REM WORKING PAPER SERIES

Stochastic differential equations death rates models: the Portuguese case

Daniel dos Santos Baptista, Nuno M. Brites, Alfredo D. Egídio dos Reis

REM Working Paper 0268-2023

April 2023

REM – Research in Economics and Mathematics
Rua Miguel Lúpi 20,
1249-078 Lisboa,
Portugal

ISSN 2184-108X

REM – Research in Economics and Mathematics

Rua Miguel Lupi, 20
1249-078 LISBOA
Portugal

Telephone: +351 - 213 925 912
E-mail: rem@iseg.ulisboa.pt

https://rem.rc.iseg.ulisboa.pt/

https://twitter.com/ResearchRem

https://www.linkedin.com/company/researchrem/

https://www.facebook.com/researchrem/

Stochastic differential equations death rates models: the Portuguese case

Daniel dos Santos Baptista[1*], Nuno M. Brites[1†] and Alfredo D. Egídio dos Reis[1†]

[1*]REM - Research in Economics and Mathematics, CEMAPRE, ISEG – School of Economics and Management, Rua do Quelhas 6, Lisbon, 1200-781, Portugal.

*Corresponding author(s). E-mail(s): l55083@aln.iseg.ulisboa.pt;
Contributing authors: nbrites@iseg.ulisboa.pt;
alfredo@iseg.ulisboa.pt;
†These authors contributed equally to this work.

Abstract

In recent years, the increasing life expectancy of the world's population, due to increased availability of prescribed medication, quality of health care services, quantity of health care institutions and quality of life, combined with a sharp decrease in birth rates over time, has proven to be a challenging problem for governments worldwide (particularly in developed countries). 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 available workforce and tax revenue (including social benefit contributions) in the near future. With the tendency for these problems to worsen in the next decades (severity varies between countries), it is of paramount importance to estimate the extension of human life in order to analyse the severity of this phenomenon. Stochastic differential equations have been used recently to model the evolution of death rates. In fact, such models have some advantages when compared to the deterministic ones since we can input random environmental fluctuations and evaluate the uncertainty in forecasts. The main goal of this paper is to apply and compare stochastic differential equations death rate models separately for each age and sex and forecast Portuguese death rates until the year 2030.

Keywords: Death rates, Geometric Brownian motion, Stochastic Gompertz model, Stochastic differential equations, Forecasting, Life insurance

1 Introduction

In Portugal, like in the majority of western countries, the age structure of the population has been changing, marked by an ageing population due to the combined effect of decreasing birth rates and increasing life expectancy throughout the years. According to some projections for the Portuguese resident population in the years of 2018 – 2080, the ageing population (individuals aged 65 years or more) will represent about 37% of the resident population in 2080, considering the expected scenario, see Instituto Nacional de Estatística (2020).

However, if it's certain that the mortality risk increases 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. Various 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 and Tickle (2008), Aro and Pennanen (2011) and Shryock and Siegel (1976)). 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 work is to apply models of stochastic differential equations (briefly, SDE) that, through cross-sectional analysis of the mortality data over time, allow us to estimate the future trend of the decreasing death rate phenomenon for all age groups and for each sex, and to compute step-by-step (SS) and long-term (LT) forecasts.

The data related to the Portuguese death rates and used throughout this paper was obtained from the 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 (corresponding to the population exposed to death risk in the same age interval). In this manuscript 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 sexes.

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 who 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 to have a very high number of parameters (often more than eight for each cohort, because the mortality trajectory is very irregular).

For the purpose of this approach, see the data representation in Figure 1, where the evolution of mortality in the various phases of the life curve is described. In this case, the year 1994 was fixed, but the shape, usually described in the literature as a "J-shaped curve," has not changed significantly over time despite the reduction in infant mortality and greater longevity in the last few decades.

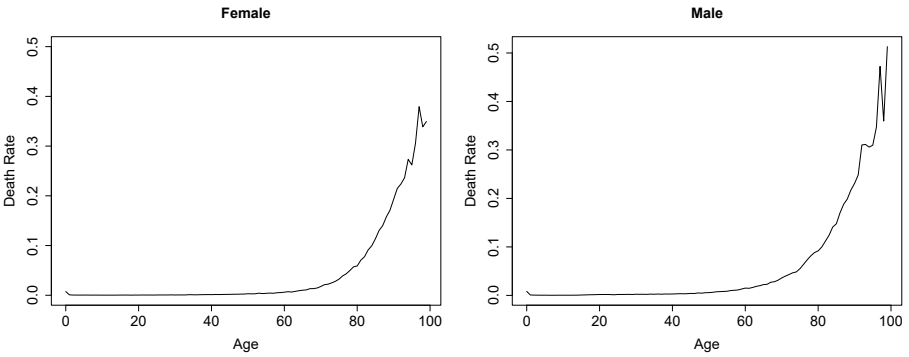

Fig. 1 Death rates of the Portuguese population (longitudinal representation)

Alternatively, the cross-sectional approach we follow makes sense, as we consider events that affect all ages. Among others, we highlight, on the positive side, changes in living conditions over time of a socio-economic nature, advances in medicine, increased quality of health care services and number of health care institutions. Also, climate change that generates 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 trend in the period under analysis, as seen in Figure 2. At almost all ages, the death rates are higher in males than in females, although with a different evolution at each age. Furthermore, throughout this paper, we divided each time series related to the observed death rates of the Portuguese population[1] into two subsets: observed death rates between the years 1940 – 2009 for model adjustment and between the years 2010 – 2020 for forecast validation.

