Statistics and Application 统计学与应用, 2025, 14(12), 295-305
Published Online December 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.1412365

Hans汉斯

成分数据视角下的死亡率建模与长寿风险度量

伍思嘉，肖鸿民

西北师范大学数学与统计学院，甘肃 兰州

收稿日期：2025年11月18日；录用日期：2025年12月9日；发布日期：2025年12月22日

摘　要

该研究旨在通过引入成分数据分析(Compositional Data, CoDa)方法，对经典的Renshaw-Haberman (RH)死亡率模型进行改进，以更精确地度量和预测人口老龄化背景下的长寿风险。论文核心方法是利用中心对数比(Centered Log-Ratio, CLR)变换，将具有总和约束的死亡人数分布数据转化到无约束的欧氏空间中进行建模。通过对西班牙和澳大利亚男性死亡率数据的实证分析，研究表明，与传统的Lee-Carter (LC)模型和RH模型相比，基于成分数据框架的CoDa-RH模型在平均绝对误差(MAE)和艾奇逊距离(AD)等评估指标上表现更优。该模型不仅提高了预测精度，其预测结果还显示出与历史数据更好的拟合度，并呈现出更快的死亡率改善趋势，从而得出较高的预期寿命预测值。最终，论文将模型应用于终身生存年金精算现值的测算，结果显示新模型评估的长寿风险敞口更大，为养老金和保险行业提供了更可靠的量化管理工具。

关键词

死亡率模型，成分数据，长寿风险，Lee-Carter模型，Renshaw-Haberman模型

Mortality Modeling and Longevity Risk Measurement from a Compositional Data Perspective

Sijia Wu, Hongmin Xiao

College of Mathematics and Statistics, Northwest Normal University, Lanzhou Gansu

Received: November 18, 2025; accepted: December 9, 2025; published: December 22, 2025

Abstract

This research aims to refine the classical RH mortality model by incorporating a Compositional Data approach, enabling more precise measurement and forecasting of longevity risk against the backdrop

文章引用: 伍思嘉, 肖鸿民. 成分数据视角下的死亡率建模与长寿风险度量[J]. 统计学与应用, 2025, 14(12): 295-305.
DOI: 10.12677/sa.2025.1412365

of population aging. The methodological core involves applying CLR transformation to convert constrained mortality distribution data—characterized by a fixed-sum constraint—into an unconstrained Euclidean space for subsequent modeling. An empirical analysis of male mortality data from Spain and Australia demonstrates that the CoDa-RH model achieves superior performance relative to both the traditional LC and standard RH models, as measured by key evaluation metrics including MAE and Aitchison distance. The proposed model not only enhances predictive accuracy but also exhibits improved alignment with historical data and reflects a more pronounced trend of mortality improvement, leading to higher life expectancy forecasts. In its final application, the model is employed to estimate the actuarial present value of life annuities. The results indicate a larger longevity risk exposure under the CoDa-RH framework, offering pension and insurance sectors a more reliable quantitative tool for risk management.

Keywords

Mortality Modeling, Compositional Data, Longevity Risk, Lee-Carter Model, Renshaw-Haberman Model

1. 引言

在人类社会发展的进程中，人口结构演变始终是影响社会、经济、文化发展的关键因素。近年来，随着全球人口死亡率的持续下降与预期寿命的显著延长，人口结构正经历着前所未有的深刻变化，老龄化进程在不断加速。据联合国《世界人口展望 2022》显示，到 2022 年，全球 65 岁及以上人口约占 10%。预计到 2030 年，全世界老龄人口比例将接近 12%，2050 年将达到 16%。这意味着，基于历史死亡率经验的定价与储备，可能严重低估未来的实际支付负担。因此，发展更为精准的死亡率预测模型，是量化并管理这一系统性风险的关键。

