Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

ACTUARIAL SCIENCE

MASTER´S FINAL WORK

DISSERTATION

STOCHASTIC DIFFERENTIAL EQUATIONS DEATH RATES MODELS: THE PORTUGUESE CASE

DANIEL DOS SANTOS BAPTISTA

OCTOBER - 2022

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
ACTUARIAL SCIENCE

MASTER´S FINAL WORK
DISSERTATION

STOCHASTIC DIFFERENTIAL EQUATIONS DEATH RATES MODELS: THE PORTUGUESE CASE

DANIEL DOS SANTOS BAPTISTA

SUPERVISION:

NUNO MIGUEL BAPTISTA BRITES

OCTOBER - 2022

Abstract

In recent years, the increasing life expectancy of the world population (which is most commonly seen in first-world countries), due to increased availability to prescribed medication, quality of health care services and quantity of health care institutions, combined with a sharp decrease in birth rates along time has proven to be a challenging problem for governments worldwide. Both of these factors put at risk the sustainability of state funded welfare programs (e.g. social security) and also lead to a decrease in productivity, available workforce and tax revenue in the near future. With the tendency for these problems to worsen in the next decades, it is of paramount importance to estimate the extension of human life in order to analyse the severity of this phenomena.

Stochastic differential equations have been used recently to model the evolution of death rates. In fact, such models have some advantages compared to the deterministic ones since we can input random environmental fluctuations and evaluate the uncertainty in forecasts.

Instead of the usual cohort analysis, we propose a cross-sectional analysis of mortality by applying stochastic differential equations models, which we wish to model to the Portuguese population, describing death rates trends for all ages and for both genders.

The main goal of this work is to apply and compare stochastic differential equations death rates models (Geometric Brownian motion, Stochastic Gompertz model) separately for each age and gender with independent standard Wiener processes and forecast Portuguese death rates until the year 2030.

Keywords: Death rates; Geometric Brownian motion; Stochastic Gompertz model; Stochastic Differential Equations; Forecasting.

Resumo

Nos últimos anos, o aumento da esperança média de vida da população mundial (que é habitualmente observada em países de primeiro mundo), devido ao aumento do acesso a medicamentos com receita médica, qualidade dos serviços relacionados com cuidados de saúde e quantidade de instituições de saúde, em conjunto com um forte decréscimo registado nas taxas de natalidade ao longo do tempo provou ser um problema desafiante para governos de todo o mundo. Estes fatores colocam em risco a sustentabilidade dos programas de assistência social financiados pelo Estado (como por exemplo a segurança social) e podem causar uma descida na produtividade, mão-de-obra disponível e receita fiscal no futuro próximo. Com a tendência destes problemas agravarem-se nas próximas décadas, é da maior importância estimar o prolongamento da vida humana com a finalidade de analisar a gravidade dos fenómenos referidos.

As equações diferenciais estocásticas têm vindo a ser usadas recentemente para modelar a evolução de taxas de mortalidade. De facto, este tipo de modelos apresentam algumas vantagens quando comparados aos modelos determinísticos, visto que podemos introduzir flutuações ambientais aleatórias e avaliar a incerteza nas previsões.

Em vez da habitual análise por coorte, propomos uma análise transversal da mortalidade através da aplicação de modelos de equações diferenciais estocásticas, que desejamos modelar para a população portuguesa, descrevendo as tendências das taxas de mortalidade para todas as idades e para ambos os géneros.

O principal objectivo deste trabalho é aplicar e comparar modelos de equações diferenciais estocásticas de taxas de mortalidade (movimento Browniano Geométrico, modelo de Gompertz estocástico) separadamente para cada idade e género com processos de Wiener padrão independentes e efetuar as previsões das taxas de mortalidade portuguesas até ao ano 2030.

Palavras-Chave: Taxas de Mortalidade; Movimento Browniano Geométrico; Modelo de Gompertz estocástico; Equações Diferenciais Estocásticas; Previsões.

Table of Contents

Acknowledgements

After finishing my Master's dissertation, it's my duty to thank the following people which, without their support, could not have been possible for me to conclude it:

First and foremost, i would like to thank my supervisor Nuno Miguel Baptista Brites. Without his extensive knowledge in the fields of stochastic calculus and R programming this dissertation would never have left stage one. Furthermore, i would also like to thank him for his endless patience in replying to my numerous emails regarding the state of my dissertation and for all the small talk and advice related to working in research and pursuing a PhD in Portugal.

Second, i would like to thank all the faculty related to the Actuarial Science Masters, in particular the Master's coordinators: João Manuel de Sousa Andrade e Silva, Alexandra Bugalho de Moura and Onofre Alves Simões, for all their commitment into teaching in the fields of actuarial mathematics and statistics, which was essential for the writing of my thesis.

Third, i would like to thank all the personnel associated with the Human Mortality Database (HMD) for providing detailed high-quality harmonized mortality data about the Portuguese population. Without this vital information my work could not have been concluded.

Fourth, i would like to thank all my family members, especially my mother (who cooks a divine bacalhau de brás), my father (for providing for both of us, and supporting me throughout my studies) and my aunt (for the fun talk and for all the free panoche drinks she bought me in the cafe near her home she usually goes to), for all their support and caring throughout the writing of my master's dissertation.

And last, but not least, i would like to thank all my friends from ISEG, ISCAL and High School for always believing in me and supporting me all the way until the end.

Thank you all so much!

Glossary

GBM - Geometric Brownian motion

HMD - Human Mortality Database

LT - long term

ML - maximum likelihood

MSE - mean squared error

p.d.f - probability density function

r.v - random variable

SDE - stochastic differential equation

SGM - Stochastic Gompertz model

SS - step-by-step

s.p - stochastic process

WN - white noise

1 Introduction

1.1 Evolution of human mortality in Portugal

In Portugal, and in the majority of western countries, the age structure of the population has been changing, marked by an increase of ageing population due to the combined effect of decreasing birth rates and increasing life expectancy throughout the years. According to the projections of the Portuguese resident population between the years of 2018 – 2080, the ageing population (individuals aged 65 years or higher) will represent about 37% of the resident population in 2080, considering the expected scenario (*Instituto Nacional de Estatística* (2020)). After analysing the data obtained from the Portuguese *census* performed in the year 2021 (*PORDATA* (2021)), the ageing population currently represents about 23% of the resident population, meaning that in the next decades the proportion of the ageing population in relation with the resident one has the tendency to increase over time (even doubling in some regions of the country).

However, if it's certain that the mortality risk increases in relation with the age of the individual, mortality rates have been plummeting worldwide. This fact has led to the study of factors, both intrinsic and extrinsic, that can explain this evolution. Types of models, deterministic or more recently, stochastic models, have been tested giving rise, namely, to comparative studies to assess which is the best model to apply in this context (see Booth & Tickle (2008) and George et al. (2003)). For all these reasons, and despite the fact that human mortality is a demographic variable that has been studied exhaustively, the main objective of this dissertation is to apply models of stochastic differential equations (SDEs) that, through cross-sectional analysis of the mortality data over time, allows us to estimate the future tendency of the decreasing death rates phenomenon for all age groups and for each gender, and to compute step-by-step (SS) forecasts and long term (LT) forecasts.

The data related with the Portuguese death rates and used throughout this work was obtained from the HMD (*Human Mortality Database* (2022)), which corresponds to the gross death rates and represents the division between the number of deaths (total for a country in a given time period for all causes of death) and an estimate of the resident population (which corresponds to the population exposed to death risk in the same age interval). In this work we will be using 200 time series, with an annual frequency, available for the years 1940 – 2020 for 100 annual age groups (ages 0 – 99) and for both genders. For example, age 0 (which corresponds to the first age analysed) refers to individuals who died in the first year of life and we denote, respectively, as F0 and M0 the death rates of females and males at age 0 (this method is used similarly for the remaining ages).

In Demography, it's common for data to be available by cohort (in a longitudinal perspective through time). A cohort represents a set of individuals born in the same year and that are followed throughout their lives. In this case, where a longitudinal approach is used over time, there is no distinction between age and calendar year. Therefore, it's very difficult to model all ages of the human life span, as it's necessary a very high number of parameters for this purpose[1].

For the purpose of this approach, see the data representation in Figure 1. The curve describes the evolution of mortality in the various phases of the arch of life. In this case, the year 1994

[1] Often more than eight for each cohort, because the mortality trajectory is very irregular.

was fixed, but the shape, usually described in the literature as "bathtub-shaped curve", has not changed significantly over time despite the reduction in infant mortality and greater longevity in the last few decades.

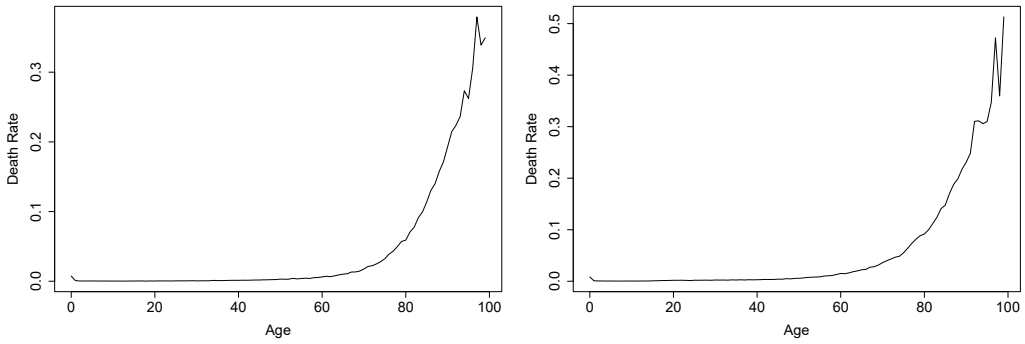

Figure 1: Death rates of the Portuguese population (female gender on the left and male gender on the right), longitudinal representation (ages 0 to 99) for the year 1994.

Alternatively, the cross-sectional approach we follow makes sense, as we consider events that, over time, affect all ages. Among others, we highlight, on the positive side, changes in living conditions of a socio-economic nature or advances in medicine and increased quality of health care services and number of health care institutions (such as hospitals, clinics and among others). Also, climate changes that generate extreme phenomena or other catastrophic situations can globally affect the Portuguese population, in this case increasing mortality risk.

The phenomenon thus described has a strong decreasing tendency in the period under analysis as seen in Figure 2. In almost all ages, the death rates are higher in males than in females, although with a different evolution at each age.

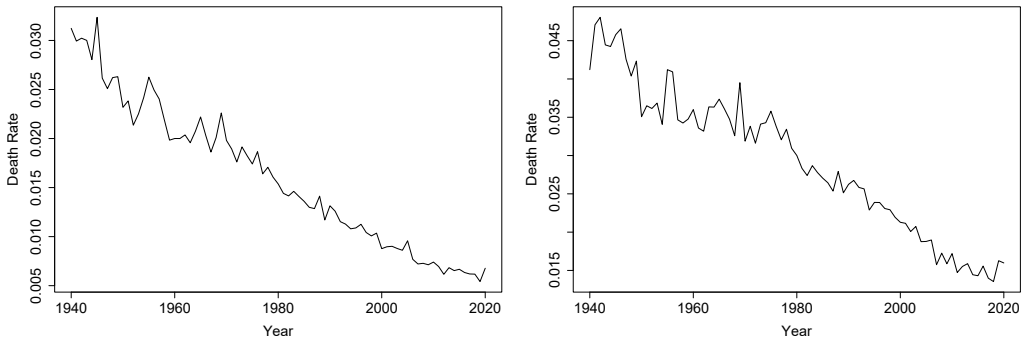

Figure 2: Death rates of individuals aged 66 along time (on the left for the female gender and on the right for the male gender) from 1940 to 2020.

The results and methods are illustrated by the death rates of the Portuguese population. We consider that they reflect the behaviour of mortality in countries that have already undergone the demographic transition (regarding the evolution of mortality in the context of the demographic transition in Portugal and worldwide see, for instance, Morais (2002)).

Furthermore, throughout this work, we divided each time series related with the observed death

rates of the Portuguese Population (which have 81 observations and are related with the observed death rates documented in each year of analysis, from 1940 to 2020), into two subsets: Observed death rates between the years 1940 – 2009 for model adjustment and between the years 2010 – 2020 for forecast validation.

Before concluding this subsection, we also call attention to the fact that, in Demography, the variable "force of mortality" is often the object of study, in most cases represented as μ. Being i a certain age, we have that $\mu_i = -\ln(1 - q_i)$, with q_i denoting the death rate of an individual aged i of a given gender (these questions are frequent in the construction of life tables and are described exhaustively, for instance, in Namboodiri & Suchindran (1987) and Preston et al. (2004)). If we consider that the death rate is constant between the exact ages of i and $i + 1$ and in a given timeline (annually, for example), we approximate q_i by the value of the force of mortality μ_i (on the subject of mortality statistics measures see also Keyfitz & Caswell (2005) and Mcgehee (2003)). In reality, the average deviation between μ_i and q_i is quite small (estimated to be around 10^{-5}) in most of the ages (in the period under analysis, it increases only from age 85 above). In Figure 3 both values corresponding to the observed death rate and force of mortality (μ), between the time period of 1940 – 2020, for an individual aged 23 of the male gender. Since, as previously stated, the difference between the values of the death rate and the force of mortality of a given individual is extremely small, in Figure 3 it's not easily perceptible the lines of the analysed variables, since they overlap one another in most of the years, from 1940 to 2020.

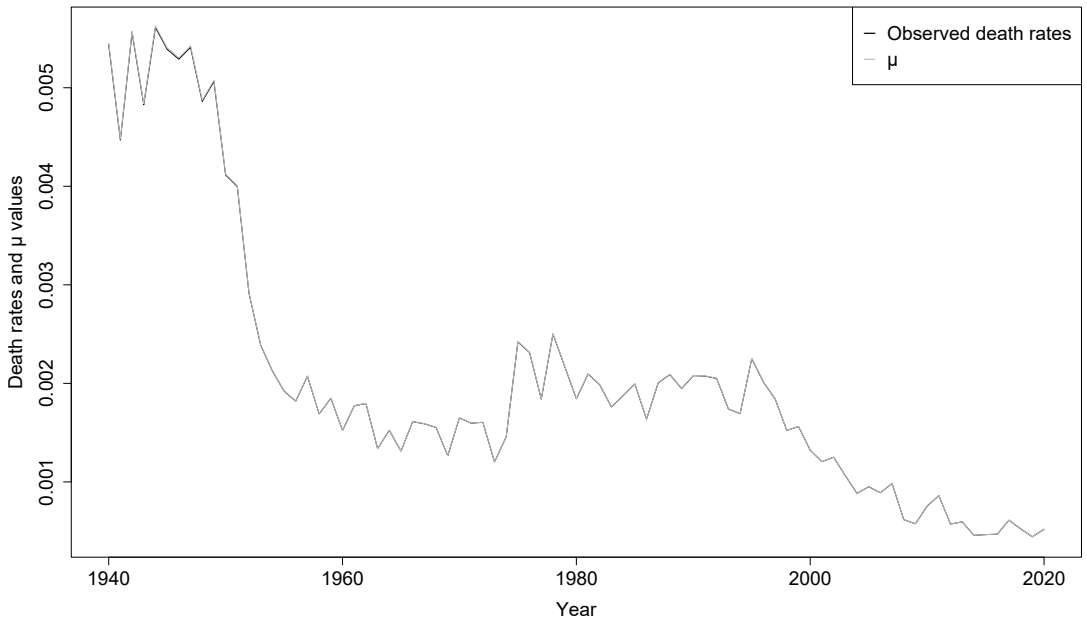

Figure 3: Observed death rates *vs* force of mortality (μ) of an individual aged 23 of the male gender, from 1940 to 2020.

