Malaysian Journal of Computing, 11 (2): 2578-2599, 2026
eISSN: 2600-8238

MODELLING MALAYSIAN MORTALITY WITH CAUSE-OF-DEATH: A COMPOSITIONAL DATA APPROACH TO LEE-CARTER EXTENSION

Saiful Azril Ishak[1], Norazliani Md Lazam[2*], Shamshimah Shamsuddin[3], Nurin Haniah Asmuni[4] and Nordiana Rosdi[5]

[1,2*,3,4,5]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia*

[1]saifulazril@uitm.edu.my, [2*]norazliani@uitm.edu.my, [3]shams611@uitm.edu.my, [4]nurinhaniah@uitm.edu.my, [5]nordianarosdi@uitm.edu.my

(indicates the corresponding author)*

ABSTRACT

Forecasting cause-specific mortality is at the core of actuarial valuation, annuity pricing, and long-term public health planning. In traditional extrapolation frameworks such as the Lee-Carter and Li-Lee models, probability leakage is a structural flaw which causes the independently projected cause death rates to not sum to total mortality and results in implausible trajectory crossovers. To overcome these shortcomings, this study explores the application of Compositional Data Analysis (CoDA) extensions namely CoDA Lee-Carter (CDLC) and CoDA Li-Lee (CDLL) to historical Malaysian cause of death data. We transform cause proportions to the unit simplex using centred log-ratio coordinates. This transformation imposes exact aggregate sum coherence but preserves competing-risk substitution dynamics. Short-term structural changes are captured by standard CDLL, but its unweighted common-factor structure overfits noise in the volatile "Others" residual category, extrapolating temporary administrative changes into an implausible expansion of aggregate mortality over multi-decade horizons. We propose a Weighted CoDA Li-Lee (WCDLL) framework with inverse-variance weights to mitigate this log-ratio sensitivity. Baseline noise is buffered by lower structural weights on high volatility residual components, protecting the common temporal index. Empirical evaluation verifies that WCDLL eliminates aggregate trajectory divergence, recovering a demographically plausible secular decline consistent with national all-cause benchmarks, while maintaining exact unit-simplex sum-coherence. For multi-cause mortality modelling, the suggested WCDLL model offers an actuarially sound and mathematically consistent framework.

Keywords: *Cause Specific Mortality, Coherent Mortality Forecasting, Compositional Data Analysis, Lee-Carter Extension, Mortality Modelling.*

Received for review: 03-09-2026; Revised: 25-09-2026; Accepted: 25-09-2026; Published: 01-10-2026
DOI: 10.24191/mjoc.v11i2.13718



This is an open access article under the CC BY-SA license (https://creativecommons.org/licenses/by-sa/3.0/).

1. Introduction

Understanding how mortality changes over time has become more and more important as ageing populations and changing health profiles become more relevant. Mortality projections have traditionally depended on models such as the Lee and Carter (1992) and Li and Lee (2005) frameworks that estimate future death rates based on aggregate mortality data. While widely adopted, these approaches often overlook cause-specific mortality dynamics, which can provide critical insights into the drivers of mortality changes across populations. In particular, aggregate models are limited in their ability to account for the heterogeneity of mortality trends arising from different diseases or external causes, potentially leading to over-simplified or misleading conclusions.

The role of cause-of-death (CoD) information has become more prominent in mortality research, especially as countries experience epidemiological transitions, where the burden of disease shifts from infectious diseases to chronic, non-communicable conditions such as heart disease, cancer, and diabetes. Malaysia is no exception to this trend. For instance, between 2010 and 2016, the country recorded a significant decline in HIV-related mortality, while deaths from diabetes and cancer increased considerably. The most recent statistics show that ischemic heart disease and pneumonia continue to be the leading causes of death, but long-term increases in deaths due to cancer and diabetes have been noted, owing to demographic changes following the pandemic (DOSM, 2025). These shifting patterns highlight the limitations of aggregate mortality models and point to the value of integrating cause-specific mortality data into forecasting frameworks. Such integration can enhance the accuracy and interpretability of mortality projections and provide a clearer understanding of the factors driving mortality improvements.

However, there are methodological difficulties with modelling cause-of-death data. There are primarily two issues: competing risks, where a person can die from just one cause, making each cause mutually exclusive, and incoherence, where the sum of the individual cause-specific death rates does not match the observed overall mortality. Ignoring these complications can lead to projections that are skewed and mortality estimates that are biased. Furthermore, traditional statistical models frequently portray cause-specific death rates as independent, ignoring the compositional structure of the data, which means that each cause-specific rate must be interpreted in relation to the others because it represents a portion of the total.

In this study, we tackle these issues using Compositional Data Analysis (CoDA), a statistical framework for modelling data representing pieces of a whole. CoDA enables one to model the relative proportions of fatalities by cause in the setting of mortality ensuring that the sum is meaningful and the relationships between causes are preserved. We employ CoDA to transform cause-specific mortality data into log-ratio coordinates for statistical modelling in their study. These coordinates are then used in modified versions of the Li-Lee and Lee-Carter models. This permits consistent cause-specific death forecasts to be produced that add up to overall mortality estimates. Integrating CoDA with conventional mortality frameworks has multiple advantages. First, it improves the internal coherence and consistency of the forecasts by linking cause-specific projections to aggregate mortality. Second, it accounts for the competing risks in CoD data by considering the proportions of the causes as parts of a whole, instead of treating them as independent quantities. Third, cause-specific coherence is essential for actuarial pricing and pension valuation, since these calculations depend on precise aggregate survival probabilities. If independent cause projections suffer from probability leakage then the combined mortality rate may be systematically over or under-estimated. This incoherence leads to mispriced life insurance premiums, insufficient reserving and more solvency risk for long-term demographic products.

In this paper, we analyse the Malaysian mortality data and investigate the performance of models enhanced by CoDA relative to the original Lee-Carter and Li-Lee models to improve the forecasting performance. The research shows that the integration of cause-specific death

structure within popular mortality models results in a more robust and flexible approach, able to adapt to changes in mortality trends. The results are expected to contribute to methodological developments in mortality modelling as well as to practical improvements in demographic and actuarial applications, providing tools more adapted to a world where mortality is more and more driven by diverse and changing causes of death.

2. Background of Studies

Mortality forecasting is crucial for actuarial risk management, public health policy and pension planning. General mortality trends are described by aggregate stochastic models; however, disaggregation to causes of death (CoD) reveals significant underlying health transitions (Ishak *et al*., 2025). Modelling specific causes, however, is methodologically challenging because of competing risks and mathematical incoherence when projections of individual causes do not sum to overall mortality rates (Wilmoth, 1995; Shang & Haberman, 2024). In the following subsections, standard stochastic mortality models are reviewed, the evolving demographic and CoD transition landscape in Malaysia is discussed, and the applicability of Compositional Data Analysis (CoDA) as a coherent multi-cause forecasting solution is examined.

2.1 Malaysia Cause-of-Death Mortality

Understanding dynamics of cause-of-death (CoD) is important for assessing national health transitions and forecasting longevity. Mortality trends in Malaysia have undergone an epidemiological transition over the past few decades, moving from infectious and communicable diseases to non-communicable diseases (NCDs) such as ischaemic heart disease, malignant neoplasms and diabetes mellitus. Department of Statistics Malaysia (DOSM, 2025) data show that circulatory and cardiovascular conditions consistently represent the leading cause of death, alongside growing contributions from respiratory illnesses and cancers.

Prior Malaysian mortality studies were largely based on either overall mortality rates or single-cause trends separately. However, Ishak *et al.* (2025) proved that the application of univariate Lee-Carter models on the disaggregated Malaysian CoD data without structural constraints can distort all-cause mortality projections which is also in line with Wilmoth's (1995) findings on issue of disintegrated data summation in mortality forecasting. Trends in leading causes of mortality, including declining rates of transport accidents and increasing proportions of NCDs (DOSM, 2025), illustrate the limitation of modelling causes in isolation. Competing risks are a consideration in causes of death, where a change in one cause directly impacts the relative magnitude of other causes, because an individual can only die from one cause at a time.

Recent applications have demonstrated the efficacy of multi-population models in capturing varied lifestyle impacts across different demographic cohorts (Varga, 2025). One established approach involves modelling each cause independently, forecasted individually and combined together. Research by Lakman *et al.* (2021) employed the Lee-Carter model for independent estimation of specific causes of death, rather than relying on a single mortality index. However, mortality rate forecasting was always pessimistic when disaggregated by cause of death (Wilmoth, 1995).

2.2 The Lee-Carter and Li and Lee Models

The Lee-Carter model is a fundamental and popular statistical method for predicting mortality and longevity. When Lee and Carter first presented it in 1992, it was a major methodological advancement in demography, offering a straightforward, reliable, and data-driven stochastic approach for modelling and projecting age-specific mortality rates over time. The model looks at four main components to reflect the logged central death rate at age x in year t. First, the model examines the overall pattern of death by age. The subsequent element, which accurately

represents the total level of mortality at time t, is the age-specific sensitivity to temporal variations in mortality multiplied by the time-varying mortality index (Basellini *et al*., 2023). It is assumed that the final component of the residual error term, which accounts for extra variability, has a mean of zero.