在死亡率模型的研究中，LC 模型被视为现代随机死亡率模型的里程碑。该模型的核心假设是对数死亡率可分解为年龄效应和时期效应的线性组合，模型参数估计通常使用奇异值分解(SVD)进行，预测是通过固定年龄参数并利用单变量时间序列模型将时间依赖的指数外推到未来获得的[1]。后续许多开发死亡率模型的尝试都是从 LC 模型中获得了灵感，包括但不限于 Brouhns 等人[2]、Currir 等人[3]、Renshaw 和 Haberman [4]、Cairns、Blake 和 Dowd 等人[5]和 Plat [6]。值得注意的是，Renshaw 和 Haberman 通过引入一个队列分量来扩展 LC 模型，以捕捉同一出生队列特定死亡率趋势，该模型被简称为“RH 模型”。RH 模型在处理包含队列效应的历史数据时，其拟合效果往往优于未纳入队列效应的同类模型[7]。然而，传统死亡率在长期预测中往往呈现出一种系统性的“悲观”倾向，即预测的未来死亡率改善速度慢于实际观察到的趋势，从而导致预测的预期寿命偏于保守。为克服这一局限，Oeppen [8]开创性提出，生命表中的死亡人数分布应被定义为成分数据，因其本质上满足成分数据(CoDa)的特性。基于这一理论，作者在 CoDa 框架内利用了 LC 模型的结构并运用 Aitchison [9]的成分数据分析方法，通过对数比变换将死亡人数年龄分布从单纯形空间映射至标准的欧氏空间。这使得可以使用多元统计方法进行建模和预测，然后进行逆变换。这种方法解决了传统模型预测中经常遇到的悲观问题。在此基础之上，Bergeron-Boucher 等人[10]实现了方法论的进一步融合，将 Oeppen 的成分数据框架与 Li 和 Lee [11]用于多人口预测的协调性思想相结合，构建了首个“成分数据协调预测模型”。

综上，本文将 RH 模型的队列效应参数与成分数据框架相结合扩展为一个死亡率新模型，该模型可以捕获历史队列趋势，同时确保预测满足死亡人数分布的成分数据约束。

2. 理论基础与方法

2.1. 数据来源与预处理

本文使用的数据来源于人类死亡率数据库。它提供了 41 个国家经过严格验证的高质量死亡率和有关人口数据。本文使用的数据是死亡率模型中广为选取的西班牙数据和澳大利亚的男性生命表数据，时间跨度为 1960 年之后且年龄范围为 0~110 岁。由于高龄死亡人数容易为零，而零值无法直接进行成分数据的对数比变换，故在这里需要对数据进行预处理。为避免这个问题，这里采用 Martín-Fernández、Barceló-Vidal 和 Pawlowsky-Glahn [12]提出的方法来处理零值问题。

2.2. 传统死亡率模型

2.2.1. Lee-Carter 模型

1992 年引入的 LC 模型被定义为：

$$\ln\left(m_{x,t}\right)=a_x+b_xk_t+\varepsilon_{x,t} \tag{1}$$

其中，$x\in\{0,1,2,\cdots,\omega\}$，$t\in\{0,1,2,\cdots,T\}$，$m_{x,t}$ 是年份为 t 年龄为 x 岁人群中心死亡率；a_x 是年龄效应，表示年龄别死亡率的一个平均水平；b_x 是年龄别敏感度，衡量了不同年龄组对 k_t 的敏感程度；k_t 是时间效应，反映死亡率随时间变化的一个趋势；$\varepsilon_{x,t}$ 为误差项。其中，a_x，b_x，k_t 都为待估参数，为了解决模型可识别性的问题，还有两个约束条件，$\sum_x b_x=1$，$\sum_t k_t=0$。

为估计 LC 模型里面的系数，Lee 和 Careter 在 1992 年的研究中建立了一个经典且权威的参数框架即对中心化死亡率矩阵进行奇异值分解。具体而言，首先基于历史死亡率数据，通过计算各年龄组的平均对数死亡率来获取基准水平参数。随后，对中心化对数死亡率矩阵进行奇异值分解以提取第一主成分，从而构建时间项参数和年龄敏感性参数。参数估计完成后，模型的核心步骤是运用 ARIMA 等时间序列方法对时间成分进行外推。最终得到的预测值将与估算的年龄参数重新组合，用于计算未来各时期所有年龄组的综合预测死亡率。

2.2.2. Renshaw-Haberman 模型

RH 模型是 LC 模型的重要扩展。通过纳入队列效应参数捕获不同出生时期所经历的独特死亡率趋势，从而增强模型的解释力。RH 模型常用的形式为：

$$\ln\left(m_{x,t}\right)=a_x+b_xk_t+\gamma_{t-x}+\varepsilon_{x,t} \tag{2}$$

该模型的约束条件为，$\sum_x b_x=1$，$\sum_t k_t=0$，$\sum_{t-x}\gamma_{t-x}=0$。RH 模型的预测因纳入队列效应增添了新的维度，该模型不仅需要外推时间效应项，还需外推同一出生队列的队列效应项。最终，预测的中心死亡率由外推得到的时间效应项、队列效应项，与固定的年龄效应参数 a_x 和 b_x 共同构成。

2.3. 基于成分数据的改进模型

成分数据受到“部分–整体”关系的严格约束，其中，它的各个分量的比例之和总是等于 1。生命表中的死亡人数年龄分布符合成分数据的定义，其中每个年龄对应的死亡人数都是大于零的且每一年所有年龄总和的死亡人数等于生命表基数(通常为 100,000 或 1)。这种约束意味着特定年龄的死亡人数之间存在相互依赖性，第一个年龄组的死亡人数减少必然会导致至少其他一个年龄组的死亡人数增加。为在标准的欧氏空间中进行有效的统计分析，成分数据框架提供了一套全面的对数比变换。这里采用中心对数

