基于夏普利值的方法：利用集成学习提升死亡率预测

Giovanna Bimonte Maria Russolillo

经济与统计系 萨勒诺大学，教宗若望保禄二世街*132*号，菲夏诺，*84084*，意大利

Yang Yang

信息与物理学院 纽卡斯尔大学数理科学系，大学大道，卡拉汉，新南威尔士州*2308*，澳大利亚

Han Lin Shang

精算学系 与商业分析系，麦考瑞大学，东路*4*号，悉尼，新南威尔士州*2109*，澳大利亚 *a*

arXiv:2603.03789v1 [stat.AP] 4 Mar 2026

摘要

死亡率预测领域一项公认的见解是，与依赖单一最佳模型相比，综合一组模型的预测能够提高准确性。本文提出一种基于Shapley值的新颖集成方法。Shapley值是一种博弈论度量，用于衡量每个模型对预测的边际贡献。我们进一步逐年龄计算基于夏普利加性解释（SHAP）的权重，从而捕捉每个模型在各年龄上的具体贡献。此外，我们引入一种阈值机制，剔除贡献可忽略不计的模型，从而有效降低预测方差。利用24个经合组织国家的数据，我们证明，SHAP 集成能够提升样本外预测性能，尤其是在较长预测期下。通过发挥不同死亡率模型的互补优势，并筛除预测能力贡献甚微的模型，我们的方法为改进死亡率预测提供了一种稳健且易于解释的方案。

关键词：死亡率模型；集成学习；夏普利值；预测准确度；预测组合。

1. 引言

死亡率预测在多个领域都至关重要，包括社会保障规划、保险定价和公共卫生政策制定。多年来，精算学和人口统计学文献提出了许多随机死亡率模型，用以捕捉单个人群内死亡率的动态变化。尽管单个模型往往在特定情境下表现出优势，但没有任何模型能在所有数据集或预测期限上始终优于其他模型

(Makridakis 等人，2020年）。为解决这一局限，许多研究建议结合多个模型的预测（例如，参见Bates and Granger，1969年；Li 和 Lee，2005年；Shang 等人，2011；Samuels 和 Sekkel，2017），因为集成方法通常比依赖单一“最佳”模型能得出更准确的预测。然而，并非所有模型组合都能带来更好的结果，因为多种因素可能导致某些组合比其他组合更成功。理解某些集成表现更优的内在因素仍是一项关键挑战，相关问题已得到广泛研究，以确定最佳模型组合。

理想的集成策略应考虑各模型的相对贡献，并根据预测价值动态调整其权重。构建集成的一种常见方法是为所有模型分配等权重，实质上就是对它们的预测取平均值。尽管这种方法简单易行，但研究表明，在许多情况下，它的表现优于单个模型。然而，这种方法没有考虑不同数据集和时间段中模型性能的固有差异，因此可能导致次优结果。从偏差—方差权衡的角度来看，等权重的预测组合往往存在偏差，但方差较低；而使用估计权重的组合通常无偏，但方差较高。

我们提出一种新颖的集成方法，利用 SHAP值（Lanctot 等人，2017），为每个死亡率模型确定最优权重。SHAP值是一种博弈论度量，可解释任何机器学习模型的输出。它借助博弈论中的经典 Shapley值，将最优信用分配与局部解释联系起来，并根据每个模型的边际性能，将其贡献分配给集成。在预测组合中使用 SHAP值，早已受到预测学界的关注。例如，Franses 等人（2024）提出一种新的预测组合方法，通过 Mincer-Zarnowitz回归估计 Shapley值权重。Qin 等人（2022年）在中国城市燃气需求的预测组合中使用了 Shapley值。虽然估计权重通常是静态的，但模型的预测能力会随时间变化。因此，考虑自适应权重至关重要。例如，Elliott 和 Timmermann（2005年）提出了一种新的预测组合方法，使组合权重受潜在状态变量的机制转换驱动。我们的方法能够适应模型预测能力的变化，从而提升死亡率预测的可靠性和精确度。

借鉴近期文献，我们聚焦于组合预测方法时面临的两个重要问题：准确性和多样性（Lichtendahl Jr 和 Winkler，2020年）。人们普遍认为，在集成中纳入表现不佳的方法会损害预测准确度。同时，即使表现良好的方法，如果缺乏足够的多样性，也可能无法取得出色效果。为同时解决这两个问题，我们提出一种方法，将近几十年来统计学家和精算师最常用的时间序列外推方法所采用的 15 种不同死亡率预测模型组合起来。随后，我们通过考察两种策略来优化该集成：一种是不设任何筛选阈值，直接使用 SHAP值；另一种则设置阈值，排除贡献较低的模型。后一种做法保留了集成内部的多样性，并提升了点预测和区间预测的准确性，凸显了在死亡率预测中进行策略性模型选择的价值。

本文结构安排如下。在第 2 节中，我们回顾不同的死亡率模型和几种集成方法（模型方程和约束条件见 附录A）。在第 3 节中，我们介绍 SHAP 加权方案的理论基础及其实现方法。在第 4 节中，我们详细介绍实证分析中所提方法和基准方法的实现。在第 5 节中，我们使用大量死亡率数据（24 个经合组织国家）评估所提方法，并将其与传统集成方法进行比较。第6节讨论基于 SHAP 和传统方法构建的集成预测区间。结论见第7节，并提出了进一步拓展本文方法的一些思路。从偏差-方差权衡的角度出发，我们在 附录B 中考察了各种加权方案的均方误差分解。

2. 文献综述

2.1. 符号说明

我们介绍各个入选死亡率模型所依据的符号和关键要素，并阐明不同参数结构如何刻画死亡率随年龄、时间和队列变化的差异。具体而言：

- $m_{x,t}$ 是年龄 x 和日历年 t的中心死亡率。
- $D_{x,t}$ 表示观察到的死亡人数，假设其服从泊松分布。

- $E_{x,t}$ 表示年龄 x 和年份 t的中心暴露量。
- $\alpha(x)$ 表示整个拟合期内对数死亡率的基线水平，反映平均年龄结构。
- $\kappa_t^{(i)}$ 是随日历年 t演变的潜在时间因子。
- $\beta^{(i)}(x)$ 调节每个 $\kappa_t^{(i)}$ 对年龄为 x时对数死亡率的影响。
- γ_{t-x} （或带有 $c = t - x$的 γ_c ）反映队列特异性变异（即按出生年份划分）。
- 求和约束（例如 $\sum_t \kappa_t^{(i)} = 0$ 或 $\sum_x \beta^{(i)}(x) = 1$）可消除模型参数中冗余的平移或缩放。

2.2. 个体年龄别死亡率预测模型

我们的方法首先整合了多种个体死亡率模型，每种模型都捕捉死亡率动态的不同方面，例如趋势、死亡率波动性、线性特征和队列效应。这些模型构成了评估我们所提出的集成框架的关键基础。基于 Shang andHaberman(2018，我们聚焦于一类成熟的离散时间死亡率模型（汇总于 A.1– A.4 中的 附录A），并结合了近期文献中使用的 15 种最流行的外推预测方法。尽管我们承认分析中可以纳入更广泛的模型，但也认识到，实施大量模型会带来显著的计算挑战。因此，这一经过精心筛选的模型集合兼顾了广度与可处理性，使我们能够严谨地评估集成方法的性能优势。

基于这一框架，表A.1 和A.2 突出了两大主要模型族，即 Renshaw-Haberman（RH，模型 1–3）和 Cairns-Blake-Dowd（CBD，模型4–8），它们已被广泛用于刻画不同的死亡率模式，包括线性趋势和年龄别死亡率改善。我们采用死亡率表的泊松对数双线性形式及其在伦肖-哈伯曼族中的队列增强扩展，并将其与年龄–时期–队列传统联系起来（Brouhns 等人，2002年；Renshaw 和 Haberman，2006年；Holford，1983年）。在伦肖-哈伯曼族中，我们保留基本的泊松 Lee–Carter 模型M1及其队列增强扩展M2和M3，它们已经提供了一系列年龄–时期–队列

结构。更一般的多因子 RH 规格，例如本着年龄特定增强模型的思路增加第二个奇异向量，参见 Renshaw 和 Haberman（2003年），原则上是可行的，也代表了该模型族的自然发展。然而，我们在此不探讨这些更丰富的 RH 变体，以避免相对于其他模型族使高度相近的规格数量膨胀，并保持整体模型集合的可处理性。

对于 Cairns–Blake–Dowd 家族，我们采用基准双因子 CBD 模型以及广泛应用的队列扩展与年龄二次项扩展（M6–M8），并将 Plat 规格作为一种灵活的年龄－时期－队列变体（Cairns 等人，2006年，2009年；Plat，2009年）。最初的 Cairns–Blake–Dowd（CBD）家族模型主要针对老年死亡率开发，通常从约55岁及以上开始；在该年龄段，简约的年龄线性或二次规格能够很好地描述对数死亡率曲线。然而，在我们的实证设定中，全部15个模型均拟合于相同的年龄范围 0–100，以确保集成具有同质的输入集合，并允许对预测表现进行完全可比的评估。因此，对于 CBD 型规格（M4–M6），这应被理解为对原本面向老年年龄设计的模型的外推。我们不以结构性方式解释年轻年龄上的 CBD 参数；相反，这些模型被用作额外的基学习器，其对最终预测的贡献由其样本外表现通过 SHAP 加权方案事后决定。

在表 A.3中，LC（模型 9–12）系列在死亡率建模中采用参数结构。具体而言，Booth-Maindonald-Smith（BMS）模型是对 LC 模型的改进版本，旨在提高死亡率预测的稳定性和可靠性。加入预期寿命调整的 LC 模型通过引入修正来考虑预期寿命的变化，从而增强其捕捉长期趋势的能力。相反，不进行得分调整的 LC 模型省去了参数的事后校准，采用更直接的估计过程，无需额外调整。对于 Lee–Carter 变体，我们采用原始的奇异值分解形式、Booth–Maindonald–Smith 的校准改进方法以及 Lee–Miller 的水平调整程序；近期综述讨论了有关得分调整的具体操作选择（Lee 和 Carter，1992年；Booth 等人，2002年；Lee 和 Miller，2001年；Basellini 等人，2022年）。

模型13至15（见表 A.4），被归类为函数型时间序列模型，采用非参数平滑和函数型数据分析技术，以产生准确的死亡率预测，同时

确保对离群值的稳健性。具体而言，函数型数据模型（FDM）将函数型主成分分析（FPCA）应用于整条死亡率曲线，从而更全面地刻画死亡率趋势随时间的变化。稳健函数数据模型扩展了FDM，增强其稳健性并减轻离群值的影响，从而得到更稳定的预测。最后，由Hyndman等人（2013年）提出的乘积比模型采用一种基于各亚组（如女性和男性数据）死亡率函数之比的死亡率建模方法，从而有效刻画不同人群之间的相对死亡率模式。对于函数型数据方法，我们考虑用于年龄别死亡率的标准FDM、用于处理离群值的稳健变体，以及通过乘积比法进行的协调的多人口预测（Hyndman和Booth，2008年；Hyndman和Ullah，2007年；Hyndman等人，2013年）。

上述模型广泛应用于死亡率建模和预测，各自具有独特的优势与局限性。LC模型因其简洁性和可解释性而广泛应用于精算和人口统计应用中。函数型时间序列模型假设存在一个平滑的潜在随机过程，用于刻画分年龄死亡率的变化，从而有效捕捉随时间变化的死亡率模式。

因此，每个模型族都为死亡率变化建模提供了不同的视角——有些强调参数假设（RH、CBD和LC变体），而另一些则采用平滑方法（FDM）。尽管这些模型在方法论上存在差异，但这15个模型都依赖于对历史模式的外推，因此其预测结果在一定程度上具有相关性。尽管如此，每个模型都具有独特的优势与劣势，这凸显了通过集成方法将它们结合起来的价值。以这一组多样化的模型为基础，我们可以更好地评估每个模型对预测准确度的贡献，进而评价所提出集成方法的优势。

2.3. 预测集成方法

集成建模已成为现代预测应用的基石，包括死亡率预测，这得益于其结合多个模型并利用它们多样化预测优势的能力。集成方法不依赖单个“最佳”模型，而是能够减轻个别模型的弱点、降低方差并减少总体偏差，最终改善高度复杂情境下的预测表现（例如，参见 Bimonte等人，2025年）。

集成构建的一项核心挑战，是为组合

各构成模型选择适当的权重。简单模型平均（SMA）等方法为所有模型赋予相同的权重，虽然简单易行，却往往无法反映模型性能的差异。贝叶斯模型平均（BMA）是一种更复杂的方法，它利用后验概率计算权重，并将各模型对数据的拟合情况纳入考量（例如，见 Benchimol等人，2016；Shang and Haberman，2018）。在死亡率研究中，Benchimol等人(2016)使用基于赤池信息准则（AIC）的权重和贝叶斯模型平均，对标准的单人口模型进行组合；而De Mori等人(2024)则分析了双人口设定下的若干模型平均方案，其权重基于各模型在训练期内的预测表现，在概念上与本文考察的基于 AIC 的加权基准相近。虽然BMA考虑了模型性能，但在使用高度复杂的模型或后验分布对先验分布敏感时，仍可能出现过拟合。

尽管这些技术能够提供可靠的基线，但其静态特性可能会限制对不断变化的数据模式以及不同预测期限的适应能力，从而可能导致过拟合，或无法捕捉替代模型所带来的改进。

机器学习的最新进展催生了更具适应性的集成方案，包括堆叠泛化（Wolpert，1992年；Kessy等人，2022年）和基于 SHAP 的加权（Lundberg和Lee，2017年）。二者同属集成方法族，但在评估和组合各个模型输出的方式上有所不同。

堆叠泛化（通常称为堆叠）会在基模型的预测结果上训练一个“元学习器”，并调整其权重，以最小化验证集上的预测误差。该方法明确以预测准确度为导向，使能够降低总体误差的模型发挥更大影响。

基于 SHAP 的加权运用合作博弈论来量化每个模型对集成的边际贡献。它并非直接以误差最小化为优化目标，而是根据每个模型对集成的可解释性或预测能力的提升程度赋予“功劳”。该过程可针对不同的预测期限或数据结构重复进行，从而更动态地了解哪些模型在特定情境中贡献最大。

在死亡率预测中，这些集成技术尤其有价值。死亡率数据往往呈现出年龄、时间和队列维度上的复杂模式。依赖单一模型

可能会忽略其他模型所捕捉到的宝贵信息，而在未仔细权衡各模型贡献的情况下将多个模型结合起来，则可能导致过拟合或计算效率低下。通过评估各模型的预测性能和边际贡献，集成方法能够捕捉更广泛的趋势，例如新出现的队列效应、年龄特异性改善以及整体死亡率水平的变化，同时减弱噪声和特异性误差。

我们采用基于SHAP的集成方法，在多个预测期优化死亡率预测。与对所有预测采用同一组权重的静态加权方案不同，我们的方法会在每个预测期重新计算各模型的 SHAP值，从而动态适应模型性能随时间的变化。通过设定 SHAP值的最小阈值，筛除预测效用较低的模型，进而降低计算复杂度并增强稳定性。

本文旨在探讨两个问题：1）我们考察基于 SHAP 得出的最优权重构建的集成，能否在死亡率预测中胜过传统的模型组合技术。2）我们通过将基于SHAP的加权与等权重等更简单的加权策略进行比较，探讨采用基于SHAP的加权时偏差-方差权衡如何演变。

