Gaussian process regression method for forecasting of mortality rates

Ruhao Wu, Bo Wang*

Department of Mathematics, University of Leicester, Leicester LE1 7RH, UK

Abstract

Gaussian process regression (GPR) has long been shown to be a powerful and effective Bayesian nonparametric approach, and has been applied to a wide range of fields. In this paper we present a new application of Gaussian process regression methods for the modelling and forecasting of human mortality rates. The age-specific mortality rates are treated as time series and are modelled by four conventional Gaussian process regression models. Furthermore, to improve the forecasting accuracy we propose to use a weighted mean function and the spectral mixture covariance function in the GPR model. The numerical experiments show that the combination of the weighted mean function and the spectral mixture covariance function provides the best performance in forecasting long term mortality rates. The performance of the proposed method is also compared with three existing models in the mortality modelling literature, and the results demonstrate that the GPR model with the weighted mean function and the spectral mixture covariance function provides a more robust forecast performance.

Key words: Gaussian process regression, Lee-Carter model, mortality forecasting, spectral mixture, weighted mean function

*Corresponding author. Tel: +44 116 252 2162; Email: bo.wang@le.ac.uk.

1. Introduction

The idea of Gaussian process models can date back to Krige (1951), and had subsequently been further developed in spatial statistics (where it is more widely known as kriging) (Cressie, 1993). O'Hagan (1978) proposed to use Gaussian processes to define prior distributions over functions and applied the theory to one-dimensional curve fitting. Over the last few decades, Gaussian process methods have widely been adopted and further developed in machine learning community; see for example MacKay (1992, 2003), Neal (1992, 1996, 1997), Seeger (2003, 2004), Snelson et al (2004), Quiñonero-Candela and Rasmussen (2005), Rasmussen and Williams (2006), Shi and Choi (2010), and the references therein. Neal (1996) has shown that many Bayesian regression models based on neural networks converge to Gaussian processes as the number of hidden units tends to infinity, and the hyperparameters of the neural network model determine the characteristic length scales of the Gaussian process. Therefore, Gaussian processes have been suggested as a replacement for supervised neural networks in nonlinear regression and classification.

Gaussian process models as a type of nonparametric method have been applied in various fields due to many desirable properties, such as the existence of explicit forms, the ease of obtaining and expressing uncertainty in predictions, the ability to capture a wide variety of behaviour through covariance functions, and a natural Bayesian interpretation. They have been shown to be effective and powerful for the problems of regression, classification, interpolation and extrapolation (forecasting). For the problem of forecasting, Girard et al (2003) studied multiple-step ahead prediction for nonlinear dynamic systems. Brahim-Belhouari and Bermak (2004) proposed to use Gaussian process regression for time series forecasting problem. Mori and Ohmi (2005) used Gaussian process for short-term load forecasting in smart grids. Banerjee et al (2008) proposed Gaussian predictive process models for large spatial data sets. Alamaniotis et al (2011) and Alamaniotis and Tsoukalas (2016) performed short-term load forecasting using an ensemble of Gaussian processes. Wu et al (2012) studied tourism demand forecasting in Hong Kong. Claveria et al (2016) applied Gaussian process regression to the study of Spain's tourism markets. Other examples of Gaussian process models in forecasting include Chapados and Bengio (2007), Ahmed et al (2010), Andrawis et al (2011), Ben Taieb et al (2012), Roberts et al (2013), among others.

In this paper we present a new application of Gaussian process method in the modelling and forecasting of human mortality rates. The growing aged population, especially in developed countries, has given rise to significant changes in both social structures and economic conditions. Assessing and forecasting the demographic mortality trends is hence of great interests to researchers due to its considerable impact on social welfare, resource allocation and governmental budgeting. In addition to biological, medical and behavioural methods, statisticians have developed very different and purely mathematical methods to model the mortality patterns. Lee and Carter (1992) first introduced their statistical model which was then named after them as Lee-Carter model. Lee-Carter model has further been developed by a series of extensions and modifications, including Bell (1997), Lee and Miller (2001), Booth et al (2002), Renshaw and Haberman (2003), Liu and Yu (2011). Hyndman and Ullah (2007) generalised the Lee-Carter model by treating the mortality rates in each year as a curve and applying functional data analysis approach (Ramsay and Silverman, 2005). Chiou and Müller (2009) further extended this method and introduced a moving window approach to collect observed data curves with respect to the birth year of cohorts falling into that window. Lee-Carter model has also been extended to study the mortality for multiple populations; see, for example, Li and Lee (2005), Oeppen (2008), Cairns et al (2011), Hyndman et al (2013), de Jong et al (2016), among others. Booth et al (2006) compared the forecasting performance of some of the variants and extensions of Lee-Carter model.

In this paper we propose to use Gaussian process regression (GPR) method to model and forecast mortality rates. Unlike most of the existing methods in mortality modelling which treat the mortality rates for all ages in a year as a whole and study their evolution over time, we consider the mortality rates for any specified age over time as a time series and assume that they have Gaussian process priors. The advantages of this treatment include that it can capture different patterns in mortality evolution over time for different ages and make use of GPR’s ability in probabilistic forecasting. We consider four conventional Gaussian process regression models and also propose to incorporate a weighted mean function with the spectral mixture covariance function for the problem of mortality modelling. The weighted mean function models the long term trend, and the spectral mixture covariance function enables that various covariance structures in mortality rates for different age groups can be captured and hence mitigates the difficulty in choosing suitable covariance functions in GPR. The combination of these two provides better forecasting results, compared with the conventional GPR models. The performance of this model is also compared with Lee-Miller model (Lee and Miller, 2001), Booth-Maindonald-Smith model (Booth et al, 2002) and the functional demographic model (Hyndman and Ullah, 2007). The numerical results demonstrate that the

GPR model with the weighted mean function and the spectral mixture covariance function provides a more robust forecast performance.

