STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, June 2026, pp 4789–4804.
Published online in International Academic Press (www.IAPress.org)

Research on Mortality Models Incorporating Cohort Effects from a Compositional Data Perspective

Sijia Wu*, Hongmin Xiao

College of Mathematics and Statistics, Northwest Normal University, Lanzhou, Gansu 730070, China

Abstract Against the backdrop of accelerating global population aging, traditional mortality models are often sensitive to outliers and struggle to capture tail characteristics, leading to an underestimation of future trends. It is worth noting that mortality models with cohort effects, by introducing cohort effects on top of the age-period only framework, can capture mortality patterns more comprehensively. To enhance the robustness and accuracy of predictions, this study embeds the Renshaw–Haberman (RH) model with cohort effects into the framework of Compositional Data Analysis (CoDa) to construct a CoDa-RH model. This study uses data from six countries in the Human Mortality Database (HMD), including four developed countries (Australia, Spain, the United States, and the United Kingdom) and two developing countries (Chile and Bulgaria). The centered log-ratio (clr) transformation is applied to the age distribution structure of death counts, and the bootstrap method is used to construct prediction intervals for life expectancy. Empirical results show that, compared with the traditional Lee-Carter (LC) model and the RH model, in terms of indicators based on log mortality rates, the CoDa-based models exhibit lower Aitchison Distance (AD) and Mean Absolute Error (MAE) on the test set. In terms of life expectancy prediction, this study randomly selects three of the six countries (Spain, UK and Chile) and provides their life expectancy forecasts up to 2035.

Keywords Mortality, Life Expectancy, Compositional Data Analysis, RH Model, Bootstrap Method

DOI: 10.19139/soic-2310-5070-3375

1 Introduction

Throughout human societal development, demographic evolution has consistently been a pivotal factor influencing social, economic, and cultural progress. In recent years, alongside the sustained decline in global mortality rates and the significant extension of life expectancy, population structures are undergoing unprecedented transformation, with the ageing process accelerating continuously. According to the United Nations' World Population Prospects 2024 report, global mortality rates continue to decline, with average life expectancy rising from 70.9 years during the pandemic peak (2020–2021) to 73.3 years in 2024, further accelerating the pace of population ageing.Currently, the proportion of the global population aged 65 and above has reached 10.1%, and is projected to rise to 16.6% by 2050 [1].This implies that pricing and reserve calculations based on historical mortality experience may significantly underestimate future actual payment burdens. Consequently, developing more precise mortality prediction models is crucial for quantifying and managing this systemic risk.

Within mortality modelling research, the LC model stands as a landmark in modern stochastic mortality modelling. Its core assumption posits that log mortality rates decompose into a linear combination of age effects and period effects. Model parameter estimation typically employs singular value decomposition (SVD), while projections are derived by fixing age parameters and extrapolating time-dependent exponentials into the future

*Correspondence to: Sijia Wu (Email: wusijia59420@outlook.com). College of Mathematics and Statistics, Northwest Normal University, Lanzhou, Gansu 730070, China.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

using univariate time series models [2]. Subsequent attempts to develop mortality models drew inspiration from the LC model and extended it in various directions. In terms of model assumptions and fundamental extensions, Brouhns et al. broke through the singular value decomposition framework of the LC model, which was based on the Gaussian error assumption, and proposed to model the number of deaths as following a Poisson distribution, thereby incorporating the model into the framework of generalized linear models (GLM). This provides a more robust theoretical foundation for handling heteroscedasticity in death count data [3].Meanwhile, Renshaw and Haberman approached the issue from the perspective of model structure. By introducing an additive cohort effect term into the age-period two-factor structure of the LC model, they developed the so-called RH model, thereby addressing the limitation of the LC model in capturing cohort differences [4]. However, this model often suffers from estimation instability and convergence difficulties due to an excessive number of parameters. To address the parameter redundancy and identifiability issues of the RH model, Currie et al. pointed out that a more parsimonious Age-Period-Cohort (APC) model framework can be regarded as a reasonable simplification and effective alternative to the RH model [5]. Among contemporaneous model innovations, Cairns, Blake, and Dowd proposed the CBD model, which models mortality through a logistic transformation within an age-period two-factor structure. Compared with the traditional LC model, the CBD model is more suitable for mortality forecasting at advanced ages, thereby broadening the application scenarios of mortality models [6]. Carins et al. introduced three extended models—M6, M7, and M8—based on data from the United Kingdom and the United States [7]. To further resolve the parameter redundancy issue of the RH model, Renshaw and Haberman optimized the model structure by compressing the original two-factor age-cohort adjustment structure into a single-factor form. While retaining the core improvement of the cohort effect, this modification reduces the number of parameters to be estimated and effectively avoids identifiability problems caused by parameter proliferation, thereby significantly enhancing the applicability of the RH framework in empirical research [8]. Furthermore, Palt attempted to integrate the core features of the classical APC model, the LC model, and the CBD model, proposing a comprehensive model that covers the entire age range and incorporates a cohort effect. However, the structure of this model includes two separate period effect terms plus a cohort effect term. It not only inherits the inherent parameter identification problem of the APC model but also may exhibit high correlation between the two period effects, requiring the imposition of strict parameter constraints during estimation [9].

In recent years, with the rise of data-driven approaches, research hotspots in the field of mortality modeling have gradually shifted from traditional parametric stochastic models to machine learning and ensemble learning methods. Kessy et al., building upon traditional combination methods such as simple averaging and Bayesian model averaging, proposed a stacked regression ensemble method that achieves optimal model fusion for different forecast horizons through a meta-learning mechanism, thereby significantly improving prediction accuracy [10]. It is worth noting that the base learners in such ensemble methods are still primarily derived from the aforementioned traditional stochastic mortality models. In terms of directly applying machine learning models, Wang et al. introduced the Transformer architecture into mortality time series forecasting tasks, leveraging its self-attention mechanism to effectively capture the complex nonlinear characteristics in the evolution of mortality rates, achieving superior predictive performance compared to traditional neural network models [11]. Cheng et al., addressing the joint modeling of multiple population groups, proposed a shared-private deep learning structure that not only learns common mortality trends across groups but also retains and captures group-specific deviations, enabling coherent mortality forecasting for multiple populations [12]. However, such methods based on complex machine learning and deep learning generally face the challenge of insufficient model interpretability.

In summary, whether in structural extensions of classical models or in the exploration of emerging methods, the fitting of existing mortality models is almost invariably based on the conditional mean regression framework. The mean regression modeling approach exhibits certain limitations, including sensitivity to outliers and inadequacy in capturing tail risks in the distribution of mortality rates. Therefore, exploring new methods that go beyond the mean regression framework to more comprehensively characterize the intrinsic complex structure of mortality data has become an important research direction for achieving robust mortality forecasting and accurate quantification of longevity risk.