通过探讨这些问题，本文从三个方面对现有文献作出贡献。1）我们提出一种自适应集成方法，根据各死亡率模型在不同预测期限内的估计贡献，为其分配动态权重。2）我们通过实证评估基于 SHAP 的集成加权相较于传统方法对预测准确度和方差缩减的影响。3 ）最后，我们对死亡率模型集成的偏差-方差分解进行形式化阐述，说明基于 SHAP 的加权如何提升预测稳定性和稳健性。

在第 3 节中，我们介绍这一集成框架的理论基础，并说明它如何平衡准确性、多样性和可解释性；这些都是实现可靠且稳健的死亡率预测的关键属性。

3. 模型集成准则

样本外评估技术在设计有效的集成时发挥着关键作用。这些方法确保在未用于训练的数据上评估模型性能，从而更可靠地估计泛化误差。交叉验证、滚动窗口评估和基于区块的验证，为比较模型在不同预测期限上的表现并优化其对集成的贡献提供了框架。

整合样本外表现指标，使集成能够应对模型不确定性，并有效提升不同预测期限上的预测准确性。

我们提出的 SHAP 集成方法通过动态计算权重并为每个模型分配边际贡献，从而优化准确性和多样性，与现有方法形成了鲜明区别。引入阈值参数后，可以排除相关性较低的模型，在不牺牲准确性的前提下降低总体方差。

我们探讨了构建有效集成的关键标准，重点关注如何确定最优权重、阈值设定在提升模型质量方面的作用，以及偏差与方差之间至关重要的平衡。本分析表明，SHAP 将准确性与可解释性独特地结合起来，在短期和长期预测期限内均能带来更优的预测表现。

3.1. 最优加权策略

选择最优权重对于提升死亡率预测集成模型的准确性和稳健性至关重要。加权方案的首要目标是在各个模型之间实现适当的平衡，利用它们互补的优势，同时减轻其不足。文献中已探讨了多种权重计算方法，每种方法都有各自的优势和局限性。

实证证据表明，与简单模型平均（SMA）和赤池信息准则权重等较简单的方法相比，采用最优加权方案的方法（如堆叠回归集成和基于SHAP的集成）能够取得更优的长期预测表现。这些先进方法具有更强的可解释性和灵活性，因此特别适用于应对死亡率数据中的挑战，最终有助于生成准确、可靠的预测，供精算学和公共卫生领域使用。

SHAP值通过量化每个模型对整体预测的贡献，为权重计算提供了一个博弈论框架，且量化粒度非常精细，从而确保了较强的可解释性。此外，引入阈值参数可确保排除贡献极小的模型，避免集成受到表现不佳的方法过度影响。这种方法兼顾了计算效率和适应不同数据结构的灵活性。

3.2. 基于 *SHAP* 的集成优化：阈值设定

在集成加权方案中引入阈值机制，旨在提升预测模型的稳定性和效率。尽管 SHAP 框架能够根据每个模型的边际贡献有效分配权重，但它本身无法区分对集成有实质贡献的模型与贡献微不足道甚至有害的模型。因此，未经筛选便纳入所有模型，可能导致不必要的复杂性、方差增加以及整体预测可靠性下降。

使用阈值参数可以有选择地排除贡献微乎其微的模型，从而增强集成的稳健性。这一筛选过程可防止方差较高但预测能力较低的模型对最终 forecast 产生过度影响，从而降低过拟合风险。此外，通过减少活跃模型的数量，设置阈值能够提高集成的计算效率，进而有助于扩展至高维预测问题。

从理论上讲，应用阈值可以进一步优化偏差与方差之间的权衡。虽然排除贡献较低的模型可能会略微增加偏差，但预计这将通过显著降低方差，最终改善均方误差（MSE）。这在死亡率预测中尤为重要，因为噪声较大的模型可能会导致长期预测不稳定。在不同预测期限内动态应用阈值机制，可确保加权方案既易于解释，又能适应不同的数据结构。这一观点与最优预测组合相关文献以及关于存在不稳定性时收缩和加权的实践见解一致（Elliott 和 Timmermann，2004年，2005年；Timmermann，2006年；Smith 和 Wallis，2009年；Genre 等人，2013）。

设来自 N 个死亡率模型 $(M_1, \ldots, M_N)$ 的 h年期死亡率预测值由 $(\widehat{m}^{(1)}_{x,g,t+h}, \ldots, \widehat{m}^{(N)}_{x,g,t+h})$ 给出，对应于时间 $t+h$的年龄 x 和性别 g。SHAP 加权集成预测构造如下：

$$\widehat{m}^{(c)}_{\text{Shap}}(x, g, t+h) = \sum_{i=1}^{N} \widehat{\omega}_{i,\text{Shap}} \widehat{m}^{(i)}_{x,g,t+h}, \tag{1}$$

其中 $\widehat{\omega}_{i,\mathrm{Shap}}$ 表示第 i^{th} 个模型的 SHAP 导出权重，并满足 s:

$$\widehat{\omega}_{i,\mathrm{Shap}} = \frac{\widetilde{\phi}_i(\nu; g)}{\sum_{i=1}^{N} \widetilde{\phi}_i(\nu; g)}. \tag{2}$$

其中 $\widetilde{\phi}_i(\nu;\ g)$ 表示分配给该 i^{th} forecasting method 的归一化平均 SHAP 值，确保模型根据其对预测准确度的估计贡献进行加权。

为减轻贡献可忽略的模型所带来的影响，我们引入一种截断机制，排除平均 SHAP 值低于预设阈值 α 的模型。令$\sim S(\alpha) = \{i :\ \phi_i(\nu;\ g) > \alpha\}$为选定集合，且 $N^* = |S(\alpha)|$。截断后的 SHAP 权重为

$$\widehat{\omega}_{i,\mathrm{Shap}\,\alpha} = \frac{\widetilde{\phi}_i(\nu; g)\, \mathbb{1}\{i \in S(\alpha)\}}{\sum_{j \in S(\alpha)} \widetilde{\phi}_j(\nu; g)}, \tag{3}$$

其中 $\mathbb{1}\{\cdot\}$ 表示二元指示函数。带截断的 SHAP 加权集成为:

$$\widehat{m}^{(c)}_{\mathrm{Shap},\alpha}(x, g, t+h) = \sum_{i=1}^{N^*} \widehat{\omega}_{i,\mathrm{Shap}\,\alpha} \widehat{m}^{(i)}_{x,g,t+h}. \tag{4}$$

阈值 α 决定了模型对~集成做出贡献所需的最低平均 SHAP 值，其中 N^* 表示满足 $\phi_i(\nu;\ g) > \alpha$ 的模型数量。

阈值 α 的选择由数据驱动。对于每个国家、性别和预测期 h，我们在验证样本上评估一组较小的候选值（例如，$\{0.05、0.10、0.15、0.20、0.50\}$）；对于每个 α，我们计算验证均方误差并选取使其最小的值。

截断水平 α决定集成的偏差—方差权衡。较小的 α （弱过滤）保留多样性，并通常稳定短期预测方差；较大的 α （更强过滤）优先选择条件信号更强的模型，并可在更长的预测步长上降低偏差，代价是多样性减少。从这个意义上说，阈值化是一种收缩：适度的截断可稳定结果，而激进的截断会提高偏差，并且——如果它剔除了相互抵消的模型——甚至可能增加方差。

3.3. 偏差-方差权衡

与单个模型相比，集成方法能更有效地权衡偏差与方差，从而提高预测准确度。在平方误差损失下，性能加权组合

在权重估计可靠时，可以逼近回归基准；而在有限样本中，估计权重的变异性可能抵消条件增益，这解释了为何简单平均在短期预测区间或不稳定情形下仍具竞争力（Elliott 和 Timmermann，2004年；Smith 和 Wallis，2009年；Genre 等人，2013年）。当模型的相对性能随时间变化时，自适应加权或状态感知加权可能更具优势（Elliott 和 Timmermann，2005年）；而实用指南和收缩方法则见 Timmermann （2006年）。

本文采用基于SHAP的回归集成，用于量化各模型对组合预测的边际贡献，并通过阈值调节多样性和稳定性（见第 3.2 节）。较小的阈值能够保留多样性，并倾向于稳定短期预测方差；较大的阈值则强调强条件信号，可能以牺牲多样性为代价减少长期预测偏差，这与组合文献中讨论的收缩原则一致。

为求完整，形式化定义和有限样本偏差—方差分解见 附录B （参见命题 1 和备注 1）。

4. 实现细节

4.1. 经合组织国家的死亡率数据

我们考察了从人类死亡率数据库（人类死亡率数据库，2025年）获取的24个经合组织国家的分年龄死亡率：奥地利、比利时、捷克、丹麦、爱沙尼亚、芬兰、法国、匈牙利、冰岛、爱尔兰、意大利、日本、拉脱维亚、立陶宛、卢森堡、荷兰、新西兰、挪威、波兰、西班牙、瑞典、瑞士、英国和美国。由于数据可用性限制，我们选取各国女性和男性1960年至2019年的死亡率观测值。年龄方面，我们考虑了从0岁到100岁的单岁死亡率。尽管各国可用数据有所不同，我们仍对分析进行标准化，以确保这一时期内的一致性。具体而言，我们将60年的数据划分为训练集（1960—1999年）、验证集（2000—2009年）和测试集（2010—2019年）。

4.2. 死亡率集成方法

结合多个模型的预测，是一种不同于依赖单一预测技术的替代方案。个体死亡率模型采用第3节所述的方法进行组合。这些方法通过实证方式整合不同模型的预测，

从而降低与模型选择相关的不确定性，并提高点预测准确性（例如，参见Bates and Granger, 1969年）。

*4.2.1. SHAP*加权方法

估计每个模型后，我们会汇总各项预测，得出最终的组合预测。我们采用SHAP方法，将表A.1- A.4所列的15种方法生成的多项预测进行组合。具体而言，对于每个年龄 x 和每个预测期限 h，我们计算SHAP值以衡量每个模型的边际贡献，然后将其组合成加权预测。

令 $x = x_1, \ldots, x_n$ 表示给定性别 g对应的所有整数年龄。对于每个年龄 x:

1) 计算基础预测：将15个死亡率模型分别拟合到训练集，并生成预测期为 $h = 1, 2, \ldots, 10$的预测。将第 i^{th} 种方法对年龄 x、性别 g在年份 t 的forecast记为 $\hat{m}_x^{(i)}{}_{,\ g,\ t}$。将所有生成的预测（涵盖 $i = 1, \ldots, 15$种方法和 $h = 1, \ldots, 10$个预测期）汇总为一组特征，在每个年龄 x处均有15个特征。

2) 计算 SHAP 值：对于每种方法 i、每个年龄 x和预测期 h，计算其 SHAP值 $\phi_i(v, x, g, h)$，该值量化了模型 i的 forecast 相较于其他模型 forecast 的“重要性”或边际贡献。

3) 计算平均 SHAP 值：设 $\widetilde{\phi}_i(v;\ g)$ 为与 i^{th}预测方法相关的平均SHAP值，并自动将其归一化到范围 $[0, 1]$。

4) 通过计算能够最优量化各模型边际贡献的基于SHAP的权重，构建 SHAP 集成 forecast，具体依据为（1）。

5) 通过引入一种机制提高计算效率，构建带截断的 SHAP-组合 forecast，具体依据为（4）。

4.2.2. 基准加权方法

为评估我们新提出的 SHAP 集成在死亡率预测方面的准确性，我们将其与以下加权方案进行比较:

- 简单模型平均：我们对每个死亡率模型得到的单项预测应用$\widehat{\text{平均值}}_i(v$的等权重；$g) = 1/15$

- 赤池模型平均：我们针对各方法在验证集上产生的样本外预测计算赤池信息准则。随后计算 AIC 权重为

$$\widehat{\omega}_{i,\mathrm{AIC}} = \frac{\widehat{|\mathrm{AIC}_i(v;g)|}}{\sum_{i=1}^{\bar{N}} \widehat{|\mathrm{AIC}_i(v;g)|}} \mathbb{1}\left\{\widehat{\mathrm{AIC}}_i(v;g) < 0\right\},$$

其中 $\widehat{\text{赤池信息准则}}_i(v;\ g)$ 表示针对 i^{th} 方法，在验证集上估算的赤池信息准则，$\bar{N}$ 表示 $\widehat{\text{赤池信息准则估计值为负的预测方法数量}}_i(v;\ g)$。

使用这些基准权重，我们参照（1）组合各单项预测。

5. 点预测准确性比较分析

在使用验证集选择权重后，我们使用测试集通过均方误差和平均绝对误差（MAE）评估模型性能。具体而言，为评估点预测准确性，我们计算采用 SHAP 加权权重的集成的均方误差和平均绝对误差：

$$\mathrm{MSE}_{\mathrm{Shap}} = \frac{1}{101 \times (11-h)} \sum_{\xi=h}^{10} \sum_{x=0}^{100} \left[y_{x,g,2009+\xi} - \widehat{m}^{(c)}_{\mathrm{Shap}}(x,g,2009+\xi)\right]^2,$$

$$\mathrm{MAE}_{\mathrm{Shap}} = \frac{1}{101 \times (11-h)} \sum_{\xi=h}^{10} \sum_{x=0}^{100} \left|y_{x,g,2009+\xi} - \widehat{m}^{(c)}_{\mathrm{Shap}}(x,g,2009+\xi)\right|,$$

其中 $y_{x,\ g,\ 2009+\xi}$ 表示 $2009+\xi$年的年龄 x、性别为 g 对应的死亡率对数　　.