比变换：

$$clr\left(d_{x,t}\right)=\ln\left(\frac{d_{x,t}}{g_t}\right) \tag{3}$$

其中，g_t 表示 t 时刻年龄构成的几何平均值。clr 变换通过除以几何平均值将成分数据从单纯形空间映射到标准的欧式空间。这消除了数据的约束，允许数据自由变化，满足了大所述经典多元统计方法的先决条件。

2.3.1. CoDa-LC 模型

Oeppen 在 LC 模型的基础上，在成分数据框架下开发了以下死亡率模型：

$$clr\left(d_{x,t}\ominus a_x\right)=b_x k_t+\varepsilon_{x,t} \tag{4}$$

其中，a_x 表示特定年龄死亡人数 $d_{x,t}$ 的几何平均值，k_t 和传统 LC 模型参数一样是时间效应，b_x 是通过 SVD 得到的年龄别模式，表示死亡密度在不同年龄之间的转移方向。$\varepsilon_{x,t}$ 为误差项，$\ominus$ 属于标准 CoDa 算子，其定义为一种扰动处理方法。该模型的参数估计过程如下：

首先，计算整个时期所有死亡人数的几何平均值：

$$a_x=\exp\left(\frac{1}{T}\sum_{t=1}^{T}\ln\left(d_{x,t}\right)\right) \tag{5}$$

然后，通过数据中的扰动程序，对原始数据相对于该基线进行中心化处理，得到：

$$f_{x,t}=d_{x,t}\ominus a_x \tag{6}$$

为克服数据的约束，对中心化后的数据进行 clr 变换：

$$h_{x,t}=clr\left(f_{x,t}\right)=\ln\left(\frac{f_{x,t}}{g_t}\right),g_t=\left(\prod_{x=1}^{X}f_{x,t}\right)^{\frac{1}{X}} \tag{7}$$

对变换后的矩阵 $\boldsymbol{H}$ (其元素为 $h_{x,t}$)执行奇异值分解，相应的模型的核心参数估计为：

$$k_t=u_{t,1}s_1,b_x=v_{x,1} \tag{8}$$

2.3.2. CoDa-RH 模型

在此 CoDa-LC 模型的基础上，将 RH 死亡率模型也融入 CoDa 框架，构建 CoDa-RH 死亡率模型，核心表达式为：

$$clr\left(d_{x,t}\ominus a_x\right)=b_x k_t+\gamma_{t-x}+\varepsilon_{x,t} \tag{9}$$

其中，γ_{t-x} 代表队列效应，捕捉不同年份出生的特定时代所经历的独立死亡风险。其余参数解释与 CoDa-LC 模型解释相同。参数估计的初始步骤与 CoDa-LC 模型相同，一直到 $\boldsymbol{H}$ 执行奇异值分解。其中第一主成分用于初始化时间效应 $k_t^{(0)}$ 和年龄模式 $b_x^{(0)}$。在此初步估计的基础上，通过残差矩阵 $\boldsymbol{R}=\boldsymbol{H}-b_x^{(0)}k_t^{(0)}$ 计算同一出生队列中所有人数的均值来初始化队列效应 γ_{t-x}。随后，采用迭代算法交替更新参数，以最小化拟合误差，直到实现收敛。

2.4. 模型评估指标体系

为了在拟合优度和预测精度方面评估模型性能，选择了以下具有代表性的统计指标。这两个指标从不同维度衡量模型与实际数据的一致性，形成一个系统性的评估框架。

平均绝对误差(MAE)衡量的是预测值与实际值之间绝对误差的平均值。该指标对离群值不敏感，其

物理意义与预测目标保持一致，因此可以直观地反映预测的总体误差水平。

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right|, \tag{10}$$

艾奇逊距离(Aitchison Distance)是成分数据分析中专门的距离度量，该度量基于中心对数比变换，有效地处理了成分数据的相对结构和闭合效应，准确地衡量了两个成分向量之间的差异。

$$AD(x,y)=\sqrt{\sum_{i=1}^{M}\left[clr(x_i)-clr(y_i)\right]^2}, \tag{11}$$

其中 M 表示成分的维度，$clr(\cdot)$ 表示中心对数比变换。

3. 实证分析

3.1. 模型参数估计结果对比分析

传统死亡率模型和 CoDa 模型的参数采用类似的符号表示，共享类似的解释，但不完全相同。图 1 是利用澳大利亚男性数据拟合 LC 模型和 CoDa-LC 模型的参数估计，图 2 是利用澳大利亚男性数据拟合 RH 模型和 CoDa-RH 模型的参数估计。