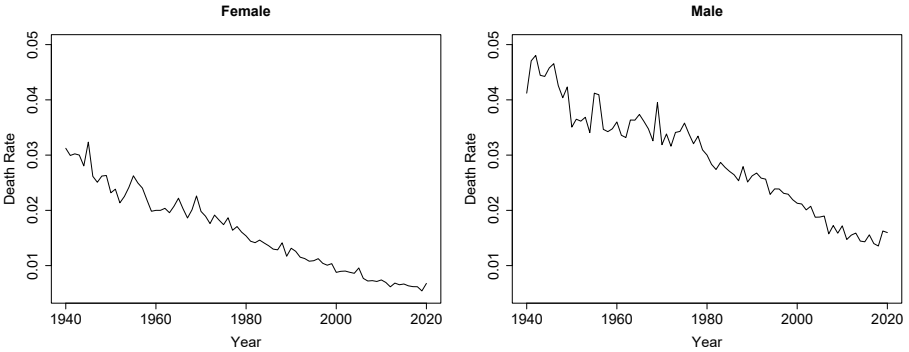

Fig. 2 Death rates of individuals aged 66 along time (cross-sectional representation)

[1] which have 81 observations and are related to the observed death rates documented in each year of analysis, from 1940 to 2020

This paper is organized as follows. In Section 2, we apply both the Geometric Brownian motion and the Stochastic Gompertz model to the Portuguese mortality data, in order to compute adjustments and forecasts. Furthermore, the statiscal aspects of parameter estimation and validation for both models are also analised and model comparison is also performed in order to conclude which model is the best to forecast Portuguese mortality rates. The main conclusions from this research are stated in Section 3.

2 Stochastic differential equations death rates models

2.1 The Geometric Brownian motion

The Geometric Brownian motion (GBM) is a stochastic process usually used to model the price of stocks and other economic variables (as in, for instance, Black and Scholes (1973) and Garcin and Grasselli (2022)). This is also the solution to the stochastic differential equation 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 the volatility. The stochastic differential equation representing the GBM is

$$dX(t) = \mu X(t)dt + \sigma X(t)dW(t),\ \sigma > 0,\ X(0) = x_0, \tag{1}$$

with $W(t)$ representing a standard Wiener process at time t. In this case, $X(t)$ represents 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) and Braumann (2019), as well as other variables in various areas of science. The solution of Equation (1) is:

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

Let's consider that the death rates of the Portuguese population follow a GBM. In this regard, notice that, in fact, when observing the death rates of the Portuguese population throughout time, they appear to have a decreasing linear trend, as was previously seen in Figure 2. Assume $X(t) = X_k(t)$ is the death rate of a given individual aged i with $i = 0, \ldots, 99$ and sex j, with $j = 1$ if female and $j = 2$ if male, at instant t and with $k = i + 100(j-1)$ to cover all ages in the life curve for both sexes. To make reading easier, we use $X(t)$ instead of $X_k(t)$ throughout this section, applying the model to each age and sex. Assume also that the initial condition $X(0) = x_0$ is known. If we denote $Y(t) = h(t, X(t)) = \ln\left(\frac{X(t)}{x_0}\right)$, then $h(t,x) = \ln\left(\frac{x}{x_0}\right)$ is a strictly increasing class C^2 function in x. Applying the Itô's formula we can obtain the SDE

$$dY(t) = Rdt + \sigma dW(t), \quad Y(0) = 0, \tag{3}$$

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

The solution for Equation (3), for each age and gender in instant t, is given by

$$Y(t) = Rt + \sigma W(t), \tag{4}$$

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

$$Y(t) \sim \mathcal{N}\left(Rt, \sigma^2 t\right), \tag{5}$$

where $X(t)$ has a log-normal distribution with expected value $E[X(t)] = x_0 \exp\{Rt\}$. Therefore, we can write Equation (4) in its original form as

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

2.1.1 Estimation

From (5) we obtain the probability density function, $f(t, y)$, of $Y(t)$ which is given by

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

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

$$Y(t_n) = Y(t_{n-1}) + Rt_{n-1}^n + \sigma(W(t_n) - W(t_{n-1})), \tag{6}$$

where $t_{n-1}^n = t_n - t_{n-1}$, the process $Y(t_n)$ conditioned by $Y(t_{n-1})$ has normal distribution with mean $Y(t_{n-1}) + Rt_{n-1}^n$ and variance Vt_{n-1}^n, since $Y(t_{n-1})$ is independent from $W(t_n) - W(t_{n-1})$. Thus, the transition probability density function of $Y(t)$ from t_{n-1} to t_n is given by

$$f_{Y(t_n)|Y(t_{n-1})=y_{n-1}}(y_n) = \frac{1}{\sqrt{2\pi V t_{n-1}^n}} \exp\left\{-\frac{(y_n - y_{n-1} - Rt_{n-1}^n)^2}{2Vt_{n-1}^n}\right\}. \tag{7}$$

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 the maximum likelihood method. 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

$$L(p|Y(t_1), \ldots, Y(t_N)) = \sum_{n=1}^{N} \ln\left(f_{Y(t_n)|Y(t_{n-1})=y_{n-1}}(y_n)\right)$$

$$= -\frac{N}{2}\ln(2\pi V) - \frac{1}{2}\sum_{n=1}^{N}\ln(t_{n-1}^{n}) - \frac{1}{2V}\sum_{n=1}^{N}\frac{(y_n - y_{n-1} - Rt_{n-1}^{n})^2}{2Vt_{n-1}^{n}}.$$

Furthermore, we can obtain the explicit expressions of the maximum likelihood estimators of p (see Brites (2010) and Braumann (2019)), by solving the system

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

obtaining, for t_{n-1}^{n},

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

and

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

Since, here, the death rates of the Portuguese population are annual rates, we can therefore assume that $t_{n-1}^{n} = 1$, 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 maximum likelihood estimators. According to Casella and Berger (2002), the Fisher information matrix for this case is

$$F = \begin{bmatrix} \frac{t_N}{V} & 0 \\ 0 & \frac{N}{2V^2} \end{bmatrix}.$$

On the other hand, 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 its parameters replaced by the maximum likelihood estimates. More specifically, the respective asymptotic CI, for R and V, are given by

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

and

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

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) and Braumann (2019), 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 $t_{(N-1)}$ and $\chi^2_{(N-1)}$ represent the t-Student and Chi-squared distributions, respectively, 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 $t_{q;\,N-1}$ and $\chi^2_{q;\,N-1}$ represent the q-quantile of the t-Student and Chi-squared distributions, respectively, in both cases 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

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

and from Equation (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 long term (LT) forecasts in each age, for $t > t_N$,

$$\widehat{Y}(t) = \widehat{E}[Y(t)|Y(t_N) = y_{t_N}] = y_{t_N} + \widehat{R}(t - t_N), \tag{8}$$

where $\widehat{E}(\cdot)$ represents the approximation value of the mathematical expectation. Since we do not know the exact value of R, we replace it by its maximum likelihood estimate, $\widehat{R}$.

