人口统计学研究

基于机器学习的 Lee-Carter 模型死亡率预测方法研究

陶祥兴，杨 峥，季彦颋
（浙江科技学院 理学院，浙江 杭州 310023）

摘 要：世界各国人口死亡率不断降低，预期寿命变得难以预测。改进死亡率预测方法，准确预测未来人口的数量变化有着重要意义。传统的 Lee-Carter 模型通过年龄组平均死亡率、时间项以及年龄因子随时间变化的敏感度这三个参数来刻画死亡率的变化，模型中的时间项采用 ARIMA 方法进行预测。但该方法并不能解决死亡率数据具有长记忆性的问题，并且现有研究很少将传统人口学方法与大数据背景下机器学习方法相结合。因此本文引入 LSTM（长短期记忆深度学习神经网络）和分数布朗运动驱动的 O-U 过程来对死亡率预测进行改进。由于中国大陆有关死亡率的数据样本量少且不完整，选用中国香港男性分年龄组死亡率数据，分别采用时间序列 ARIMA 方法、时间序列与机器学习相结合的 ARIMA-LSTM 方法以及分数 O-U 过程来拟合和预测模型中的时间项，通过残差图和三种评价指标值来比较三种方法的短期预测效果。结果表明，ARIMA-LSTM 方法的短期预测效果最好，证明了引入机器学习方法对死亡率预测方法改进的可行性，为政府预测未来死亡率提供新思路，也为相关机构研究长寿风险提供依据。

关键词：Lee-Carter 模型；ARIMA 方法；ARIMA-LSTM 方法；分数 O-U 过程；死亡率预测

中图分类号：C921 **文献标识码**：A **文章编号**：1000-4149（2022）06-0047-11

DOI：10.3969/j.issn.1000-4149.2022.00.032

收稿日期：2022-03-13；**修订日期**：2022-08-03

基金项目：国家自然科学基金面上项目“若干非标准核奇异积分及相关分数次非线性方程的研究”（11771399）；2020 年度浙江省哲学社会科学重点研究基地课题“基于随机时滞模型的长寿风险量化理论与催化从策略研究”（20JDZD071）。

作者简介：陶祥兴，数学博士，浙江科技学院理学院教授，博士生导师；杨峥，浙江科技学院理学院硕士研究生；季彦颋，哲学博士，浙江科技学院理学院副教授。

一、引言

随着社会经济的迅速发展和医疗水平的大幅进步，世界人口死亡率逐渐降低，老龄化问题日益严重。人口寿命的非预期延长导致养老事业发展滞后于人口老龄化的进程，给养老金机构和人寿保险公司带来巨大压力，进而可能影响整个社会经济的发展。因此，提高死亡率预测精度对于政府制定未来人口政策、经济政策以及养老政策至关重要。

纵观学者对死亡率模型的研究，影响力最大的当属 Lee-Carter 模型。该模型是由李（Lee）和卡特（Carter）提出的一个对数双线性模型，首次考虑时间因素和年龄因素对对数中心死亡率的影响，并通过时间序列 ARIMA 模型来外推预测死亡率[1]。Lee-Carter 模型由于其形式简单、计算方便、参数可解释性强的特点被各国学者广泛应用。随着研究的不断深入，Lee-Carter 模型也暴露出一定缺陷，如模型对参数的假设条件过高等。为了提高预测的精确性，国内外学者从不同方向对经典 Lee-Carter 模型进行了改进研究，其改进方面主要有：一是放宽了模型的假设条件。在经典 Lee-Carter 模型中关于死亡率误差的假设是独立同分布的，针对这一假设条件过强的情况，勃朗恩斯（Brouhns）等假设死亡人数服从泊松分布，提出泊松对数双线性模型，并对死亡率进行了预测[2]。二是对经典 Lee-Carter 模型参数估计方法的改进。经典 Lee-Carter 模型采用奇异值分解（SVD）法[3-4] 进行参数估计，吴晓坤等学者采用加权最小二乘（WLS）法[5-7]，极大似然（ML）法[8-9] 和贝叶斯马尔科夫蒙特卡洛（MCMC）法[10] 进行参数估计，发现这三种方法在提高参数估计的拟合优度和死亡率预测的精度上都有很好的表现。三是对有限死亡率数据下的预测方法的改进。Lee-Carter 模型对死亡率数据的连续性有着较高的要求，大大限制了该模型在有限数据国家的应用。考虑到中国死亡率数据量较小，王晓军和黄顺林通过改进时序中的波动性来提高模型预测的精度[11]。四是对刻画 Lee-Carter 模型中的时间项参数 k_t 的改进。传统的 Lee-Carter 模型运用 ARIMA 方法拟合并预测 k_t 的值，但该方法并不能捕捉到死亡率数据中的长记忆性特征，存在一定的缺陷。有学者采用一个带漂移项的随机游走模型来刻画 Lee-Carter 模型中的时间项 k_t，并证明该方法同样适用于稀疏数据[12]。而对于死亡率中存在的跳跃性变化，田梦和邓颖璐采用双指数跳跃扩散模型来描述这一特征，并得到了较好的预测效果[13]。