To overcome the limitations of traditional mortality forecasting methods, Oeppen first proposed treating the age-specific death distribution in life tables (i.e., the proportion of deaths at each age to total deaths) as compositional

data, and based on this framework, performing LC modeling and forecasting on CLR-transformed data [13]. Compositional Data Analysis (CoDa) is an analytical framework specifically designed for compositional data, which are non-negative vectors whose components represent parts of a whole with a fixed sum. Standard statistical analysis requires first mapping such data from the Aitchison simplex to the real space. The centered log-ratio (CLR) transformation, owing to its strong interpretability and distance-preserving properties, has become the most commonly used transformation in the CoDa framework [20]. The age-specific death counts in life tables are non-negative, range between zero and the life table radix, and the sum of deaths across all ages in each year is exactly equal to the life table radix, naturally satisfying the core characteristics of compositional data. Traditional log-linear methods for forecasting death counts assume independence among age-specific predictions, thus failing to satisfy the life table radix constraint. In contrast, the CoDa framework, by leveraging the constant-sum property, establishes a natural covariance structure among different age groups, effectively addressing this issue. Building on this theoretical foundation, Oeppen further incorporated the classical LC model into the CoDa framework, using the CLR transformation to map death counts from the Aitchison simplex to real space for modeling and forecasting [13]. Bergeron-Boucher et al. introduced the CoDa framework into the field of multi-population coherent forecasting, integrating the coherence ideas of Li and Lee to develop a "compositional data coherent forecasting model" [14, 15]. More recently, the same authors extended the CoDa approach beyond general mortality forecasting to the joint and coherent modeling of health status and mortality, proposing the CoDAS and CoDAM models based on the Sullivan method and multistate life table to forecast healthy life expectancy (HLE/DFLE), thus expanding the application of compositional data analysis to population health research[16]. Moreover, the application of this analytical framework has also been extended to the study of cause-of-death structures. For example, Oeppen himself, as well as Stefanucci and Mazzuco, have used this method to conduct systematic and in-depth analyses of cause-specific mortality data [17]. It is worth noting that among various mortality models, Renshaw and Haberman extended the classical LC model by incorporating a cohort effect to capture mortality trends specific to the same birth cohort; this extended model is commonly referred to as the "RH model". When analyzing historical data with pronounced cohort effects, the RH model generally exhibits better fitting performance compared to baseline models that do not account for cohort effects [18]. Bergeron-Boucher et al. further emphasized that the CoDa approach is not only applicable to the classical LC structure, but its covariance advantages derived from the constant-sum constraint can also be extended to other modeling environments that incorporate cohort effects [14].

In summary, this study aims to integrate the RH model with the Compositional Data Analysis (CoDa) framework to construct a mortality forecasting model that incorporates cohort effects within the CoDa framework. To further quantify the uncertainty of future trends, a bootstrap method is employed for interval estimation of mortality forecasts, thereby constructing prediction intervals for life expectancy.

The remainder of this paper is structured as follows: Section 2 introduces traditional mortality models and their parameter estimation methods. Section 3 elaborates on the basic concepts of compositional data, within which the mortality model and its corresponding parameter estimation approach are developed. Section 4 presents an empirical analysis, providing a comprehensive comparison and evaluation of four models along with life expectancy prediction results. Section 5 concludes with a summary of findings and directions for future research.

2 Traditional Mortality Models

2.1 LC Model

The LC model decomposes the log mortality rate into an age-specific baseline mortality rate, a common temporal trend, and a factor representing the sensitivity of different ages to this temporal trend. The model is expressed as follows:

$$\ln(m_{x,t}) = a_x + b_x k_t + \epsilon_{x,t}, \tag{1}$$

where $m_{x,t}$ denotes the central mortality rate for individuals aged x at time t, a_x represents the age-dependent parameter, reflecting the average level of the natural logarithm of age-specific mortality rates; k_t constitutes the time-dependent parameter, commonly termed the time index, indicating the rate of change in mortality over time;

b_x signifies the sensitivity of k_t to the age factor, and $\epsilon_{x,t}$ constitutes the error term. Here a_x, b_x and k_t are all parameters to be estimated. Research indicates that if $\{\tilde{a}_x, \tilde{b}_x, \tilde{k}_t\}$ is a set of solutions satisfying equation (1), then for any scalar c and d, the following transformation:

$$\left\{\tilde{a}_x, \tilde{b}_x, \tilde{k}_t\right\} = \left\{\tilde{a}_x - c\tilde{b}_x, \frac{\tilde{b}_x}{d}, d(\tilde{k}_t + \tilde{c})\right\} \tag{2}$$

will also yield invariant fitted values.
Therefore, two constraints are added to formula (1) to ensure the uniqueness of the parameter estimation results:

$$\sum_t k_t = 0, \quad \sum_x b_x = 1 \tag{3}$$

To estimate coefficients in the LC model, Lee and Carter established a classic and authoritative parameter estimation framework in their pioneering 1992 study—specifically, the singular value decomposition of the centred mortality matrix. Specifically, first, based on historical mortality data, the benchmark level parameters $\hat{a}_x$ are obtained by calculating the average logarithmic mortality rates for each age group. Subsequently, singular value decomposition is applied to the centred logarithmic mortality matrix to extract the first principal component, thereby constructing the time-term parameter $\hat{k}_t$ and the age-sensitivity parameter $\hat{b}_x$. Following parameter estimation, the model's core step involves extrapolating the temporal component using time series methods such as Auto-Regressive Integrated Moving Average (ARIMA). The resulting forecast values are then recombined with the estimated age parameters $\hat{a}_x$ and $\hat{b}_x$ to compute comprehensive projected mortality rates $\hat{m}_{x,t}$ for all age groups in future periods.

2.2 RH Model

The RH model represents a significant extension of the LC model. By incorporating cohort effects, it captures the unique mortality trends experienced by cohorts born in different periods, thereby enhancing the model's explanatory power.

$$\ln(m_{x,t}) = a_x + b_x^{(1)} k_t + b_x^{(2)} \gamma_{t-x} + \varepsilon_{x,t} \tag{4}$$

Within this model, a_x, k_t and $b_x^{(1)}$ retain the same interpretative meanings as in the LC model. The cohort effect γ_{t-x} reflects the influence of birth year on mortality rates, while $b_x^{(2)}$ measures the intensity and pattern of this cohort effect across different age groups throughout the life course. The model's constraints are:

$$\sum_x b_x^{(1)} = 1, \quad \sum_t k_t = 0, \quad \sum_{t-x} \gamma_{t-x} = 0 \tag{5}$$

Its parameter estimates are obtained using SVD to derive initial values in the first stage, identical to LC model parameter estimation, followed by iterative methods for parameter updating.

$$\hat{\gamma}_{t-x} = \frac{\sum_{x,t} E_{x,t} \left(\ln(m_{x,t}) - \hat{a}_x - \hat{b}_x \hat{k}_t\right)}{\sum_{x,t} E_{x,t}} \tag{6}$$

Where $E_{x,t}$ represents the number of exposed individuals, serving as a weight to enhance estimation stability. Following each cohort parameter update, k_t and b_x are adjusted until predetermined criteria are met.
The RH model exhibits three variants applicable when age adjustment yields insignificant period or cohort effects:

$$\begin{aligned} &H_0 : b_x^{(1)} = 1,\ b_x^{(2)} = 1 \\ &H_1 : b_x^{(2)} = 1 \\ &H_2 : b_x^{(1)} = 1 \end{aligned} \tag{7}$$

When $b_x^{(1)} = 1$ (it is set to 1), it corresponds to the LC model. Haberman and Renshaw proposed a simplified single-factor form of the RH model, based on the original RH model, which is the previously mentioned H_1 [8]. The specific model form is as follows:

$$\ln(m_{x,t}) = a_x + b_x k_t + \gamma_{t-x} + \varepsilon_{x,t} \tag{8}$$

The RH model used subsequently is the aforementioned single-factor model.