The step-by-step (SS) forecasts are estimated following the same reasoning as to obtain (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 our work, one year).

Finally, using the Monte Carlo simulation method, we obtain an approximation distribution of the forecast error, $\widehat{Y}(t) - Y(t)$, as well as the forecasting confidence intervals. From (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 maximum likelihood estimates for p and simulated a sufficiently large number of trajectories, say r (in this case, we used $r = 2000$), represented by a vector $\boldsymbol{Y(t)}$. This way, we obtained up to a certain year t_N the maximum likelihood 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 CI for $Y(t)$, for a certain age and gender considered, considering

$$CI_{(1-\alpha)\times 100\%}(Y(t)) = \left(M_t \pm z_{1-\frac{\alpha}{2}} \sqrt{V_t} \right).$$

2.1.2 Results

We adjusted the GBM to the observed death rates of the Portuguese population, for each one of the ages selected from the life curve (ages 0 to 99) and for each sex. We used the variable $Y(t) = \ln\left(\frac{X(t)}{X(0)}\right)$ for this purpose, with $X(t)$ representing the expected death rate at time t and $X(0)$ representing the first observed death rate of a given individual.

Figures 3 and 4 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 quite noticeable during 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 a 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.

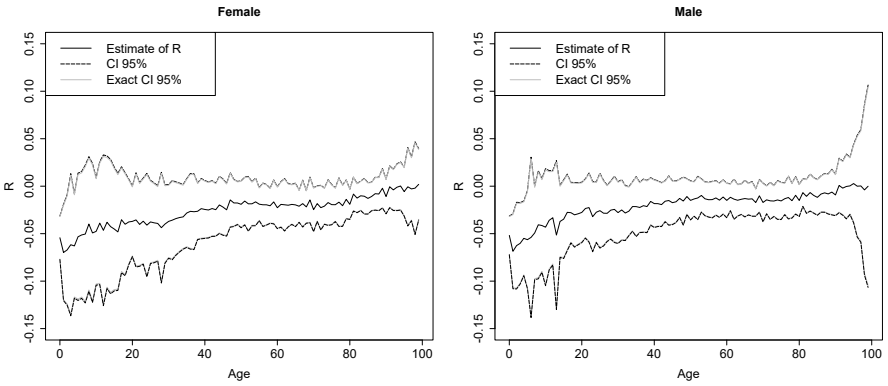

Fig. 3 Estimates $\widehat{R}$, $CI_{95\%}$ and $CI^{e}_{95\%}$ for the GBM

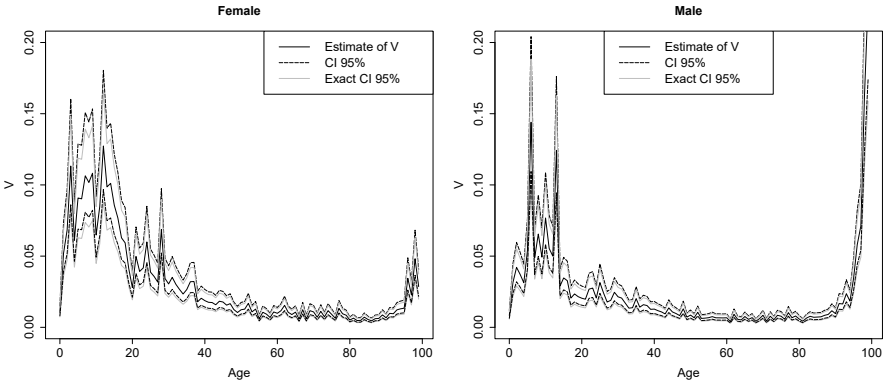

Fig. 4 Estimates $\widehat{V}$, $CI_{95\%}$ and $CI^{e}_{95\%}$ for the GBM

For both parameters, the asymptotic and exact confidence intervals have identical values (in Figures 3 and 4, the representation of both confidence intervals almost overlap each other in most ages and both sexes), therefore, there are no substantial benefits related with the use of exact confidence intervals.

The confidence interval range of R and V are, respectively, approximately proportional to $\sqrt{V}$ and to V. This fact explains the higher range in the confidence intervals of R compared to the confidence intervals of V. Furthermore, with respect to R, it explains also the massive range in the confidence intervals when analysing male individuals aged 95 or more.

Results related with adjustments and forecasts of the death rates where reversed to its original scale, $X(t)$, instead of $Y(t)$. Figure 5 shows the adjustment (fixing $\sigma = 0$ in Equation (4) and replacing its parameters with the maximum likelihood estimates) and forecasts for a 15 year old male.

We recall that we used for the adjustment the observed death rates obtained for the years 1940 – 2009, setting aside the remaining ones (2010 –

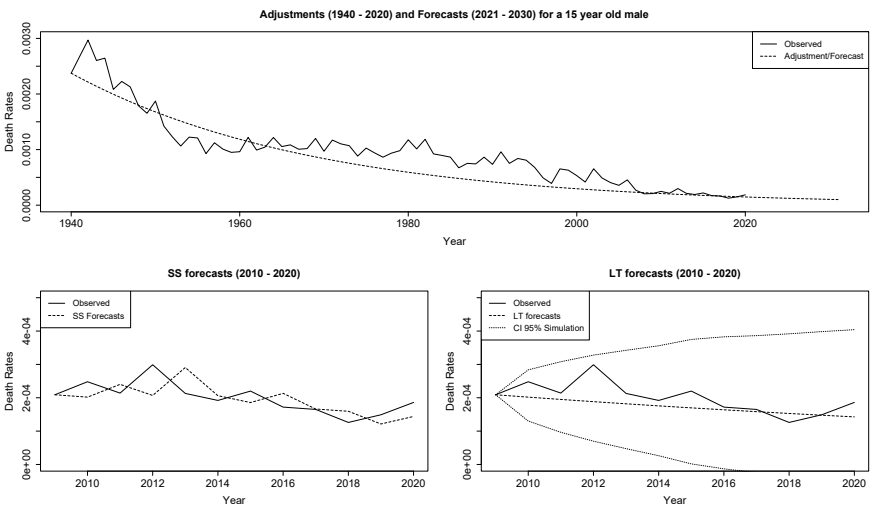

Fig. 5 GBM adjustments and forecasts for a 15 year old male

2020) for forecasting. Also, we have chosen to represent these values in Figure 5 (top) related with adjusted and forecasted values, since they reflect additional information to the error estimate, which stems from the comparison of the tendency and forecasts of the GBM. Generally speaking, the results obtained from the application of the GBM are quite good, since the model fits well the observed death rates and provides reliable forecasts.