近年来，我国对 Lee-Carter 模型中的时间项 k_t 的改进研究较少，大多都是单独利用离散方法或是连续方法对时间项 k_t 进行拟合和预测，很少在此基础上引入机器学习方法对时间项 k_t 的拟合和预测进行改进。因此本文在 Lee-Carter 模型预测死亡率的基础上提出对时间项 k_t 进行拟合和预测的两种创新方法：第一，引入机器学习中的长短期记忆网络模型（Long Short-Term Memory，LSTM），在 Lee-Carter 模型中时间项 k_t 用离散 ARIMA 模型拟合的基础上，运用 LSTM 模型修正其残差并对死亡率进行短期预测；第二，引入分数布朗运动驱动的 O-U 过程来刻画 Lee-Carter 模型中的时间项 k_t，通过机器学习中的遗传算法来估计分数布朗运动驱动的 O-U 过程中的未知参数并进行死亡率的短期预测。本文通过残差图和三种回归指标将两种创新方法和 ARIMA 方法的死亡率预测效果进行对比，确定出一个短期预测精度最高的死亡率预测方法，为政府预测未来死亡率提供了一种新的思路，也为相关机构研究长寿风险提供了一定的依据。

二、基础模型

1. 模型简述

Lee-Carter 模型考虑了年龄因素和时间因素对模型的影响，具体模型表达式如下：

$$\ln(m_{x,t}) = \alpha_x + k_t\beta_x + \varepsilon_{x,t} \tag{1}$$

其中，$m_{x,t}$ 表示 t 时刻 x 岁人群的中心死亡率；α_x 为影响死亡率的年龄参数，表示 x 岁人口对数死亡率的平均水平，即 $\hat{\alpha}_x = \frac{1}{T}\sum_t \ln(m_{x,t})$（$T$ 为待估计死亡率数据中包含的日历年总数）；β_x 为年龄因子对时间的敏感度；k_t 通常被称为死亡指数，表示时间 t 对死亡率的影响程度，通常被认为是一个 ARIMA 过程或随机游走过程；$\varepsilon_{x,t}$ 表示模型的随机误差项，且 $\varepsilon_{x,t} \sim N(0, \sigma^2)$。

由模型可知 $\{\alpha_x, \beta_x, k_t\}$ 是模型的一组解，取任意一个不为 0 的常数 c，将参数变换成 $\{\alpha_x, \beta_x/c, ck_t\}$ 或 $\{\alpha_x - c\beta_x, \beta_x, k_t + c\}$，都可使得原模型保持不变。因此本文对参数增加以下的约束条件使得模型满足唯一的参数估计结果：

$$\sum_x \beta_x = 1, \qquad \sum_t k_t = 0 \tag{2}$$

2. 参数估计

由于现实中死亡率误差独立同分布的假设通常不成立，故本文采用勃朗恩斯等去除了 $\varepsilon_{x,t}$ 同方差的假定[2]，提出用极大似然法来进行参数估计。该方法假设死亡人数 $d_{x,t}$ 服从参数为 $\lambda_{x,t}$ 的泊松分布，即 $d_{x,t} \sim Poisson(m_{x,t}E_{x,t})$，其中 $\lambda_{x,t} = m_{x,t}E_{x,t}$，$m_{x,t} = exp(\alpha_x + k_t\beta_x)$，$E_{x,t}$ 表示 t 时刻年龄为 x 的暴露人数。传统 Lee-Carter 模型的极大似然函数可表达为：

$$L(\alpha_x, \beta_x, k_t) = \sum_{x,t}[d_{x,t}(\alpha_x + k_t\beta_x) - E_{x,t}e^{\alpha_x + k_t\beta_x}] + C \tag{3}$$

本文根据约束条件，借助李志生等运用的初始值选定方案来实现模型中三类参数的估计[14]，从参数初始值 $\hat{\alpha}_x^{(0)} = 0$，$\hat{\beta}_x^{(0)} = 1$ 和 $\hat{k}_t^{(0)} = 0$ 开始，按照以下步骤迭代参数：

$$\hat{\alpha}_x^{(n+1)} = \hat{\alpha}_x^{(n)} + \frac{\sum_t d_{x,t} - \hat{d}_{x,t}^{(n)}}{\sum_t \hat{d}_{x,t}^{(n)}}, \quad \hat{\beta}_x^{(n+1)} = \hat{\beta}_x^{(n)}, \quad \hat{k}_t^{(n+1)} = \hat{k}_t^{(n)} \tag{4}$$