该分析同时关注男性和女性人口，并根据提前一年至提前10年的均方误差和平均绝对误差，展示样本外死亡率预测准确性的提升百分比。为简洁起见，在表1中，我们仅报告 $h=1$、 $h=6$和 $h=10$年前预测的经合组织国家死亡率结果，这些结果通过均方误差的平均值（$\times 100$）进行衡量。结果分别展示了未进行截断以及进行50%截断的SHAP集成。为评估经过截断和未经过截断的SHAP集成相较于现有组合方法的竞争力，我们将其与SMA和赤池信息准则进行比较。

表 1：经济合作与发展组织死亡率的 MSE forecast 准确性，采用集成方法 Average、赤池信息准则、SHAP 和 SHAP $\alpha = 50\%$，分别进行一步（h = 1）、六步（h = 6）和十步（h = 10）ahead forecast

	SMA		AIC		SHAP		SHAP $\alpha = 50\%$	
国家	F	M	F	M	F	M	F	M
h = 1								
奥地利	0.0049	**0.0137**	**0.0044**	0.0139	0.0048	0.0138	0.0052	0.0160
比利时	0.0053	**0.0223**	0.0059	0.0233	**0.0041**	0.0254	0.0055	0.0484
捷克	0.0112	0.0573	**0.0086**	**0.0517**	0.0092	0.0647	0.0102	0.0755
丹麦	0.0046	0.0237	0.0037	0.0240	**0.0033**	**0.0231**	0.0045	0.0243
爱沙尼亚	**0.0353**	0.1364	0.0354	**0.1347**	0.0360	0.1446	0.0380	0.1759
芬兰	0.0112	**0.0532**	0.0108	0.0535	**0.0104**	0.0551	0.0108	0.0544
法国	**0.0162**	0.0159	0.0168	**0.0159**	0.0230	0.0217	0.0274	0.0260
匈牙利	0.0092	0.0515	**0.0086**	**0.0422**	0.0100	0.0440	0.0114	0.0530
冰岛	0.0887	**0.2617**	0.0874	0.2640	**0.0867**	0.2628	0.0877	0.2727
爱尔兰	**0.0111**	0.0615	0.0115	0.0598	0.0131	**0.0586**	0.0149	0.0621
意大利	0.0035	0.0050	0.0032	0.0047	**0.0029**	0.0046	0.0035	**0.0042**
日本	0.0012	0.0024	0.0012	**0.0018**	**0.0010**	0.0027	0.0016	0.0033
拉脱维亚	0.0271	0.1028	**0.0249**	0.1036	0.0320	0.1032	0.0317	**0.0920**
立陶宛	0.0213	0.0808	0.0230	0.0765	**0.0194**	0.0706	0.0250	**0.0625**
卢森堡	0.0962	**0.3146**	0.0979	0.3180	0.0978	0.3202	**0.0874**	0.3229
荷兰	0.0038	**0.0074**	0.0037	0.0077	**0.0032**	0.0089	0.0039	0.0099
新西兰	0.0182	0.0361	0.0162	0.0353	0.0102	0.0355	**0.0085**	**0.0335**
挪威	0.0073	**0.0264**	0.0086	0.0268	**0.0065**	0.0275	**0.0065**	0.0307
波兰	**0.0033**	0.0121	0.0051	0.0117	0.0057	**0.0114**	0.0074	0.0128
西班牙	0.0033	0.0050	**0.0027**	0.0059	0.0030	0.0046	0.0036	**0.0040**
瑞典	0.0032	**0.0104**	**0.0029**	0.0108	0.0034	0.0106	0.0088	0.0144
瑞士	0.0049	**0.0169**	0.0044	0.0171	**0.0037**	0.0171	0.0038	0.0293
UK	0.0031	0.0052	0.0026	0.0047	**0.0023**	**0.0038**	0.0025	0.0049
USA	0.0031	0.0029	0.0020	0.0026	0.0019	**0.0016**	**0.0019**	0.0019
Mean	0.0166	0.0552	**0.0163**	**0.0546**	0.0164	0.0557	0.0172	0.0598
h = 6								
奥地利	0.0106	0.0368	**0.0073**	**0.0160**	0.0134	0.0222	0.0250	0.0169
比利时	0.0098	0.0364	0.0163	**0.0274**	0.0059	0.0464	**0.0041**	0.0783
捷克	0.0199	0.0528	**0.0071**	**0.0393**	0.0180	0.0420	0.0201	0.0519
丹麦	0.0046	0.0375	**0.0032**	**0.0347**	0.0089	0.0401	0.0303	0.0529
爱沙尼亚	0.1200	0.1521	0.1208	**0.1443**	**0.0779**	0.1673	0.0863	0.2216

表 1: （续）

	SMA		AIC		SHAP		SHAP $\alpha = 50\%$	
国家	F	M	F	M	F	M	F	M
芬兰	0.0085	0.0647	0.0117	**0.0559**	**0.0079**	0.0631	0.0099	0.0569
法国	0.0218	0.0316	**0.0086**	**0.0226**	0.0325	0.0318	0.0323	0.0424
匈牙利	0.0360	0.0892	**0.0143**	**0.0441**	0.0229	0.0712	0.0293	0.1062
冰岛	0.0721	**0.2996**	0.0662	0.3088	**0.0638**	0.3310	0.0696	0.3578
爱尔兰	0.0288	0.1311	**0.0177**	0.0866	0.0242	0.0844	0.0414	**0.0780**
意大利	0.0084	0.0225	**0.0043**	**0.0071**	0.0047	0.0184	0.0047	0.0148
日本	**0.0022**	0.0090	0.0061	**0.0055**	0.0025	0.0126	0.0059	0.0138
拉脱维亚	0.0232	0.0939	**0.0196**	0.0911	0.0236	0.0883	0.0407	**0.0816**
立陶宛	0.1045	**0.0838**	0.1095	0.1220	**0.0716**	0.0959	0.0844	0.0998
卢森堡	0.1094	**0.2107**	**0.0924**	0.2127	0.0959	0.2252	0.0939	0.2713
荷兰	0.0117	0.0336	0.0086	**0.0143**	**0.0079**	0.0326	0.0126	0.0457
新西兰	0.0316	0.0476	0.0393	**0.0369**	0.0199	0.0411	**0.0172**	0.0377
挪威	0.0165	0.0298	0.0168	**0.0213**	0.0091	0.0237	**0.0073**	0.0292
波兰	0.0350	0.0956	**0.0137**	**0.0184**	0.0258	0.0373	0.0180	0.0227
西班牙	0.0099	0.0219	0.0046	0.0099	0.0058	0.0103	**0.0039**	**0.0043**
瑞典	**0.0043**	0.0153	0.0103	**0.0098**	0.0083	0.0133	0.0207	0.0204
瑞士	0.0078	0.0420	0.0143	**0.0145**	0.0073	0.0286	**0.0038**	0.0492
UK	0.0054	0.0215	0.0073	**0.0070**	**0.0037**	0.0141	0.0129	0.0192
USA	0.0047	0.0204	**0.0025**	**0.0024**	0.0029	0.0043	0.0026	0.0031
Mean	0.0295	0.0700	0.0259	**0.0564**	**0.0235**	0.0644	0.0282	0.0740
$h = 10$								
奥地利	0.0424	0.0415	0.0184	**0.0201**	0.0133	0.0286	**0.0109**	0.0478
比利时	0.0268	0.0559	0.0244	0.0393	0.0087	0.0325	**0.0026**	**0.0290**
捷克	0.0394	0.1398	**0.0056**	**0.0570**	0.0273	0.0968	0.0282	0.0932
丹麦	0.0189	0.0559	**0.0071**	**0.0520**	0.0150	0.0693	0.0342	0.1355
爱沙尼亚	0.0559	**0.2116**	**0.0305**	0.2166	0.0626	0.2683	0.0615	0.4564
芬兰	0.0241	0.1014	0.0179	0.0819	**0.0133**	0.1000	0.0153	**0.0697**
法国	0.0141	0.0302	**0.0030**	**0.0166**	0.0233	0.0282	0.0343	0.0437
匈牙利	0.0337	0.0460	**0.0146**	**0.0264**	0.0225	0.0355	0.0356	0.0656
冰岛	0.0541	0.6384	**0.0442**	**0.5906**	0.0605	0.6538	0.0856	0.7037
爱尔兰	0.0422	0.0983	**0.0207**	0.0987	0.0505	0.0821	0.1392	**0.0801**
意大利	0.0198	0.0604	0.0119	**0.0140**	**0.0068**	0.0291	0.0085	0.0202
日本	0.0122	0.0357	0.0082	**0.0048**	0.0028	0.0147	**0.0022**	0.0068

表 1：（续）

	SMA		AIC		SHAP		SHAP $\alpha = 50\%$	
国家	F	M	F	M	F	M	F	M
拉脱维亚	0.0281	0.2381	0.0278	0.1573	**0.0231**	0.1756	0.0325	**0.1208**
立陶宛	0.2487	0.2680	0.2108	0.2578	0.1455	0.2269	**0.1140**	**0.1876**
卢森堡	0.1436	0.2221	0.0858	**0.0897**	0.0995	0.1129	**0.0803**	0.1069
荷兰	0.0252	0.0380	0.0171	**0.0212**	**0.0143**	0.0724	0.0225	0.1638
新西兰	0.0546	0.1119	0.0584	**0.0366**	**0.0316**	0.0753	0.0382	0.0492
挪威	0.0480	0.0498	0.0417	**0.0218**	0.0235	0.0374	**0.0101**	0.0773
波兰	0.0436	0.0858	**0.0231**	**0.0293**	0.0277	0.0455	0.0312	0.0359
西班牙	0.0215	0.0403	**0.0056**	0.0078	0.0114	0.0061	0.0061	**0.0056**
瑞典	0.0287	0.0445	0.0220	**0.0232**	**0.0203**	0.0385	0.0310	0.0557
瑞士	0.0285	0.0165	0.0314	**0.0087**	0.0155	0.0134	**0.0064**	0.0351
UK	0.0207	0.0408	**0.0076**	**0.0075**	0.0096	0.0159	0.0186	0.0175
USA	0.0158	0.0710	0.0072	**0.0028**	**0.0058**	0.0113	0.0086	0.0039
Mean	0.0455	0.1142	0.0310	**0.0784**	**0.0306**	0.0946	0.0357	0.1088

5.1. SHAP 与 SMA 和赤池信息准则基准的比较

在考察 24 个国家以及预测期限 $h = 1$、……、10 内男性和女性的均方误差与平均绝对误差结果时，我们首先关注基于 SHAP 的集成，以及截尾至 50% 的基于 SHAP 的集成。在表 1 中，我们根据均方误差将二者与 SMA 和赤池信息准则进行比较。作为说明，表 1 报告了使用四组组合权重得到的提前 1 年、6 年和 10 年的年龄别死亡率预测的均方误差 $\times 100$。粗体数字突出显示了所评估集成方法中最低的均方误差。下文将按预测期限、性别和国家总结主要规律。

在较短的预测期（$h \approx 1$–3）内，基于 SHAP 的集成在 24 个女性国家案例中的 16 个（67%）以及 24 个男性案例中的 10 个（42%）优于 SMA。相较于赤池信息准则，优势略弱（分别为 58 % 和 50%）。截断版本仅在 10 个女性国家和 7 个男性国家中表现最佳，这表明在最短预测期内，大幅削减模型的收益有限。丹麦（男性）、芬兰（女性）和奥地利（女性）是例外，在这些案例中 SMA 或赤池信息准则仍略占优势。

预测期长度（h ≈ 4–6）进一步凸显了 SHAP 权重的优势。与 SMA 相比，SHAP 权重在 18/24（75%）的女性案例和 16/24（67%）的男性案例中取得了更低的均方误差。与赤池信息准则相比，结果呈现不对称性：在 12 个国家中，SHAP 在女性案例中胜出，而在男性案例中仅在 3 个国家胜出。对于两种性别，基于 SHAP 的集成在约三分之二的国家中优于截断变体；在西班牙和挪威等国家，截断变体仍有用，因为其额外的方差缩减抵消了偏差的小幅增加。

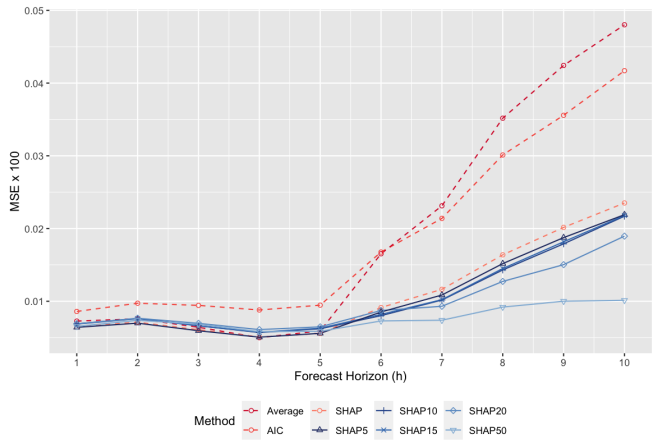
在最长预测期（h ≈ 7–10），基于 SHAP 的集成几乎全面优于 SMA（女性为 20/24，男性为 20/24 ）。对女性而言，其相较于赤池信息准则的优势有所扩大（ 13/24 ），但对男性而言则大幅缩小（4/24），这反映出赤池信息准则在男性死亡率长期预测中偶尔表现强劲（例如法国、丹麦和冰岛）。截尾集成如今很少成为最佳选择。不过，它在相当一部分情况下仍能取得最低均方误差（16 个国家的女性和 14 个国家的男性），比利时和瑞士尤为明显。

就国家间的异质性而言，在比利时、意大利、挪威、西班牙、瑞典和美国，改进最为稳定；在这些国家，无论预测期和性别如何，SHAP（标准型或截尾型）的表现都优于两种平均方案。相反，在死亡率波动较大或结构性变化较剧烈的国家（爱沙尼亚、冰岛、立陶宛），四种组合的相对排名波动明显，凸显了自适应加权的重要性。

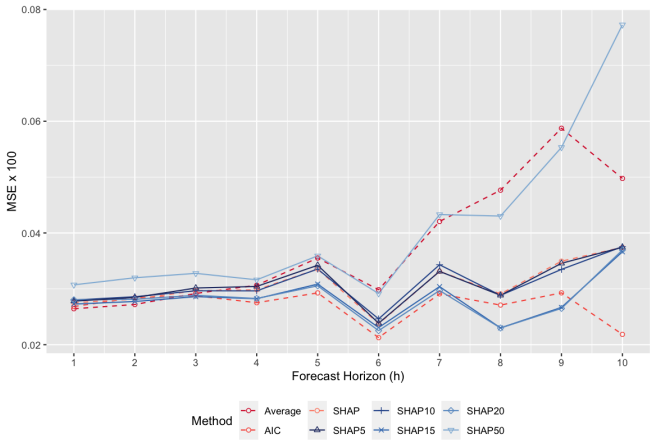
与Kessy等人（2022年）的研究类似，我们的结果表明，SHAP 集成准则在选择最优死亡率模型时，能够有效考虑预测期的影响。基于 SHAP 的集成相较于等权重，能够系统性地提高准确性，而且随着预测期延长，收益也会增加。相比之下，赤池信息准则和 SMA 并不针对特定预测期，因此在较长预测期限内选择死亡率模型时效果较差。此外，尽管减少贡献较低的模型可能有所帮助，尤其是在极短预测期内，但在所考察的不同预测期、性别以及大多数经合组织国家中，基于 SHAP 的集成是最稳健的配置。

图1 展示了挪威、西班牙和美国的所有 SHAP 集成预测，证实了基于 SHAP 的策略在不同实现方式下具有一致性和稳健性。其他经合组织国家也呈现出类似结果，平均绝对误差指标亦是如此；相关结果见 Shiny 应用，网址为 https://github.com/YangANU/Ensemble_死亡率_模型。

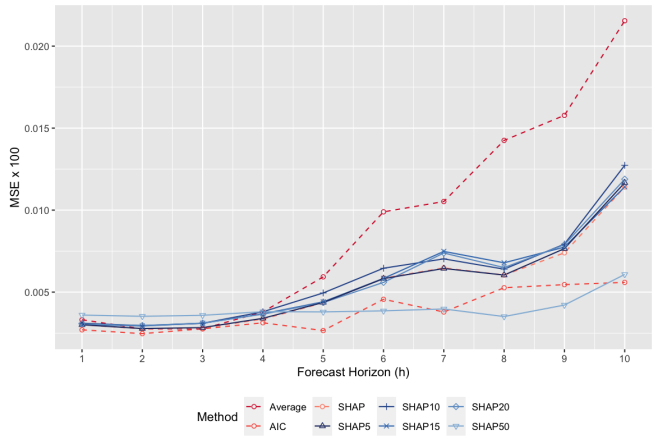
我们还进行 Diebold-Mariano检验，以评估 SHAP、SMA 的点预测准确性，