The rest of the paper is organised as follows. In Section 2, we briefly introduce Gaussian process regression models, followed by a detailed description on how this method can be applied to mortality modelling and forecasting. In Section 3, the GPR models are applied to the French total mortality data and are compared with some existing models in the mortality modelling literature. Conclusion and discussions are given in Section 4.

2. Methodology

2.1. Gaussian process regression (GPR)

Let $y \in \mathbb{R}$ be a response variable and $t \in \mathbb{R}$ the covariate variable. Consider the following nonlinear regression model with noise:

$$y = f(t) + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$ represents the measurement error and $f(\cdot): \mathbb{R} \rightarrow \mathbb{R}$ is an unknown function. By Gaussian process method, $f(\cdot)$ is treated as a random function and is assumed to have a Gaussian process prior with a mean function $\mu(\cdot)$ and a covariance function $k(\cdot,\cdot)$. The covariance function relates one point to another and is defined as:

$$k(t, t'; \theta) = Cov\big(f(t), f(t')\big),$$

where θ denotes the set of hyper-parameters which need to be estimated.

Therefore, given the observed data $\mathcal{D} = \{(t_1, y_1), \dots, (t_n, y_n)\}$, we have

$$y_i = f(t_i) + \varepsilon_i$$

where $\{\varepsilon_i\}_{i=1,\cdots,n}$ are independent and identically distributed normal random noises with mean 0 and variance σ^2. Hence the joint distribution of $y_1, y_2, \dots, y_n$ is multivariate normal:

$$\boldsymbol{y} = (y_1, y_2, \dots, y_n)^T \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Psi}),$$

where the mean $\boldsymbol{\mu}$ has entries $\mu_i = \mu(t_i)$ and $\boldsymbol{\Psi}$ is an $n \times n$ matrix whose (i, j)th element is given by

$$\Psi_{ij} = Cov(y_i, y_j) = k(t_i, t_j; \theta) + \sigma^2 \delta_{ij}, \tag{1}$$

where δ_{ij} is the Kronecker delta.

Suppose that t^* is a test point and y^* is the corresponding response value. Then, given the training data $\mathcal{D}$, the conditional distribution of y^* is normal with the following mean and variance (Rasmussen and Williams, 2006):

$$E(y^*|\mathcal{D}) = \mu(t^*) + \boldsymbol{\psi}^T(t^*)\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \,, \tag{2}$$

$$Var(y^*|\mathcal{D}) = k(t^*, t^*; \theta) + \sigma^2 - \boldsymbol{\psi}^T(t^*)\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}(t^*) \,, \tag{3}$$

where $\boldsymbol{\psi}(t^*) = (k(t^*, t_1; \theta), \ldots, k(t^*, t_n; \theta))^T$ is the covariance between $f(t^*)$ and $\boldsymbol{f} = (f(t_1), \ldots, f(t_n))^T$, and $\boldsymbol{\Psi}$ is the covariance matrix of $(y_1, y_2, \ldots, y_n)^T$ defined in (1).

The unknown parameters in the GPR model include the hyper-parameters θ in the covariance function, the noise variance σ^2 and any parameters (denoted generically by β) in the mean function $\mu(\cdot)$. They can be estimated by maximising the following marginal log-likelihood

$$l(\theta, \sigma^2, \beta|\mathcal{D}) = -\frac{1}{2}\log(\det(\boldsymbol{\Psi})) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^T\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) - \frac{n}{2}\log(2\pi). \tag{4}$$

2.2. Gaussian process regression models for mortality forecasting

Let $y_x(t)$ denote the log of the mortality rate for age x in year t. We assume that there is an underlying function $f_x(t)$ that we are observing with error at discrete points of t. Suppose that our observations are $\{t_i, y_x(t_i)\},\ x = 0, \ldots, m,\ i = 1, \ldots, n,$ where m is the maximum age of interest and n is the number of years. Then

$$y_x(t_i) = f_x(t_i) + \varepsilon_{i,x},$$

where $\varepsilon_{i,x}$ represents the random observation error and is assumed to be independent and identically distributed normal random variable $N(0, \sigma_x^2)$ for a given x and $i = 1, \ldots, n$. Based on the observations we are interested in forecasting $y_x(t)$ for any $x \in \{0, \ldots, m\}$ and $t = t_{n+1}, \ldots, t_{n+h}$.

For a given age x, we assume that $f_x(\cdot)$ follows a Gaussian process prior with a mean function $\mu_x(\cdot)$ and a covariance function $k_x(\cdot,\cdot)$. Therefore, for each x we can build a GPR model for the unknown function $f_x(\cdot)$ as discussed in the previous subsection, and the forecast mean and variance of the mortality rate at a future year t^*, $y_x(t^*)$, can then be obtained by the equations (2) and (3).

Since the log mortality rates over time for most ages display an overall decreasing trend, a linear function in t is used as the mean function in the above GPR models, that is:

$$\mu_x(t) = \alpha_x + \beta_x t, \tag{5}$$

where α_x and β_x are constants for the given x. In Gaussian process regression, the covariance function plays an important role in the predictive mean and variance. Covariance functions contain our presumptions about the function we wish to learn and define the closeness and similarity between data points. As a result, the choice of covariance function has a profound impact on the performance of GPR models. A wide range of covariance functions have been proposed and discussed in the literature; see for example Rasmussen and Williams (2006) and Shi and Choi (2011). In this paper we consider three commonly used stationary covariance functions, namely squared exponential (SE), Matern (MA) (with degree of freedom equal to 3/2) and rational quadratic (RQ), and the spectral mixture covariance function (SM) introduced by Wilson and Adams (2013).