1.2 Modelling human mortality with stochastic differential equations: brief overview of the literature

The future evolution of life expectancy is uncertain, due to external factors and to the uncertainty itself in the evolutionary trend of death rates as a demographic phenomenon. Since the 19th century, with the first studies by Gompertz, much has changed in the approach to this problem, which has been extensively investigated throughout the last few decades.

Originally, such models did not incorporate uncertainty - it was introduced through the construction of mortality tables as seen in Mendes (2004) and Mexia & Corte-Real (1995), studying a generation or cohort. It was only when this need was recognized, that the first stochastic (or probabilistic) models were developed, which emerged especially since the 1990s, mainly from the perspective of actuaries, economists and investment banks (Li (2007)). Of these, the Lee-Carter model, in Lee & Carter (1992), is undoubtedly the most well known, with many applications and variations (see, for example, Lee (2000) or *Life Office Mortality Comitee* (2007), in which a summary of the results of it's application is made, or Bravo et al. (2010) related to mortality predictions in Portugal).

With a long application in the study of financial markets behavior, SDEs, whose Black-Scholes model (1970s) has stimulated research and the development of applications to other areas of science, have been widely used in modelling population growth (see Braumann (2008) and references contained in Brites (2017), Brites & Braumann (2019) and Brites & Braumann (2020)). Recently, SDEs models variant of the Ornstein-Uhlenbeck model started to be applied in Portugal, which also incorporate a term with an environmental random component, to demographic data, namely in the longitudinal study of mortality or in the construction of dynamic mortality tables. See, for example, their use in the construction of prospective mortality tables, actuarial applications and longevity risk coverage in Bravo (2007) and Bravo & Braumann (2007) or, along the same lines, studies on dynamic tables, applied to death rates in Spain on Debón et al. (2008).

These models allow the implementation of randomness, which translates the effects of environmental variations in the coefficients (thus, they are more realistic), and it's possible, from the solution of the equation, to infer on it's probability distribution. From the few references found in the literature related to the use of SDEs to model human mortality, and from the perspective of the study of cohorts, we highlight the recent model by Jevtic et al. (2013), for a mortality surface and using factor analysis, also the model by Park (2008), in which, to obtain the probability of survival, the mortality force is estimated using factor analysis of mortality through a diffusion process with jumps, and the model of Yashin et al. (2007), in which mortality is a function of risk factors, which changes with age and are translated by a SDE with jumps translated by standard independent Wiener processes.

Although the set of discrete-time models of the Lee-Carter type, which generally incorporate the stochastic component in a single term, proves to be good in the short run (the parameters generally need to be readjusted for medium-long term predictions), the recent SDEs models bring additional advantages, as they associate uncertainty with the dynamics of the process. Their construction is based on deterministic models of ordinary differential equations, while incorporating the effect of environmental variability in the evolution of death rates.

The SDEs models that we propose to apply are intended to be simple and flexible (although

with different parameters by age and gender). Assuming that the demographic system does not evolve independently from the economic and environmental systems (Mishra (2008)), death rates have stochastic fluctuations as a function of the "environment" in a broad sense (as we have mentioned previously). In addition, to the environmental (or systematic) randomness, observed death rates also have an associated sampling error (demographic randomness), which is not object of study in this dissertation. This is an error which, in relative terms, is small and therefore is not treated, since it has some significance only at older ages (because the "sample", meaning, the population at risk, has a small size in comparison with younger ages).

In a previous approach, already mentioned in Bravo (2007), longitudinal models of SDEs were used, in order to explain the evolution of a fictional cohort (age and time evolve together) and good results were obtained for older ages. However, the longitudinal approach has limitations, because a restricted time/age period has to be selected from the outset, given the very complex behaviour of the death rate relative to the age of a given individual when considering the entire human life span. The cross-sectional over time approach, used throughout this work, on the contrary, models the evolution of the death rate of a certain age (fixed over time), which has a relatively constant behaviour.

Therefore, as the data shows a dynamic evolution of death rates over time (and not merely in a sample), it makes sense to build and apply models with a random environmental component, hence the use of SDEs models. About the potential use of these models, where one seeks to explain mortality variability in a simple and credible way for planning purposes (e.g., pensions, savings, health plans, or insurance), we can also convert the results into derived variables, such as life expectancy or survival rates, random variables that also depend on the environmental conditions, studying complementary or related problems and even introducing explanatory variables outside the scope of the mortality system.

1.3 Dissertation's objectives and structure

Considering the problem that was the starting point for this research (as explained in subsection 1.1), this work seeks to answer the following questions: what is the future medium/long term trend of the Portuguese population death rates by age and gender?, and, in particular, how do the forecasts given by SDEs models behave if we consider the correlation effect between death rates of different genders, for the same age, and between different ages considering the same gender?

As for the dissertation's structure, this work is comprised of 4 chapters. In the first chapter, we present the motivational aspects that led to the identification of the problem and selection of the methodology, then provided a brief overview of the literature related to stochastic mortality models in order to contextualize the topic.

On the second chapter, we briefly present a conceptual and methodological exposition about the theory behind SDEs, which is necessary for the development on the subsequent chapter.

The third chapter, which is the core of this work, is related with modelling the death rates of the Portuguese population through SDEs models. The models used are the Geometric Brownian motion (GBM) and the Stochastic Gompertz model (SGM). In addition, will be treated, using examples, the statistical aspects of model selection, estimation and forecasting, as well as their confidence intervals. As for the calibration of the SDEs models, the maximum likelihood (ML)

method will be used to estimate the parameters (by age and gender). Considering model validation, beyond comparison between both models used, measures of performance evaluation and the study of predictive capability will be used (the mean squared error (MSE) for comparison and validation purposes).

In Chapter 4, we summarize the main conclusions from this dissertation and make some considerations about future work. Furthermore, the data related to the observed death rates of the Portuguese population, by age and gender, between the years of 1940 to 2020 is stored in an excel file titled "Death_Rates_1X1.xlsx" which is available for download here:[2] https://drive.google.com/drive/folders/1fTLCkfCstzivHGvA6ezvHXMB_s9ZE9is?usp=sharing.

Also all the results obtained in this dissertation were computed using the R programming language (available, with free access, in http://www.r-project.org). The R code related with this dissertation is publicly available and can be found here: https://github.com/DanielBaptista99/Stochastic-Differential-Equations-Death-Rates-Models-The-Portuguese-Case.git.

[2]In order to use the data provided by the HMD we had to do a few adjustments. First we had to remove the ages 100 – 115 since these are not analysed in this dissertation, second we had to copy-transpose each of the death rates in sheet "Folha1" because in this format it's easier to run the R code. Furthermore, there seems to be some errors in the data provided by the HMD since in some cases the death rates will be equal to zero (which is highly unlikely). In these cases we computed the average between the previous and next year in order to obtain a more suitable value of the death rate (the death rates in which this adjustment was done are marked with the colour yellow).

2 Topics on stochastic differential equations

2.1 Introduction

We present below a brief exposition of concepts, properties and numerical aspects concerning the theory related to SDEs. All these topics, some of which are based on probability theory or originated from mathematical analysis, are exhaustively stated and demonstrated in the reference bibliography, namely in Karlin & Taylor (1981), Arnold (1992), Nicolau (2001), Øksendal (2003), Braumann (2005), Bravo (2007), Li (2007), Müller (2007), Skiadas (2010), Lagarto (2014) and Brites (2017).

2.2 Stochastic processes

Beforehand, we consider that the phenomenon we are going to study is not purely deterministic, because by observing, in our case, the death rates (by age and gender) and their variations over time, we find that they suffer random fluctuations that we cannot predict, meaning that they feature stochastic behavior. These processes can be modelled using sets of r.vs that describe the system under study at each moment of time, t, with $t \in T$ (usually, $T = \mathbb{R}_0^+$ or $T = \mathbb{N}_0$, i.e, in continuous or discrete way), and which also depend on chance, ω, with $\omega \in \Omega$, where Ω represents the set of all possible outcomes of an event (or random event) or possible states of nature (in a broad sense), susceptible of disturbing this same phenomenon. Our main goal is to insert a noise source in our models in order to capture or better explain the random variability of a given process over time. The phenomenon described in this way, which translates the evolution of a set of r.vs, $\{X(t)\}$, with $t \in T$, is a stochastic process (s.p) indexed by T, which we denote by $X(t)$. A s.p is also a function $X(t, \omega)$ defined on $T \times \Omega$. Setting ω as a fixed value results in a non-random function of t, which we call trajectory or sample path of the process (meaning different values of ω generate different trajectories). Following the same pattern as the literature mentioned in the previous subchapter, we will denote throughout this dissertation $X(t, \omega)$ as $X(t)$. From now on, we assume $T = [0, +\infty[$, so the s.p, is in continuous time, and also the state variable $X(t)$ is continuous (since the variable can change it's value at any instant of time and can take any real value). A s.p indexed by T is a group of r.vs, all of them defined on the same probability space $(\Omega, \mathcal{F}, \mathcal{P})$, with $\mathcal{P}$ denoting the probability measure and $\mathcal{F}$ denoting a σ-algebra on Ω.

A filtration $\mathcal{F}_t$, with $t \in T$, is a set of σ-algebras of $\mathcal{F}$ such that $s \leq t \implies \mathcal{F}_s \subseteq \mathcal{F}_t$. A s.p $X(t)$ with $t \in T$, is adapted to this filtration if $X(t)$ is $\mathcal{F}_t$-mensurable for all $t \in T$. The process $X(t)$ is adapted to it's natural filtration, $\mathcal{F}_t = \sigma(X(s) : 0 \leq s \leq t)$, where $\mathcal{F}_t$ is the σ-algebra generated by the present and past of $X(t)$. Furthermore, we say that $X(t)$ is a s.p with:

- independent increments if and only if for all $n \in \mathbb{N}_0$ and for all $t = 0, 1, \ldots, n \in T$, the random variables $X(1) - X(0)$,..., $X(n) - X(n-1)$ are independent.
- stationary increments if and only if for all $s, t \in T$ such that $s < t$, the distribution of $X(t) - X(s)$ depends only on the duration $t - s$.

In addition, $X(t)$ is a second order process if and only if for all $t \in T : E[X(t)^2] < +\infty$. Further-

more, $X(t)$ is a $\mathcal{F}_t$ martingale[3] if:

- $X(t)$ is adapted to the filtration $\mathcal{F}_t$.
- $E[|X(t)|] < +\infty$
- $\forall s \leq t$: $E[X(t)|\mathcal{F}_s] = X(s)$ almost surely.

There are several classifications for s.ps, depending on the characteristics of the defining r.vs, the set T considered, and the state space. Therefore, it should be noted that all s.ps we will use in this dissertation, as well as the solutions of the presented SDEs, can be considered Markov processes. The standard Wiener process, $W(t)$, fundamental for the construction of SDEs (since it can describe the accumulated effect of environmental fluctuations of a given phenomenon, up to a certain time t considered) is an homogeneous Markov process.

The standard Wiener process was first discovered by the English botanist Robert Brown[4], when he observed the random movements of small particles of pollen immersed in liquid in the year 1828. Later, in 1900, Louis Bachelier, in his thesis "Théorie de la Spéculation" (Bachelier (1900)) used the Brownian motion to model the evolution in the price of financial assets along time. However Bachelier's remarkable work was far ahead of his time and was not appreciated during his lifetime, since, in the eyes of the French mathematical elite, Bachelier was considered of lesser importance. In 1905, Albert Einstein justified this movement with the constant collision between the particles and the surrounding liquid molecules and characterized it by a stochastic process that would come to be called Wiener process. Finally, in the year 1918, the first mathematical definition of the term appeared through the mathematician Norbert Wiener. A very interesting description on the history of the standard Wiener process can be found in Nelson (2021).

Let's denote B as a Borel set, such that $B \in \mathcal{B}$, with $\mathcal{B}$ denoting the Borel σ-algebra which represents the smallest σ-algebra that contains the intervals in T. $X(t)$ is a Markov process if, for all $s, t \in T$ with $s < t$ and for any Borel set,

$$P[X(t) \in B|X(u), 0 \leq u \leq s] = P[X(t) \in B|X(s)].$$

This property, usually known as the Markov property, states that, knowing the present value of the process, it's future values are independent from past values. If a Markov process has stationary transition probabilities (in time), this is,

$$P[X(t+\tau) \in B|X(s+\tau) = x] = P[X(t) \in B|X(s) = x],$$

then it's called an homogeneous Markov process.

A s.p $X(t)$ with second order moments is called a diffusion process if it verifies the Markov property and if, additionally, almost certainly exhibits continuous trajectories, for $\epsilon > 0, x \in \mathbb{R}$ and $s \in [0, t] \subset T$, with uniform convergences, the limits

$$\lim_{\Delta \to 0^+} \frac{P_{s,x}[|X(s+\Delta) - x| > \epsilon]}{\Delta} = 0,$$

[3]When the considered filtration coincides with the natural one, $X(t)$ is simply called martingale.

[4]For this reason, the standard Wiener process is described in some literature as the standard Brownian motion, or simply Brownian motion, and is often denoted as $B(t)$ instead of $W(t)$.

$$\lim_{\Delta \to 0^+} E_{s,x}\left[\frac{X(s+\Delta)-x}{\Delta}\right] = a(s,x),$$

$$\lim_{\Delta \to 0^+} E_{s,x}\left[\frac{(X(s+\Delta)-x)^2}{\Delta}\right] = b(s,x),$$

where $P_{s,x}$ denotes the conditional probability to which $X(s) = x$ and $E_{s,x}$ denotes the conditional mathematical expectation to which $X(s) = x$. This definition can be generalized to second-order processes. The functions $a(s,x)$ and $b(s,x)$, which correspond, respectively, to the infinitesimal moments of first and second order, are called the drift coefficient (or infinitesimal average) and the diffusion coefficient (or infinitesimal variance). If these coefficients do not depend on t, then the diffusion process is said to be homogeneous.

The standard Wiener process $(W(t))$ is an homogeneous diffusion process and has the following properties:

- $W(0) = 0$ almost certainly;
- $W(t)$ has a normal distribution with mean zero and variance t ,with $t \in T$ (meaning, $W(t) \sim \mathcal{N}(0,t)$);
- the increments $W(t) - W(s)$ (with $0 \leq s < t$ and $t \in T$) have a normal distribution with mean zero and variance $t - s$ (meaning, $(W(t) - W(s)) \sim \mathcal{N}(0, t-s)$);
- the increments $W(t) - W(s)$ (with $0 \leq s < t$ and $t \in T$), on non-overlapping time intervals, are independent;
- $Cov[W(s), W(t)] = min(s,t)$;
- $W(t)$ is a martingale;
- $W(t)$ is a Markov process.

2.3 Stochastic differential equations

Ordinary differential equations have been extensively used to model the behaviour of dynamical time-dependent phenomena in various scientific areas. Such dynamics can often be characterized by the rate of change of a variable $X(t)$ and denoted as

$$dX(t) = f(t, X(t))dt, \quad X(0) = X_0. \tag{2.1}$$

Usually, a SDE is obtained from an ordinary differential equation, such as Equation (2.1), to which we add a noise term, in order to describe the random fluctuations that affect the phenomenon under study. Assuming that the accumulated effects of these random fluctuations up until time t can be described by a standard Wiener process, $W(t)$, then the SDE can be denoted as

$$dX(t) = f(t, X(t))dt + g(t, X(t))dW(t), \quad X(0) = X_0. \tag{2.2}$$

We assume X_0 to be a r.v independent from $W(t)$, and that f and g are real functions. A solution

for $X(t) = X(t, \omega)$ from Equation (2.2) is a s.p that solves the integral equation

$$X(t) = X(0) + \int_0^t f(s, X(s))ds + \int_0^t g(s, X(s))dW(s), \tag{2.3}$$

more explicitly

$$X(t, \omega) = X(0, \omega) + \int_0^t f(s, X(s, \omega))ds + \int_0^t g(s, X(s, \omega))dW(s, \omega), \tag{2.4}$$

with the defined integrals as we will describe next.