This model achieves a high level of dimensionality reduction in which complex mortality patterns across ages and years are captured by a dominant time-varying index (κ_t) and two age-related profiles. It is purely data-driven and demands no subjective expert judgement, requiring only minimal subjective judgement in the estimation. The LC model can offer probabilistic forecasts by incorporating time-series methods, allowing for the estimation of prediction intervals for various mortality measures. But there are limitations to the model. Limitations include possible misfits at some ages, oversimplification in the sensitivity profile and a tendency for mortality forecasts for similar populations to diverge unrealistically over the long term. These have led to a large number of extensions including the addition of cohort effects, multi-population coherent approaches such as the Li-Lee model, alternative smoothing methods for parameter estimation, and versions based on Poisson or Bayesian formulations (Basellini *et al*., 2023). Lately, scholars have expanded these frameworks by incorporating fuzzy logic to address demographic uncertainty (Koissi & Shapiro, 2006) and utilizing deep learning convolutional autoencoders for detailed forecasting (Tanaka & Matsuyama, 2025). Further recent extensions include the incorporation of exogenous environmental and lifestyle predictors in multi-population architectures (Dimai, 2025).

Another extension of the traditional Lee-Carter model that may be applied to several populations is the Li-Lee model. It is a model that enhances the Lee-Carter model by adding the common factor of overall mortality trends in the model to solve the long-term divergence problem (Li & Lee, 2005). The primary drawback of the Lee-Carter is addressed by this extension. Although the Lee-Carter model is effective at predicting mortality in a single population, when applied independently to several populations, it frequently produces long-term, implausible, divergent mortality patterns that are at odds with observed tendencies of convergence across similar populations (Varga, 2025). Populations that are closely related, such as those that share socioeconomic, environmental, or medical characteristics, should not exhibit infinite mortality variation. This is taken into consideration in a coherent model, which increases accuracy and realism.

2.3 The Lee-Carter and Li and Lee Models

Compositional Data Analysis (CoDA), initiated by Aitchison (1986), resolved these limitations by treating cause-of-death fractions as relative proportions of a closed total. Instead of evaluating raw death rates in Euclidean space, CoDA uses logarithmic transformations such as the Centred Log-Ratio (CLR) to map constrained compositional proportions into unconstrained real space. Compositional data consist of vectors of positive values where the total sum holds no absolute meaning, and the key information lies in the ratios between components. Because of the constant sum requirement and the interdependence it creates across components, applying conventional statistical techniques directly to the raw component values may result in inaccurate or erroneous findings (Greenacre, 2021). CoDA analysis allows for rigorous, interpretable, and statistically valid analysis by focusing on relation information such as the ratios between parts.

Some issues that may arise when using standard statistical analysis on raw proportions are misleading correlation, spurious correlation, biased regression coefficient, and many more. The CoDA analysis will resolve these by working on an appropriate relative scale combined with transformations that account for the mathematical structure of the data. CoDA analysis is widely used across multidisciplinary when data are compositional by nature. Classic applications include allocation of time or resources among competing activities or groups in sociology or behavioral sciences, budget shares, financial ratios or market share distributions in economics and finance (Greenacre, 2021), and movement behaviors in health and time-use epidemiology (Dumuid *et al*., 2020).

CoDA analysis can be effectively integrated with mortality models to analyze and forecast mortality rates across causes of death. This integration leverages the relative nature of mortality components such as age-specific death counts or cause-specific mortality proportions, which naturally form compositional data, as they sum to a constant total and convey information primarily through their ratios rather than their absolute values. Shang and Haberman (2024) show that direct application of traditional statistical models to compositional mortality data can lead to biased or misleading conclusions due to the constant-sum constraint and the relative importance of information between categories. CoDA provides the appropriate mathematical and statistical toolbox to deal with such constraints by using ratios between components, reducing spurious correlations, and ensuring interpretable outcomes.

Bergeron-Boucher *et al.* (2017) demonstrated that principal component modelling within the CoDA framework can accurately capture trends in the age-at-death distribution and is robust to the compositional nature of the data. CoDA can be used to analyze the proportionate structure and trends of cause of death data by modelling the cause specific death rates for each age or period as a composition. Furthermore, recent compositional innovations have successfully addressed the challenge of zero counts in fine-grained mortality datasets by utilizing advanced power transformations (Dong *et al*., 2025). This innovative approach, rooted in the statistical framework for analyzing data that represents parts of a whole, offers a distinct advantage over traditional methods when dealing with compositional cancer data.

This study builds upon the work of Bergeron-Boucher *et al.* (2017) and Ishak *et al.* (2025) by incorporating CoDA transformations directly into modified Lee-Carter (CDLC) and Li-Lee (CDLL) frameworks. The models are built by treating the Malaysian cause-of-death distributions as an interrelated compositional system, thereby maintaining the internal consistency of the models, accounting for competing health risks, and coherently relating cause-specific projections to aggregate mortality forecasts.

3. Data and Method

Causes of death are typically classified according to the International Classification of Diseases (ICD). This system is widely used across data sources to assign causes of death and to code and classify information from death certificates. Since different data providers and countries may employ varying categorizations of causes of death, the World Health Organization, together with ten international centers, developed the ICD to ensure that causes of death are systematically classified and standardized for the international reporting of mortality statistics.

For this study, we utilized data on age, population and cause-specific mortality in Malaysia from 2000 to 2019, sourced from the Department of Statistics Malaysia (DOSM). The observation period is limited to 20 years as these are the only detailed cause-specific data currently available from DOSM. The risk of a precarious 1:1 ratio in conventional unconstrained time-series forecasting, which arises when projecting a 20-year horizon (2020–2039) from a 20-year baseline, is mitigated by the structural stability of the CoDA framework. The unit-simplex closure of CoDA makes the exponential divergence of long horizons a natural mitigation of the typical criticism of classical models trained on limited historical windows. The dataset is structured by age groups in interval form (e.g., 0, 1–5, 6–10, . . . , 95+), which is applied in the estimation of cause-specific mortality rates. For the construction of life tables, the radix was set at 100,000 at age 0, representing a hypothetical cohort of individuals alive at the beginning of life.

The selection and grouping of the 9 leading causes of death in Malaysia follow a methodology similar to that used in studies by Arnold and Sherris (2015) and Varga (2025). In this approach, the 9 main cause-of-death groups that account for more than 80% of total deaths combined are chosen. For the Malaysian dataset, this particular nine-category classification offers the best possible balance between mathematical stability and epidemiological detail.

Adding more rarer causes to the classification runs the risk of creating zero-count sparsity, which essentially destabilizes log-ratio transformations. In contrast, a more aggregated grouping would obscure the divergent paths between chronic non-communicable diseases like neoplasms and circulatory conditions versus the very volatile external causes.

Table 1. Cause of Death Categories

Cause of Death Categories	Example of Common Diseases
Circulatory	Ischemic heart disease, cerebrovascular disease
Respiratory	Pneumonia, influenza, respiratory diseases like asthma and bronchitis
Neoplasms (cancer)	Lung, colon, breast, prostate and other related diseases to cancer
Infectious	Gastroenteritis, tuberculosis, HIV, Viral fever, bacteria diseases
Septicaemia	Death due to Septicaemia
Transport accident	Death due to accident involving vehicles
External	Death due to injury other than transport accident
Diabetes	Death due to diabetes
All other diseases	Death due to other causes than above

3.1 The Lee-Carter Model

The Lee and Carter (1992) model specifies that the natural log of central mortality at age x, has a linear relationship with age x and year t. The equation describing this is as below:

$$ln(m_{x,t}) = \alpha_x + \beta_x \kappa_t + \epsilon_{x,t} \quad (1)$$

where $m_{x,t}$ is the central death rate with age x in year t, α_x is the average age of log mortality, β_x is the sensitivity of the log-mortality to changes κ_t. The κ_t is the measure of mortality index and $\epsilon_{x,t}$ is the error term. To ensure identifiability, the parameter β_x and κ_t are estimated via Singular Value Decomposition (SVD). They are derived from the first left singular vector u_t, the first right singular vector v_x and the largest singular value, s as follows:

$$\kappa_t = u_t s \quad (2)$$

and

$$\beta_x = \frac{v_x}{\sum v_x} \quad (3)$$

As suggested by Lee and Carter (1992), a random walk with drift can be used to forecast the κ_t.

3.2 The Li and Lee Model

The Li and Lee (2005) model is an extension of the Lee and Carter (1992) model that addresses the issue of incoherent mortality forecasts when modelling multiple related populations or sub-groups. The Li and Lee model introduces a shared mortality trend across all populations where

in this research we use the total mortality rates as the benchmark in modelling the cause of death mortality. The model is specified as follows:

$$ln(m_{x,t,j}) = \alpha_{x,j} + \beta_x \kappa_t + b_{x,j}\kappa_{t,j} + \epsilon_{x,t,j} \quad (4)$$

where $m_{x,t,j}$ is the mortality rate at age x, year t and cause j, $\alpha_{x,j}$ is the average log mortality at age x and cause j, the $\beta_x \cdot \kappa_t$ is the common factor for all causes of death which can be found by applying the Lee-Carter model to the average mortality for all the causes. The $b_{x,j} \cdot \kappa_{t,j}$ is the same as in the Lee-Carter model but only for the cause of death j and $\epsilon_{x,t,j}$ is the error term for the cause of death j.