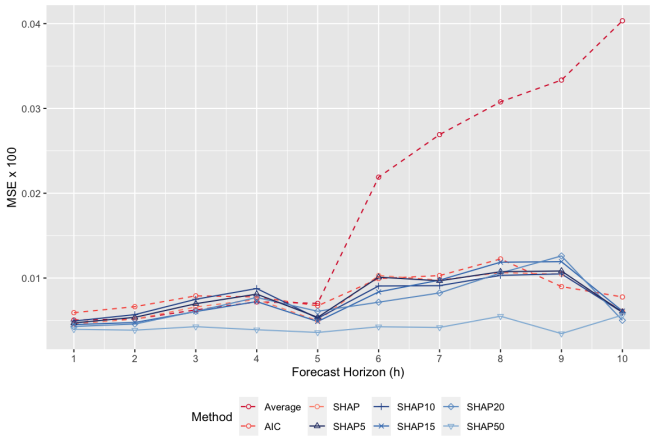
(a) 挪威女性



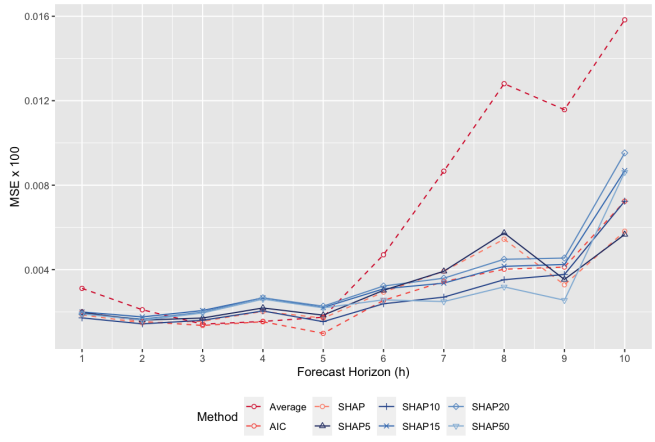
(b) 挪威男性



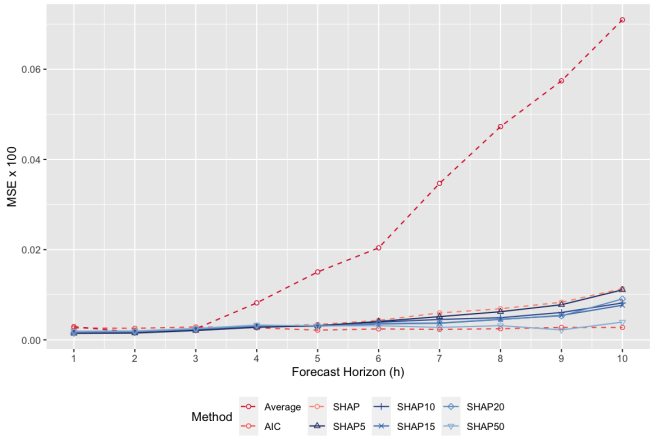
(c) 西班牙女性



(d) 西班牙男性



(e) 美国女性



(f) 美国男性

图1：挪威、西班牙和美国女性与男性死亡率数据在不同预测步长下的 SHAP集成均方误差。

以及 AIC 集成。对于每个整数年龄 x 介于 0 到 100 之间的情况，将样本外点预测误差汇总为一个向量 $e_{\text{ensemble}} = y_{x,\ g,\ 2009+\xi} - \hat{m}^{(c)}_{\text{集成}}(x, g, 2009+\xi)$，其中 $h = 1, \ldots..10$，$\xi = h, \ldots..,$ 10，并且集成 $\in$ {SHAP、SMA、AIC}。随后，使用 'forecast' 软件包（Hyndman等人，2025年）中的 'dm.test' 函数，并采用平方损失，检验以下假设：

H_0 : 两个集成具有相同的预测准确度。

H_A : SHAP 集成比基准更准确。

采用常用的 5% 显著性水平，我们统计经合组织国家 101 个整数年龄中 p-值小于 0.05 的比例，并在图 2 中展示结果。

热图的前两列显示了在两个性别中，SHAP 集成在经合组织国家的哪些年龄范围内优于SMA方法。总体而言，SHAP方法带来了广泛且稳健的改进：在大多数国家，至少对一个性别而言，它在超过一半的年龄段上提供了更准确的预测。最后两列报告了 SHAP 集成优于 AIC集成的年龄段占比，表明 SHAP 也能在基于 AIC 的加权之外带来改进。这些提升在波罗的海国家（爱沙尼亚、立陶宛和拉脱维亚）尤为显著。

5.2. 年龄分层均方误差

死亡率模式的异质性会影响所考察集成的预测性能。例如，已有研究证实，平坦模型适用于日本死亡率数据（Shang andHaberman，2018）；而函数型时间序列（即 FDM、稳健 FDM 和乘积比）模型则适用于意大利的死亡率（Bimonte等人，2025年）。为突出 SHAP 集成表现较好或较差的年龄段，我们对日本和意大利不同集成的年龄别死亡率预测进行了标准化，并在图 3中以热图展示结果。其他经合组织国家的年龄分层均方误差热图可在 GitHub 页面查看。

HMD 年龄组（即 0、1-4、5-9、10-14、.....、91-94、95-99、100）（人类死亡率数据库，2025年）用于绘制年龄分层均方误差图。对于任一年龄组，所获得的最大均方误差

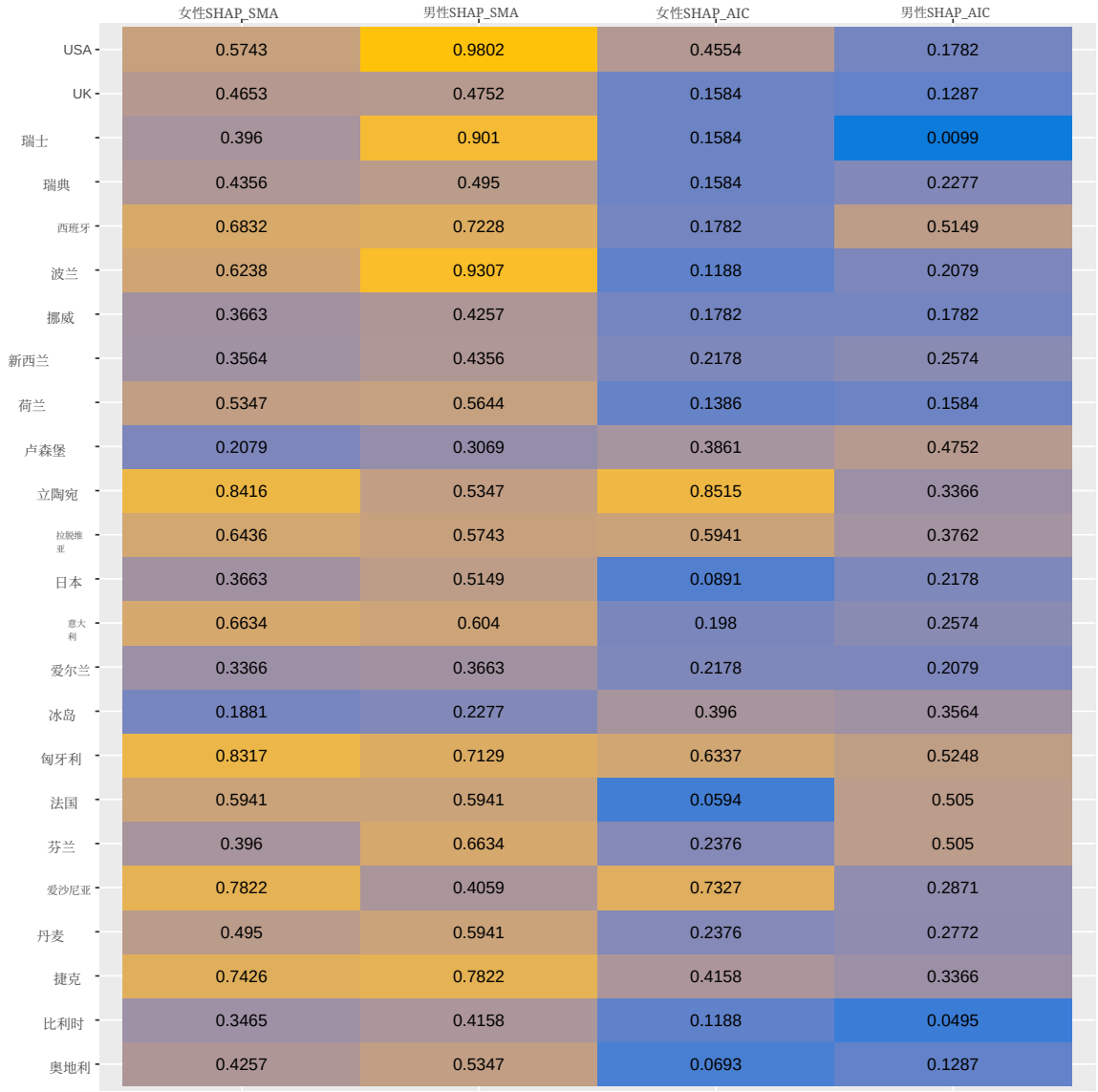


图2：在经合组织国家中，基于 Diebold-Mariano检验评估的 0 至 100 岁整数年龄范围内，SHAP 集成优于 SMA 和 AIC 方法的比例。蓝色数值小于黄色数值。

八种集成方法之一的均方误差将被标准化为 1，并以黄色显示。同样，每个年龄的最小均方误差以蓝色标示。比较图 3 中的“黄-蓝”分布模式可见，对于日本数据，SMA 和 AIC 集成在为大多数 60 岁以下人群做出准确预测方面表现不佳。阈值参数超过 10% 的 SHAP集成在这些年龄段表现良好。SMA方法能够为 80 多岁的日本老年女性以及 75 至 90 岁的男性提供准确预测。相比之下，基于 SHAP 的集成方法与基准集成方法在意大利死亡率上呈现出略有不同的模式。SMA

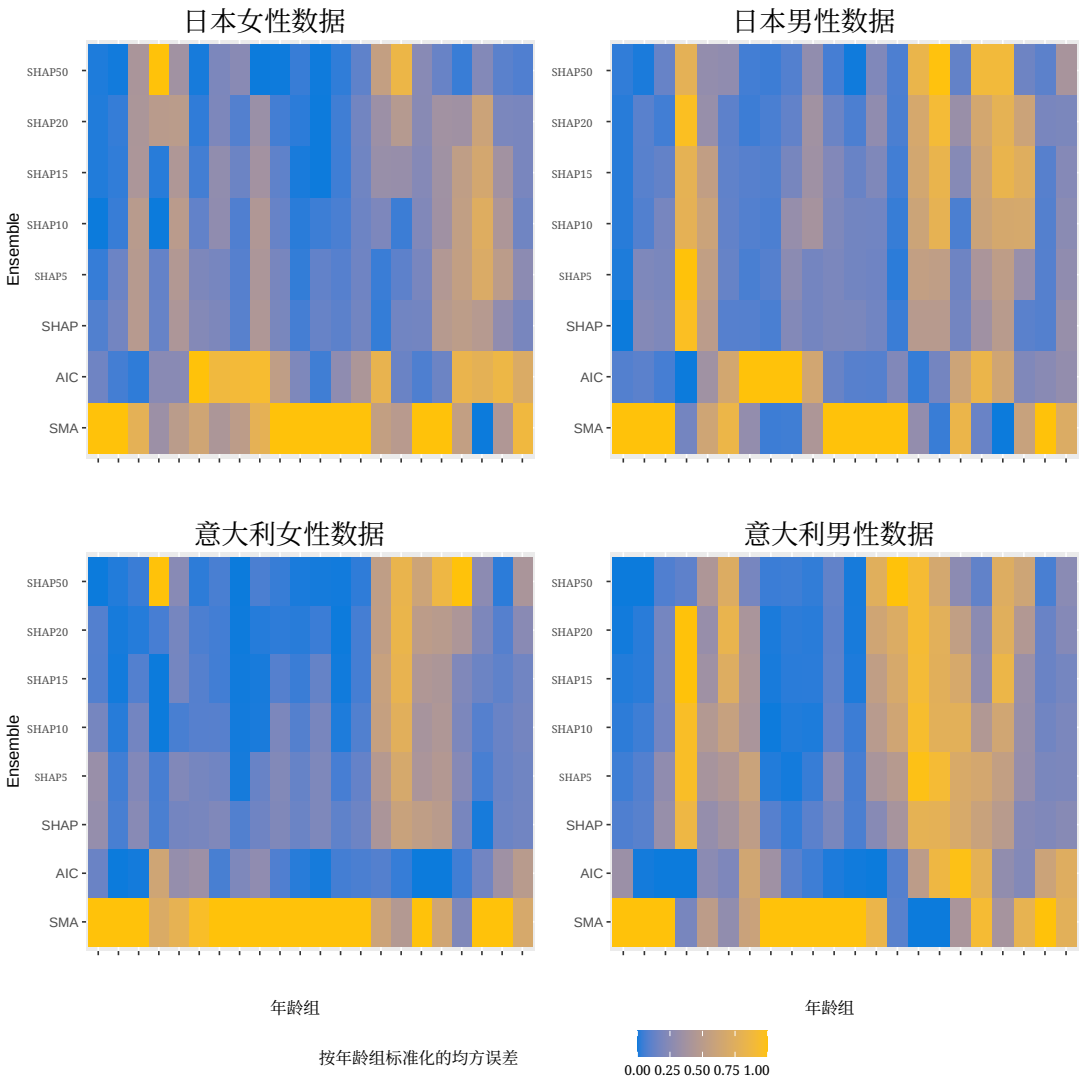


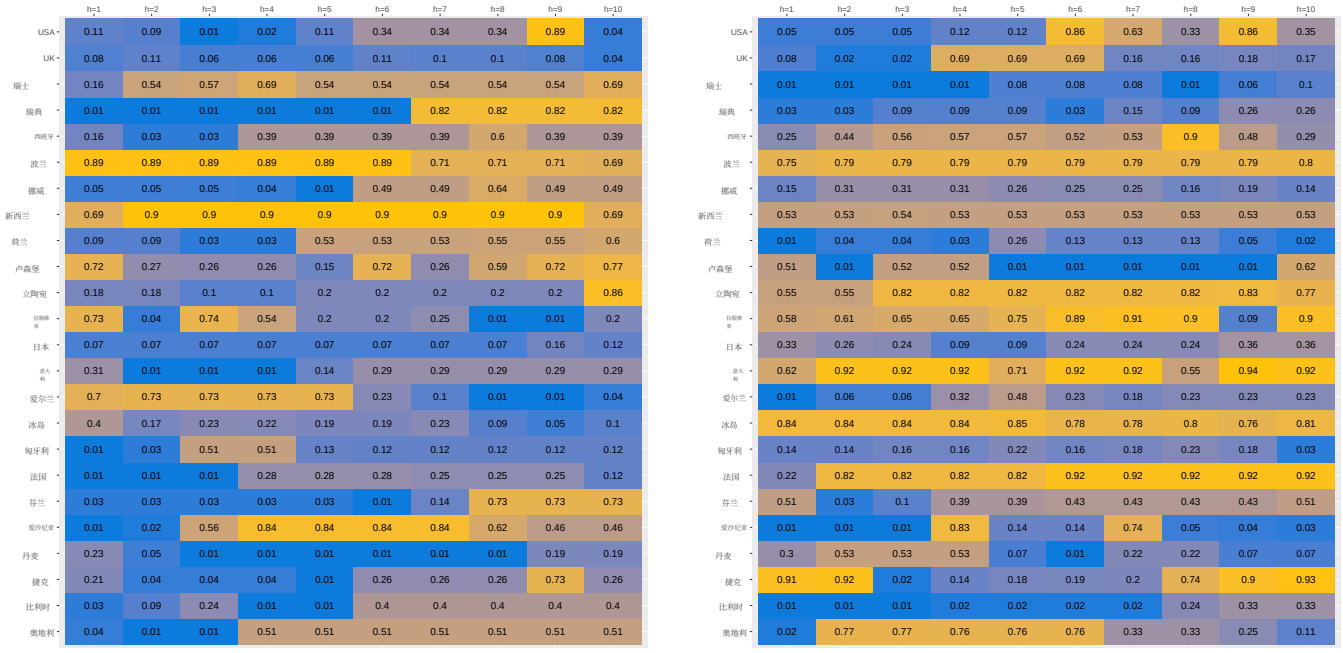
图 3：日本和意大利女性及男性死亡率的年龄分层均方误差（按年龄组标准化）。

集成在意大利女性的大多数年龄段以及意大利男性的 0–20、27–55 和 80–83 岁年龄段表现不佳。AIC集成在 0–20 岁年龄段对两种性别的预测效果更佳，在女性 60–80 岁年龄段以及男性 45–55 岁年龄段也表现更优。在意大利，带有$\alpha \geqslant 5\%$ 的 SHAP集成在 28–45 岁女性和男性中表现更好。