Let's denote $F(s, \omega) = f(s, X(s, \omega))$ and $G(s, \omega) = g(s, X(s, \omega))$. The integral $\int_0^t F(s, \omega)ds$ can be considered, for each fixed ω, as a Riemann integral. However, the integral $\int_0^t G(s, \omega)dW(s, \omega)$ cannot be considered as a Riemann-Stieltjes integral since different sums of Riemann-Stieltjes converge to different limits. We work with non-anticipative functions $G(s, \omega)$ with a finite norm L^2, that is, $(||G||_2)^2 = E[\int_0^t |G(s, \omega)|^2 ds] < +\infty$. The function $G(s, \omega)$ is non-anticipative if it's jointly measurable in s and ω and it's independent from future increments of the standard Wiener processes. For $G \in L^2$ we use the Itô's integral, which is defined as the limit in quadratic mean of the sums of Riemann-Stieltjes, that is,

$$\underset{n \to +\infty}{l.i.m.} \sum_{k=1}^{n} G(t_{k-1})(W(t_k) - W(t_{k-1})),$$

where $0 = t_{0,n} \leq t_{1,n} \leq \ldots \leq t_{n,n} = t$, with $n \geq 1$, are breakdowns of the interval $[0, t]$ whose range tends to 0 when $n \to +\infty$. Notice that the Riemann-Stieltjes sums use as an intermediate point the initial point of each breakdown interval. Other choices of intermediate points would generate other integral types, but the choice that was made (non-anticipative), which leads to the Itô's integral, has the main advantage of generating rather interesting properties of the integral. This definition can be extended to non-anticipative functions of class G such that $\int_0^t |G(s)^2| ds < +\infty$ almost certainly.

The leading researcher, either for the definitions or to what has come to be known as stochastic calculus, was Kiyoshi Itô, Japanese mathematician who developed, in the 1940s, the basis for the SDEs theory. Identifying almost identical functions, L^2 is an Hilbert space. Of the properties related with stochastic integrals, we highlight the following, considering the integration interval $[0, t]$, $a, b \in \mathbb{R}$ and $G, G_1, G_2 \in L^2$:

- $\int_0^t dW(s) = W(t) - W(0)$;
- $\int_0^t (aG_1(s) + bG_2(s))\, dW(s) = a \int_0^t G_1(s)dW(s) + b \int_0^t G_2(s)dW(s)$;
- $E\left[\int_0^t G(s)dW(s)\right] = 0$;
- $E\left[\left(\int_0^t G(s)dW(s)\right)^2\right] = E\left[\int_0^t G^2(s)ds\right]$;
- $E\left[\int_0^t G_1(s)dW(s) \int_0^t G_2(s)dW(s)\right] = E\left[\int_0^t G_1(s)G_2(s)ds\right]$.

Analysing Equations (2.3) and (2.4), if f and g satisfy the adequate properties (see, for example, Braumann (2008)), the solution exists, is unique and is a diffusion process with drift coefficient $a(s,x) = f(s,x)$ and diffusion coefficient $b(s,x) = g^2(s,x)$. When f and g do not depend on time, as it occurs in this dissertation, the SDE is said to be autonomous and it's solution is an Itô's diffusion.

2.4 Itô's formula

A process $X(t)$ presented as a variation of Equation (2.4), in which $f(s, W(s,\omega)) = F(s,\omega)$ and $g(s, W(s,\omega)) = G(s,\omega)$, meaning,

$$X(t,\omega) = X(0,\omega) + \int_0^t F(s,\omega)ds + \int_0^t G(s,\omega)dW(s,\omega),\ X(0,\omega) = x_0,$$

independent from $W(t)$ and F, G mensurable in s and ω, which verify, almost certainly,

- $\int_0^t G^2(s)ds < +\infty$,
- $\int_0^t |F(s)|ds < +\infty$,

is called an Itô process. Let's assume, from onwards, that $X(t)$ is an Itô process. If $Y(t) = h(t, X(t))$, with $h(t,x)$ being a function of class $C^{1,2}$ (that is, with first order partial derivative continuous at t and second order partial derivative continuous at x), then $Y(t) = Y(t,\omega)$ is still an Itô process. The Itô's formula (which refers to the differentiation of a composite function rule or chain rule), is given, relative to $Y(t)$, as:

$$dY(t) = \frac{\partial h(t,X(t))}{\partial t}dt + \frac{\partial h(t,X(t))}{\partial x}dX(t) + \frac{1}{2}\frac{\partial^2 h(t,X(t))}{\partial x^2}\ (dX(t))^2. \tag{2.5}$$

Furthermore, the following properties are related with Itô's formula and are used:

- $(dt)^2 = 0$,
- $dtdW(t) = dW(t)dt = 0$,
- $(dW(t))^2 = dt$.

3 Stochastic differential equations death rates models

3.1 Introduction

In the cross-sectional analysis of human mortality, we consider that one must take into account the random fluctuations of the environmental conditions, to which SDEs are used to model the death rates of the Portuguese population, considering the age and gender of the individuals under analysis in this work.

From the merely preliminary analysis of the observed death rates of the Portuguese population, in the sense of finding the first results to make the dissertation's plan feasible, we observed that relatively simpler models (with two or three parameters) allow us to obtain encouraging results (which, even, portray the variability of death rates at older ages, generally more difficult to model, and enable us to measure forecasts). In this work, we adjusted two stochastic differential equations death rates models, the GBM and the SGM, to the observed death rates, which are analysed in the following subsections.

3.2 The Geometric Brownian motion (GBM)

The GBM is a s.p usually used to model the price of stocks and other economic variables. This is also the solution for the SDE commonly known as the Black-Scholes model (also, in some literature, designated as the diffusion equation of Black-Scholes), with μ and σ representing, respectively, the mean growth rate and volatility of a given r.v. The SDE representing the GBM is

$$dX(t) = \mu X(t)dt + \sigma X(t)dW(t),\ \sigma > 0,\ X(0) = x_0. \tag{3.1}$$

In this case, $X = X(t)$ can represent the price of a given financial asset along time t, but this equation has various applications, not limited to only modelling economic variables, since it can also be used to model population growth, as seen in Brites (2010), as well as other variables in various areas of science. Integrating Equation (3.1) we obtain

$$X(t) = X(0) + \mu \int_0^t X(s)ds + \sigma \int_0^t X(s)dW(s).$$

Furthermore, we can solve the SDE (3.1) in order to obtain the equation that defines $X(t)$, by applying the Itô's formula (2.5). Let's assume, for this case, that $Z(t) = \ln(X(t))$ and subsequently that $h(t, y) = \ln(y)$ and $Z(t) = h(t, X(t))$. After applying Itô's formula, we get

$$\begin{aligned}
dZ(t) &= \frac{\partial \ln(X(t))}{\partial t}dt + \frac{\partial \ln(X(t))}{\partial X(t)}dX(t) + \frac{1}{2}\frac{\partial^2 \ln(X(t))}{\partial X(t)^2}\ (dX(t))^2 \\
&= \frac{\partial \ln(X(t))}{\partial X(t)}dX(t) + \frac{1}{2}\frac{\partial^2 \ln(X(t))}{\partial X(t)^2}\ (dX(t))^2 \\
&= \frac{1}{X(t)}dX(t) - \frac{1}{2}\frac{1}{(X(t))^2}(dX(t))^2 \\
&= \frac{1}{X(t)}(\mu X(t)dt + \sigma X(t)dW(t)) - \frac{1}{2}\frac{1}{(X(t))^2}\sigma^2(X(t))^2dt \\
&= \mu dt + \sigma dW(t) - \frac{1}{2}\sigma^2 dt \\
&= \left(\mu - \frac{1}{2}\sigma^2\right)dt + \sigma dW(t).
\end{aligned}$$

From the above, we conclude that $dZ(t) = (\mu - \frac{1}{2}\sigma^2)dt + \sigma dW(t)$. Then, we can solve this equation in order to obtain $X(t)$ by integrating $dZ(s)$ in the time interval $[0,t]$ with $\int_0^t dZ(s) = \ln(X(t)) - \ln(X(0))$, i.e.,

$$\begin{aligned}
\ln(X(t)) - \ln(X(0)) &= \int_0^t \left(\mu - \frac{1}{2}\sigma^2\right)ds + \int_0^t \sigma dW(s) \\
&= \left(\mu - \frac{1}{2}\sigma^2\right)\int_0^t ds + \sigma\int_0^t dW(s).
\end{aligned}$$

Following these computations, we notice that $\ln(X(t)) - \ln(X(0))$ can be rewritten as

$$\ln\left(\frac{X(t)}{X(0)}\right) = \left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W(t).$$

From this equation we can finally have a solution for the s.p $X(t)$ (known as the GBM) which will be equal to

$$X(t) = X(0)\exp\left\{\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W(t)\right\}, \quad X(0) = x_0. \tag{3.2}$$

Let's consider that the death rates of the Portuguese population follow a GBM and assume, as a starting point for modelling, Equations (3.1) and (3.2). In this regard, notice that, in fact, when observing the death rates of the Portuguese population throughout time, it appears to have a decreasing linear trend, as was previously seen in Figure 2. From onwards, let's assume $X_k(t)$ to be the death rate of a given individual aged $i-1$ with $i = \{1,\dots,100\}$ and gender j with $j = 1$ if the individual in question is female and $j = 2$ if it's male, on instant t, with $k = i + 100(j-1)$ in order to cover all ages in the arch of life for both genders. To simplify the reading we use throughout this section $X(t) = X_k(t)$, applying the model to each age and gender. Assume also that the initial condition $X(0) = x_0$ is known. If we denote $Y(t) = h(t, X(t)) = \ln\left(\frac{X(t)}{x_0}\right)$, with $X(t)$ denoting the same result shown in Equation (3.2), $h(t,x) = \ln\left(\frac{x}{x_0}\right)$ is a strictly increasing function of class C^2 in x. Applying the Itô's formula in (2.5) we can obtain the SDE

$$dY(t) = Rdt + \sigma dW(t), \quad Y(0) = 0, \tag{3.3}$$

with $R = \mu - \frac{\sigma^2}{2}$. Notice that since we are using $X(t)$ instead of $X_k(t)$, the same reasoning can be implemented to the model's parameters, which we could have denoted as R_k and σ_k representing, respectively, the average growth rate of $Y_k(t)$ and the effect of random fluctuations on mortality dynamics.

The solution for Equation (3.3), for each age and gender in instant t, is given by

$$Y(t) = Rt + \sigma W(t), \tag{3.4}$$

which follows a normal distribution with mean Rt and variance $\sigma^2 t$, that is,

$$Y(t) \sim \mathcal{N}\left(Rt, \sigma^2 t\right), \tag{3.5}$$

where $X(t)$ has a log-normal distribution with expected value $E[X(t)] = x_0 \exp\{Rt\}$. Therefore, we can write Equation (3.4) in it's original form as

$$X(t) = X(0)\exp\{Rt + \sigma W(t)\}, \quad X(0) = x_0.$$

Furthermore, notice that Equation (3.3) is an autonomous SDE and that it's solution (3.4) is an Itô's diffusion and an homogeneous diffusion process with drift coefficient R and diffusion coefficient σ^2.

3.2.1 Estimation

From (3.5) we obtain the probability density function (p.d.f), $f(t, y)$, of $Y(t)$ which is given by

$$f(t,y) = \frac{1}{\sqrt{2\pi V t}} \exp\left\{-\frac{1}{2}\frac{(y - Rt)^2}{Vt}\right\}, \quad V = \sigma^2.$$

Let $t_n = t_0 + n,\ n = 0, 1, \ldots, N$, represent the years in which the death rates of the Portuguese population were observed, for each age and gender (in this case, all series have the same dimension). Considering $Y(t_0) = 0$ and

$$Y(t_n) = Y(t_{n-1}) + R(t_n - t_{n-1}) + \sigma(W(t_n) - W(t_{n-1})), \tag{3.6}$$

the process $Y(t_n)$ conditioned by $Y(t_{n-1})$ has normal distribution with mean $Y(t_{n-1}) + R(t_n - t_{n-1})$ and variance $V(t_n - t_{n-1})$, since $Y(t_{n-1})$ is independent from $W(t_n) - W(t_{n-1})$. Thus, the transition p.d.f of $Y(t)$ between t_{n-1} and t_n is given by

$$f(Y(t_n)|Y(t_{n-1})) = \frac{1}{\sqrt{2\pi V(t_n - t_{n-1})}} \exp\left\{-\frac{1}{2}\frac{(Y(t_n) - Y(t_{n-1}) - R(t_n - t_{n-1}))^2}{V(t_n - t_{n-1})}\right\}. \tag{3.7}$$

Notice that R and V are, respectively, the mean and variance of the logarithm of the death rates returns, $\ln\left(\frac{X(t_n)}{X(t_{n-1})}\right) = Y(t_n) - Y(t_{n-1})$. The parameter vector denoted as $p = (R, V)$ can be estimated by applying the ML methodology. Since $Y(t)$ is a Markov process, the log-likelihood

function, L, given the observed values $Y(t_1), \ldots, Y(t_N)$, can be written as

$$\begin{aligned} L(p|Y(t_1), \ldots, Y(t_N)) &= \sum_{n=1}^{N} \ln\left(f(Y(t_n)|Y(t_{n-1})\right) \\ &= -\frac{N}{2}\ln\left(2\pi V\right) - \frac{1}{2}\sum_{n=1}^{N}\ln\left(t_n - t_{n-1}\right) \\ &- \frac{1}{2V}\sum_{n=1}^{N}\frac{(Y(t_n) - Y(t_{n-1}) - R(t_n - t_{n-1}))^2}{2V(t_n - t_{n-1})}. \end{aligned}$$

Furthermore, we can obtain the explicit expressions of the ML estimators of p (see Brites (2010)), by solving the following system of equations

$$\begin{cases} \frac{\partial L(y;p)}{\partial R}|_{\widehat{R},\widehat{V}} = 0 \\ \frac{\partial L(y;p)}{\partial V}|_{\widehat{R},\widehat{V}} = 0, \end{cases}$$

obtaining, for $t_n - t_{n-1}$,

$$\widehat{R} = \frac{Y(t_N)}{t_N},$$

and

$$\widehat{V} = \frac{1}{N}\sum_{n=1}^{N}\frac{(Y(t_n) - Y(t_{n-1}) - \widehat{R}(t_n - t_{n-1}))^2}{t_n - t_{n-1}}.$$

Since, here, the death rates of the Portuguese population are annually rates, we can therefore assume that $t_n - t_{n-1} = 1$, which simplifies significantly the computations. This simplification is valid for all models applied to the data set and displayed in the following subsections.

