", ylab="log(mortality)",
  type="o", col= "gold3")
  lines(40:89,log(raw_mort_2[,53]),xlab="Age", ylab="log(mortality)",
  type="o", col= "orange")
  lines(40:89,log(raw_mort_2[,54]),xlab="Age", ylab="log(mortality)",
  type="o", col= "orange2")
  lines(40:89,log(raw_mort_2[,55]),xlab="Age", ylab="log(mortality)",
  type="o", col= "darkorange")
  lines(40:89,log(raw_mort_2[,56]),xlab="Age", ylab="log(mortality)",
  type="o", col= "darkorange3")
  lines(40:89,log(raw_mort_2[,57]),xlab="Age", ylab="log(mortality)",
  type="o", col= "tomato")
  lines(40:89,log(raw_mort_2[,58]),xlab="Age", ylab="log(mortality)",
  type="o", col= "tomato3")
  lines(40:89,log(raw_mort_2[,59]),xlab="Age", ylab="log(mortality)",
  type="o", col= "red")
  lines(40:89,log(raw_mort_2[,60]),xlab="Age", ylab="log(mortality)",
  type="o", col= "red3")

  plot(40:94,log(qx_todos[,1]),xlab="Age", ylab="log(mortality)",
  main = "Log(mortaity) per age - Sample", type="o", col="gold")
  lines(40:94,log(qx_todos[,2]),xlab="Age", ylab="log(mortality)",
  type="o", col= "gold3")
  lines(40:94,log(qx_todos[,3]),xlab="Age", ylab="log(mortality)",
  type="o", col= "orange")
  lines(40:94,log(qx_todos[,4]),xlab="Age", ylab="log(mortality)",
  type="o", col= "orange2")
  lines(40:94,log(qx_todos[,5]),xlab="Age", ylab="log(mortality)",
  type="o", col= "darkorange")
  lines(40:94,log(qx_todos[,6]),xlab="Age", ylab="log(mortality)",
  type="o", col= "darkorange3")
  lines(40:94,log(qx_todos[,7]),xlab="Age", ylab="log(mortality)",
  type="o", col= "tomato")
  lines(40:94,log(qx_todos[,8]),xlab="Age", ylab="log(mortality)",
  type="o", col= "tomato3")
  lines(40:94,log(qx_todos[,9]),xlab="Age", ylab="log(mortality)",
  type="o", col= "red")
  lines(40:94,log(qx_todos[,10]),xlab="Age", ylab="log(mortality)",
  type="o", col= "red3")

  transp<-t(log(qx_todos))
  plot(2007:2016,transp[,1],xlab="Year", ylab="log(mortality)",
  main = "Log(mortaity) per year- Sample", type="o", col="gold",
  ylim=c(min(transp),-6.4))
    for (i in 2:5) {lines(2007:2016,transp[,i],xlab="Age", ylab="log(mortality)",
    type="o", col= "gold")}
    for (i in 6:10) {lines(2007:2016,transp[,i],xlab="Age", ylab="log(mortality)",
    type="o", col= "gold3")}
    for (i in 11:15) {lines(2007:2016,transp[,i],xlab="Age", ylab="log(mortality)",
    type="o", col= "orange")}
    for (i in 16:20) {lines(2007:2016,transp[,i],xlab="Age", ylab="log(mortality)",
    type="o", col= "orange2")}
    for (i in 21:25) {lines(2007:2016,transp[,5],xlab="Age", ylab="log(mortality)",
    type="o", col= "darkorange")}
    for (i in 26:30) {lines(2007:2016,transp[,6],xlab="Age", ylab="log(mortality)",
    type="o", col= "darkorange3")}
    for (i in 31:35) {lines(2007:2016,transp[,7],xlab="Age", ylab="log(mortality)",
    type="o", col= "tomato")}
    for (i in 36:40) {lines(2007:2016,transp[,8],xlab="Age", ylab="log(mortality)",
    type="o", col= "tomato3")}
    for (i in 41:45) {lines(2007:2016,transp[,9],xlab="Age", ylab="log(mortality)",
    type="o", col= "red")}
    for (i in 46:50) {lines(2007:2016,transp[,10],xlab="Age", ylab="log(mortality)",
    type="o", col= "red3")}
    for (i in 51:55) {lines(2007:2016,transp[,10],xlab="Age", ylab="log(mortality)",
    type="o", col= "red4")}