3 Mortality Models under Component Data Frameworks

3.1 Component Data

Component data is strictly constrained by the "part-whole" relationship, wherein the sum of the proportions of its components always equals a fixed constant (e.g., 1 or 100%). The proportion of deaths in each age group within a life table relative to the total number of deaths constitutes a typical component data vector, inherently satisfying the "sum-to-one constraint". This constraint implies interdependence among the number of deaths at specific ages; the number of deaths in each age group is not independent, and their distributions collectively form a whole with a fixed sum.

When traditional multivariate statistical methods are applied directly to compositional data, they are subject to interference from the constant-sum constraint. This leads to spurious correlations among variables and distorts both the sample space and distance metrics. The underlying reason is that compositional data intrinsically reside in the simplex space rather than Euclidean space, ultimately compromising the reliability of conventional analytical techniques such as correlation analysis and regression modeling. To address the issues described above, Aitchison proposed the use of log-ratio transformations to map compositional data from the simplex space into standard Euclidean space, thereby relaxing the constant-sum constraint and enabling the application of classical statistical methods [19]. For a compositional data vector $x = (x_1, x_2, \ldots, x_D)$ comprising D components, common transformations include:

Additive Logratio (ALR) transformation, selecting a reference component (e.g., the Dth component):

$$\text{alr}(x) = \left(\frac{\ln x_1}{x_D}, \frac{\ln x_2}{x_D}, \ldots, \frac{\ln x_{D-1}}{x_D} \right) \tag{9}$$

Centred Logratio (CLR) transformation, using the geometric mean as the reference:

$$\text{clr}(x) = \left(\frac{\ln x_1}{g(x)}, \frac{\ln x_2}{g(x)}, \ldots, \frac{\ln x_D}{g(x)} \right) \tag{10}$$

where $g(x) = \left(\prod_{i=1}^{D} x_i \right)^{\frac{1}{D}}$ denotes the geometric mean of each component.

Subsequent sections of this paper employ the centred logratio transformation for relevant components. Specifically, the age distribution vector $d_{i,t} = (d_{1,t}, d_{2,t}, \ldots, d_{t,D})$ of deaths at time is treated as a component, whose clr transformation is:

$$\text{clr}(d_{x,t}) = \left(\frac{\ln d_{1,t}}{g(d_t)}, \frac{\ln d_{2,t}}{g(d_t)}, \ldots, \frac{\ln d_{D,t}}{g(d_t)} \right) \tag{11}$$

where $g(d_t)$ denotes the geometric mean of the age composition at time t.

This transformation effectively maps component data into Euclidean space by dividing by a common geometric mean, eliminating fixed constraints and enabling free variation to satisfy classical statistical method prerequisites. Upon analysis completion, results may be interpreted back into the original component space via inverse transformation, thus concluding the entire modelling process.

3.2 CoDa-LC Model

Building upon the LC model framework, Oeppen developed the following mortality model within the compositional data (CoDa) paradigm [13]:

$$\operatorname{clr}\left(d_{x,t} \ominus a_x\right) = b_x k_t + \varepsilon_{x,t} \tag{12}$$

Here, a_x denotes the geometric mean of the age-specific death counts $d_{x,t}$, k_t represents the time effect parameter consistent with traditional LC models, and b_x is the age-specific pattern derived via singular value decomposition (SVD), which indicates the direction of mortality density transfer across different ages. $\varepsilon_{x,t}$ constitutes the random error term, while $\ominus$ is a standard CoDa operator, defined as a perturbation processing method. The parameter estimation process for this model is outlined as follows:
First, calculate the geometric mean of all deaths across the entire period:

$$a_x = \exp\left(\frac{1}{T}\sum_{t=1}^{T}\ln(d_{x,t})\right) \tag{13}$$

Subsequently, the original data undergoes centring relative to this baseline via the data perturbation procedure, yielding:

$$f_{x,t} = d_{x,t} \ominus a_x \tag{14}$$

To overcome data constraints, the centred data undergoes a transformation:

$$h_{x,f} = \operatorname{clr}(f_{x,t}) = \ln\left(\frac{f_{x,t}}{g_t}\right), \quad g_t = \left(\prod_{x=1}^{X} f_{x,t}\right)^{\frac{1}{X}} \tag{15}$$

Perform singular value decomposition on the transformed matrix $\mathbf{H}$ (whose elements are $h_{x,t}$) and retain the first principal component to capture the primary variation model within the data. Accordingly, the core parameters of the model are estimated as:

$$k_t = u_{t,1}s_1, \quad b_x = v_{x,1} \tag{16}$$

Here $u_{t,1}$ and $v_{x,1}$ denote the first left and right singular vectors respectively, and represents the largest singular value. Unlike the LC model, owing to the dual-centre property of the matrix clr , no additional constraints are required on k_t and b_x . These automatically satisfy the condition that their sum equals zero, thereby guaranteeing the uniqueness of the model.

3.3 CoDa-RH Model

Building upon the CoDa-LC model, the CoDa-RH mortality model is constructed by integrating the RH mortality model into the CoDa framework. Its core expression is:

$$\operatorname{clr}\left(d_{t,x} \ominus a_x\right) = k_t b_x + \gamma_{t-x} + \varepsilon_{t,x} \tag{17}$$

Here γ_{t-x} represents cohort effects, capturing the unique mortality risks experienced by specific generations born in different years. The interpretation of the remaining parameters is the same as in the CoDa-LC model. The initial parameter estimation steps mirror those of the CoDa-LC model, continuing until singular value decomposition is performed on $\mathbf{H}$. The first principal component is used to initialise the time effect k_t^0 and the age pattern b_x^0. Building upon this preliminary estimation, the cohort effect γ_{t-x} is initialised by calculating the mean of the residual matrix $\boldsymbol{R} = \boldsymbol{H} - b_x^{(0)}k_t^{(0)}$ across all individuals within the same birth cohort. Subsequently, an iterative algorithm alternately updates the parameters to minimise the fitting error until convergence is achieved.
The specific update formulae are as follows:

$$\hat{k}_t = \frac{\sum_x \left(\operatorname{clr}\left(d_{t,x} \ominus \hat{a}_x\right) - \hat{\gamma}_c\right)\hat{b}_x}{\sum_x \hat{b}_x^2} \tag{18}$$

$$\hat{b}_x = \frac{\sum_x \left(\text{clr}\left(d_{t,x} \ominus \hat{a}_x\right) - \hat{\gamma}_c\right) \hat{k}_t}{\sum_t \hat{k}_t^2} \tag{19}$$

At each iteration, update $\hat{k}_t$ and $\hat{b}_x$, then update the cohort effect parameter $\hat{\gamma}_(t-x)$ based on the residual matrix. Repeat this iterative process until the convergence criteria are satisfied. Given the high dimensionality of this model, which may induce parameter identifiability issues, consistent with the RH model, constraints must be imposed on the parameters after each iteration.