To obtain the confidence intervals, CI, for R and V, we can take into account the asymptotic properties of the ML estimators. According to Pestana & Velosa (2002), the Fisher information matrix is

$$F = \begin{bmatrix} -E\left[\frac{\partial^2 L}{\partial R^2}\right] & -E\left[\frac{\partial^2 L}{\partial R \partial V}\right] \\ -E\left[\frac{\partial^2 L}{\partial V \partial R}\right] & -E\left[\frac{\partial^2 L}{\partial R^2}\right] \end{bmatrix} = \begin{bmatrix} \frac{t_N}{V} & 0 \\ 0 & \frac{N}{2V^2} \end{bmatrix}.$$

In turn, the variance of each one of the parameters in $\hat{p}$ are given by the diagonal values of the inverse of F. For each parameter in p we can then obtain an approximation of the confidence interval limits assuming a confidence level $(1-\alpha)\times 100\%$, denoted by $CI_{(1-\alpha)\times 100\%}$, using $\left(\hat{p} \pm z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var}[\hat{p}]}\right)$, where $\widehat{Var}[\hat{p}]$ represents the estimated variance of p with it's parameters replaced by the ML estimates. More specifically, the respective asymptotic CI, for R and V, are given by the following expressions

$$CI_{(1-\alpha)\times 100\%}(R) = \left(\widehat{R} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{V}}{t_N}}\right),$$

and

$$CI_{(1-\alpha)\times 100\%}(V) = \left(\widehat{V} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{2\widehat{V}^2}{N}} \right),$$

where z_q denotes the q-quantile of the standard normal distribution. In this case, we can also compute the exact confidence intervals, $CI^e_{(1-\alpha)\times 100\%}$, using the exact distributions, as shown in Brites (2010), which are defined as

$$(\widehat{R} - R)\sqrt{\frac{N-1}{N}\frac{t_N}{\widehat{V}}} \sim t_{(N-1)}$$

and

$$\frac{N\widehat{V}}{V} \sim \chi^2_{(N-1)},$$

where $t_{(N-1)}$ represents the t-student distribution and $\chi^2_{(N-1)}$ represents the chi-squared distribution, in both cases with $N-1$ degrees of freedom. Thus, the exact confidence intervals for both R and V are given by the following expressions

$$CI^e_{(1-\alpha)\times 100\%}(R) = \left(\widehat{R} \pm t_{1-\frac{\alpha}{2};(N-1)} \sqrt{\frac{N}{(N-1)}\frac{\widehat{V}}{t_N}} \right)$$

and

$$CI^e_{(1-\alpha)\times 100\%}(V) = \left(\frac{N\widehat{V}}{\chi^2_{1-\frac{\alpha}{2};(N-1)}}, \frac{N\widehat{V}}{\chi^2_{\frac{\alpha}{2};(N-1)}} \right),$$

where $t_{q;(N-1)}$ represents the q-quantile of the t-student distribution with $N-1$ degrees of freedom and $\chi^2_{q;(N-1)}$ represents the q-quantile of the chi-squared distribution also with $N-1$ degrees of freedom.

If we have observed values up to a given time t_N, with $Y(t_N) = y_{t_N}$, and want to obtain a forecast for a given time $t > t_N$, considering that $Y(t)$ is a Markov process, we have

$$E[Y(t)|Y(t_1), \ldots, Y(t_N)] = E[Y(t)|Y(t_N)].$$

From Equation (3.6), we get

$$Y(t)|Y(t_N) \sim \mathcal{N}\Big(Y(t_N) + R(t - t_N), V(t - t_N) \Big).$$

Therefore, we can use for the LT forecasts in each age, for $t > t_N$,

$$\widehat{Y}(t) = \widehat{E}[Y(t)|Y(t_N) = y_{t_N}] = y_{t_N} + \widehat{R}(t - t_N), \tag{3.8}$$

where $\widehat{E}(\cdot)$ represents the approximated value of the mathematical expectation. Since, we do not know the exact value of R, we replace it by it's ML estimate, $\widehat{R}$.

The step-by-step (SS) forecasts are estimated following the same reasoning as in (3.8). However, we update t and the last observed value, as well as the parameter estimates, each time we progress one step in time (in the case of this work, one year).

Finally, using the Monte Carlo simulation method, we obtain an approximation of the forecast error distribution, $\widehat{Y}(t) - Y(t)$, as well as the forecasting confidence intervals. From (3.7), we get the mean and variance of $Y(t_n)|Y(t_{n-1}) = y_{t_{n-1}}$. We used, for each age and gender, the ML estimates in p and simulated a sufficiently large number of trajectories $\boldsymbol{Y(t)}$, say r (in this case, we used $r = 2000$). This way, we obtained up to a certain year t_N the ML estimates, for each one of the r replicas simulated, a new parameter vector $\boldsymbol{p}$, the forecasts $\widehat{\boldsymbol{Y}}(\boldsymbol{t})$ (for $t > t_N$), the forecasting errors $\widehat{\boldsymbol{Y}}(\boldsymbol{t}) - \boldsymbol{Y(t)}$, as well as the empirical mean and variance of these in the group of the r replicas, in order to estimate the mean and variance of the forecasting error.

Let's denote M_t and V_t the respective empirical means and variances. We can obtain an approximation of the limits of $CI_{(1-\alpha)\times 100\%}$, for a certain age and gender considered, by applying

$$\left(M_t \pm z_{1-\frac{\alpha}{2}} \sqrt{V_t}\right).$$

3.2.2 Results

We adjusted the GBM to the observed death rates of the Portuguese population, for each one of the ages selected from the arch of life (ages 0 to 99) and for each gender. For this purpose, we used the variable $Y(t) = \ln\left(\frac{X(t)}{X(0)}\right)$, with $X(t)$ denoting the expected death rate at time t and $X(0)$ denoting the first observed death rate of a given individual (in this case, the death rate of year $t = 0$ is the one related to the year 1940, which was the first year analysed in this work).

Figures 4 and 5 illustrate the estimated parameters of the model used, respectively $\widehat{R}$ and $\widehat{V}$, which represent a different estimated parameter for each age and gender, as well as the asymptotic confidence intervals, CI, and exact confidence intervals, CI^e, associated with each parameter. If we analyse the behaviour of the estimated parameters, we conclude that parameter $\widehat{R}$ has a small increasing tendency, which is more noticeable in the first ages analysed, increasing at a very slow pace after age 20. Furthermore, we also conclude that, although the values of $\widehat{R}$ have an almost similar pattern (increasing tendency in relation with age of the individual), the same cannot be said when considering the estimated parameter $\widehat{V}$, since it displays more fluctuations between each age, which is most noticeable when analysing the ages between 18 and 30 and after age 95 (particularly in individuals of the male gender), thus displaying a totally different pattern when compared to $\widehat{R}$. As for the asymptotic confidence intervals, CI, and exact confidence intervals, CI^e, for each parameter R and V, we used a confidence level of 95% in order to compute their values. For each parameter, the results of $CI_{95\%}$ and $CI^e_{95\%}$ are also depicted in the figures below.

For both parameters, the asymptotic and exact confidence intervals have identical values (on the figures shown in the next page, the representation of both confidence intervals almost overlap each other in most ages for both genders), therefore, there are no substantial advantages related with the use of exact confidence intervals in this work. The confidence intervals range of R and V are approximately proportional to $\sqrt{V}$ and to V. This fact explains the bigger range in the confidence intervals of R compared to the confidence intervals of V. Furthermore, considering parameter R, it explains also the massive range in the confidence intervals when analysing individuals aged 95, or more, of the male gender.

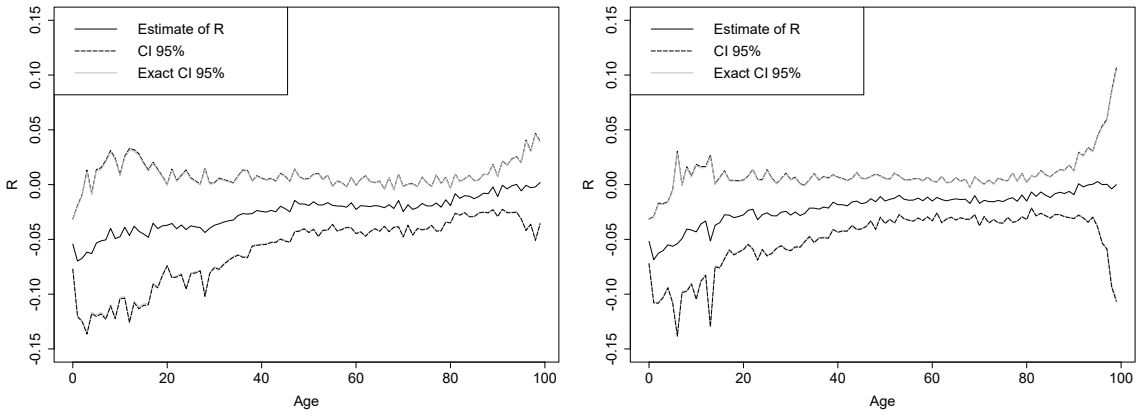

Figure 4: Estimates of R ($\widehat{R}$) of the GBM, $CI_{95\%}$ and $CI^{e}_{95\%}$ values, for each age and gender (female gender on the left and male gender on the right).

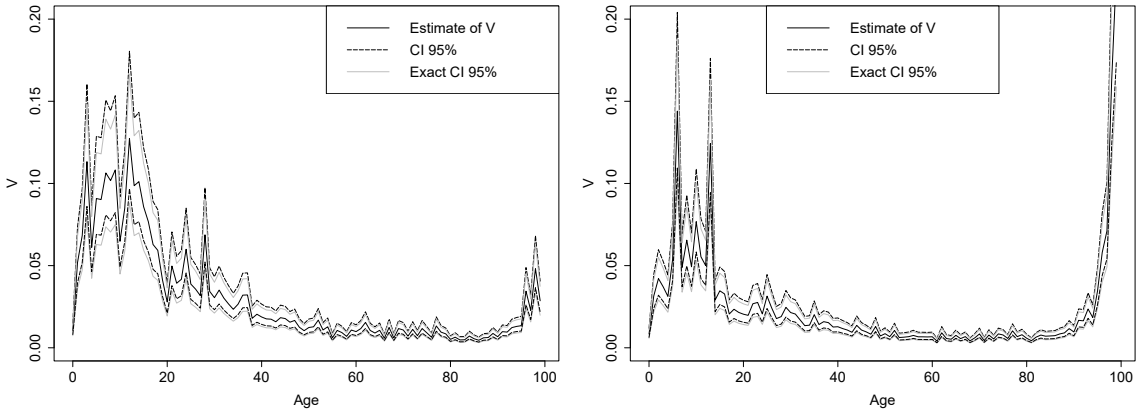

Figure 5: Estimates of V ($\widehat{V}$) of the GBM, $CI_{95\%}$ and $CI^{e}_{95\%}$ values, for each age and gender (female gender on the left and male gender on the right).

The results related with adjustments and forecasts of the death rates where reversed to it's original scale, $X(t)$, instead of $Y(t)$. In Figure 6 we illustrate the adjustment (fixing $\sigma = 0$ in Equation (3.4) and replacing it's parameters with the ML estimates) and forecast results, in this case, for a 15 year old male. Figure 7 shows, respectively, the simulated death rates between the periods of 1940 – 2020 and 2009 – 2020. These simulated death rates where used in order to compute the asymptotic confidence intervals, CI, with a confidence level of 95% for the long term adjustments, as seen in Figure 6.

We recall that we used for the adjustment the observed death rates obtained between the years 1940 – 2009, and set aside the remaining ones (2010 – 2020) for forecasting. Notice that we have chosen to also represent these values in Figure 6 (top) related with adjusted and forecasted values, since they reflect additional information to the error estimate, which stems from the comparison between the tendency and forecasts of the GBM. Generally speaking, the results obtained from the application of the GBM are quite satisfactory, since the model fits well the observed death rates and provides plausible forecasts.

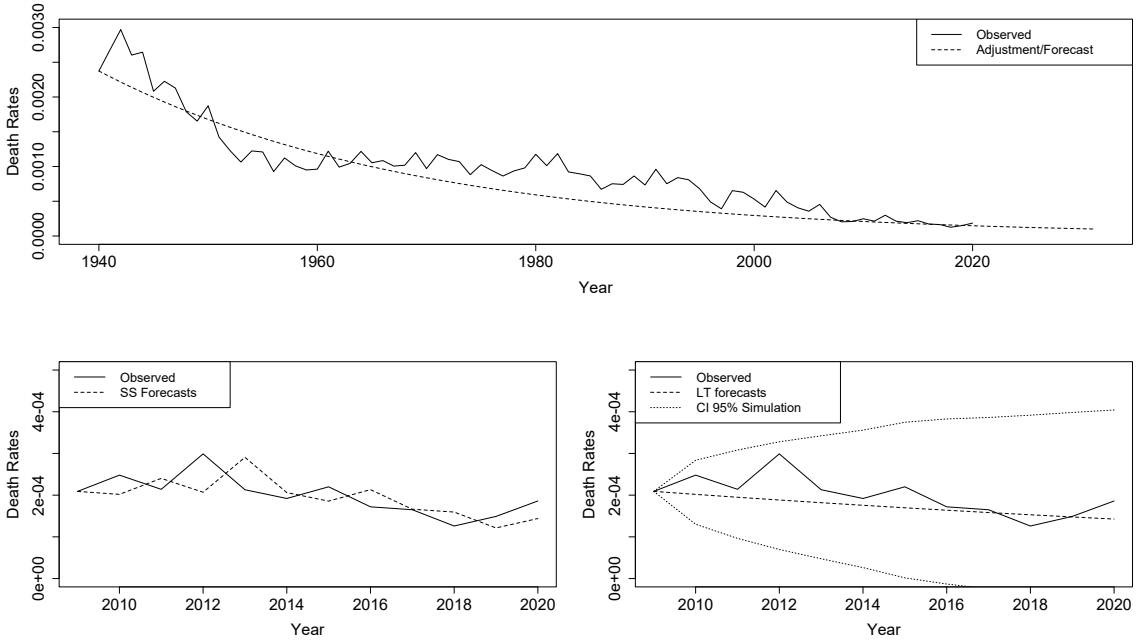

Figure 6: GBM adjustments (for 1940 – 2020) and forecasts (for 2021 – 2030) for a 15 year old male (shown on top); SS and LT forecasts (for 2010 – 2020) with asymptotic $CI_{95\%}$ (on the bottom).

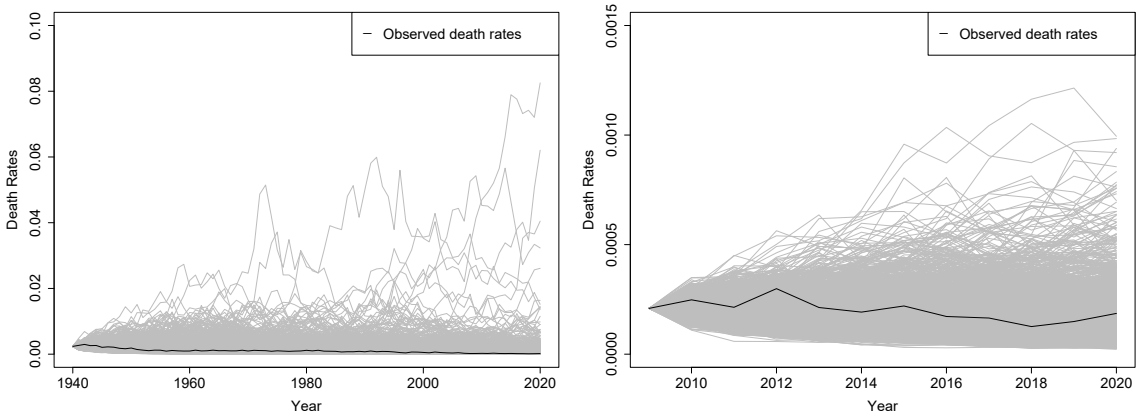

Figure 7: Simulated death rates between the years 1940 – 2020 (on the left side) and between the years 2009 – 2020 (on the right side), using the GBM (with $r = 2000$) for a 15 year old male.