Furthermore, in order to measure the "goodness of fit" for the 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 sex. In Figures 6, 7 and 8 we illustrate the respective MSE for each age and sex, also for each method used (LT and SS).

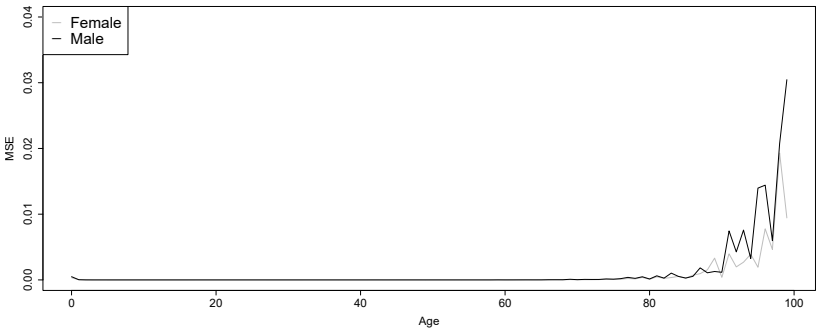

Fig. 6 MSE of the adjusted death rates (1940 – 2020) obtained from the GBM

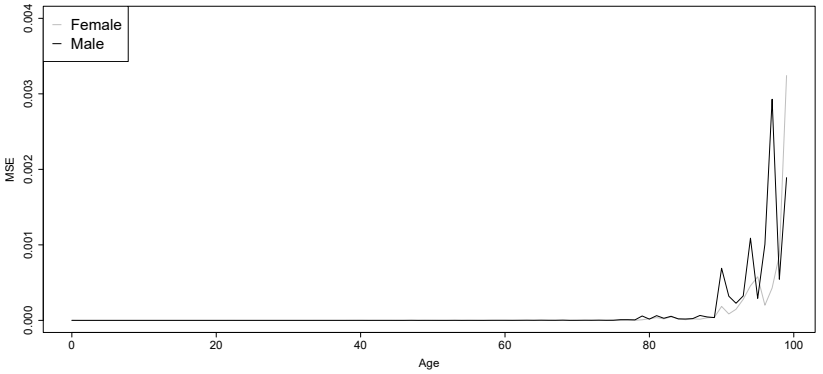

Fig. 7 MSE of the LT forecasts (2010 – 2020) obtained from the GBM

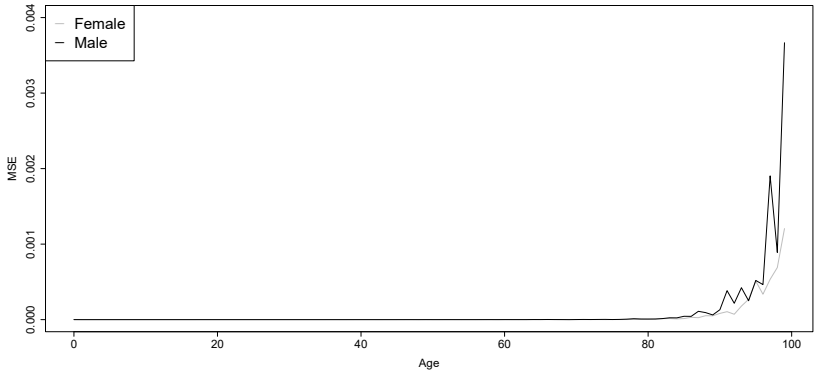

Fig. 8 MSE of the SS forecasts (2010 – 2020) obtained from the GBM

The difference in the performance of the model between genders is more noticeable[2] after the age of 40. Also, after the age of 90, in both sexes, yet more significant in the male sex, the model is not capable of replicating the variability of the death rate time series and of obtaining an adequate adjustment, hence the sharp increase in the MSE values, as illustrated in Figure 6. 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 7 and 8), the model can still provide some forecasts to be considered, since they tend strongly towards the observed death rate series averages (see Figure 9).

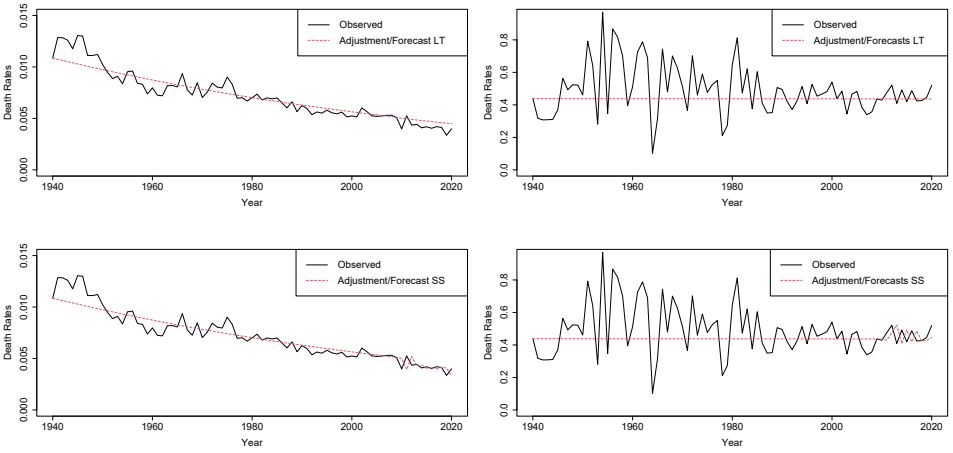

Fig. 9 Adjustment of the GBM with LT (on top) and SS (on the bottom) forecasts (2010 – 2020) for the ages 49 (on the left) and 99 (on the right) of the male sex

2.2 The Stochastic Gompertz model

An example of a deterministic model that can translate the Gompertz law for mortality can be denoted as

$$dX(t) = bX(t) \ln\left(\frac{a}{X(t)}\right) dt, \tag{9}$$

where $X(t)$ represents the death rate (varying throughout time) of a group of individuals of a given age and gender, a the asymptotic death rate and b an approach rate to the asymptotic regimen.