$$\hat{k}_t^{(n+2)} = \hat{k}_t^{(n+1)} + \frac{\sum_t (d_{x,t} - \hat{d}_{x,t}^{(n+1)})\hat{\beta}_x^{(n+1)}}{\sum_x \hat{d}_{x,t}^{(n+1)}(\hat{\beta}_x^{(n+1)})^2}, \quad \hat{\alpha}_x^{(n+2)} = \hat{\alpha}_x^{(n+1)}, \quad \hat{\beta}_x^{(n+2)} = \hat{\beta}_x^{(n+1)} \tag{5}$$

$$\hat{\beta}_x^{(n+3)} = \hat{\beta}_x^{(n+2)} + \frac{\sum_t (d_{x,t} - \hat{d}_{x,t}^{(n+2)})\hat{k}_x^{(n+2)}}{\sum_x \hat{d}_{x,t}^{(n+2)}(\hat{k}_t^{(n+2)})^2}, \quad \hat{\alpha}_x^{(n+3)} = \hat{\alpha}_x^{(n+2)}, \quad \hat{k}_t^{(n+3)} = \hat{k}_t^{(n+2)} \tag{6}$$

其中，$\hat{d}_{x,t}^{(n)}$ 表示的是第 n 步迭代得到的死亡人数的估计值，即 $\hat{d}_{x,t}^{(n)} = E_{x,t}e^{\hat{\alpha}_x^{(n)} + \hat{\beta}_x^{(n)}\hat{k}_t^{(n)}}$。

三、ARIMA 方法预测死亡率

1. 数据来源与参数估计结果

本文采用 1971—2020 年香港男性分年龄组死亡率、死亡人口以及暴露人口的数据，参

照简易生命表和已有文献的年龄分组，将香港男性数据分为 19 个年龄组（即 0—4 岁、5—9 岁、……、85—89 岁以及 90 岁及以上），其中每个年龄组的死亡率数据采用 5 年死亡率数据的算数平均。数据来源于人类死亡率数据库（Huamn Mortality Database）和香港特别行政区政府统计处。根据这些数据绘制了分年龄组死亡率三维图，如图 1 所示。

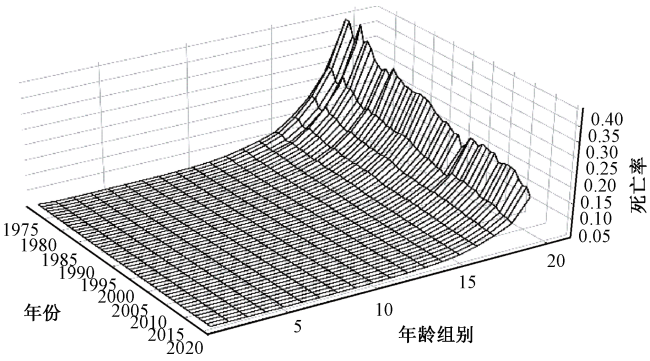

图 1　1971—2020 年中国香港男性死亡率三维图

从图 1 中能够看出每一个年龄组的死亡率都呈现出普遍下降的趋势，这符合死亡率正在逐渐降低的社会环境，但降低的程度在每个年龄组都有着差异。接下来将对死亡率模型的参数进行估计、拟合和预测。

选取 1971—2018 年香港男性完整死亡率数据，采用极大似然法运用 R 软件进行参数估计，得到 Lee-Carter 模型的参数估计值如图 2 所示。

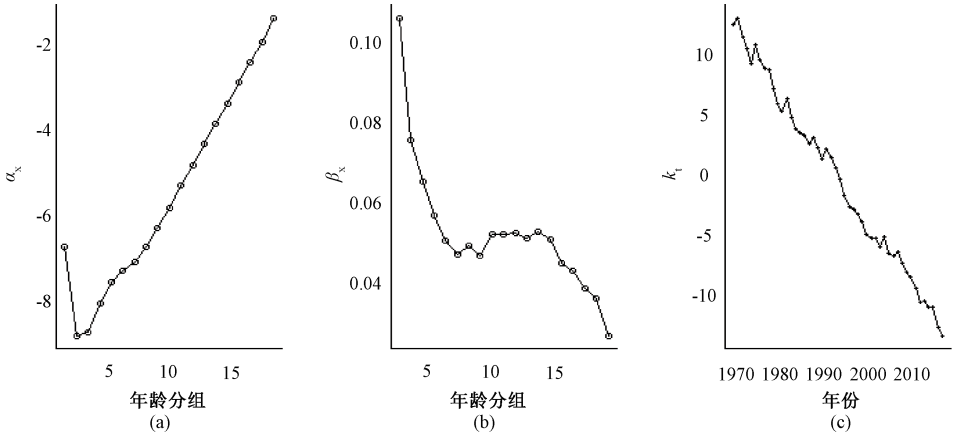

图 2　Lee-Carter 模型参数的估计值

从图 2 可以看出，参数 α_x 表示年龄因素对死亡率的影响，这一影响随着年龄的增长呈现出了先下降后上升的趋势；β_x 随着年龄的增加而逐渐下降，且低年龄组和高年龄组 β_x 值下降幅度相对较大，表明这两个年龄段对死亡率时间因子变化的敏感度相对较高；而参数 k_t 的估计值随着时间的推移表现出显著的递减趋势。