参数 k_t 反映了总体死亡率随时间的变化，从图 1 和图 2 中可以发现 k_t 随着时间的推移呈现出近似线性趋势。然后对于传统模型而言，这些值随时间增加而减少，而成分数据框架下的模型表现出相反的趋势(这里值得注意的是，在成分数据框架下的模型，参数 b_x 和 k_t 不是唯一确定的，有时两组参数都随时间增加，有时会减少[10])。对于 k_t 的预测，采用 R 软件中的 auto.arima()函数进行时间序列建模。该函数通过对 AIC 和 BIC 等信息准则的优化，确定四个模型 k_t 序列的最优 ARIMA 模型结构，对于参数 γ_{t-x} 的预测也采用这种方法。

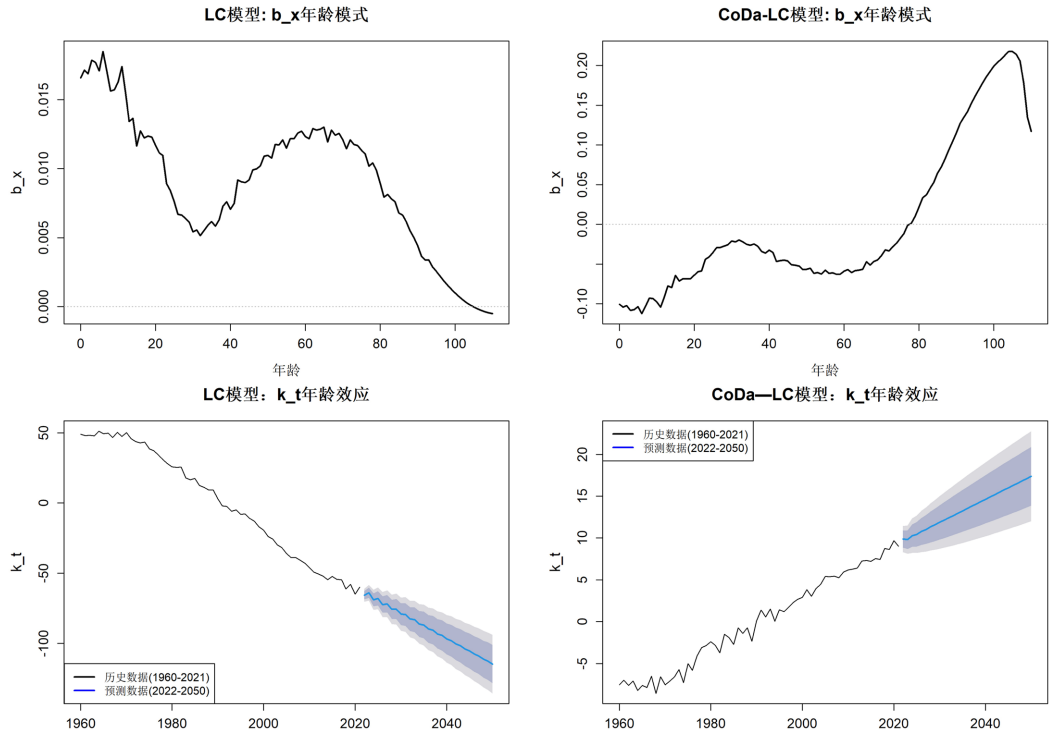


Figure 1. Parameter estimation of the LC model and CoDa-LC model for Australian males
图 1. 澳大利亚男性 LC 模型与 CoDa-LC 模型参数估计