The SE, MA and RQ covariance functions have the following forms:

$$k_x(t,t') = k_{SE}(\tau) = \sigma_{SE}^2 \exp\{-\tau^2/(2l_{SE}^2)\},$$

$$k_x(t,t') = k_{MA}(\tau) = \sigma_{MA}^2\left(1+\sqrt{3}\tau/l_{MA}\right)\exp\left(-\sqrt{3}\tau/l_{MA}\right),$$

$$k_x(t,t') = k_{RQ}(\tau) = \sigma_{RQ}^2(1+\tau^2/(2\alpha l_{RQ}^2))^{-\alpha}, \quad (\alpha > 0)$$

where $\tau = t - t'$. Note that in the above covariance functions the dependence of the hyper-parameters on age x is omitted for the sake of simplicity in notations.

The spectral mixture (SM) covariance function is derived by modelling a spectral density − the Fourier transform of a kernel − with a Gaussian mixture. Considering a mixture of Q Gaussians on $\mathbb{R}$, where the qth component has mean μ_q and variance υ_q^2, and letting $\tau = t - t'$, then the spectral mixture covariance function is expressed as

$$k_{SM}(\tau) = \sum_{q=1}^{Q} \omega_q \exp\{-2\pi^2\tau^2\,\upsilon_q^2\}\cos(2\pi\tau\mu_q), \tag{6}$$

where the weights $\{\omega_q\}_{q=1,\cdots,Q}$ specify the contribution of each component. The hyper-parameters $\{\omega_q,\ \mu_q,\ \upsilon_q^2\}_{q=1,\cdots,Q}$ can be estimated by maximising the log-likelihood (4). Q is the number of mixture components and in our numerical examples we have found that $Q = 4$ is sufficient.

2.2.1. GPR model with weighted mean function and SM kernel

In the GPR models, the prior mean function has a significant impact on the forecast performance since the extrapolation tends to move to the prior mean in the long run. In mortality modelling, it is often the case that more recent data tend to have more impact on the results than those in the distant past: the more recent the data point is, the greater influence it tends to have on the future mortality rates. However, the mean function defined by (5) is modelled by a linear regression on the training data, which means each data point in the past carries equal weight on the mean function. Therefore, we propose to model the prior mean function by assigning different weights to the training data points, that is, using weighted least squares (WLS) method to estimate the parameters in the mean function. Furthermore, it can be seen from the numerical examples presented later that the mortality rates for different age groups exhibit very different patterns over time. Therefore different covariance functions may be needed for different age groups, which is not straightforward and is time consuming. However, it is noted that the spectral mixture covariance function can support a broad class of stationary covariance functions and enables that various covariance structures in mortality rates for different age groups can be captured, and hence mitigates the difficulty in choosing suitable covariance functions in the GPR models. Therefore we propose to use a weighted mean function and the spectral mixture covariance function in the Gaussian process regression for the modelling and forecasting of mortality rates. This model makes use of the strengths of both mean function and covariance function and provides a unified method and improved performance for both short term and long term mortality forecasts, as shown in our numerical examples.

For a given age x, the parameters of the linear mean function (5), α_x and β_x, are estimated by minimising the error:

$$e = \sum_{i=1}^{n} w_i [y_x(t_i) - \alpha_x - \beta_x t_i]^2,$$

where w_i is the weight for the ith year. Here we assume the weights to be equal to the inverse of the time distance to the first year to be forecasted, namely t_0 (in the numerical example later on, t_0 =1991), therefore $w_i = 1/(t_0 - t_i)$ for $i = 1, \ldots, n$. It is noted that other weights can be used for the mean function, and if the weights involve tuning parameters, they can be determined by cross validation.

Let $\boldsymbol{W} = \mathrm{diag}\{w_1, \ldots, w_n\}$ and

$$\boldsymbol{X} = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}, \qquad \boldsymbol{y}_x = (y_x(t_1), \ldots, y_x(t_n))^T,$$

then, α_x and β_x are estimated by

$$\begin{pmatrix} \widehat{\alpha_x} \\ \widehat{\beta_x} \end{pmatrix} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y}_x \, .$$

The hyper-parameters in the spectral mixture covariance function (6), $\{\omega_q,\ \mu_q,\ \upsilon_q^2\}_{q=1,\cdots,Q}$, and the noise variance σ_x^2 can then be estimated by maximising the marginal log-likelihood (4).

3. Applications of GPR methods in French mortality rates

In this section we apply the GPR models to French total mortality rates and compare their forecasting performances. The performance of the proposed GPR model with the weighted mean function and the SM covariance function is also compared with three existing models in the literature.

The data are obtained from the Human Mortality Database (2010), consisting of the observed French total mortality rates for ages 0-100 from the year 1950 to 2010. As demonstration Figure 1 shows the log mortality rates for ages 0, 10, 20, 30, 40, 50 and 100. It can be observed that, although the mortality rates generally rise with the increase of age, the mortality rates at birth are relatively high due to babies being born with illness or complications and also vulnerable to illness before their immune system develops. The mortality rates at age 20 are also relatively high because accidental deaths rise during late teen and early twenties. Overall, the log mortality rates for all ages show a downward trend over time, but it can be seen that for different age groups the mortality rates exhibit different patterns, especially for ages 20 and 30.

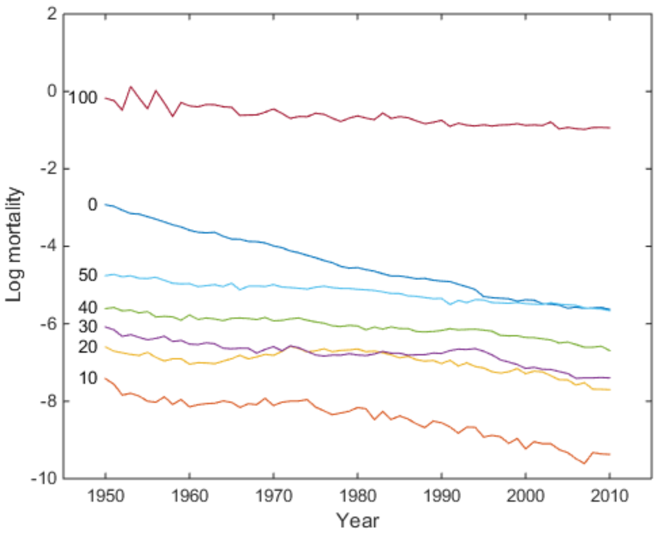

Figure 1. Log French total mortality rates for ages 0, 10, 20, 30, 40, 50 and 100 observed from 1950 to 2010.