2. 死亡率预测

本文利用 1971—2018 年的香港男性死亡率数据，对模型进行拟合，并根据上述参数估计结果，采用 ARIMA（0，1，1）模型预测 2019、2020 年的 k_t 值（见图 3）。

根据 k_t 的预测值和式（1），可以得到 2019—2020 年香港男性分年龄组人口的死亡率预测值，结果如表 1 所示。

四、ARIMA-LSTM 方法预测死亡率

1. LSTM 模型介绍

长短期记忆深度学习神经网络（Long Short-Term Memory，LSTM）是一种改良过的循环神经网络，因其能够有效解决长时间依赖问题，被广泛应用于时间序列的预测。LSTM 神经网络的数据流向和传递过程如图 4 所示。

图中 X_t 表示 t 时刻输入的训练样本数据，h_t 表示神经网络当前单元的输出值，h_{t-1} 表示 $t-1$ 时刻神经网络单元的输出值，f_t 表示遗忘门、i_t 表示输入门，O_t 表示输出门，$\tilde{C}_t$ 表示前一时刻的单元状态，C_t 表示当前时刻的单元状态，σ 和 $tanh$ 表示神经网络内的激活函数。LSTM 神经网络通过遗忘门、输入门和输出门三种门控机制来完成数据的丢弃、增加、保存和传递，从而实现有用信息的长期记忆。在神经单元内的数据传递满足以下公式：

$$
\begin{aligned}
f_t &= \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \\
i_t &= \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \\
\tilde{C}_t &= tanh\left(W_{\tilde{c}}\left[h_{t-1}, x_t\right] + b_{\tilde{c}}\right) \\
C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\
O_t &= \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \\
h_t &= O_t * tanh\left(C_t\right)
\end{aligned}
\tag{7}
$$

其中，W 表示权重，b 表示模型的偏置。

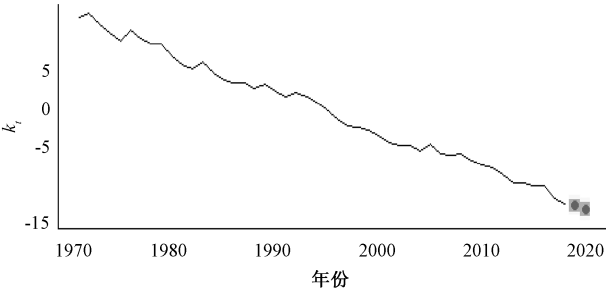

图 3　时间项 k_t 的预测值

表 1　基于 ARIMA 方法的 2019—2020 年死亡率预测值

年龄组（岁）	2019 年	2020 年
0—4	0. 000307	0. 000289
5—9	0. 000059	0. 000057
10—14	0. 000074	0. 000072
15—19	0. 000161	0. 000156
20—24	0. 000290	0. 000282
25—29	0. 000389	0. 000379
30—34	0. 000473	0. 000461
35—39	0. 000680	0. 000662
40—44	0. 000992	0. 000964
45—49	0. 001601	0. 001556
50—54	0. 002621	0. 002546
55—59	0. 004336	0. 004216
60—64	0. 006903	0. 006705
65—69	0. 011324	0. 011010
70—74	0. 019587	0. 019110
75—79	0. 032960	0. 032190
80—84	0. 056429	0. 055240
85—89	0. 092267	0. 090451
90 及以上	0. 183658	0. 180977

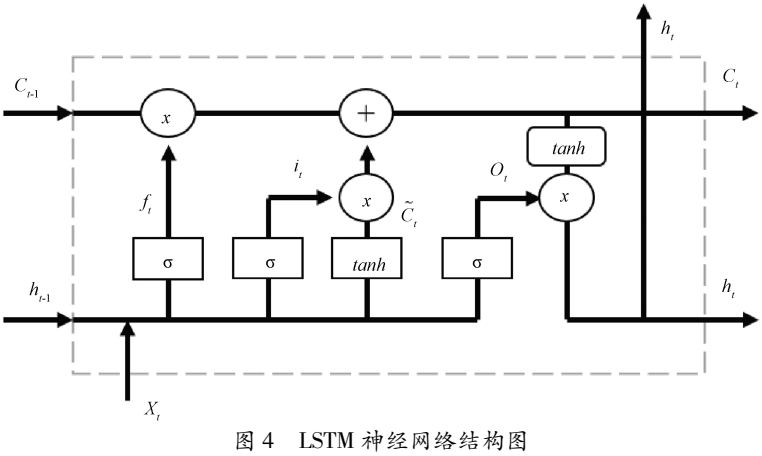

图 4　LSTM 神经网络结构图

2. 建模步骤

本文引用具有长记忆性并且适用于时间序列建模的 LSTM 神经网络对 ARIMA 模型进行修正，具体步骤如下。

（1）对获得到的 k_t 序列进行平稳性检验。首先对 k_t 序列作图来判断序列是否平稳，即是否需要经过差分处理。对差分后的序列进行单位根（ADF）检验，若通过 ADF 检验表明差分后的序列已经平稳，则不需要进行差分处理，否则需要在此基础上再进行差分处理。