Furthermore, in order to measure the "goodness" of the adjusted values, we used as a quantitative criterion the mean squared error (MSE). In an overall analysis of the results obtained, both adjusted and forecasted values are better fitted (according to the criterion mentioned above), in data series related with the female gender. In Figures 8, 9 and 10 we illustrate the respective MSE for each age and gender and also for each method used (LT and SS).

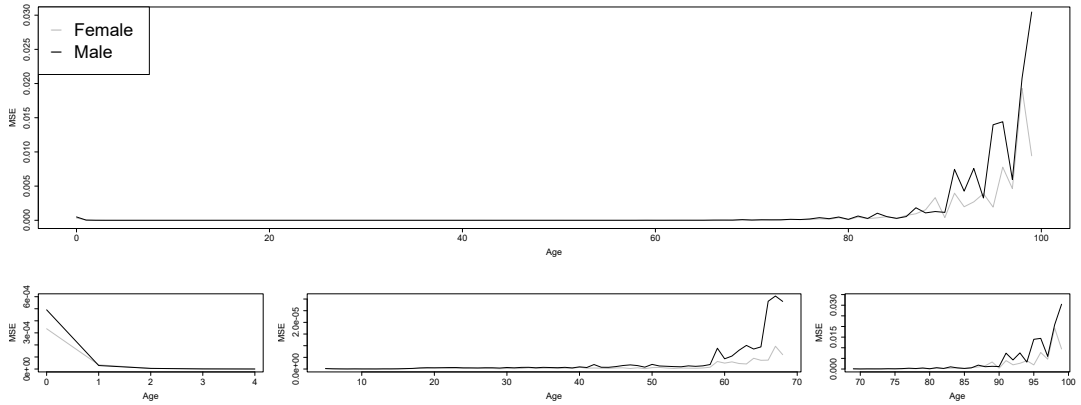

Figure 8: MSE of the adjusted death rates obtained from the GBM, for each age and gender. On the top: representation for all ages. On the bottom: amplification for the groups of ages (0 – 4), (5 – 68) and (69 – 99).

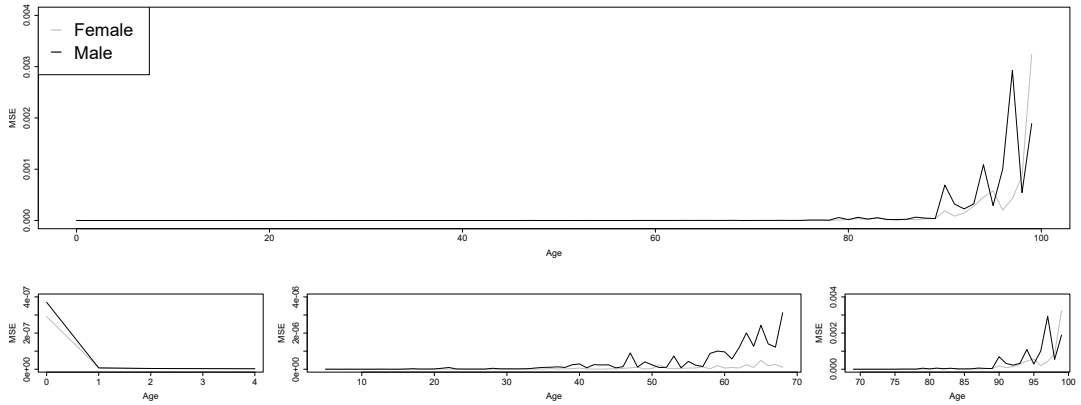

Figure 9: MSE of the LT forecasts obtained from the GBM, for the time period between 2010 – 2020, for each age and gender. On the top: representation for all ages. On the bottom, amplification for the groups of ages (0 – 4), (5 – 68) and (69 – 99).

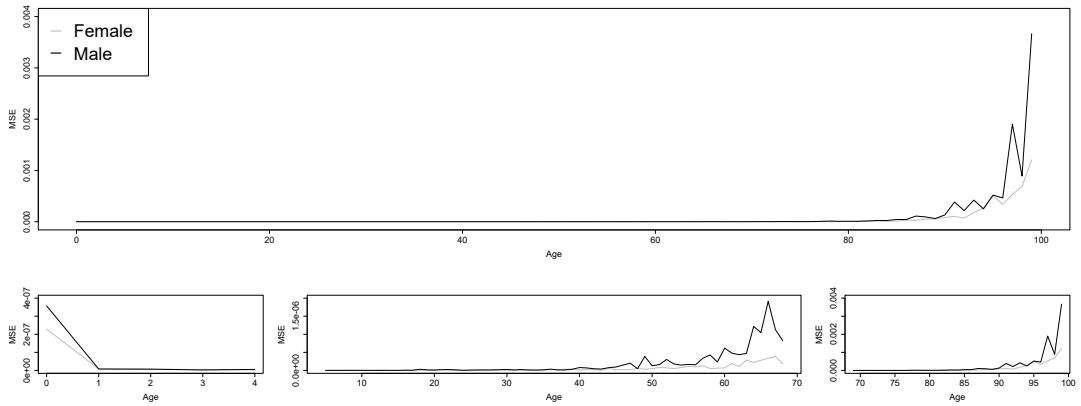

Figure 10: MSE of the SS forecasts obtained from the GBM, for the time period between 2010 – 2020, for each age and gender. On the top: representation for all ages. On the bottom: amplification for the groups of ages (0 – 4), (5 – 68) and (69 – 99).

The difference in the performance of the model between genders is more noticeable after the age of 40 (which corresponds to a set of ages where, throughout time, the mortality pattern of the male gender undergoes an inflexion relative to the prevailing overall downward trend). Also after the age of 90, in both genders, yet more significant in the male gender, the model is not capable of replicating the variability of the death rates time series and of obtaining an adequate adjustment, hence the sharp increase in the MSE values, as illustrated in Figure 8. However, and despite the MSE of the forecasts being extremely high when considering older ages (90+ years) in comparison to other ages (as seen in Figures 9 and 10), the model can still, when dealing with older ages, provide some forecasts to be considered, since they tend strongly towards the observed death rates series averages (see Figure 11).

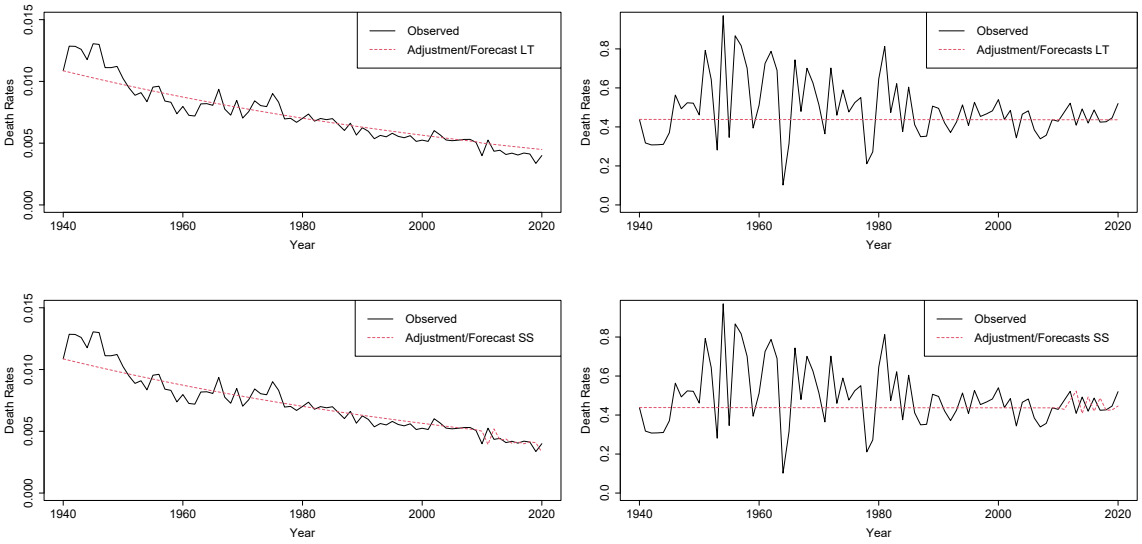

Figure 11: Adjustment of the GBM with LT (on top) and SS (on the bottom) forecasts (2010 – 2020) for the ages 49 (on the left) and 99 (on the right) of the male gender.

3.3 The Stochastic Gompertz model (SGM)

Considering the research topic related to this work, an example of a deterministic model, which can translate the Gompertz law for mortality, can be denoted by

$$dX(t) = bX(t) \ln\left(\frac{a}{X(t)}\right) dt, \tag{3.9}$$

where $X(t)$ represents the death rate (which changes throughout time) of a group of individuals of a given age and gender (which, for now, are fixed), a denotes the asymptotic death rate and b is an approach rate to the asymptotic regimen.

For calculation convenience, let's use $Y(t) = \ln(X(t))$ and $A = \ln(a)$. Thus, we can obtain an equivalent equation from (3.9)

$$dY(t) = -b(A - Y(t))dt. \tag{3.10}$$

To obtain the SGM, we add in (3.10) a noise source, $\epsilon(t)$, where $dW(t) = \epsilon(t)dt$ is a white noise (WN). The standard Wiener process, $W(t)$, reflects the accumulated effect of the "environ-

mental" disruptions which are present in the mortality phenomenon up until a given time t, and the coefficient σ measures the intensity of the environmental variability arising from the random disruptions which affect the variable Y around it's dynamic tendency. This way, we obtain the autonomous SDE

$$dY(t) = -b(A - Y(t))dt + \sigma\epsilon(t)dt = -b(A - Y(t))dt + \sigma dW(t), \tag{3.11}$$

with initial value $Y(0) = \ln(X(0)) = y_0$, which we assume to be known.

Considering the generic form of a SDE presented in (2.2), in this case, for a s.p $Y(t)$, let $f(t, y) = -b(y - A)$, $g(t, y) = \sigma$, and write $Z(t) = e^{bt}(Y(t) - A)$. The solution of (3.11) is obtained by applying the Itô's formula (shown in (2.5)), with $h(t, y) = e^{bt}(y - A)$ and $Z(t) = h(t, Y(t))$, to which we get

$$\begin{aligned}
dZ(t) &= \frac{\partial h(t, Y(t))}{\partial t}dt + \frac{\partial h(t, Y(t))}{\partial Y(t)}dY(t) + \frac{1}{2}\frac{\partial^2 h(t, Y(t))}{\partial Y(t)^2}\ (dY(t))^2 \\
&= \frac{\partial e^{bt}(Y(t) - A)}{\partial t}dt + \frac{\partial e^{bt}(Y(t) - A)}{\partial Y(t)}dY(t) + \frac{1}{2}\frac{\partial^2 e^{bt}(Y(t) - A)}{\partial Y(t)^2}\ (dY(t))^2 \\
&= be^{bt}(Y(t) - A)dt + e^{bt}dY(t) + \frac{1}{2}0(dY(t))^2 \\
&= be^{bt}(Y(t) - A)dt + e^{bt}(-b(Y(t) - A)dt + \sigma dW(t)) \\
&= be^{bt}(Y(t) - A)dt - be^{bt}(Y(t) - A)dt + e^{bt}\sigma dW(t) \\
&= e^{bt}\sigma dW(t).
\end{aligned}$$

Integrating, in the interval $[0, t]$, we get

$$\int_0^t dZ(s) = \int_0^t e^{bs}\sigma dW(s),$$

to which the result of the integral will be

$$Z(t) = Z(0) + \sigma\int_0^t e^{bs}dW(s).$$

Inverting the transformation $Z(t) = e^{bt}(Y(t) - A)$, we get

$$e^{bt}(Y(t) - A) = y_0 - A + \sigma\int_0^t e^{bs}dW(s),$$

to which the result of the following integral will be

$$Y(t) = A + (y_0 - A)e^{-bt} + \sigma e^{-bt}\int_0^t e^{bs}dW(s).$$

Considering that the integral function is deterministic, $\int_0^t e^{bs}dW(s)$ has a normal distribution with mean 0 and variance $\int_0^t (e^{bs})^2$, i.e,

$$\mathcal{N}\left(0, \int_0^t (e^{bs})^2 ds\right) = \mathcal{N}\left(0, \frac{\sigma^2}{2b}(1 - e^{-2bt})\right),$$

wherefore

$$Y(t) \sim \mathcal{N}\left(A + (y_0 - A)e^{-bt}, \frac{\sigma^2}{2b}(1 - e^{-2bt})\right).$$

From $Y(t) = \ln(X(t))$, the solution for $X(t)$ is

$$X(t) = \exp\left\{A + (\ln(x_0) - A)e^{-bt} + \sigma e^{-bt}\int_0^t e^{bs}dW(s)\right\}.$$

The deterministic monomolecular model (from the early 19th century), originally proposed in order to describe a first order irreversible chemical reaction, also known as the Mitscherlich model in the scientific fields of plant nutrition and soil fertilisation, can be represented as

$$dY(t) = b(A - Y(t))dt. \tag{3.12}$$

It's solution, assuming $Y(t_0) = y_0$ (known) is

$$Y(t) = A - (A - y_0)\exp\{-b(t - t_0)\}.$$

When considering the scientific field of plant nutrition, the evolution of the growth rate throughout time, $dY(t)/dt$, where we assume that $Y(t)$ is a measurement of biomass, proportional to the difference between a maximum, or asymptotic, biomass and the biomass formed in the meantime, with b denoting the proportionality constant. If we want, the model in (3.12) can be rewritten in a generalized way, setting $\frac{dY^c(t)}{dt} = b(a^c - Y^c(t))$ (in the following examples, we always use $c = 1$). This model is the source of other models (as seen in Brites (2010)), that can be obtained through transformations, by using a variable $Y(t) = h(X(t))$, in which h is a monotonous function of class C^1. An example, is the deterministic Gompertz model, that can be represented as

$$d\ln(X(t)) = b\left(\ln(a) - \ln(X(t))\right)dt,$$

considering Equation (3.12), and setting $Y(t) = h(X(t)) = \ln(X(t))$ and $A = h(a) = \ln(a)$.

The SGM, which we will discuss in more detail in the next pages, is obtained by adding to the deterministic Gompertz model, represented in (3.12), a WN process that approximates the random fluctuation of the Portuguese population death rates (by age and gender).

Let's consider, as in Subsection 3.2, the simplification of notation $X(t) = X_k(t)$, for the death rates of individuals of a given age $i - 1$ and gender j, with $k = i + 100(j - 1)$, on time instant t. Let $Y(t)$ and A defined as in (3.12), that is, $Y(t) = \ln(X(t))$ and $A = \ln(a)$, the SGM can be represented as

$$dY(t) = b(A - Y(t))dt + \sigma dW(t), \quad Y(t_0) = y_0, \tag{3.13}$$

with y_0 denoting the assumed known initial condition, $W(t)$ is the standard Wiener process and $A = \ln(a)$, where a denotes the mean rate of asymptotic mortality, b denotes the velocity of approximation to asymptotic regimen and σ denotes the intensity of the random environmental fluctuations. Note that several h transformations were experimented, according to the recommendations in the reference literature (see, for example, Sokal & Rohlf (1998)), in order to reduce the variance of the observed death rates series and to try to obtain series with a more linear or smooth curved pattern, so as to facilitate modelling, but in fact the logarithmic transformation, used more

frequently to model the growth rates of several variables related with the scientific field of biology, proved to be the most advantageous for this dataset.