3.1. Forecasting comparison of the GPR models

To compare the performances of the different GPR models, we select 20 age groups, namely, ages 0, 1, 2, 5, 10, 12, 15, 18, 20, 22, 25, 28, 30, 40, 50, 60, 70, 80, 90 and 100, to carry out analysis. The data for each age group are split into two parts: the data from 1950 to 1990 are used as training data and those from 1991 to 2010 as testing data. Five GPR models, that is, the models with the linear mean function (5) and the four different covariance functions (squared exponential, Matern, rational quadratic and spectral mixture, denoted by SE, MA, RQ and SM respectively) as well as the proposed GPR with the weighted mean function and the SM covariance function (denoted by WM-SM), are fitted to the training data for each age group separately. The parameters in each model are estimated using 100 random initial values and the ones that give the largest marginal likelihood are used as the estimates. Then the mortality forecasts are made for the years from 1991 to 2010 and are compared with the actual values. As demonstration Figures 2 illustrates the forecasting results by different models for the 40-year age group. The root mean squared errors (RMSE) between the forecasted values and the actual values for each age group are reported in Table 1.

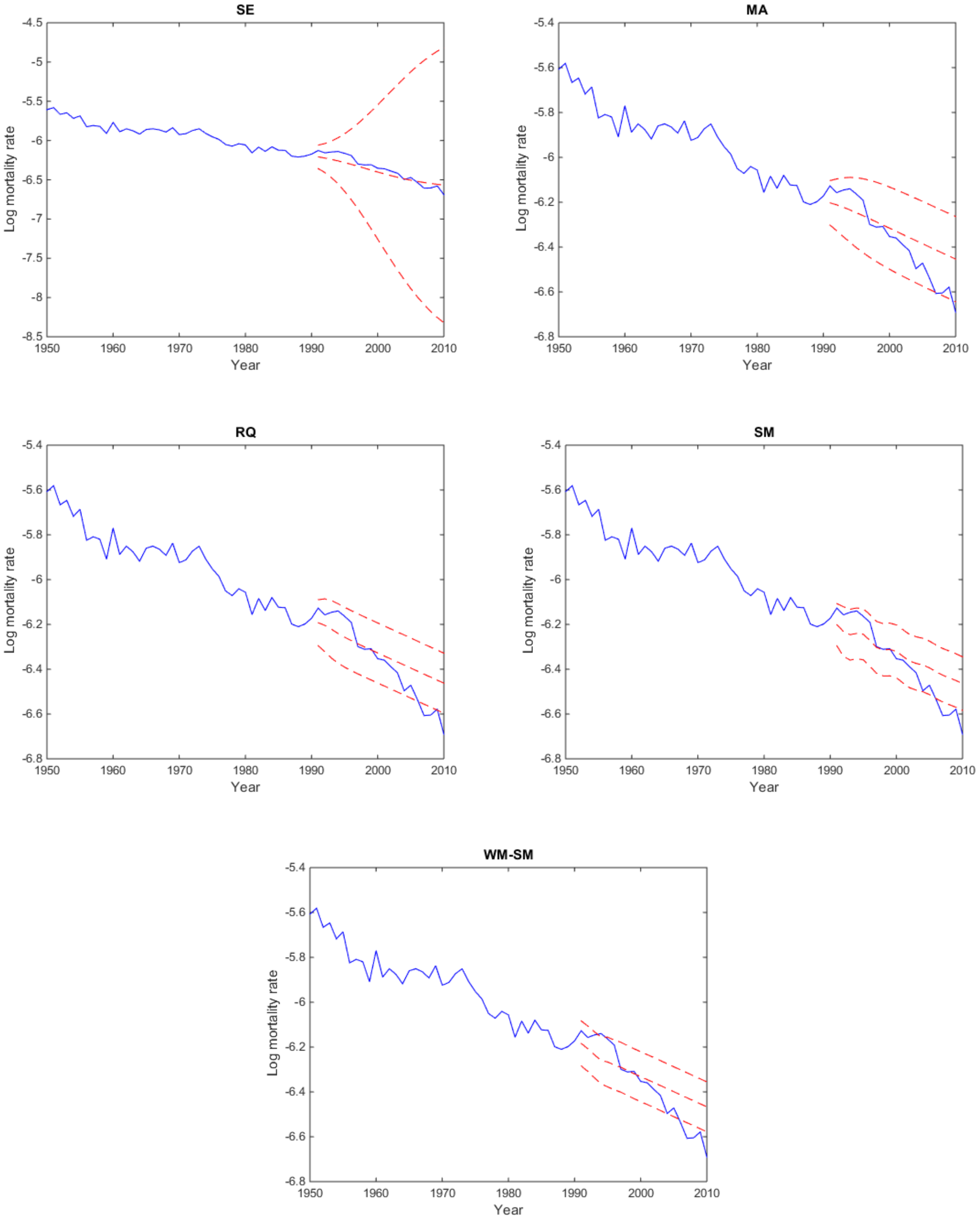

Figure 2. The forecasted mortality rates for 40-year age group using different GPR models. The solid lines are the observed and the dashed lines are the forecasted means and the 95% confidence intervals.