4 Empirical Analysis

4.1 Data

The data used in this study are obtained from the Human Mortality Database (HMD). To evaluate the generalizability of the proposed model, six countries are selected as the analysis sample, including four representative developed countries (Spain, Australia, the United Kingdom, and the United States) and two developing countries (Chile and Bulgaria). Due to differences in the starting years of the data across countries and the potential disturbance of early data by factors such as wars, this study uniformly adopts 1960 as the cutoff point, retaining only records after 1960. The age range is restricted from 0 to 110 years. In life tables, death counts at advanced ages may be zero, and zero values cannot be directly used for subsequent log-ratio transformations. To address this issue, a multiplicative replacement strategy is employed to preprocess the zero counts.

4.2 Parameter Estimation Results

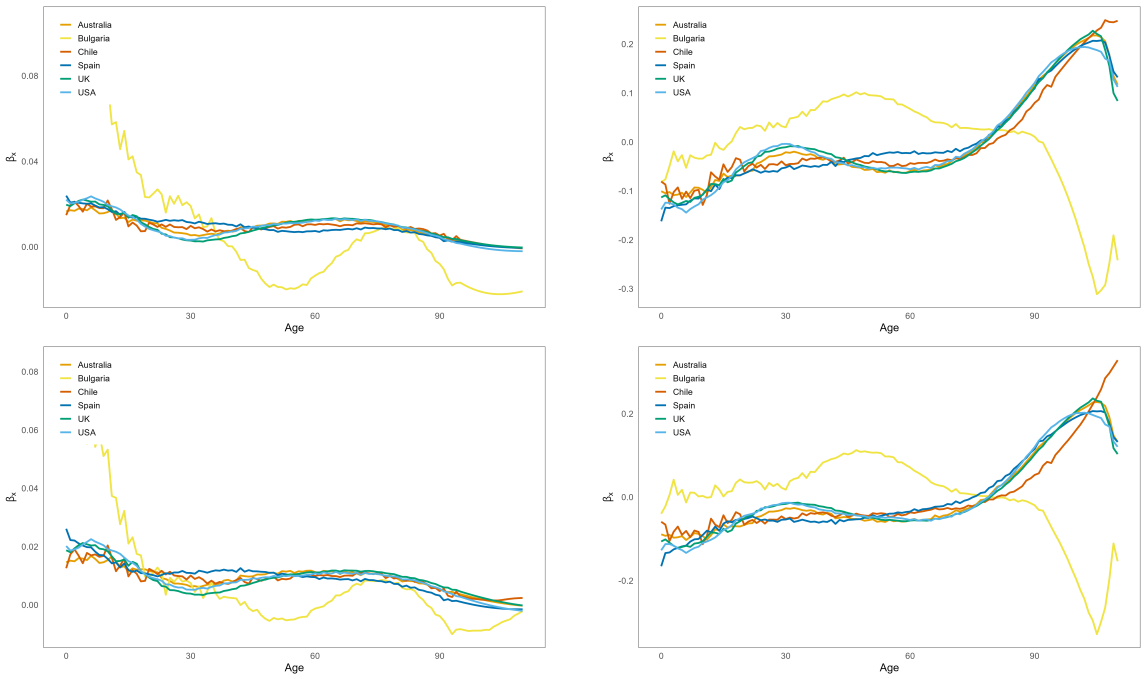

Figure 1. Model-Based Estimates of Age Parameters for Males Across Five Countries
(Top-left: LC Model, Top-right: CoDa-LC Model, Bottom-left: RH Model, Bottom-right: CoDa-RH Model)

Parameters in the traditional mortality models and those under the Compositional Data (CoDa) framework are denoted using similar symbols. While not entirely identical, their interpretations share certain parallels. As shown in Figure 1, the age effect parameter b_x captures the specific pattern of mortality variation with age. In traditional models, when multiplied by the temporal factor, it reflects the pace of mortality improvement across different age groups. Here b_x reaches its maximum at birth, indicating that male newborns and children exhibit higher mortality rates and are more sensitive to changes in the temporal factor. Within the CoDa framework, this parameter represents the transfer of mortality from one age group to another [14]. When b_x takes a negative value for a specific age group, the mortality density shifts relatively toward age groups where b_x is positive [13].

Furthermore, it can be observed that among the six countries, the age-effect parameters show broadly consistent trends across all nations except Bulgaria. Under traditional models, Bulgaria exhibits greater sensitivity to age effects, whereas under the component framework model, its age-effect parameters decline sharply at around age 70. This distinctive pattern is likely associated with Bulgaria being one of the EU countries with the most severe population aging and the fastest population decline, as its mortality age structure is increasingly concentrated at advanced ages.