[2] which corresponds to a set of ages where, throughout time, the mortality pattern of the male sex undergoes an inflexion relative to the prevailing overall downward trend

For calculation convenience, let's use $Y(t) = \ln(X(t))$ and $A = \ln(a)$. Thus, we can obtain an equivalent equation from (9)

$$dY(t) = -b(A - Y(t))dt. \tag{10}$$

According to Brites and Braumann (2019a) and Brites and Braumann (2019b), in order to obtain the Stochastic Gompertz model (SGM), we add in (10) a noise source, $\epsilon(t)$, such that $dW(t) = \epsilon(t)dt$. The standard Wiener process, $W(t)$ with parameter σ, reflects the accumulated effect of the "environmental" disruptions which are present in the mortality phenomenon up until a given time t, where the coefficient σ measures the intensity of the environmental variability arising from the random disruptions which affect the variable $Y(t)$ around its dynamic tendency. This way, we obtain the autonomous stochastic differential equation

$$dY(t) = -b(A-Y(t))dt+\sigma\epsilon(t)dt = -b(A-Y(t))dt+\sigma dW(t), Y(t_0) = y_0 \tag{11}$$

with $Y(t_0) = y_0$ denoting the known initial condition, and where a denotes the mean rate of asymptotic mortality, b denotes the velocity of approximation to asymptotic regimen and σ represents the intensity of the random environmental fluctuations.

Let's consider, as before, a simplification of notation $X(t) = X_k(t)$ for the death rates of individuals of a given age i and sex j, with $k = i + 100(j - 1)$, on time instant t.

Note that several h transformations were experimented, according to the recommendations in the reference literature (see for example Sokal and Rohlf (1998)), in order to reduce the variance of the observed death rates series and to try to obtain series with a better linear or smooth curved pattern to facilitate modelling. In fact the logarithmic transformation is used more frequently in modelling the growth rates of several variables in the field of biology, proven to be the most favorable for this dataset.

The solution of (11) for each age and sex is

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

For $t_0 = 0$ we get

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

taking its expectation and variance we obtain

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

2.2.1 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 sex were observed. The transient probability density function of $Y(t_n)$ given $Y(t_{n-1})$ is

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

where

$$\mu = E[Y(t_n) \mid Y(t_{n-1})] = A + (Y(t_{n-1}) - A)\exp\left\{-bt_{n-1}^n\right\},$$

and

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

The parameter vector, $\boldsymbol{p} = (A, b, \sigma)$, can also be estimated. Hence,

$$\begin{aligned} L(\boldsymbol{p}|Y(t_1), \ldots, Y(t_N)) &= \sum_{n=1}^{N} \ln\left(f_{Y(t_n)|Y(t_{n-1})=y_{n-1}}(y_n)\right) \\ &= -\frac{N}{2}\left(\ln(2\pi) + \ln(s^2)\right) \\ &\quad - \frac{1}{2}\sum_{n=1}^{N}\frac{(Y(t_n)-\mu)^2}{s^2}. \end{aligned} \tag{12}$$

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

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

and fixing $\widehat{b}$ (following the same reasoning as in Brites (2010)), we get

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

and

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

Without loss of generality, assume that $t_{n-1}^n = t_n - t_{n-1} = 1$, since the observed death rates of the Portuguese population obtained from Human Mortality Database (2022), are analysed on an annual basis. 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, denoted as L^*, with the same optimal values as the log likelihood function defined in (12), but depending solely on the parameter b, as

$$L^*(b|Y(t_1), \ldots, Y(t_N)) = -\frac{N}{2} \ln\left(\frac{\zeta_2(\zeta_1(b), b)^2}{2b}\right) - \frac{1}{2}\sum_{n=1}^{N} \ln\left(1 - E^2\right) - \frac{b}{\zeta_2(\zeta_1(b), b)^2} \sum_{n=1}^{N} \left(\frac{(Y(t_n) - \zeta_1(b) - (Y(t_{n-1}) - \zeta_1(b))E)^2}{1 - E^2}\right),$$

where $E = \exp\{-bt_{n-1}^n\}$.

We get the maximum likelihood estimator of b, for each age and gender, obtained by minimizing the symmetric of $L^*(\cdot)$ using the R built-in function *optimize*. This method, described in Franco (2003), and applied on Brites (2010), uses L^* instead of L to compute the maximum likelihood estimators of the parameter vector $\boldsymbol{p}$, 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 other complicated numerical methods). Once we obtain $\widehat{b}$, the maximum likelihood estimators $\widehat{A}$ and $\widehat{\sigma}$ are obtained from $\widehat{A} = \zeta_1(\widehat{b})$ and $\widehat{\sigma} = \zeta_2(\widehat{A}, \widehat{b})$, respectively.

To obtain an approximation of the confidence intervals for the parameters, we assume that we are in an asymptotic situation, considering the maximum likelihood estimation properties. 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 vector $\boldsymbol{p}$ and its maximum likelihood estimator $\widehat{p}$, an approximation of the confidence interval, $CI_{(1-\alpha)\times 100\%}$, can be obtained the same way as for 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 seek forecasts until a certain time t, with $t > t_N$, we have that, since we have a Markov Process

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

Since

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

where $t_N^* = t - t_N$, we can use for the LT forecasts, considering each age and sex,

$$\widehat{Y}(t) = \widehat{E}[Y(t)|Y(t_N) = y_{t_N}] = \widehat{A} + (y_{t_N} - \widehat{A})\exp\left\{-\widehat{b}t_N^*\right\}, \tag{13}$$

where $\widehat{E}(\cdot)$ is the approximated value of the mathematical expectation, replacing the exact values of A and b by its maximum likelihood estimates, respectively $\widehat{A}$ and $\widehat{b}$.

The SS forecasts are estimated in the same way as in (13) however, we update t, the last observed value, as well as the parameter estimates, each time we progress one step in time (in this case one year).