Age	SE	MA	RQ	SM	WM-SM
0	0.2094	0.1825	0.1818	0.1435	0.0919
1	0.1475	0.3491	0.2332	0.3163	0.1685
2	0.2757	0.1023	0.0945	0.0996	0.2047
5	0.1385	0.3759	0.3706	0.3743	0.1850
10	0.4997	0.4289	0.4275	0.4223	0.3610
12	0.3307	0.3086	0.3091	0.3125	0.2132
15	0.3498	0.3023	0.3371	0.3341	0.2356
18	0.8013	0.5920	0.5112	0.4384	0.3251
20	0.3516	0.4245	0.4808	0.5661	0.3878
22	0.6957	0.4796	0.4533	0.4628	0.4125
25	0.2426	0.2485	0.3061	0.3080	0.3367
28	0.3870	0.1888	0.2067	0.1929	0.3052
30	0.4752	0.2810	0.2445	0.2074	0.3217
40	0.1017	0.1111	0.1201	0.1026	0.1003
50	0.3026	0.0499	0.0540	0.0784	0.0386
60	0.0403	0.0899	0.1025	0.1182	0.0705
70	0.0863	0.0376	0.1021	0.1396	0.0622
80	0.1267	0.1438	0.1398	0.1378	0.0528
90	0.0796	0.0797	0.0797	0.0798	0.0588
100	0.1158	0.1158	0.1158	0.1526	0.0695
Average	0.2879	0.2446	0.2435	0.2493	0.2001

Table 1. The RMSEs between the forecasted values and the actual values for each age group by different GPR models.

It can be seen from Table 1 that, although the GPR models with the linear mean function and different covariance functions perform similarly for some age groups (for example the 40-year age group), the choices of the covariance functions still have a significant impact on forecasting accuracy for many other age groups. Taking SE as an example, it performs the best for the 20-year age group, but does poorly for the 50-year group. On the other hand, WM-SM may not provide the best prediction accuracy for some age groups (for example the 25 and 30 age groups), it does significantly improve the overall forecasting accuracy. The average RMSE by WM-SM is the smallest, while RQ comes the second smallest. Wilcoxon signed rank test for the median RMSE by WM-SM being smaller than that of RQ gives a *p*-value of 0.0191, which indicates the former does outperform the latter. The above results show that for a given age group the performance of GPR models largely depend on the choice of covariance functions, but the WM-SM model mitigates this difficulty and provides better overall performance in terms of forecasting accuracy.

Additionally, we also consider the metric of standardised negative log Gaussian predictive density, which is defined as follows (Rasmussen and Williams, 2006):

$$\mathrm{SLL} = \frac{1}{2}\log(2\pi\sigma_*^2) + \frac{(y-y_*)^2}{2\sigma_*^2} - \frac{1}{2}\log(2\pi\bar{\sigma}^2) - \frac{(y-\bar{y})^2}{2\bar{\sigma}^2},$$

where y is the true observation, y_* and σ_*^2 are the predictive mean and variance by the model of interest, and $\bar{y}$ and $\bar{\sigma}^2$ are the sample mean and variance of the training data. Hence the SLL will be zero for the trivial model which predicts using a Gaussian distribution with the mean and variance of the training data and negative for better methods. The average SLLs over the forecasting period (1991 to 2010) for all 20 age groups by the five models (SE, MA, RQ, SM and WM-SM) are -4.6377, -2.3908, -2.9271, -3.1939, -4.0299, respectively. Only SE has smaller average SLL than WM-SM, but we have found that it is due to the very large predictive variances produced by SE, as can be seen in Figure 2.

3.2. Comparison of forecasted mortality curves

We now compare the forecast accuracy of the WM-SM GPR model with some existing models in the literature, namely, Lee-Miller model (Lee and Miller, 2001), Booth-Maindonald-Smith model (Booth et al, 2002) and the functional demographic model (Hyndman and Ullah, 2007).

To construct the forecasted mortality curves for a future year, we select 20 age groups (0, 1, 2, 5, 10, 12, 15, 18, 20, 22, 25, 28, 30, 40, 50, 60, 70, 80, 90 and 100), and fit a WM-SM GPR model for each of them. The mortality curve for all ages (0-100) is then obtained by piecewise cubic spline interpolation. The rationale for age selection is that we want to have dense points in the areas with large variation and sparse points in those with small variation. Our experiment shows that there is no significant difference in the results if more or slightly different age groups are used. Lee-Miller model (LM in short), Booth-Maindonald-Smith model (BMS) and the functional demographic model (FDM) are implemented using R package '*demography'*.

The four models (WM-SM GPR, LM, BMS and FDM) are applied to the French total mortality data using a rolling window approach. That is, we use the data for years from 1950 to Z (where $Z = 1981, \dots, 1990$) to train the models and forecasts are then made for up to 20-year horizon, i.e., to forecast the mortality rates for $Z + 1, \dots, Z + 20$. The forecasts are compared with the actual mortality rates (on log scale) and the RMSEs between the forecast mortality curves and the actual ones over the 20-year horizon for $Z = 1981, \dots, 1990$ are calculated. As illustration Figure 3 presents the forecasted mortality curves and the 95% confidence intervals for the year 1995 (5-year horizon), 2000 (10-year horizon), 2005 (15-year horizon) and 2010 (20-year horizon) by WM-SM GPR model, based on the training data from 1950 to 1990 (i.e. $Z = 1990$). The means and the standard deviations of the ten RMSEs (corresponding to $Z = 1981, \dots, 1990$) for all 20 forecasting horizons are reported in Table 2, and the means at different horizons by the four models are also plotted in Figure 4.