5.3. 敏感性分析

为评估截断SHAP集成的敏感性，我们针对24个经合组织国家，考察了常用参数 $\alpha \in 0.05$、0.1、0.15、0.2、0.5这四个取值，以及通过网格搜索选出的参数。网格搜索法旨在为每个国家确定最优的 α ，其

点预测准确度。具体而言，我们对$\alpha \in \{0.01、0.02、\ldots、0.99\}$逐一重复计算，并评估所得的 SHAP 集成。图4 展示了 α的最优 $h = 1$ 从 $h = 10$。



a - 女性　　b - 男性

图4：网格搜索法为各国家选出的、具有最佳点预测准确度的最优截断参数 α，适用于从 $h = 1$ 到 $h = 10$ 的预测期。

可以看出，预测期限越长，通常所需的截断参数也越大。这是因为对遥远未来的预测包含更多不确定性。此外，男性的 α 通常高于女性，因为男性死亡率波动更大，方差也更高。

我们分别对男性和女性重复计算，并在表 $h = 1$中呈现2的预测均方误差（MSE）。所有预测步长（$h = 1$至 10）的完整结果（包括平均绝对误差指标）可在 GitHub 上的 Shiny 应用中查看。我们发现，在较短的预测期（$h \approx 1$–3）下，将 SHAP值的阈值 $\alpha = 0$设为 .05 时，24 个女性序列中有 16 个、24 个男性序列中有 14 个取得最低均方误差。换言之，保留几乎所有模型，同时剔除贡献最弱的模型，可带来最大的短期收益；这种做法不仅优于不设阈值的 SHAP 集成，也优于 SMA、赤池信息准则和阈值为 50% 的 SHAP。

表2：经济合作与发展组织死亡率的一步超前（$h = 1$）预测准确度，适用于由网格搜索法选取 α且 $\alpha \in \{0.05、0.1、0.15、0.2、0.5\}$的SHAP 集成。

	网格搜索		**SHAP** $\alpha = 5\%$		**SHAP** $\alpha = 10\%$	
国家	F	M	F	M	F	M
奥地利	**0.0047**	**0.0138**	0.0049	0.0139	0.0051	0.0142
比利时	**0.0041**	**0.0254**	0.0042	0.0256	0.0044	0.0265
捷克	**0.0090**	**0.0511**	0.0091	0.0650	0.0092	0.0651
丹麦	**0.0032**	**0.0228**	0.0033	0.0231	0.0033	0.0233
爱沙尼亚	**0.0360**	**0.1446**	0.0361	0.1451	0.0362	0.1478
芬兰	**0.0104**	**0.0542**	0.0104	0.0551	0.0105	0.0553
法国	**0.0230**	**0.0207**	0.0236	0.0224	0.0236	0.0226
匈牙利	**0.0100**	**0.0416**	0.0100	0.0442	0.0104	0.0440
冰岛	**0.0863**	**0.2523**	0.0867	0.2628	0.0868	0.2631
爱尔兰	**0.0119**	**0.0586**	0.0132	0.0588	0.0134	0.0592
意大利	**0.0027**	**0.0037**	0.0029	0.0046	0.0029	0.0046
日本	**0.0009**	**0.0022**	0.0010	0.0028	0.0009	0.0026
拉脱维亚	**0.0238**	**0.0906**	0.0322	0.1036	0.0340	0.1014
立陶宛	**0.0187**	**0.0612**	0.0193	0.0704	0.0191	0.0702
卢森堡	**0.0855**	**0.3199**	0.0975	0.3235	0.1012	0.3246
荷兰	**0.0032**	**0.0089**	0.0032	0.0090	0.0032	0.0091
新西兰	**0.0073**	**0.0330**	0.0099	0.0355	0.0095	0.0355
挪威	**0.0064**	**0.0273**	0.0064	0.0278	0.0069	0.0280
波兰	**0.0045**	**0.0099**	0.0058	0.0114	0.0060	0.0118
西班牙	**0.0030**	**0.0039**	0.0030	0.0047	0.0031	0.0049
瑞典	**0.0034**	**0.0105**	0.0034	0.0105	0.0034	0.0108
瑞士	**0.0034**	**0.0171**	0.0037	0.0171	0.0037	0.0173
UK	**0.0021**	**0.0037**	0.0022	0.0040	0.0022	0.0037
USA	**0.0017**	**0.0014**	0.0020	0.0014	0.0017	0.0015
Mean	**0.0152**	**0.0533**	0.0164	0.0559	0.0167	0.0561

	SHAP $\alpha = 15\%$		**SHAP** $\alpha = 20\%$		**SHAP** $\alpha = 50\%$	
国家	F	M	F	M	F	M
奥地利	0.0053	0.0143	0.0053	0.0147	0.0052	0.0160
比利时	0.0045	0.0280	0.0048	0.0285	0.0055	0.0484
捷克	0.0090	0.0649	0.0090	0.0650	0.0102	0.0755
丹麦	0.0033	0.0240	0.0033	0.0239	0.0045	0.0243

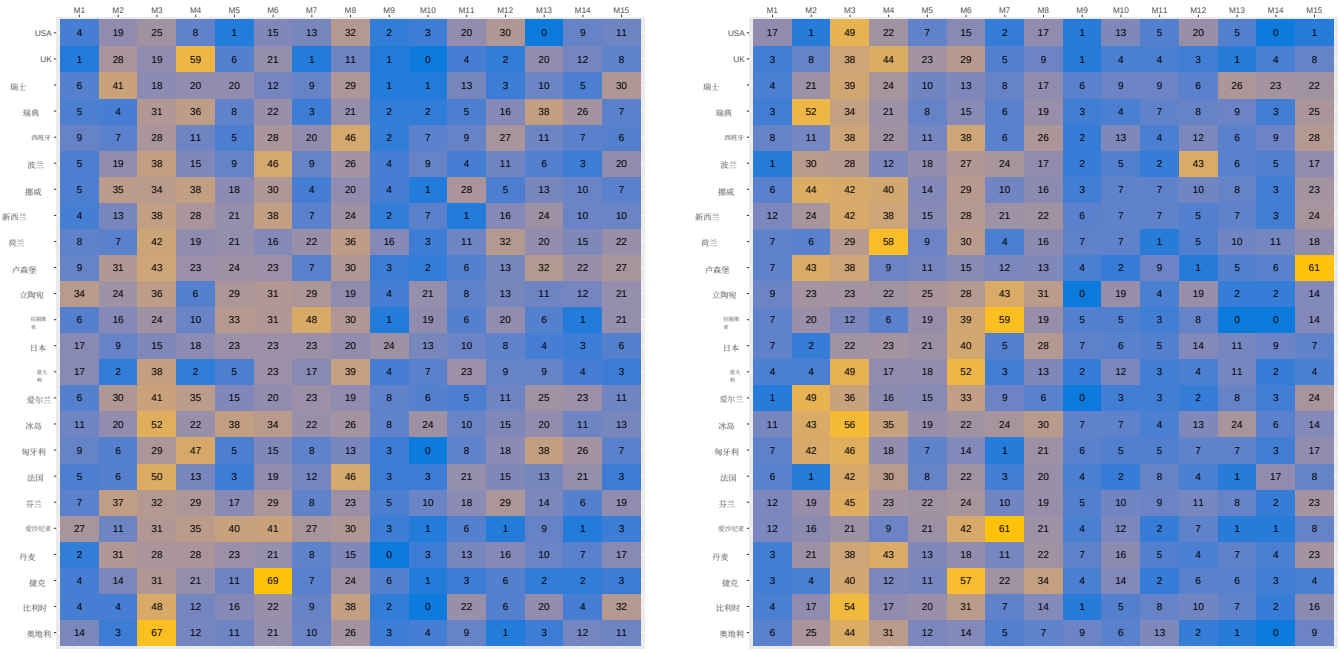
爱沙尼亚	0.0371	0.1464	0.0374	0.1475	0.0380	0.1759
芬兰	0.0105	0.0559	0.0105	0.0561	0.0108	0.0544
法国	0.0244	0.0217	0.0248	0.0207	0.0274	0.0260
匈牙利	0.0107	0.0418	0.0122	0.0428	0.0114	0.0530
冰岛	0.0866	0.2638	0.0867	0.2649	0.0877	0.2727
爱尔兰	0.0128	0.0598	0.0128	0.0618	0.0149	0.0621
意大利	0.0029	0.0046	0.0030	0.0047	0.0035	0.0042
日本	0.0009	0.0026	0.0012	0.0027	0.0016	0.0033
拉脱维亚	0.0360	0.1023	0.0363	0.1021	0.0317	0.0920
立陶宛	0.0196	0.0695	0.0195	0.0690	0.0250	0.0625
卢森堡	0.1005	0.3256	0.1009	0.3250	0.0874	0.3229
荷兰	0.0032	0.0097	0.0036	0.0100	0.0039	0.0099
新西兰	0.0088	0.0355	0.0088	0.0351	0.0085	0.0335
挪威	0.0069	0.0273	0.0069	0.0280	0.0065	0.0307
波兰	0.0061	0.0127	0.0060	0.0124	0.0074	0.0128
西班牙	0.0030	0.0045	0.0031	0.0043	0.0036	0.0040
瑞典	0.0035	0.0111	0.0035	0.0110	0.0088	0.0144
瑞士	0.0036	0.0174	0.0036	0.0176	0.0038	0.0293
UK	0.0022	0.0038	0.0022	0.0041	0.0025	0.0049
USA	0.0020	0.0016	0.0019	0.0016	0.0019	0.0019
Mean	0.0168	0.0562	0.0170	0.0564	0.0172	0.0598

然而，在较长的预测步长下（$h \geqslant 4$），截断水平为 50% 的 SHAP 集成在许多国家表现最佳，优于或接近新测试的截断 SHAP 方法和基于平均的方法。尽管不同国家和性别之间存在一些细微差异，例如在意大利和瑞士，$\alpha = 0.05$ 截断 SHAP 集成在较短预测步长下对女性表现突出。相比之下，$\alpha = 0.10$ 变体在丹麦和西班牙对男性表现更好；没有任何一个截断水平在所有情境下都占据优势。值得注意的是，网格搜索选出的最优参数相比常用的 $\alpha \in 0.05$、0.1、0.15、0.2 仅带来 marginal improvements，却以较慢的计算速度为代价。

这些发现凸显了通过采用不同的截断阈值优化 SHAP 集成方法所带来的灵活性提升和稳健的预测准确度，同时也表明，在长期预测中，截断 50% 的 SHAP 具有竞争力。

5.4. 死亡率模型选择的挑战

图5 展示了男性和女性人群中排除 SHAP值低于 50% 的方法后，各基础预测方法的热图。为简洁起见，两个热图中的数字表示每种方法在 101 次迭代中被选中的次数，也就是说，对 0 至 100 岁的每个年龄分别运行一次 SHAP50算法。如果某种基础预测方法表现良好，其出现频率将接近 101；否则则趋近于零。



a - 女性　　b - 男性

图 5：剔除 SHAP值处于最低 50% 的方法后，剩余的基础预测方法。"M1" 至 "M15" 分别代表 "lc"、"rh"、"apc"、"cbd"、"m6"、"m7"、"m8"、"plat"、"lca_dt"、"lca_dxt"、"lca_e0"、"lca_none"、"fdm"、"robust_fdm" 和 "pr" 这些基础预测方法。两个热图中的数字表示每种方法在 101 次迭代中被选中的次数；每个单元格中的最大值为 101，对应整数年龄 0-100。

对于女性人口，分析揭示了死亡率建模中显著的地理模式：各国往往在特定方法上表现出集中优势，而在其他方法上的结果则较弱或更为分散。例如，在北欧国家（瑞典、挪威、芬兰、冰岛、丹麦）中，M3 始终表现出较强的效果。然而，丹麦和挪威还分别通过伦肖-哈伯曼 M2 对其优势进行了补充，而冰岛则在 M5 和 M6 上表现出色。这五个国家的 M12 拟合效果均较差。

波罗的海国家（Est　尼亚、拉脱维亚和立陶宛）更青睐CBD的队列敏感变体

族。爱沙尼亚将其高频选择分布在M3、M4、M5、M6和M8；拉脱维亚几乎完全聚焦于M5–M8这一组；而立陶宛则将李–卡特基准模型M1与M3及M6相结合。这三个国家几乎均未赋予函数模型M13任何权重。

中欧（波兰、捷克共和国、匈牙利）明显偏好Plat扩展模型M6，其次是M3；匈牙利单独将函数模型M13纳入组合，使模型组合更加丰富。

在西欧（法国、比利时、荷兰、卢森堡），M3和M8在模型选择中占主导地位。卢森堡和比利时还赋予函数模型M13和乘积比率模型M15较高权重，而法国几乎不给M15分配权重。

在英语国家（美国、英国、爱尔兰和新西兰）中，APC模型M3普遍占据显著地位。不过，每个国家偏好的次要模型各不相同：美国提高了得分调整的李–卡特模型M12的权重，英国主要采用CBD模型M4，爱尔兰将M2与M4搭配使用，而新西兰则明显偏好Plat变体M6。

日本表现突出，在大多数方法下拟合度始终较低，这表明其具有独特特征，或这些方法并不适用于日本。

对于男性，M3是核心模型，但呈现出两种明显的模式。在北欧大部分地区（奥地利、丹麦、冰岛、荷兰、挪威、瑞典、英国），基本CBD模型M4的表现与M3相当，甚至更胜一筹。此外，队列驱动特征显著的变体M6和M7在波罗的海三国（爱沙尼亚、拉脱维亚、立陶宛）占据主导地位，并且常与M3一同出现在比利时、捷克、意大利和西班牙。卢森堡是唯一一个将M15乘积比率模型与M2和M3一同纳入领先组的国家；而波兰则将M2与按预期寿命调整的M12搭配使用。