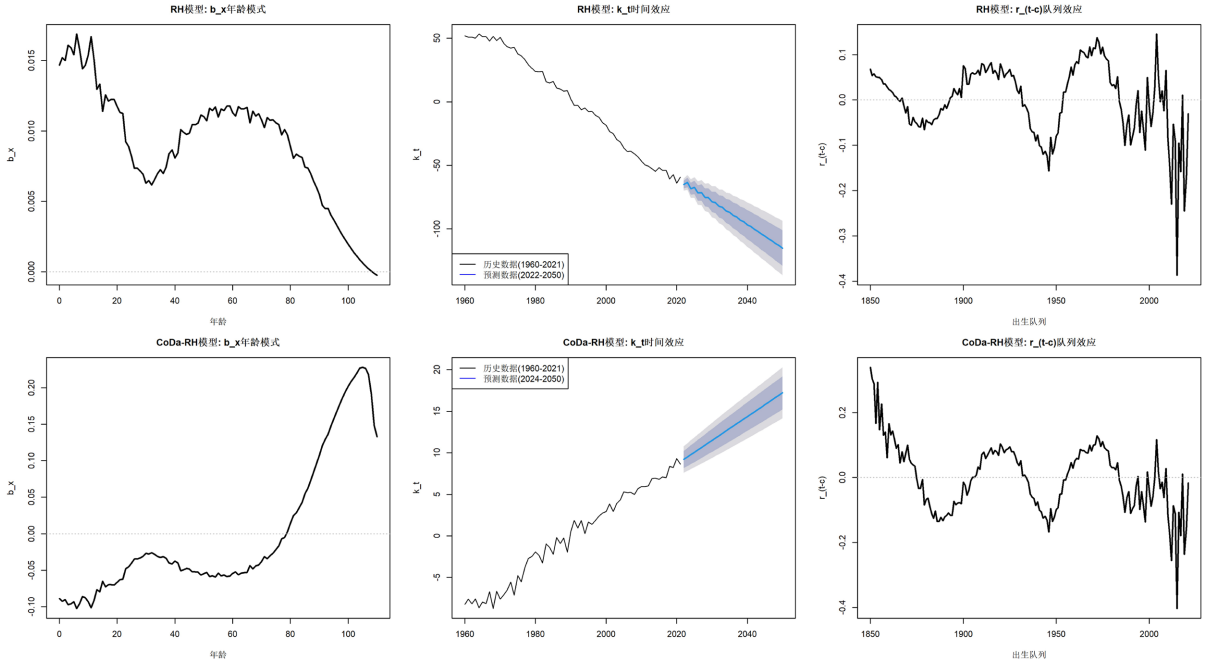


Figure 2. Parameter estimation of the RH model and CoDa-RH model for Australian males
图 2. 澳大利亚男性 RH 模型与 CoDa-RH 模型参数估计

3.2. 模型拟合与预测精度比较

表 1 评估了四种死亡率模型在澳大利亚和西班牙 2000 年后的预测性能。所有模型均基于 1960~2000 年数据进行训练，并在 2000 年后的数据上进行验证。基于 MAE 的结果表明，不论是 LC 模型还是 RH 模型，结合成分数据的方法都提高了模型的预测性能。特别是对于澳大利亚的数据来说，CoDa-RH 模型展示出最小的误差，同样对于西班牙国家的数据来说，尽管结果之间的差异较小，但也体现出成分数据方法的优势。从分布拟合角度看，基于艾奇逊距离(AD)的结果显示，考虑队列效应的 RH 模型系列在捕捉死亡率分布方面明显优于 LC 模型系列。

Table 1. Comparison of error metrics under four models
表 1. 四种模型下误差指标比较

国家	指标	LC	CoDa-LC	RH	CoDa-RH
澳大利亚	MAE	0.19	0.18	0.14	0.11
	AD	2.68	2.54	2.10	1.85
西班牙	MAE	0.28	0.25	0.26	0.26
	AD	3.85	3.65	3.22	3.13

3.3. 预期寿命估计

在拟合四种模型后，就可以进行死亡率的预测了。建立完善、准确、高效的生命表是保险精算、人口预测及养老金规划等领域的基础[13]。生命表分为静态生命表与动态生命表，但静态生命表没有考虑未来死亡率随时间改善的趋势，通常会有低估预期寿命的倾向。故在这里使用动态生命表来进行预测，以

更准确地反映人口的未来生存轨迹。

长寿风险指因人口预期寿命的实际改善幅度持续超出预期而引发的财务风险，对养老金体系与保险公司的长期偿付能力构成严峻挑战。为评估不同模型所揭示的风险敞口，本研究选取 55、60 和 65 岁这三个退休规划与政策制定中的关键年龄作为分析节点。不同死亡率模型因结构差异，对未来改善趋势的捕捉能力不同，其预测结果所隐含的长寿风险水平亦存在显著区别。

表 2 对比了各模型在相应年龄点上的预期寿命预测结果。总体而言，CoDa-RH 模型在多数情况下给出最高的预期寿命估计，尤其在澳大利亚数据中表现突出；相比之下，传统 LC 模型的预测值普遍最低，提示其可能系统性地低估未来的生存改善程度，从而隐含更大的未被充分认知的长寿风险。从方法层面看，成分数据(CoDa)模型通过引入对数比变换，将年龄别死亡率作为构成性数据进行建模，有效约束于概率空间之内，从而更稳健地捕捉其内在结构与协同变化。因此，该类模型能够反映出更为持续而稳健的死亡率下降趋势[10]。一个具有一致性的发现是，与传统模型相比，基于 CoDa 框架的模型在绝大多数情境下都给出了更高的剩余预期寿命预测值，呈现出更为乐观的长期生存改善前景。