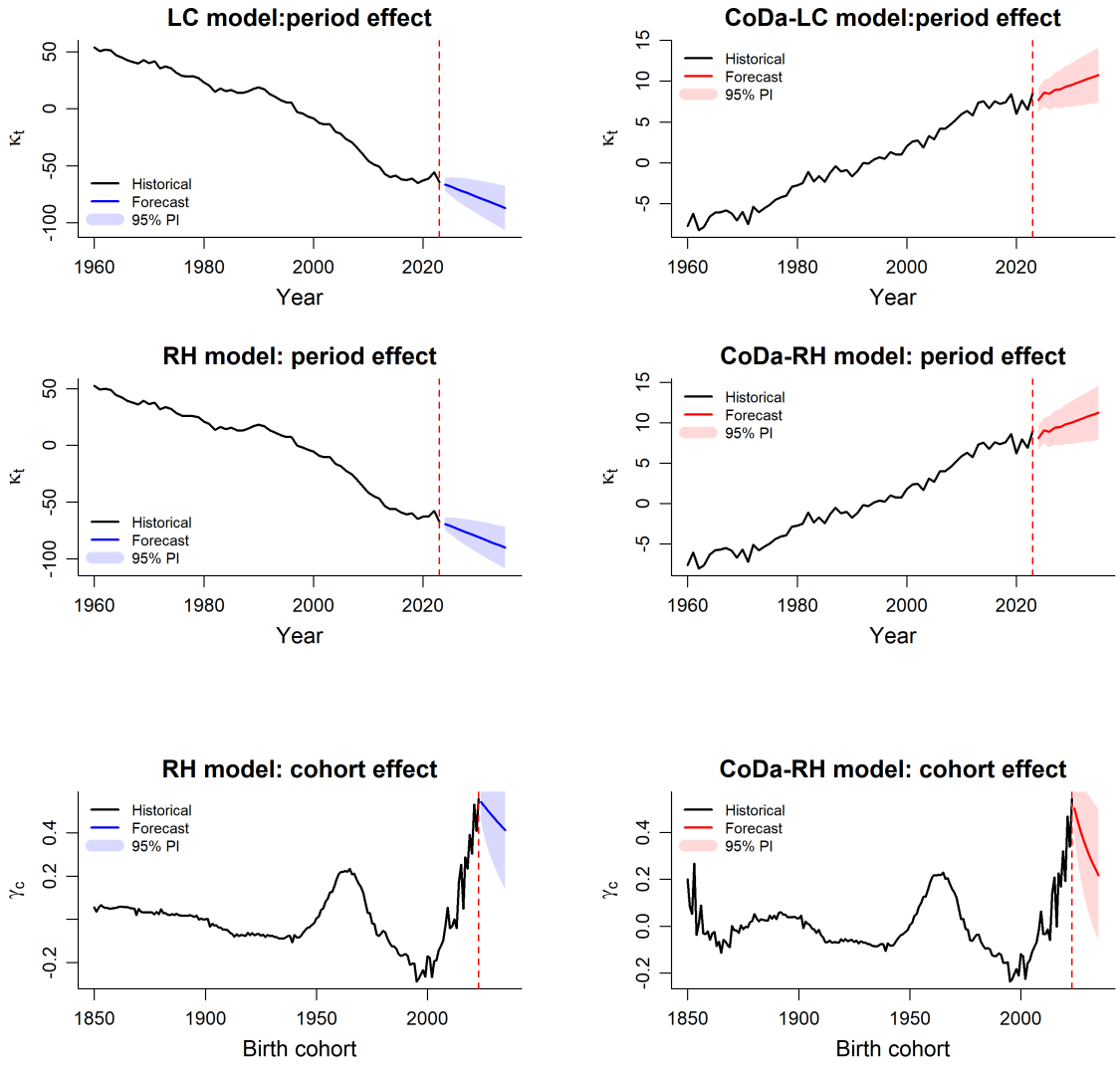

Figure 2. Predicted Period and Cohort Parameters for Spain

The parameter k_t reflects the overall change in mortality over time.As shown in Figure 2 (visualized for one randomly selected country from the six, with similar trends observed in the remaining countries), kt exhibits an approximately linear trend over the observation period. However, for traditional models, these values show a decreasing trend over time, whereas under the compositional data framework, they display an opposite pattern, potentially increasing or decreasing over time [14].

For the forecasting of the time effect k_t in the four models, the auto.arima() function from the forecast package in R was employed to automatically fit and forecast the k_t series of each model using the ARIMA methodology. This function performs automatic identification and global search to select the optimal ARIMA model structure for each series based on information criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), thereby achieving effective prediction of the time effect k_t. Furthermore, since the RH model and the CoDa RH model also include a cohort effect parameter, which reflects the long-term, slowly evolving mortality characteristics of different birth cohorts resulting from shared historical experiences, this parameter typically exhibits strong persistence and should not contain complex short-term fluctuations. If forecasted using the same method as for the time effect, random noise or measurement errors in the cohort effect series could easily be modeled as higher-order ARMA terms, causing the model to overfit incidental patterns in the sample period. Therefore, an AR(1) model with a fixed order is employed to fit and forecast this parameter.

$$\gamma_t = \phi\gamma_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d.}(0, \sigma^2) \tag{20}$$

This approach not only avoids the risk of overfitting caused by information criteria over-selecting complex models under small sample conditions, but also ensures consistent forecasting structures across populations and genders, thereby improving extrapolation stability and interpretability.

4.3 Fitting Error Analysis

In singular value decomposition (SVD), the first principal component corresponds to the direction in the data matrix that explains the maximum variance. The overall goodness-of-fit (R^2) of a model measures the proportion of variance explained when approximating the data using the first principal component. Table 1 presents the overall goodness-of-fit of the four models across the six countries.

Table 1. Goodness of Fit for Rank-1 Approximation by Model

Country	LC	CoDa-LC	RH	CoDa-RH
Spain	90.11%	92.24%	94.11%	94.89%
Chile	74.91%	81.55%	76.91%	84.83%
UK	94.48%	96.22%	96.30%	97.28%
USA	92.23%	91.72%	94.42%	94.54%
Bulgaria	60.79%	63.30%	66.20%	76.46%
Australia	92.57%	94.44%	94.64%	96.01%

Table 1 presents the overall goodness-of-fit results for the first-order (rank-1) approximation of each model. Among the six countries under investigation, the four models exhibit certain differences in their explanatory power for the dataset.

First, all models achieve relatively high goodness-of-fit values for the United Kingdom, Australia, Spain, and the United States, generally exceeding 90%. This indicates that the mortality data structures in these countries conform more closely to the rank-1 approximation assumption of the models, with clear and stable dominant patterns of period and age effects. Notably, the United Kingdom records the highest goodness-of-fit of 97.28% under the CoDa RH model, suggesting that this model possesses strong explanatory power for the variation in British mortality.

Second, the overall goodness-of-fit for the two developing countries, Bulgaria and Chile, is noticeably lower than that of the aforementioned four nations; nevertheless, the lowest goodness-of-fit among the four models still surpasses 60%. Furthermore, a cross-model comparison reveals that the CoDa models within the compositional

data framework—particularly the CoDa-RH model—achieve higher goodness-of-fit across nearly all countries, with a general improvement of 1–10 percentage points relative to conventional models. This enhancement is especially pronounced in countries with more heterogeneous data such as Bulgaria. This demonstrates that the CoDa framework based on the clr transformation better captures the intrinsic structure of mortality data, while the inclusion of a cohort effect term in the model further improves its ability to explain data variation.

The Aitchison distance(AD) is a specialized distance metric in compositional data analysis. Based on the principle of log-ratio transformation, it effectively handles the relative structure and closure effect of compositional data, providing an accurate measure of the difference between two compositional vectors. A smaller value indicates greater similarity in the internal proportional structure between the two compositional vectors.

$$AD(x,y) = \sqrt{\sum_{i=1}^{M} [clr(x_i) - clr(y_i)]^2}, \tag{21}$$

Where M denotes the dimension of the component, and $clr(\cdot)$ represents the centred logarithmic ratio transformation.

The life table data for each country were divided into training and test sets in an 7:3 ratio.

Table 2. Aitchison distances on the test set

Country	LC	CoDa-LC	RH	CoDa-RH
Spain	3.70362	3.55996	3.09005	3.06191
Chile	2.03959	1.94509	2.05006	2.01770
UK	2.22577	2.15626	1.72100	1.63336
USA	1.61089	1.72157	2.04547	1.64051
Bulgaria	10.46148	6.17955	8.49933	6.04313
Australia	2.66715	2.50722	2.12283	1.84861

The life table data of each country were split into training and testing sets at a ratio of 7:3 (post-1960 data). Table 2 presents the fitting results of various models on the testing set, along with the corresponding Aitchison distance (AD) values calculated on the testing set. It can be observed that among the six countries, except for Chile (where CoDa-LC performs slightly better), the remaining five countries achieve the lowest AD values under the CoDa-RH model. Specifically, the mean AD of the CoDa-RH model is only 2.71, which is substantially lower than those of the LC model (3.79), the CoDa-LC model (3.01), and the RH model (3.26). Notably, for Bulgaria, the most challenging country to fit, the AD of CoDa-RH is 42% lower than that of the LC model.