（2）对差分后平稳的序列做自相关（ACF）图和偏自相关（PACF）图，初步确定模型 p、q 的值并根据 AIC 最小准则确定最优模型。根据最优模型构造残差序列 $\{e_{it}\}$，有 $e_{it} = K_{it} - \hat{k}_{it}$，$K_{it}$ 为死亡率数据运用极大似然法估计出的参数 k_t 的值，$\hat{k}_{it}$ 为时间 t 下 ARIMA 模型的拟合值。

（3）对残差序列 $\{e_{it}\}$ 进行白噪声（Ljung-Box）检验，若通过白噪声检验说明拟合模型充分提取了原序列中的相关信息。

（4）将残差序列 $\{e_{it}\}$ 作为输入值，利用 LSTM 神经网络对其进行训练，其输出结果为时间 t 下模型残差的预测值 $\hat{l}_{it}$。根据 ARIMA 模型对 k_t 序列的预测值和 LSTM 模型对其残差的预测值得到组合模型的 k_t 预测值，即：$\hat{K}_{it} = \hat{k}_{it} + \hat{l}_{it}$。本文提出的组合模型流程见图 5。

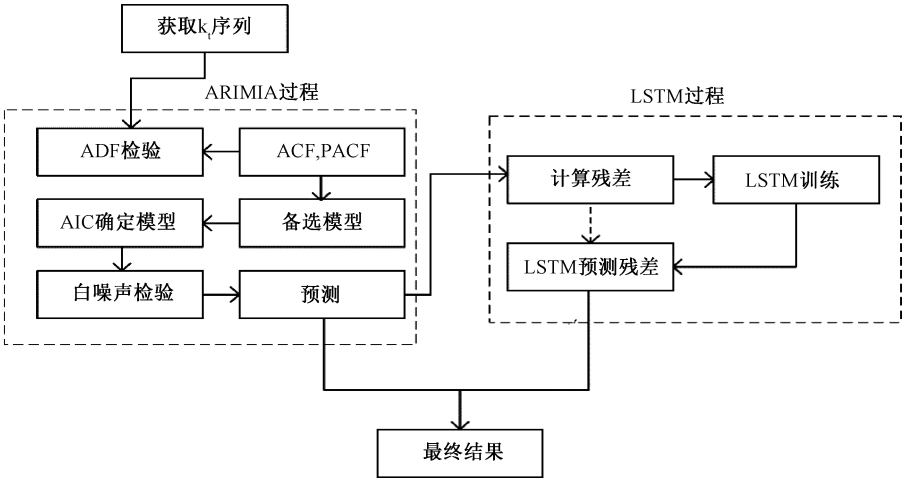

图 5　流程建模图

3. 死亡率预测

根据上述建模步骤可以得到未来两年 k_t 的预测值，进一步运用 Lee-Carter 死亡率预测模型可得到未来两年死亡率的预测值，即 $m_{x,t} = exp(\alpha_x + \hat{K}_{it}\beta_x)$，其中 $\hat{K}_{it}$ 为组合模型 k_t 的预测值。2019—2020 香港男性分年龄组人口死亡率的预测值如表 2 所示。

五、分数 O-U 过程对 Lee-Carter 模型时间项的拟合与预测

1. 分数 O-U 过程

分数 O-U 过程是由分数布朗运动驱动的 O-U 过程，对于处理实际问题中的长记忆性有着十分普遍的应用。我们考虑以下形式的分数 O-U 过程来刻画 Lee-Carter 模型中的时间项 k_t：

$$dk_t = -\lambda k_t dt + \sigma dB_t^H, \quad t \geqslant 0 \tag{8}$$

将式（8）离散化得到如下表达式：

$$k_{t+\Delta t} = (1 - \lambda \Delta t) k_t + \sigma (B_{t+\Delta t}^H - B_t^H) \tag{9}$$

其中，$\{B_t^H,\ t \geqslant 0\}$ 是 Hurst 参数 $H > 1/2$ 的分数布朗运动，具有长记忆性的特点，所以本文引入其来刻画时间因子 k_t；H、λ 和 σ 是需要估计的未知参数。

2. 参数估计

根据 Hurst 参数的研究成果[15]，本文得到分数布朗运动驱动的分数 O-U 过程中 Hurst 参数 H 的估计：

$$\hat{H} = \frac{1}{2} - \frac{1}{2\ln 2} ln \left[\frac{\sum_{t=1}^{N-1} (k_{t+1} - k_t)^2}{\sum_{t=1}^{\frac{N}{2}-1} (k_{2(t+1)} - k_{2t})^2} \right] \tag{10}$$