Table 2. Prediction results of remaining life expectancy
表 2. 剩余预期寿命预测结果

国家	Year	Age	LC	CoDa-LC	RH	CoDa-RH
澳大利亚	2020	55	27.75	28.89	28.54	29.53
		60	23.31	24.41	24.12	25.11
		65	19.19	20.12	19.94	20.88
	2025	55	27.89	29.03	28.71	29.47
		60	23.50	24.54	24.29	25.07
		65	19.31	20.23	20.06	20.87
西班牙	2020	55	26.45	26.49	27.39	26.71
		60	22.21	22.23	23.10	22.57
		65	18.19	18.19	19.01	18.64
	2025	55	26.58	27.41	27.60	27.61
		60	22.33	23.06	23.28	23.30
		65	18.30	18.91	19.18	19.26

4. 长寿风险度量应用分析

长寿分险的度量和动态死亡率的预测密不可分，对人口寿命延长导致的长寿风险问题，一般使用生存年金现值的方法对长寿分险进行度量，基于精算模型的理论知识，最经典的终身生存年金精算现值表达式如下：

$$\ddot{a}_x = \sum_{t=1}^{\omega-x} {}_tP_x v^t \tag{12}$$

其中，ω 为极限年龄(这里设为 110 岁)，${}_tP_x$ 为 x 岁的人活过 t 年的生存概率，v 为折线因子，表达式为：

$v=(1+i)^{-1}$，i 为折现率，$\ddot{a}_x$ 表示期初付终身生存年金的精算现值。

图 3 和图 4 展示了在四种模型下对西班牙和澳大利亚男性三个关键年龄、对应三个不同利率的终身生存年金精算现值。分析可见，在固定利率下，2010 至 2025 年间的精算现值均呈现明确上升趋势，该现象在所有贴现率与死亡率模型下保持一致，直观揭示了长寿风险的加剧。从模型差异看，LC 模型因预测相对乐观，估算的精算现值通常最低；而 CoDa-RH 模型因考虑了队列效应，预测更为保守，故得出的精算现值最高，对长寿风险的度量也更为充分。