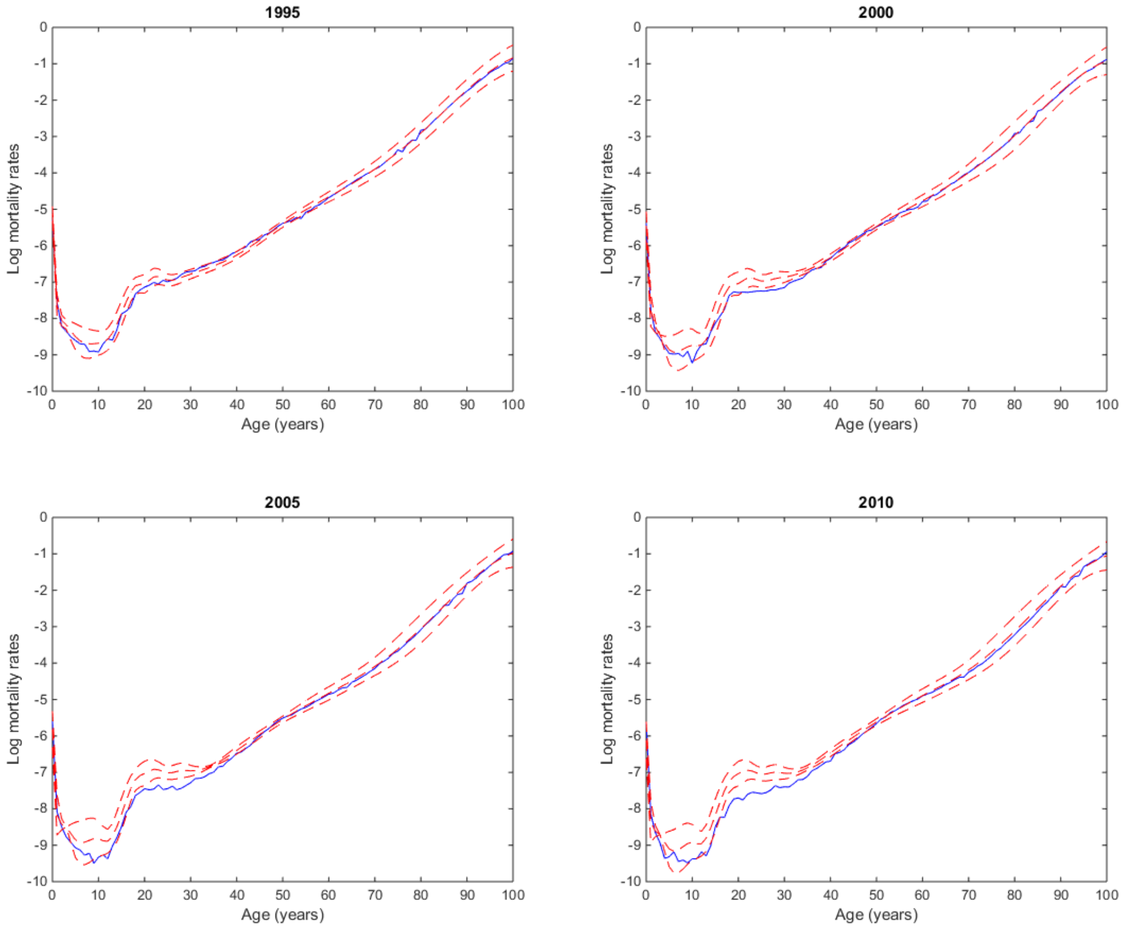

Figure 3. The forecasted mortality curves for 1995, 2000, 2005 and 2010 by WM-SM GPR model, based on the data from 1950 to 1990. The solid lines are the actual values and the dashed lines are the forecasted values and the 95% confidence intervals.

Horizon	LM	BMS	FDM	WM-SM
1	0.0962 (0.0172)	0.0779 (0.0126)	0.0574 (0.0052)	0.0553 (0.0105)
2	0.1096 (0.0185)	0.0906 (0.0166)	0.0652 (0.0062)	0.0686 (0.0114)
3	0.1217 (0.0202)	0.1044 (0.0164)	0.0782 (0.0070)	0.0801 (0.0079)
4	0.1351 (0.0192)	0.1174 (0.0221)	0.0929 (0.0096)	0.0911 (0.0128)
5	0.1475 (0.0183)	0.1287 (0.0264)	0.1089 (0.0095)	0.1030 (0.0149)
6	0.1590 (0.0150)	0.1399 (0.0341)	0.1276 (0.0112)	0.1245 (0.0129)
7	0.1674 (0.0185)	0.1484 (0.0401)	0.1452 (0.0146)	0.1338 (0.0173)
8	0.1747 (0.0216)	0.1571 (0.0443)	0.1637 (0.0122)	0.1441 (0.0202)
9	0.1821 (0.0249)	0.1673 (0.0507)	0.1822 (0.0149)	0.1553 (0.0264)
10	0.1879 (0.0257)	0.1757 (0.0523)	0.1987 (0.0112)	0.1633 (0.0269)
11	0.1937 (0.0269)	0.1857 (0.0552)	0.2128 (0.0165)	0.1740 (0.0308)
12	0.1983 (0.0253)	0.1936 (0.0565)	0.2237 (0.0157)	0.1820 (0.0297)
13	0.2074 (0.0219)	0.2054 (0.0573)	0.2378 (0.0193)	0.1921 (0.0316)
14	0.2177 (0.0181)	0.2175 (0.0557)	0.2528 (0.0335)	0.2078 (0.0301)
15	0.2271 (0.0138)	0.2286 (0.0528)	0.2646 (0.0447)	0.2205 (0.0274)
16	0.2393 (0.0117)	0.2417 (0.0464)	0.2788 (0.0508)	0.2343 (0.0211)
17	0.2537 (0.0154)	0.2562 (0.0463)	0.2931 (0.0531)	0.2496 (0.0172)
18	0.2689 (0.0196)	0.2709 (0.0515)	0.3064 (0.0523)	0.2646 (0.0101)
19	0.2834 (0.0204)	0.2860 (0.0596)	0.3185 (0.0468)	0.2795 (0.0154)
20	0.2978 (0.0219)	0.3008 (0.0678)	0.3305 (0.0392)	0.2932 (0.0213)

Table 2. The means and the standard deviations (in bracket) of the ten RMSEs by LM, BMS, FDM and WM-SM GPR.