The solution of (3.13), for each age and gender, on time instant t, is

$$Y(t) = A + (y_{t_0} - A)\exp\{-b(t-t_0)\} + \sigma\exp\{-bt\}\int_{t_0}^{t}\exp\{bs\}\,dW(s). \tag{3.14}$$

Considering $t_0 = 0$, the equation above can be rewritten as

$$Y(t) = A + (y_0 - A)\exp\{-bt\} + \sigma\exp\{-bt\}\int_{0}^{t}\exp\{bs\}\,dW(s),$$

meaning $Y(t)$ follows a normal distribution with expected value

$$A + (y_0 - A)\exp\{-bt\}$$

and variance

$$\sigma^2\left(\frac{1-\exp\{-2bt\}}{2b}\right),$$

that is,

$$Y(t) \sim \mathcal{N}\left(A + (y_0 - A)\exp\{-bt\}, \sigma^2\left(\frac{1-\exp\{-2bt\}}{2b}\right)\right).$$

Equation (3.13) is an autonomous SDE. In turn, it's solution, (3.14), is an Itô's diffusion and an homogeneous diffusion process with drift coefficient $a(y) = b(A-y)$ and diffusion coefficient $b(y) = \sigma^2$.

3.3.1 Estimation

Assume that $t_0 = 0$ and let $t_n = n,\ n = 0, 1, 2, \ldots, N$, denote the years in which the death rates of Portuguese population, by age and gender, were observed. The transient p.d.f of $Y(t_n)$ given $Y(t_{n-1})$ is

$$f_{Y(t_n)|Y(t_{n-1})}(y_n) = \frac{1}{\sqrt{2\pi s^2}}\exp\left\{-\frac{1}{2}\frac{(y_n-\mu)^2}{s^2}\right\},$$

where

$$\mu = E[Y(t_n) \mid Y(t_{n-1})] = A + (Y(t_{n-1}) - A)\exp\{-b(t_n - t_{n-1})\}$$

and

$$s^2 = Var[Y(t_n) \mid Y(t_{n-1})] = \sigma^2\left(\frac{1-\exp\{-2b(t_n-t_{n-1})\}}{2b}\right).$$

The parameter vector, $p = (A, b, \sigma)$, can also be estimated using the ML methodology. Hence,

$$\begin{aligned} L(p|Y(t_1),\ldots,Y(t_N)) &= \sum_{n=1}^{N}\ln\left(f_{Y(t_n)|Y(t_{n-1})}(y_n)\right) \\ &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(s^2) - \frac{1}{2}\sum_{n=1}^{N}\frac{(Y(t_n)-\mu)^2}{s^2}. \end{aligned} \tag{3.15}$$

To obtain $\hat{p}$ one needs to compute

$$\begin{cases} \frac{\partial L(y;p)}{\partial A}|_{\widehat{A},\widehat{b},\widehat{\sigma}} = 0 \\ \frac{\partial L(y;p)}{\partial b}|_{\widehat{A},\widehat{b},\widehat{\sigma}} = 0 \\ \frac{\partial L(y;p)}{\partial \sigma}|_{\widehat{A},\widehat{b},\widehat{\sigma}} = 0. \end{cases}$$

Fixing $\widehat{b}$ (and following the same reasoning as in Brites (2010)), since it's not possible to obtain explicitly the expressions for the three parameters, we get

$$\widehat{A} = \sum_{n=1}^{N} \left(\frac{Y(t_n) - Y(t_{n-1}) \exp\left\{-\widehat{b}(t_n - t_{n-1})\right\}}{1 + \exp\left\{-\widehat{b}(t_n - t_{n-1})\right\}} \right) \sum_{n=1}^{N} \left(\frac{1 - \exp\left\{-\widehat{b}(t_n - t_{n-1})\right\}}{1 + \exp\left\{-\widehat{b}(t_n - t_{n-1})\right\}} \right)^{-1}$$

and

$$\widehat{\sigma} = \left(\frac{2\widehat{b}}{N} \sum_{n=1}^{N} \left(\frac{\left(Y(t_n) - \widehat{A} - (Y(t_{n-1}) - \widehat{A}) \exp\left\{-\widehat{b}(t_n - t_{n-1})\right\}\right)^2}{1 - \exp\left\{-2\widehat{b}(t_n - t_{n-1})\right\}} \right) \right)^{\frac{1}{2}}$$

Assume, without loss of generality, that $t_n - t_{n-1} = 1$, since the observed death rates of the Portuguese population, obtained from the HMD, are analysed on an annual basis. It follows, from the equations shown above, defining $\widehat{A}$ as a function of $\widehat{b}$, such that $\widehat{A} = \zeta_1(\widehat{b})$, and defining $\widehat{\sigma}$ as a function of both $\widehat{A}$ and $\widehat{b}$, such that $\widehat{\sigma} = \zeta_2(\widehat{A}, \widehat{b})$. Thus we obtain a new function, L^*, with the same optimal values as the log likelihood function defined in (3.15), but which depends solely on the parameter b and can be written in the following way

$$\begin{aligned} L^*(b|Y(t_1), \ldots, Y(t_N)) = &-\frac{N}{2} \ln\left(\frac{\zeta_2(\zeta_1(b), b)^2}{2b} \right) - \frac{1}{2} \sum_{n=1}^{N} \ln(1 - \exp\{-2b(t_n - t_{n-1})\}) \\ &- \frac{b}{\zeta_2(\zeta_1(b), b)^2} \sum_{n=1}^{N} \left(\frac{(Y(t_n) - \zeta_1(b) - (Y(t_{n-1}) - \zeta_1(b)) \exp\{-b(t_n - t_{n-1})\})^2}{1 - \exp\{-2b(t_n - t_{n-1})\}} \right). \end{aligned} \tag{3.16}$$

The ML estimator of b, for each age and gender, is obtained by minimizing the symmetric of (3.16), using, for that effect, the R built-in function *optimize*. This method, described in Franco (2003), and applied in the same way on Brites (2010), uses L^* instead of L to compute the ML estimators of the parameter vector p, and is particularly useful when it's difficult to find an explicit expression for the estimators, with the main advantage of being computationally efficient (without resorting to more complicated numerical methods of implementation). Once we obtain $\widehat{b}$, the ML estimators $\widehat{A}$ and $\widehat{\sigma}$ are obtained, respectively, from $\widehat{A} = \zeta_1(\widehat{b})$ and $\widehat{\sigma} = \zeta_2(\widehat{A}, \widehat{b})$.

To obtain an approximation of the confidence intervals for the parameters, we assume that we are in an asymptotic regimen, considering the ML estimation properties, and we also do, an approximation of the Fisher information matrix by computing the symmetric of the inverse of the Hessian matrix, from whose diagonal we obtain an approximation of the variances related with the estimated parameters. Considering a parameter p and it's ML estimator, $\widehat{p}$, an approximation of the confidence interval, $CI_{(1-\alpha)\times 100\%}$, can be obtained the same way as described in the GBM

case, by using

$$\left(\widehat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var}[\widehat{p}]}\right),$$

where $\widehat{Var}[\widehat{p}]$ represents an estimate of the parameter variance obtained from the inverse of the Hessian matrix using the method described above. If we have observations up until a given time t_N and want to obtain forecasts until a certain time t, with $t > t_N$, considering that $Y(t)$ is a Markov process, we have that

$$E[Y(t)|Y(t_1), \ldots, Y(t_N)] = E[Y(t)|Y(t_N)].$$

Since

$$Y(t)|Y(t_N) \sim \mathcal{N}\left(A + (Y(t_N) - A)\exp\{-b(t-t_N)\}, \sigma^2\left(\frac{1-\exp\{-2b(t-t_N)\}}{2b}\right)\right),$$

we can use for the LT forecasts, considering each age and gender,

$$\widehat{Y}(t) = \widehat{E}[Y(t)|Y(t_N) = y_{t_N}] = \widehat{A} + (y_{t_N} - \widehat{A})\exp\Big\{-\widehat{b}(t-t_N)\Big\}, \tag{3.17}$$

where $\widehat{E}(\cdot)$ represents the approximated value of the mathematical expectation, replacing the exact values of A and b by it's ML estimates, respectively, $\widehat{A}$ and $\widehat{b}$.

The SS forecasts are estimated the same way as in (3.17) however, we update t and the last observed value, as well as the parameter estimates, each time we progress one step in time (in the case of this work, one year).

With the forecasts obtained, if we wish to compute the confidence intervals of the forecasting errors, given by $\widehat{Y}(t) - Y(t)$, we can use, as an alternative to the Monte Carlo simulation method presented in the previous subsection, the Delta method to estimate the variance of the forecasting errors (see Casella & Berger (2002) and Pestana & Velosa (2002)). This method is used in order to estimate the expected value and variance of the parameter functions, using, for that effect, the estimates of the expected value and variance of the parameters. In (3.13), we denoted $A = h(a) = \ln(a)$, whence, in order to reverse to it's initial parameter a, we, here too, can use the Delta method, by using a function g, such that, $g(A) = h^{-1}(A) = \exp\{A\}$, to obtain, in particular, the limits of $CI_{(1-\alpha)\times 100\%}$ from the expressions $g(\widehat{A}) \pm z_{1-\frac{\alpha}{2}}\dot{g}(\widehat{A})\sqrt{\widehat{Var}[g(\widehat{A})]}$, where $\dot{g}$ represents the derivative of g.

Notice that, for the application of this method, function g has to be differentiable, since this method is based on the Taylor series expansion of that same function (in this case, we use only the linear term), through a generalization of the central limit theorem. Concerning the forecasting errors of the GSM, applying to each age and gender, we make (as shown in Brites (2010))

$$\begin{aligned}\widehat{Y}(t) - Y(t) = g_t(\widehat{A}, \widehat{b}, S) = \widehat{A} + (Y(t_N) - \widehat{A})\exp\Big\{-\widehat{b}(t-t_N)\Big\} \\ - A - (Y(t_N) - A)\exp\{-b(t-t_N)\} - \sigma\exp\{-bt\}S\end{aligned} \tag{3.18}$$

where $S = \int_{t_N}^{t} \exp\{bs\}dW(s)$ has a normal distribution, with null mean and variance equal to $\left(\frac{\exp\{2bt\}-\exp\{2bt_N\}}{2b}\right)$. From the application of the Delta method it follows that

$$E[\widehat{Y}(t) - Y(t)] \approx g_t(A, b, 0) = 0,$$

and

$$Var[\widehat{Y}(t) - Y(t)] \approx E\left[\left((\widehat{A} - A)\frac{\partial g_t(A,b,0)}{\partial A} + (\widehat{b} - b)\frac{\partial g_t(A,b,0)}{\partial b} + S\frac{\partial g_t(A,b,0)}{\partial S}\right)^2\right],$$

whereas the variance can be broken down in two terms: a term V_P, which corresponds to the variability of the parameter estimation errors, and a term V_E, associated to the variability due to the random environmental fluctuations (through the stochastic integral). We can then obtain an approximation of the variance in the form of

$$Var[\widehat{Y}(t) - Y(t)] \approx V_P + V_E,$$

with

$$V_P = (1 - E_N^b)^2\ Var[\widehat{A}] + (Y(t_N) - A)(t - t_N)^2 E_N^{2b}\ Var[\widehat{b}] - 2(Y(t_N) - A)(t - t_N)E_N^b\ Cov[\widehat{A}, \widehat{b}],$$

where

$$E_N = \exp\{-b(t - t_N)\},$$

and

$$V_E = \frac{\sigma^2}{2b}\left(1 - E_N^{2b}\right).$$

The variances and covariance in V_P can be obtained (approximated values) using the inverse matrix of the symmetric of the empirical Fisher information matrix.

3.3.2 Results

Just like in the previous subsection, related to the GBM, it was possible to adjust the SGM to the data regarding the observed death rates of the Portuguese population, for each age selected from the arch of life (0 – 99 years) and for each gender. For this purpose, we used, in this specific case, the variable $Y(t) = \ln(X(t))$, with $X(t)$ denoting the expected death rate at time t.

In Figure 12, it's illustrated the values of the SGM parameters, a, b and σ, for each age and gender. Recall that we estimated the value $A = \ln(a)$, but we choose to display the parameter in it's original scale, a, which represents the average asymptotic death rate (geometric average). Furthermore, in the same figure, we illustrate the values of the SGM parameters with the last 10 ages excluded (the plots related with these are easily identified, since the age axis only takes values between 0 and 90, while in the first case it takes values between 0 and 100), in order to show in more detail the behaviour of each estimated parameter when analysing adult ages and make it possible to better understand the shape described in each graph (mainly with regard to parameter b).

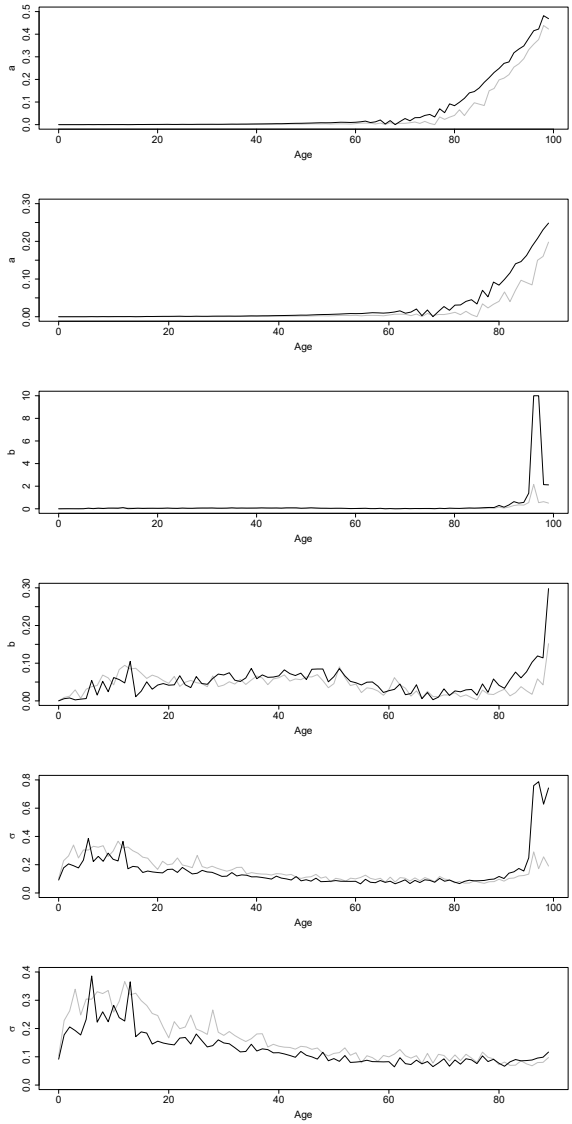

Figure 12: SGM parameter estimates (a, b and σ) for each age and gender (female in grey and male in black), including plots in which the last 10 ages are excluded.

In fact, the results obtained regarding the model's estimated parameters are not surprising, considering the knowledge, obtained from past research projects and articles, about the phenomenon under study. Hence, a, which represents the asymptotic death rate, increases in relation with the age of the individual, presenting much higher values when analysing the last ages from the arch of life, in which the probability of death is higher.

Parameter b can informally be translated as a measure of the speed of approach to the asymptotic regimen. In this case, displays an upward trend, when analysing the first ages of the arch of life (from 0 to 14), followed by a sharp decrease at age 15, and representing several increases and decreases between the years of 16 – 80 but remaining at a level fluctuating, on an average basis, around the value of 0.05 for both genders. After age 80, the estimated values of b increase up to, respectively, twenty and six times it's average values, for the male and female gender.

As for σ, this parameter is associated with the stochastic integral term of the model and measures the intensity of random fluctuations of the environment upon observed death rates. The estimated values present an upward trend in the first ages analysed (which concern children and young people). After age 18, there is a slow decrease in these values, stabilising only between the ages of 60 and 80, after which the pattern described by the parameter shows a new increasing tendency, which translates the susceptibility of the last ages analysed from the arch of life, in which any random event may cause death.