2.2.2 Results

Like in Section 2.1.2 related to the GBM, we could adjust the SGM to the observed death rates of the Portuguese population, for each age selected from the life curve (0 – 99 years) and for each sex. For this purpose, we used, in this specific case, the variable $Y(t) = \ln(X(t))$.

Figure 10 illustrates the values of the SGM parameters, a, b and σ, for each age and sex. Recall that we estimated the value $A = \ln(a)$, but we choose to display the parameter in its original scale, a, which represents the average asymptotic death rate (geometric mean). 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).

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 increases in relation with the age of the individual, presenting much higher values when analysing the last ages from the life curve for which the probability of death is higher.

Parameter b displays an upward trend when analysing the first ages of the life curve followed by a sharp decrease at age 15, representing several increases and decreases between the years of 16 – 80 and remaining at a level fluctuating, on an average basis, around the value of 0.05 for both sexes. After age 80, the estimated values of b increase up to twenty and six times its average values for the male and female sexes, respectively.

As for σ, parameter that is associated with the stochastic integral term of the model and measures the intensity of random fluctuations of the environment upon observed death rates, we can say the following: The estimated

values present an upward trend in the younger ages analysed (concerning 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 life curve, in which any random event may cause death.

Figure 10 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 sexes, in a and b, the estimated values are higher in males when compared to females, while the opposite occurs in parameter σ.

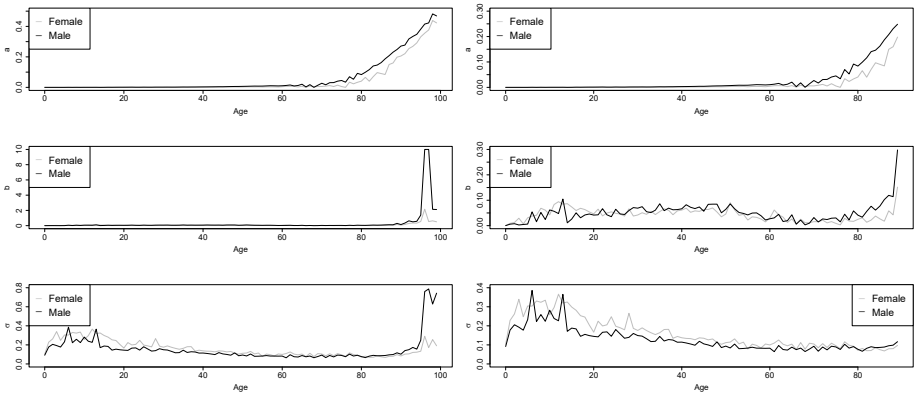

Fig. 10 SGM parameter estimates (a, b and σ) for each age and sex

In Figure 11 we illustrate the estimates of the adjustment (by fixing $\sigma = 0$ and replacing the model parameters by its maximum likelihood estimates) and forecasts for a 29 year old female.

In general, the results of the application of the SGM are quite good. Indeed, both the adjustment itself and the forecasts are generally better in the female sex (like in the GBM). This difference between sexes is higher after the age of 80 (as seen in Figures 12, 13 and 14). Therefore, like 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.

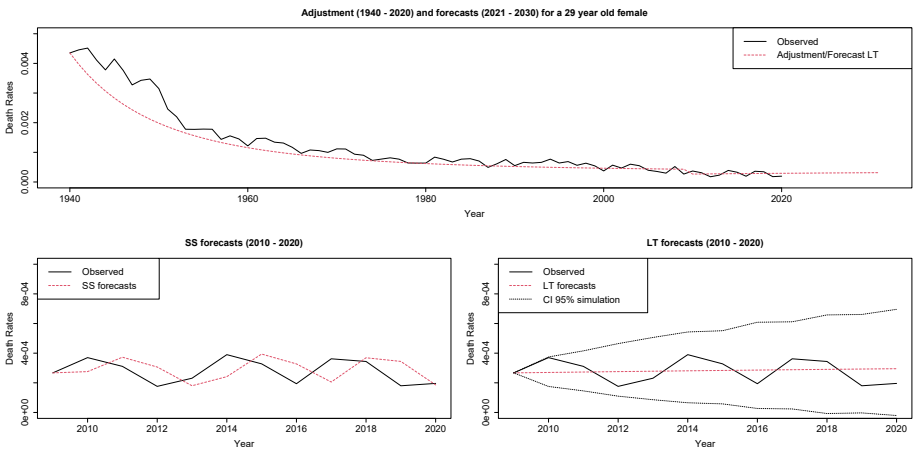

Fig. 11 SGM adjustments and forecasts for a 29 year old female

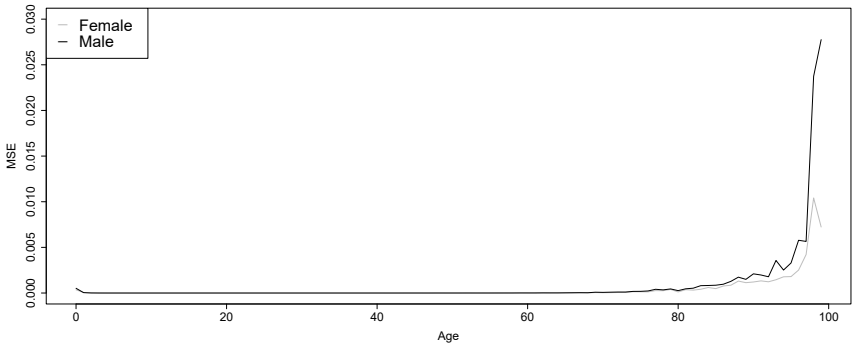

Fig. 12 MSE of the adjusted death rates obtained from the SGM

2.3 Comparing the results from both models

In this subsection, we compare the results of the two stochastic differential equations models applied in the previous subsections, the GBM and the SGM.

We consider that both models present realistic forecasts with values in the same order of magnitude and with close MSE, which do not allow us to state, in a preliminary analysis, that one model is generally better than the other. Figure 15 illustrates the application of the two stochastic differential equations models for age 23 and for both sexes (results are presented at the original data scale).