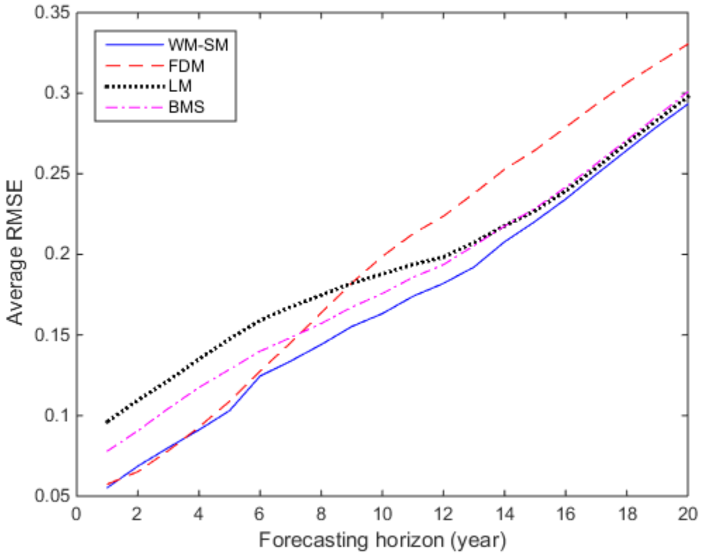

Figure 4. The average of the ten RMSEs by LM, BMS, FDM and WM-SM GPR.

It can be seen from the table and the figure that, the performances of WM-SM GPR and FDM are indistinguishable for short term forecasting (from 1- to 6-year forecasting horizons). But from 7-year horizon onwards, the GPR model substantially outperforms the FDM model. On the other hand, the performance of WM-SM GPR model is almost equal to those of LM and BMS for long term forecasting (for 16- to 20-year horizons). However, the former has much better accuracy for 1- to 15-year horizons than the latter two, particularly than LM model. In the mid-term (from 7- to 15-year horizons), the WM-SM GPR model is systematically better than the other three. Therefore, the WM-SM GPR model provides robust performance for both short term and long term forecasting with improved forecasting accuracy for mid-term, when compared with the other three models.

4. Conclusion and discussion

We have introduced Gaussian process regression as a new approach for modelling and forecasting mortality rates. We considered four commonly used Gaussian process regression models and also proposed to incorporate a weighted mean function with the spectral mixture covariance function for the problem of mortality modelling, which mitigated the difficulty in choosing suitable covariance functions in GPR modelling. The numerical examples showed that the proposed GPR model improved the overall forecasting accuracy of mortality rates, compared with the conventional GPR models. The performance of this model was also compared with Lee-Miller model, Booth-Maindonald-Smith model and the functional demographic model using the French total mortality rates. The numerical results demonstrated that the proposed GPR model provided a more robust forecast performance.

In contrast to Lee-Carter model and most of its variants, which directly act on the historical mortality curves for forecasting, the GPR models provide a different angle to handle this forecasting problem. We treat the mortality rates for each age group over time as a time series and assume that it follows a Gaussian process. After forecasts are made for some age groups for a future year, the mortality rates at the other ages can be obtained by interpolating the forecasted mortality rates to all age groups. The forecasting accuracy may depend on the choice of the age groups to be modelled. In our example, 20 age groups were selected, including 0, 1, 2, 5, 10, 12, 15, 18, 20, 22, 25, 28, 30, 40, 50, 60, 70, 80, 90, 100-year groups. The reason for choosing these age groups is that, the mortality rates tend to be very variable from age 0 to age 30 while they increase almost linearly from age 30 to age 100. Hence we need denser grids for interpolation in the interval from 0 to 30 and fewer points from the age 30 onwards. Our experiment showed that there was no significant difference in the results if more or slightly different age groups were used. Although the topic of this paper concerns human mortality modelling and forecasting, the idea can also be used in similar demographic problems such as fertility and migration modelling.

Another issue to be discussed is the weights chosen for the historical data in the mean function. When forecasts are made for long horizons, the correlations between the future points and the historical points become very low and the predictions by Gaussian process regression will converge to the mean function in long term. The weights for the historical data partially determine the mean function and therefore can also impact the accuracy of forecasts. In this paper we used the inverse of the time distances as the weights assigned to historical data for all the age groups. It is of course possible to use other weights, and if the weights involve tuning parameters, they can be determined by the cross validation. Furthermore, in this paper the mean function and the covariance function used for each age group are independent of the other age groups. However, as can be expected the mortality rates for different ages are closely correlated, especially between the neighbouring ages, it is therefore worth further investigating how to build mean functions and covariance functions taking the correlations into account.

Acknowledgements

This work was jointly funded by the Institute and Faculty of Actuaries (IFoA) and the College of Science and Engineering of the University of Leicester (UoL) through a PhD studentship. The authors thank the IFoA and the UoL for their financial support. They also thank the reviewers for their constructive suggestions and helpful comments which have substantially improved the manuscript.

References

Ahmed, N. K., Atiya, A. F., El Gayar, N. & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. Econometric Reviews 29(5): 594-621.

Alamaniotis, M., Ikonomopoulos, A. & Tsoukalas, L. H. (2011). A Pareto optimization approach of a Gaussian process ensemble for short-term load forecasting. In *the 16th International Conference on Intelligent System Application to Power Systems (ISAP),* pages 1-6. IEEE.

Alamaniotis, M. & Tsoukalas, L. H. (2016). Fusion of Gaussian process kernel regressors for fault prediction in intelligent energy systems. International Journal on Artificial Intelligence Tools 25(4): 1650023.

Andrawis, R. R., Atiya, A. F. & El-Shishiny, H. (2011). Forecast combinations of computational intelligence and linear models for the NN5 time series forecasting competition. International Journal of Forecasting 27(3): 672-688.