3.3 Compositional Data Analysis Mortality Model

The transformation of mortality data by causes of death follows the foundation laid out by Aitchison (1986) and Greenacre (2021). The cause-specific mortality rates can be derived from all-cause mortality rates where the number of deaths for each cause j at age x and time t can be represented by $d_{x,t,j}$ (Bergeron-Boucher *et al*., 2017). We analyze nine primary cause of death categories to produce the life table of death of $d_{x,t,j}$ with an $N \times DJ$ matrix. The rows in the life table are still in compositions and can be summed to the radix in which this case we choose 100,000 as the radix. Thus, the total number of deaths represented by the number of deaths by causes of death can be defined as:

$$d_{x,t} = \sum_{j=1}^{D} d_{x,t,j} \quad (5)$$

where $d_{x,t}$ denotes the total all-cause death counts. The death data are then transformed into composition components where the cause of death data will be used to formulate the composition structure as $d_{x,t}$. We then find the centered log-ratio transformation of $d_{x,t}$ (Dumuid *et al*., 2020) by taking the log of each death count as:

$$clr(d_{x,t,j}) = log\left(\frac{d_{x,t,j}}{g(x,t)}\right) \quad (6)$$

with $g(x,t)$ is the geometric mean by the age of year t of the number of death such:

$$g(x,t) = \left(\prod_{j=1}^{D} d_{x,t,j}\right)^{\frac{1}{D}} \quad (7)$$

The clr is a centered log-ratio where it can be used to represent a composition as a real vector so that standard multivariate statistical can be used in the analysis. This operator is used to ensure that the matrix can retain its constant summation. The composition of the range of selected causes of death and coherence are analyzed in terms of their mean vector, covariance and variation matrix (Aitchison, 1986).

3.4 CoDA Lee-Carter (CDLC) and CoDA Li-Lee (CDLL)

The construct of CoDA Lee-Carter is by first constructing the $N \times DJ$ matrix of $d_{x,t,j}$ where N and DJ represent the row in the mortality table. The forecasting of $d_{x,t,j}$ is based on Principal Component Analysis similar to Lee-Carter but adopted within the CoDA framework for the number of deaths where:

$$clr(d_{x,t,j} \ominus \alpha_{x,j}) = b_{x,j}\kappa_{t,j} + \epsilon_{x,t,j} \quad (8)$$

where $\alpha_{x,j}$ is the age-specific geometric mean of the $d_{x,t,j}$ over time, while the time index $\kappa_{t,j}$ and age pattern $b_{x,j}$ can be found by using SVD. The $\ominus$ is an operator in the CoDA framework

and defined as perturbation procedure composition by the inverse element of another composition. The parameter $\kappa_{t,j}$ and $b_{x,j}$ are estimated on first left singular vectors of years and first right singular vectors ages, respectively according to low-rank approximation. Once the parameters are estimated, the $d_{x,t,j}$ can be found by:

$$d_{x,t,j} = \alpha_{x,j} \oplus \mathcal{C}\left[\exp(b_{x,j}\kappa_{t,j} + \epsilon_{x,t,j})\right] \tag{9}$$

where using $C[]$ as the closing procedure to transform the estimated value into compositional data by ensuring each composition entry is multiplied by the number of survivors at each age so that it meets the restriction of death that sums into a constant. The CoDA Lee-Carter model can be extended by including the common factor (Li & Lee, 2005). The common factor is the same as in Lee-Carter β_x and K_t where it can be used to find the average mortality trends for all causes of death. We follow the CoDA coherence population model proposed by Bergeron-Boucher *et al.* (2017) but modified to causes of death, the model can be written as:

$$clr(d_{x,t,j}) \ominus \alpha_{x,j} \ominus \mathcal{C}[\exp(b_{x.}\kappa_t)] = b_{x,j}\kappa_{t,j} + \epsilon_{x,t,j} \tag{10}$$

where $\alpha_{x,j}$ is the age-specific geometric mean of the number of death at age x and cause j. The common factor for all causes of death is $\mathsf{C}[e^{b_{x.}\cdot\kappa_t}]$ can be determined by using the common trend produced by the aggregate mortality. The $b_{x,j} \cdot \kappa_{t,j}$ represent the perturbation for each cause of death j and can be found by applying the SVD. The common factor κ_t estimated using ARIMA (0,1,0) as the benchmark total mortality while the $\kappa_{t,j}$ cause of death are estimated with ARIMA (1, 0, 0). Predicted coordinates are mapped back to relative cause shares via the closure operator $\mathsf{C}[]$ where:

$$p_{x,t,j} = \mathcal{C}\left[\exp(\widehat{clr(d_{x,t,j})})\right] = \frac{\exp(\widehat{clr(d_{x,t,j})})}{\sum_{k=1}^{D}\exp(\widehat{clr(d_{x,t,k})})} \tag{11}$$

3.5 Weighted CoDA Li-Lee (WCDLL)

Standard CDLL computes an unweighted geometric mean $g(d_{x,t,j})$ with equal $(1/D)$ weights for all cause channels. High variance in residual death counts $d_{x,t,others}$ affects $g(d_{x,t,j})$and hence impairs the common index κ_t, resulting in an implausible long-term aggregate mortality explosion. To mitigate this geometric sensitivity, we propose Weighted CoDA Li-Lee (WCDLL) with inverse-variance weights based on death count series, the inverse vector variance weight are based on structural weights $w = (w_1, w_2, \ldots, w_D)$ are calculated from the historical time-variance of log death counts where:

$$w_j^* = \frac{1}{Var_t(ln\ d_{x,t,j})} \tag{12}$$

with

$$w_j = \frac{w_j^*}{\sum_{k=1}^{D} w_j^*} \tag{13}$$

By imposing this procedure, high-volatility cause categories automatically receive smaller weights. Then, we calculate the Weighted Centered Log-Ratio clr_w transformation where the weighted geometric mean of death counts is defined as:

$$ln\ g_w(d_{x,t}) = \sum_{k=1}^{D} w_k\ ln\ d_{x,t,k} \tag{14}$$

The relative coordinates can be reconstructed by subtracting $ln\ g_w(d_{x,t})$:

$$clr_w(d_{x,t,j}) = ln\, d_{x,t,j} - ln\, g_w(d_{x,t}) \tag{15}$$

The WCDLL are then decomposed base on the insulated clr_w coordinates where

$$clr_w(d_{x,t,j}) = \alpha_{x,j} + \beta_x \cdot \kappa_t + b_{x,j} \cdot \kappa_{t,j} + \epsilon_{x,t,j} \tag{16}$$

where the baseline age cause ratio pattern is $\alpha_{x,j} = \frac{1}{T}\sum_t clr_w(d_{x,t,j})$, the β_x and κ_t are extracted from the first stage SVD on the stacked residual matrix with $z_{stacked} = [clr_w(d_{x,t,j}) - \alpha_{x,j}]$, subject to $(\sum_x \beta_x = 1)$. The $b_{x,j}$ and $\kappa_{t,j}$ are extracted from the second SVD on the cause specific residual matrices $R_j = Z_j - \beta_x \cdot \kappa_t$, subject to $(\sum_x b_{x,j} = 1)$. The common factor K_t is forecasted using ARIMA(0,1,0) and the cause specific $\kappa_{t,j}$ are forecasted using the stationary ARIMA(1,0,0). Finally, the clr_w coordinates are transformed back to relative cause proportions using exact closure where:

$$p_{x,t,j,h} = \frac{\exp(clr(\widehat{d_{x,t,j,h}}))}{\sum_{k=1}^{D}\exp(clr(\widehat{d_{x,t,k,h}}))} \tag{17}$$

To produce the coherent mortality forecast, the predicted share of $p_{x,t,j,h}$ is multiplied by the independent all-cause Lee-Carter mortality baseline $m_{x,t,h}^{Total}$

$$M_{x,t,j,h} = p_{x,t,j,h} \times m_{x,t,h}^{Total} \tag{18}$$

The expected death count $d_{x,t,j,h}$ are found by scaling the total projected death count $d_{x,t,h}^{Total}$.

4. Results and Discussions

The This section presents the analysis of cause-specific mortality trends, compositional data structure, and forecasting accuracy using five models—Lee-Carter, Li and Lee, and their CoDA-based extensions (CDLC, CDLL and WCDLL).