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 sexes, the adjustment can be represented by an image similar to that of the left side of Figure 15, since the observed death rates present

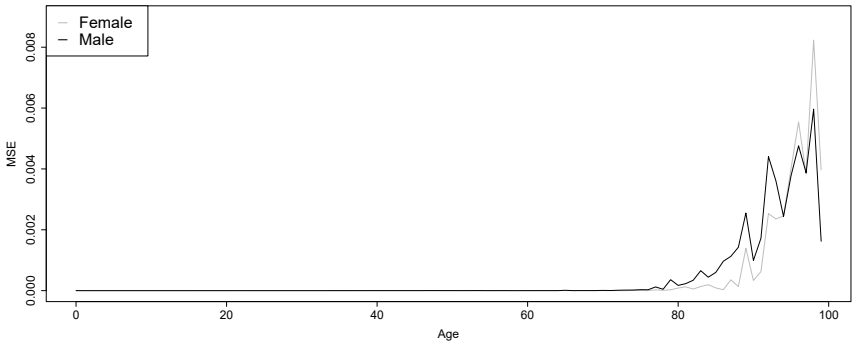

Fig. 13 MSE of the LT forecasts (2010 – 2020) obtained from the SGM

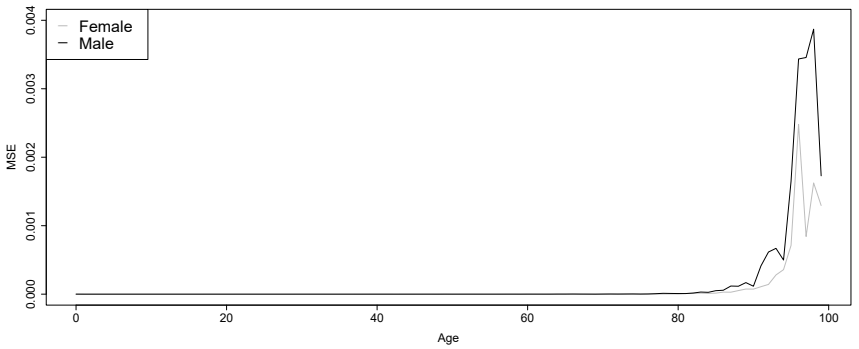

Fig. 14 MSE of the SS forecasts (2010 – 2020) obtained from the SGM

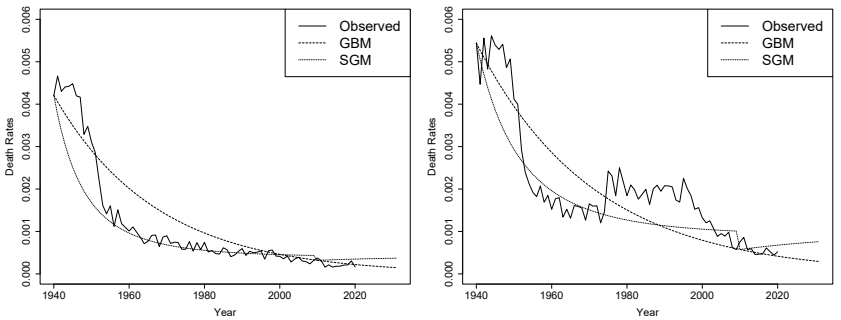

Fig. 15 Comparison between the GBM and SGM adjustments with LT forecasts for the age 23 of the female sex (on the left side) and for the male sex (on the right side).

a near constant downward trend. This is the 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 at the end of the adjustment period, whereas the SGM, although not following 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 15, the exception to this behaviours is noticeable. Sensitively between the ages of 17 and 37 a "hump" effect occurs in the male sex which reflects an increase in mortality in this age group and which causes the main difference in the pattern of mortality between sexes.

In terms of forecasts, for most ages the GBM underestimates with a decreasing trend while the SGM overestimates with an increasing trend (as can be seen in Figure 15).

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 sex, 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 sex) and SS or LT forecasts, there is a tendency that the error associated to the GBM is lower than the one associated to the SGM.

Figures 16 to 21 depict the differences, for all ages and for each sex, 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.

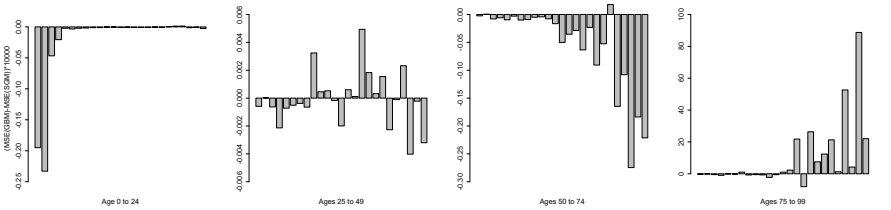

Fig. 16 Difference (×10000) between the MSEs associated with the death rates adjustment of the GBM and SGM, for each age of the female sex.

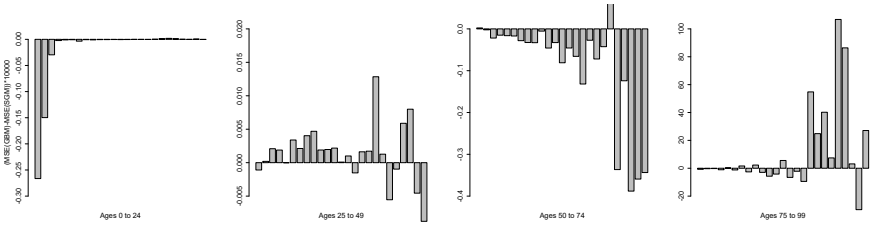

Fig. 17 Difference (×10000) between the MSEs associated with the death rates adjustment of the GBM and SGM, for each age of the male sex.

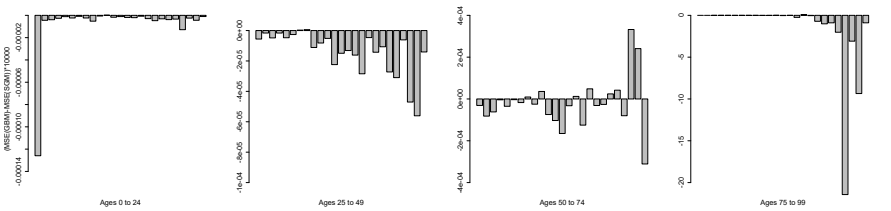

Fig. 18 Difference (×10000) between the MSEs associated with the SS forecasts (from 2010 to 2020) of the GBM and SGM, for each age of the female sex.