值得强调的是，模型M4–M6属于Cairns–Blake–Dowd 家族，该模型家族最初是为老年死亡率设计的。本文将这些CBD型模型应用于完整的0–100岁年龄范围，以保持基模型之间的可比性以及集成构建的一致性。因此，这些模型设定在成年及老年年龄段最具解释意义。相比之下，在较年轻年龄段，集成中其他模型家族的存在减轻了它们可能存在的设定偏误。因此，在某些国家，M4–M6表现良好，应理解为：尽管这些模型最初针对老年年龄段设计，但仍能为组合预测提供有用信息，尤其是在老年年龄段；而其他模型（例如Lee–Carter 变体以及函数模型M13和M14）往往在较年轻年龄段贡献更大。

各国模型之间观察到的差异凸显了识别死亡率建模中一致偏好的难题。同样，也很难找到一种能在所有情境下稳定发挥作用的单一死亡率预测方法。此外，所选模型的预测表现会因拟合数据时采用的性别而异，没有任何单一模型能始终在男性和女性群体中表现最佳。这些跨国家和性别的差异凸显了找到一种普遍最优的死亡率预测方法所面临的挑战。事实上，预测表现因国家和性别而异，没有任何单一模型能在所有条件下始终优于其他模型。这种复杂性说明了采用模型集成的必要性，它能够综合各种方法各自的优势，从而提高整体预测准确度。

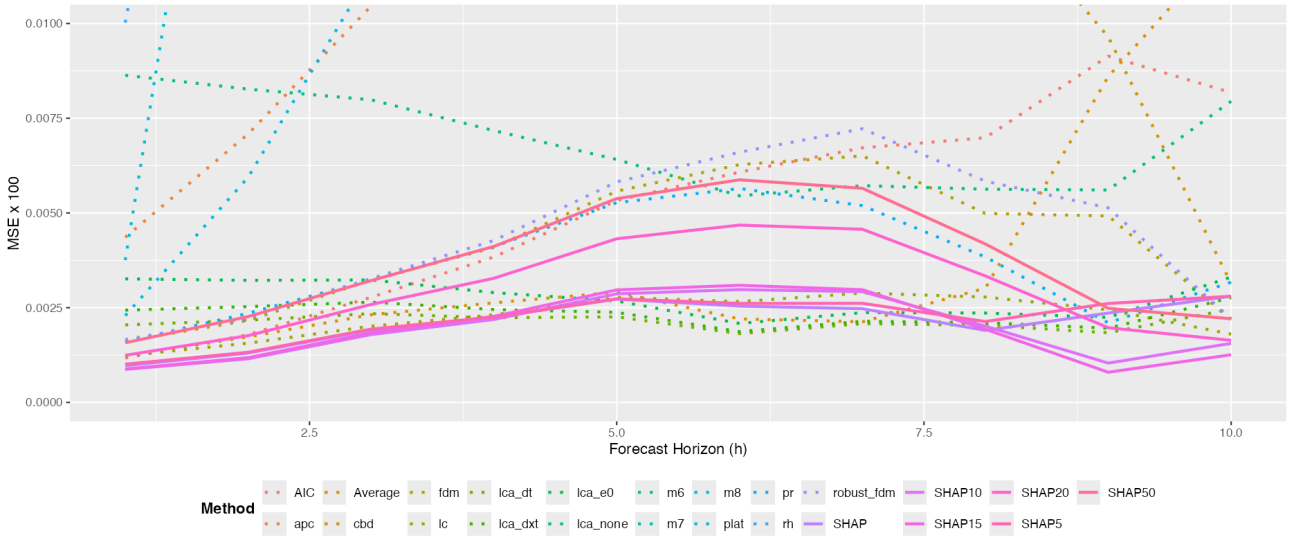
5.5. 集成与单个模型的比较

基于均方误差（MSE）对 SHAP 集成与单个模型进行比较后发现，对于男性和女性而言，集成模型在多个国家通常优于单个模型。值得注意的是，在新西兰、挪威、西班牙和美国，这一优势十分明显：基于 SHAP 的预测在所有预测步长和性别中均能持续提升预测准确度。

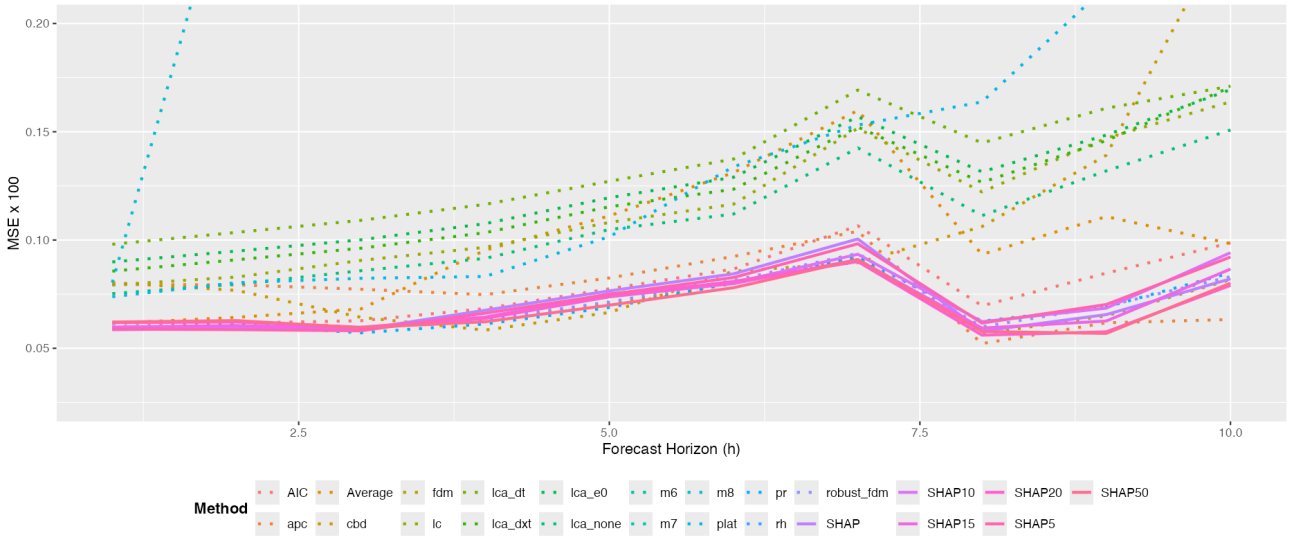
相比之下，在其他一些国家，SHAP集成的优势似乎因性别而异。例如，在比利时、意大利、日本（考虑到日本数据的特殊性，这一结果尤其引人注目）、立陶宛和瑞士，SHAP集成对女性的表现明显更优。相比之下，捷克共和国、芬兰、爱尔兰、拉脱维亚和英国则表现出针对男性的优势：集成方法显著改善了男性死亡率的预测。

研究结果强调，尽管 SHAP 集成能够显著提升性能，且提升幅度往往很大，但对于所有国家和两种性别而言，没有任何单一模型或模型组合能够普遍达到最优。这些结果与Kessy等人（2022年）的见解一致。他们指出，要找到一种在不同人口和地理背景下始终表现良好的主导性死亡率预测方法，本身就面临着固有挑战。观察到的异质性凸显了模型集成的重要性，也说明采用集成策略是审慎之举，尤其是在存在国家和性别特定动态的情况下。图6展示了这种比较，其中将基于 SHAP 的集成预测与单个模型的预测进行了对比

以两个具有代表性的案例为例：日本（女性）和爱尔兰（男性）。在这两个案例中，集成方法的表现均显著优于基准。基于均方误差对各模型进行分析的结果，可在 GitHub 上的 Shiny 应用中查看。



(a) 日本女性数据



(b) 爱尔兰男性数据

图 6：日本（女性）和爱尔兰（男性）的 SHAP集成与单模型预测对比。

6. 集成预测区间

在实际应用中，量化死亡率预测方法相关的不确定性至关重要。我们遵循 Shang and Haberman (2018) 的方法计算样本外预测区间

针对2010年至2019年的测试数据，采用扩展窗口方法。与点预测分析类似，该方法会生成10个一步预测区间、9个两步预测区间，……以及1个十步预测区间。随后，我们按照下一节介绍的方法构建 h步预测集成预测区间。

6.1. 集成区间构建

我们考虑 Grushka-Cockayne 和 Jose (2020年) 以及 Li等人(2025年) 提出的四种预测区间组合方法：简单平均（SA）、内部修剪（IT）、赤池信息准则和均方误差，以及一种基于SHAP权重的集成区间构建方法。

令 (L_i, U_i) 表示某个死亡率预测模型的中心 $100(1-\alpha)\%$ 预测区间，其中 $i = 1$、2、……、N。令 (L_C, U_C) 表示采用方法 $C \in \{SA, IT, AIC, MSE, SHAP\}$，基于单个预测模型计算出的 N 个预测区间构建的集成 $100(1-\alpha)\%$ 预测区间。

6.1.1. 简单平均

简单平均法计算所有待考虑模型的下端点和上端点。实际上，对每个死亡率预测模型都赋予相同的权重 $1/N$。尽管形式简单，这种集成区间构建方法已被证明在许多应用中都很有用。

$$L_{SA} = \frac{1}{N}\sum_{i=1}^{N} L_i, \qquad U_{SA} = \frac{1}{N}\sum_{i=1}^{N} U_i.$$

6.1.2. 内部修剪

内部修剪法会排除 d 个最大的下端点和 d 个最小的上端点，然后对剩余的预测区间计算简单平均值。令下标 (j) 表示第 j个次序统计量，则集成区间的计算方式为

$$L_{IT} = \frac{1}{N-d}\sum_{j=1}^{N-d} L_{(j)}, \qquad U_{IT} = \frac{1}{N-d}\sum_{j=d+1}^{N} U_{(j)}.$$

实际上，该方法将权重 0 分配给截尾模型，并为其余每个预测区间分配相等的权重。我们遵循 Liet al. (2025年)和 Grushka-Cockayne和 Jose (2020年) 的做法，排除 20% 的模型，即 $d = 15 \times 0。2 = 3$ 个模型参与本次分析。

6.1.3. 基于 *AIC* 的加权平均

该方法在验证集上，针对具有负赤池信息准则估计值的模型$\widehat{\ }$，即 $\mathrm{AIC}_i(\nu; g) < 0$，计算其下端点和上端点的加权平均值。由于较低的赤池信息准则值表示模型拟合效果更好，因此取绝对值可确保这些模型获得更高的权重。

$$L_{AIC} = \sum_{i=1}^{N} w_{i,AIC} L_i, \qquad U_{AIC} = \sum_{i=1}^{N} w_{i,AIC} U_i,$$

其中 N^* 是$\widehat{\ }$具有负估计 $\mathrm{AIC}_i(\nu; g)$ 值的预测方法的数量，并且

$$w_{i,AIC} = \frac{|\widehat{\mathrm{AIC}}_i(\nu; g)|}{\sum_{i=1}^{N^*} |\widehat{\mathrm{AIC}}_i(\nu; g)|} \mathbb{1}\left\{\widehat{\mathrm{AIC}}_i(\nu; g) < 0\right\}.$$

6.1.4. 基于 *MSE* 的加权平均

除了像上文那样根据拟合表现进行平均，也可以使用验证集上的预测均方误差来选择权重 w_i，如下所示。

$$L_{MSE} = \sum_{i=1}^{N} w_{i,MSE} L_i, \qquad w_{i,MSE} = \frac{\exp(-\mathrm{MSE}_i)}{\sum_{i=1}^{N} \exp(-\mathrm{MSE}_i)},$$
$$U_{MSE} = \sum_{i=1}^{N} w_{i,MSE} U_i.$$

指数运算可确保权重随着均方误差的增加而适度衰减，从而为点预测准确性较高的模型赋予更高的权重。

6.1.5. 基于 *SHAP* 的加权平均

基于 SHAP 的加权方法也可用于构建集成预测区间。SHAP 加权方案根据验证集上的归一化平均 SHAP 值来选择权重，并为偏差和方差较低的模型分配更大的权重。对于特定年龄，我们计算 SHAP 权重以及由下式给出的预测区间

$$L_{SHAP} = \sum_{i=1}^{N} w_{i,SHAP} L_i, \qquad U_{SHAP} = \sum_{i=1}^{N} w_{i,SHAP} U_i,$$

其中 N^* 表示　　表示 SHAP 值高于阈值 α 的模型数量，并且

$$w_{i,\mathrm{SHAP}} = \frac{\widetilde{\phi}_i(\nu;g)}{\sum_{i=1}^{N^*}\widetilde{\phi}_i(\nu;g)}\mathbb{1}\left\{\widetilde{\phi}_i(\nu;g) > \alpha\right\}.$$

6.2. 区间预测的评估

我们通过Gneiting 和 Raftery (2007年) 提出的区间评分来评估区间预测准确性（另见 Gneiting 和 Katzfuss，2014年）。对数尺度下的逐点预测区间评分定义为

$$\begin{aligned} S_\theta\left[L_C(x,g,t+h), U_C(x,g,t+h), y_{x,g,t+h}\right] = {} & \left[U_C(x,g,t+h) - L_C(x,g,t+h)\right] \\ & + \frac{2}{\theta}\left[L_C(x,g,t+h) - \ln y_{x,g,t+h}\right]\mathbb{1}\left\{\ln y_{x,g,t+h} < L_C(x,g,t+h)\right\} \\ & + \frac{2}{\theta}\left[\ln y_{x,g,t+h} - U_C(x,g,t+h)\right]\mathbb{1}\left\{\ln y_{x,g,t+h} > U_C(x,g,t+h)\right\}. \end{aligned}$$

当留出观测值位于预测区间内时，区间评分会对较窄的预测区间给予奖励。真实观测值位于 $L_C(x,g,t+h)$ 和 $U_C(x,g,t+h)$之间，且 $L_C(x,g,t+h)$和 $U_C(x,g,t+h)$ 之间的逐点距离最小时，可获得最优区间评分。参数 θ 表示显著性水平，通常选取为 $\theta = 0。2$。对于不同的预测期限 h 以及死亡率曲线上的不同时间点，平均区间评分定义为

$$\overline{S}_\theta(h) = \frac{1}{101\times(11-h)}\sum_{\xi=h}^{11}\sum_{x=0}^{101} S_\theta\left[L_C(x,g,t+\xi), U_C(x,g,t+\xi), y_{x,g,t+\xi}\right].$$

在表3中，我们给出使用简单平均、内部修剪、赤池信息准则、均方误差和基于 SHAP 的加权方案所得的一步预测的平均区间评分。作为对 SHAP 集成进行阈值处理的示例，选择适中的截断水平 $\alpha = 10\%$。两步至十步预测区间的评估结果可在 GitHub 上的 Shiny 应用中查看。

表3：经济合作与发展组织死亡率国家集成区间预测的一步前（h = 1）平均区间评分。本表报告了简单平均、未截断的 SHAP、采用α = 10%截断的 SHAP、内部修剪、基于赤池信息准则和基于均方误差的集成的相关结果。