Bell, W. R. (1997). Comparing and assessing time series methods for forecasting age-specific fertility and mortality rates. Journal of Official Statistics 13: 279-303.

Ben Taieb, S., Bontempi, G., Atiya, A. F. & Sorjamaa, A. (2012). A review and comparison of strategies for multiple-step ahead time series forecasting based on the NN5 forecasting competition. Experts Systems with Applications 39(8): 1950-1957.

Banerjee, S., Gelfand, A. E., Finley, A. O. & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. Journal of the Royal Statistical Society Series B (Statistical Methodology) 70: 825-848.

Booth, H., Hyndman, R., Tickle, L., & De Jong, P. (2006). Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. Demographic Research 15: 289-310.

Booth, H., Maindonald, J. & Smith, L. (2002). Applying Lee-Carter under conditions of variable mortality decline. Population Studies 56: 325-336.

Brahim-Belhouari, S. & Bermak, A. (2004). Gaussian process for nonstationary time series prediction. Computational Statistics and Data Analysis 47 (4): 705–712.

Cairns, A., Blake, D., Dowd, K., Coughlan, G. & Khalaf-Allah, M. (2011). Bayesian stochastic mortality modelling for two populations. ASTIN Bulletin 41: 29-59.

Chapados, N. & Bengio, Y. (2007). Augmented functional time series representation and forecasting with Gaussian processes. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems* 19: 457-464. The MIT Press, Cambridge, MA.

Chiou, J. & Müller, H.G. (2009). Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. Journal of American Statistical Association 104: 572-585.

Claveria, O., Monte, E. & Torra, S. (2016). Modelling cross-dependencies between Spain's regional tourism markets with an extension of the Gaussian process regression model. SERIEs 7(3): 341-357.

Cressie, N (1993). *Statistics for Spatial Data*. Wiley.

de Jong, P., Tickle, L. & Xu, J. (2016). Coherent modeling of male and female mortality using Lee–Carter in a complex number framework. Insurance: Mathematics and Economics 71: 130-137.

Girard, A., Rasmussen, C., Quiñonero-Candela, J. & Murray-Smith, R. (2003). Multiple-step ahead prediction for nonlinear dynamic systems - a Gaussian process treatment with propagation of the uncertainty. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems* 15. The MIT Press, Cambridge, MA.

Human Mortality Database. (2010). Univeristy of California, Berkeley (USA), and Max Plank Institute for Demographic Research (Germany). http://www.mortality.org/.

Hyndman, R. J., Booth H. & Yasmeen, F. (2013). Coherent mortality forecasting: The product-ratio method with functional time series models. Demography 50: 261-283.

Hyndman, R. J. & Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. Computational Statistics & Data Analysis 51: 4942–4956.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. Journal of the Southern African Institute of Mining and Metallurgy 52(6): 119-139.

Lee, R. & Carter, L. (1992). Modeling and forecasting the time series of U.S. mortality. Journal of the American Statistical Association 87: 659-671.

Lee, R. & Miller, T. (2001). Evaluating the performance of the Lee-Carter method for forecasting mortality. Demography 38: 537-549.

Li, N. & Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. Demography 42: 575-594.

Liu, X. & Yu, H. (2011). Assessing and extending the Lee-Carter model for long-term mortality prediction. Living to 100 Symposium.

MacKay, D. J. C. (1992). Bayesian Methods for Adaptive Models. PhD Thesis, California Institute of Technology.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Mori, H. & Ohmi, M. (2005). Probabilistic short-term load forecasting with Gaussian processes. In *Proceedings of the 13th International Conference on Intelligent Systems Application to Power Systems*. IEEE.

Neal, R. M. (1992). Bayesian training of backpropagation networks by the hybrid Monte Carlo method. Tech. Rep. CRG-TR-92-1, Dept. of Computer Science, Univ. of Toronto.

Neal, R. (1996). Bayesian Learning for Neural Networks. Lecture Notes in Statistics, No 118. Springer-Verlag.

Neal, R. (1997). Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Tech. Rep. CRG-TR-97-2, Dept. of Computer Science, Univ. of Toronto.

Oeppen, J. (2008). Coherent forecasting of multiple-decrement life tables: A test using Japanese cause of death data. Technical report. Rostock, Germany: Max Planck Institute for Demographic Research.

O’Hagan, A. (1978). Curve fitting and optimal design for prediction. J. Roy. Statist. Soc. Ser. B 40 (1): 1–42.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*. Springer.

Rasmussen, C. & Williams, C. (2006). *Gaussian Process for Machine Learning*. MIT Press.

Renshaw, A. & Haberman, S. (2003). Lee-Carter mortality forecasting: A parallel generalized linear modelling approach for England and Wales mortality projections. Applied Statistics 52(1): 119-137.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N. & Aigrain, S. (2013). Gaussian processes for time-series modelling. Phil. Trans. R. Soc. A, 371(1984): 20110550.

Quiñonero-Candela, J. & Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. Journal of Machine Learning Research 6: 1939-1959.

Seeger, M. (2003). Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations. PhD thesis, University of Edinburgh.

Seeger, M. (2004). Gaussian processes for machine learning. International Journal of Neural Systems 14(2): 1–38.

Shi, J. & Choi, T. (2010). *Gaussian Process Regression Analysis for Functional Data*. CRC

Press.

Snelson, E., Rasmussen, C. E. & Ghahramani, Z. (2004). Warped Gaussian processes. In *Advances in Neural Information Processing Systems* 16: 337–344.

Wilson, A. G. & Adams, R. P. (2013). Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, Georgia, USA.

Wu, Q., Law, R. & Xu, X. (2012). A sparse Gaussian process regression model for tourism demand forecasting in Hong Kong. Expert Systems with Applications 39(5): 4769-4774.