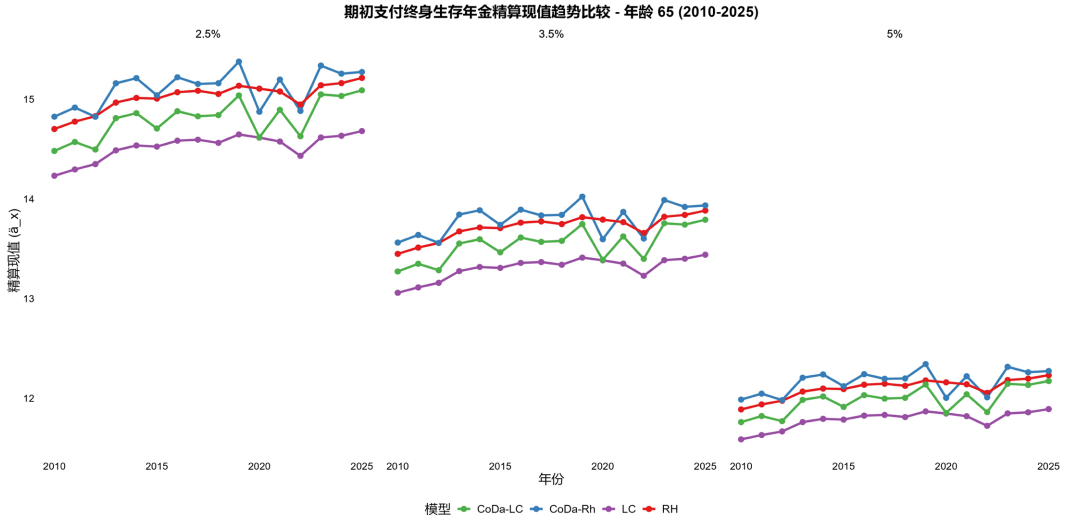
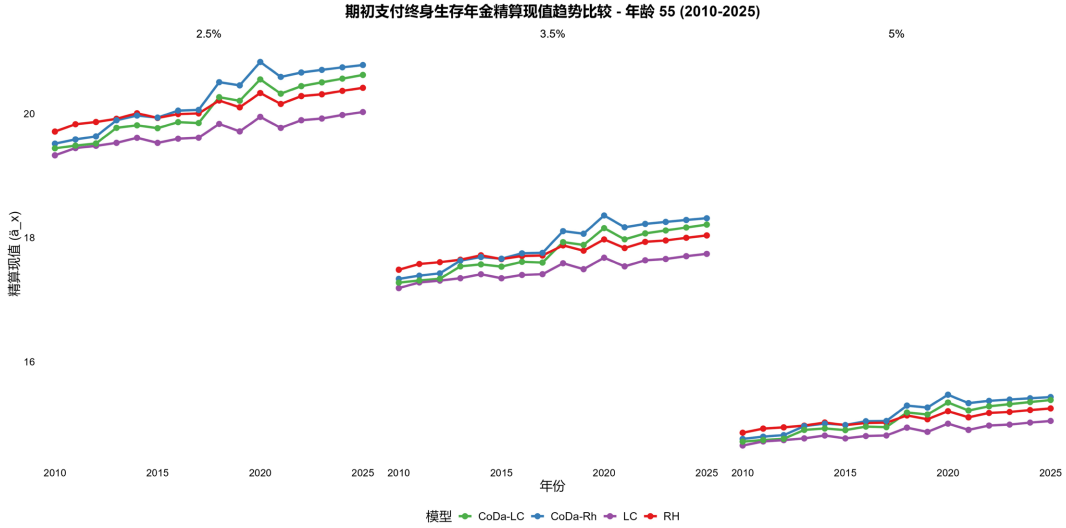
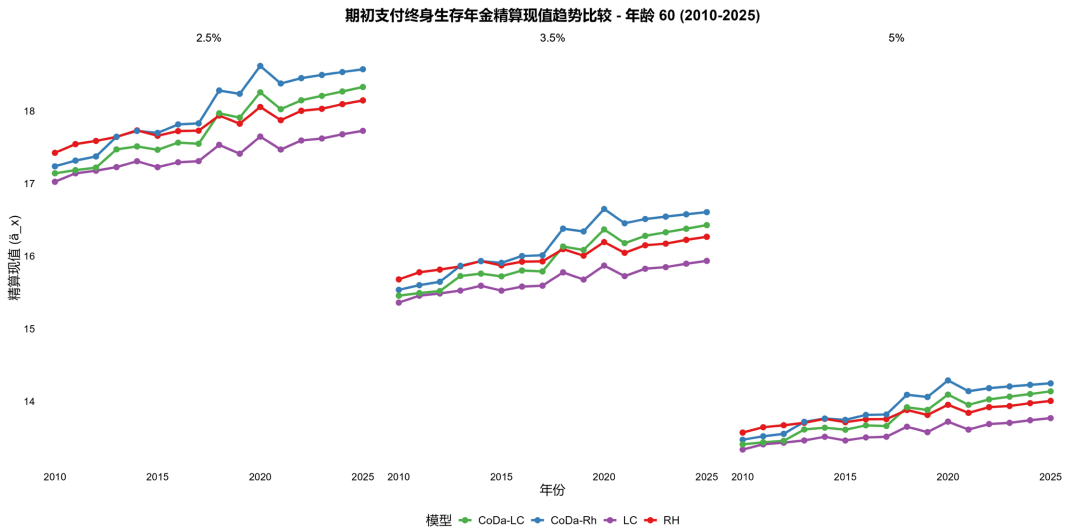


Figure 3. Predicted results of actuarial present value of lifetime survival annuities for Spanish males
图 3. 西班牙男性终身生存年金精算现值预测结果