	简单平均		**SHAP**		**SHAP** $\alpha = 10\%$	
国家	F	M	F	M	F	M
奥地利	1.6864	1.5280	1.6700	1.5548	1.7342	1.7093
比利时	1.6788	1.4757	1.6132	1.1635	1.6001	1.0953
捷克	1.3585	1.4590	1.3957	1.5281	1.5512	1.5946
丹麦	2.1237	1.7903	1.7273	1.5503	1.7465	1.5655
爱沙尼亚	1.8764	1.7977	1.7750	2.0704	1.8542	2.1110
芬兰	1.6775	1.5688	1.6757	1.6257	1.7041	1.6412
法国	1.2599	1.2731	0.8499	0.8595	0.8113	0.8075
匈牙利	2.0375	2.3011	1.4091	1.5939	1.4158	1.6545
冰岛	1.9836	2.2088	2.2782	2.3339	2.3502	2.3652
爱尔兰	2.0303	1.7180	1.8859	1.8936	1.8862	1.7325
意大利	1.3691	1.1925	1.8284	1.1631	1.9453	1.1845
日本	1.0573	1.0303	1.1275	0.9539	1.1946	1.0273
拉脱维亚	2.7536	1.9807	2.8336	2.2892	2.8907	2.3254
立陶宛	1.7458	2.3068	1.6090	2.2937	1.6261	2.2891
卢森堡	1.7800	2.0453	1.9208	2.0565	2.0115	2.0951
荷兰	1.6632	1.4559	1.3372	1.3839	1.2902	1.6189
新西兰	1.6344	1.5191	1.8735	1.7124	1.9834	1.7703
挪威	1.7496	1.4510	1.5948	1.5384	1.6026	1.5721
波兰	1.5216	1.5150	1.6181	1.3784	1.8231	1.3634
西班牙	1.3867	7.8200	1.3840	5.2977	1.3233	3.0897
瑞典	1.4622	1.5585	1.3448	1.3330	1.3817	1.3352
瑞士	1.9896	1.9386	1.8089	1.7400	1.8106	1.8183
UK	1.3987	1.4396	1.6277	1.6355	2.0662	1.7464
USA	1.4975	1.4247	1.0614	0.8744	0.9944	0.6934
Mean	1.6967	1.9083	1.6354	1.7427	1.6589	1.7101
	内部修剪		**AIC**		**MSE**	
国家	F	M	F	M	F	M
奥地利	**1.0222**	**0.8532**	1.2664	0.9903	1.4321	1.2379
比利时	**1.0125**	**0.8491**	1.2066	1.0017	1.4340	1.2243

捷克	**0.8057**	**0.8487**	1.0327	0.9047	1.1333	1.1743
丹麦	**1.3153**	**1.1100**	1.6290	1.2535	1.8411	1.5491
爱沙尼亚	**1.5692**	**1.2700**	1.8948	1.6387	1.8776	1.7259
芬兰	**1.2332**	**1.0829**	1.4799	1.1800	1.5812	1.1945
法国	**0.5193**	**0.4959**	0.6084	0.5143	0.9942	0.9589
匈牙利	**1.1081**	**1.2734**	1.3699	1.6906	1.6848	2.0493
冰岛	**1.7801**	**1.9743**	1.8074	2.0974	1.9474	2.1992
爱尔兰	**1.3275**	**1.0769**	1.5350	1.2967	1.7415	1.5085
意大利	0.6039	0.5598	**0.5811**	**0.5537**	0.9152	0.8334
日本	**0.5504**	**0.5068**	0.5883	0.5644	0.8506	0.8603
拉脱维亚	**2.0929**	**1.3091**	2.6729	1.6556	2.7725	1.8484
立陶宛	**1.1972**	**1.6432**	1.5601	2.0245	1.6413	2.2152
卢森堡	**1.5977**	**1.8470**	1.7749	1.9408	1.7865	2.0165
荷兰	**0.8442**	**0.7229**	1.0810	0.9137	1.3713	1.1944
新西兰	**1.0994**	**0.9682**	1.2688	1.1654	1.4546	1.3561
挪威	**1.2384**	**0.9516**	1.5097	1.1664	1.6277	1.2927
波兰	0.6349	0.6588	**0.6279**	**0.5800**	0.8487	0.9788
西班牙	**0.6833**	4.2268	0.7108	**0.8513**	1.0697	8.0521
瑞典	**0.9078**	**0.9066**	1.1516	1.0646	1.2940	1.3398
瑞士	**1.2758**	1.2596	1.5985	**1.2532**	1.8206	1.8009
UK	**0.5125**	**0.5604**	0.5407	0.5748	0.9693	1.0729
USA	**0.4941**	**0.4785**	0.5939	0.5278	1.1482	1.1103
Mean	**1.0594**	1.1431	1.2538	**1.1418**	1.4682	1.6997

从平均区间评分来看，内部修剪和 AIC 方法通常能取得最佳的预测区间准确性。基于 SHAP 的集成预测区间构造方法在大多数国家优于简单平均法；在丹麦、法国、匈牙利、瑞典和美国，该方法显著降低了两种性别的平均区间评分，在比利时和西班牙则显著降低了男性的平均区间评分。尽管基于 SHAP 的方法仅在爱沙尼亚女性死亡率预测方面优于基于 AIC 的集成方法，但在法国、匈牙利、瑞士和美国，其区间评分均低于基于 MSE 的方法。

根据 SHAP 估计值截去低贡献模型，可提高比利时、法国、西班牙和美国的区间预测准确性。具体而言，剔除归一化 SHAP 值低于 0.1 的模型后，基于 SHAP 的集成区间的平均区间评分低于

内部修剪法。在表A.1 中的 Renshaw-Haberman模型以及表A.2 中的 M6、M7、M8 方法，对西班牙男性死亡率给出的区间预测不如其他方法。

7. 结论

我们提出了一种基于 SHAP 的集成方法，用于预测多个国家的死亡率。该方法根据多种预测方法的边际贡献动态分配权重，将不同模型的互补优势结合起来，无论这些模型来自 Renshaw-Haberman、Cairns-Blake-Dowd、Lee-Carter，还是函数型时间序列模型，从而比依赖单一模型提高了预测准确度。这些结果表明，不同的死亡率模型在不同预测期限内能够捕捉死亡率数据的不同特征。通过集成方法将它们结合起来，既能发挥各自优势，也能减轻每种模型固有的缺点。

此外，对24个经济合作与发展组织国家数据的实证分析表明，在均方误差和平均绝对误差方面，基于SHAP的集成优于较简单的组合方案（例如简单模型平均或基于AIC的权重）。该框架的另一项优势是可解释性：SHAP值能够显示在不同年龄组、预测期限和性别中，哪些模型对整体预测准确度的贡献最大。

此外，对 SHAP值设置阈值，排除贡献微不足道或有损整体表现的模型，能够在不影响整体准确性的前提下有效降低预测方差。我们还发现，没有任何单一死亡率模型能在所有国家或两种性别中始终表现突出，这表明该领域存在固有的模型风险。SHAP 集成通过自适应地选择并加权模型，适应人口统计模式和数据可用性的差异，从而缓解了这一问题。

未来研究可以将所提出的方法扩展至涵盖更广泛的死亡率模型，考察自适应阈值选择，或探索与元学习技术（如堆叠回归）的结合，进一步提升死亡率预测的稳健性和可解释性。不同国家可能具有相似的死亡率特征，对同质人群进行聚类或许能够提高预测准确度。由于聚类归属会影响SHAP值计算，如何利用SHAP值确定最佳聚类数量仍是一个尚未探索的研究问题。

将经济合作与发展组织国家的死亡率进行部分汇聚，并将其纳入SHAP 集成的构建，可能是未来研究的一个有前景的方向。

致谢

作者感谢编辑、副编辑和三位审稿人提出的富有洞察力的评论与建议。作者感谢阿姆斯特丹2025年Gompertz 200 会议参与者提出的意见。Han Lin Shang 感谢澳大利亚研究理事会未来学者奖（FT240100338）提供的资金支持。

资金支持

本研究未获得来自公共部门、商业部门或非营利部门资助机构的任何专项资助。

作者贡献

所有作者对该稿件的贡献均等。

利益冲突

作者声明，他们不存在利益冲突。

附录A。15个模型及其约束条件

表A.1：Renshaw-Haberman族模型

标签	名称	模型	限制条件
M1	Lee-Carter（LC）模型 - 泊松误差	$\ln m_{x,t} = \alpha_x + \beta_x^{(1)}\kappa_t^{(1)}$	$\sum_t \kappa_t^{(1)} = 0,\ \sum_x \beta_x^{(1)} = 1$
M2	Renshaw-Haberman	$\ln m_{x,\ t} = \alpha_x + \beta_x^{(1)}\kappa_t^{(1)} + \beta_x^{(0)}\gamma_{t-x}$	$\sum_t \kappa_t^{(1)} = 0,\ \sum_x \beta_x^{(1)} = 1,\ \sum_x \beta_x^{(0)} = 1$
M3	年龄-时期-队列	$\ln m_{x,\ t} = \alpha_x + \kappa_t^{(1)} + \gamma_{t-x}$	$\sum_t \kappa_t^{(1)} = 0,\ \sum_c \gamma_c = 0,\ \sum_c c\,\gamma_c = 0$

表A.2：Cairns-Blake-Dowd（CBD）系列模型

标签	名称	模型	约束
M4	CBD	$\ln m_{x,\ t} = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)}$	无需约束
M5	M6	$\ln m_{x,\ t} = \kappa_t^{(1)} + (x-\bar{x})\,\kappa_t^{(2)} + (x_c - x)\,\gamma_{t-x}$	与 M7 类似，但不含二次项
M6	M7	$\ln m_{x,\ t} = \kappa_t^{(1)} + (x-\bar{x})\,\kappa_t^{(2)} + ((x-\bar{x})^2 - \sigma_x^2)\,\kappa_t^{(3)} + \gamma_{t-x}$	$\sum_t \kappa_t^{(1)} = 0,\ \sum_t \kappa_t^{(2)} = 0,\ \sum_t \kappa_t^{(3)} = 0,$ $\sum_c \gamma_c = 0,\ \sum_c c^2\gamma_c = 0$
M7	M8	$\ln m_{x,\ t} = \alpha_x + \kappa_t^{(1)} + (x-\bar{x})\,\kappa_t^{(2)} + (x-\bar{x})^+\kappa_t^{(3)} + \gamma_{t-x}$	类似于平坦模型
M8	Plat	$\ln m_{x,\ t} = \alpha_x + \kappa_t^{(1)} + (x-\bar{x})\,\kappa_t^{(2)} + (x-\bar{x})^+\kappa_t^{(3)} + \gamma_{t-x}$	$\sum_t \kappa_t^{(1)} = 0,\ \sum_t \kappa_t^{(2)} = 0,\ \sum_t \kappa_t^{(3)} = 0,$ $\sum_c \gamma_c = 0,\ \sum_c c^2\gamma_c = 0$

表 A.3：LC 模型及其变体

标签	名称	模型	限制因素
M9	Lee–Carter模型 - 高斯误差	$\ln m_{x,t} = \alpha_x + \beta_x^{(1)}\kappa_t^{(1)} + \varepsilon_{x,t}$	$\sum_t \kappa_t^{(1)} = 0,\ \sum_x \beta_x^{(1)} = 1$
M10	BMS	$\ln m_{x,\ t} = \alpha_x + \beta_x^{(1)}\kappa_t^{(1)} + \beta_x^{(2)}\kappa_t^{(2)}$	$\sum_t \kappa_t^{(1)} = 0,\ \sum_x \beta_x^{(1)} = 1,\ \sum_x \beta_x^{(2)} = 1$
M11	Lee–Carter模型 - 预期寿命调整	$\ln m_{x,\ t} = \alpha_x + \beta_x^{(1)}\tilde{\kappa}_t^{(1)}$	$\sum_t \tilde{\kappa}_t^{(1)} = 0,\ \sum_x \beta_x^{(1)} = 1$
M12	Lee–Carter模型 - 不调整得分	$\ln m_{x,\ t} = \alpha_x + \beta_x^{(1)}\kappa_t^{(1)}$	无事后校准约束

表 A.4：函数型时间序列模型

标签	名称	模型	限制条件
M13	FDM	$m_{x,t} = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + \varepsilon_{x,t}$	对前 K 个主成分的约束
M14	稳健函数数据模型	$m_{x,t} = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + \varepsilon_{x,t}$	与函数型数据模型类似，但对离群值具有稳健性
M15	乘积比	$p_t(x) = \sqrt{m_M(x,t)\,m_F(x,t)},\quad r_t(x) = \sqrt{\frac{m_M(x,t)}{m_F(x,t)}}$	无特定限制

附录B. 均方误差分解：SHAP权重与等权重 s

既有研究强调了偏差、方差和多样性如何在不同预测步长之间相互作用。Taieb 和 Atiya (2015) 表明，多步策略呈现出特定于预测步长的偏差/方差行为；

Wood 等人（2023）强调，等权重往往能够保留多样性，而在短期预测区间内，这可以成为一种有效的方差缩减手段。因此，在我们的设定中，等权重在短期内可能具有竞争力；随着预测期延长，性能加权方案（包括 SHAP）可能会逐渐占据优势，因为条件偏差缩减的重要性相对提高。

令 $\mathcal{G}_h$ 表示预测期 h 的样本外评估集，即我们据以计算损失的单元格集合 j（例如，年龄 $\times$、队列 $\times$、性别）。对于每个 $j \in \mathcal{G}_h$，令 $f_{h,\,j} \in \mathbb{R}^N$ 为基础预测的（中心化）向量；我们逐个单元格构造组合预测，如下所示：

$$\widehat{y}_{h,j} = w' f_{h,j}.$$

我们通过均方误差评估准确性：

$$\mathrm{MSE}_h \;=\; \mathbb{E}\Big[\big(y_{h,j} - \widehat{y}_{h,j}\big)^2\Big] \;=\; \underbrace{\big(\mathbb{E}[\widehat{y}_{h,j}] - y_{h,j}\big)^2}_{\mathrm{Bias}_h^2} + \underbrace{\mathrm{Var}(\widehat{y}_{h,j})}_{\mathrm{Var}_h} + \sigma^2,$$

其中，σ^2 表示不可约噪声。用我们的符号表示，$f_{h,\,j} = (\widehat{m}^{(1)}_{x_j,g_j,t_j+h}, \ldots\ldots \widehat{m}^{(N)}_{x_j\,g_j',\,t_j+h})'$。

全文中，变量均在每一折内进行中心化（等价于纳入折特定截距），因此水平偏差由截距吸收，而权重则反映二阶矩结构（Elliott 和 Timmermann，2004年）。

为将这一分解与组合文献联系起来，考虑 $\Sigma_{f,\,h} = \mathrm{Var}(f_{h,j})$，它表示预测期 h 下基础预测的 $N \times N$折内协方差矩阵（假设其为正定，即不存在完全共线性）。

命题 1 (条件优势)。 在平方误差损失下，并将组合权重视为给定值时，预测期 h 的条件均方误差由回归解 $w_h^\star$(Elliott 和 Timmermann，2004年) 最小化。对于任意其他权重 w，

$$\mathrm{MSE}_h^{\mathrm{cond}}(w) \;=\; \sigma_h^2 \;+\; (w - w_h^\star)'\,\Sigma_{f,h}\,(w - w_h^\star),$$

其中

$$\mathrm{MSE}_h^{\mathrm{cond}}(w) \;=\; \mathbb{E}\big[(y_{h,j} - w' f_{h,j})^2 \,\big|\, w\big] \quad \text{and} \quad \sigma_h^2 = \mathrm{MSE}_h^{\mathrm{cond}}(w_h^\star).$$

因此，均方误差$_h^{\mathrm{cond}}(w_h^\star) \leqslant$ 均方误差$_h^{\mathrm{cond}}(w)$，当且仅当 $w = w_h^\star$时取等号。如果基于 SHAP 的加权

渐近地 恢复了 $w_h^\star$，则其条件均方误差弱优于等权重w_{eq} .