Table 3. Mean absolute error (MAE) of log mortality rates on the test set

Country	LC	CoDa-LC	RH	CoDa-RH
Spain	0.26753	0.25027	0.25571	0.26679
Chile	0.11438	0.10296	0.11597	0.11507
UK	0.15166	0.14919	0.10646	0.10054
USA	0.12888	0.13089	0.13584	0.11288
Bulgaria	0.31001	0.29207	0.29576	0.28545
Australia	0.18730	0.17538	0.14976	0.11516

Table 3 presents the mean absolute error (MAE) of log mortality rates calculated on the test set. It can be observed that among the six countries, CoDa-RH achieves the smallest MAE in four countries—the United Kingdom, the United States, Bulgaria, and Australia—with the most notable improvement observed for Australia. Although the

MAE of CoDa-RH is slightly higher than that of CoDa-LC for Spain and Chile, the differences are marginal. In terms of average MAE, CoDa-RH attains 0.1660, which is lower than that of CoDa-LC (0.1835), RH (0.1766), and LC (0.1933), indicating that this model has a consistent advantage in overall prediction error. Figure 3 presents a

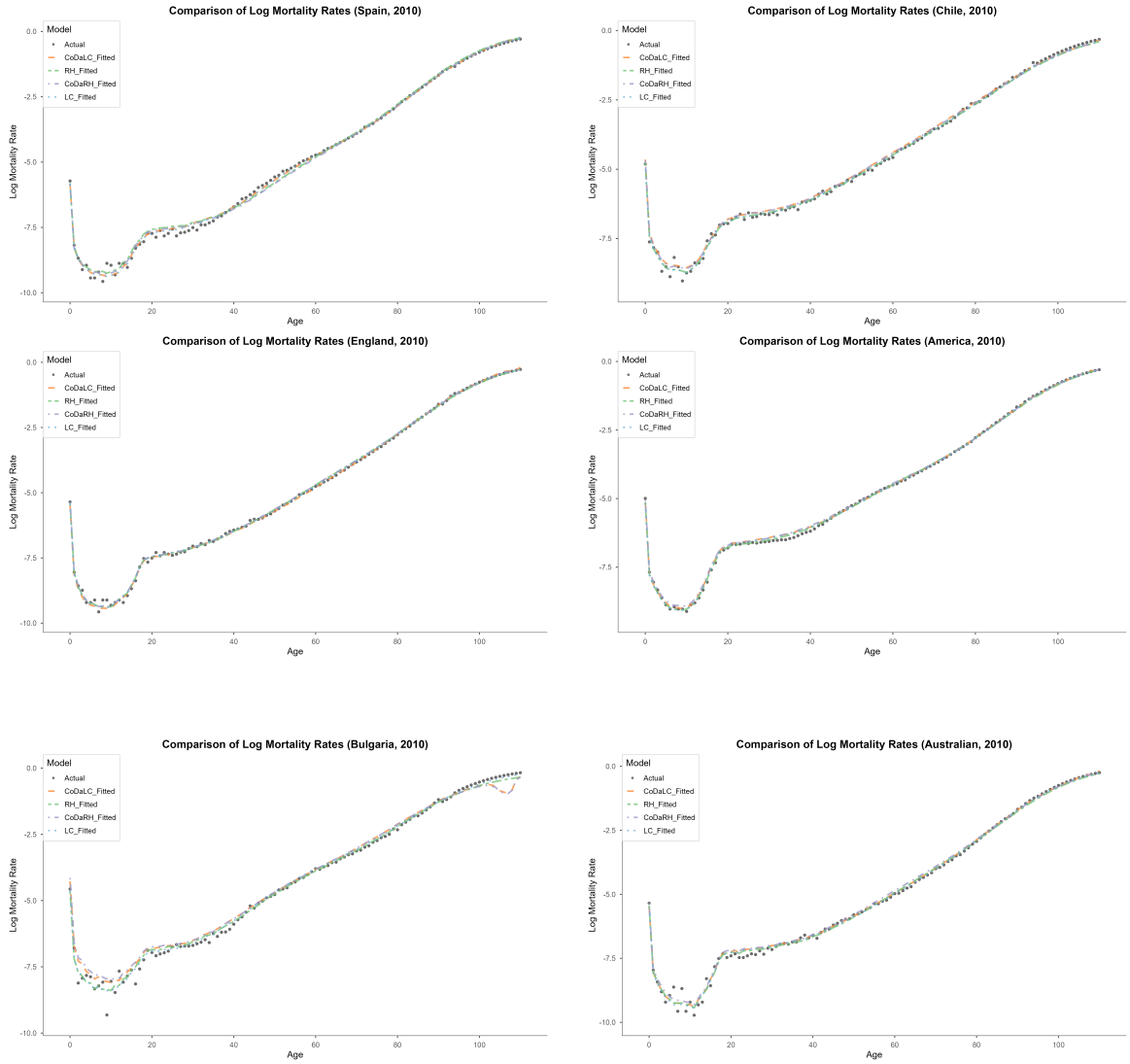

Figure 3. Comparison of Log Mortality Rates by Model, 2010

comparison between the fitted log mortality rates derived from different mortality models and the actual observed values for six countries in 2010. It can be observed that for ages below 60—i.e., among the young and middle-aged groups—the fitted curves of the models exhibit certain deviations from the observed values, with notable differences across the models. In contrast, for ages above 60 (the elderly group), the fitted curves of the four models closely coincide; all effectively capture the overall age-related increase in mortality, and the discrepancies among the models become substantially smaller.

4.4 Life Expectancy Forecasting

After fitting the four models and systematically analyzing their fitting performance, the parameter estimation and forecasting results from Section 4 can be combined to project future mortality trends. Here, a short-term forecast up to 2035 is conducted. Additionally, when forecasting mortality, constructing prediction intervals for mortality analysis is crucial for quantifying forecast uncertainty. In the quantification of mortality forecasting uncertainty, the bootstrap method has gradually become a core standard technique for building such intervals due to its efficiency. In related literature, D'Amato et al. proposed a two-stage method combining bootstrap sampling with stratified variance reduction techniques, which reduces mortality forecast uncertainty based on the LC model [21]. Bergeron-Boucher et al. employed the bootstrap method to generate confidence intervals for mortality forecasts, accounting for uncertainties in both parameter estimation and temporal index extrapolation [14]. Han proposed a nonparametric bootstrap method based on a dynamic factor model, constructing mortality prediction intervals for life tables through data transformation and a two-stage analysis [22].

Using the method described above, the core processing procedure can be summarized as follows: First, from the fitted residual matrix of the original model, 100 new residual matrices are generated by random resampling with replacement across all years and ages. Each new residual matrix is added back to the fitted values of the original model to obtain 100 bootstrap log-mortality datasets. These datasets are then re-estimated separately, yielding 100 sets of parameter estimates (a_x, b_x, κ_t). This step captures the parameter estimation uncertainty arising from random fluctuations in the historical data. Next, for each estimated time-parameter series κ_t, an optimal ARIMA model is fitted and its residuals are extracted. These ARIMA residuals are then resampled with replacement to generate, for each parameter set, 100 sequences of future random residuals, which are used to extrapolate 100 future paths of the time parameter based on the ARIMA model structure. Finally, each parameter set is combined with each of its corresponding paths and substituted into the model formula, producing a total of $100 \times 100 = 10\,000$ future mortality trajectories. Life table indicators (e.g., life expectancy at birth) are then derived from these trajectories. By taking quantiles across the 10,000 simulated outcomes, the median forecast and prediction intervals at any desired confidence level can be obtained. (For models that include a cohort effect, such as the RH model, the same ARIMA residual resampling procedure is applied to the cohort effect series to generate multiple future cohort-effect paths, which are then combined with the period-effect paths, resulting in a triple bootstrap.)