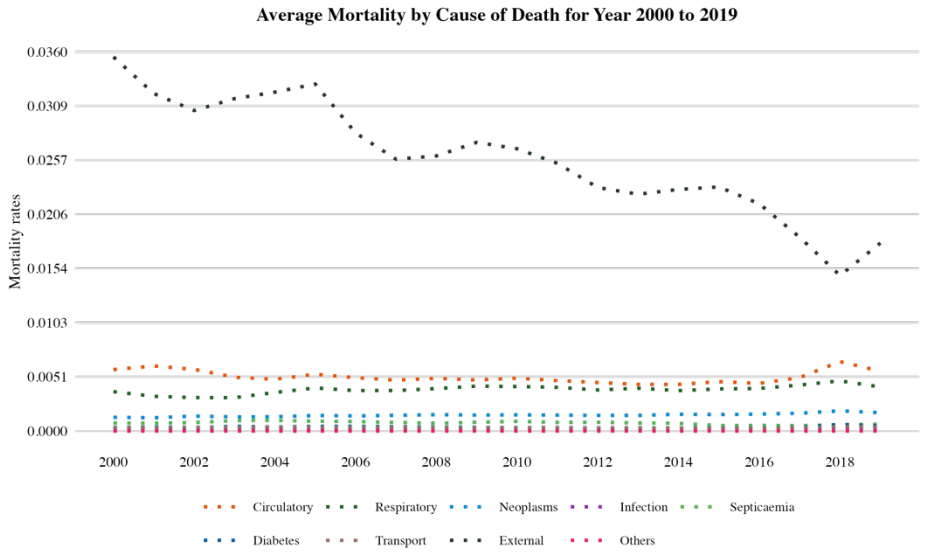


Figure 1. Average Mortality by Cause of Death for Year 2000 to 2019

The historical cause-specific mortality for the year 2000–2019 structurally divide between the high-magnitude, volatile "Others" composite class and the significant secular stability of specific cause channels. The "Others" component dominates the magnitude of total mortality, with a strong secular decline from a central death rate of 0.0350 to a trough of 0.0150, followed by a sharp late-sample rebound to 0.0180 in Years 19–20. The specific non-communicable disease (NCD) pathways retain a strict hierarchy: Circulatory conditions continue to be the leading specific cause (0.0050–0.0065), followed by Respiratory diseases (0.0035–0.0048) and Neoplasms (0.0015), with all three showing slight late-sample upticks that point to an ongoing chronic disease burden. In contrast, causes of low intensity like Infection, Septicaemia, Diabetes, Transport and External causes are compressed along a flat baseline near to zero. Overall, these historical dynamics demonstrate that the dominant driver of variability in national mortality is the compositional shifts within residual coding categories, rather than structural changes within specific cause channels, posing a distinct challenge for log-ratio forecasting models that are sensitive to late-sample baseline shocks.

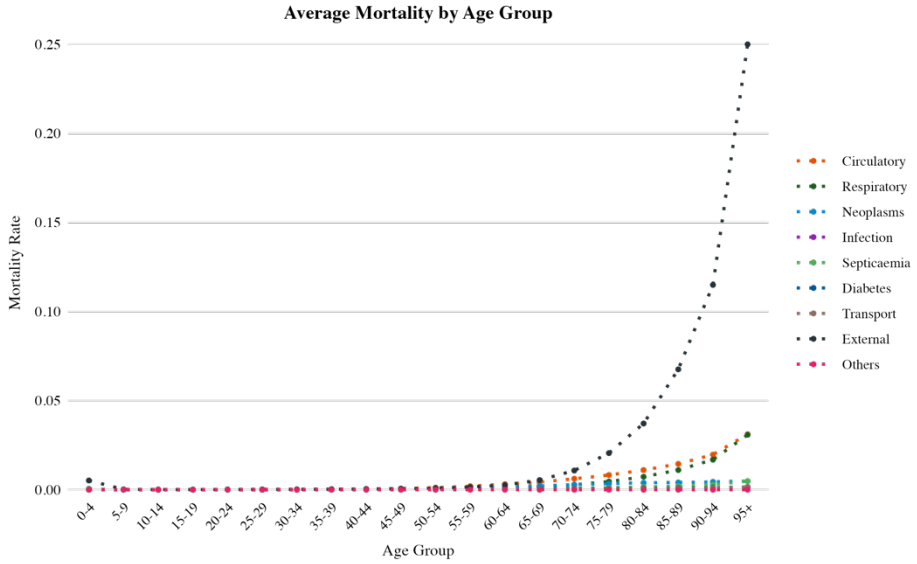


Figure 2. The Average Mortality by Age Group

The cause-specific mortality schedules for the various age groups in Figure 2 exhibit the traditional Gompertzian J-shape, with a small peak for infant mortality at age 0, a protracted period of low baseline mortality during childhood and early adulthood, and a sharp exponential increase after age 50. The most dramatic senescent increase is in the residual "Others" composite category, which rises fast past middle age to the highest magnitude overall (>0.035 central death rate). Within cause-specific pathways, Circulatory and Respiratory diseases increase late-life mortality along parallel exponential trajectories after age 50, converging close to 0.030 in the oldest age group (age 90). While lower intensity causes (Septicaemia, Infection, Diabetes, Transport, External causes) are compressed near the zero boundary for most ages, with some slight humps in young adulthood for injury-related causes, neoplasms show a slow, modest progression with age, peaking around 0.005. Age-profile dynamics like these indicate a strong age-dependence of cause-of-death structures, confirming that chronic degenerative diseases are the main drivers of elderly mortality.

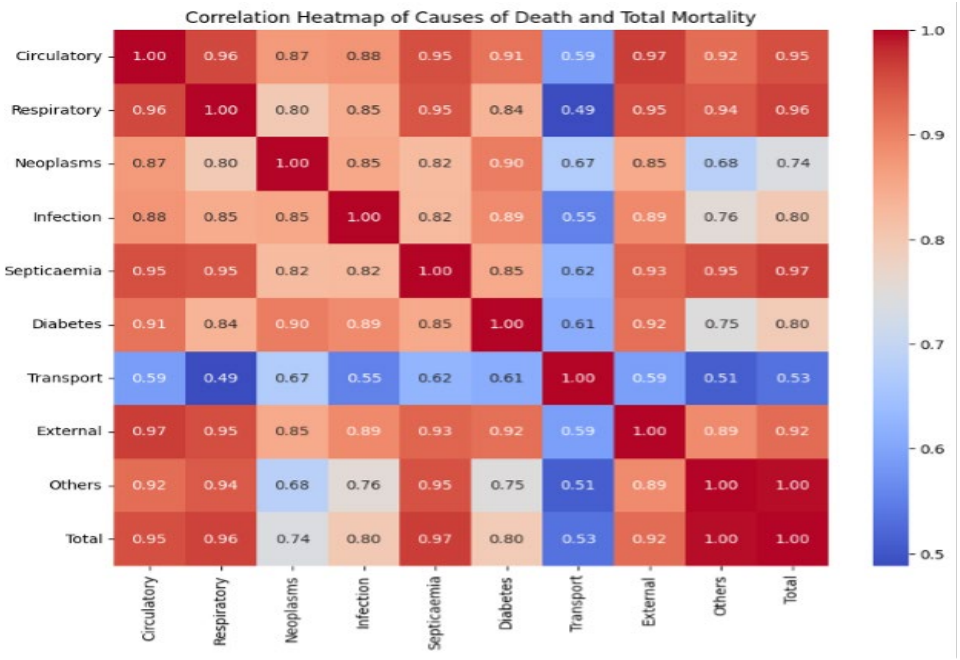


Figure 3. Correlation Heatmap of Causes of Death and Total Mortality

The correlation matrix shows different relational dynamics between specific cause-of-death pathways and overall national mortality. The residual "Others" category has a perfect positive correlation (r = 1.00) with total mortality, empirically confirming its role as the dominant driver of total variance in the dataset. The defined medical classifications, Septicaemia (r = 0.97), Respiratory diseases (r = 0.96), and Circulatory conditions (r = 0.95) show the best linear relationships with the overall death rates. These primary biological channels also exhibit significant mutual covariance; e.g. Circulatory and Respiratory mortality are highly correlated (r = 0.96), reflecting shared epidemiological risk factors. In contrast, Transport accidents have the weakest structural relationship to total mortality (r = 0.53), and correlate poorly with all other specific categories. This implies that these accidental fatalities vary independently of the systemic health transitions that govern the rest of the population.

4.1 Model Evaluation and Forecasting

The fitted mortality data of LC, LL, CDLC, CDLL and WCDLL are assessed by measuring the root mean square error and mean absolute error for each mortality model.