利用二次变差法计算出 σ^2 估计量的值：

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{N-1} [k_{2(t+1)} - k_t]^2}{N - 1} \tag{11}$$

利用机器学习方法中的遗传算法来进行最后一个未知参数 λ 的估计。遗传算法（Genetic Algorithm，GA）是一种通过模拟自然进化过程搜索最优解的方法，它克服了传统极大似然等方法容易陷入局部极值丢失最优解的缺点，其具体步骤如下。

（1）编码。随机产生一个种群作为该问题的初始解，并运用合适的编码方案对种群中的每一个个体进行编码，如二进制编码或实值编码等。本文采用的编码方式是二进制编码，其编码过程简单易行，相应的交叉算子、变异算子等操作运用位运算即可实现。

（2）适应度函数设计。本文设计第 i 个个体的回应度函数为：$F_i = \sum_{i=1}^{n} \omega_i [C_i^H(\Omega) - C_i^M]^2$，选取 $\omega_i = 1/n$，n 为样本总量，F_i 为种群中第 i 个个体所表示的模型拟合效果的优劣程度，F_i 越小，拟合效果越好。

（3）选择。根据每一个个体的适应度值进行选择，选择的原则为适应度越高的个体可能被选中的概率就越大。本文按照轮盘赌的选择方法，每一个个体按照 $P_i = F_i / \sum_{j=1}^{N} F_j$ 的概率将一个圆盘划分为 N 个扇形区域，N 为种群规模。通过转动圆盘，圆盘指针随机落在哪

表 2　基于 ARIMA-LSTM 方法的 2019—2020 年死亡率预测值

年龄组（岁）	2019 年	2020 年
0—4	0.000295	0.000279
5—9	0.000058	0.000055
10—14	0.000073	0.000070
15—19	0.000158	0.000153
20—24	0.000285	0.000278
25—29	0.000382	0.000373
30—34	0.000465	0.000453
35—39	0.000668	0.000652
40—44	0.000973	0.000948
45—49	0.001570	0.001529
50—54	0.002569	0.002502
55—59	0.004253	0.004145
60—64	0.006767	0.006589
65—69	0.011108	0.010826
70—74	0.019260	0.018829
75—79	0.032431	0.031735
80—84	0.055614	0.054538
85—89	0.091021	0.089377
90 及以上	0.181821	0.179385

个区域，哪个区域的个体被选中。

（4）交叉与变异。交叉是将随机配对的两个个体相互交换本体中的部分基因，其主要的方法有单点交叉、多点交叉等，本文采用多点交叉法，交叉点根据交叉概率随机选取。变异是指个体上的某些基因发生改变，主要的变异方法有基本位变异、均匀变异等。本文采用基本位变异法，根据变异的概率随机确定每个个体想要发生变异的基因点位，然后将二进制编码中的“1”变为“0”，“0”变为“1”。

（5）参数选择（交叉概率、变异概率、种群规模、迭代次数）。遗传算法中，参数的选择是否合理将直接影响模型的精度和有效性。本文通过多次试验后，选取交叉概率为 0.6，变异概率为 0.1，种群规模为 50，遗传算法迭代次数为 500。

利用 PYTHON 实现上述遗传算法代码编程，可得分数布朗运动驱动的 O-U 过程的参数估计值如表 3 所示。

表 3　分数布朗运动驱动的分数 O-U 过程的参数估计值

$\hat{H}$	$\hat{\lambda}$	$\hat{\sigma}$
0.592	0.010	0.907

3. 死亡率预测

将分数布朗运动驱动的 O-U 过程的参数估计值代入其离散表达式（9），可以对 2019—2020 年的香港分年龄组的死亡率进行预测，预测结果如表 4 所示。

表 4　基于分数布朗运动驱动的 O-U 过程的 2019—2020 年死亡率预测值

年龄组（岁）	2019 年	2020 年
0—4	0.000321	0.000322
5—9	0.000061	0.000061
10—14	0.000076	0.000077
15—19	0.000165	0.000165
20—24	0.000296	0.000297
25—29	0.000397	0.000398
30—34	0.000483	0.000484
35—39	0.000693	0.000694
40—44	0.001014	0.001016
45—49	0.001636	0.001640
50—54	0.002679	0.002684
55—59	0.004429	0.004438
60—64	0.007056	0.007071
65—69	0.011567	0.011590
70—74	0.019955	0.019990
75—79	0.033555	0.033611
80—84	0.057345	0.057431
85—89	0.093665	0.093796
90 及以上	0.185712	0.185904

六、短期预测结果比较

根据上述预测结果，结合 2019—2020 年真实死亡率数据的对比图（见图 6）发现，2019 年和 2020 年，80 岁以下三种模型预测的死亡率均表现出较好的预测效果，而 80 岁以上死亡率的预测值与真实值之间存在较小偏差。