Figure 12 also suggests a greater variability of parameter estimates between consecutive ages for b and σ compared to a. When we observe the pattern of these estimates as a function of age, although it's similar in both genders, in a and b, the estimated values are higher in the male gender than the female one, while the opposite occurs in parameter σ.

Similar to what we have shown for the GBM, we also can, for the SGM, obtain the confidence intervals, CI, for the parameters, considering the asymptotic properties of ML estimation and the approximation of the Fisher information matrix by the symmetrical of the Hessian matrix.

In Figure 13 we illustrate the estimated values, in the original scale of the data, of the adjustment (by fixing $\sigma = 0$ and replacing the model's parameters by it's ML estimates) and of the forecasts for 21 years (from 2010 to 2030) on top. We also show in detail the SS and LT forecasts (for 11 years, from 2009 to 2020) for the death rates, with associated CI, obtained from the Monte Carlo simulation method (which we already used and applied in the GBM case in a similar way).

Globally, the results of the application of the SGM are quite satisfactory. Considering the quality of the adjustment, both the adjustment itself and the forecasts are generally better in the female gender (as it occurred in the GBM). This difference between genders is more significant after the age of 80 (as seen in Figures 14, 15 and 16). Therefore, and similarly to the results obtained in the previous subsection regarding the GBM, the SGM also seems adequate to model this type of data, considering the promising results obtained so far.

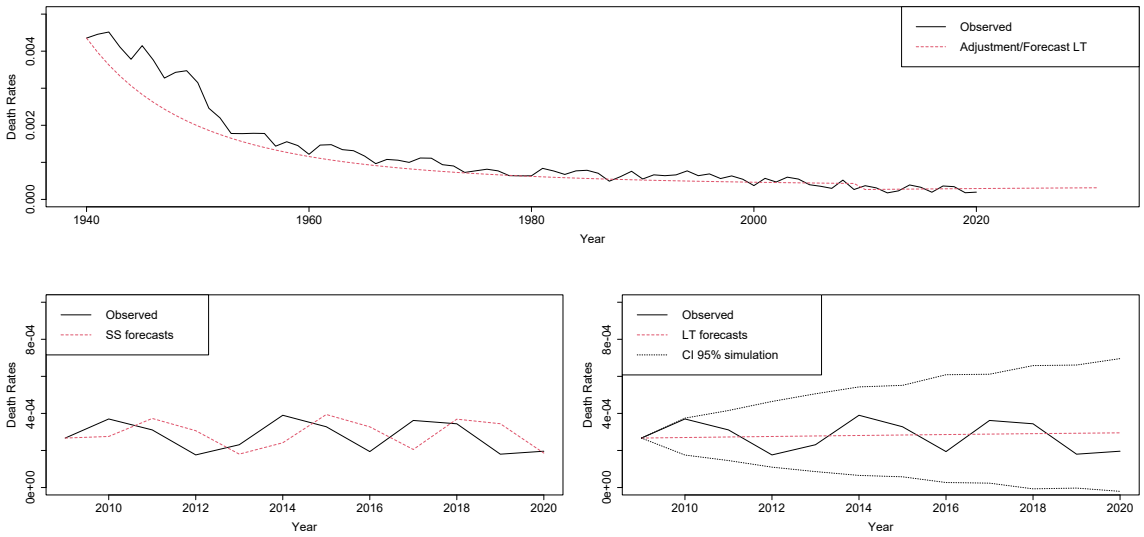

Figure 13: SGM adjustments (for 1940 – 2020) and forecasts (for 2021 – 2030) for a 29 year old female (shown on top); SS and LT forecasts (from 2010 – 2020) with asymptotic $CI_{95\%}$ (respectively, on the left and right, on the bottom).

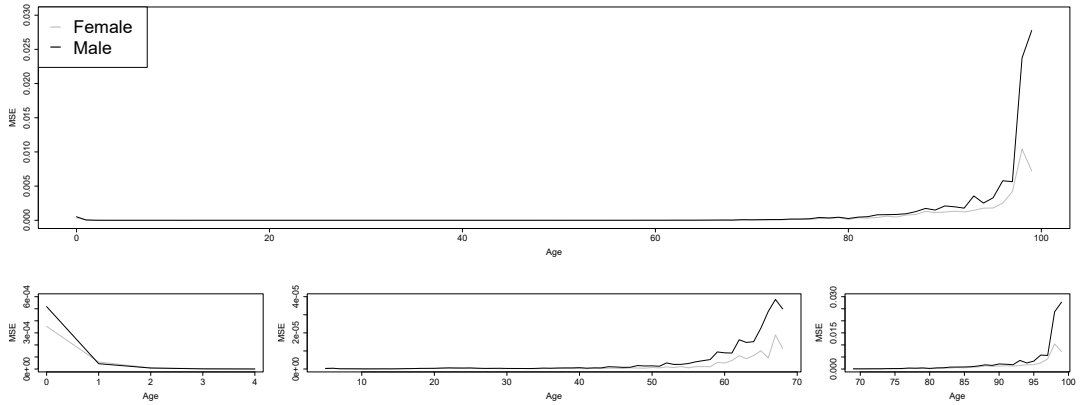

Figure 14: MSE of the adjusted death rates obtained from the SGM, for each age and gender. On the top: representation for all ages. On the bottom: amplification for the groups of ages (0 – 4), (5 – 68) and (69 – 99).

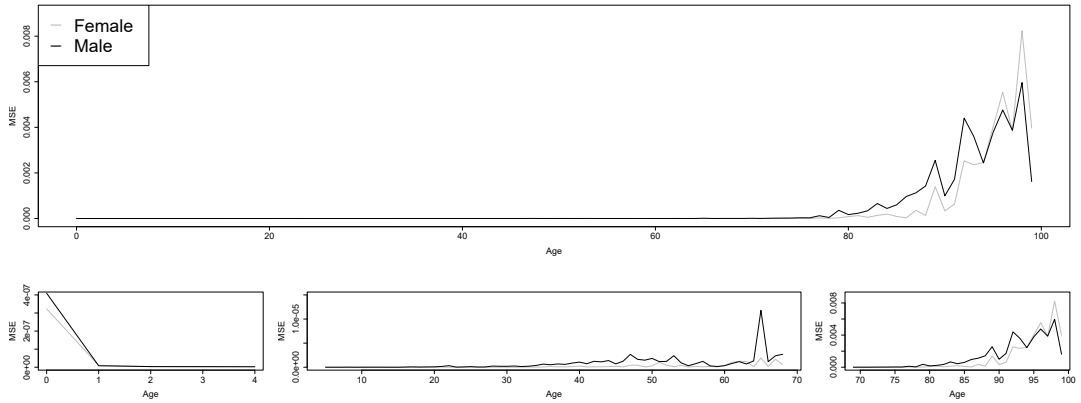

Figure 15: MSE of the LT forecasts obtained from the SGM, for the time period between 2010 – 2020 for each age and gender. On the top: representation for all ages. On the bottom: amplification for the groups of ages (0 – 4), (5 – 68) and (69 – 99).

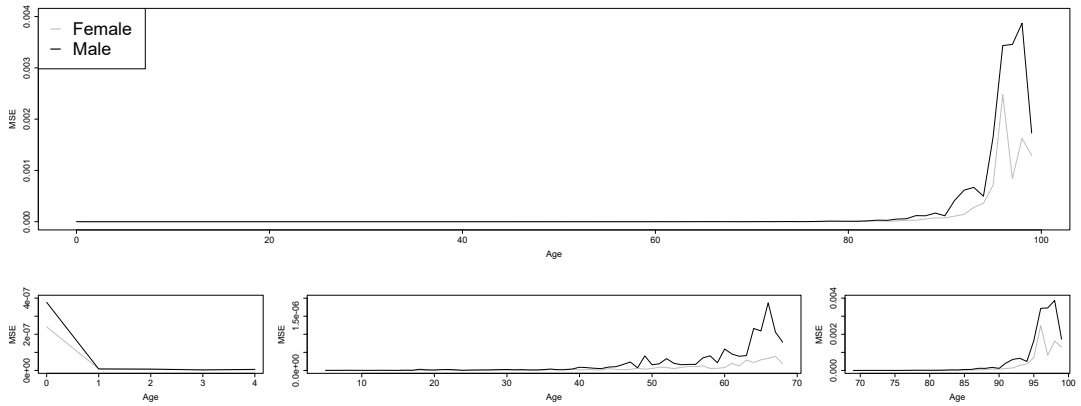

Figure 16: MSE of the SS forecasts obtained from the SGM, for the time period between 2010 – 2020 for each age and gender. On the top: representation for all ages. On the bottom: amplification for the groups of ages (0 – 4), (5 – 68) and (69 – 99).

3.4 Comparison of the results from both models

Recall that the main goal of this work is to capture evolutionary trends in the time series related with the Portuguese death rates, for each age and gender. In this subsection, we compare the results of the two SDEs models applied in the previous subsections (GBM and SGM).

In the course of this research, we have considered (and experimented) variations of GBM and SGM (other transformations of variables, reduction of the time horizon to a stable period for each series, or, instead, separating and estimating the parameters according to the possible phases of the time horizon under analysis). Not recognizing significant advantages, in terms of results, of these alternative models, we have opted to present the results obtained by applying these two SDEs models as described in the previous subsections.

We recall that it was possible to adjust both models and to compute forecasts for all annual age groups from 0 to 99 years old for both genders. We consider that both models present realistic forecasts with values in the same order of magnitude and with close MSEs, which do not allow us to state, in a preliminary analysis, that, one model is generally better than the other. Figure 17 illustrates the application of the two SDE models for age 23 and for both genders (results are presented at the original data scale).

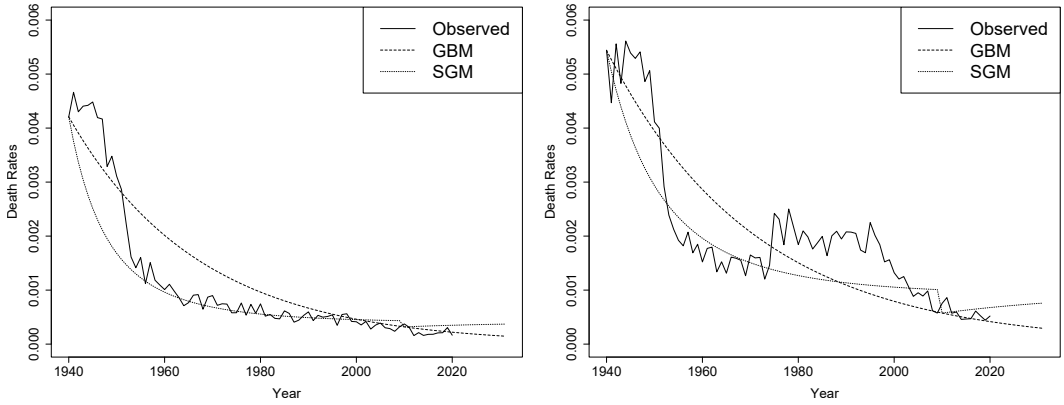

Figure 17: Comparison between the GBM and SGM adjustments with LT forecasts for the age 23 of the female gender (on the left side) and for the male gender (on the right side).

We selected this age (23 years), because it's the typical example of the behaviour of the estimated values, both in terms of adjustment and of forecasting trend, which distinguishes the GBM from the SGM. Thus, for most ages, and for both genders, the adjustment can be represented by an image similar to the one on the left side of Figure 17, since the observed death rates present a near constant downward trend (opposite to what happens in the male case). Note that the curve estimated by the GBM only follows the variability of the series at the beginning and end of the adjustment period, whereas the SGM, although it does not follow the observed death rates curve in the first years, it captures the variability of the series earlier than the GBM. On the right side of Figure 17, the exception to this behaviour is noticeable. Sensitively between the ages of 17 and 37 a "hump" effect occurs in the male gender (in this case, between 1970 and 1999, but this time period can vary depending on the age under analysis) which reflects an increase in mortality in this age group and which causes the main difference in the pattern of mortality between genders.

In terms of forecasts, for most ages (except for ages after 85, where the forecasts have no significant trend, as they tend towards the series average), the GBM underestimates with a decreasing trend (more or less strong, depending on age), while the SGM overestimates with an increasing trend (as can be seen in Figure 17).

Although the performance of neither model stands out explicitly from one another, if we analyse for both models the difference between their respective MSEs, for each age and by gender, the GBM presents advantages over the SGM. In fact, both for the adjustment (exception for some ages, mostly between 25 and 49 years old and also after 85 years old, in the male gender) and SS or LT forecasts (in this case the exceptions are even more occasional and mainly in the last two or three ages) there is a tendency that the error associated to the GBM is lower than the one associated to the SGM.

Figures 18 to 23 depict the differences, for all ages and for each gender, between the MSE associated with the GBM and the SGM, i.e, $MSE_{GBM} - MSE_{SGM}$, for the adjustments, SS forecasts and LT forecasts. Note that due to the order of magnitude of the error estimates, which are often very close and small for several ages, the differences are multiplied by 10000.

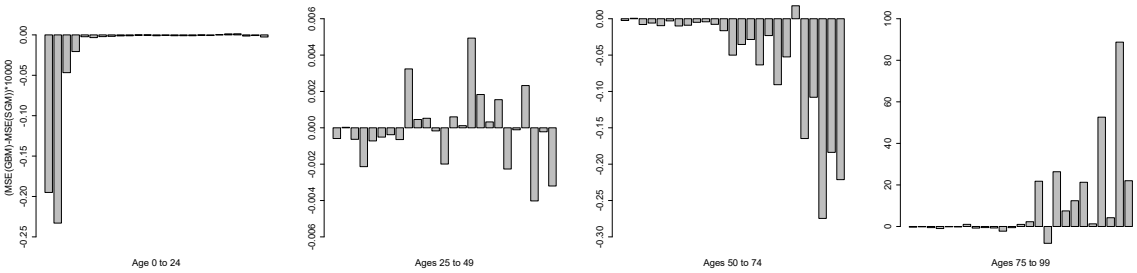

Figure 18: Difference (×10000) between the MSEs associated with the death rates adjustment of the GBM and SGM, for each age of the female gender.

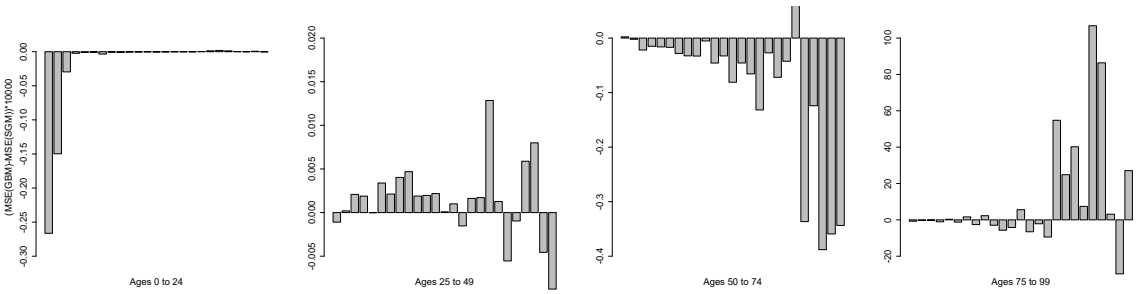

Figure 19: Difference (×10000) between the MSEs associated with the death rates adjustment of the GBM and SGM, for each age of the male gender.