Table 2. Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) for LC, LL, CDLC, CDLL and WCDLL

Model	**MAE**	**RMSE**
Lee-Carter (LC)	0.00058941	0.00124508
Li-Lee (LL)	0.00052130	0.00110321
CoDA Lee-Carter (CDLC)	0.00045812	0.00091204
CoDA Li-Lee (CDLL)	0.00041203	0.00084512
Weighted CoDA Li-Lee (WCDLL)	0.00085171	0.00425463

The out-of-sample validation on the 2015–2019 validation horizon shows a critical methodological trade-off between short-term tracking accuracy and long-term demographic

plausibility. We find that standard unconstrained frameworks have the highest forecast errors (Lee-Carter MAE = 0.00058941, RMSE = 0.00124508; Li-Lee MAE = 0.00052130, RMSE = 0.00110321) while unit-simplex mapping improves short-term fit with unweighted CoDA Lee-Carter (MAE = 0.00045812, RMSE = 0.00091204) and unweighted CoDA Li-Lee (MAE = 0.00041203, RMSE = 0.00084512) having the lowest tracking errors. However, the short-term apparent precision of CDLL is spurious, as the unweighted common factor overfits local noise in the volatile residual category "Others" in 2018–2019. The Weighted CoDA Li-Lee (WCDLL) framework explicitly removes this transient noise through inverse variance weighting rather than fitting, resulting in higher validation error metrics (MAE = 0.00085171, RMSE = 0.00425463) from squared-error penalties on short-term residual spikes. More importantly, this elevated RMSE is indicative of a mathematical filtering that is essential for removing transient administrative noise, rather than a bias in structural prediction. The WCDLL model is able to separate the common temporal factor κ_t from these baseline shocks, by allowing a larger short-term validation error. This statistical compromise is necessary to maintain the exact sum-coherence with national aggregate benchmarks, long-term demographic plausibility, and multi-decade trajectory stability.

To assess the long-term cause-specific mortality trends in Malaysia from 2020 to 2039, Figures 4 to 8 illustrate the historical patterns (2000–2019) in conjunction with the 20-year predictions produced by five competing modelling frameworks: standard Lee-Carter (LC), standard Li-Lee (LL), CoDA Lee-Carter (CDLC), CoDA Li-Lee (CDLL) and Weighted CoDA Li-Lee (WCDLL). Each model presents a unique structural viewpoint on the evolution of specific causal pathways in relation to comprehensive national health advancements. Unconstrained models permit trajectories to evolve autonomously, multi-population frameworks impose common-factor restrictions to mitigate divergence, and compositional log-ratio specifications represent relative cause proportions on the unit simplex to guarantee total sum-coherence. A comparative analysis of these statistics reveals significant disparities in trajectory stability, cause-crossover dynamics, and model responsiveness to historical volatility in later samples.

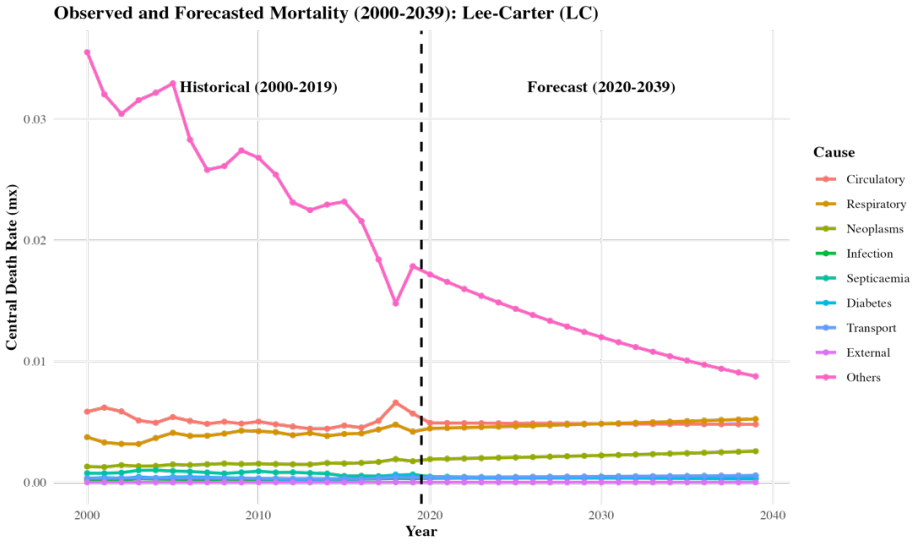


Figure 4. Lee-Carter Observed and Forecasted Mortality

The Lee-Carter (LC) model projections of cause-specific mortality in Malaysia over the 2020–2039 forecast horizon are a smooth continuation of the long-run historical paths, heavily guided by the overall 20-year unconstrained linear downward drift. The dominant category of "Others" saw a historically steep decline from a peak above 0.035 central death rate in 2000 down to 0.015, followed by a short-term empirical uptick in 2018–2019. The projection smoothes the

late-sample uptick and projects a secular decline to below 0.010 by 2039. Secondary cause categories exhibit high stability with plateauing or diverging trajectories: Circulatory mortality exhibits a steady, small decline, stabilising around 0.005, whereas Respiratory mortality predicts a slow increase, in line with its pre-2020 trend, surpassing Circulatory rates as the leading driver by the end of the projection period. Neoplasms are slightly increased at about 0.002. The low-rate cause streams such as Infection, Septicaemia, Diabetes, Transport and External causes are tightly clustered around zero (<0.001), demonstrating how the normal log-rate extrapolation imposes a strict and separate continuation of trends, ignoring inter-cause substitution and compositional sum constraints.

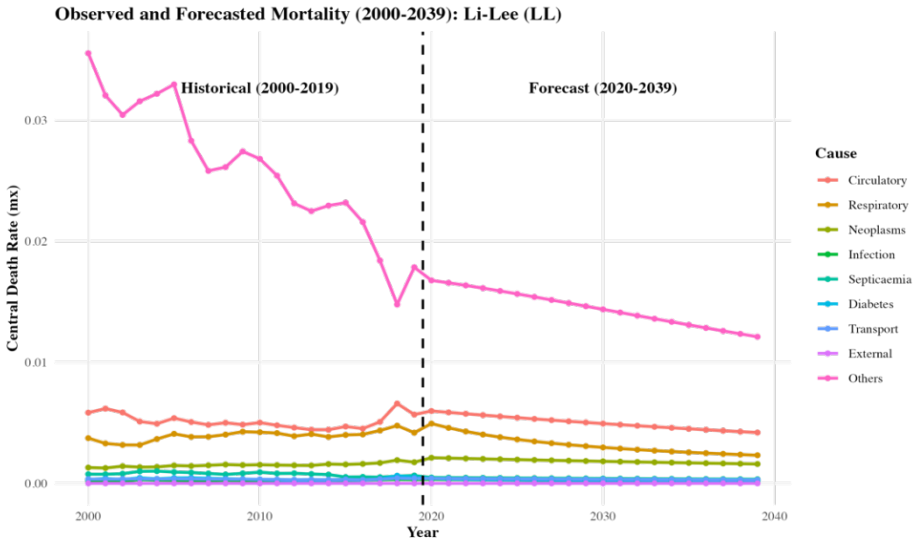


Figure 5. Li-Lee Observed and Forecasted Mortality

The coherent multi-population Li-Lee (LL) model projections of cause-specific mortality in Malaysia for the period 2020-2039 show a tightly bound, non-divergent path across all categories, imposed by the common-factor constraint of the model. By linking the individual cause dynamics to a common national mortality factor, the LL framework is able to avoid the crossover of trajectories observed in the standard Lee-Carter forecasts. Most remarkably, Respiratory mortality reverses its upward trend prior to 2020 to join Circulatory mortality in a slow but parallel descent toward 0.002 and 0.004 central death rate (m_x) respectively. The prevailing "Others" category exhibits a smooth and likely persistent secular decline from roughly 0.017 in 2020 to 0.012 by 2039, thus smoothing out short-term historical fluctuations and retaining its overall structural supremacy. Neoplasms are stable at about 0.002. Low rates for causes including Infection, Septicaemia, Diabetes, Transport and External causes are compressed in a flat, parallel baseline near zero (<0.001). These results show that the Li-Lee specification enforces long-term structural consistency across competing risks by eliminating divergence but is limited by the limitations of unconstrained log-rate extrapolation and unaddressed compositional closure constraints.