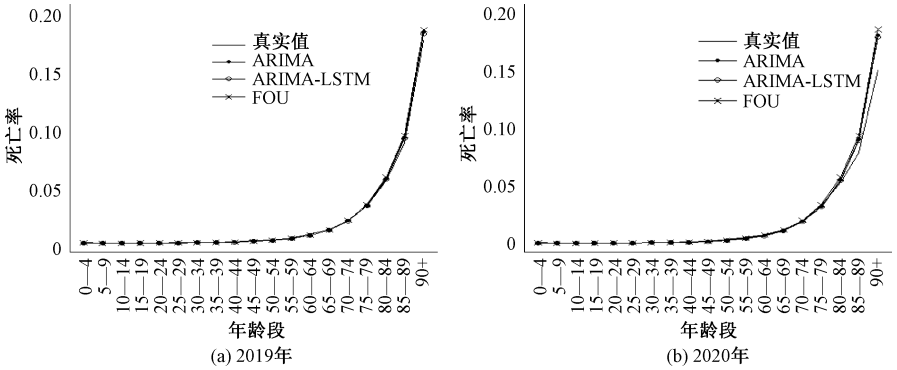

图 6　2019—2020 年死亡率数值对比图

为进一步评估三种方法的预测效果，本文选用残差图以及三个评价指标对预测值进行综合比较，并选择出相对最优的预测方法。

1. 残差图检验

由于三种预测方法得到的死亡率预测值与真实值之间均存在一定的差异，故本文通过考察分年龄组的死亡率残差图（见图 7）来初步评估三种方法的预测效果。图 7 显示，60 岁以下三种方法的死亡率残差值均位于 0 线附近，且并未表现出明显差异，表明该年龄段内，三种方法的预测效果较为稳定且精度较高；高年龄组的预测残差值差异较为明显，整体上 ARIMA-LSTM 方法的短期预测表现相对最好，传统的 ARIMA 方法次之。

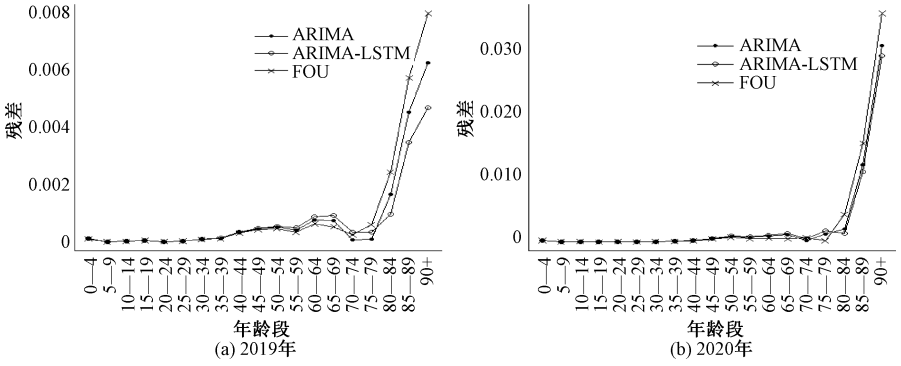

图 7　2019—2020 年死亡率残差对比图

2. 评价指标检验

为了进一步比较三种方法的预测效果，本文引入平均绝对百分比误差（MAPE）、均方误差（MSE）和平均绝对误差（MAE）三个评价指标来评价模型的预测效果。

$$MAPE(\hat{y}) = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{12}$$

$$MSE(\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{13}$$

$$MAE(\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{14}$$

其中，$\hat{y}_i$ 为预测值，y_i 为真实值，n 为预测期。三个评价指标数值越小，说明模型预测效果越好，故综合残差图和评估指标值（见表 5），本文认为 ARIMA-LSTM 方法在死亡率的短期预测中具有相对较好的预测能力。

表 5　三种方法预测精度对比

模型	MAPE	MSE	MAE
ARIMA	0. 153563	0. 000031	0. 001842
ARIMA-LSTM	0. 147006	0. 000027	0. 001715
FOU	0. 159528	0. 000043	0. 002180

七、结论

本文采用中国香港 1971—2020 年男性分年龄组数据，在 Lee-Carter 模型的基础上，提出了两种改进方法对模型中时间项进行拟合和预测。第一种是引入机器学习 LSTM 模

型，在 Lee-Carter 模型中时间项 k_t 用离散 ARIMA 模型拟合的基础上，运用 LSTM 来修正其残差并进行死亡率短期预测。第二种是引入分数布朗运动驱动的 O-U 过程来刻画 Lee-Carter 模型中时间项 k_t，通过遗传算法来估计分数 O-U 过程中的未知参数并进行死亡率短期预测。

将本文提出的两种改进方法与传统 ARIMA 方法的预测效果作对比，结合预测残差图发现，60 岁以下三种方法的死亡率残差值基本位于 0 线附近，而高年龄组的残差值表现出较大差异，整体上 ARIMA-LSTM 方法的残差值明显低于其他两种方法，表明 ARIMA-LSTM 方法的短期预测效果相对较好。