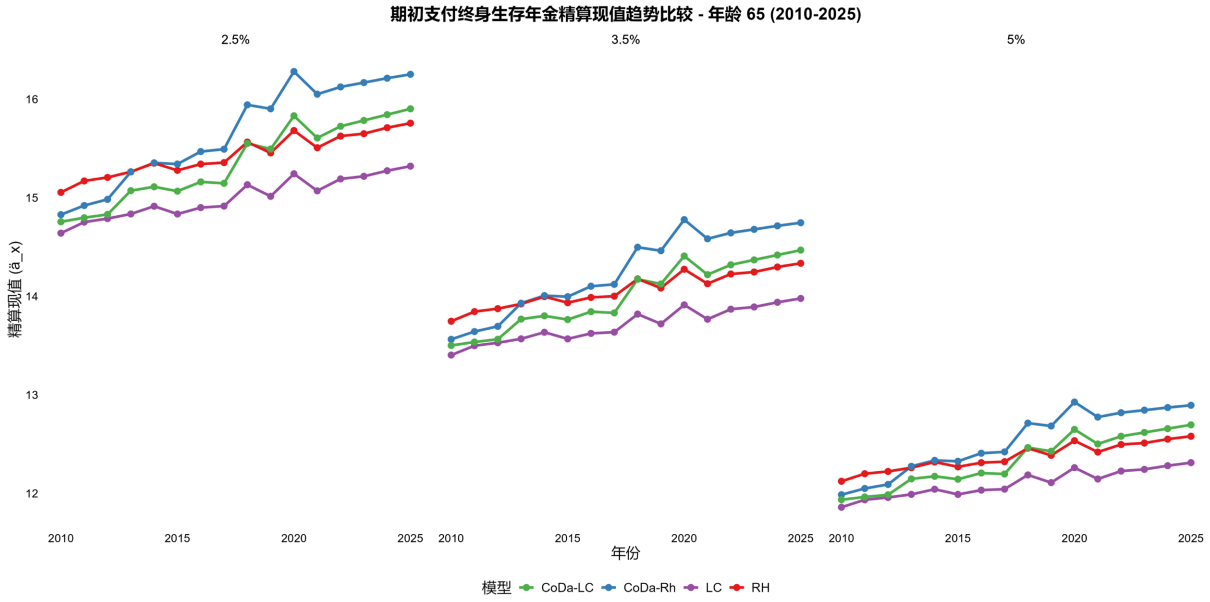


Figure 4. Predicted results of actuarial present value of lifetime survival annuities for Australia males
图 4. 澳大利亚男性终身生存年金精算现值预测结果

5. 结论与展望

本文引入成分数据框架对传统 RH 模型进行拓展，提出了一种新的死亡率预测方法。新模型不仅能有效捕捉历史数据中的时期与队列效应，还凭借其固有的成分特性，确保了死亡人数分布预测的逻辑一致性。在预测准确性上，该模型在误差指标上展现出一定优势。未来可考虑引入具有完整高质量死亡率数据的国家进行多国模型构建，以增强模型结构的泛化能力。由于中国大陆的数据量少，且缺失值多，待未来中国数据条件进一步改善后，也可进一步验证该模型在中国人口场景下的适用性与预测精度。

基金项目

国家自然科学基金(NO. 12061066)；甘肃省自然科学基金(NO. 20JR5RA528)。

参考文献

[1] Lee, R.D. and Carter, L.R. (1992) Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association*, **87**, 659-671. https://doi.org/10.1080/01621459.1992.10475265

[2] Brouhns, N., Denuit, M. and Vermunt, J.K. (2002) A Poisson Log-Bilinear Regression Approach to the Construction of Projected Lifetables. *Insurance*: *Mathematics and Economics*, **31**, 373-393. https://doi.org/10.1016/s0167-6687(02)00185-3

[3] Currie, I.D., Durban, M. and Eilers, P.H.C. (2006) Generalized Linear Array Models with Applications to Multidimensional Smoothing. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **68**, 259-280. https://doi.org/10.1111/j.1467-9868.2006.00543.x

[4] Renshaw, A.E. and Haberman, S. (2006) A Cohort-Based Extension to the Lee-Carter Model for Mortality Reduction Factors. *Insurance*: *Mathematics and Economics*, **38**, 556-570. https://doi.org/10.1016/j.insmatheco.2005.12.001

[5] Cairns, A.J.G., Blake, D. and Dowd, K. (2006) A Two-factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration. *Journal of Risk and Insurance*, **73**, 687-718. https://doi.org/10.1111/j.1539-6975.2006.00195.x

[6] Plat, R. (2009) On Stochastic Mortality Modeling. *Insurance*: *Mathematics and Economics*, **45**, 393-404. https://doi.org/10.1016/j.insmatheco.2009.08.006

[7] Guo, Y. and Li, J.S. (2025) Fast Estimation of the Renshaw-Haberman Model and Its Variants. *European Actuarial Journal*, **15**, 633-666. https://doi.org/10.1007/s13385-025-00407-w

[8] Oeppen, J. (2008) Coherent Forecasting of Multiple-Decrement Life Tables: A Test Using Japanese Cause of Death Data. *The European Population Conference 2008*, Barcelona, 9-12 July 2008, 9-12.

[9] Aitchison, J. (1986) The Statistical Analysis of Compositional Data. Chapman and Hall.

[10] Bergeron-Boucher, M., Canudas-Romo, V., Oeppen, J.E. and Vaupel, J.W. (2017) Coherent Forecasts of Mortality with Compositional Data Analysis. *Demographic Research*, **37**, 527-566. https://doi.org/10.4054/demres.2017.37.17

[11] Li, N. and Lee, R. (2005) Coherent Mortality Forecasts for a Group of Populations: An Extension of the Lee-Carter Method. *Demography*, **42**, 575-594. https://doi.org/10.1353/dem.2005.0021

[12] Martín-Fernández, J.A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (2003) Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Mathematical Geology*, **35**, 253-278. https://doi.org/10.1023/a:1023866030544

[13] 肖鸿民, 赵弘宇, 马海飞. 中国人口死亡率建模比较及长寿风险度量[J]. 经济数学, 2020, 37(4): 11-18.