#PIGGY BACK MODEL

X <- cbind(40:89,40:89,40:89,40:89,40:89,40:89,40:89,40:89,40:89,40:89)

library(StMoMo)

mxtHat_NL_link <- fitted(LCfit_NL, type = "link") #We verify that
mxtHat_NL_link=log(mxtHat_NL), therefore mxtHat_NL_link=hat(alpha)+[hat(beta)*kappa]
Estimate.Sheet <- mxtHat_NL_link[1:50,51:60]


GLM_1 <- glm(c(Claims[1:50,]) ~ c(X) + offset(log(c(Exposures[1:50,])))
+ offset(c(Estimate.Sheet)), family = poisson)
summary(GLM_1)

pred<- predict(GLM_1,type="response")

pred_2 <- (pred/c(Exposures[1:50,]))

m<-matrix(pred_2,50,10)
rownames(m)<- c(40:89)
colnames(m) <- c(2007:2016)

J <-cbind(LCfor_NL$rates[1:26,1]+(GLM_1$coefficients[2]*(40:65))/Exposures[,1],
          LCfor_NL$rates[1:26,2]+(GLM_1$coefficients[2]*(40:65))/Exposures[,2],
          LCfor_NL$rates[1:26,3]+(GLM_1$coefficients[2]*(40:65))/Exposures[,3],
          LCfor_NL$rates[1:26,4]+(GLM_1$coefficients[2]*(40:65))/Exposures[,4],
          LCfor_NL$rates[1:26,5]+(GLM_1$coefficients[2]*(40:65))/Exposures[,5],
          LCfor_NL$rates[1:26,6]+(GLM_1$coefficients[2]*(40:65))/Exposures[,6],
          LCfor_NL$rates[1:26,7]+(GLM_1$coefficients[2]*(40:65))/Exposures[,7],
          LCfor_NL$rates[1:26,8]+(GLM_1$coefficients[2]*(40:65))/Exposures[,8],
          LCfor_NL$rates[1:26,9]+(GLM_1$coefficients[2]*(40:65))/Exposures[,9],
          LCfor_NL$rates[1:26,10]+(GLM_1$coefficients[2]*(40:65))/Exposures[,10])

#GRAPHS (SNIPS) TO INCLUDE IN THE DOCUMENT.

#Final Forecast

par(mar = c(4, 4.5, 4, 2))
par(mfrow=c(3,1))
for(i in 64:66) {
  A<-(m[i-40,])
  B <- J[i-40,]-(LCfor_NL$rates[i-40,1]-A[10])

  W<-LCfor_NL$rates[i-40,]-(LCfor_NL$rates[i-40,1]-A[10])