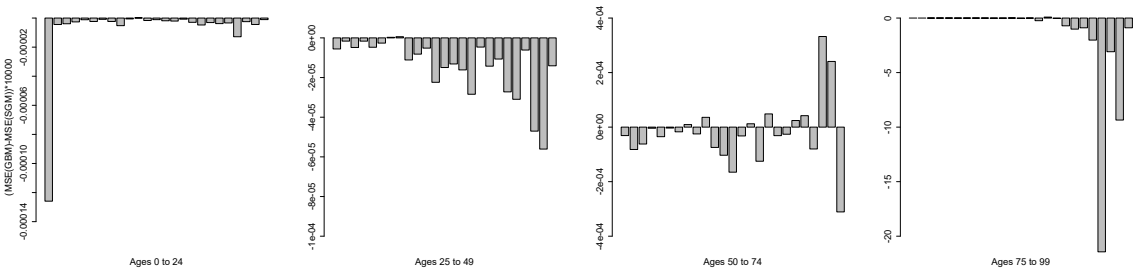

Figure 20: Difference (×10000) between the MSEs associated with the SS forecasts (from 2010 to 2020) of the GBM and SGM, for each age of the female gender.

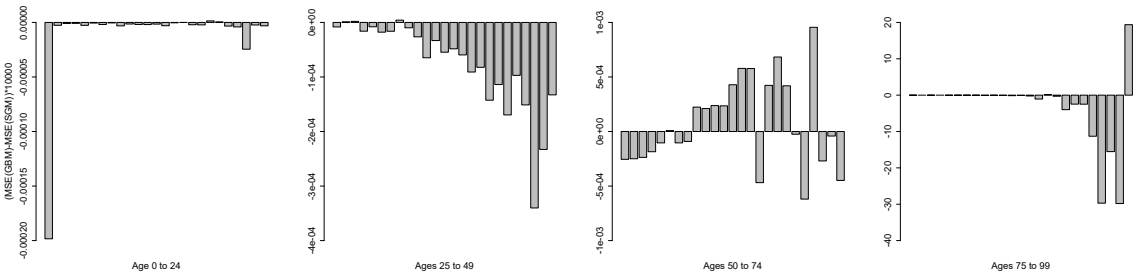

Figure 21: Difference (×10000) between the MSEs associated with the SS forecasts (from 2010 to 2020) of the GBM and SGM, for each age of the male gender.

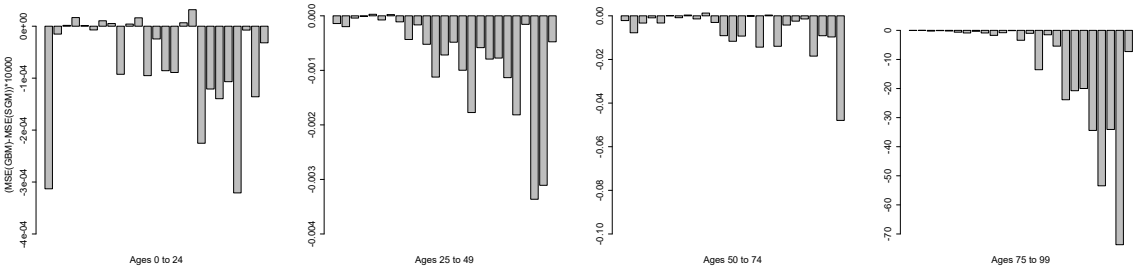

Figure 22: Difference (×10000) between the MSEs associated with the LT forecasts (from 2010 to 2020) of the GBM and SGM, for each age of the female gender.

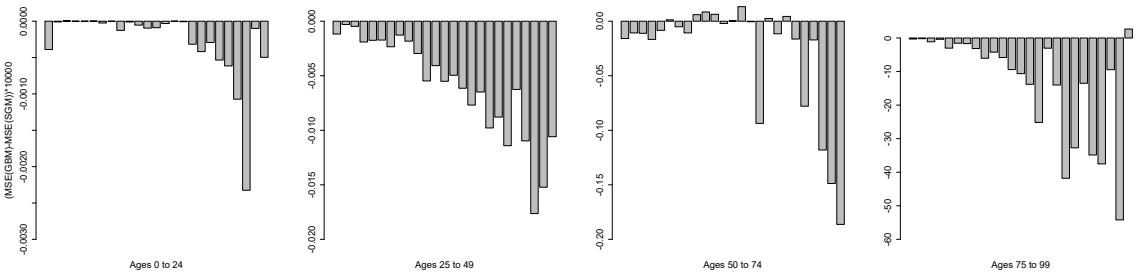

Figure 23: Difference (×10000) between the MSEs associated with the LT forecasts (from 2010 to 2020) of the GBM and SGM, for each age of the male gender.

4 Conclusions

As Benjamin Franklin once said "The only two certainties in life are death and taxes". However it happens that the probability of someone dying, although certain, has been decreasing during the last few decades for individuals belonging to any age group and any gender, at least in Portugal, according to the data provided by the HMD. The decreasing death rates phenomenon is both a blessing (assuming that one values his own life) and a curse, since these decreasing probabilities of mortality combined with the low birth rates observed in the last few decades put a strain in governments worldwide (including the Portuguese one) specially regarding the sustainability of public services and welfare programs.

Thus, explaining by means of a statistical model, the evolutionary trend of the death rates of the Portuguese population over time and computing forecasts, by age and gender, with associated error margins, seems pertinent to us, specially in the current economic and social context and considering the general phenomenon of population ageing.

It's well known that there is no consensus as to which is the best model to explain mortality. There are advantages in the use of different approaches, depending on the purpose for which they are intended. Here, we proposed to perform a cross-sectional analysis of mortality over time and model the death rates associated to each age and gender using SDEs. Furthermore, we intended to explain, in a single model, the mortality over the entire life span (and for both sexes) and compute forecasts.

Given the results obtained, we can conclude that the use of stochastic differential equations death rates models (at least the GBM and SGM) replicate, almost exactly, the decreasing death rates phenomenon observed thus far (meaning they adjust well to the data regarding the mortality in the Portuguese population). Furthermore, both models present realist forecasts with values in the same order of magnitude and with relatively small MSEs, which did not allow us to state, in a preliminary analysis, which model was generally better.

However, in Section 3.4, when the models were compared to one another, it was concluded that the GBM outperforms the SGM in almost all age groups for both genders considering the difference of the MSEs between the models in both SS and LT forecasts. In fact, even when only considering the adjustment, the GBM in most of age groups outperforms the SGM (only in individuals aged 80 or more years for both genders, the SGM outperforms the GBM).

Also, and without surprise, the SS forecasts present a smaller forecasting error when compared to the LT forecasts. This, of course, is logical since in the case of SS forecasts, we update t and the last observed value, as well as the parameter estimates, each time we progress one step in time (in our case, 1 year), meaning, the forecasts will be more accurate, given the added information available and used, than those of the LT forecasts.

In summary, our initial goal was to explain the evolutionary trend of mortality in the Portuguese population and, considering the models applied, we verify that the results of the application of this methodology are satisfactory. However, we accept the hypothesis that there may be one or more variables (we do not know which ones, because they may or may not be observable) that are likely to affect the probability of death in a group of individuals (of the same or different ages and of the same or different genders) in a certain period of time. We believe that the improvement of this

type of models involves extracting even more information from the data of the populations under study, making parameter estimation more flexible and thus improving it's overall performance.

Regarding considerations about future work, we assume that, naturally, the next step could be to apply the proposed models to population data from other countries with different time horizons and analyse their performance. Such exercise could also include variations in the models now proposed and applied, such as the introduction of regime shifts or jumps (to better capture extreme values of mortality).

Another aspect that we consider of interest would be to study the phenomenon from the perspective of magnitudes derived from or complementary to mortality, such as life expectancy, for instance, or to study the time it takes for given death rate to reach a predefined reference value.

Furthermore, and considering only the death rates of the Portuguese population, we could select and apply, to the data provided by the HMD, other types of stochastic models and compare them to the ones applied throughout this work. The stochastic models that could be applied, given the mortality data available, are the Auto Regressive Integrated Moving Average model (ARIMA) or even the Vector Auto Regressive Integrated Moving Average model (VARIMA), since, from a preliminary analysis of the mortality data regarding the Portuguese population, there is a decreasing trend throughout time (from the year 1940 until the year 2020) and these type of models are easily able to detect this trend and compute accurate forecasts.

Another type of stochastic model we could apply is the Binomial Generalized Linear Model (or GLM who belongs to the Binomial family). This type of model has the advantage of computing forecasts for the death rates of the Portuguese Population, using explanatory variables such as economic variables (GDP, unemployment rate, mean and median salary of the population), environmental variables (how much has rained during the year, mean temperature recorded, number of natural disasters and/or forest fires recorded) and even healthcare variables (number of medical appointments, number os surgeries performed, general quality of healthcare services) among several others that can be used, since the models we discussed previously cannot use this type of information. Of course the major downside for this type of models is the quantity of information it requires in order to compute the forecasts (therefore it may not be a wise choice to study the use of this type of models, without first having a solid database with all the values related with these explanatory variables).

References

Arnold, L. (1992), *Stochastic differential equations: Theory and applications*, Krieger Publishing Company, Florida.

Bachelier, L. (1900), 'Théorie de la spéculation', *Annales scientifiques de l'École Normale Supérieure* **17**, 21–86.

Booth, H. & Tickle, L. (2008), 'Mortality modelling and forecasting: a review of methods', *Annals of Actuarial Science* .

Braumann, C. A. (2005), *Introdução às Equações Diferenciais Estocásticas e Aplicações*, Edições SPE, Lisbon.

Braumann, C. A. (2008), 'Growth and extinction of populations in randomly varying environments', *Computers and Mathematics with Applications* **56**, 631–644.

Bravo, J. (2007), Tábuas de mortalidade contemporâneas e prospetivas: modelos estocásticos, aplicações actuariais e cobertura de risco de longevidade, PhD thesis, Universidade de Évora, Évora.

Bravo, J. & Braumann, C. A. (2007), 'The value of a random life: modelling survival probabilities in a stochastic environment', Bulletin of International Statistical Institute.

Bravo, J., Coelho, E. & Magalhães, M. (2010), Mortality projections in portugal, *in* 'Conference of European Statisticians'. UNECE-Eurostat.

Brites, N. M. (2010), Modelos estocásticos de crescimento individual e desenvolvimento de software de estimação e previsão, Master's thesis, Universidade de Évora, Évora.

Brites, N. M. (2017), Stochastic differential equation harvesting models: Sustainable policies and profit optimization, PhD thesis, Universidade de Évora, Évora.

Brites, N. M. & Braumann, C. A. (2019), 'Harvesting in a random varying environment: optimal, stepwise and sustainable policies for the gompertz model', *Statistics, Optimization & Information Computing* **7**, 533–544.

Brites, N. M. & Braumann, C. A. (2020), 'Stochastic differential equations harvesting policies: Allee effects, logistic-like growth and profit otimization', *Applied stochastic models in business and industry* **36**, 825–835.

Casella, G. & Berger, R. (2002), *Statistical Inference*, 2 edn, Duxbury, New Dehli.

Debón, J., Montes, F. & Puig, F. (2008), 'Modelling and forecasting mortality in spain', *European Journal of Operational Research* **189**, 624–637.

Franco, J. (2003), 'Maximum likelihood estimation of mean reverting processes', Real Options Practice - Ownward Inc.

George, M., Smith, S., Swanson, D. & Tayman, J. (2003), *Population Projections*, 2 edn, Elsevier Academic Press, chapter The Methods and Materials of Demography, pp. 561–601.

Human Mortality Database (2022), University of California and Max Planck Institute for Demographic research. [data extracted on 15-02-2022 from http://www.mortality.org].

Instituto Nacional de Estatística (2020), Projeções de população residente em Portugal 2018-2080. [data extracted on 13-07-2022 from http://www.ine.pt].

Jevtic, P., Luciano, E. & Vigna, E. (2013), 'Mortality surface by means of continuous time cohort models', *Insurance: Mathematics and Economics* **53**, 122–133.

Karlin, S. & Taylor, H. E. (1981), *A Second Course in Stochastic Processes*, Academic Press, New York.

Keyfitz, N. & Caswell, H. (2005), *Applied Mathematical Demography (Statistics for Biology and Health)*, 3 edn, Springer, New York.

Lagarto, S. (2014), Modelos estocásticos de taxas de mortalidade e aplicações, PhD thesis, Universidade de Évora, Évora.

Lee, R. (2000), 'The lee-carter method for forecasting mortality, with various extensions and applications', *North American Actuarial Journal* **4**, 80–93.

Lee, R. & Carter, L. (1992), 'Modelling and forecasting the time series of us mortality', *Journal of the American Statistical Association* **87**, 659–671.

Li, S. H. (2007), Stochastic Mortality Models with Applications in Financial Risk Management, PhD thesis, Waterloo University, Canada.

Life Office Mortality Comitee (2007), Stochastic Projection methodologies: Lee-Carter model features, example results and implications. Continuous Mortality Investigation - Working Paper 25.

Mcgehee, M. (2003), *Mortality*, 2 edn, Elsevier Academic Press, chapter The Methods and Materials of Demography, pp. 265–300.

Mendes, M. (2004), 'Mortalidade: Tábua de mortalidade', Universidade de Évora.

Mexia, T. & Corte-Real, P. (1995), 'Tabelas de mortalidade auto-corretivas: o caso português', *Boletim do Instituto de Actuários Portugueses* **35**, 1–120.

Mishra, T. (2008), 'Stochastic demographic dynamics and economic growth: An application and insights from the world data', *Historical Social Research* **33**, 9–187.

Morais, M. (2002), *Causas de Morte no Século XX: Transição e Estruturas da Mortalidade em Portugal Continental*, Edições Colibri and CIDEHUS-UE, Lisbon.

Müller, D. (2007), *Processos Estocásticos e Aplicações*, Edições Almedina, Coimbra.

Namboodiri, K. & Suchindran, C. M. (1987), *Life table techniques and their applications*, Academic Press, Orlando.

Nelson, E. (2021), *Dynamical Theories of Brownian Motion*, 2 edn, Princeton University Press.

Nicolau, J. (2001), Modelação e Estimação de Séries Financeiras através de Equações Diferenciais Estocásticas, PhD thesis, Universidade Técnica de Lisboa. Instituto Superior de Economia e Gestão, Lisbon.

Park, H. S. (2008), 'The survival probability of mortality intensity with jump-diffusion', *Journal of the Korean Statistical Society* **37**, 355–363.

Pestana, D. & Velosa, S. (2002), *Introdução à Probabilidade e à Estatística: Volume I*, Fundação Calouste Gulbenkian, Lisbon.

PORDATA (2021), População residente segundo os Censos: total e por grandes grupos etários. [data extracted on 13-07-2022 from https://www.pordata.pt].

Preston, S., Heuveline, P. & Guillot, M. (2004), *Demography: measuring and modelling population processes*, Blackwell, Oxford.

Skiadas, C. (2010), 'Exact solutions of stochastic differential equations: Gompertz, generalized logistic and revised exponential', *Methodology and Computing in Applied Probability* **12**, 261–270.

Sokal, R. & Rohlf, F. (1998), *Biometry*, 3 edn, Freeman and Company, New York.

Yashin, A. I., Arbeev, K. G., Akushevich, I., Kulminski, A., Akushevich, L. & Ukraintseva, S. V. (2007), 'Stochastic model for analysis of longitudinal data on aging and mortality', *Mathematical Biosciences* **208**, 538–551.

Øksendal, B. (2003), *Stochastic Differential Equations: An Introduction with Applications*, 6 edn, Springer-Verlag, Berlin.