Table 4. Predicted life expectancy for male newborns in Spain (2026–2035) under four models, with 80% confidence intervals

Year	LC			CoDa-LC			RH			CoDa-RH		
	Median	LB	UB	Median	LB	UB	Median	LB	UB	Median	LB	UB
2026	81.02	80.64	81.36	81.19	80.29	81.86	81.26	80.96	81.62	81.03	80.15	81.74
2027	81.19	80.82	81.52	81.54	80.65	82.20	81.38	81.06	81.76	81.42	80.61	82.10
2028	81.30	80.95	81.63	81.63	80.70	82.30	81.51	81.21	81.89	81.45	80.57	82.19
2029	81.46	81.11	81.78	81.88	80.96	82.53	81.60	81.29	81.95	81.75	80.75	82.44
2030	81.58	81.23	81.92	82.02	81.12	82.68	81.70	81.41	82.06	81.94	81.03	82.58
2031	81.73	81.39	82.05	82.24	81.34	82.90	81.81	81.47	82.21	82.05	81.13	82.72
2032	81.86	81.51	82.19	82.40	81.51	83.06	81.91	81.65	82.31	82.20	81.38	82.92
2033	82.01	81.66	82.32	82.60	81.70	83.25	82.02	81.70	82.40	82.51	81.49	83.12
2034	82.14	81.76	82.46	82.77	81.85	83.43	82.11	81.84	82.49	82.70	81.98	83.39
2035	82.27	81.93	82.59	82.97	82.07	83.61	82.24	81.94	82.59	82.89	82.11	83.52

Table 5. Predicted life expectancy for male newborns in Chile (2026–2035) under four models, with 80% confidence intervals

Year	LC			CoDa-LC			RH			CoDa-RH		
	Median	LB	UB	Median	LB	UB	Median	LB	UB	Median	LB	UB
2026	77.60	77.13	78.03	78.01	76.57	79.00	77.50	77.17	77.89	77.73	76.38	78.75
2027	77.76	77.30	78.18	78.10	76.67	79.09	77.66	77.34	78.01	77.91	76.52	78.74
2028	77.92	77.46	78.36	78.25	76.78	79.23	77.86	77.47	78.21	78.00	76.64	78.96
2029	78.08	77.62	78.52	78.41	76.92	79.34	77.99	77.52	78.35	78.13	76.58	79.16
2030	78.25	77.78	78.68	78.53	77.07	79.50	78.19	77.78	78.52	78.29	76.93	79.28
2031	78.41	77.94	78.84	78.67	77.14	79.65	78.34	77.94	78.69	78.37	76.90	79.49
2032	78.57	78.09	78.99	78.77	77.26	79.79	78.47	78.06	78.83	78.59	77.16	79.52
2033	78.73	78.25	79.15	78.95	77.38	79.97	78.62	78.26	79.05	78.68	77.17	79.66
2034	78.88	78.41	79.31	79.07	77.38	80.13	78.79	78.33	79.16	78.79	77.29	79.85
2035	79.04	78.57	79.46	79.16	77.48	80.35	78.94	78.50	79.33	78.97	77.39	80.02

Table 6. Predicted life expectancy for male newborns in the United States (2026–2035) under four models, with 80% confidence intervals

Year	LC			Codan-lc			RH			Codan-rh		
	Median	LB	UB	Median	LB	UB	Median	LB	UB	Median	LB	UB
2026	76.11	75.78	76.43	77.43	76.88	78.07	75.66	75.32	75.95	77.00	76.37	77.74
2027	76.24	75.91	76.56	77.60	77.03	78.25	75.77	75.40	76.05	77.15	76.48	77.92
2028	76.35	76.03	76.67	77.78	77.22	78.42	75.87	75.54	76.18	77.24	76.62	77.99
2029	76.47	76.14	76.78	77.96	77.39	78.59	75.99	75.61	76.33	77.43	76.81	78.18
2030	76.59	76.26	76.91	78.13	77.55	78.77	76.10	75.72	76.36	77.54	76.97	78.23
2031	76.71	76.39	77.03	78.29	77.73	78.93	76.19	75.80	76.53	77.78	77.12	78.52
2032	76.83	76.52	77.15	78.47	77.91	79.11	76.29	75.94	76.61	77.88	77.19	78.63
2033	76.95	76.62	77.27	78.64	78.07	79.28	76.41	76.04	76.72	78.13	77.39	78.88
2034	77.07	76.73	77.39	78.81	78.25	79.45	76.50	76.18	76.82	78.25	77.54	79.05
2035	77.19	76.85	77.50	78.99	78.42	79.02	76.62	76.29	76.92	78.50	77.79	79.28

Tables 4, 5 and 6 present the predicted life expectancy at birth and their 80% confidence intervals for male newborns in Spain, Chile and the United States over the 2026–2035 period, based on four mortality models. The forecasting results across the three countries show that life expectancy estimates derived from models under the compositional data framework are generally higher than those from traditional models, and their prediction intervals are also consistently wider. This is mainly attributed to the greater variability in the time index of CoDa models, which enables them to capture more uncertainty and thus provide more conservative yet information-rich interval estimates.

Here is a detailed analysis of the prediction results for Spain. According to the official forecast released by the Spanish National Statistics Institute (INE) for the period 2026–2035, the life expectancy at birth for male newborns in the country will range between 81.08 and 82.50 years [23]. Over the entire forecast horizon, the LC model shows some underestimation relative to the official interval; the RH model increases from 81.26 to 82.24 years, exhibiting a relatively conservative growth. Under the compositional data framework, both models—CoDa-LC and CoDa-RH—generally yield higher median estimates. In particular, by the end of the forecast period in 2035, their medians reach 82.97 and 82.89 years, respectively, which are significantly higher than those of the traditional LC and RH models (82.27 and 82.24 years). Moreover, in 2035, the predicted value of the CoDa-RH model is closer to the official forecast. This systematic shift suggests that the CoDa approach, through its holistic constraint on the age distribution of mortality, may capture the non-linear acceleration of mortality decline at younger ages in recent years, thereby extrapolating a steeper path of life expectancy improvement. At the same time, the 80% confidence intervals of the CoDa models are noticeably wider. This expansion of the intervals reflects a more complete propagation of random error and sampling variability from the compositional data perspective, and essentially

represents a more honest expression of the two-sided uncertainty surrounding future mortality improvements.
For the United States and Chile, according to life table data from the Human Mortality Database, the life expectancy at birth for male newborns in 2023 was 75.99 years in the United States and 77.78 years in Chile. In the projections for the United States, the median of the RH model for 2026 is only 75.66 years, which falls below the 2023 actual level, indicating a clear underestimation; the median of the LC model is 76.11 years, also lower than the predictions from the compositional data models. This underestimation tendency of the two traditional models is also observed in Chile, where their 2026 projected values are both lower than the country's actual life expectancy in 2023.
In summary, mortality models developed from the CoDa perspective can effectively mitigate the conservative estimation bias commonly observed in traditional benchmark models such as LC and RH, demonstrating high reliability and favorable applicability in short-term life expectancy forecasting. Furthermore, the wider prediction intervals obtained via bootstrap simulation should not be regarded as a model deficiency. Instead, they offer a more cautious and informative basis for formulating population policies, designing social security systems, and assessing long-term risk reserve requirements.