  plot(years_2,raw_mort_2[i-40,], xlim = c(1957,2026),
  ylim = c(min(qx_todos[i-40,],A,B),max(raw_mort_2[i-40,],A,B)),
  xlab="Year", ylab="Death probability",
  main = paste ("Population vs. company data for age", i))
  lines(2007:2016,qx_todos[i-40,], col="red", type = "p")
  lines(2007:2016, A, col="blue", type = "p")
  lines(2017:2026, LCfor_NL$rates[i-40,], col="orange", lty="dashed")
  lines(2017:2026, B, col="red", lty="dashed")
}
```

Bibliography

W. R. Bell. Comparing and assessing time series methods for forecasting age-specific fertility and mortality rates. *Journal of Official Statistics; Stockholm, Issue 13, Num. 3*, pages 279–303, 1997. URL https://search.proquest.com/openview/ba214d335e90f5679d401f22107e665b/1?pq-origsite=gscholar&cbl=105444.

W.D Bell and B.C. Monsell. Using principal components in time series modeling and forecasting of age-specific mortality rates. 1991. URL https://www.semanticscholar.org/paper/Using-principal-components-in-time-series-modeling-Bell-Monsell/c09ce204827cdb06d055037cf0081c498ba2c45f.

N. Brouhns, M. Denuit, and I Van Keilegom. Bootstrapping the poisson log-bilinear model for mortality forecasting. *Scandinavian Actuarial Journal, Vol. 3*, pages 212–224, 2005.

A.J.G. Cairns, D. Blake, and K. Dowd. A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *The Journal of Risk and Insurance, Vol. 73, No. 4*, pages 687–718, 2006.

A. Coale and G. Guo. Revised regional model life tables at very low levels of mortality. *Population Index, Vol. 55*, pages 613–643, 1989. URL https://www.jstor.org/stable/3644567?seq=1s.

I.D. Currie. Smoothing constrained generalized linear models with an application to the lee-carter model. *Statistical Modelling, Vol. 13*, pages 69–93, 2013.

I.D. Currie, M. Durban, and P.H.C. Eilers. Smoothing and forecasting mortality rates. *Statistical Modelling, No. 4*, page 279–298, 2004.

F. Denton, C. Feaver, and B. Spencer. Time series analysis and stochastic forecasting: An econometric study of mortality and life expectancy. *Journal of Population Economics, Vol. 18*, pages 203–227, 2005. URL `https://www.jstor.org/stable/20007956?seq=1`.

M. Durban. Métodos de suavizado eficientes con p-splines. 2003. URL `http://www.est.uc3m.es/durban/esp/web/cursos/Colombia/material/Pspline.pdf`. [Accessed: 2020-03-23].

T. Goicoa, A. Adin, J. Etxeberria, A.F. Militino, and M.D. Ugarte. Flexible bayesian p-splines for smoothing age-specific spatio-temporal mortality patterns. *Statistical methods in medical Research, Vol. 28*, pages 384–403, 2019. URL `https://journals.sagepub.com/doi/abs/10.1177/0962280217726802`.

R. Lee. The lee-carter method for forecasting mortality, with various extensions and applications. *North American Actuarial Journal, Vol. 4*, pages 80–91, 2000. URL `https://www.tandfonline.com/doi/abs/10.1080/10920277.2000.10595882`.

R.D. Lee and L.R. Carter. Modeling and forecasting u. s. mortality. *Journal of the American Statistical Association, Vol. 87, No. 419*, pages 659–671, 1992. URL `https://www.jstor.org/stable/2290201?seq=1`.

J. Lledó, J.M. Pavía, and Morillas F.G. The level of mortality in insured populations. *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, pages 449–454, 2018.

J. Lledó, N. Salazar, and J.M. Pavía. Mid-year estimators in life table construction. *Mathematical and Statistical Methods for Actuarial Sciences and Finance: Conference Extracts*, 2020.

J.P. Mackenbach, A.E. Kunst, and C.W.N. Looman. Cultural and economic determinants of geographical mortality patterns in the netherlands. *Journal of Epidemiology and Community Health, Vol. 45*, pages 231–237, 1991.

R. McNown and A. Rogers. Forecasting mortality: A parameterized time series approach. *Demography, Vol. 26*, pages 645–660, 1999. URL `https://www.jstor.org/stable/2061263?seq=1`.

University of California Berkeley-USA and Max Planck Institute for Demographic Research Germany. *Human Mortality Database*, 2020. http://www.mortality.org [Accessed: 2020-04-03].

A. Ornelas and M. Guillen. A comparison between general population mortality and life tables for insurance in mexico under gender proportion inequality. *Revista de Métodos Cuantitativos para la Economía y la Empresa, Vol. 6*, pages 47–67, 2013. URL https://www.redalyc.org/pdf/2331/233129568003.pdf.

A.E. Renshaw and S. Haberman. A cohort-based extension to the lee–carter model for mortality reduction factors. *Insurance: Mathematics and Economics, Volume 38, Issue 3*, pages 556–570, 2006. URL https://www.sciencedirect.com/science/article/pii/S0167668705001678.

D. Torres and W. Mayorga. Graduación de una nueva tabla de mortalidad de asegurados de vida individual. *Revista Fasecolda, Vol. 166*, pages 44–53, 2017.

S Yang, J. Yue, and H.C. Huang. Modeling longevity risks using a principal component approach: A comparison with existing stochastic mortality models. *Insurance: Mathematics and Economics, Vol. 46*, pages 254–270, 2010. URL https://www.sciencedirect.com/science/article/abs/pii/S0167668709001309.