最后引入 MAPE、MSE 和 MAE 指标进一步考察三种方法的预测效果，结果表明 ARIMA-LSTM 方法具有较好的短期预测能力。综合残差图和评估指标，本文认为 ARIMA-LSTM 方法能够更加精确地描述死亡率的短期变化，表明将机器学习方法应用于死亡率的短期预测中是切实有效的，有助于改善人口老龄化给政府和相关长寿保险机构带来的负面影响。

在进一步的研究中，将尝试运用 ARIMA-LSTM 方法对大陆的死亡率数据进行拟合预测。同时考虑到大陆死亡率数据有限且有缺失，将采用贝叶斯方法进行参数估计，该方法能够有效减少数据质量不高带来的不利影响。

参考文献：

[1] LEE R D, CARTER L R. Modeling and forecasting US mortality [J]. Journal of American Statistical Association, 1992, 87 (419): 659-671.

[2] BROUHNS N, DENUIT M, VERMUNT J. A Poisson log-linear regression approach to the construction of projected lifetables [J]. Insurance: Mathematics and Economics, 2002, 31 (3): 373-393.

[3] 黄匡时. Lee-Carter 模型在模型生命表拓展中的应用——以中国区域模型生命表为例 [J]. 人口研究, 2015 (5): 37-48.

[4] 胡仕强. 基于 Lee-Carter 模型和王变换方法的长寿债券定价研究 [J]. 商业研究, 2015 (10): 82-88.

[5] 吴晓坤, 刘江涛, 苏雨桐, 高建伟. 基于 Lee-Carter 模型的高龄人口死亡率预测 [C] //2020—2021 中国保险与风险管理国际年会论文集, 2021: 1134-1147.

[6] 赵明. 中国男性人口死亡率动态预测的方法比较——基于 Lee-Carter 模型与贝叶斯分层模型的研究 [J]. 人口与发展, 2022 (1): 40-49.

[7] 高甲. 基于 Lee-Carter 模型的香港人口死亡率预测 [D]. 济南: 山东大学, 2020: 1-51.

[8] 吴晓坤, 李姚洁. Lee-Carter 模型外推预测死亡率及偏差纠正 [J]. 统计与决策, 2016 (20): 19-21.

[9] 曹园. 基于 Lee-Cater 模型的我国死亡率预测 [J]. 统计与决策, 2018 (9): 32-36.

[10] 胡仕强. 基于贝叶斯 MCMC 方法的我国人口死亡率预测 [J]. 保险研究, 2015 (10): 70-83.

[11] 王晓军, 黄顺林. 中国人口死亡率随机预测模型的比较与选择 [J]. 人口与经济, 2011 (1): 82-86.

[12] LI N, LEE R, TULJAPURKAR S. Using the Lee-Carter method to forecast mortality for populations with limited data [J]. International Statistical Review, 2004, 72 (1): 19-36.

[13] 田梦, 邓颖璐. 我国随机死亡率的长寿风险建模和衍生品定价 [J]. 保险研究, 2013 (1): 14-26.

[14] 李志生, 刘恒甲. Lee-Carter 的估计与应用——基于中国人口数据的分析 [J]. 中国人口科学, 2010 (3): 47-56.

[15] ZHANG P, XIAO W L, ZHANG X L. Parameter identification for fractional Ornstein-Uhlenbeck processes based on discrete observation [J]. Economic Modelling, 2014, 36 (1): 198-203.

Research on the Prediction Method of Death Rate of Lee-Carter Model Based on Machine Learning

TAO Xiangxing, YANG Zheng, JI Yanting
(School of Science, Zhejiang University of Science and Technology, Hangzhou 310023, China)

Abstract: The death rate of the world's population has been continuously reduced, and the life expectancy has become unpredictable. Therefore, it is of great significance to improve the mortality prediction method to accurately predict the future population changes. The traditional Lee-Carter model describes the change of mortality through three parameters: the average mortality rate of the age group, the time term and the sensitivity of the age factor to change with time. The time term in the model is predicted by ARIMA method. However, the method can not solve the problem of long memory of mortality data, and the existing research rarely combines traditional demographic methods with machine learning methods in the context of today's big data. Therefore, this paper introduces LSTM (long-term and short-term memory deep learning neural network) and fractional Brownian motion driven O-U process to improve the prediction of mortality. Due to the small sample size and incomplete data on mortality mainland China, this paper selects the mortality data of male age groups in Hong Kong, China, and used time series ARIMA method, the ARIMA-LSTM method combined with time series and machine learning, and the fractional O-U process to fit and predict the time items in the model, and compares the short-term prediction effects of the three methods through residual diagram and three evaluation index values. The results show that ARIMA-LSTM method has the best short-term prediction effect, which proves the feasibility of introducing machine learning method to improve the mortality prediction method, provides a new idea for the government to predict future mortality, and also provides a certain basis for relevant institutions to study the risk of longevity.

Keywords: Lee-Carter model; ARIMA method; ARIMA-LSTM method; fractional O-U process; mortality prediction

[责任编辑　刘爱华]