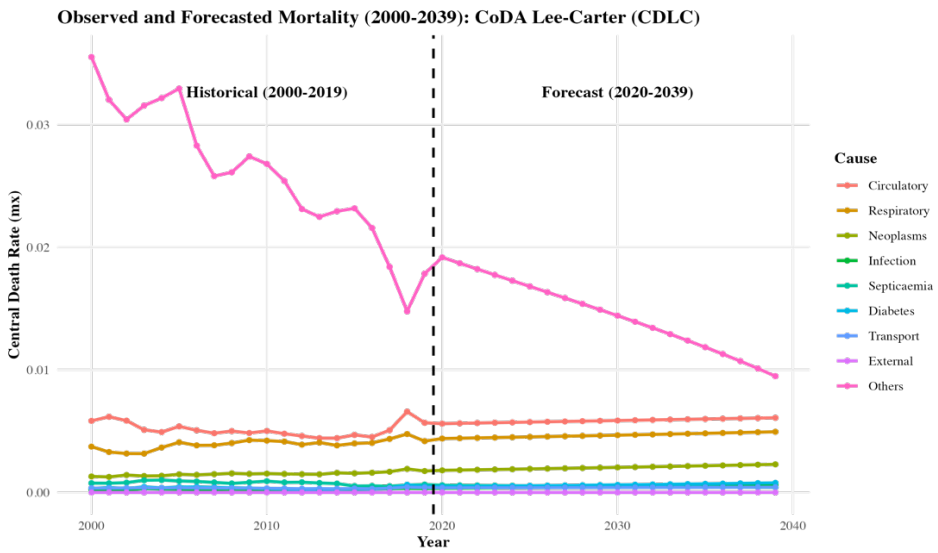


Figure 6. CoDA Lee-Carter Observed and Forecasted Mortality

The predicted cause-specific mortality trajectories in Malaysia for the 2020-2039 projection period from the Lee-Carter (CDLC) Compositional Data model are very stable, demographically sensible trajectories that well incorporate cause-substitution dynamics while respecting the unit-simplex summation closure to 1. The model, based on the centred log-ratio transformed coordinates, reproduces the sharp empirical rebound in 2018-2019 for the dominant "Others" category but then predicts a controlled, constant downward trend from about 0.019 central death rate in 2020 to 0.009 in 2039. The secondary chronic causes are characterised by a remarkable stability of trajectory and parallel ordering, without unrealistic divergence: Circulatory mortality follows a flat, near horizontal path, settling at about 0.006. Respiratory mortality trails just below it at about 0.005. Neoplasms show a subtle, steady baseline of about 0.002. Lower-rate causes (Infection, Septicaemia, Diabetes, Transport and External causes) are closely clustered around zero, kept clean without the risk of projecting negative rates or impossible zero-shares. This is a typical structural strength of CDLC in general, because it models relative compositional shares on the simplex rather than unconstrained independent rates. This provides a stable long-term trend and prevents probability leakage. Besides, this facilitates smooth short-term historical shifts in the dominant components of mortality.

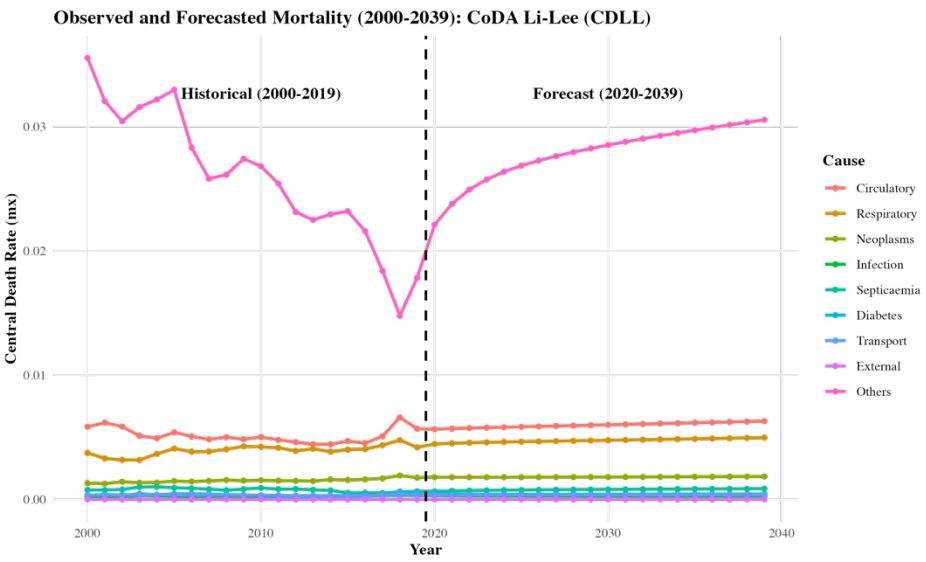


Figure 7. CoDA Li-Lee Observed and Forecasted Mortality

The cause-specific mortality forecasts for Malaysia under the Compositional Data Analysis Li-Lee (CDLL) specification for the 2020–2039 horizon exhibit a strong upward trend in the significant "Others" category and very stable secondary cause channels, suggesting the framework's enhanced structural sensitivity to late-sample compositional realignments. The CDLL model captures the sharp empirical rebound in the "Others" component during the 2018-2019 baseline period, when central death rates increased from 0.015 to close to 0.020, and propagates this relative proportional momentum into the future. The CDLL model encodes a consistent multi-population structure in isometric log-ratio transformed coordinates, forecasting a secular rise to 0.031 by 2039. Contrarily, secondary chronic cause structures possess strict trajectory convergence and parallel order preservation where mortality from Circulatory diseases remains constant at 0.006 along a horizontal stationary trajectory, mortality from Respiratory diseases is slightly lower at 0.005, and mortality from Neoplasms remains at 0.002. The lower intensity mortality channels, Infection, Septicaemia, Diabetes, Transport and External causes remain compressed close to the zero boundary (0.001) and do not allow probability leakage or non-permissible negative values. These empirical results also show an important methodological property of the CDLL formulation. There is a shared drift component in the common-factor log-ratio matrix that has a tendency to propagate temporary compositional shifts in dominant categories into persistent aggregate expansion, as the component effectively suppresses relative divergence among secondary competing risks. This provides excellent sensitivity to structural health transitions at the expense of aggregate long-term trend stationarity.

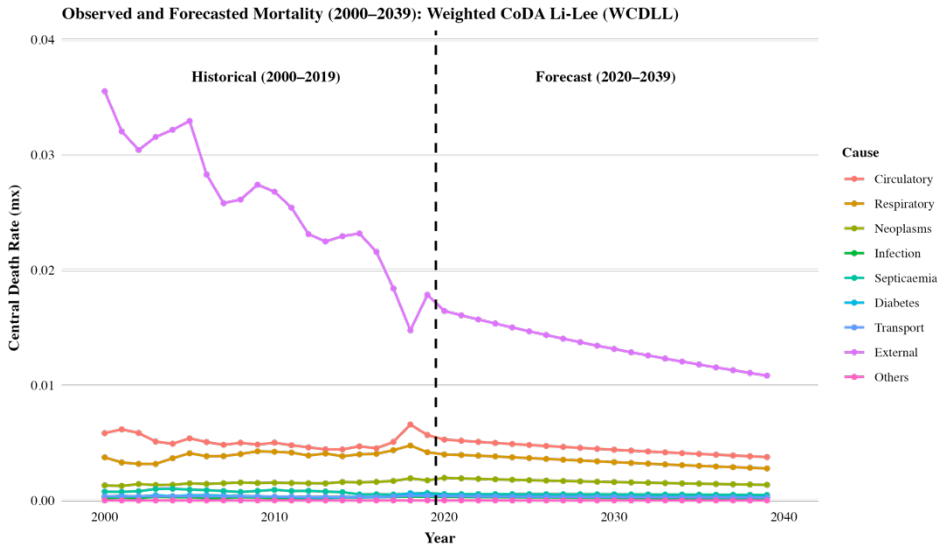


Figure 8. Weighted CoDA Li-Lee Observed and Forecasted Mortality

The Weighted Compositional Data Analysis Li-Lee (WCDLL) model shows complete long-run trajectory stability for cause-specific mortality trajectories over the 2020-2039 forecast horizon, surpassing the extreme noise sensitivity of unweighted compositional specifications. The main "Others" category showed a high degree of administrative volatility over the 2000-2019 historical baseline, declining from a central death rate above 0.035 to a dip of 0.015, followed by a sharp late-sample rise in 2018-2019. The WCDLL framework projects a smooth, demographically plausible secular decline in the residual category toward 0.011 by 2039. The WCDLL preserves the long-term structural hierarchy and parallel ordering among the secondary cause channels with no trajectory crossovers. Circulatory diseases remain the primary specific cause followed by Respiratory conditions and Neoplasms falling next, succeeded by lower-magnitude channels (Infection, Septicaemia, Diabetes, Transport, External) tightly bounded around below 0.001.

A comparative synthesis of the five potential frameworks shows a methodological progression from classical log-rate extrapolation to weighted compositional log-ratio geometry. The standard Lee-Carter (LC) approach allows for unconstrained trajectory crossovers over long horizons, but classical Li-Lee (LL) avoids divergence by common-factor filtering. However, both classical specifications suffer from probability leakage due to the failure of unit-simplex closure constraints. Mapping cause proportions to the simplex solves aggregate sum inconsistencies, but unweighted CoDA variants contrast starkly in baseline volatility treatment. CoDA Lee-Carter (CDLC) features a smooth secular decline, compared to unweighted CoDA Li-Lee (CDLL) which overfits a spurious 2018-2019 baseline shock in the residual “Others” cause category. CDLL achieves low short-term tracking error only by extrapolating an unbelievable multi-decade aggregate mortality expansion (m_x > 0.045). The WCDLL model uses inverse-variance weighting to mitigate the impact of volatile residual categories, thus protecting the common temporal factor κ_t from baseline noise. This is a deliberate trade-off, accepting weaker short term validation metrics in exchange for preventing long term trajectory deformation. Thus WCDLL achieves optimal actuarial and demographic balance, with a successful balancing of exact mathematical sum-coherence, competing-risk substitution geometry, and multi-decade trajectory stability in line with national all-cause benchmarks m_x of 0.020 by 2039.