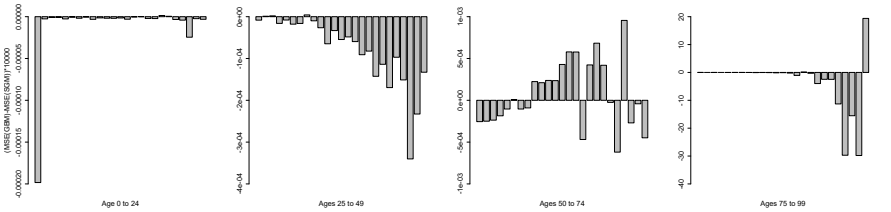

Fig. 19 Difference (×10000) between the MSEs associated with the SS forecasts (from 2010 to 2020) of the GBM and SGM, for each age of the male sex.

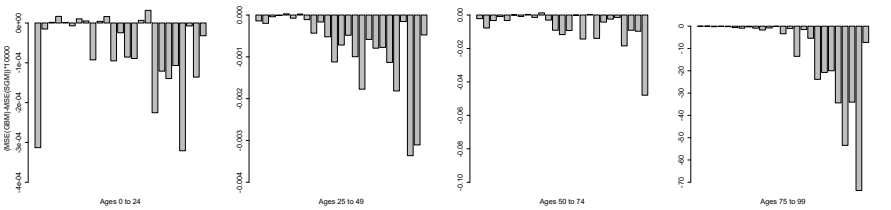

Fig. 20 Difference (×10000) between the MSEs associated with the LT forecasts (from 2010 to 2020) of the GBM and SGM, for each age of the female sex.

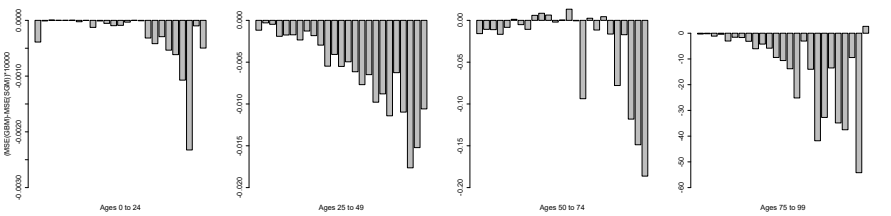

Fig. 21 Difference (×10000) between the MSEs associated with the LT forecasts (from 2010 to 2020) of the GBM and SGM, for each age of the male sex.

3 Conclusions

We can conclude that the use of stochastic differential equations death rate models (the GBM and SGM) replicates almost exactly the decreasing death rate phenomenon observed so far for the Portuguese population. Furthermore, both models present realistic forecasts with values in the same order of magnitude and with relatively small MSE, which did not allowing us to state which model was generally better.

However, in Section 2.3, where the models were compared to one another, we could state that the GBM outperforms the SGM in most of the age groups for both sexes, considering the difference in the MSE between the models in both SS and LT forecasts. Even when only considering the adjustment, the GBM in most age groups outperforms the SGM, only in individuals aged 80 or more years for both sexes, the SGM outperforms the GBM.

Without surprise, the SS forecasts present a smaller forecasting error when compared to the LT forecasts. This is of course 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 further in time. We mean, 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 we verify that the results of the application of this methodology are quite good. However, we accept that there may be one or more variables, we dont know, 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 sexes) in a certain period of time. We believe that improvement of this type of model involves extracting more information from the data of the populations under study, making parameter estimation more flexible and thus improving its overall performance.

Acknowledgments

Nuno M. Brites was partially supported by projects: i) CEMAPRE/REM - UIDB/05069/2020 and ii) EXPL/EGE-IND/0351/2021, both financed by FCT/MCTES through national funds. Alfredo D. Egídio dos Reis was partially supported by project CEMAPRE/REM - UIDB/05069/2020 financed by FCT/MCTES through national funds.

References

Aro H, Pennanen T (2011) A user-friendly approach to stochastic mortality modelling. European Actuarial Journal 1(2):151–167. https://doi.org/10.1007/s13385-011-0030-4, URL https://doi.org/10.1007/s13385-011-0030-4

Black F, Scholes M (1973) The pricing of options and corporate liabilities. Journal of Political Economy 81(3):637–654

Booth H, Tickle L (2008) Mortality modelling and forecasting: a review of methods. Annals of Actuarial Science 3(1-2):3–43. https://doi.org/10.1017/S1748499500000440

Braumann CA (2019) Introduction to Stochastic Differential Equations with Applications to Biology and Finance, 1st edn. Wiley & Hoboken NJ

Brites NM (2010) Modelos estocásticos de crescimento individual e desenvolvimento de software de estimação e previsão. Master thesis available at http://hdl.handle.net/10174/19943

Brites NM, Braumann CA (2019a) Fisheries management in randomly varying environments: Comparison of constant, variable and penalized efforts policies for the gompertz model. Fisheries Research 216:196–203. https://doi.org/https://doi.org/10.1016/j.fishres.2019.03.016, URL https://www.sciencedirect.com/science/article/pii/S0165783619300803

Brites NM, Braumann CA (2019b) Harvesting in a random varying environment: Optimal, stepwise and sustainable policies for the gompertz model. Statistics, Optimization & Information Computing 7(3):533–544. https://doi.org/10.19139/soic.v7i3.830, URL http://www.iapress.org/index.php/soic/article/view/soic.20190902

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

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]

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

Garcin M, Grasselli M (2022) Long versus short time scales: the rough dilemma and beyond. Decisions Econ Finan 45:257–278. https://doi.org/10.1007/s10203-021-00358-3

Shryock HS, Siegel JS (1976) Chapter 23 - population projections. In: Shryock HS, Siegel JS (eds) The Methods and Materials of Demography. Studies in Population, Academic Press, San Diego, p 439–482, https://doi.org/https://doi.org/10.1016/B978-0-12-641150-8.50027-5, URL https://www.sciencedirect.com/science/article/pii/B9780126411508500275

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