5 Conclusions and Outlook

This study introduces a compositional data framework based on the traditional RH model and proposes a new mortality forecasting method. The new model not only effectively captures period and cohort effects in historical data but also ensures the logical consistency of the predicted mortality distribution by leveraging the inherent constraints of compositional data. In terms of forecasting accuracy, the model shows certain advantages over conventional metrics. Future research may consider incorporating more countries with complete and high-quality mortality data to construct multi-country models, thereby enhancing the generalizability of the model structure.In addition, more complex neural network models, such as LSTM or Transformer, could be explored to forecast the cohort and period effect parameters and capture their potential nonlinear dynamics.

Acknowledgement

The authors would like to express their sincere gratitude to the Editors and anonymous reviewers for their valuable comments and suggestions, which have greatly improved this paper.This work was supported by the National Natural Science Foundation of China (Grant No. 12061066) and the Natural Science Foundation of Gansu Province (Grant No. 20JR5RA528).

Appendix

Zero-value treatment Construction of the Mortality Matrix:

$$D_{t,x} = \begin{bmatrix} d_{1,0} & d_{1,1} & \cdots & d_{1,\omega} \\ d_{2,0} & d_{2,1} & \cdots & d_{2,\omega} \\ \vdots & \vdots & \ddots & \vdots \\ d_{T,0} & d_{T,1} & \cdots & d_{T,\omega} \end{bmatrix}$$

where t denotes the year, ω denotes the age, and $d_{t,x}$ denotes the number of deaths in year t for individuals aged x.

Calculate annual total deaths and convert to a ratio matrix:

$$K_t = \sum_{x=1}^{\omega} d_{t,x}, \qquad t = 1, 2, \ldots, T$$

$$p_{t,x} = \frac{d_{t,x}}{K_t}$$

Zero-value detection and parameter calculation
For each year t, compute the number of zero values:

$$z_t = \sum_{x=1}^{\omega} \mathbb{1}(p_{t,x} = 0)$$

Calculate the global minimum positive proportional value:

$$P_{\min} = \min_{t,x}\{p_{t,x} : p_{t,x} > 0\}$$

Calculate replacement parameters:

$$\delta = \frac{P_{\min}}{2}$$

Multiplication Substitution Algorithm
Apply the multiplier substitution formula:

$$r_{t,x} = \begin{cases} \delta & \text{if } p_{t,x} = 0 \\ (1 - z_t\delta)\, p_{t,x} & \text{if } p_{t,x} > 0 \end{cases}$$

Post-processing and verification
Post-processing verification:

$$\sum_{x=1}^{\omega} r_{t,x} = z_t\delta + (1 - z_t\delta) \cdot \sum_{x:p_{t,x}>0} p_{t,x} = 1$$

Convert back to death count:

$$d_{t,x} = r_{t,x} \cdot K_t$$

REFERENCES

1. United Nations. World Population Prospects 2024: Summary of Results, 2024. Available at https://digitallibrary.un.org/record/4053940?v=pdf.
2. Lee, R. D., and Carter, L. R. Modeling and Forecasting U.S. Mortality, Journal of the American Statistical Association, vol. 87, no. 419, pp. 659–671, 1992.
3. Brouhns, N., Denuit, M., and Vermunt, J. K. A Poisson log-bilinear regression approach to the construction of projected lifetables, Insurance: Mathematics and Economics, vol. 31, no. 3, pp. 373–393, 2002.
4. Renshaw, A. E., and Haberman, S. A cohort-based extension to the Lee–Carter model for mortality reduction factors, Insurance: Mathematics and Economics, vol. 38, no. 3, pp. 556–570, 2005.
5. Currie, I. D., Durban, M., and Eilers, P. H. C. Generalized linear array models with applications to multidimensional smoothing, Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 68, no. 2, pp. 259–280, 2006.
6. Cairns, A. J. G., Blake, D., and Dowd, K. A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration, Journal of Risk and Insurance, vol. 73, no. 4, pp. 687–718, 2006.
7. Cairns G J A ,Blake D ,Dowd K , et al.A Quantitative Comparison of Stochastic Mortality Models Using Data From England and Wales and the United States[J].North American Actuarial Journal,2009,13(1):1-35.
8. Haberman, S., and Renshaw, A. A comparative study of parametric mortality projection models, Insurance: Mathematics and Economics, vol. 48, no. 1, pp. 35–55, 2010.
9. Plat, R. On stochastic mortality modeling, Insurance: Mathematics and Economics, vol. 45, no. 3, pp. 393–404, 2009.
10. R. S K ,Michael S ,M. A V , et al.Mortality forecasting using stacked regression ensembles[J].Scandinavian Actuarial Journal,2022,2022(7):591-626.
11. Jun W ,Lihong W ,Lu X , et al.Time-series forecasting of mortality rates using transformer[J].Scandinavian Actuarial Journal,2024,2024(2):109-123.
12. Cheng R ,Shi J ,Loh M J , et al.Neural networks for simultaneous modeling of multi-population mortality with coherent forecasts[J].Scandinavian Actuarial Journal,2025,2025(9):853-882.
13. Oeppen, J. Coherent forecasting of multiple-decrement life tables: A test using Japanese cause of death data, In European Population Conference 2008, Barcelona, Spain, July 9–12, 2008.
14. Bergeron-Boucher, M., Canudas-Romo, V., Oeppen, J., and Vaupel, J. W. Coherent forecasts of mortality with compositional data analysis, Demographic Research, vol. 37, pp. 527–566, 2017.

15. Li, N., and Lee, R. D. Coherent mortality forecasts for a group of populations: An extension of the Lee–Carter method, Demography, vol. 42, no. 3, pp. 575–594, 2005.
16. Boucher B P M ,Strozza C ,Simonacci V , et al.Modeling and Forecasting Healthy Life Expectancy With Compositional Data Analysis.[J].Demography,2025
17. Stefanucci, M., and Mazzuco, S. Analysing cause-specific mortality trends using compositional functional data analysis, Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 185, no. 1, pp. 61–83, 2021.
18. Guo, Y., and Li, J. S.-H. Fast estimation of the Renshaw-Haberman model and its variants, European Actuarial Journal, vol. 15, no. 2, pp. 633–666, 2025.
19. Booth, H., and Tickle, L. Mortality Modelling and Forecasting: A Review of Methods, Annals of Actuarial Science, vol. 3, no. 1-2, pp. 3–43, 2008.
20. Aitchison, J. The Statistical Analysis of Compositional Data, Chapman and Hall, London, 1986.
21. D'Amato, V., Haberman, S., and Russolillo, M. The Stratified Sampling Bootstrap for Measuring the Uncertainty in Mortality Forecasts, Methodology and Computing in Applied Probability, vol. 14, no. 1, pp. 135–148, 2012.
22. Shang, H. L. Bootstrap prediction intervals for the age distribution of life-table death counts, Mathematical Population Studies, vol. 32, no. 3, pp. 166–181, 2025.
23. Instituto Nacional de Estadística (INE). Projected Mortality Tables 2024-2073: Life expectancy by age and sex, 2024. Report No. 36775.