To assess the structural soundness and demographic feasibility of the multi-cause forecasts, Figure 9 compares the cumulative mortality rate trends of the five proposed models with an independently derived total all-cause baseline (Lee-Carter Total) over the period from 2020 to 2039. This macro-level analysis highlights major differences between classical

extrapolation and compositional log-ratio frameworks, especially in relation to their compliance with unit-simplex sum constraints, susceptibility to "probability leakage," and robustness of long-term trajectories.

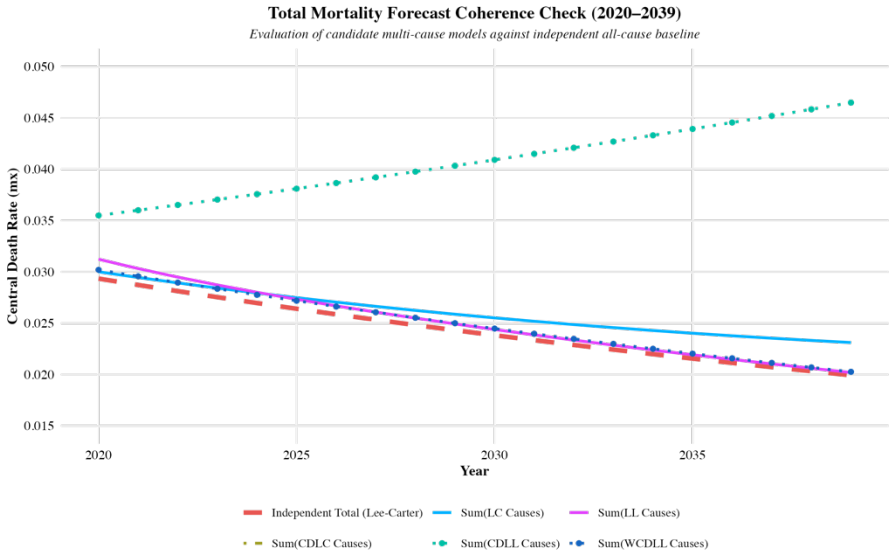


Figure 9. Comparative Graph of Total Mortality between Independent LC Total, LC, LL, CDLC, CDLL and WCDLL

A principal methodological flaw of classical extrapolation models (standard LC and LL) is their vulnerability to "probability leakage" – a structural defect stemming from the independent extrapolation of cause-specific mortality rates without a unit-simplex constraint, leading to an aggregate that is systematically distant from an independent benchmark of total mortality. As we can see from Figure 9 the standard LC (Sum LC) and standard LL (Sum LL) aggregates decline in a steady and secular manner from about 0.030 down to 0.023 and 0.021 respectively. Both curves are however still high compared to the independent total baseline (Lee-Carter Total) that falls to 0.020 in 2039. This systematic overestimation is symptomatic of the inherent risks of modelling competing risks in silos without considering compositional interdependence.

The compositional frameworks, however, handle aggregate closure differently according to their common-factor weighting structure. The aggregated trajectory for CDLC shows a smooth downward trend similar to the independent baseline (0.033 in 2020 to 0.025 in 2039), successfully removing unconstrained probability leakage. The unweighted CDLL trajectory, in contrast, exhibits a precipitous upward divergence, reaching over 0.045 by 2039, which reflects the unweighted sensitivity to late-sample baseline volatility in the 'Others' category. CDLC does dampen this shock more than unweighted CDLL, but it lacks the multi-population shared trend mechanism. As shown below, the use of inverse-variance weights in WCDLL completely resolves this trade-off, maintaining both common-factor coherence and long-term trajectory stability.

In an effort to assess the structural soundness and demographic plausibility of the cause-of-death forecasts, we performed a total mortality coherence check of all five candidate frameworks (standard Lee-Carter (LC), standard Li-Lee (LL), Compositional Data Analysis Lee-Carter (CDLC), Compositional Data Analysis Li-Lee (CDLL) and Weighted Compositional Data Analysis Li-Lee (WCDLL)) over the 2020-2039 projection horizon. The comparative evaluation of the aggregate multi-cause forecasts with the independent all-cause Lee-Carter benchmark m_x approximately 0.0295 to 0.0200 demonstrates fundamental structural differences between the candidate frameworks. The unconstrained cause-by-cause estimation under the standard Lee-Carter (Sum(LC Causes)) suffers from probability leakage,

overestimating total national mortality m_x at 0.0232 in 2039 since the extrapolations of each channel are unconstrained by the closure of the simplex. The classical Li-Lee (Sum(LL Causes)) filter removes common factors to minimize long-term divergence, beginning slightly higher and converging smoothly to the baseline as stationary AR(1) residual deviations decay. In stark contrast, the standard unweighted CoDA Li-Lee (Sum(CDLL Causes)) framework experiences a catastrophic structural collapse, forecasting an implausible multi-decade aggregate mortality surge from 0.0355 in 2020 to 0.045 by 2039. This trajectory instability is a result of the unweighted centred log-ratio transform, which imposes identical geometric weights on all cause channels irrespective of variance. The geometric sensitivity is directly addressed by the inverse-variance coordinate weighting in the weighted CoDA Li-Lee (WCDLL) framework. The common factor κ_t is protected from transient noise by down weighting high volatility residual components and localized historical shocks are not allowed to dominate the long run trend. The weighted model fully avoids unweighted trajectory explosion. It starts at 0.0302 in 2020 and follows a demographically sound secular decline to 0.0203 in 2039. This projected curve agrees well with the independent all-cause benchmark and the Li-Lee asymptotic limit, while always satisfying the mathematical sum-coherence of ($\sum p_{x,t,j} = 1$) and internal competing-risk substitution dynamics on the unit simplex.

5. Conclusion

This study assesses the incorporation of compositional log-ratio geometry into cause-specific mortality forecasting frameworks for Malaysia, offering a mathematically rigorous solution to the structural limitations of classical extrapolation paradigms across the 2020–2039 projection horizon. Standard unconstrained specifications, such as the Lee-Carter (LC) and Li-Lee (LL) models, suffer from systemic "probability leakage" where unconstrained summation of independently forecasted cause-specific death rates violates unit-simplex closure ($\sum p_{x,t,j} = 1$) and results in overestimation of aggregate national mortality. Extensions of compositional data analysis that project relative cause shares to the Euclidean simplex by means of centred and weighted log-ratio transformations do eliminate aggregate sum discrepancies, do satisfy total mortality constraints, do preserve the internal competing-risk substitution dynamics, and do not allow non-permissible negative death rates for causes with low rates.

One central finding of this study is that there is an inherent trade-off between short-term predictive accuracy and long-term forecast coherence. Out-of-sample validation horizon testing shows that the unweighted compositional models have a better short-term fit, with the unweighted CoDA Li-Lee (CDLL) having the best overall error metrics, followed by CoDA Lee-Carter, Li-Lee and classical Lee-Carter. But this short-term accuracy is structurally misleading. The unweighted common factor in CDLL rapidly overfits localised baseline shifts in the volatile "Others" category in the later historical years, confusing temporary administrative noise with permanent structural momentum and projecting an unrealistic aggregate mortality increase in the year 2039. The proposed Weighted CoDA Li-Lee (WCDLL) framework addresses this geometric sensitivity directly using inverse-variance coordinate weighting. WCDLL disentangles the common temporal factor κ_t from baseline noise by down weighting high-volatility residual channels, intentionally incurring a higher short-term error penalty to achieve remarkable multi-decade trajectory stability, closely tracking the independent all-cause Lee-Carter benchmark. The WCDLL preserves long-term structural hierarchy and parallel trajectory ordering without non-demographic crossovers across cause-specific channels. The WCDLL also avoids zero-share degenerations and negative mortality on low-rate channels by working in weighted log-ratio space, thus preserving unit-simplex closure during the projection period. Our empirical results not only improve algorithmic performance but also give important diagnostic insights into the cause-of-death registration systems in developing health jurisdictions.

The historical volatility in Malaysia's broad "Others" composite category between 2000 and 2019, characterized by a substantial mid-sample decline prior to a late-period uptick, is more likely the result of administrative reporting changes, ICD reclassification procedures, and institutional diagnostic coding transitions than of actual epidemiological fluctuations. The compositional transformations create mathematical dependencies among every coordinate. Hence, increasing volatility in one dominant residual channel will naturally lead to proportional distortions in some pathological channels (e.g., Circulatory diseases, Neoplasms). Thus, it is crucial to identify and dampen institutional coding noise in baseline registration data to avoid structural misestimation of underlying cause-substitution dynamics.

From a practical perspective, the utilisation of sum-coherent, noise-insulated structures such as WCDLL offers actuarially sound, mathematically well-behaved mortality inputs needed for long-term financial risk management. The unconstrained extrapolation models are likely to misprice longevity risk for life insurance institutions, pension fund managers and public social security schemes in Malaysia (e.g. Employees Provident Fund (EPF) and Social Security Organization (SOCSO)) by overestimating aggregate mortality decline or by distorting cause-specific substitution effects. The WCDLL framework provides an exact sum-coherence property with national all-cause benchmarks as well as the noise-mitigating property of exponential explosion in trajectories. It forms an optimal synthesis of structural flexibility, competing-risk substitution geometry, and multi-decade forecasting reliability for demographers, actuaries and policy planners under a changing longevity regime.