备注1。令 $\widehat{w}_h$ 为估计值 得到权重，并记为 $\Delta_h = \widehat{w}_h - w_h^\star$。无条件 均方误差满足 s

$$\mathbb{E}[\mathrm{MSE}_h(\widehat{w}_h)] \;=\; \sigma_h^2 \;+\; \underbrace{\mathrm{tr}(\Sigma_{f,h}\,\mathrm{Var}(\widehat{w}_h))}_{\text{estimation variance}} \;+\; \underbrace{(\mathbb{E}[\widehat{w}_h]-w_h^\star)'\,\Sigma_{f,h}\,(\mathbb{E}[\widehat{w}_h]-w_h^\star)}_{\text{weight bias}}.$$

对于等权重等固定方案,

$$\mathrm{MSE}_h^{\mathrm{cond}}(w_{eq}) = \sigma_h^2 + (w_{eq}-w_h^\star)'\Sigma_{f,h}(w_{eq}-w_h^\star).$$

因此，只有当相对于 w_{eq} 的条件误差降幅超过估计方差（以及任何权重偏差）带来的惩罚时，基于 SHAP 的权重才会改善样本外表现。这种偏差—方差权衡是预测组合难题的根源（Elliott 和 Timmermann，2004年、2005年；Smith 和 Wallis，2009年；Genre 等人，2013年）。

我们的条件性论述与 Elliott 和 Timmermann (2004年)的观点一致：在平方误差损失下，使条件均方误差最小的权重与回归解相同；而中心化（或折特定截距）可吸收水平偏差，使权重反映二阶矩。当相对表现随时间变化时，Elliott 和 Timmermann (2005年)阐释了为何恒定的最小二乘权重可能表现不佳，以及自适应组合或依赖状态的组合为何能够占优。这与预测期延长时，基于 Shapley 值的权重日益受到青睐的现象一致。

备注 1 中的限定条件与组合预测难题直接相关：估计权重的抽样变异性可能抵消条件增益，使简单平均在短样本、高共线性或协方差不稳定的情况下具有竞争力（Smith 和 Wallis，2009年）。大型比较研究同样发现，等权重很难被稳健地超越（Genre 等人，2013年）。在我们的设定中，当基于 SHAP 的权重接近回归解时，它们能改善条件拟合，但其有限样本优势取决于估计方差惩罚。

附录 *B.1. SHAP*权重的均方误差分解

对于采用 SHAP 权重的集成,

$$\widehat{m}^{(c)}_{h,\mathrm{Shap}}(x,g,t+h) = \sum_{i=1}^{N^*} \widehat{\omega}_{i,\mathrm{Shap}}\,\widehat{m}^{(i)}_{x,g,t+h},$$

均方误差可分解为

$$\mathrm{MSE}_{h,\mathrm{Shap}} = \mathrm{Bias}^2_{h,\mathrm{Shap}} + \mathrm{Var}_{h,\mathrm{Shap}} + \sigma^2,$$

with

$$\mathrm{Bias}^2_{h,\mathrm{Shap}} = \Big(\sum_{i=1}^{N^*} \widehat{\omega}_{i,\mathrm{Shap}}\, \mathbb{E}\left[\widehat{m}^{(i)}_{x,g,t+h}\right] - y_{t+h} \Big)^2$$

$$\mathrm{Var}_{h,\mathrm{Shap}} = \sum_{i} \widehat{\omega}^2_{i,\mathrm{Shap}}\, \mathrm{Var}(\widehat{m}^{(i)}) + \sum_{i \neq j} \widehat{\omega}_{i,\mathrm{Shap}} \widehat{\omega}_{j,\mathrm{Shap}}\, \mathrm{Cov}(\widehat{m}^{(i)}, \widehat{m}^{(j)}).$$

附录 *B.2.* 等权重的均方误差分解

对于等权重 $\widehat{\omega}_i = 1/N$,

$$\mathrm{MSE}_{h,\mathrm{equal}} = \mathrm{Bias}^2_{h,\mathrm{equal}} + \mathrm{Var}_{h,\mathrm{equal}} + \sigma^2,$$

with

$$\mathrm{Bias}^2_{h,\mathrm{equal}} = \Big(\tfrac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\widehat{m}^{(i)}_{x,g,t+h}\right] - y_{t+h} \Big)^2,$$

$$\mathrm{Var}_{h,\mathrm{equal}} = \tfrac{1}{N^2} \sum_{i} \mathrm{Var}(\widehat{m}^{(i)}) + \tfrac{1}{N^2} \sum_{i \neq j} \mathrm{Cov}(\widehat{m}^{(i)}, \widehat{m}^{(j)}).$$

结合命题 1 和备注 1，这些分解表明，SHAP 与等权重的相对表现会通过偏差—方差权衡以及估计权重的抽样变异性而依赖于预测期：在较短预测期内，多样性驱动的方差降低可能使等权重具有竞争力；随着预测期延长，基于 SHAP 的加权可能更好地逼近回归解，并降低条件均方误差。

附录 *B.3.* 基于 *SHAP* 的阈值下的均方误差分解

为提高稳定性，我们引入阈值 α，排除贡献可忽略的模型：

$$S(\alpha) = \{ i : \widetilde{\phi}_i(v; g) > \alpha \}, \qquad \widehat{m}^{(c)}_{h,\mathrm{Shap}\,\alpha} = \sum_{i \in S} \widehat{\omega}_{i,\mathrm{Shap}\,\alpha}\, \widehat{m}^{(i)}_{x,g,t+h}.$$

由于 $S(\alpha)$ 是嵌套的 d 在 α 中，增大 α 会略微缩小剩余集合。代入 th e

截断权重得出：

$$\mathrm{Bias}^2_{h,\mathrm{Shap}\,\alpha} = \Big(\sum_{i \in S(\alpha)} \widehat{\omega}_{i,\mathrm{Shap}\,\alpha}\, \mathbb{E}[\widehat{m}^{(i)}_{x,g,t+h}] - y_{t+h} \Big)^2, \tag{B.1}$$

$$\mathrm{Var}_{h,\mathrm{Shap}\,\alpha} = \sum_{i \in S(\alpha)} \widehat{\omega}^2_{i,\mathrm{Shap}\,\alpha}\, \mathrm{Var}(\widehat{m}^{(i)}) \;+\; \sum_{\substack{i \neq j \\ i,j \in S(\alpha)}} \widehat{\omega}_{i,\mathrm{Shap}\,\alpha} \widehat{\omega}_{j,\mathrm{Shap}\,\alpha}\, \mathrm{Cov}(\widehat{m}^{(i)}, \widehat{m}^{(j)}). \tag{B.2}$$

过滤低贡献模型通常会降低方差——尤其是当被排除的模型不稳定时——同时保持偏差不变。然而，过于激进的阈值会降低多样性，可能增大偏差；若负协方差被去除，甚至会抬高方差。在实践中，阈值起到一种收缩作用：适中的取值可改善短视界稳健性，而过度过滤则会因限制多样性而损害长视界性能。

参考文献

Basellini, U., Camarda, C. G. and Booth, H. (2022年), ‘三十年回顾：李–卡特死亡率预测方法综述’, 《国际预测学期刊》**38**(3), 915–944.

Bates, J. M. and Granger, C. W. J. (1969年), ‘预测的组合’, 《运筹学季刊》**20**(4), 451–468.

Benchimol, A. G., Alonso-González, P. J., Marín-Díazaraque, J. M. and Albarrán-Lozano, I. (2016), 采用模型组装方法进行死亡率投影中的模型不确定性方法，技术报告 16-06，马德里卡洛斯三世大学。

Bimonte, G., Russolillo, M., Shang, H. L. and Yang, Y. (2025年), ‘基于夏普利值的死亡率模型集成’, 《经济与金融决策》**48**, 1131–1159.

Booth, H.、Maindonald, J. 和 Smith, L.（2002年），《在死亡率可变下降的情况下应用 Lee–Carter》，《人口研究》**56**(3)，325–336。

Brouhns, N.、Denuit, M. 和 Vermunt, J. K.（2002年），《构建预测生命表的泊松对数双线性回归方法》，《保险：数学与经济学》**31**(3)，373–393。

Cairns, A. J. G.、Blake, D. 和 Dowd, K.（2006年），《具有参数不确定性的随机死亡率双因子模型：理论与校准》，《风险与保险杂志》**73**(4)，687–718。

Cairns, A. J. G.、Blake, D.、Dowd, K.、Coughlan, G. D.、Epstein, D.、Ong, A. 和 Balevich, I.（2009年），《使用英格兰和威尔士及美国数据对随机死亡率模型进行定量比较》，《北美精算学杂志》**13**(1)，1–35。

De Mori, L.、Millossovich, P.、Zhu, R. 和 Haberman, S.（2024），《双人口死亡率预测：一种基于模型平均的方法》，《风险》**12**(4)，60。

Elliott, G. 和 Timmermann, A.（2004），《一般损失函数和预测误差分布下的最优预测组合》，《计量经济学杂志》**122**(1)，47–79。

Elliott, G. 和 Timmermann, A.（2005年），‘机制转换下的最优预测组合’，《国际经济评论》46(4)，1081–1102。

Franses, P. H.、Zou, J. 和 Wang, W.（2024），‘基于沙普利值的预测组合’，《预测学期刊》43(8)，3194–3202。

Genre, V.、Kenny, G.、Meyler, A. 和 Timmermann, A.（2013年），“组合专家预测：有什么能胜过简单平均？”，《国际预测学期刊》**29**(1)，108–121。

Gneiting, T. 和 Katzfuss, M.（2014年），“概率预测”，《统计学及其应用年度评论》**1**(1)，125–151。

Gneiting, T. 和 Raftery, A. E.（2007年），“严格适当评分规则、预测与估计”，《美国统计协会期刊：评论文章》**102**(477)，359–378。

Grushka-Cockayne, Y. 和 Jose, V. R. R.（2020年），“M4竞赛中的预测区间组合”，《国际预测学期刊》**36**(1)，178–185。

Holford, T. R.（1983），“生命统计率的年龄、时期和队列效应估计”，《生物统计学》**39**(2)，311–324。

人类死亡率数据库（2025年），加州大学伯克利分校（美国）和马克斯·普朗克人口研究所（德国）。数据下载于2024年10月2日。网址：www.mortality.org。

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O’Hara-Wild, M., Petropou- los, F., Razbash, S., Wang, E. 和 Yasmeen, F.（2025年），*forecast*：用于时间序列和线性模型的预测函数。R 软件包版本 8.24.0。

网址：https://pkg.robjhyndman.com/forecast/

Hyndman, R. J. 和 Booth, H.（2008年），《使用死亡率、生育率和迁移的函数型数据模型进行随机人口预测》，《国际预测学期刊》**24**(3), 323–342。

Hyndman, R. J., Booth, H. 和 Yasmeen, F.（2013年），《一致性死亡率预测：采用函数型时间序列模型的乘积比率法》，《人口学》**50**(1), 261–283。

Hyndman, R. J. 和 Ullah, M. S.（2007年），“死亡率和生育率的稳健预测：一种函数型数据方法”，《计算统计与数据分析》 **51**(10)，4942–4956。

Kessy, S. R.、Sherris, M.、Villegas, A. M. 和 Ziveyi, J.（2022年），“使用堆叠回归集成进行死亡率预测”，《斯堪的纳维亚精算期刊》**2022年**(7)，591–626。

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Perolat, J., Silver, D. 和 Graepel, T.（2017），多智能体强化学习的统一博弈论方法，收录于I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R.Fergus, S.Vishwanathan 和 R. Garnett 编，《神经信息处理系统进展》，第 30 卷，Curran Associates公司。

网址： https://proceedings.neurips.cc/paper_files/paper/2017/file/3323fe11e9595c09af38fe67567a9394-Paper.pdf

Lee, R.D. 和 Carter, L. R.（1992年），《美国死亡率的建模与forecast》，《美国统计协会期刊》：应用与案例研究**87**（419），659–671。

Lee，R. D. 和 Miller，T.（2001年），“评估 Lee–Carter方法预测死亡率的表现”，《人口学》**38**（4），537–549。

Li，J.、Wang，M.、Liu，J. 和 Tickle，L.（2025年），“死亡率的集成区间预测”，《斯堪的纳维亚精算期刊》**2025年**（6），598–616。

Li，N. 和 Lee，R. D.（2005年），“一组人群的一致性死亡率预测：Lee-Carter方法的扩展”，《人口学》**42**（3），575–594。

Lichtendahl Jr，K. C. 和 Winkler，R. L.（2020年），“为什么某些组合的表现优于其他组合？”，《国际预测学期刊》**36**（1），142–149。

Lundberg，S. M. 和 Lee，S.-I.（2017），解读模型预测的统一方法，载于《神经信息处理系统进展30（NIPS 2017）》。**网址：** https://dl.acm.org/doi/10.5555/3295222.3295230

Makridakis, S., Spiliotis, E. 和 Assimakopoulos, V.（2020年），《M4 竞赛：结果、发现、结论与未来方向》，《国际预测学期刊》**36(1)**, 54–74。

Plat, R.（2009年），《随机死亡率建模》，《保险：数学与经济学》**45(3)**，393–404。

Qin, B.、Zhang, J.、Li, L.、Zeng, L.、Wen, J. 和 Han, M.（2022年），《夏普利值在中国城市燃气需求组合预测中的应用》，《天然气技术与经济》**16**(2)，50–55。

Renshaw, A. E. 和 Haberman, S.（2003年），《采用年龄别增强的 Lee–Carter死亡率预测》，《保险：数学与经济学》**33**(2), 255–272。

Renshaw, A. E. 和 Haberman, S.（2006年），《Lee–Carter模型中死亡率降低因子的基于队列的扩展》，《保险：数学与经济学》**38**(3), 556–570。

Samuels, J. D. 和 Sekkel, R. M.（2017），《模型置信集与预测组合》，《国际预测学期刊》**33**(1), 48–60。

Shang, H.L.、Booth,H. 和Hyndman, R. J.（2011），《死亡率与预期寿命的点预测与区间预测：十种主成分方法的比较》，人口研究**25**, 173–214。

Shang, H. L. 和 Haberman, S.（2018），《模型置信集与预测组合：年龄别死亡率的应用》，*Genus***74**(1), 文章编号 19。

Smith, J. 和 Wallis, K. F.（2009年），《对预测组合之谜的简单解释》，牛津经济与统计公报 **71**(3)，331–355。

Taieb, S. B. 和 Atiya, A. F.（2015），《多步超前时间序列预测的偏差与方差分析》，*IEEE*神经网络与学习系统汇刊 **27**(1)，62–76。

Timmermann, A.（2006年），预测组合，载于 G. Elliott、C. W. J. Granger 和 A. Timmermann 编，《经济预测手册》，第1卷，爱思唯尔，第135–196页。

Wolpert, D. H.（1992年），《堆叠泛化》，神经网络 **5**(2)，241–259。

Wood, D.、Mu, T.、Webb, A. M.、Reeve, H. W.、Lujan, M. 和 Brown, G.（2023），《集成学习中多样性的统一理论》，机器学习研究杂志 **24**(359)，1–49。