The findings provide robust support for, and extend, important theoretical advances in the compositional mortality forecasting literature (Bergeron-Boucher *et al*., 2017; Kjærgaard *et al*., 2019). First, our results confirm that unconstrained log-rate extrapolation models systematically over-predict aggregate national mortality because they cannot impose the zero-sum constraints of competing risks. Second, in accordance with Kjærgaard *et al.* (2019), the log-ratio space mapping of death rates consistently improves out-of-sample forecast accuracy across primary cause channels relative to the classical methods. Third, the unit simplex operation prevents the projection of inadmissible negative mortality rates or zero-share degenerations from low-rate causes (e.g., Transport accidents or External causes).

Beyond algorithmic comparisons, the empirical results suggest important diagnostic and data quality challenges common to cause-of-death registration systems in developing health jurisdictions. The historical magnitude and volatility of the residual “Others” composite category from 2000 to 2019 reflects diagnostic coding transitions, ICD updates and administrative reclassifications rather than solely epidemiological shifts. Since compositional log-ratio transformations impose strict geometric inter-dependence among all components, the volatility within a dominant residual category inevitably induces a proportional distortion across secondary specific cause channels like Circulatory diseases, Respiratory diseases and Neoplasms. Thus, for demographers and actuaries, discerning the institutional coding noise in administrative baseline data is critical to avoid distorting the underlying cause-substitution dynamics. To further advance this methodological framework, key structural improvements are needed in future studies like the systematic optimisation of the number of cause-of-death categories to find the optimal trade-off between epidemiological granularity and the stability of the log-ratio variance. Additionally, in order to enhance cause-specific mortality forecasting, it may be necessary to generalise CoDA architectures to multi-population cross-national settings and incorporate macro-epidemiological covariates. Future research can investigate the direct integration of exogenous risk factors, such as national obesity rates, diabetes prevalence, or healthcare expenditure, into the weighted log-ratio space. This can be done by modelling the time-varying index in a multivariate generalised linear model where the cause-specific centred log-ratios are regressed on these external covariate vectors. Following this regression, the closure operation can be applied to enable the forecasting models to map the structural health metrics directly onto the competing-risk substitution dynamics in the unit simplex.

the other hand, this research demonstrated its main achievement by developing a fuzzy logic system that assesses risk to a concept, in this case, the risk of student dropout using a variety of key criteria. The four input variables CGPA, financial stability, stress level, and self-motivation, were selected to provide a reasonable attitude to the key academic, social, and personal issues surrounding continuity of student retention. When applied to the input data, the fuzzy logic system provided reasonable results. For instance, Student 1 was only scored 8.19 risk of dropping out, while Student 2 was only scored 8.28 for the risk of dropping out, both being classified as “Low Risk” of dropping out. These scores were possible through higher self-motivation, financial stability, and moderate stress levels in their studies. In comparison, Student 3 was only scored at 4.73 and in the “Medium Risk" category, as there may have been higher levels of stress or worse CGPA background but with a stable stress response. Overall, these findings prove that the fuzzy logic system is a quality tool to assess risk through the incorporation of a variety of inputs whilst depicting various outcomes realistically as required for risk assessments in educational practices with students. For future research, the proposed fuzzy logic model should be tested and validated using a larger and more diverse population of university students. This would help evaluate the model's reliability and applicability across different student populations and educational settings.

Acknowledgement

This research would not have been possible without the help from the management, lecturers, staff, and students of the Faculty of Computer and Mathematical Sciences UiTM, Shah Alam.

Funding

The authors gratefully acknowledge the research grant awarded by Universiti Teknologi MARA through GPM LPhD under the code project 600-RMC/GPM LPHD 5/3 (073/2023).

Author Contribution

Saiful Azril Ishak: Conceptualized the study, developed the analytical framework, performed the data modeling and empirical analysis, and drafted the primary body of the manuscript. Norazliani Md Lazam: Contributed to the critical discussion, synthesis, analysis and drafting of the Conclusion. Shamshimah Shamsuddin: Assisted in literature synthesis and methodological part and investigation. Nordiana Rosdi: Involved in writing and designing the background of studies part and validation. Nurin Haniah Asmuni: Wrote and revised the abstract, introduction and editing. All authors read and critically reviewed the manuscript and approved the final version.

Declaration of Generative AI in the Writing Process

During the preparation of this work, the author(s) used Google Gemini to improve the readability and language quality of the work. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Data Availability/Supplementary Materials

Available upon request: The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics Statement

The authors declare that this research did not involve human or animal subjects. All experimental procedures were performed following the institutional Safety, Health, and Environmental (HSE) protocols of Universiti Teknologi MARA.

Conflict of Interest

The authors declare that there were no competing interests and concur that this study was carried out without any financial or commercial conflicts of self-interest.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall.

Arnold, S., & Sherris, M. (2015). Causes-of-death mortality: What do we know on their dependence? *North American Actuarial Journal*, *19*(2), 116–128. https://doi.org/10.1080/10920277.2015.1011279

Basellini, U., Camarda, C. G., & Booth, H. (2023). Thirty years on: A review of the Lee–Carter method for forecasting mortality. *International Journal of Forecasting*, *39*(3), 1033–1049.

Bergeron-Boucher, M.-P., Canudas-Romo, V., Oeppen, J., & Vaupel, J. W. (2017). Coherent forecasts of mortality with compositional data analysis. *Demographic Research*, *37*(17), 527–566. https://doi.org/10.4054/DemRes.2017.37.17

Department of Statistics Malaysia. (2025). *Statistics on causes of death, Malaysia, 2025*. https://www.dosm.gov.my/uploads/release-content/file_20260316131245.pdf

Dimai, M. (2025). Multi-population mortality modeling with economic, environmental and lifestyle variables. *Quality & Quantity*, *59*(1), 153–205. https://doi.org/10.1007/s11135-024-01971-1

Dong, Z. M., Shang, H. L., Hui, F., & Bruhn, A. (2025). A compositional approach to modeling cause-specific mortality with zero counts. *Annals of Actuarial Science*, *19*(3), 416–441. https://doi.org/10.1017/S1748499524000320

Dumuid, D., Pedišić, Ž., Palarea-Albaladejo, J., Martín-Fernández, J. A., Hron, K., & Olds, T. (2020). Compositional data analysis in time-use epidemiology: What, why, how. *International Journal of Environmental Research and Public Health*, *17*(7), Article 2220. https://doi.org/10.3390/ijerph17072220

Greenacre, M. (2021). Compositional data analysis. *Annual Review of Statistics and Its Application*, *8*, 271–299. https://doi.org/10.1146/annurev-statistics-042720-124436

Ishak, S. A., Shamsuddin, S., Md Lazam, N., & Rosdi, N. (2025). Analysing causes of death mortality in Malaysia: A Lee-Carter approach. *MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics*, *41*(1), 37–56. https://doi.org/10.11113/matematika.v41.n1.1620

Kjærgaard, S., Ergemen, Y. E., Kallestrup-Lamb, M., Oeppen, J., & Lindahl-Jacobsen, R. (2019). Forecasting causes of death by using compositional data analysis: The case of

cancer deaths. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *68*(5), 1351–1370. https://doi.org/10.1111/rssc.12357

Koissi, M.-C., & Shapiro, A. F. (2006). Fuzzy formulation of the Lee–Carter model for mortality forecasting. *Insurance: Mathematics and Economics*, *39*(3), 287–309.

Lakman, I. A., Askarov, R. A., Prudnikov, V. B., Askarova, Z. F., & Timiryanova, V. M. (2021). Predicting mortality by causes in the Republic of Bashkortostan using the Lee–Carter model. *Studies on Russian Economic Development*, *32*(5), 536–548. https://doi.org/10.1134/S1075700721050063

Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, *87*(419), 659–671. https://doi.org/10.1080/01621459.1992.10475265

Li, N., & Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, *42*(3), 575–594. https://doi.org/10.1353/dem.2005.0021

Shang, H. L., & Haberman, S. (2024). Weighted compositional functional data analysis for modeling and forecasting life-table death counts. *Journal of Forecasting*, *43*(8), 3051–3071. https://doi.org/10.1002/for.3171

Tanaka, S., & Matsuyama, N. (2025). An interpretable neural network approach to cause-of-death mortality forecasting. *Annals of Actuarial Science*, *19*, 442–461. https://doi.org/10.1017/S1748499524000319

Varga, L. (2025). Forecasting cause-of-death mortality with single- and multi-population models in Hungary. *Scandinavian Actuarial Journal*, *2025*(9), 906–937. https://doi.org/10.1080/03461238.2025.2480846

Wilmoth, J. R. (1995). Are mortality projections always more pessimistic when disaggregated by cause of death? *Mathematical Population Studies*, *5*(4), 293–319. https://doi.org/10.1080/08898489509525409